VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka.

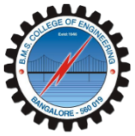
LAB REPORT
on

Big Data Analytics (22CS6PEBDA)

Submitted by

Polu Rajeswari Vinuthna (1BM22CS193)

in partial fulfillment for the award of the degree of
BACHELOR OF ENGINEERING
in
COMPUTER SCIENCE AND ENGINEERING



B.M.S. COLLEGE OF ENGINEERING
(Autonomous Institution under VTU)
BENGALURU-560019
Feb-2025 to July-2025

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

CERTIFICATE

This is to certify that the Lab work entitled "**Big Data Analytics**" carried out by **Polu Rajeswari Vinuthna (1BM22CS193),** who is a 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 2024-25. 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.

Vikranth B M
Assistant Professor
Department of CSE
BMSCE, Bengaluru

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

`

Index Sheet

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/VINUTHNA193/BDA

Lab 1:

I. MongoDB- CRUD Demonstration

1. Create a collection by name Customers with the following attributes.

Cust_id, Acc_Bal, Acc_Type

2. Insert at least 5 values into the table

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Customers.insertMany([
...   { Cust_id: 1, Acc_Bal: 1500, Acc_Type: 'Z' },
...   { Cust_id: 2, Acc_Bal: 1100, Acc_Type: 'Z' },
...   { Cust_id: 3, Acc_Bal: 2000, Acc_Type: 'X' },
...   { Cust_id: 4, Acc_Bal: 3000, Acc_Type: 'Z' },
...   { Cust_id: 5, Acc_Bal: 1800, Acc_Type: 'Z' }
... ]);
{
  acknowledged: true,
  insertedIds: {
    '0': ObjectId('67cff661b9683e22f4fa4214'),
    '1': ObjectId('67cff661b9683e22f4fa4215'),
    '2': ObjectId('67cff661b9683e22f4fa4216'),
    '3': ObjectId('67cff661b9683e22f4fa4217'),
    '4': ObjectId('67cff661b9683e22f4fa4218')
  }
}
```

3. Write a query to display those records whose total account balance is greater than 1200 of account type ‘Z’ for each customer_id.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Customers.find({
...   Acc_Bal: { $gt: 1200 },
...   Acc_Type: 'Z'
... });
[
  {
    _id: ObjectId('67cff661b9683e22f4fa4214'),
    Cust_id: 1,
    Acc_Bal: 1500,
    Acc_Type: 'Z'
  },
  {
    _id: ObjectId('67cff661b9683e22f4fa4217'),
    Cust_id: 4,
    Acc_Bal: 3000,
    Acc_Type: 'Z'
  },
  {
    _id: ObjectId('67cff661b9683e22f4fa4218'),
    Cust_id: 5,
    Acc_Bal: 1800,
    Acc_Type: 'Z'
  }
]
```

4. Determine Minimum and Maximum account balance for each customer_i

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Customers.aggregate([
...   {
...     $group: {
...       _id: "$Cust_id",           // Group by Cust_id
...       min_balance: { $min: "$Acc_Bal" },  // Minimum balance
...       max_balance: { $max: "$Acc_Bal" }   // Maximum balance
...     }
...   }
... ]);
[
  { _id: 2, min_balance: 1100, max_balance: 1100 },
  { _id: 3, min_balance: 2000, max_balance: 2000 },
  { _id: 4, min_balance: 3000, max_balance: 3000 },
  { _id: 5, min_balance: 1800, max_balance: 1800 },
  { _id: 1, min_balance: 1500, max_balance: 1500 }
]
Atlas atlas-9myyh8-shard-0 [primary] test>
```

II. You are developing an e-commerce platform where users can browse and purchase products. Each product has a unique identifier, a name, a category, a price, and available quantity. Additionally, users can add products to their cart and place orders. Design a MongoDB schema to efficiently handle product information, user carts, and orders.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Products.insertMany([
...   { product_id: "P001", name: "Smartphone XYZ", category: "Electronics", price: 599.99, quantity: 50 },
...   { product_id: "P002", name: "Laptop ABC", category: "Electronics", price: 999.99, quantity: 30 },
...   { product_id: "P003", name: "Wireless Headphones", category: "Electronics", price: 150.0, quantity: 100 },
...   { product_id: "P004", name: "Air Fryer", category: "Home Appliances", price: 89.99, quantity: 60 },
...   { product_id: "P005", name: "Gaming Chair", category: "Furniture", price: 199.99, quantity: 25 }
... ]);
{
  acknowledged: true,
  insertedIds: {
    '0': ObjectId('67cffafab9683e22f4fa4219'),
    '1': ObjectId('67cffafab9683e22f4fa421a'),
    '2': ObjectId('67cffafab9683e22f4fa421b'),
    '3': ObjectId('67cffafab9683e22f4fa421c'),
    '4': ObjectId('67cffafab9683e22f4fa421d')
  }
}
Atlas atlas-9myyh8-shard-0 [primary] test> db.Carts.insertMany([
...   { user_id: "789ghi123", products: [{ product_id: "P001", quantity: 2 }, { product_id: "P003", quantity: 1 }] },
...   { user_id: "456def456", products: [{ product_id: "P002", quantity: 1 }, { product_id: "P004", quantity: 3 }] },
...   { user_id: "123abc789", products: [{ product_id: "P005", quantity: 1 }] }
... ]);
{
  acknowledged: true,
  insertedIds: {
    '0': ObjectId('67cffb26b9683e22f4fa421e'),
    '1': ObjectId('67cffb26b9683e22f4fa421f'),
    '2': ObjectId('67cffb26b9683e22f4fa4220')
  }
}
Atlas atlas-9myyh8-shard-0 [primary] test> db.Orders.insertMany([
...   { user_id: "789ghi123", products: [{ product_id: "P001", quantity: 2, price: 599.99 }, { product_id: "P003", quantity: 1, price: 150.0 }], total_price: 1349.98, order_date: ISODate("2025-03-10T15:30:00Z") },
...   { user_id: "456def456", products: [{ product_id: "P002", quantity: 1, price: 999.99 }, { product_id: "P004", quantity: 3, price: 89.99 }], total_price: 1368.95, order_date: ISODate("2025-03-09T12:00:00Z") },
...   { user_id: "123abc789", products: [{ product_id: "P005", quantity: 1, price: 199.99 }], total_price: 199.99, order_date: ISODate("2025-03-08T10:00:00Z") }
... ]);
{
  acknowledged: true,
  insertedIds: {
    '0': ObjectId('67cffb44b9683e22f4fa4221'),
    '1': ObjectId('67cffb44b9683e22f4fa4222'),
    '2': ObjectId('67cffb44b9683e22f4fa4223')
  }
}
```

Query to Retrieve All Products.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Products.find();
[
  {
    _id: ObjectId('67cffafab9683e22f4fa4219'),
    product_id: 'P001',
    name: 'Smartphone XYZ',
    category: 'Electronics',
    price: 599.99,
    quantity: 50
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421a'),
    product_id: 'P002',
    name: 'Laptop ABC',
    category: 'Electronics',
    price: 999.99,
    quantity: 30
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421b'),
    product_id: 'P003',
    name: 'Wireless Headphones',
    category: 'Electronics',
    price: 150,
    quantity: 100
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421c'),
    product_id: 'P004',
    name: 'Air Fryer',
    category: 'Home Appliances',
    price: 89.99,
    quantity: 60
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421d'),
    product_id: 'P005',
    name: 'Gaming Chair',
    category: 'Furniture',
    price: 199.99,
    quantity: 25
  }
]
```

Retrieve Products in a Specific Category (e.g., Electronics).

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Products.find({ category: "Electronics" });
[
  {
    _id: ObjectId('67cffafab9683e22f4fa4219'),
    product_id: 'P001',
    name: 'Smartphone XYZ',
    category: 'Electronics',
    price: 599.99,
    quantity: 50
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421a'),
    product_id: 'P002',
    name: 'Laptop ABC',
    category: 'Electronics',
    price: 999.99,
    quantity: 30
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421b'),
    product_id: 'P003',
    name: 'Wireless Headphones',
    category: 'Electronics',
    price: 150,
    quantity: 100
  }
]
```

Retrieve Products with Quantity Greater Than 0.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Products.find({ quantity: { $gt: 0 } });
[
  {
    _id: ObjectId('67cffafab9683e22f4fa4219'),
    product_id: 'P001',
    name: 'Smartphone XYZ',
    category: 'Electronics',
    price: 599.99,
    quantity: 50
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421a'),
    product_id: 'P002',
    name: 'Laptop ABC',
    category: 'Electronics',
    price: 999.99,
    quantity: 30
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421b'),
    product_id: 'P003',
    name: 'Wireless Headphones',
    category: 'Electronics',
    price: 150,
    quantity: 100
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421c'),
    product_id: 'P004',
    name: 'Air Fryer',
    category: 'Home Appliances',
    price: 89.99,
    quantity: 60
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421d'),
    product_id: 'P005',
    name: 'Gaming Chair',
    category: 'Furniture',
    price: 199.99,
    quantity: 25
  }
]
```

Retrieve Products Sorted by Price in Ascending Order.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Products.find().sort({ price: 1 });
[
  {
    _id: ObjectId('67cffafab9683e22f4fa421c'),
    product_id: 'P004',
    name: 'Air Fryer',
    category: 'Home Appliances',
    price: 89.99,
    quantity: 60
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421b'),
    product_id: 'P003',
    name: 'Wireless Headphones',
    category: 'Electronics',
    price: 150,
    quantity: 100
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421d'),
    product_id: 'P005',
    name: 'Gaming Chair',
    category: 'Furniture',
    price: 199.99,
    quantity: 25
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa4219'),
    product_id: 'P001',
    name: 'Smartphone XYZ',
    category: 'Electronics',
    price: 599.99,
    quantity: 50
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421a'),
    product_id: 'P002',
    name: 'Laptop ABC',
    category: 'Electronics',
    price: 999.99,
    quantity: 30
  }
]
```

Retrieve Products with Price Less Than or Equal to $100.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Products.find({ price: { $lte: 100 } });
[
  {
    _id: ObjectId('67cffafab9683e22f4fa421c'),
    product_id: 'P004',
    name: 'Air Fryer',
    category: 'Home Appliances',
    price: 89.99,
    quantity: 60
  }
]
```

Retrieve Products Added to a User's Cart (User with ID "789ghi...")

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Carts.findOne({ user_id: "789ghi123" }, { products: 1 });
{
  _id: ObjectId('67cffb26b9683e22f4fa421e'),
  products: [
    { product_id: 'P001', quantity: 2 },
    { product_id: 'P003', quantity: 1 }
  ]
}
```

Retrieve Orders Placed by a User (User with ID "123abc...")

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Orders.find({ user_id: "123abc789" });
[
  {
    _id: ObjectId('67cffb44b9683e22f4fa4223'),
    user_id: '123abc789',
    products: [ { product_id: 'P005', quantity: 1, price: 199.99 } ],
    total_price: 199.99,
    order_date: ISODate('2025-03-08T10:00:00.000Z')
  }
]
```

Retrieve Total Price of Orders Placed by a User (User with ID "123abc...")

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Orders.aggregate([ { $match: { user_id: "123abc789" } }, { $group: { _id: "$user_id", total_order_price: { $sum: "$total_pr
[ { _id: '123abc789', total_order_price: 199.99 } ]
```

Additional Aggregation queries based on Assignment-3 design:

1 .Calculate Total Number of Products in Each Category.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Products.aggregate([
...   { $group: { _id: "$category", total_products: { $sum: 1 } } }
... ]);
[
  { _id: 'Electronics', total_products: 3 },
  { _id: 'Home Appliances', total_products: 1 },
  { _id: 'Furniture', total_products: 1 }
]
```

2. Calculate Total Price of Products in Each Category.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Products.aggregate([
...   { $group: { _id: "$category", total_price: { $sum: { $multiply: ["$price", "$quantity"] } } } }
... ]);
[
  { _id: 'Furniture', total_price: 4999.75 },
  { _id: 'Electronics', total_price: 74999.2 },
  { _id: 'Home Appliances', total_price: 5399.4 }
]
```

3. Find Average Price of Products.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Products.aggregate([
...   { $group: { _id: null, average_price: { $avg: "$price" } } }
... ]);
[ { _id: null, average_price: 407.992 } ]
```

4. Find Products with Quantity Less Than 10.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Products.find({ quantity: { $lt: 10 } });
```

5. Sort Products by Price in Descending Order.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Products.find().sort({ price: -1 });
[
  {
    _id: ObjectId('67cffafab9683e22f4fa421a'),
    product_id: 'P002',
    name: 'Laptop ABC',
    category: 'Electronics',
    price: 999.99,
    quantity: 30
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa4219'),
    product_id: 'P001',
    name: 'Smartphone XYZ',
    category: 'Electronics',
    price: 599.99,
    quantity: 50
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421d'),
    product_id: 'P005',
    name: 'Gaming Chair',
    category: 'Furniture',
    price: 199.99,
    quantity: 25
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421b'),
    product_id: 'P003',
    name: 'Wireless Headphones',
    category: 'Electronics',
    price: 150,
    quantity: 100
  },
  {
    _id: ObjectId('67cffafab9683e22f4fa421c'),
    product_id: 'P004',
    name: 'Air Fryer',
    category: 'Home Appliances',
    price: 89.99,
    quantity: 60
  }
]
```

6. Calculate Total Price of Orders Placed by Each User.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Orders.aggregate([
...   { $group: { _id: "$user_id", total_order_price: { $sum: "$total_price" } } }
... ]);
[
  { _id: '123abc789', total_order_price: 199.99 },
  { _id: '789ghi123', total_order_price: 1349.98 },
  { _id: '456def456', total_order_price: 1368.95 }
]
```

7. Find Users with the Highest Total Price of Orders.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Orders.aggregate([
...   { $group: { _id: "$user_id", total_order_price: { $sum: "$total_price" } } },
...   { $sort: { total_order_price: -1 } },
...   { $limit: 1 }
... ]);
[ { _id: '456def456', total_order_price: 1368.95 } ]
```

8. Find Average Total Price of Orders.

```
Atlas atlas-9myyh8-shard-0 [primary] test> db.Orders.aggregate([
...   { $group: { _id: null, average_order_price: { $avg: "$total_price" } } }
... ]);
[ { _id: null, average_order_price: 972.9733333333334 } ]
```

Lab 2: Cassandra

Working with Cassandra

Create KeySpace :

CREATE KEYSPACE Student WITH REPLICATION =

{'class':'SimpleStrategy','replication_factor':1};

Describe the existing Keyspaces:

DESCRIBE KEYSPACES;

```
AlreadyExists: Keyspace 'students' already exists
cqlsh> CREATE KEYSPACE Student WITH REPLICATION = {'class':'SimpleStrategy','replication_factor':1};
cqlsh> DESCRIBE KEYSPACES;

employees  students1    system_distributed  system_views
student    system       system_schema       system_virtual_schema
students   system_auth  system_traces
```

For More details on existing keyspaces:

SELECT * FROM system_schema.keyspaces;

```
Specify keyspace.tablename
cqlsh> SELECT * FROM system_schema.keyspaces;

 keyspace_name      | durable_writes | replication
--------------------+----------------+--------------------------------------------------------------------------------
            student |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '1'}
          employees |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '1'}
        system_auth |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '1'}
      system_schema |           True |                                {'class': 'org.apache.cassandra.locator.LocalStrategy'}
           students1 |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '1'}
 system_distributed |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '3'}
             system |           True |                                {'class': 'org.apache.cassandra.locator.LocalStrategy'}
      system_traces |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '2'}
           students |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '1'}

(9 rows)
cqlsh>
```

use the keyspace “Student”:

USE Student;

To create table (column family) by name Student_Info:

CREATE TABLE Student_Info (Roll_No int PRIMARY KEY, StudName text, DateOfJoining timestamp, last_exam_Percent double);

Lookup the names of all tables in the current keyspaces

DESCRIBE TABLES;

Describe the table information

DESCRIBE TABLE <Table_Name>;

```
(9 rows)
cqlsh> USE Student;
cqlsh:student> CREATE TABLE Student_Info (Roll_No int PRIMARY KEY, StudName text
, DateOfJoining timestamp, last_exam_Percent double);
cqlsh:student> DESCRIBE TABLES;

student_info

cqlsh:student>
cqlsh:student> DESCRIBE TABLE <Table_Name>;
```

CRUD

Insert :

BEGIN BATCH

INSERT INTO Student_Info(Roll_No, StudName, DateOfJoining, last_exam_Percent)

VALUES (1,'Asha','2012-03-12',79.9)

INSERT INTO Student_Info(Roll_No, StudName, DateOfJoining, last_exam_Percent)

VALUES (2,'Krian','2012-03-12',89.9)

INSERT INTO Student_Info(Roll_No, StudName, DateOfJoining, last_exam_Percent)

VALUES (3,'Tarun','2012-03-12',78.9)

INSERT INTO Student_Info(Roll_No, StudName, DateOfJoining, last_exam_Percent)

VALUES (4,'Samrth','2012-03-12',90.9)

INSERT INTO Student_Info(Roll_No, StudName, DateOfJoining, last_exam_Percent)

VALUES (5,'Smitha','2012-03-12',67.9)

INSERT INTO Student_Info(Roll_No, StudName, DateOfJoining, last_exam_Percent)

VALUES (6,'Rohan','2012-03-12',56.9)

APPLY BATCH;

View data from the table "Student_Info"

SELECT * FROM Student_Info;

```
cqlsh:student> SELECT * FROM  Student_Info;

 roll_no | dateofjoining                   | last_exam_percent | studname
---------+---------------------------------+-------------------+----------
       5 | 2012-03-11 18:30:00.000000+0000 |              67.9 |   Smitha
       1 | 2012-03-11 18:30:00.000000+0000 |              79.9 |     Asha
       2 | 2012-03-11 18:30:00.000000+0000 |              89.9 |    Krian
       4 | 2012-03-11 18:30:00.000000+0000 |              90.9 |   Samrth
       6 | 2012-03-11 18:30:00.000000+0000 |              56.9 |    Rohan
       3 | 2012-03-11 18:30:00.000000+0000 |              78.9 |    Tarun

(6 rows)
cqlsh:student> SELECT * FROM  Student_Info WHERE Roll_No IN (1,2,3);

 roll_no | dateofjoining                   | last_exam_percent | studname
---------+---------------------------------+-------------------+----------
       1 | 2012-03-11 18:30:00.000000+0000 |              79.9 |     Asha
       2 | 2012-03-11 18:30:00.000000+0000 |              89.9 |    Krian
       3 | 2012-03-11 18:30:00.000000+0000 |              78.9 |    Tarun
```

View data from the table "Student_Info" where Rollo column either has a value 1 or 2 or 3

SELECT * FROM Student_Info WHERE Roll_No IN (1,2,3);

```
(6 rows)
cqlsh:student> SELECT * FROM  Student_Info WHERE Roll_No IN (1,2,3);

 roll_no | dateofjoining                   | last_exam_percent | studname
---------+---------------------------------+-------------------+----------
       1 | 2012-03-11 18:30:00.000000+0000 |              79.9 |     Asha
       2 | 2012-03-11 18:30:00.000000+0000 |              89.9 |    Krian
       3 | 2012-03-11 18:30:00.000000+0000 |              78.9 |    Tarun
```

To execute a non primary key - will throw an error

select * from Student_info where Studname= 'Asha';

So create an INDEX on the Column as below:

To create an INDEX on StudName Column of the Student_Info column family

CREATE INDEX ON Student_Info (StudName);

Now execute the query based on the INDEXED Column:

select * from Student_info where Studname= 'Asha';

```
cqlsh:student> SELECT * FROM Student_info WHERE Studname = 'Asha' ALLOW FILTERIN
G;

 roll_no | dateofjoining                   | last_exam_percent | studname
---------+---------------------------------+-------------------+----------
       1 | 2012-03-11 18:30:00.000000+0000 |              79.9 |     Asha
```

To specify the number of rows returned in the output

select Roll_No, StudName from Student_info LIMIT 2;

```
(1 rows)
cqlsh:student> select Roll_No, StudName from Student_info LIMIT 2;

 roll_no | studname
---------+----------
       5 |   Smitha
       1 |     Asha
```

Alias for Column:

Select Roll_No as “USN” from Student_info;

```
cqlsh:student> SELECT Roll_No FROM Student_info;

 roll_no
---------
       5
       1
       2
       4
       6
       3

(6 rows)
cqlsh:student> ALTER TABLE Student_info RENAME Roll_No TO USN;
cqlsh:student> UPDATE Student_info SET StudName='David Sheen' WHERE RollNo=2;
```

UPDATE

UPDATE Student_info SET StudName='David Sheen' WHERE RollNo=2;

Lets try to update the primary key

UPDATE Student_info SET Roll_No=6 WHERE Roll_No=3;

DELETE

DELETE LastExamPercent FROM Student_info WHERE USN=2;

Delete a Row

DELETE FROM student_info WHERE USN=2;

Set Collection

A column of type set consists of unordered unique values. However, when the column is queried, it returns, it returns the values in sorted order. For example, for text values, it sorts in alphabetical order.

ALTER TABLE Student_info ADD hobbies set<text>

List Collection

When the order of elements matter, one should go for a list collection.

ALTER TABLE Student_info ADD language list<text>;

UPDATE Student_info

SET hobbies=hobbies+{'Chess,Table Tennis'}

WHERE USN=1;

SELECt * from Student_info WHERE USN=1;

```
cqlsh:student> SELECt * from Student_info WHERE USN=1;

 usn | dateofjoining                   | last_exam_percent | studname
-----+---------------------------------+-------------------+----------
   1 | 2012-03-11 18:30:00.000000+0000 |              79.9 |     Asha
```

UPDATE Student_info

SET langusge=language+['Hindi,English']

WHERE USN=1;

Note: You can remove an element from a set using the subtraction(-) operator.

USING A COUNTER

A counter is a special column that is changed in increments. For example, we may need a counter column to count the number of times a particular book is issued from the library bythe student.

CREATE TABLE library_book(counter_value counter, book_name varchar, stud_name varchar, PRIMARY KEY(book_name,stud_name));

Load data into the counter column

UPDATE library_book SET counetr value=couner_vale+1 WHERE book_name='Big data Analytics' AND stud_name='jeet';

```
cqlsh:student> UPDATE library_book SET counter_value = counter_value + 1
           ... WHERE book_name='Big data Analytics' AND stud_name='jeet';
cqlsh:student> CREATE TABLE userlogin(userid int PRIMARY KEY, password text);
cqlsh:student>
cqlsh:student> INSERT INTO userlogin(userid, password) VALUES (1,'infy') USING T
TL 30;
cqlsh:student>
cqlsh:student> SELECT TTL(password) FROM userlogin WHERE userid=1;

 ttl(password)
---------------
            30
```

TIME TO LIVE

CREATE TABLE userlogin(userid int PRIMARY KEY, password text);

INSERT INTO userlogin(userid, password) VALUES (1,'infy') USING TTL 30;

SELECT TTL(password) FROM userlogin WHERE userid=1;

IMPORT and EXPORT

Export to CSV

COPY elearninglists(id,course_order, course_id,courseowner,title) TO 'd:\elearninglists.csv';

```
cqlsh:student> COPY Student_info(USN, StudName, DateOfJoining, last_exam_Percent
) TO 'd:\student_info.csv';
Using 16 child processes

Starting copy of student.student_info with columns [usn, studname, dateofjoining
, last_exam_percent].
Processed: 4 rows; Rate:      53 rows/s; Avg. rate:      53 rows/s
4 rows exported to 1 files in 0.085 seconds.
```

Import from CSV

COPY elearninglists(id,course_order, course_id,courseowner,title) FROM 'd:\elearninglists.csv';

```
cqlsh:student> COPY Student_info (USN, StudName, DateOfJoining, last_exam_Percen
t)
           ... FROM 'd:\student_info.csv';
Using 16 child processes

Starting copy of student.student_info with columns [usn, studname, dateofjoining
, last_exam_percent].
Processed: 4 rows; Rate:       8 rows/s; Avg. rate:      11 rows/s
4 rows imported from 1 files in 0.363 seconds (0 skipped).
```

```
cqlsh:student> SELECT * FROM Student_info;

 usn | dateofjoining                   | hobbies                  | language            | last_exam_percent | studname
-----+---------------------------------+--------------------------+---------------------+-------------------+----------
   5 | 2012-03-11 18:30:00.000000+0000 |                     null |                null |              67.9 |   Smitha
   1 | 2012-03-11 18:30:00.000000+0000 | {'Chess', 'Table Tennis'} | ['Hindi', 'English'] |              79.9 |     Asha
   4 | 2012-03-11 18:30:00.000000+0000 |                     null |                null |              90.9 |   Samrth
   6 | 2012-03-11 18:30:00.000000+0000 |                     null |                null |              78.9 |    Tarun

(4 rows)
```

Import FROM STDIN

COPY persons(id,fname,lnmae)FROM STDIN;

Export to STDOUT

COPY elearninglists(id,course_order, course_id,courseowner,title) TO STDOUT;

Lab 3: Cassandra

Step 1: Create Keyspace Library

```
CREATE KEYSPACE Library WITH replication = {
  'class': 'SimpleStrategy',
  'replication_factor': 1
};
```

Step 2: Create Column Family Library_Info

```
CREATE TABLE Library.Counter_Table (
        Stud_Id int PRIMARY KEY,
        Counter_value counter
);
CREATE TABLE Library.Library_Info (
        Stud_Id int,
        Stud_Name text,
        Book_Name text,
        Book_Id text,
        Date_of_issue date,
        PRIMARY KEY (Stud_Id, Book_Name)
);
```

```
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.0.0 | Cassandra 4.0.5 | CQL spec 3.4.5 | Native protocol v5]
Use HELP for help.
cqlsh> CREATE KEYSPACE Library WITH replication = {
   ...   'class': 'SimpleStrategy',
   ...   'replication_factor': 1
   ... };

cqlsh>
cqlsh> CREATE TABLE Library.Counter_Table (
   ...     Stud_Id int PRIMARY KEY,
   ...     Counter_value counter
   ... );
cqlsh> CREATE TABLE Library.Library_Info (
   ...     Stud_Id int,
   ...     Stud_Name text,
   ...     Book_Name text,
   ...     Book_Id text,
   ...     Date_of_issue date,
   ...     PRIMARY KEY (Stud_Id, Book_Name)
   ... );
```

Step 3: Insert Values into the Table in Batch

```
BEGIN BATCH
```

INSERT INTO Library.Library_Info (Stud_Id, Stud_Name, Book_Name, Book_Id, Date_of_issue)
VALUES (112, 'Alice', 'BDA', 'B101', '2025-04-01');
INSERT INTO Library.Library_Info (Stud_Id, Stud_Name, Book_Name, Book_Id, Date_of_issue)
VALUES (113, 'Bob', 'DBMS', 'B102', '2025-04-02');
APPLY BATCH;

```
cqlsh> BEGIN BATCH
   ... INSERT INTO Library.Library_Info (Stud_Id, Stud_Name, Book_Name, Book_Id,
 Date_of_issue)
   ... VALUES (112, 'Alice', 'BDA', 'B101', '2025-04-01');
   ...
   ... INSERT INTO Library.Library_Info (Stud_Id, Stud_Name, Book_Name, Book_Id,
 Date_of_issue)
   ... VALUES (113, 'Bob', 'DBMS', 'B102', '2025-04-02');
   ... APPLY BATCH;
```

UPDATE Library.Counter_Table SET Counter_value = Counter_value + 1 WHERE Stud_Id = 112;

```
cqlsh> UPDATE Library.Counter_Table SET Counter_value = Counter_value + 1 WHERE
Stud_Id = 112;
```

Step 4: Display Table Details and Update Counter

SELECT * FROM Library.Library_Info;

SELECT * FROM Library.Counter_Table;

```
cqlsh> SELECT * FROM Library.Library_Info;

 stud_id | book_name | book_id | date_of_issue | stud_name
---------+-----------+---------+---------------+-----------
     113 |      DBMS |    B102 |    2025-04-02 |       Bob
     112 |       BDA |    B101 |    2025-04-01 |     Alice

(2 rows)
cqlsh> SELECT * FROM Library.Counter_Table;

 stud_id | counter_value
---------+---------------
     112 |             1

(1 rows)
```

To increase counter value

UPDATE Library.Counter_Table SET Counter_value = Counter_value + 1 WHERE Stud_Id = 112;

```
(1 rows)
cqlsh> UPDATE Library.Counter_Table SET Counter_value = Counter_value + 1 WHERE
Stud_Id = 112;
```

Step 5: Show a Student with ID 112 has Taken Book "BDA" 2 Times

SELECT * FROM Library.Counter_Table WHERE Stud_Id = 112;

```
cqlsh> SELECT * FROM Library.Counter_Table WHERE Stud_Id = 112;

 stud_id | counter_value
---------+---------------
     112 |             2
```

Step 6: Export Table to CSV and Import the CSV

1. Exit cqlsh:

   cqlsh> exit

2. **Then in your system terminal run:**
3. Export data to CSV:

   cqlsh -e "COPY Library.Library_Info TO 'Library_Info.csv' WITH HEADER = true;"

4. Import data from CSV:

   cqlsh -e "COPY Library.Library_Info FROM 'Library_Info.csv' WITH HEADER = true;"

```
cqlsh> exit
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cqlsh -e "COPY Library.Lib
rary_Info TO 'Library_Info.csv' WITH HEADER = true;"
Using 16 child processes

Starting copy of library.library_info with columns [stud_id, book_name, book_id,
 date_of_issue, stud_name].
Processed: 2 rows; Rate:      24 rows/s; Avg. rate:      24 rows/s
2 rows exported to 1 files in 0.092 seconds.
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cqlsh -e "COPY Library.Lib
rary_Info FROM 'Library_Info.csv' WITH HEADER = true;"
Using 16 child processes

Starting copy of library.library_info with columns [stud_id, book_name, book_id,
 date_of_issue, stud_name].
Processed: 2 rows; Rate:       4 rows/s; Avg. rate:       5 rows/s
2 rows imported from 1 files in 0.380 seconds (0 skipped).
```

Step 7: Show Imported Data

1. Open cqlsh:

cqlsh

```
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.0.0 | Cassandra 4.0.5 | CQL spec 3.4.5 | Native protocol v5]
Use HELP for help.
```

2. Use your keyspace:

   USE Library;

3. Run a SELECT query to view the data:

   SELECT * FROM Library_Info;

```
cqlsh> USE Library;
cqlsh:library> SELECT * FROM Library_Info;

 stud_id | book_name | book_id | date_of_issue | stud_name
---------+-----------+---------+---------------+-----------
     113 |      DBMS |    B102 |    2025-04-02 |       Bob
     112 |       BDA |    B101 |    2025-04-01 |     Alice

(2 rows)
cqlsh:library>
```

Lab 4: Cassandra

Execution of HDFS Commands for interaction with Hadoop Environment.

Example Scenario for Execution:

1. **Start Hadoop (make sure to use the correct user, in this case hduser):**

   $ start-all.sh

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 15924.  Stop it first and ensure /tmp/
hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 16070.  Stop it first and ensure /tmp/
hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-600-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: secondarynamenode is running as proce
ss 16276.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid fil
e is empty before retry.
Starting resourcemanager
resourcemanager is running as process 16601.  Stop it first and ensure /tmp/hado
op-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 16731.  Stop it first and ensure /t
mp/hadoop-hadoop-nodemanager.pid file is empty before retry.
```

2. **Create Directory in HDFS (create a directory named /bda_hadoop):**

   $ hdfs dfs -mkdir /bda_hadoop

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -mkdir /bda_hadoop
```

3. **List Contents of HDFS Root (list the files and directories at the root of HDFS):**

   $ hadoop fs -ls /

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -ls /
Found 15 items
drwxr-xr-x   - hadoop supergroup          0 2023-04-27 12:49 /Hadoop
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:27 /abc
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:39 /bda_hadoop
drwxr-xr-x   - hadoop supergroup          0 2023-05-17 10:00 /count
drwxr-xr-x   - hadoop supergroup          0 2023-05-22 10:14 /deept
drwxr-xr-x   - hadoop supergroup          0 2023-05-15 14:51 /dir
drwxr-xr-x   - hadoop supergroup          0 2023-04-27 12:44 /hadoop
drwxr-xr-x   - hadoop supergroup          0 2023-05-04 12:59 /inputbda
drwxr-xr-x   - hadoop supergroup          0 2023-05-22 14:53 /output3
drwxr-xr-x   - hadoop supergroup          0 2023-05-04 13:01 /outputbda
drwxr-xr-x   - hadoop supergroup          0 2023-05-15 14:36 /pankaj
drwxr-xr-x   - hadoop supergroup          0 2023-04-27 12:20 /sample
drwxr-xr-x   - hadoop supergroup          0 2023-05-19 11:49 /temp01input
drwxr-xr-x   - hadoop supergroup          0 2023-05-19 11:57 /temp02input
drwxr-xr-x   - hadoop supergroup          0 2023-05-12 12:02 /tempinput
```

4. **Put Local File into HDFS (copy a local file from /home/hduser/Desktop/bda_local.txt to HDFS as /bda_hadoop/file.txt):**

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

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -put /home/hadoop/D
esktop/Welcome.txt /bda_hadoop/file.txt
```

5. **Display Contents of the File (display the contents of the file /bda_hadoop/file.txt on HDFS):**

   $ hdfs dfs -cat /bda_hadoop/file.txt

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -cat /bda_hadoop/fi
le.txt
Welcome
```

6. **Copy File from Local File System to HDFS (using copyFromLocal to copy a local file into HDFS):**

   $ hdfs dfs -copyFromLocal /home/hadoop/Desktop/Tee/bda_hadoop/file_cp_local.txt

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -copyFromLocal /hom
e/hadoop/Desktop/Tee /bda_hadoop/file_cp_local.txt
```

7. **Retrieve a File from HDFS to Local (use get to copy a file from HDFS to the local filesystem):**

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

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -get /bda_hadoop/fi
le.txt /home/hadoop/Downloads/downloaded_file.txt
```

8. **Move Directory in HDFS (move the directory /bda_hadoop to /abc in HDFS):**

   $ hadoop fs -mv /bda_hadoop /abc

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -mv /bda_hadoop /a
bc
```

9. **Verify Directory Move (check that the directory /bda_hadoop has been successfully moved to /abc):**

   $ hadoop fs -ls /abc

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -ls /abc
Found 3 items
-rw-r--r--   1 hadoop supergroup         70 2023-05-11 14:35 /abc/Tee
-rw-r--r--   1 hadoop supergroup          9 2025-04-15 14:27 /abc/WC.txt
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:42 /abc/bda_hadoop
```

10. **Copy Directory in HDFS:**

Copy a file from /abc/bda_hadoop/file.txt to another location, say /abc/file_copy.txt:

$ hadoop fs -cp /abc/bda_hadoop/file.txt /abc/file_copy.txt

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -cp /abc/bda_hadoo
p/file.txt /abc/file_copy.txt
```

Copy an entire directory (e.g., /abc/bda_hadoop) to /copy_bda_hadoop:

$ hadoop fs -cp /abc/bda_hadoop /copy_bda_hadoop

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -cp /abc/bda_hadoo
p /copy_bda_hadoop
```

Verify Copy:

Directory:

$ hadoop fs -ls /copy_bda_hadoop

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -ls /copy_bda_hado
op
Found 2 items
-rw-r--r--   1 hadoop supergroup          9 2025-04-15 14:47 /copy_bda_hadoop/fi
le.txt
-rw-r--r--   1 hadoop supergroup         70 2025-04-15 14:47 /copy_bda_hadoop/fi
le_cp_local.txt
```

File:

$ hadoop fs -cat /abc/file_copy.txt

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -cat /abc/file_cop
y.txt
Welcome
```

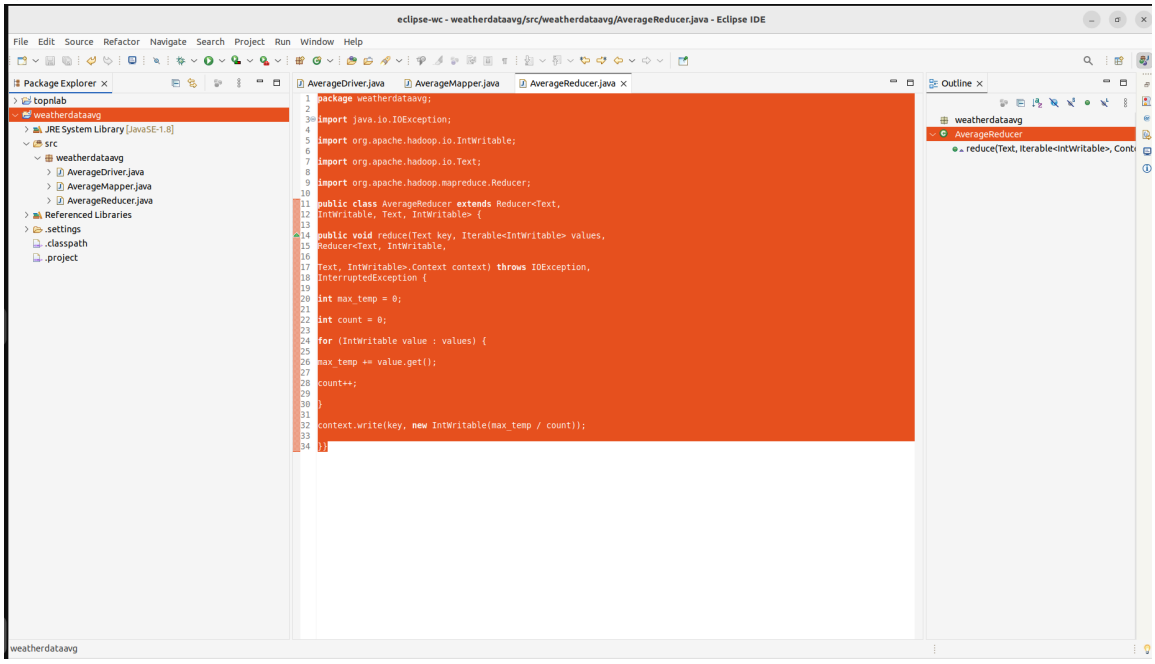
Lab 5: Hadoop

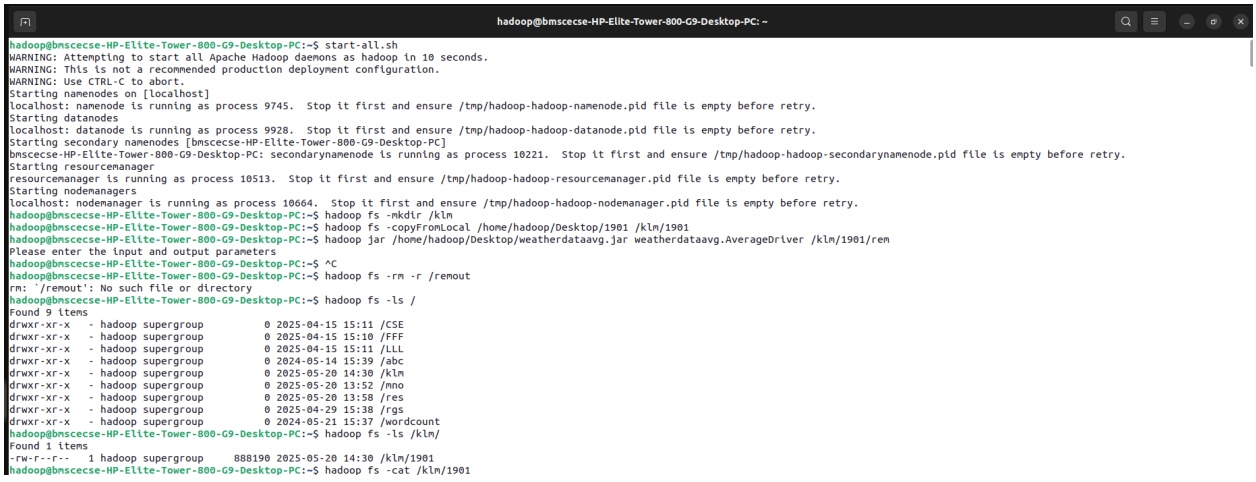
Question: From the following link extract the weather data https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all Create a Map Reduce program to

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

Code with Output:

a)





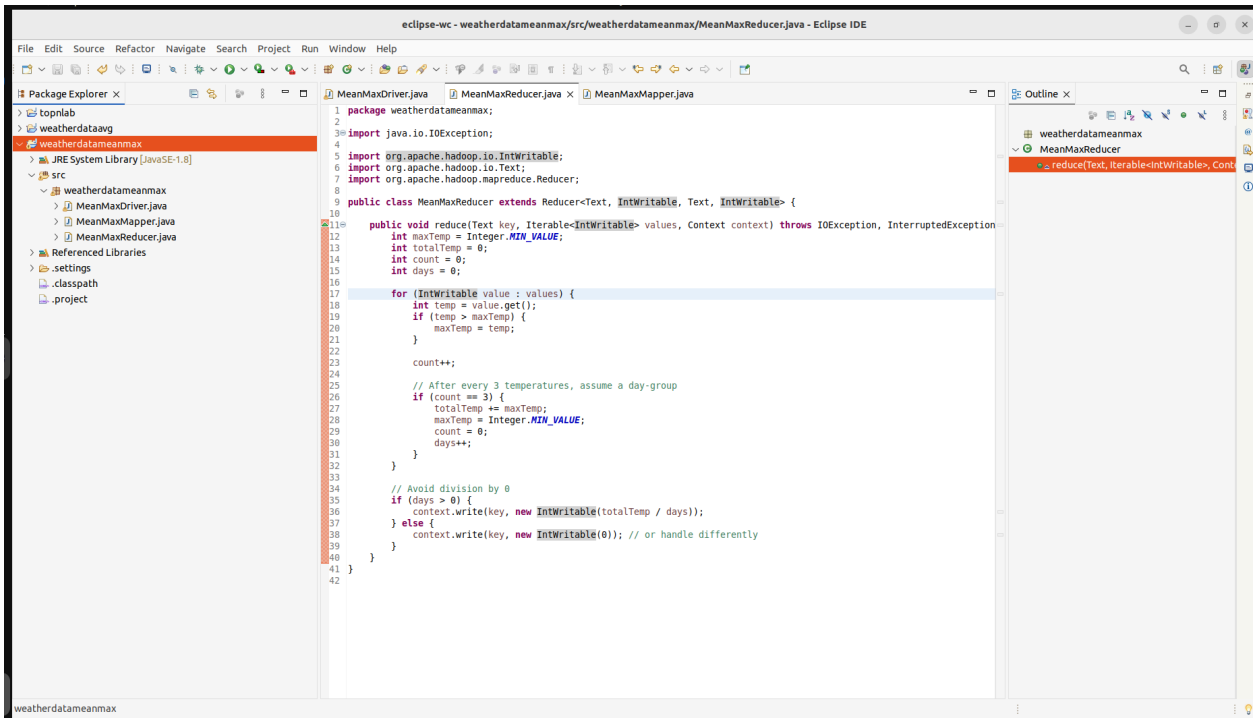
```
hadoop@bnscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/weatherdataavg.jar weatherdataavg.AverageDriver /kln/1901 /rem
2025-05-20 14:34:21,074 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-20 14:34:21,116 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-20 14:34:21,116 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-20 14:34:21,176 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-20 14:34:21,230 INFO input.FileInputFormat: Total input files to process : 1
2025-05-20 14:34:21,262 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-20 14:34:21,334 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local261815800_0001
2025-05-20 14:34:21,334 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-20 14:34:21,408 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-20 14:34:21,401 INFO mapreduce.Job: Running job: job_local261815800_0001
2025-05-20 14:34:21,402 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-20 14:34:21,405 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21,405 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:34:21,406 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:34:21,406 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-05-20 14:34:21,446 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-20 14:34:21,447 INFO mapred.LocalJobRunner: Starting task: attempt_local261815800_0001_m_000000_0
2025-05-20 14:34:21,459 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21,460 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:34:21,460 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:34:21,469 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-20 14:34:21,472 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/kln/1901:0+888190
2025-05-20 14:34:21,511 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-20 14:34:21,511 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-20 14:34:21,511 INFO mapred.MapTask: soft limit at 83886080
2025-05-20 14:34:21,511 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-20 14:34:21,511 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-20 14:34:21,513 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-20 14:34:21,586 INFO mapred.LocalJobRunner:
2025-05-20 14:34:21,587 INFO mapred.MapTask: Starting flush of map output
2025-05-20 14:34:21,587 INFO mapred.MapTask: Spilling map output
2025-05-20 14:34:21,587 INFO mapred.MapTask: bufstart = 0; bufend = 59076; bufvoid = 104857600
2025-05-20 14:34:21,587 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188144(104752576); length = 26253/6553600
2025-05-20 14:34:21,595 INFO mapred.MapTask: Finished spill 0
2025-05-20 14:34:21,600 INFO mapred.Task: Task:attempt_local261815800_0001_m_000000_0 is done. And is in the process of committing
2025-05-20 14:34:21,602 INFO mapred.LocalJobRunner: map
2025-05-20 14:34:21,602 INFO mapred.Task: Task 'attempt_local261815800_0001_m_000000_0' done.
2025-05-20 14:34:21,605 INFO mapred.Task: Final Counters for attempt_local261815800_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=4430
                FILE: Number of bytes written=713998
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=888190
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=6565
                Map output records=6564
```

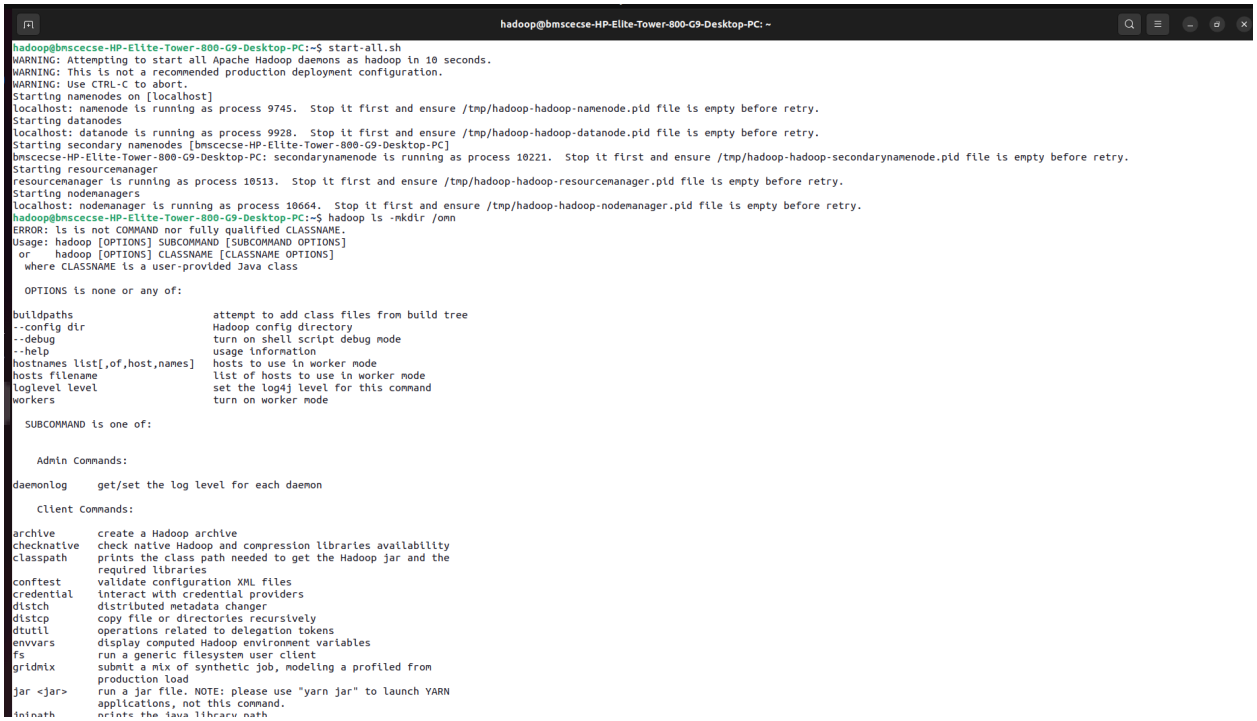
```
                HDFS: Number of bytes written=8
                HDFS: Number of read operations=10
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=72210
                Reduce input records=6564
                Reduce output records=1
                Spilled Records=6564
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=633339904
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=8
2025-05-20 14:34:21,803 INFO mapred.LocalJobRunner: Finishing task: attempt_local261815800_0001_r_000000_0
2025-05-20 14:34:21,803 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-20 14:34:22,404 INFO mapreduce.Job: Job job_local261815800_0001 running in uber mode : false
2025-05-20 14:34:22,405 INFO mapreduce.Job:  map 100% reduce 100%
2025-05-20 14:34:22,407 INFO mapreduce.Job: Job job_local261815800_0001 completed successfully
2025-05-20 14:34:22,414 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=153312
                FILE: Number of bytes written=1500206
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1776380
                HDFS: Number of bytes written=8
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=6565
                Map output records=6564
                Map output bytes=59076
                Map output materialized bytes=72210
                Input split bytes=95
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=72210
                Reduce input records=6564
```

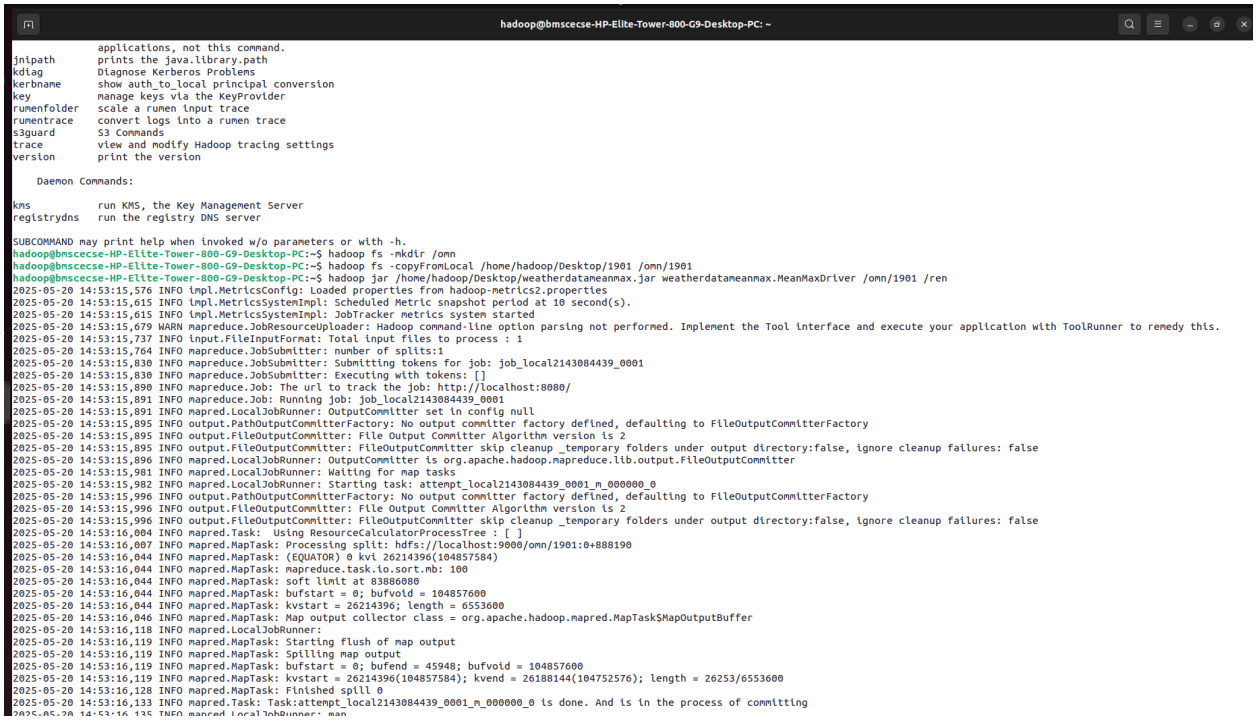
```
                Map input records=6565
                Map output records=6564
                Map output bytes=59076
                Map output materialized bytes=72210
                Input split bytes=95
                Combine input records=0
                Spilled Records=6564
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=633339904
        File Input Format Counters
                Bytes Read=888190
2025-05-20 14:34:21,605 INFO mapred.LocalJobRunner: Finishing task: attempt_local261815800_0001_m_000000_0
2025-05-20 14:34:21,606 INFO mapred.LocalJobRunner: map task executor complete.
2025-05-20 14:34:21,607 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-05-20 14:34:21,608 INFO mapred.LocalJobRunner: Starting task: attempt_local261815800_0001_r_000000_0
2025-05-20 14:34:21,612 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21,612 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:34:21,612 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:34:21,612 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-20 14:34:21,613 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@c4d4167
2025-05-20 14:34:21,614 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-20 14:34:21,622 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5829453312, maxSingleShuffleLimit=1457363328, mergeThreshold=3847439360, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-20 14:34:21,623 INFO reduce.EventFetcher: attempt_local261815800_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-05-20 14:34:21,635 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local261815800_0001_m_000000_0 decomp: 72206 len: 72210 to MEMORY
2025-05-20 14:34:21,637 INFO reduce.InMemoryMapOutput: Read 72206 bytes from map-output for attempt_local261815800_0001_m_000000_0
2025-05-20 14:34:21,637 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 72206, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->72206
2025-05-20 14:34:21,638 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2025-05-20 14:34:21,638 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:34:21,638 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2025-05-20 14:34:21,641 INFO mapred.Merger: Merging 1 sorted segments
2025-05-20 14:34:21,641 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 72199 bytes
2025-05-20 14:34:21,644 INFO reduce.MergeManagerImpl: Merged 1 segments, 72206 bytes to disk to satisfy reduce memory limit
2025-05-20 14:34:21,645 INFO reduce.MergeManagerImpl: Merging 1 files, 72210 bytes from disk
2025-05-20 14:34:21,645 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2025-05-20 14:34:21,645 INFO mapred.Merger: Merging 1 sorted segments
2025-05-20 14:34:21,645 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 72199 bytes
2025-05-20 14:34:21,646 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:34:21,679 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2025-05-20 14:34:21,767 INFO mapred.Task: Task:attempt_local261815800_0001_r_000000_0 is done. And is in the process of committing
2025-05-20 14:34:21,771 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:34:21,771 INFO mapred.Task: Task attempt_local261815800_0001_r_000000_0 is allowed to commit now
2025-05-20 14:34:21,800 INFO output.FileOutputCommitter: Saved output of task 'attempt_local261815800_0001_r_000000_0' to hdfs://localhost:9000/rem
2025-05-20 14:34:21,802 INFO mapred.LocalJobRunner: reduce > reduce
2025-05-20 14:34:21,802 INFO mapred.Task: Task 'attempt_local261815800_0001_r_000000_0' done.
2025-05-20 14:34:21,803 INFO mapred.Task: Final Counters for attempt_local261815800_0001_r_000000_0: Counters: 30
        File System Counters
                FILE: Number of bytes read=148882
                FILE: Number of bytes written=786208
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=888190
                HDFS: Number of bytes written=8
                HDFS: Number of read operations=10
```

```
                Reduce shuffle bytes=72210
                Reduce input records=6564
                Reduce output records=1
                Spilled Records=13128
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=1266679808
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=888190
        File Output Format Counters
                Bytes Written=8
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /rem/part-00000
cat: '/rem/part-00000': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /rem/part-r-00000
1901    46
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ package weatherdataavg;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Reducer;

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

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

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

int max_temp = 0;

int count = 0;

for (IntWritable value : values) {

max_temp += value.get();

count++;

}

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

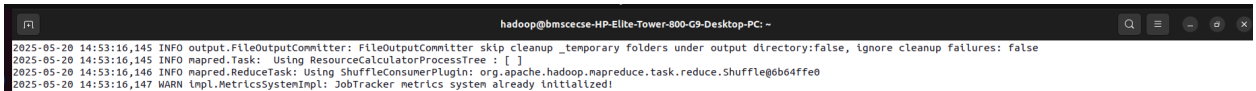
b)


eclipse-wc - weatherdatameanmax/src/weatherdatameanmax/MeanMaxReducer.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer
toplab
weatherdataavg
weatherdatameanmax
JRE System Library [JavaSE-1.8]
src
weatherdatameanmax
MeanMaxDriver.java
MeanMaxMapper.java
MeanMaxReducer.java
Referenced Libraries
.settings
.classpath
.project
MeanMaxDriver.java
MeanMaxReducer.java
MeanMaxMapper.java
1 package weatherdatameanmax;
2
3 import java.io.IOException;
4
5 import org.apache.hadoop.io.IntWritable;
6 import org.apache.hadoop.io.Text;
7 import org.apache.hadoop.mapreduce.Reducer;
8
9 public class MeanMaxReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
10
11 public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException
12 int maxTemp = Integer.MIN_VALUE;
13 int totalTemp = 0;
14 int count = 0;
15 int days = 0;
16
17 for (IntWritable value : values) {
18 int temp = value.get();
19 if (temp > maxTemp) {
20 maxTemp = temp;
21 }
22
23 count++;
24
25 // After every 3 temperatures, assume a day-group
26 if (count == 3) {
27 totalTemp += maxTemp;
28 maxTemp = Integer.MIN_VALUE;
29 count = 0;
30 days++;
31 }
32 }
33
34 // Avoid division by 0
35 if (days > 0) {
36 context.write(key, new IntWritable(totalTemp / days));
37 } else {
38 context.write(key, new IntWritable(0)); // or handle differently
39 }
40 }
41 }
42
Outline
weatherdatameanmax
MeanMaxReducer
reduce(Text, Iterable<IntWritable>, Cont
weatherdatameanmax

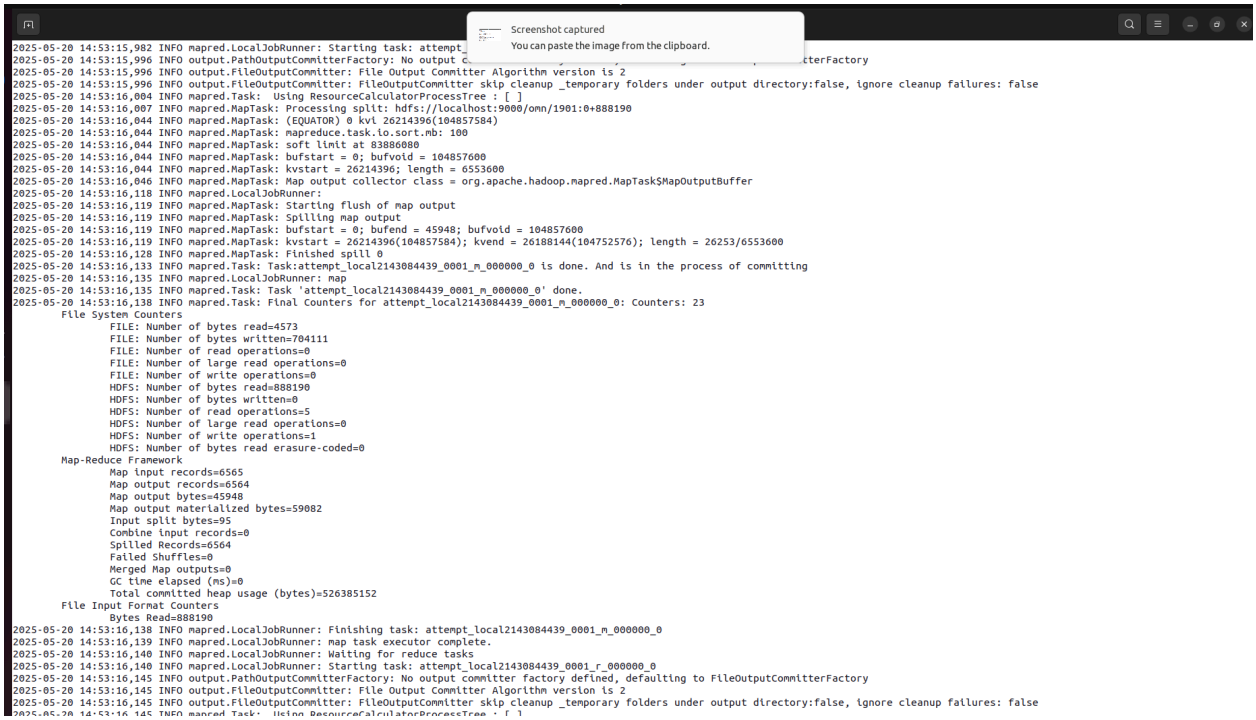

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 9745. Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 9928. Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10221. Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
resourcemanager is running as process 10513. Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10664. Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop ls -mkdir /omn
ERROR: ls is not COMMAND nor fully qualified CLASSNAME.
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
or hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
where CLASSNAME is a user-provided Java class
OPTIONS is none or any of:
buildpaths attempt to add class files from build tree
--config dir Hadoop config directory
--debug turn on shell script debug mode
--help usage information
hostnames list[,of,host,names] hosts to use in worker mode
hosts filename list of hosts to use in worker mode
loglevel level set the log4j level for this command
workers turn on worker mode
SUBCOMMAND is one of:
Admin Commands:
daemonlog get/set the log level for each daemon
Client Commands:
archive create a Hadoop archive
checknative check native Hadoop and compression libraries availability
classpath prints the class path needed to get the Hadoop jar and the
required libraries
conftest validate configuration XML files
credential interact with credential providers
distch distributed metadata changer
distcp copy file or directories recursively
dtutil operations related to delegation tokens
envvars display computed Hadoop environment variables
fs run a generic filesystem user client
gridmix submit a mix of synthetic job, modeling a profiled from
production load
jar <jar> run a jar file. NOTE: please use "yarn jar" to launch YARN
applications, not this command.

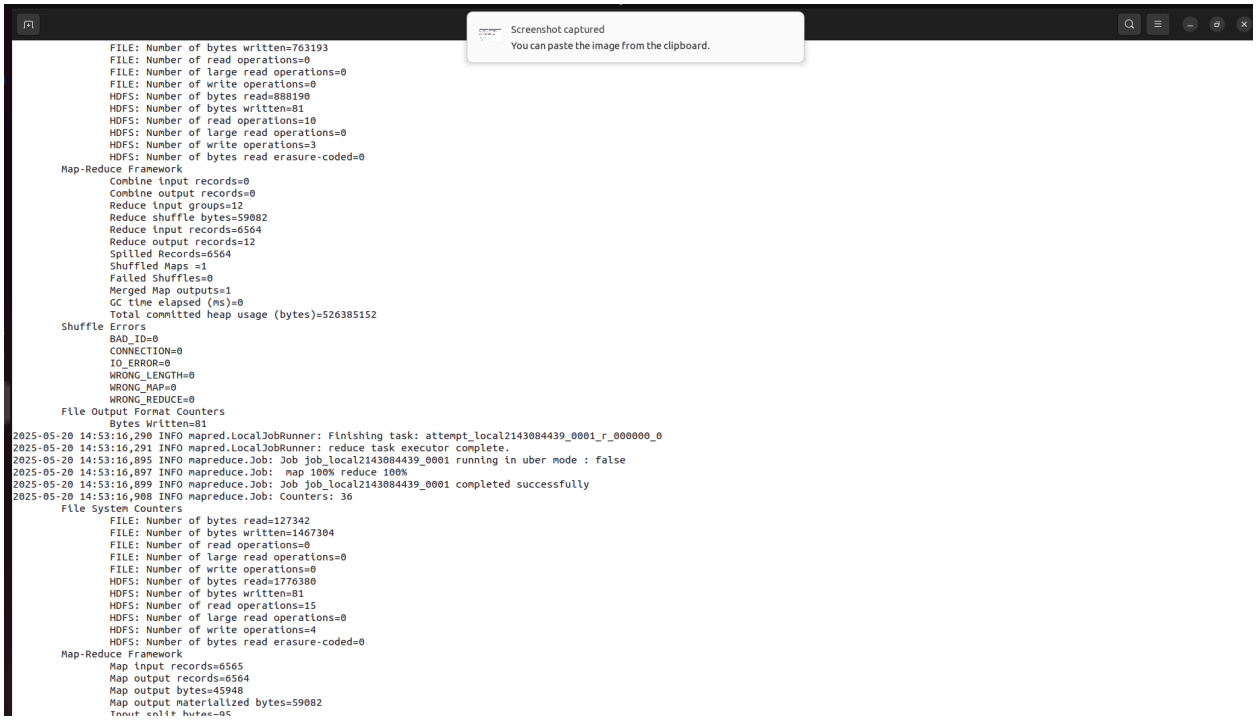



```
            applications, not this command.
jnipath     prints the java.library.path
kdiag       Diagnose Kerberos Problems
kerbname    show auth_to_local principal conversion
key         manage keys via the KeyProvider
rumenfolder scale a rumen input trace
rumentrace  convert logs into a rumen trace
s3guard     S3 Commands
trace       view and modify Hadoop tracing settings
version     print the version

    Daemon Commands:

kms         run KMS, the Key Management Server
registrydns   run the registry DNS server

SUBCOMMAND may print help when invoked w/o parameters or with -h.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /omn
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/1901 /omn/1901
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/weatherdatameanmax.jar weatherdatameanmax.MeanMaxDriver /omn/1901 /ren
2025-05-20 14:53:15,576 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-20 14:53:15,615 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-20 14:53:15,615 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-20 14:53:15,679 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-20 14:53:15,737 INFO input.FileInputFormat: Total input files to process : 1
2025-05-20 14:53:15,764 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-20 14:53:15,830 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2143084439_0001
2025-05-20 14:53:15,830 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-20 14:53:15,890 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-20 14:53:15,891 INFO mapreduce.Job: Running job: job_local2143084439_0001
2025-05-20 14:53:15,891 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-20 14:53:15,895 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:53:15,895 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:53:15,895 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:15,896 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-05-20 14:53:15,981 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-20 14:53:15,982 INFO mapred.LocalJobRunner: Starting task: attempt_local2143084439_0001_m_000000_0
2025-05-20 14:53:15,996 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:53:15,996 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:53:15,996 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:16,004 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-20 14:53:16,007 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/omn/1901:0+888190
2025-05-20 14:53:16,044 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-20 14:53:16,044 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-20 14:53:16,044 INFO mapred.MapTask: soft limit at 83886080
2025-05-20 14:53:16,044 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-20 14:53:16,044 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-20 14:53:16,046 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-20 14:53:16,118 INFO mapred.LocalJobRunner:
2025-05-20 14:53:16,119 INFO mapred.MapTask: Starting flush of map output
2025-05-20 14:53:16,119 INFO mapred.MapTask: Spilling map output
2025-05-20 14:53:16,119 INFO mapred.MapTask: bufstart = 0; bufend = 45948; bufvoid = 104857600
2025-05-20 14:53:16,119 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188144(104752576); length = 26253/6553600
2025-05-20 14:53:16,128 INFO mapred.MapTask: Finished spill 0
2025-05-20 14:53:16,133 INFO mapred.Task: Task:attempt_local2143084439_0001_m_000000_0 is done. And is in the process of committing
2025-05-20 14:53:16,135 INFO mapred.LocalJobRunner: map
```




```
2025-05-20 14:53:16,145 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:16,145 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-20 14:53:16,146 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@6b64ffe0
2025-05-20 14:53:16,147 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-20 14:53:16,155 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5829453312, maxSingleShuffleLimit=1457363328, mergeThreshold=3847439360, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-20 14:53:16,156 INFO reduce.EventFetcher: attempt_local2143084439_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-05-20 14:53:16,170 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local2143084439_0001_m_000000_0 decomp: 59078 len: 59082 to MEMORY
2025-05-20 14:53:16,172 INFO reduce.InMemoryMapOutput: Read 59078 bytes from map-output for attempt_local2143084439_0001_m_000000_0
2025-05-20 14:53:16,172 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 59078, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->59078
2025-05-20 14:53:16,173 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2025-05-20 14:53:16,174 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:53:16,174 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2025-05-20 14:53:16,176 INFO mapred.Merger: Merging 1 sorted segments
2025-05-20 14:53:16,176 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 59073 bytes
2025-05-20 14:53:16,180 INFO reduce.MergeManagerImpl: Merged 1 segments, 59078 bytes to disk to satisfy reduce memory limit
2025-05-20 14:53:16,181 INFO reduce.MergeManagerImpl: Merging 1 files, 59082 bytes from disk
2025-05-20 14:53:16,181 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2025-05-20 14:53:16,181 INFO mapred.Merger: Merging 1 sorted segments
2025-05-20 14:53:16,182 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 59073 bytes
2025-05-20 14:53:16,182 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:53:16,212 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2025-05-20 14:53:16,273 INFO mapred.Task: Task:attempt_local2143084439_0001_r_000000_0 is done. And is in the process of committing
2025-05-20 14:53:16,275 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:53:16,275 INFO mapred.Task: Task attempt_local2143084439_0001_r_000000_0 is allowed to commit now
2025-05-20 14:53:16,289 INFO output.FileOutputCommitter: Saved output of task 'attempt_local2143084439_0001_r_000000_0' to hdfs://localhost:9000/ren
2025-05-20 14:53:16,290 INFO mapred.LocalJobRunner: reduce > reduce
2025-05-20 14:53:16,290 INFO mapred.Task: Task 'attempt_local2143084439_0001_r_000000_0' done.
2025-05-20 14:53:16,290 INFO mapred.Task: Final Counters for attempt_local2143084439_0001_r_000000_0: Counters: 30
        File System Counters
                FILE: Number of bytes read=122769
                FILE: Number of bytes written=763193
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=888190
                HDFS: Number of bytes written=81
                HDFS: Number of read operations=10
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=59082
                Reduce input records=6564
                Reduce output records=12
                Spilled Records=6564
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=526385152
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
```


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

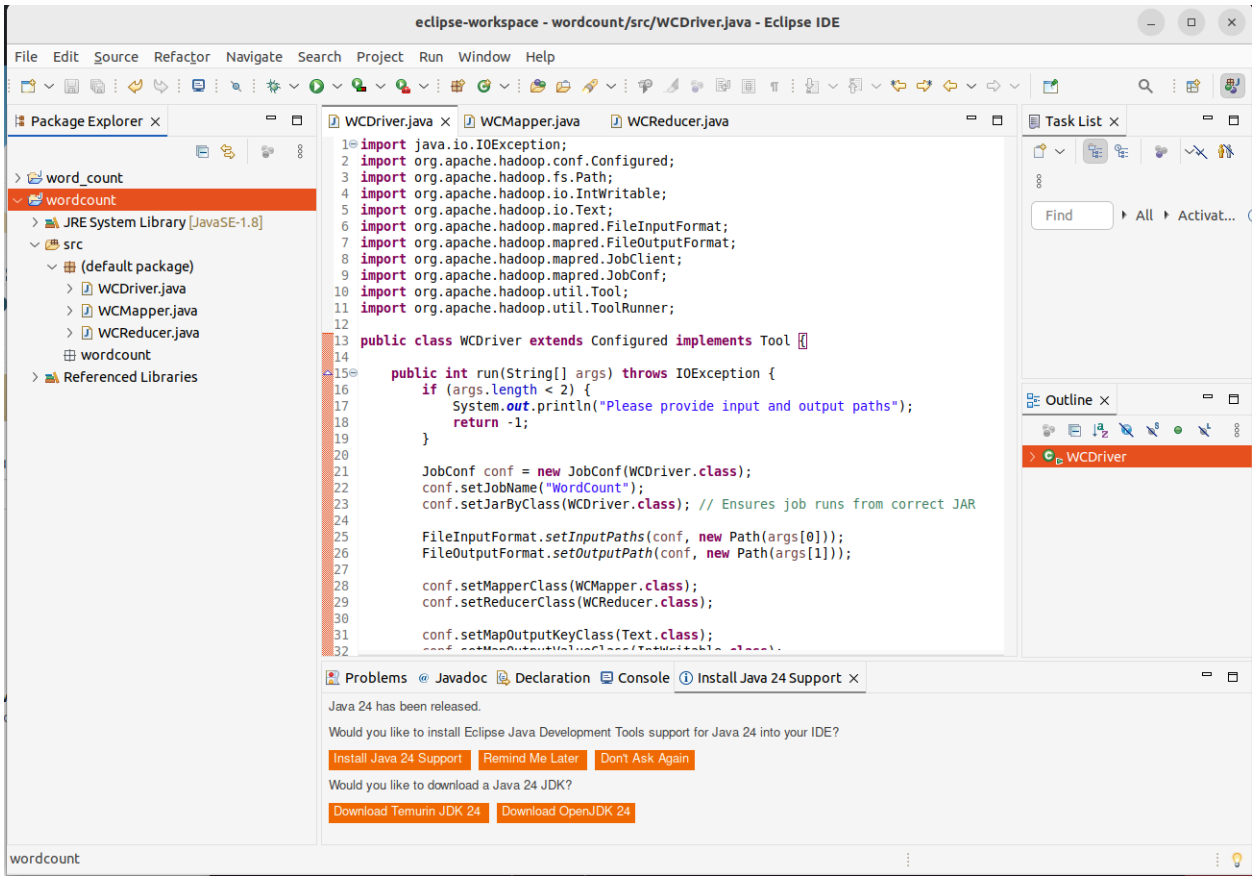

Screenshot captured
You can paste the image from the clipboard.
FILE: Number of bytes written=763193
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=888190
HDFS: Number of bytes written=81
HDFS: Number of read operations=10
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Combine input records=0
Combine output records=0
Reduce input groups=12
Reduce shuffle bytes=59082
Reduce input records=6564
Reduce output records=12
Spilled Records=6564
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=526385152
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
Bytes Written=81
2025-05-20 14:53:16,290 INFO mapred.LocalJobRunner: Finishing task: attempt_local2143084439_0001_r_000000_0
2025-05-20 14:53:16,291 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-20 14:53:16,895 INFO mapreduce.Job: Job job_local2143084439_0001 running in uber mode : false
2025-05-20 14:53:16,897 INFO mapreduce.Job: map 100% reduce 100%
2025-05-20 14:53:16,899 INFO mapreduce.Job: Job job_local2143084439_0001 completed successfully
2025-05-20 14:53:16,908 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=127342
FILE: Number of bytes written=1467304
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1776380
HDFS: Number of bytes written=81
HDFS: Number of read operations=15
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=6565
Map output records=6564
Map output bytes=45948
Map output materialized bytes=59082


hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
FILE: Number of bytes written=1467304
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1776380
HDFS: Number of bytes written=81
HDFS: Number of read operations=15
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=6565
Map output records=6564
Map output bytes=45948
Map output materialized bytes=59082
Input split bytes=95
Combine input records=0
Combine output records=0
Reduce input groups=12
Reduce shuffle bytes=59082
Reduce input records=6564
Reduce output records=12
Spilled Records=13128
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=1052770304
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=888190
File Output Format Counters
Bytes Written=81
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /omn/part-r-00000
cat: `/omn/part-r-00000': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /ren/part-r-00000
01 -13
02 -66
03 -15
04 43
05 100
06 168
07 219
08 198
09 141
10 100
11 1
12 -61
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$

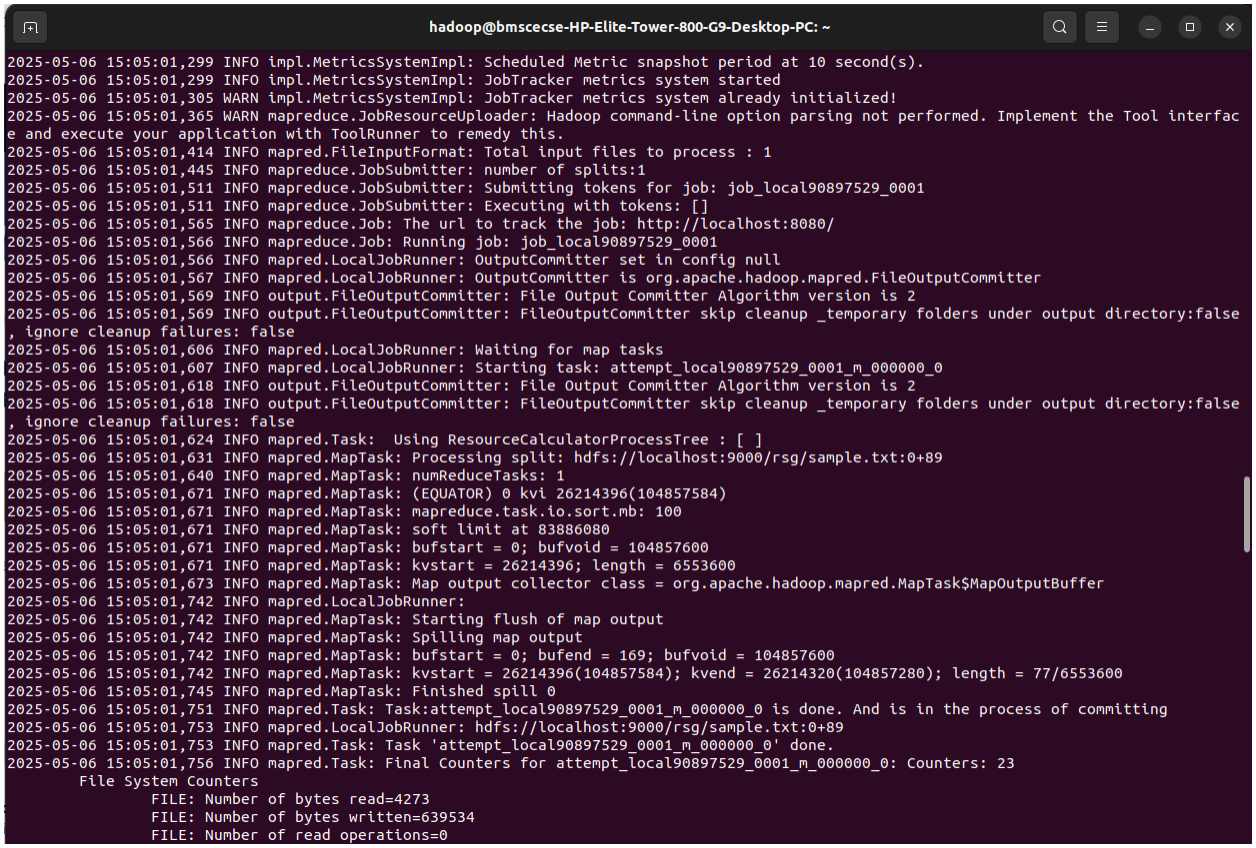
Lab 6: Hadoop

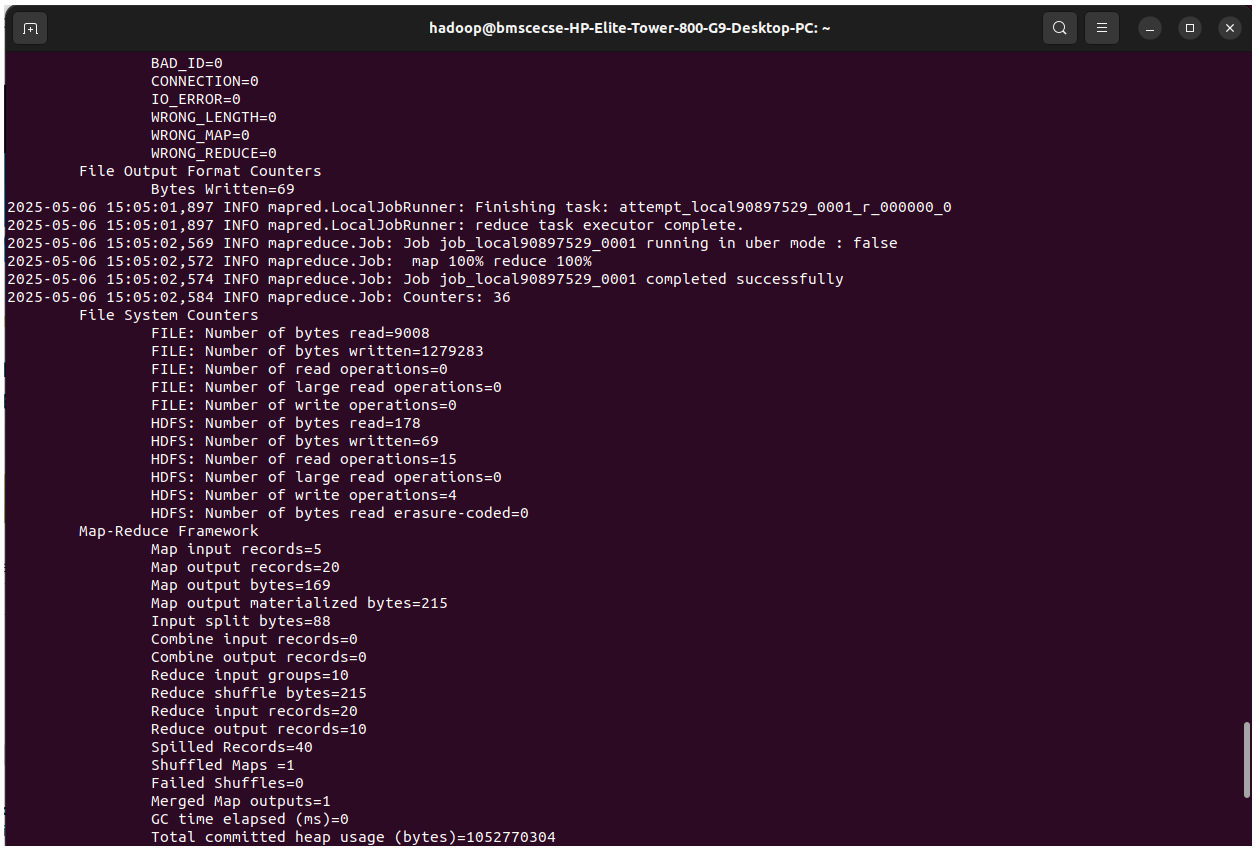
Question: Implement Wordcount program on Hadoop framework

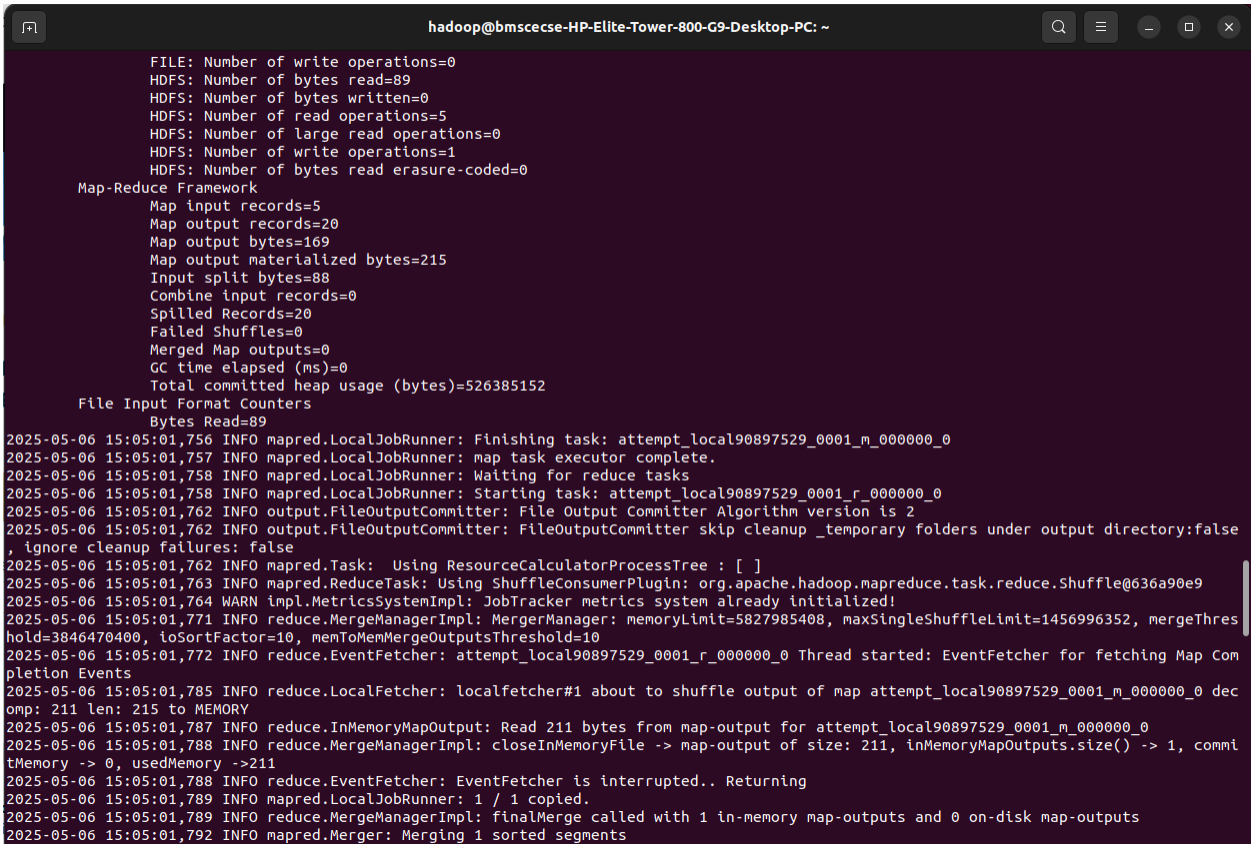
Code with Output:

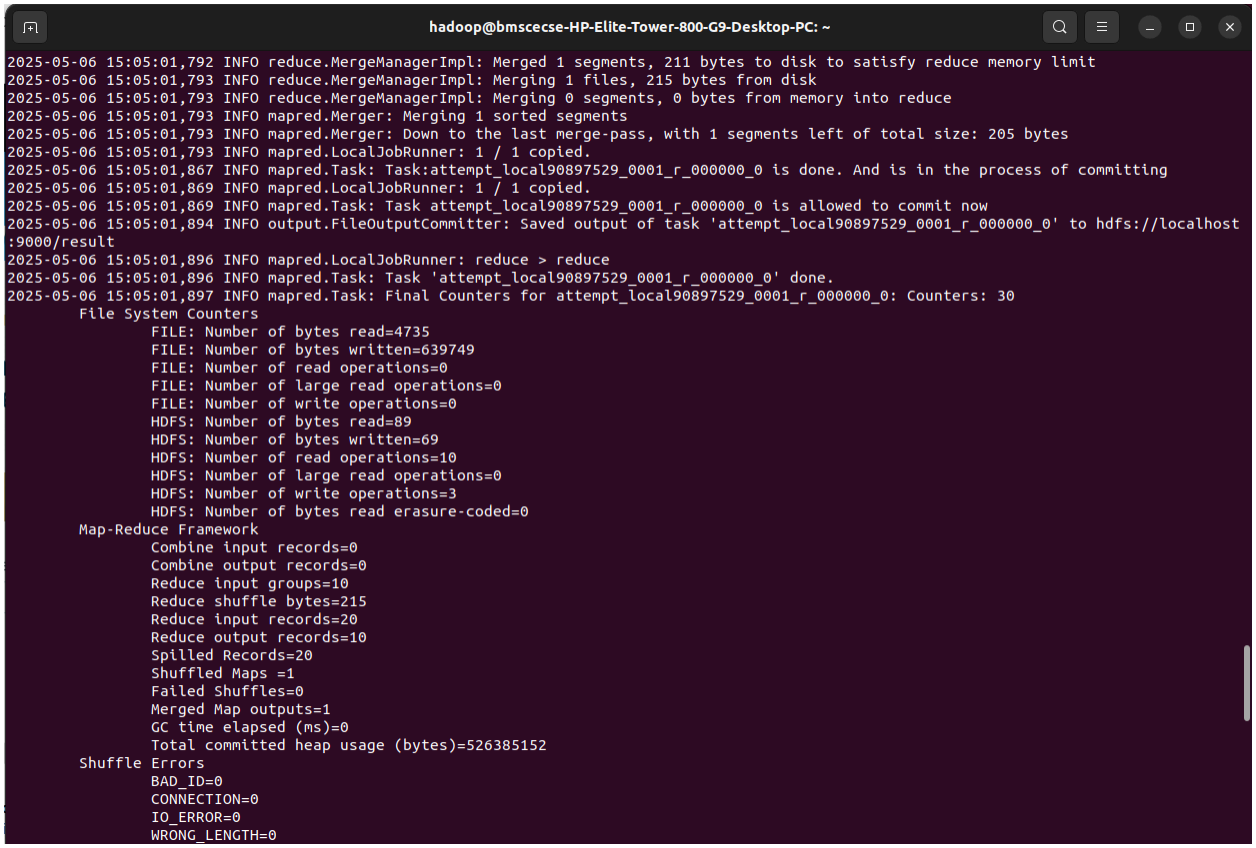


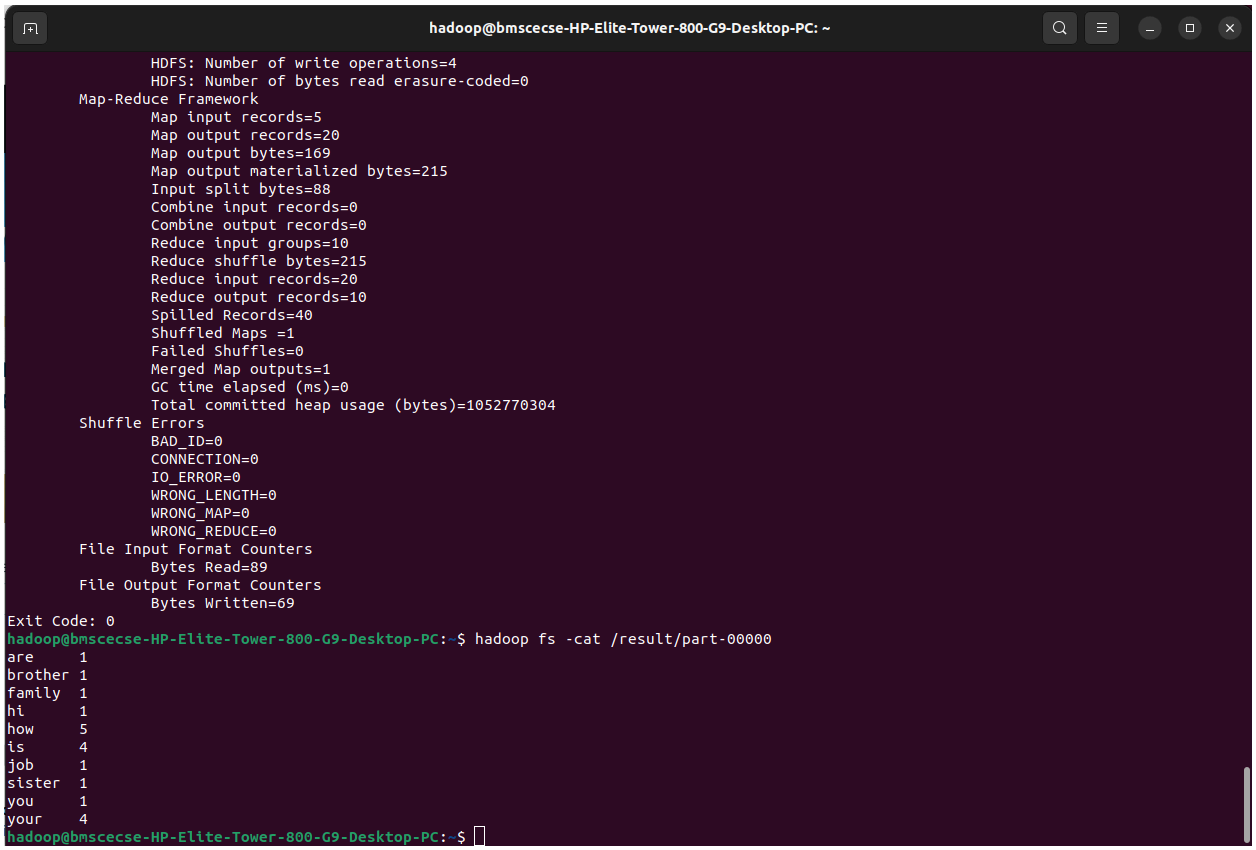

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /rsg
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/sample.txt /rsg/sample.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/wordcount.jar WCDriver /rsg/sample.txt /result
2025-05-06 15:05:01,260 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-06 15:05:01,299 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-06 15:05:01,299 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-06 15:05:01,305 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-06 15:05:01,365 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interfac
e and execute your application with ToolRunner to remedy this.
2025-05-06 15:05:01,414 INFO mapred.FileInputFormat: Total input files to process : 1
2025-05-06 15:05:01,445 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-06 15:05:01,511 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local90897529_0001
2025-05-06 15:05:01,511 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-06 15:05:01,565 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-06 15:05:01,566 INFO mapreduce.Job: Running job: job_local90897529_0001
2025-05-06 15:05:01,566 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-06 15:05:01,567 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2025-05-06 15:05:01,569 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
```


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


hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
Bytes Written=69
2025-05-06 15:05:01,897 INFO mapred.LocalJobRunner: Finishing task: attempt_local90897529_0001_r_000000_0
2025-05-06 15:05:01,897 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-06 15:05:02,569 INFO mapreduce.Job: Job job_local90897529_0001 running in uber mode : false
2025-05-06 15:05:02,572 INFO mapreduce.Job: map 100% reduce 100%
2025-05-06 15:05:02,574 INFO mapreduce.Job: Job job_local90897529_0001 completed successfully
2025-05-06 15:05:02,584 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=9008
FILE: Number of bytes written=1279283
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=178
HDFS: Number of bytes written=69
HDFS: Number of read operations=15
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=5
Map output records=20
Map output bytes=169
Map output materialized bytes=215
Input split bytes=88
Combine input records=0
Combine output records=0
Reduce input groups=10
Reduce shuffle bytes=215
Reduce input records=20
Reduce output records=10
Spilled Records=40
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=1052770304


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

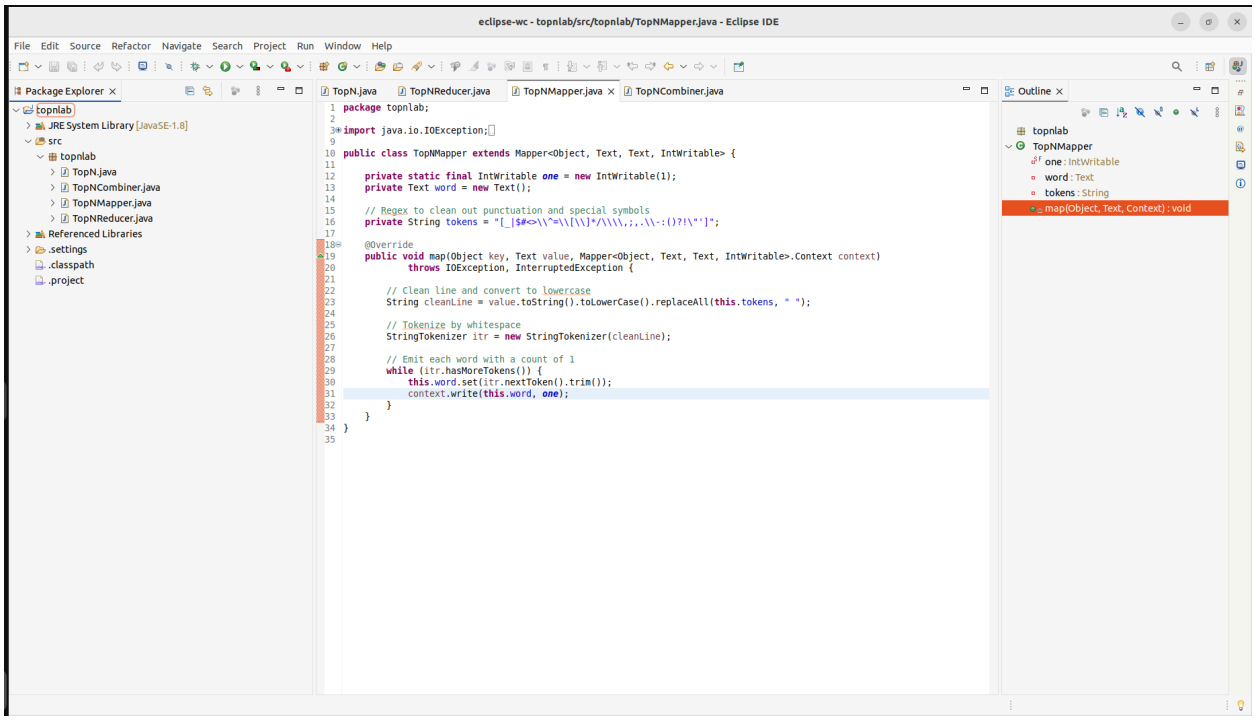

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
2025-05-06 15:05:01,792 INFO reduce.MergeManagerImpl: Merged 1 segments, 211 bytes to disk to satisfy reduce memory limit
2025-05-06 15:05:01,793 INFO reduce.MergeManagerImpl: Merging 1 files, 215 bytes from disk
2025-05-06 15:05:01,793 INFO reduce.MergeManagerImpl: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

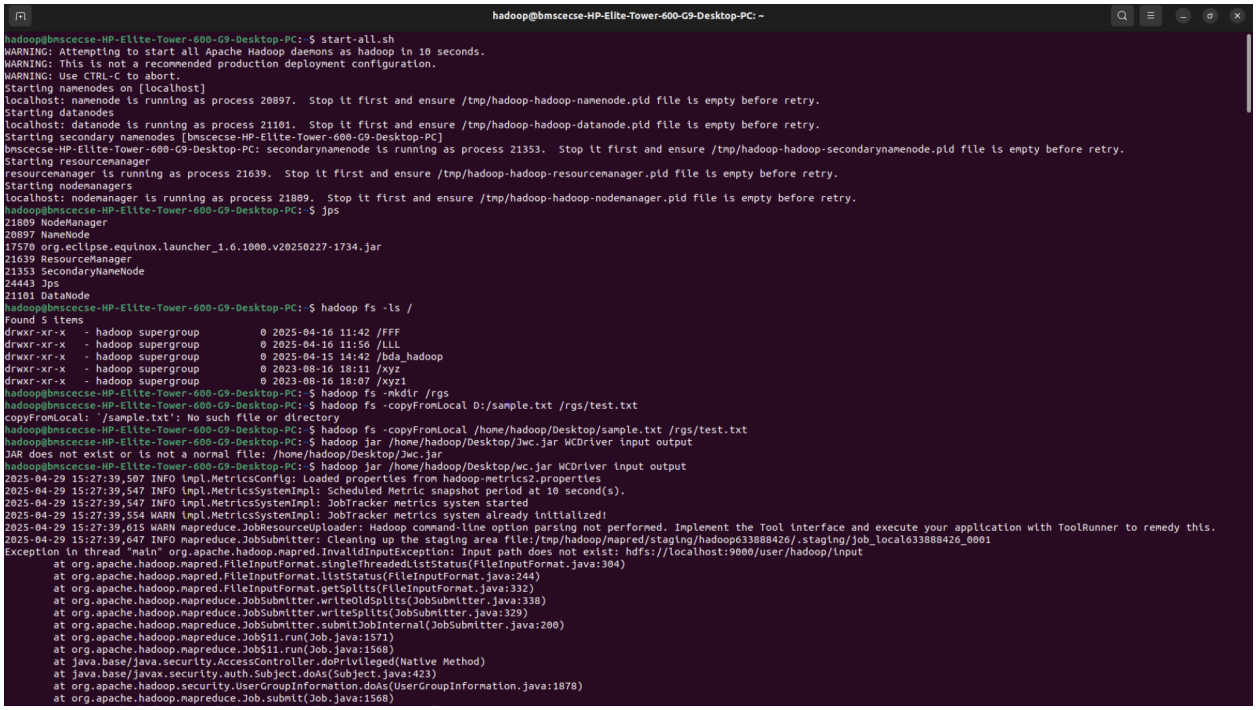

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
HDFS: Number of write operations=4
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=5
Map output records=20
Map output bytes=169
Map output materialized bytes=215
Input split bytes=88
Combine input records=0
Combine output records=0
Reduce input groups=10
Reduce shuffle bytes=215
Reduce input records=20
Reduce output records=10
Spilled Records=40
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=1052770304
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=89
File Output Format Counters
Bytes Written=69
Exit Code: 0
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /result/part-00000
are 1
brother 1
family 1
hi 1
how 5
is 4
job 1
sister 1
you 1
your 4
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$

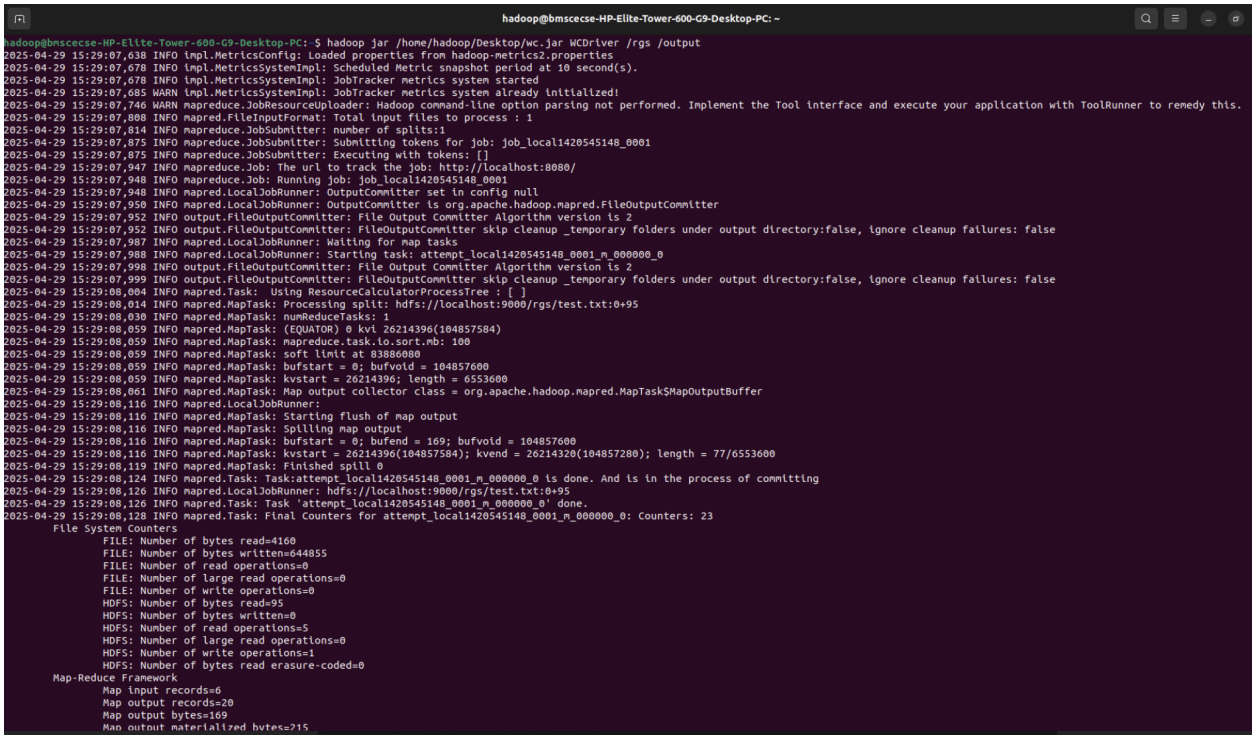
Lab 7: Hadoop

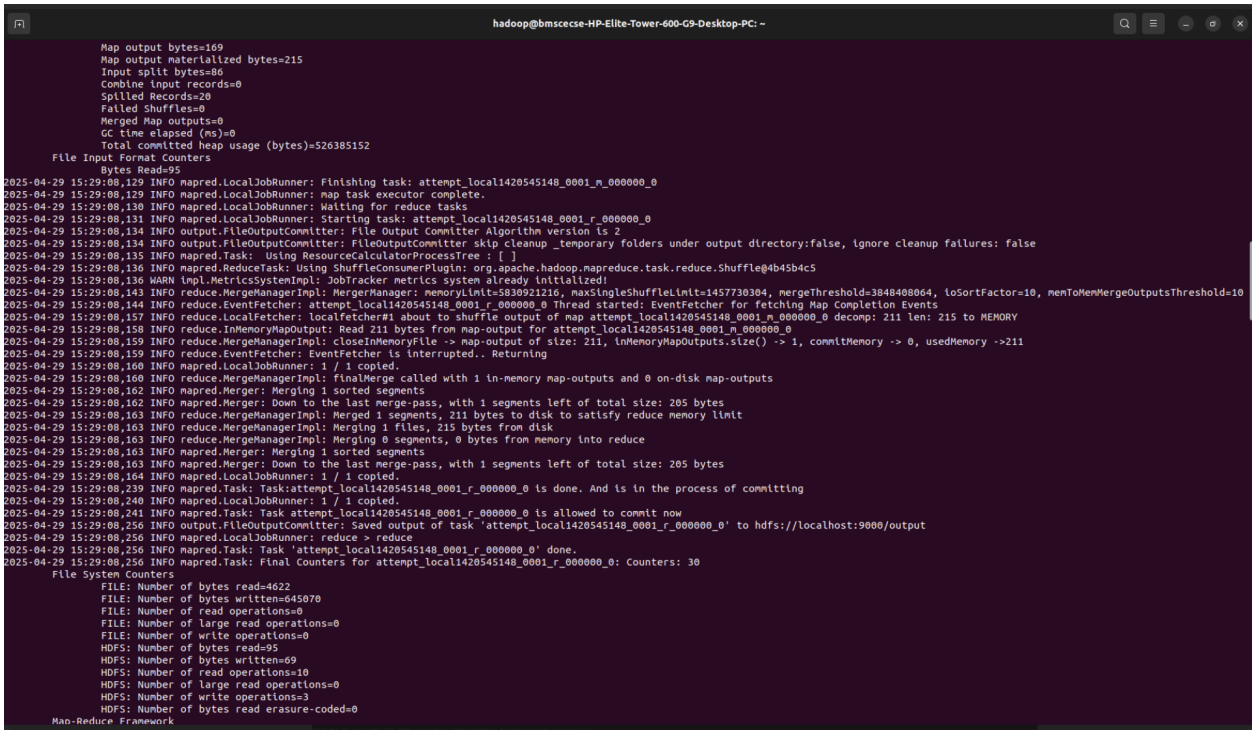
Question: For a given Text file, Create a Map Reduce program to sort the content in an alphabetic order listing only top 10 maximum occurrences of words.

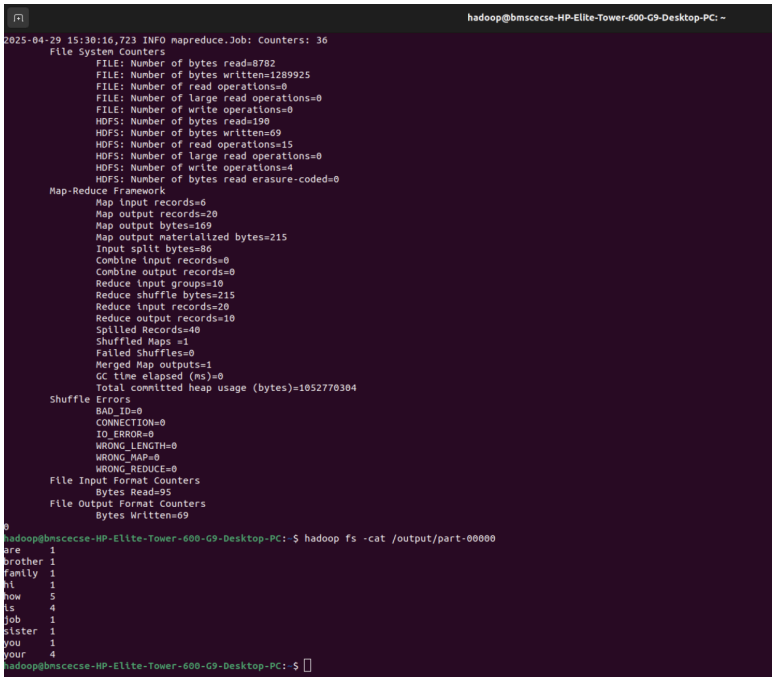
Code with Output:




hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 20897. Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 21101. Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-600-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: secondarynamenode is running as process 21353. Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
resourcemanager is running as process 21639. Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 21809. Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ jps
21809 NodeManager
20897 NameNode
17570 org.eclipse.equinox.launcher_1.6.1000.v20250227-1734.jar
21639 ResourceManager
21353 SecondaryNameNode
24443 Jps
21101 DataNode
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -ls /
Found 5 items
drwxr-xr-x   - hadoop supergroup   0 2025-04-16 11:42 /FFF
drwxr-xr-x   - hadoop supergroup   0 2025-04-16 11:56 /LLL
drwxr-xr-x   - hadoop supergroup   0 2025-04-15 14:42 /bda_hadoop
drwxr-xr-x   - hadoop supergroup   0 2023-08-16 18:11 /xyz
drwxr-xr-x   - hadoop supergroup   0 2023-08-16 18:07 /xyz1
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -mkdir /rgs
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -copyFromLocal D:/sample.txt /rgs/test.txt
copyFromLocal: `/sample.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/sample.txt /rgs/test.txt
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/Jwc.jar WCDriver input output
JAR does not exist or is not a normal file: /home/hadoop/Desktop/Jwc.jar
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/wc.jar WCDriver input output
2025-04-29 15:27:39,507 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-04-29 15:27:39,547 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-04-29 15:27:39,547 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-04-29 15:27:39,554 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-29 15:27:39,615 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-04-29 15:27:39,647 INFO mapreduce.JobSubmitter: Cleaning up the staging area file:/tmp/hadoop/mapred/staging/hadoop633888426/.staging/job_local633888426_0001
Exception in thread "main" org.apache.hadoop.mapred.InvalidInputException: Input path does not exist: hdfs://localhost:9000/user/hadoop/input
at org.apache.hadoop.mapred.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:304)
at org.apache.hadoop.mapred.FileInputFormat.listStatus(FileInputFormat.java:244)
at org.apache.hadoop.mapred.FileInputFormat.getSplits(FileInputFormat.java:332)
at org.apache.hadoop.mapreduce.JobSubmitter.writeOldSplits(JobSubmitter.java:338)
at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java:329)
at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:200)
at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1571)
at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1568)
at java.base/java.security.AccessController.doPrivileged(Native Method)
at java.base/javax.security.auth.Subject.doAs(Subject.java:423)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1878)
at org.apache.hadoop.mapreduce.Job.submit(Job.java:1568)


hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/wc.jar WCDriver /rgs /output
2025-04-29 15:29:07,638 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-04-29 15:29:07,678 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-04-29 15:29:07,678 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-04-29 15:29:07,685 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-29 15:29:07,746 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-04-29 15:29:07,808 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-29 15:29:07,814 INFO mapreduce.JobSubmitter: number of splits:1
2025-04-29 15:29:07,875 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1420545148_0001
2025-04-29 15:29:07,875 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-29 15:29:07,947 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-04-29 15:29:07,948 INFO mapreduce.Job: Running job: job_local1420545148_0001
2025-04-29 15:29:07,948 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-04-29 15:29:07,950 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2025-04-29 15:29:07,952 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-04-29 15:29:07,952 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-04-29 15:29:07,987 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-04-29 15:29:07,988 INFO mapred.LocalJobRunner: Starting task: attempt_local1420545148_0001_m_000000_0
2025-04-29 15:29:07,998 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-04-29 15:29:07,999 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-04-29 15:29:08,004 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-04-29 15:29:08,014 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/rgs/test.txt:0+95
2025-04-29 15:29:08,030 INFO mapred.MapTask: numReduceTasks: 1
2025-04-29 15:29:08,059 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-04-29 15:29:08,059 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-04-29 15:29:08,059 INFO mapred.MapTask: soft limit at 83886080
2025-04-29 15:29:08,059 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-04-29 15:29:08,059 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-04-29 15:29:08,061 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-04-29 15:29:08,116 INFO mapred.LocalJobRunner:
2025-04-29 15:29:08,116 INFO mapred.MapTask: Starting flush of map output
2025-04-29 15:29:08,116 INFO mapred.MapTask: Spilling map output
2025-04-29 15:29:08,116 INFO mapred.MapTask: bufstart = 0; bufend = 169; bufvoid = 104857600
2025-04-29 15:29:08,116 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214320(104857280); length = 77/6553600
2025-04-29 15:29:08,119 INFO mapred.MapTask: Finished spill 0
2025-04-29 15:29:08,124 INFO mapred.Task: Task:attempt_local1420545148_0001_m_000000_0 is done. And is in the process of committing
2025-04-29 15:29:08,126 INFO mapred.LocalJobRunner: hdfs://localhost:9000/rgs/test.txt:0+95
2025-04-29 15:29:08,126 INFO mapred.Task: Task 'attempt_local1420545148_0001_m_000000_0' done.
2025-04-29 15:29:08,128 INFO mapred.Task: Final Counters for attempt_local1420545148_0001_m_000000_0: Counters: 23
File System Counters
FILE: Number of bytes read=4160
FILE: Number of bytes written=644855
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=95
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=6
Map output records=20
Map output bytes=169
Map output materialized bytes=215


hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~
Map output bytes=169
Map output materialized bytes=215
Input split bytes=86
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=95
2025-04-29 15:29:08,129 INFO mapred.LocalJobRunner: Finishing task: attempt_local1420545148_0001_m_000000_0
2025-04-29 15:29:08,129 INFO mapred.LocalJobRunner: map task executor complete.
2025-04-29 15:29:08,130 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-04-29 15:29:08,131 INFO mapred.LocalJobRunner: Starting task: attempt_local1420545148_0001_r_000000_0
2025-04-29 15:29:08,134 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-04-29 15:29:08,134 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-04-29 15:29:08,135 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-04-29 15:29:08,136 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@4b45b4c5
2025-04-29 15:29:08,136 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-29 15:29:08,143 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5830921216, maxSingleShuffleLimit=1457730304, mergeThreshold=3848408064, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-04-29 15:29:08,144 INFO reduce.EventFetcher: attempt_local1420545148_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-04-29 15:29:08,157 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1420545148_0001_m_000000_0 decomp: 211 len: 215 to MEMORY
2025-04-29 15:29:08,158 INFO reduce.InMemoryMapOutput: Read 211 bytes from map-output for attempt_local1420545148_0001_m_000000_0
2025-04-29 15:29:08,159 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 211, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->211
2025-04-29 15:29:08,159 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2025-04-29 15:29:08,160 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-04-29 15:29:08,160 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2025-04-29 15:29:08,162 INFO mapred.Merger: Merging 1 sorted segments
2025-04-29 15:29:08,162 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 205 bytes
2025-04-29 15:29:08,163 INFO reduce.MergeManagerImpl: Merged 1 segments, 211 bytes to disk to satisfy reduce memory limit
2025-04-29 15:29:08,163 INFO reduce.MergeManagerImpl: Merging 1 files, 215 bytes from disk
2025-04-29 15:29:08,163 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2025-04-29 15:29:08,163 INFO mapred.Merger: Merging 1 sorted segments
2025-04-29 15:29:08,163 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 205 bytes
2025-04-29 15:29:08,164 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-04-29 15:29:08,239 INFO mapred.Task: Task:attempt_local1420545148_0001_r_000000_0 is done. And is in the process of committing
2025-04-29 15:29:08,240 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-04-29 15:29:08,241 INFO mapred.Task: Task attempt_local1420545148_0001_r_000000_0 is allowed to commit now
2025-04-29 15:29:08,256 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1420545148_0001_r_000000_0' to hdfs://localhost:9000/output
2025-04-29 15:29:08,256 INFO mapred.LocalJobRunner: reduce > reduce
2025-04-29 15:29:08,256 INFO mapred.Task: Task 'attempt_local1420545148_0001_r_000000_0' done.
2025-04-29 15:29:08,256 INFO mapred.Task: Final Counters for attempt_local1420545148_0001_r_000000_0: Counters: 30
File System Counters
FILE: Number of bytes read=4622
FILE: Number of bytes written=645070
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=95
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


hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~
2025-04-29 15:30:16,723 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=8782
FILE: Number of bytes written=1289925
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=190
HDFS: Number of bytes written=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=6
Map output records=20
Map output bytes=169
Map output materialized bytes=215
Input split bytes=86
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=95
File Output Format Counters
Bytes Written=69
0
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: $ hadoop fs -cat /output/part-00000
are 1
brother 1
family 1
hi 1
how 5
is 4
job 1
sister 1
you 1
your 4
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: $

Lab 8:The Scala Interpreter

Install Scala.

- curl -s "https://get.sdkman.io" | bash
- source "$HOME/.sdkman/bin/sdkman-init.sh"
- sdk install scala
- scala -version

Run "Hello World"

Create a Scala File

- nano Hello.scala

Scala Code

- object Hello {

```
def main(args: Array[String]): Unit = {
    println("Hello, Scala!")
}
}
```

Save and Exit Nano

- Ctrl + O to write the file
- Enter to confirm the filename
- Ctrl + X to exit

Compile the Scala Program

- scalac Hello.scala

Run the Program

- scala Hello

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ scala -version
Scala code runner version 2.11.12 -- Copyright 2002-2017, LAMP/EPFL
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ scalac Hello.scala
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ scala Hello
Hello, Scala!
```

Experiment with Scala Basics

- scala

Data types & Variables

- val x: Int = 5
- var y = "Scala"

Operators & Conditionals

- if (x > 3) println("Greater")

Loops

- for (i <- 1 to 5) println(i)

```
scala> :quit
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ scala
Welcome to Scala 2.11.12 (OpenJDK 64-Bit Server VM, Java 11.0.26).
Type in expressions for evaluation. Or try :help.

scala> val x: Int = 5
x: Int = 5

scala> var y = "Scala"
y: String = Scala

scala> if (x > 3) println("Greater")
Greater

scala> for (i <- 1 to 5) println(i)
1
2
3
4
5
```

Work with Functions

- def add(a: Int, b: Int): Int = a + b
- println(add(3, 4))

Collections

- val list = List(1, 2, 3)
- list.foreach(println)
- val map = Map("a" -> 1, "b" -> 2)
- println(map("a"))

```
scala> def add(a: Int, b: Int): Int = a + b
add: (a: Int, b: Int)Int

scala> println(add(3, 4))
7

scala> val list = List(1, 2, 3)
list: List[Int] = List(1, 2, 3)

scala> list.foreach(println)
1
2
3

scala>

scala> val map = Map("a" -> 1, "b" -> 2)
map: scala.collection.immutable.Map[String,Int] = Map(a -> 1, b -> 2)

scala> println(map("a"))
1
```

Object-Oriented Programming

- class Person(name: String) {
  def greet() = println(s"Hello, $name")
  }

- val p = new Person("Alice")

p.greet()

Advanced Features

- **Traits**
  - trait Greeter {
    def greet(): Unit
    }
  - class EnglishGreeter extends Greeter {
    def greet() = println("Hello")
    }
  - val g = new EnglishGreeter()
  - g.greet()
- **Pattern Matching**
  - def describe(x: Any): String = x match {
    case 1 => "One"
    case "two" => "Two"
    case _ => "Something else"
    }

```
scala> class Person(name: String) {
     |   def greet() = println(s"Hello, $name")
     | }
defined class Person

scala> val p = new Person("Alice")
p: Person = Person@79f82fc4

scala> p.greet()
Hello, Alice

scala> trait Greeter {
     |   def greet(): Unit
     | }
defined trait Greeter

scala>

scala> class EnglishGreeter extends Greeter {
     |   def greet() = println("Hello")
     | }
defined class EnglishGreeter

scala>

scala> val g = new EnglishGreeter()
g: EnglishGreeter = EnglishGreeter@64c79b69

scala> g.greet()
Hello

scala> def describe(x: Any): String = x match {
     |   case 1 => "One"
     |   case "two" => "Two"
     |   case _ => "Something else"
     | }
describe: (x: Any)String
```

Lab 9: Scala

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

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

Lab 10: Spark

Question: Using RDD and FlatMap count how many times each word appears in a file and write out a list of words whose count is strictly greater than 4 using Spark.

Code with Output:

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

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

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

scala>

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

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

scala>    .map(word => (word, 1))
```

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

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

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

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

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

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

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

scala>

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

scala>

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

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

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

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

scala>
```