

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

“JnanaSangama”, Belgaum -590014, Karnataka.



LAB REPORT on

Big Data Analytics (22CS6PEBDA)

Submitted by

BHUVANA M (1BM22CS071)

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-2024 to July-2024

**B. M. S. College of Engineering,
Bull Temple Road, Bangalore 560019**
(Affiliated To Visvesvaraya Technological University, Belgaum)
Department of Computer Science and Engineering



CERTIFICATE

This is to certify that the Lab work entitled "**Big Data Analytics**" carried out by **Bhuvana M (1BM22CS071)**, who is bonafide student of **B. M. S. College of Engineering**. It is in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visvesvaraya Technological University, Belgaum during the year 2024. The Lab report has been approved as it satisfies the academic requirements in respect of a **Big Data Analytics Lab (22CS6PEBDA)** work prescribed for the said degree.

Srushti C S
Assistant Professor
Department of CSE
BMSCE, Bengaluru

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

Index Sheet

Sl. No.	Experiment Title	Page No.
1	MongoDB- CRUD Demonstration	1
2	<p>Perform the following DB operations using Cassandra.</p> <ol style="list-style-type: none"> 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. 	11
3	<p>Perform the following DB operations using Cassandra.</p> <ol style="list-style-type: none"> 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 	12
4	Execution of HDFS Commands for interaction with Hadoop Environment. (Minimum 10 commands to be executed)	14
5	<p>From the following link extract the weather data https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all</p> <p>Create a Map Reduce program to</p> <ol style="list-style-type: none"> a) find average temperature for each year from NCDC data set. b) find the mean max temperature for every month 	15

6	Implement Wordcount program on Hadoop framework	22
7	For a given Text file, Create a Map Reduce program to sort the content in an alphabetic order listing only top 10 maximum occurrences of words.	26
8	Write a Scala program to print numbers from 1 to 100 using for loop.	35
9	Using RDD and FlatMap count how many times each word appears in a file and write out a list of words whose count is strictly greater than 4 using Spark	36

Course Outcome

CO1	Apply the concept of NoSQL, Hadoop, 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.

Github Link : <https://github.com/Bhu06/BDA-LAB.git>

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, Con
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: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, Con
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, Con
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, Con
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, Co
nt: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'})
[{"_id": 5, "StudName": "Vamsi", "Grade": "VI", "Hobbies": "Skating", "Age": 11, "RollNo": 11, "Email": "rea.de9@gmail.com", "Cont": 2276, "Status": "Active"}, {"_id": 6, "StudName": "Aryan David", "Grade": "VII", "Hobbies": "Skating", "Age": 12, "RollNo": 12, "Email": "rea.de9@gmail.com", "Cont": 2277, "Status": "Active"}]

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}, {$set:{Location:'Network'}})
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
[{"_id": "4", "Name": "Abhinav", "RollNo": 10, "Email": "abhinav@gmail.com", "Address": "Kharar", "Phone": 9876543210, "Location": "Network"}]
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.update({_id:4}, {$unset:{Location:'Network'}})
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
[{"_id": "4", "Name": "Abhinav", "RollNo": 10, "Email": "abhinav@gmail.com", "Address": "Kharar", "Phone": 9876543210, "Location": null}]

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 2: Cassandra

Question: Perform the following DB operations using Cassandra.

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

Code with Output:

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

cqlsh> ALTER TABLE Employee.Employee_Info ADD Projects SET<TEXT>;
cqlsh> UPDATE Employee.Employee_Info SET Projects = ['Project A', 'Project B'] WHERE Emp_Id = 121 AND Salary = 60000;
cqlsh> INSERT INTO Employee.Employee_Info (Emp_Id, Salary, Emp_Name, Designation, Date_of_Joining, Dept_Name) VALUES (124, 30000, 'Temp Employee', 'Intern', '2023-10-01', 'Temp Dept') USING TTL 15;
cqlsh> SELECT * FROM Employee.Employee_Info;
emp_id | salary | date_of_joining | dept_name | designation | emp_name | projects
-----+-----+-----+-----+-----+-----+-----+
  123 | 55000 | 2021-11-10 | Finance | Analyst | Alice Johnson | null
  122 | 80000 | 2022-05-20 | HR | Manager | Jane Smith | null
  121 | 60000 | 2023-01-15 | Engineering | Developer | Johnathan Doe | {'Project A', 'Project B'}
(3 rows)
cqlsh> 
```

Lab 3: Cassandra

Question: Perform the following DB operations using Cassandra.

1. Create a keyspace by name Library
2. Create a column family by name Library-Info with attributes Stud_Id Primary Key, Counter_value of type Counter, Stud_Name, Book-Name, Book-Id, Date_of_issue
3. Insert the values into the table in batch
4. Display the details of the table created and increase the value of the counter
5. Write a query to show that a student with id 112 has taken a book “BDA” 2 times.
6. Export the created column to a csv file
7. Import a given csv dataset from local file system into Cassandra column family

Code with Output:

```

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

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

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

  count
-----
  2

(1 rows)

```

```

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

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

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

```

Lab 4: Cassandra

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

Code with Output:

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~ cd ./Desktop/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
Starting resourcemanager
Starting nodemanagers
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -mkdir /Lab05
```

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

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

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

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

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

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

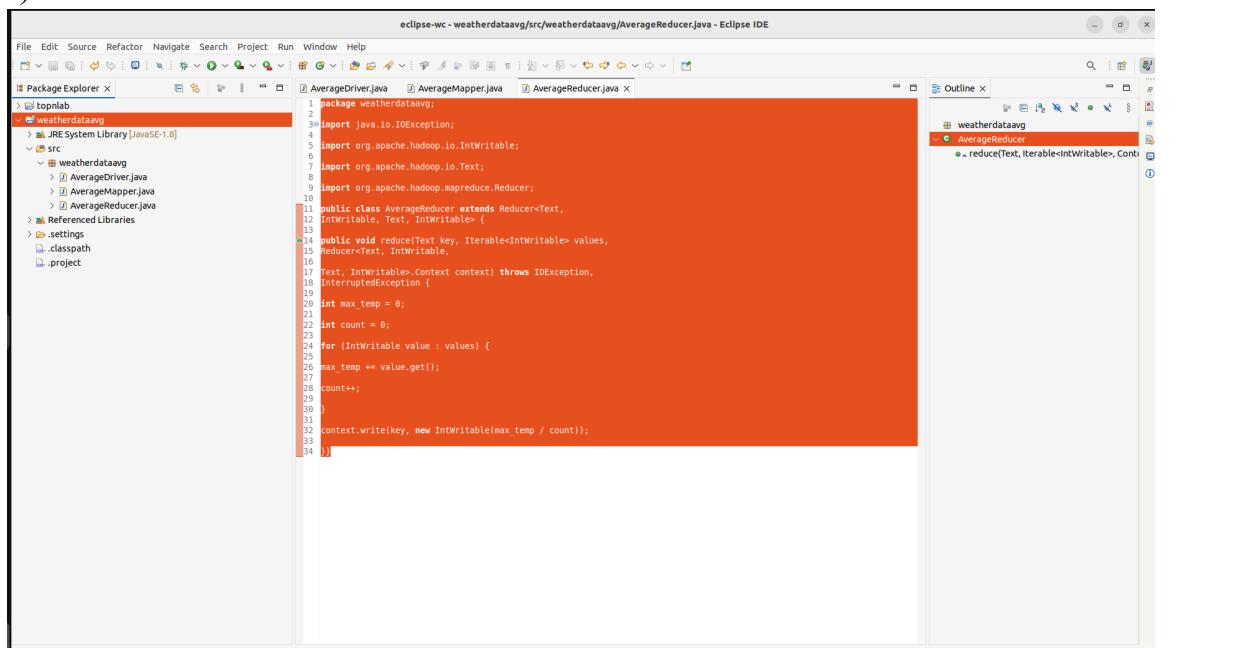
Lab 5: Hadoop

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

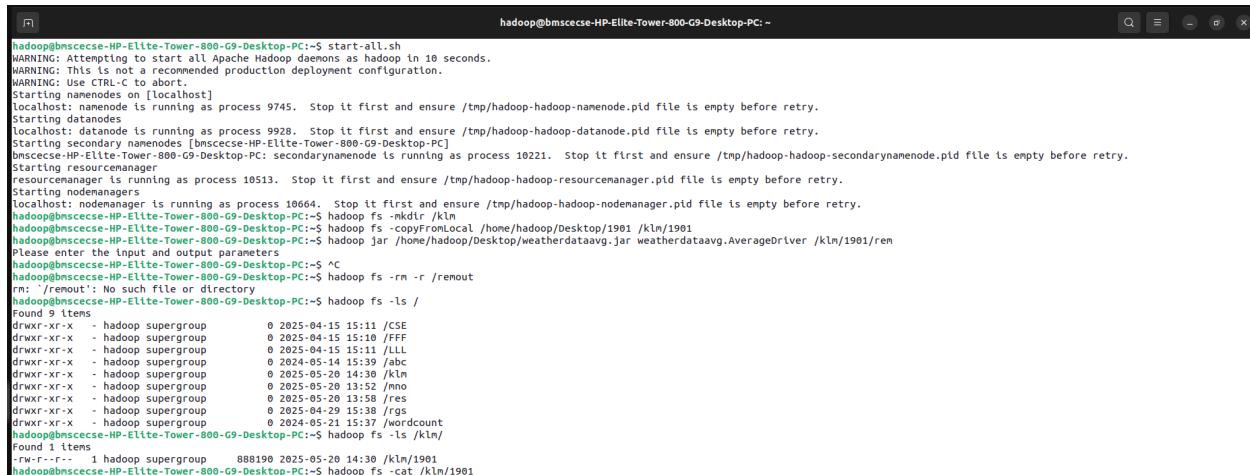
- find average temperature for each year from NCDC data set.
- find the mean max temperature for every month

Code with Output:

a)



```
File Edit Source Refactor Navigate Search Project Run Window Help
eclipse-wc - weatherdataavg/src/weatherdataavg/AverageReducer.java - Eclipse IDE
Package Explorer X
src
  AverageDriver.java
  AverageMapper.java
  AverageReducer.java
weatherdataavg
  AverageDriver.java
  AverageMapper.java
  AverageReducer.java
Referenced Libraries
settings
  classpath
  project
Outline X
weatherdataavg
  AverageReducer
    reduce(Text, Iterable<IntWritable>, Context)
AverageReducer.java
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 public class AverageReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
10
11     public void reduce(Text key, Iterable<IntWritable> values,
12                        Text, IntWritable>.Context context) throws IOException,
13                        InterruptedException {
14         int max_temp = 0;
15         int count = 0;
16         for (IntWritable value : values) {
17             max_temp += value.get();
18         }
19         count++;
20         context.write(key, new IntWritable(max_temp / count));
21     }
22 }
23
24 }
```

```
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenode on [localhost]
localhost: namenode is running as process 9745. Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 9928. Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10221. Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
resourcemanager is running as process 10513. Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10664. Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /klm
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/1901 /klm/1901
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/weatherdataavg.jar weatherdataavg.AverageDriver /klm/1901/rem
Please enter the input and output parameters
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -rm -r /remout
rm: '/remout': No such file or directory
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /
Found 9 items
drwxr-xr-x  - hadoop supergroup          0 2025-04-15 15:11 /CSE
drwxr-xr-x  - hadoop supergroup          0 2025-04-15 15:10 /FFF
drwxr-xr-x  - hadoop supergroup          0 2025-04-15 15:11 /JBL
drwxr-xr-x  - hadoop supergroup          0 2025-05-20 14:39 /abc
drwxr-xr-x  - hadoop supergroup          0 2025-05-20 14:39 /klm
drwxr-xr-x  - hadoop supergroup          0 2025-05-20 13:52 /mno
drwxr-xr-x  - hadoop supergroup          0 2025-05-20 13:58 /res
drwxr-xr-x  - hadoop supergroup          0 2025-04-29 15:38 /rgs
drwxr-xr-x  - hadoop supergroup          0 2024-05-21 15:37 /wordcount
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /klm/
Found 1 items
-rw-r--r--  1 hadoop supergroup  888190 2025-05-20 14:30 /klm/1901
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /klm/1901
```

```

hadoop@bnsccse-HP-Elite-Tower-800-09-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/weatherdataavg.jar weatherdataavg.AverageDriver /klm/1901 /rem
2025-05-20 14:34:21.074 INFO impl.MetricConfig: Loaded properties from hadoop-metrics2.properties
2025-05-20 14:34:21.116 INFO impl.MetricSystemImpl: 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 14:34:21.233 INFO input.FileInputFormat: Total input files to process : 1
2025-05-20 14:34:21.262 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-20 14:34:21.332 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local261815800_0001
2025-05-20 14:34:21.370 INFO mapreduce.Job: Job tracking url: http://localhost:8080/jobs/1
2025-05-20 14:34:21.408 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-20 14:34:21.409 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 14:34:21.233 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21.405 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:34:21.406 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:34:21.407 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _attempt folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:34:21.446 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-20 14:34:21.447 INFO mapred.LocalJobRunner: Starting task: attempt_local261815800_0001_m_000000_0
2025-05-20 14:34:21.459 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21.460 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:34:21.460 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:34:21.465 INFO mapred.Tasks: Using ResourceCalculatorProcessTree : []
2025-05-20 14:34:21.501 INFO mapred.Task: attempt_local261815800_0001_m_000000_0
2025-05-20 14:34:21.511 INFO mapred.MapTask: (REDATOR) kv=26188144(104857584)
2025-05-20 14:34:21.511 INFO mapred.MapTask: mapred.task.io.sort.mbo_100
2025-05-20 14:34:21.511 INFO mapred.MapTask: soft limit at 83884080
2025-05-20 14:34:21.511 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-20 14:34:21.511 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-20 14:34:21.511 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-20 14:34:21.584 INFO mapred.LocalJobRunner: map task attempt_local261815800_0001_m_000000_0
2025-05-20 14:34:21.587 INFO mapred.Task: 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 14:34:21.600 INFO mapred.Task: Task attempt_local261815800_0001_m_000000_0 is done. And is in the process of committing
2025-05-20 14:34:21.602 INFO mapred.LocalJobRunner: map
2025-05-20 14:34:21.605 INFO mapred.Task: Task attempt_local261815800_0001_m_000000_0 is done. And is in the process of committing
2025-05-20 14:34:21.605 INFO mapred.Task: Final Counters for attempt_local261815800_0001_m_000000_0: Counters: 23
  File System Counters
    FILE: Number of bytes read=4430
    FILE: Number of bytes written=713998
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=8808198
    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
    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=0
    HDFS: Number of bytes read erasure-coded=0
  Map-Reduce Framework
    Combine input records=0
    Combine output records=0
    Reduce input groups=1
    Reduce shuffle bytes=2220
    Reduce input records=6564
    Reduce output records=1
    Spilled Records=6564
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=0
    Total committed heap usage (bytes)=633339904
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
  File Output Format Counters
    Bytes Written=8
2025-05-20 14:34:21.803 INFO mapred.LocalJobRunner: Finishing task: attempt_local261815800_0001_r_000000_0
2025-05-20 14:34:21.803 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-20 14:34:22.404 INFO mapred.Task: Job job_local261815800_0001 running in uber mode : false
2025-05-20 14:34:22.405 INFO mapred.Job: map 100% reduce 100%
2025-05-20 14:34:22.405 INFO mapred.Task: Job job_local261815800_0001 completed successfully
2025-05-20 14:34:22.414 INFO mapred.Job: Counters: 36
  File System Counters
    FILE: Number of bytes read=153312
    FILE: Number of bytes written=1500266
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=1776380
    HDFS: Number of bytes written=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 materialized bytes=72210
Input split bytes=95
Combiner Input records=0
Spilled Records=6564
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=0
Total committed heap usage (bytes)=633339904
File Input Format Counters
  Bytes Read=888190
2025-05-20 14:34:21,685 INFO mapred.LocalJobRunner: Flushing task: attempt_local261815800_0001_m_000000
2025-05-20 14:34:21,687 INFO mapred.LocalJobRunner: map task is complete.
2025-05-20 14:34:21,687 INFO mapred.LocalJobRunner: Waiting for reduce tasks.
2025-05-20 14:34:21,688 INFO mapred.LocalJobRunner: Starting task: attempt_local261815800_0001_r_000000
2025-05-20 14:34:21,698 INFO output.PathOutputCommitterFactory: No output Committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21,612 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:34:21,612 INFO output.FileOutputCommitter: skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:34:21,612 INFO mapred.Tasks: Using ResourceCalculatorProcessTree: []
2025-05-20 14:34:21,614 WARN impl.MetricsSystemImpl: Used org.apache.hadoop.mapreduce.task.reduce.Shuffle@cd4d107
2025-05-20 14:34:21,622 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5829453312, maxSinglesShuffleLimit=1457363328, mergeThreshold=3847439360, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-20 14:34:21,622 INFO reduce.MergeManagerImpl: attempt_local261815800_0001_r_000000 Thread started: EventFetcher for fetching Map Completion Events
2025-05-20 14:34:21,635 INFO reduce.LocalFetcher: attempt_local261815800_0001 about to shuffle output of map attempt_local261815800_0001_m_000000_0
2025-05-20 14:34:21,637 INFO reduce.InMemoryMapOutput: Read 72206 bytes from map-output for attempt_local261815800_0001_m_000000_0 decomp: 72206 len: 72210 to MEMORY
2025-05-20 14:34:21,637 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 72206, InMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory -> 72206
2025-05-20 14:34:21,638 INFO reduce.MergeManagerImpl: attempt_local261815800_0001_r_000000_0 decomposed
2025-05-20 14:34:21,638 INFO mapred.LocalJobRunner: EventFetcher is interrupted.. Returning
2025-05-20 14:34:21,638 INFO reduce.MergeManagerImpl: findMerge 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 14:34:21,645 INFO reduce.MergeManagerImpl: Merged 1 segments, 72206 bytes to disk to satisfy reduce memory limit
2025-05-20 14:34:21,645 INFO reduce.MergeManagerImpl: Merging 1 files, 72199 bytes from disk
2025-05-20 14:34:21,645 INFO reduce.MergeManagerImpl: Merged 1 segments, 0 bytes from memory into reduce
2025-05-20 14:34:21,645 INFO reduce.MergeManagerImpl: Down to the last merge-pass, with 1 segments left of total size: 72199 bytes
2025-05-20 14:34:21,645 INFO mapred.Merger: 1 / 1 copied.
2025-05-20 14:34:21,679 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2025-05-20 14:34:21,707 INFO mapred.Task: Task:attempt_local261815800_0001_r_000000_0 is done. And is in the process of committing
2025-05-20 14:34:21,771 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:34:21,771 INFO mapred.Task: Task:attempt_local261815800_0001_r_000000_0 is allowed to commit now
2025-05-20 14:34:21,800 INFO mapred.Task: Saved report of task 'attempt_local261815800_0001_r_000000_0' to hdfs://localhost:9000/re
2025-05-20 14:34:21,802 INFO mapred.Task: reduce done
2025-05-20 14:34:21,803 INFO mapred.Task: Task 'attempt_local261815800_0001_r_000000_0' done.
2025-05-20 14:34:21,803 INFO mapred.Task: Final Counters for attempt_local261815800_0001_r_000000_0: Counters: 30
  File System Counters
    FILE: Number of bytes read=148882
    FILE: Number of bytes written=786208
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=888190
    HDFS: Number of bytes written=8
    HDFS: Number of read operations=10

    Reduce shuffle bytes=72210
    Reduce input records=6564
    Reduce output records=0
    Spilled Records=13128
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=0
    Total committed heap usage (bytes)=1266679808
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_POS=0
    WRONG_SIZE=0
    WRONG_SEQ=0
  File Input Format Counters
    Bytes Read=888190
  File Output Format Counters
    Bytes Written=8
hadoop@bnsccece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /rem/part-00000
cat : /rem/part-00000: No such file or directory
hadoop@bnsccece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /rem/part-r-00000
190 46
hadoop@bnsccece-HP-Elite-Tower-800-G9-Desktop-PC:~$ package weatherdataavg;

import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
public class AverageReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
public void reduce(Text key, Iterable<IntWritable> values,
Reducer<Text, IntWritable,
Text, IntWritable>.Context context) throws IOException,
InterruptedException {
int max_temp = 0;
int count = 0;
for (IntWritable value : values) {
max_temp += value.get();
count++;
}
context.write(key, new IntWritable(max_temp / count));
}
}

```

b)

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

- Eclipse IDE Window:** The title bar reads "eclipse-wc - weatherdatameanmax/src/weatherdatameanmax/MeanMaxReducer.java - Eclipse IDE".
- Package Explorer:** Shows the project structure under "weatherdatameanmax".
- MeanMaxReducer.java:** The code implements a Reducer for Hadoop. It calculates the mean and maximum temperature over a 3-day window. The code uses `IntWritable` for values and context.
- Outline View:** Shows the class structure, including the `reduce` method.
- Terminal Window:** Titled "hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC: ~", it displays the output of the "start-all.sh" command. The log includes messages about starting daemons like namenode, datanode, secondarynamenode, and resourcemanager.

```

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

  applications, not this command.
jnpipath  prints the java.library.path
kdtag   Diagnose Kerberos Problems
kername  show auth_to_local principal conversion
key    manage keys via the KeyProvider
rumenfolder scale a runen input trace
runentrace  convert tracings into a runen trace
sguard   S3 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@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /omn
hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/1901 /omn/1901
hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/weatherdatameanmax.jar weatherdatameanmax.MeanMaxDriver /omn/1901 /ren
2025-05-20 14:53:15.615 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Input paths from /omn/1901/metrics2.properties
2025-05-20 14:53:15.615 INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl: Scheduled metrics snapshot period 10 second(s).
2025-05-20 14:53:15.615 INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-20 14:53:15.674 WARN mapred.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-20 14:53:15.737 INFO input.FileInputFormat: Total input files to process : 1
2025-05-20 14:53:15.833 INFO mapred.JobSubmitter: number of splits:1
2025-05-20 14:53:15.833 INFO mapred.JobSubmitter: Submitting tokens for job: job_local2143084439_0001
2025-05-20 14:53:15.881 INFO mapred.JobSubmitter: The url to track the job: http://localhost:8080/
2025-05-20 14:53:15.891 INFO mapred.Job: Running job: job_local2143084439_0001
2025-05-20 14:53:15.895 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:53:15.895 INFO output.FileOutputCommitter: File Output Committer Algorith 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 14:53:15.895 INFO mapred.LocalJobunner: FileOutputCommitter is org.apache.hadoop.mapred.lib.output.FileOutputCommitter
2025-05-20 14:53:15.895 INFO mapred.LocalJobunner: Waiting for map tasks
2025-05-20 14:53:15.982 INFO mapred.LocalJobunner: 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 Algorith 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 14:53:16.008 INFO mapred.Tasks: Using ResourceCalculatorForProcessTree: [ ]
2025-05-20 14:53:16.084 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/omn/1901:0+888190
2025-05-20 14:53:16.084 INFO mapred.MapTask: (Equation) kvCount=1000000 (104857584)
2025-05-20 14:53:16.084 INFO mapred.MapTask: source=tcp://localhost:100
2025-05-20 14:53:16.084 INFO mapred.MapTask: soft limit at 83884800
2025-05-20 14:53:16.084 INFO mapred.MapTask: bufstart = 0; bufvold = 104857600
2025-05-20 14:53:16.084 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-20 14:53:16.118 INFO mapred.LocalJobunner: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-20 14:53:16.118 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
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; bufvold = 104857600
2025-05-20 14:53:16.119 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188144(104752576); length = 26253/6553600
2025-05-20 14:53:16.133 INFO mapred.Task: Task@attempt_local2143084439_0001_m_000000_0 is done. And ls in the process of committing
2025-05-20 14:53:16.135 INFO mapred.Task: attempt_local2143084439_0001_m_000000_0 Thread started: EventFetcher for fetching Map Completion Events
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: Using ResourceCalculatorForProcessTree: [ ]
2025-05-20 14:53:16.145 INFO mapred.ReduceTask: Using ResourceCalculatorForProcessTree: [ ]
2025-05-20 14:53:16.147 WARN org.apache.hadoop.mapred.Task.reduce.Shuffle@6b4ffe0: JobManagerImpl: JobTracker metrics system already initialized!
2025-05-20 14:53:16.155 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5829453312, maxSingleShuffleLimit=1457363328, mergeThreshold=3847439360, toSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-20 14:53:16.155 INFO reduce.MergeManagerImpl: MergerManager: attempt_local2143084439_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-05-20 14:53:16.172 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local2143084439_0001_m_000000_0 decomps: 59078 len: 59082 to MEMORY
2025-05-20 14:53:16.172 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 59078, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory -> 59078
2025-05-20 14:53:16.172 INFO reduce.MergeManagerImpl: EventFetcher is interrupted.. Returning
2025-05-20 14:53:16.174 INFO mapred.LocalJobunner: 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 14:53:16.181 INFO reduce.MergeManagerImpl: Merged 1 segments, 59078 bytes to disk to satisfy reduce memory limit
2025-05-20 14:53:16.181 INFO reduce.MergeManagerImpl: Merging 1 files, 59082 bytes from disk
2025-05-20 14:53:16.181 INFO reduce.MergeManagerImpl: Merged 1 segments, 0 bytes from memory into reduce
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.LocalJobunner: 1 / 1 copied.
2025-05-20 14:53:16.212 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2025-05-20 14:53:16.275 INFO mapred.LocalJobunner: Task@attempt_local2143084439_0001_m_000000_0 is done. And is in the process of committing
2025-05-20 14:53:16.275 INFO mapred.Task: attempt_local2143084439_0001_m_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-05-20 14:53:16.290 INFO mapred.Task: attempt_local2143084439_0001_m_000000_0 is allowed to commit now
2025-05-20 14:53:16.290 INFO mapred.Task: Task attempt_local2143084439_0001_m_000000_0 saved output of task "attempt_local2143084439_0001_m_000000_0" to hdfs://localhost:9000/ren
2025-05-20 14:53:16.290 INFO mapred.Task: attempt_local2143084439_0001_m_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-05-20 14:53:16.290 INFO mapred.Task: attempt_local2143084439_0001_m_000000_0 done.
2025-05-20 14:53:16.290 INFO mapred.Task: Task attempt_local2143084439_0001_m_000000_0 done.
2025-05-20 14:53:16.290 INFO mapred.Task: Final Counters for attempt_local2143084439_0001_m_000000_0: Counters: 30
  File System Counters
    FILE: Number of bytes read=122769
    FILE: Number of bytes written=763193
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=888190
    HDFS: Number of bytes written=81
    HDFS: Number of read operations=10
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=3
    HDFS: Number of bytes read erasure-coded=0
  Map-Reduce Framework
    Combine input records=0
    Combine output records=0
    Reduce input groups=12
    Reduce shuffle bytes=59082
    Reduce time records=5644
    Reduce output records=12
    Spilled Records=664
    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
    RACKER=0

```

Screenshot captured You can paste the image from the clipboard.

```

2025-05-20 14:53:15,982 INFO mapred.LocalJobRunner: Starting task: attempt_12345678901234567890_0001_m_000000_0
2025-05-20 14:53:15,996 INFO output.PathOutputCommitterFactory: No output C
2025-05-20 14:53:15,996 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:53:15,996 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:16,004 INFO mapred.Task: Using ResourceCalculatorProcessTree : []
2025-05-20 14:53:16,004 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/omn/1901:0+888190
2025-05-20 14:53:16,004 INFO mapred.MapTask: (Equivocation ID: 12345678901234567890)
2025-05-20 14:53:16,044 INFO mapred.MapTask: source.task.io.sort.mbo: 100
2025-05-20 14:53:16,044 INFO mapred.MapTask: soft limit at 83884960
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 14:53:16,118 INFO mapred.LocalJobRunner:
2025-05-20 14:53:16,118 INFO mapred.LocalJobRunner: Doing flush of map output
2025-05-20 14:53:16,119 INFO mapred.MapTask: Spilling map output
2025-05-20 14:53:16,119 INFO mapred.MapTask: bufstart = 0; bufend = 45948; bufvoid = 104857600
2025-05-20 14:53:16,119 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188144(104752576); length = 26253/6553600
2025-05-20 14:53:16,124 INFO mapred.MapTask: Finished spill 0
2025-05-20 14:53:16,133 INFO mapred.Task: Task:attempt_local2143084439_0001_m_000000_0 is done. And is in the process of committing
2025-05-20 14:53:16,135 INFO mapred.LocalJobRunner: map
2025-05-20 14:53:16,135 INFO mapred.Task: Task 'attempt_local2143084439_0001_m_000000_0' done.
2025-05-20 14:53:16,136 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
    Combined Input records=0
    Spilled Records=6564
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ns)=0
    Total committed heap usage (bytes)=526385152
  File Input Format Counters
    Bytes Read=888190
2025-05-20 14:53:16,138 INFO mapred.LocalJobRunner: Flushing task: attempt_local2143084439_0001_m_000000_0
2025-05-20 14:53:16,140 INFO mapred.LocalJobRunner: map task executor complete.
2025-05-20 14:53:16,140 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-05-20 14:53:16,140 INFO mapred.LocalJobRunner: Starting task: attempt_local2143084439_0001_r_000000_0
2025-05-20 14:53:16,145 INFO output.PathOutputCommitterFactory: No output Committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:53:16,145 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:53:16,145 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:16,145 INFO mapred.Task: Using ResourceCalculatorProcessTree : []

```

Screenshot captured
You can paste the image from the clipboard.

```

FILE: Number of bytes written=763193
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=888198
HDFS: Number of bytes written=81
HDFS: Number of large 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=0
  Reduce shuffle bytes=59082
  Reduce input records=6564
  Reduce output records=12
  Spilled Records=6564
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ns)=0
  Total committed heap usage (bytes)=526385152

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  NETWORK_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0

File Output Format Counters
  Bytes Written=81
2025-05-20 14:53:16,290 INFO mapred.LocalJobRunner: Finishing task: attempt_local2143084439_0001_r_000000_0
2025-05-20 14:53:16,291 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-20 14:53:16,295 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 14:53:16,902 INFO mapreduce.Job: Counters
  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=1776388
    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
  Total committed heap usage (bytes)=36

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=1776388
HDFS: Number of bytes written=81
HDFS: Number of read operations=15
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
HDFS: Number of bytes read erasure-coded=0

Map-Reduce Framework
  Map input records=6565
  Map output records=6564
  Map output bytes=45948
  Map output materialized bytes=59082
  Input split bytes=95
  Combine input records=0
  Combine output records=0
  Reduce input groups=12
  Reduce shuffle bytes=59082
  Reduce input records=6564
  Reduce output records=12
  Spilled Records=13128
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ns)=0
  Total committed heap usage (bytes)=1052770304

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  NETWORK_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=888198
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      108
06      168
07      219
08      198
09      141
10      100
11      1
12      -61
hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ 
```

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 Structure:** The Package Explorer view shows a project named "wordcount" containing a "src" folder with three files: WCDriver.java, WCMapper.java, and WCReducer.java.
- Code Editor:** The main editor window displays the WCDriver.java code. The code implements a Tool interface and defines a run method that checks for input arguments, sets up a JobConf, and configures the FileInputFormat and FileOutputFormat.
- Output Console:** The bottom terminal window shows the command-line execution of the Hadoop wordcount program. It includes commands like "hadoop fs -mkdir /rsg", "hadoop fs -copyFromLocal", and "hadoop jar". The log output shows standard Hadoop metrics and job progress.

```

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

2025-05-06 15:05:01,299 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-06 15:05:01,299 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-06 15:05:01,305 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-06 15:05:01,365 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-06 15:05:01,414 INFO mapred.FileInputFormat: Total input files to process : 1
2025-05-06 15:05:01,445 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-06 15:05:01,511 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local90897529_0001
2025-05-06 15:05:01,511 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-06 15:05:01,565 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-06 15:05:01,566 INFO mapreduce.Job: Running job: job_local90897529_0001
2025-05-06 15:05:01,566 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-06 15:05:01,567 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2025-05-06 15:05:01,569 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:05:01,569 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false , ignore cleanup failures: false
2025-05-06 15:05:01,606 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-06 15:05:01,607 INFO mapred.LocalJobRunner: Starting task: attempt_local90897529_0001_m_000000_0
2025-05-06 15:05:01,618 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:05:01,618 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false , ignore cleanup failures: false
2025-05-06 15:05:01,624 INFO mapred.Task: Using ResourceCalculatorProcessTree : []
2025-05-06 15:05:01,631 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/rsg/sample.txt:0+89
2025-05-06 15:05:01,640 INFO mapred.MapTask: numReduceTasks: 1
2025-05-06 15:05:01,671 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-06 15:05:01,671 INFO mapred.MapTask: mapreduce.task.io.sort.nb: 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; bufend = 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@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC: ~

BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
    Bytes Written=69
2025-05-06 15:05:01,897 INFO mapred.LocalJobRunner: Finishing task: attempt_local90897529_0001_r_000000_0
2025-05-06 15:05:01,897 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-06 15:05:02,569 INFO mapreduce.Job: Job job_local90897529_0001 running in uber mode : false
2025-05-06 15:05:02,572 INFO mapreduce.Job: map 100% reduce 100%
2025-05-06 15:05:02,574 INFO mapreduce.Job: Job job_local90897529_0001 completed successfully
2025-05-06 15:05:02,584 INFO mapreduce.Job: Counters: 36
File System Counters
    FILE: Number of bytes read=9008
    FILE: Number of bytes written=1279283
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=178
    HDFS: Number of bytes written=69
    HDFS: Number of read operations=15
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=4
    HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
    Map input records=5
    Map output records=20
    Map output bytes=169
    Map output materialized bytes=215
    Input split bytes=88
    Combine input records=0
    Combine output records=0
    Reduce input groups=10
    Reduce shuffle bytes=215
    Reduce input records=20
    Reduce output records=10
    Spilled Records=40
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=0
    Total committed heap usage (bytes)=1052770304
```

```
hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC: ~
FILE: Number of write operations=0
HDFS: Number of bytes read=89
HDFS: Number of bytes written=0
HDFS: Number of read operations=5
HDFS: Number of large read operations=0
HDFS: Number of write operations=1
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=5
  Map output records=20
  Map output bytes=169
  Map output materialized bytes=215
  Input split bytes=88
  Combine input records=0
  Spilled Records=20
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=0
  Total committed heap usage (bytes)=526385152
File Input Format Counters
  Bytes Read=89
2025-05-06 15:05:01,756 INFO mapred.LocalJobRunner: Finishing task: attempt_local90897529_0001_m_000000_0
2025-05-06 15:05:01,757 INFO mapred.LocalJobRunner: map task executor complete.
2025-05-06 15:05:01,758 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-05-06 15:05:01,758 INFO mapred.LocalJobRunner: Starting task: attempt_local90897529_0001_r_000000_0
2025-05-06 15:05:01,762 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:05:01,762 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false , ignore cleanup failures: false
2025-05-06 15:05:01,762 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-05-06 15:05:01,763 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@636a90e9
2025-05-06 15:05:01,764 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-06 15:05:01,771 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5827985408, maxSingleShuffleLimit=1456996352, mergeThreshold=3846470400, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-06 15:05:01,772 INFO reduce.EventFetcher: attempt_local90897529_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-05-06 15:05:01,785 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local90897529_0001_m_000000_0 dec
omp: 211 len: 215 to MEMORY
2025-05-06 15:05:01,787 INFO reduce.InMemoryMapOutput: Read 211 bytes from map-output for attempt_local90897529_0001_m_000000_0
2025-05-06 15:05:01,788 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 211, inMemoryMapOutputs.size() -> 1, commitMemory -> 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 reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2025-05-06 15:05:01,793 INFO mapred.Merger: Merging 1 sorted segments
2025-05-06 15:05:01,793 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 205 bytes
2025-05-06 15:05:01,793 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-06 15:05:01,867 INFO mapred.Task: Task@attempt_local90897529_0001_r_000000_0 is done. And is in the process of committing
2025-05-06 15:05:01,869 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-06 15:05:01,869 INFO mapred.Task: Task attempt_local90897529_0001_r_000000_0 is allowed to commit now
2025-05-06 15:05:01,894 INFO output.FileOutputCommitter: Saved output of task 'attempt_local90897529_0001_r_000000_0' to hdfs://localhost:9000/result
2025-05-06 15:05:01,896 INFO mapred.LocalJobRunner: reduce > reduce
2025-05-06 15:05:01,896 INFO mapred.Task: Task 'attempt_local90897529_0001_r_000000_0' done.
2025-05-06 15:05:01,897 INFO mapred.Task: Final Counters for attempt_local90897529_0001_r_000000_0: Counters: 30
File System Counters
    FILE: Number of bytes read=4735
    FILE: Number of bytes written=639749
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=89
    HDFS: Number of bytes written=69
    HDFS: Number of read operations=10
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=3
    HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
    Combine input records=0
    Combine output records=0
    Reduce input groups=10
    Reduce shuffle bytes=215
    Reduce input records=20
    Reduce output records=10
    Spilled Records=20
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=0
    Total committed heap usage (bytes)=526385152
Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
```

```
hadoop@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: ~ $
```

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 title "eclipse-wc - topnlab/src/topnlab/TopNMapper.java - Eclipse IDE". The left pane displays the "Package Explorer" showing a project named "topnlab" with four files: TopN.java, TopNReducer.java, TopNMapper.java, and TopNCombiner.java. The right pane shows the "Outline" view with the class structure of TopNMapper.java. The code itself is as follows:

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

```
hadoop@bmscsece-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@bmscsece-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 [bmscsece-HP-Elite-Tower-800-G9-Desktop-PC]
Starting resourcemanager
Starting nodemanagers
hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /mno
mkdir: Cannot create directory /mno. Name node is in safe mode.
hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ ^C
hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfsadmin -safemode get
Safe mode is OFF
hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /mno
hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/sample.txt /mno/sample.txt
hadoop@bmscsece-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@bmscsece-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@bmscsece-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@bmscscse-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@bmscscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar topnlab /mnosample.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@bmscscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar TopN /mnosample.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$1.run(URLClassLoader.java:452)
at java.base/java.security.AccessController.doPrivileged(Native Method)
at java.base/java.net.URLClassLoader.findClass(URLClassLoader.java:451)
at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:594)
at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:527)
at java.base/java.lang.Class.forName0(Native Method)
at java.base/java.lang.Class.forName(Class.java:398)
at org.apache.hadoop.util.RunJar.run(RunJar.java:321)
at org.apache.hadoop.util.RunJar.main(RunJar.java:241)
hadoop@bmscscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar topnlab.TopN /mnosample.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 FileOutputCommitterFactory
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

```

```
hadoop@bmscscse-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
0
2025-05-20 13:58:10,010 INFO mapred.MapTask: Finished spill 0
2025-05-20 13:58:10,014 INFO mapred.Task: Task:attempt_local408680812_0001_m_000000_0 is done. And is in the process of 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=751
        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@bmscse-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@bmscse-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=1457363328, 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@bmscsece-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@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /res/part-00000
cat: `/res/part-00000': No such file or directory
hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /res
```

```
hadoop@bmscsece-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@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /res/part-00000
cat: '/res/part-00000': No such file or directory
hadoop@bmscsece-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@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ ^
hadoop@bmscsece-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@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ 
```

Lab 8: Scala

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

Code with Output:

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

The screenshot shows a terminal window titled "GNU nano 6.2" with the file name "pi.scala" at the top. The code inside the file is:

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

At the bottom of the terminal window, there is a menu bar with various keyboard shortcuts for file operations like Help, Exit, Write Out, Read File, Where Is, Replace, Cut, Paste, Execute, Justify, Location, and Go To Line. A status bar at the bottom indicates "[Read 7 lines]".

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:

```
bmscsece@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC:~$ spark-shell
25/05/20 15:32:38 WARN Utils: Your hostname, bmscsece-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/bmscsece/Desktop/sparkdata.txt")
textFile: org.apache.spark.rdd.RDD[String] = /home/bmscsece/Desktop/sparkdata.txt MapPartitionsRDD[1] at textFile at <console>:2
4

scala>

scala> val counts = textFile
counts: org.apache.spark.rdd.RDD[String] = /home/bmscsece/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> █
```