VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka.

LAB REPORT on

BIG DATA ANALYTICS (23CS6PCBDA)

Submitted by

CHANDRAKALA K M (1BM23CS403)

in partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

B.M.S. COLLEGE OF ENGINEERING BENGALURU-560019 Mar-2025 to June-2025

(Autonomous Institution under VTU)

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 **CHANDRAKALA K M (1BM23CS403),** who is a bonafide student of **B. M. S. College of Engineering.** It is in partial fulfilment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visvesvaraya Technological University, Belgaum during the year 2025. The Lab report has been approved as it satisfies the academic requirements in respect of a **Big Data Analytics- (23CS6PCBDA)** work prescribed for the said degree.

Spoorthi D M
Assistant Professor
Department of CSE
BMSCE, Bengaluru

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

`

Index Sheet

LAB 1 - MongoDB- CRUD Operations Demonstration (Practice and Self Study)

```
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.insertOne({_id: 1, StudName: "MichelleJacintha",
 Grade: "VII", Hobbies: "InternetSurfing"})
{ acknowledged: true, insertedId: 1 }
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.updateOne({_id: 2, StudName: 'AryanDavid', Grade
: 'VII'}, {$set: {Hobbies: "Skating"}}, {upsert: true})
{
  acknowledged: true,
  insertedId: 2,
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 1
}
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.insertMany([{_id: 3, StudName: 'Charan', Grade:
'VII'}, {_id: 4, StudName: 'Vibinn', Grade: 'VII'}])
{ acknowledged: true, insertedIds: { '0': 3, '1': 4 } }
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.updateOne({_id: 3}, {$set: {Hobbies: 'Drawing'}}
)
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
```

```
Atlas atlas-ru5tdz-shard-0 [primary] test> use myDB
switched to db myDB
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db
myDB
Atlas atlas-ru5tdz-shard-0 [primary] myDB> show dbs
admin  232.00 KiB
local   18.01 GiB
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.show()
TypeError: db.show is not a function
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.show
myDB.show
Atlas atlas-ru5tdz-shard-0 [primary] myDB> show collections

Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.createCollection('Student')
{ ok: 1 }
Atlas atlas-ru5tdz-shard-0 [primary] myDB> show collections
Student
```

```
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.updateOne({_id: 2, StudName: 'Charan', Grade: 'V
II'}, {$set: {Hobbies: 'Drawing'}}, {upsert: false})
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 0
}
```

```
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.updateMany({_id: 4}, {$set: {Hobbies: 'Drawing'}
})
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 0,
  upsertedCount: 0
}
```

```
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.deleteOne({_id: 1})
{ acknowledged: true, deletedCount: 1 }
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.deleteMany({Hobbies: 'Drawing'})
{ acknowledged: true, deletedCount: 2 }
Atlas atlas-ru5tdz-shard-0 [primary] myDB>
```

```
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.find()
[
  {
    _id: 1,
    StudName: 'MichelleJacintha',
    Grade: 'VII',
    Hobbies: 'InternetSurfing'
  },
  { _id: 2, Grade: 'VII', StudName: 'AryanDavid', Hobbies: 'Skating' },
  { _id: 3, StudName: 'Charan', Grade: 'VII', Hobbies: 'Drawing' },
  { _id: 4, StudName: 'Vibinn', Grade: 'VII', Hobbies: 'Drawing' }
]
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.find({StudName: 'DavidAryan'})

Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.find({StudName: 'AryanDavid'})
[
  { _id: 2, Grade: 'VII', StudName: 'AryanDavid', Hobbies: 'Skating' }
]
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.find({StudName: 'AryanDavid'},{_id: 0})
[ { Grade: 'VII', StudName: 'AryanDavid', Hobbies: 'Skating' } ]
Atlas atlas-ru5tdz-shard-0 [primary] myDB> db.Student.find({Grade: {$eq :'VII'}})
[
  {
    _id: 1,
    StudName: 'MichelleJacintha',
    Grade: 'VII',
    Hobbies: 'InternetSurfing'
  },
  { _id: 2, Grade: 'VII', StudName: 'AryanDavid', Hobbies: 'Skating' },
  { _id: 3, StudName: 'Charan', Grade: 'VII', Hobbies: 'Drawing' },
  { _id: 4, StudName: 'Vibinn', Grade: 'VII', Hobbies: 'Drawing' }
]
```

```
Microsoft Windows [Version 10.0.22631.4890]
(c) Microsoft Corporation. All rights reserved.

C:\Users\STUDENT>mongosh "mongodb+srv://cluster0.neu24.mongodb.net/" --apiVersion 1 --username charancs22
Enter password: *********
Current Mongosh Log ID: 67cff65547feb36d8a893bf7
Connecting to:          mongodb+srv://<credentials>@cluster0.neu24.mongodb.net/?appName=mongosh+2.3.4
Using MongoDB:          8.0.5 (API Version 1)
Using Mongosh:          2.3.4
mongosh 2.4.2 is available for download: https://www.mongodb.com/try/download/shell

For mongosh info see: https://www.mongodb.com/docs/mongodb-shell/

Atlas atlas-ru5tdz-shard-0 [primary] test> use Student
switched to db Student
Atlas atlas-ru5tdz-shard-0 [primary] Student> db.Student.find()

Atlas atlas-ru5tdz-shard-0 [primary] Student> db.createCollection('Students')
{ ok: 1 }
Atlas atlas-ru5tdz-shard-0 [primary] Student> show collections
Students
Atlas atlas-ru5tdz-shard-0 [primary] Student> Student.drop()
ReferenceError: Student is not defined
Atlas atlas-ru5tdz-shard-0 [primary] Student> db.Student.drop()
true
Atlas atlas-ru5tdz-shard-0 [primary] Student> db.dropDatabase()
{ ok: 1, dropped: 'Student' }
Atlas atlas-ru5tdz-shard-0 [primary] Student> use Customer
switched to db Customer

Atlas atlas-ru5tdz-shard-0 [primary] Student> use Customer
switched to db Customer
Atlas atlas-ru5tdz-shard-0 [primary] Customer> db.createCollection("Customers")
{ ok: 1 }
Atlas atlas-ru5tdz-shard-0 [primary] Customer> show collections
Customers
Atlas atlas-ru5tdz-shard-0 [primary] Customer> db.Customers.insertOne({cust_id: 1, acc_bal: 2000, acc_type: 'X'})
{
  acknowledged: true,
  insertedId: ObjectId('67cffb9d47feb36d8a893bf8')
}
Atlas atlas-ru5tdz-shard-0 [primary] Customer> db.Customers.insertMany([])
MongoInvalidArgumentError: Invalid BulkOperation, Batch cannot be empty
Atlas atlas-ru5tdz-shard-0 [primary] Customer> db.Customers.insertMany([{}])
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67cffbb247feb36d8a893bf9') }
}
Atlas atlas-ru5tdz-shard-0 [primary] Customer> db.Customers.find()
[
  {
    _id: ObjectId('67cffb9d47feb36d8a893bf8'),
    cust_id: 1,
    acc_bal: 2000,
    acc_type: 'X'
  },
  { _id: ObjectId('67cffbb247feb36d8a893bf9') }
]
Atlas atlas-ru5tdz-shard-0 [primary] Customer> db.Customers.deleteMany({cust_id: {$ne: 1}})
{ acknowledged: true, deletedCount: 1 }
```

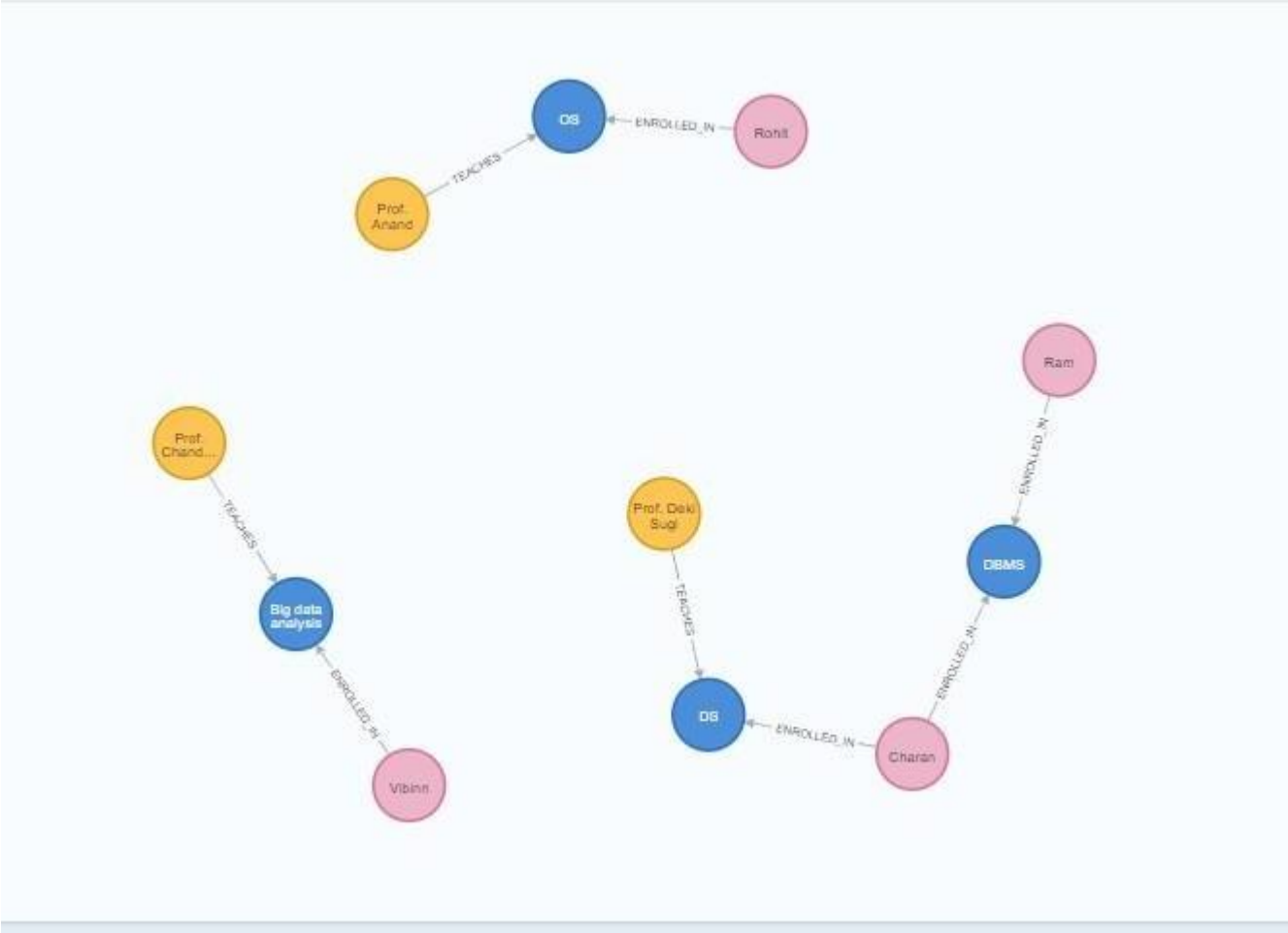
```
Atlas atlas-ru5tdz-shard-0 [primary] Customer> db.Customers.deleteMany({cust_id: {$ne: 1}})
{ acknowledged: true, deletedCount: 1 }
Atlas atlas-ru5tdz-shard-0 [primary] Customer> db.Customers.find()
[
  {
    _id: ObjectId('67cffb9d47feb36d8a893bf8'),
    cust_id: 1,
    acc_bal: 2000,
    acc_type: 'X'
  }
]
```

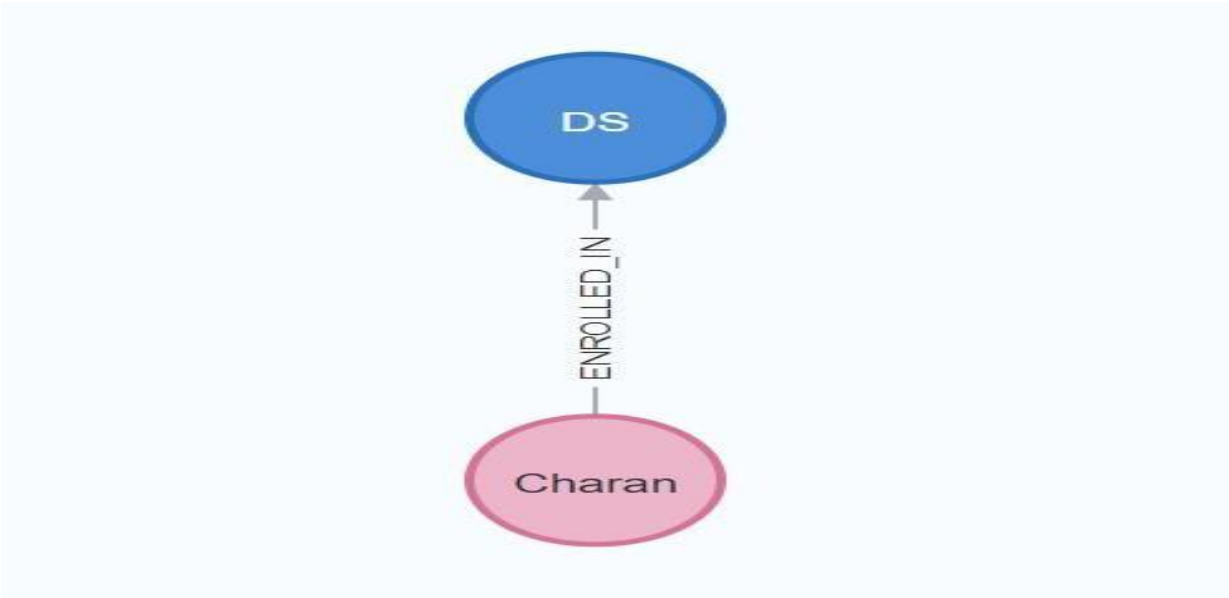
```
Atlas atlas-ru5tdz-shard-0 [primary] Customer> db.Customers.insertMany([{cust_id: 1, acc_bal: 1000, acc_type: 'Y'}, {cust_id: 2, acc_bal: 2000, acc_type: 'Y'}, {cust_id: 2, acc_bal:
100, acc_type: 'Z'}, {cust_id: 3, acc_bal: 500, acc_type: 'Z'}, {cust_id: 3, acc_bal: 3000, acc_type: 'X'}])
{
  acknowledged: true,
  insertedIds: {
    '0': ObjectId('67cffd9547feb36d8a893bfa'),
    '1': ObjectId('67cffd9547feb36d8a893bfb'),
    '2': ObjectId('67cffd9547feb36d8a893bfc'),
    '3': ObjectId('67cffd9547feb36d8a893bfd'),
    '4': ObjectId('67cffd9547feb36d8a893bfe')
  }
}
Atlas atlas-ru5tdz-shard-0 [primary] Customer> db.Customers.find()
[
  {
    _id: ObjectId('67cffb9d47feb36d8a893bf8'),
    cust_id: 1,
    acc_bal: 2000,
    acc_type: 'X'
  },
  {
    _id: ObjectId('67cffd9547feb36d8a893bfa'),
    cust_id: 1,
    acc_bal: 1000,
    acc_type: 'Y'
  },
  {
    _id: ObjectId('67cffd9547feb36d8a893bfb'),
    cust_id: 2,
    acc_bal: 2000,
    acc_type: 'Y'
```

```
Atlas atlas-ru5tdz-shard-0 [primary] Customer> db.accounts.aggregate([
...   {
...     $match: {
...       account_type: 'Z'  // Filter records where account_type is 'Z'
...     }
...   },
...   {
...     $group: {
...       _id: "$customer_id",  // Group by customer_id
...       total_balance: { $sum: "$account_balance" }  // Sum up the account balance for each customer_id
...     }
...   },
...   {
...     $match: {
...       total_balance: { $gt: 1200 }  // Filter for records where the total_balance is greater than 1200
...     }
...   },
...   {
...     $project: {
...       customer_id: "$_id",  // Display the customer_id
...       total_balance: 1,  // Display the total_balance
...       _id: 0  // Exclude the _id field from the final output
...     }
...   }
... ]);
```

```
[
  { _id: 1, min_balance: 1000, max_balancd: 2000 },
  { _id: 2, min_balance: 100, max_balancd: 2000 },
  { _id: 3, min_balance: 500, max_balancd: 3000 }
]
```

Neo 4J DB


DS
ENROLLED_IN
Rohit
TEACHES
Prof. Anand
Ram
Prof. Chand...
Prof. Deki Sugi
DBMS
TEACHES
Big data analysis
ENROLLED_IN
DS
ENROLLED_IN
Charan
Vibinn


DS
ENROLLED_IN
Charan

LAB 2 – CASSANDRA

Perform the following DB operations using Cassandra. a) Create a keyspace by name Employee b) Create a column family by name Employee-Info with attributes Emp_Id Primary Key, Emp_Name, Designation, Date_of_Joining, Salary, Dept_Name c) Insert the values into the table in batch d) Update Employee name and Department of Emp-Id 121 e) Sort the details of Employee records based on salary f) Alter the schema of the table Employee_Info to add a column Projects which stores a set of Projects done by the corresponding Employee. g) Update the altered table to add project names. h) Create a TTL of 15 seconds to display the values of Employees.

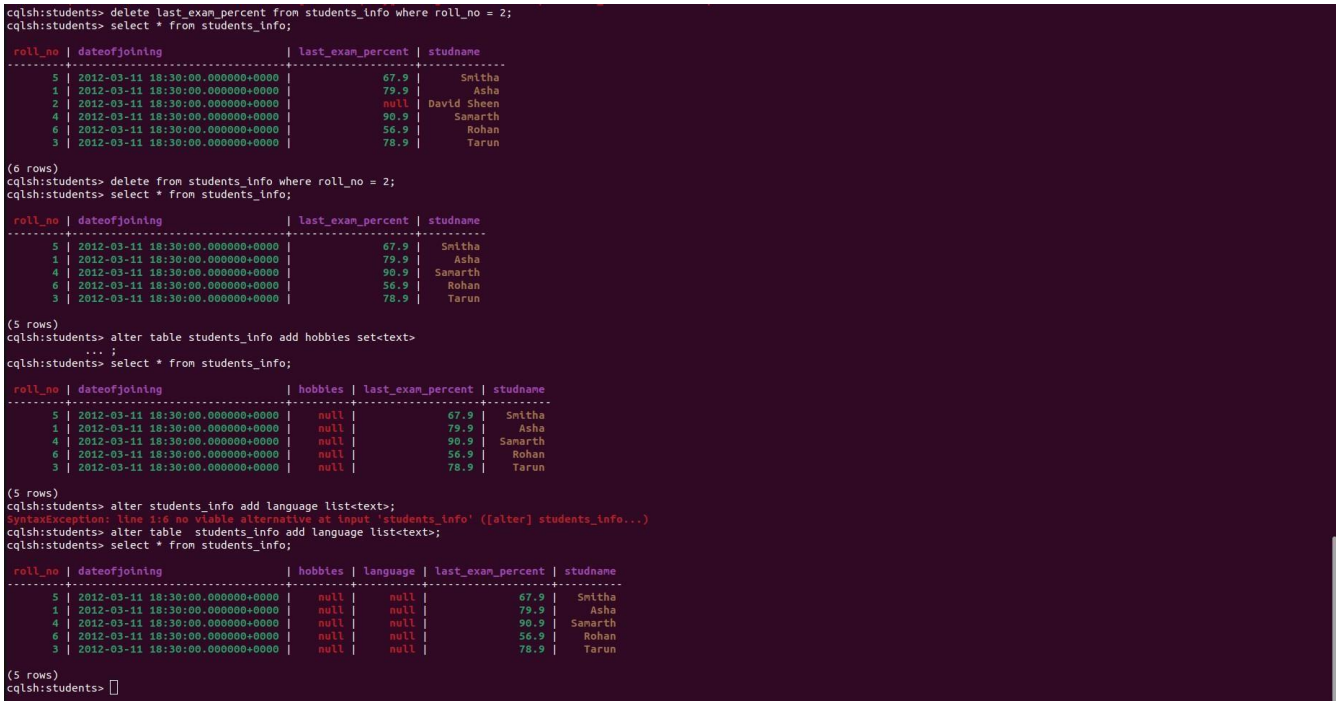
Screenshots:

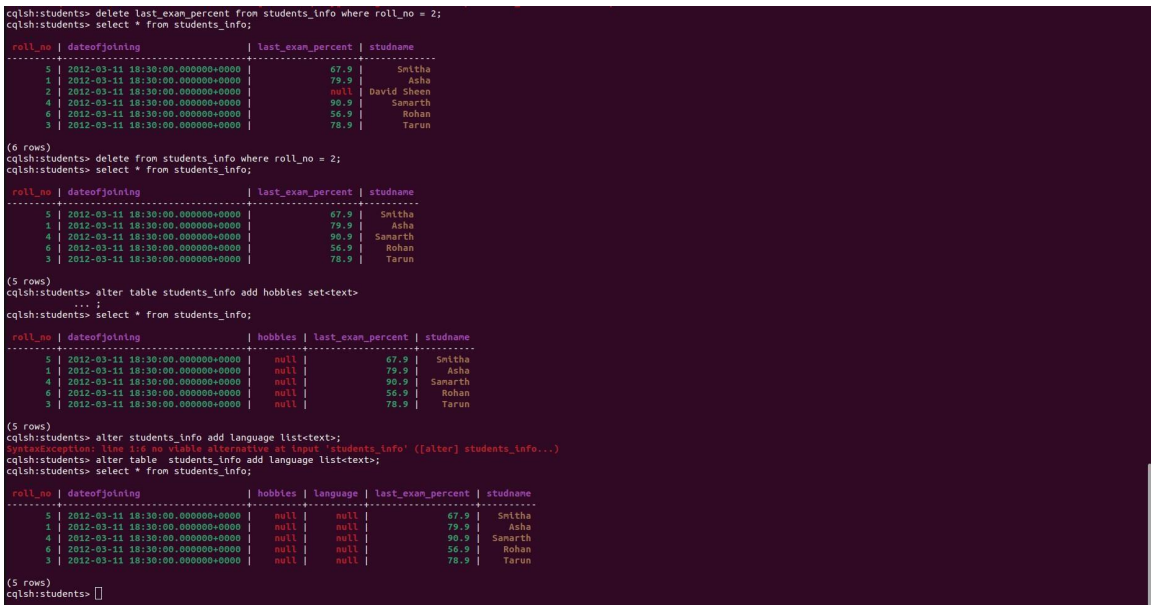
```
cqlsh> DESCRIBE KEYSPACES;

bookstore  employees  system_auth         system_schema  system_views
employee   system     system_distributed  system_traces  system_virtual_schema

cqlsh> SELECT * FROM system_schema.keyspaces;

 keyspace_name      | durable_writes | replication
--------------------+----------------+--------------------------------------------------------------------------------------
          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'}
 system_distributed |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '3'}
             system |           True |                                {'class': 'org.apache.cassandra.locator.LocalStrategy'}
           bookstore |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '1'}
      system_traces |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '2'}
           employee |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '1'}

(8 rows)
cqlsh> CREATE KEYSPACE Students WITH REPLICATION = {'class':'SimpleStrategy','replication_factor':1};
cqlsh> DESCRIBE KEYSPACES;

bookstore  students     system_distributed  system_views
employee   system       system_schema       system_virtual_schema
employees  system_auth  system_traces

cqlsh> SELECT * FROM system_schema.keyspaces;

 keyspace_name      | durable_writes | replication
--------------------+----------------+--------------------------------------------------------------------------------------
          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'}
 system_distributed |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '3'}
             system |           True |                                {'class': 'org.apache.cassandra.locator.LocalStrategy'}
           bookstore |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '1'}
      system_traces |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '2'}
           students |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '1'}
           employee |           True | {'class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication_factor': '1'}
```

```
cqlsh> use students
   ... ;
cqlsh:students> CREATE TABLE Students_Info(Roll_No int PRIMARY KEY, StudName text, DateOfJoining timestamp, last_exam_Percent double);
cqlsh:students> Describe tables

students_info

cqlsh:students> DESCRIBE TABLE students_info

CREATE TABLE students.students_info (
    roll_no int PRIMARY KEY,
    dateofjoining timestamp,
    last_exam_percent double,
    studname text
) WITH additional_write_policy = '99p'
    AND bloom_filter_fp_chance = 0.01
    AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
    AND cdc = false
    AND comment = ''
    AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactionStrategy', 'max_threshold': '32', 'min_threshold': '4'}
    AND compression = {'chunk_length_in_kb': '16', 'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}
    AND memtable = 'default'
    AND crc_check_chance = 1.0
    AND default_time_to_live = 0
    AND extensions = {}
    AND gc_grace_seconds = 864000
    AND max_index_interval = 2048
    AND memtable_flush_period_in_ms = 0
    AND min_index_interval = 128
    AND read_repair = 'BLOCKING'
    AND speculative_retry = '99p';
cqlsh:students>
```

LAB 3 – CASSANDRA

Perform the following DB operations using Cassandra. a) Create a keyspace by name Library b) Create a column family by name Library-Info with attributes Stud_Id Primary Key, Counter_value of type Counter, Stud_Name, Book-Name, Book-Id, Date_of_issue c) Insert the values into the table in batch d) Display the details of the table created and increase the value of the counter e) Write a query to show that a student with id 112 has taken a book “BDA” 2 times. f) Export the created column to a csv file g) Import a given csv dataset from local file system into Cassandra column family





```
cqlsh:students> BEGIN BATCH INSERT INTO Students_Info(Roll_No, StudName, DateOfJoining, last_exam_Percent) VALUES (1,'Asha','2012-03-12',79.9)INSERT INTO Students_Info(Roll_No, StudName, DateOfJoining, la
st_exam_Percent) VALUES (2,'Kiran','2012-03-12',89.9)INSERT INTO Students_Info(Roll_No, StudName, DateOfJoining, last_exam_Percent) VALUES (3,'Tarun','2012-03-12',78.9)INSERT INTO Students_Info(Roll_No, S
tudName, DateOfJoining, last_exam_Percent) VALUES (4,'Samarth','2012-03-12',90.9)INSERT INTO Students_Info(Roll_No, StudName, DateOfJoining, last_exam_Percent) VALUES (5,'Smitha','2012-03-12',67.9)INSERT
INTO Students_Info(Roll_No, StudName, DateOfJoining, last_exam_Percent) VALUES (6,'Rohan','2012-03-12',56.9)APPLY BATCH;
cqlsh:students> select * from students_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 |    Kiran
       4 | 2012-03-11 18:30:00.000000+0000 |              90.9 |  Samarth
       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:students>
```

LAB 4 – HDFS COMMANDS

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

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop version
Hadoop 3.3.6
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da9266a4f88195058f08fd012bf
9c
Compiled by ubuntu on 2023-06-18T08:22Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum 5652179ad55f76cb287d9c633bb53bbd
This command was run using /home/hadoop/hadoop/share/hadoop/common/hadoop-common-3.3.6.jar
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -ls /
ls: Call From bmscecse-HP-Elite-Tower-800-G9-Desktop-PC/127.0.1.1 to localhost:9000 failed on connect
ion exception: java.net.ConnectException: Connection refused; For more details see:  http://wiki.apac
he.org/hadoop/ConnectionRefused
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ ^C
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ jps
5800 Jps
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ ^[[200~start-dfs.sh
start-dfs.sh: command not found
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ jps
5938 ResourceManager
6166 NodeManager
6653 Jps
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 5938.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemana
ger.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 6166.  Stop it first and ensure /tmp/hadoop-hadoop-nodem
anager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ jps
5938 ResourceManager
7731 Jps
6835 NameNode
6166 NodeManager
7263 SecondaryNameNode
6991 DataNode
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:27 /abc
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:38 /bda_hadoop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls -R /
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:27 /abc
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:38 /bda_hadoop
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:40 /bda_hadoop/file.txt
-rw-r--r--   1 hadoop supergroup       9421 2024-05-13 14:40 /bda_hadoop/file.txt/bda_local.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:27 /abc
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:38 /bda_hadoop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd abc
bash: cd: abc: No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd /abc
bash: cd: /abc: No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -touchz /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /abc
Found 1 items
-rw-r--r--   1 hadoop supergroup          0 2025-04-15 14:43 /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ echo "Hello Charan" > /abc/f1.txt
bash: /abc/f1.txt: No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ echo "Hello Charan" > f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -cat /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:43 /abc
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:38 /bda_hadoop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /abc
Found 1 items
-rw-r--r--   1 hadoop supergroup          0 2025-04-15 14:43 /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -cat /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ echo "Hello Charan" > sample.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -put sample.txt /abc/f1.txt
put: `/abc/f1.txt': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -cat /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /abc
Found 1 items
-rw-r--r--   1 hadoop supergroup          0 2025-04-15 14:43 /abc/f1.txt
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -cat /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ echo "Hello Charan" > sample.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -put sample.txt /abc/f1.txt
put: `/abc/f1.txt': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -cat /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /abc
Found 1 items
-rw-r--r--   1 hadoop supergroup          0 2025-04-15 14:43 /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -chmod 644 /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /abc
Found 1 items
-rw-r--r--   1 hadoop supergroup          0 2025-04-15 14:43 /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -put sample.txt /abc/f1.txt
put: `/abc/f1.txt': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -cat /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -put -f sample.txt /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -cat /abc/f1.txt
Hello Charan
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ echo "Charan G" > /abc/f1.txt
bash: /abc/f1.txt: No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -getfacl /abc/
# file: /abc
# owner: hadoop
# group: supergroup
user::rwx
group::r-x
other::r-x

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mv /abc /FFF
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /FF
ls: `/FF': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /FFF
Found 1 items
-rw-r--r--   1 hadoop supergroup         13 2025-04-15 14:50 /FFF/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cp /abc/ /LLL
cp: `/abc/': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cp /abc /LLL
cp: `/abc': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /abc
ls: `/abc': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:50 /FFF
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:38 /bda_hadoop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cp /FFF/ /abc
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls abc
ls: `abc': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /abc
Found 1 items
-rw-r--r--   1 hadoop supergroup         13 2025-04-15 15:13 /abc/f1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$
```

LAB 5 – WORDCOUNT ON HADOOP

Implement Wordcount program on Hadoop framework

SCREENSHOTS:

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
Starting resourcemanager
Starting nodemanagers
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ jps
16928 NodeManager
16482 SecondaryNameNode
16035 NameNode
12837 org.eclipse.equinox.launcher_1.6.1000.v20250227-1734.jar
16202 DataNode
17421 Jps
16766 ResourceManager
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /
Found 5 items
drwxr-xr-x   - hadoop supergroup          0 2025-04-21 11:50 /FFF
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:34 /Hadoop
drwxr-xr-x   - hadoop supergroup          0 2025-04-21 12:22 /LLL
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:23 /abc
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:49 /bda_hadoop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /rgs
```

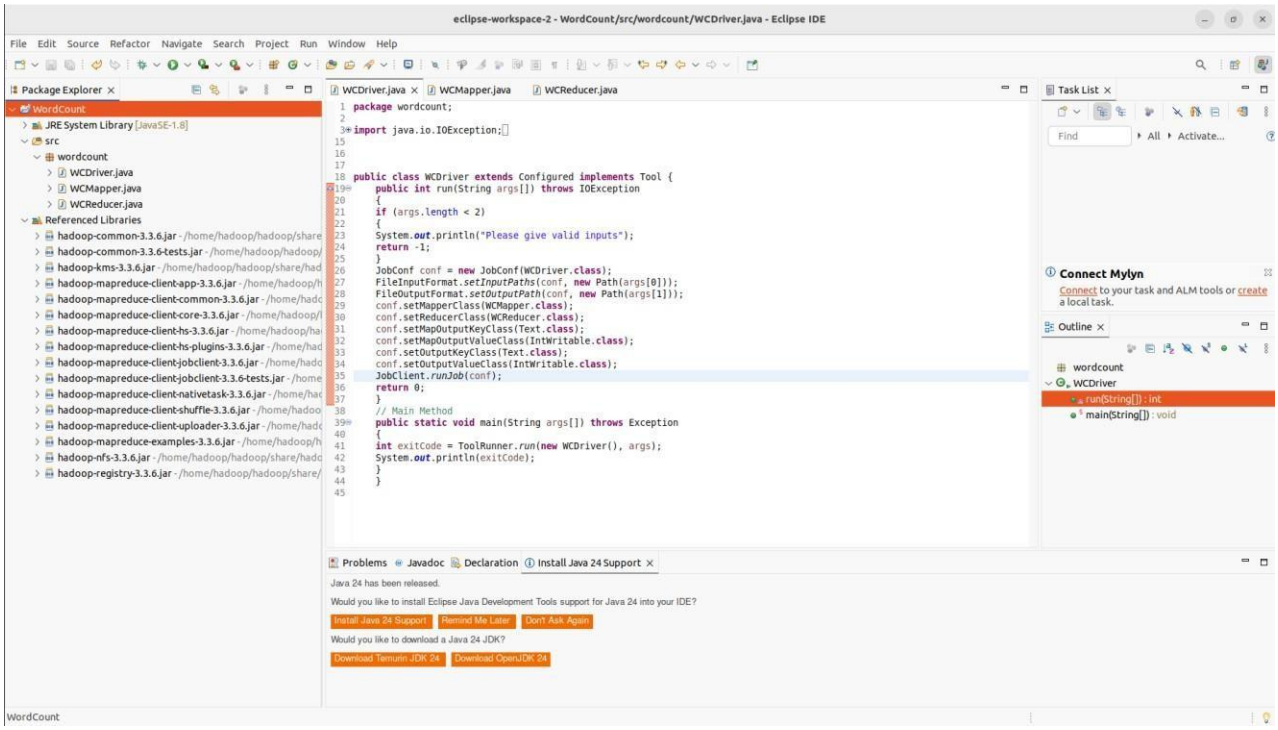
```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /
Found 6 items
drwxr-xr-x   - hadoop supergroup          0 2025-04-21 11:50 /FFF
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:34 /Hadoop
drwxr-xr-x   - hadoop supergroup          0 2025-04-21 12:22 /LLL
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:23 /abc
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:49 /bda_hadoop
drwxr-xr-x   - hadoop supergroup          0 2025-04-29 15:24 /rgs
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ jar tf WordCount3.jar
META-INF/MANIFEST.MF
.classpath
.project
wordcount/
wordcount/WCDriver.class
wordcount/WCReducer.class
wordcount/WCMapper.class
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar WordCount3.jar wordcount.WCDriver /rgs
/test.txt /rgs/output
2025-04-29 15:32:09,761 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager
 at /0.0.0.0:8032
2025-04-29 15:32:09,829 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager
 at /0.0.0.0:8032
2025-04-29 15:32:09,918 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not pe
rformed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-04-29 15:32:09,944 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/h
adoop-yarn/staging/hadoop/.staging/job_1745919848818_0003
2025-04-29 15:32:10,138 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-29 15:32:10,227 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-29 15:32:10,318 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745919848818_000
3
2025-04-29 15:32:10,318 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-29 15:32:10,405 INFO conf.Configuration: resource-types.xml not found
2025-04-29 15:32:10,405 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-29 15:32:10,556 INFO impl.YarnClientImpl: Submitted application application_1745919848818_000
3
2025-04-29 15:32:10,574 INFO mapreduce.Job: The url to track the job: http://bmscecse-HP-Elite-Tower-
800-G9-Desktop-PC:8088/proxy/application_1745919848818_0003/
2025-04-29 15:32:10,575 INFO mapreduce.Job: Running job: job_1745919848818_0003
2025-04-29 15:32:15,652 INFO mapreduce.Job: Job job_1745919848818_0003 running in uber mode : false
2025-04-29 15:32:15,654 INFO mapreduce.Job:  map 0% reduce 0%
2025-04-29 15:32:18,772 INFO mapreduce.Job:  map 100% reduce 0%
2025-04-29 15:32:22,799 INFO mapreduce.Job:  map 100% reduce 100%
2025-04-29 15:32:23,824 INFO mapreduce.Job: Job job_1745919848818_0003 completed successfully
2025-04-29 15:32:23,882 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=215
                FILE: Number of bytes written=829242
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=306
                HDFS: Number of bytes written=69
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=2555
                Total time spent by all reduces in occupied slots (ms)=1281
                Total time spent by all map tasks (ms)=2555
                Total time spent by all reduce tasks (ms)=1281
                Total vcore-milliseconds taken by all map tasks=2555
                Total vcore-milliseconds taken by all reduce tasks=1281
                Total megabyte-milliseconds taken by all map tasks=2616320
                Total megabyte-milliseconds taken by all reduce tasks=1311744
```

```
        Map-Reduce Framework
                Map input records=5
                Map output records=20
                Map output bytes=169
                Map output materialized bytes=221
                Input split bytes=172
                Combine input records=0
                Combine output records=0
                Reduce input groups=10
                Reduce shuffle bytes=221
                Reduce input records=20
                Reduce output records=10
                Spilled Records=40
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=18
                CPU time spent (ms)=1090
                Physical memory (bytes) snapshot=1007276032
                Virtual memory (bytes) snapshot=8417542144
                Total committed heap usage (bytes)=1572864000
                Peak Map Physical memory (bytes)=373477376
                Peak Map Virtual memory (bytes)=2806251520
                Peak Reduce Physical memory (bytes)=266080256
                Peak Reduce Virtual memory (bytes)=2808696832
        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=134
        File Output Format Counters
                Bytes Written=69
0
```

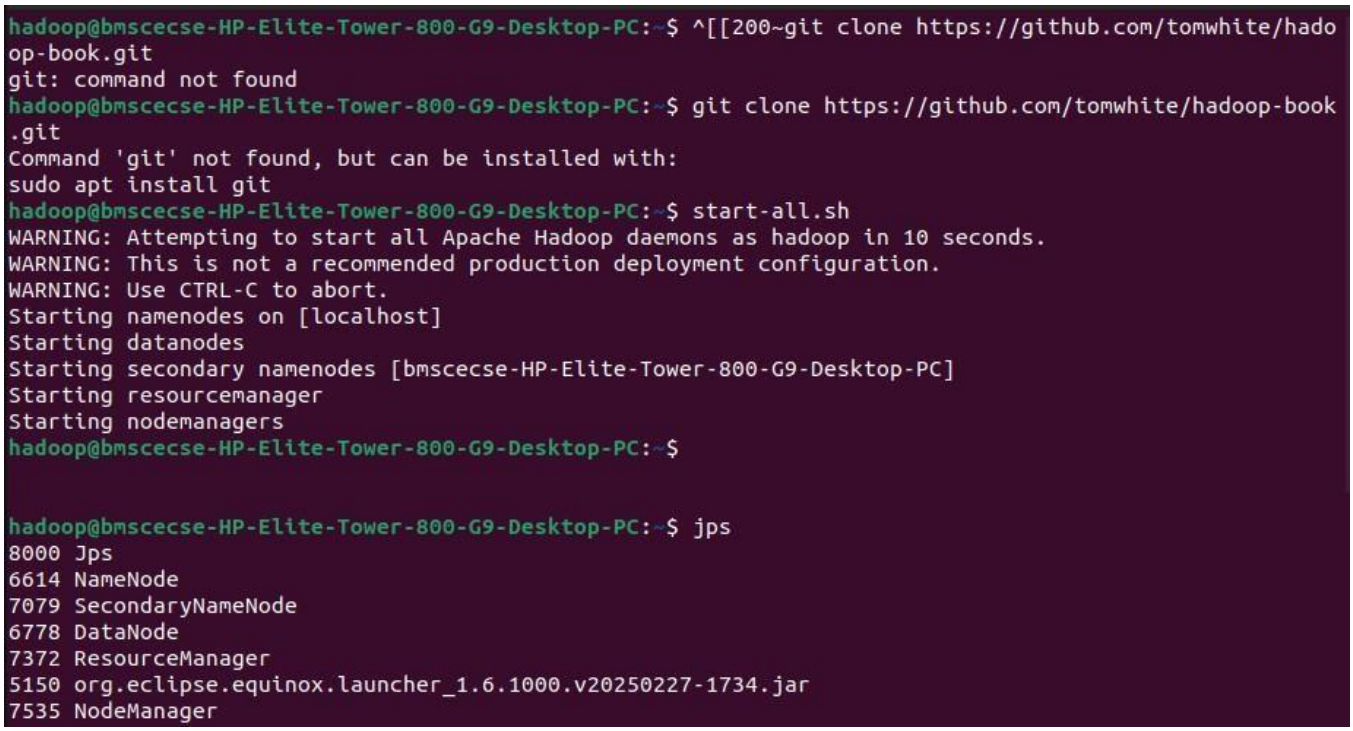
```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /rgs/output
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2025-04-29 15:32 /rgs/output/_SUCCESS
-rw-r--r--   1 hadoop supergroup         69 2025-04-29 15:32 /rgs/output/part-00000
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /rgs/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-800-G9-Desktop-PC:~$
```

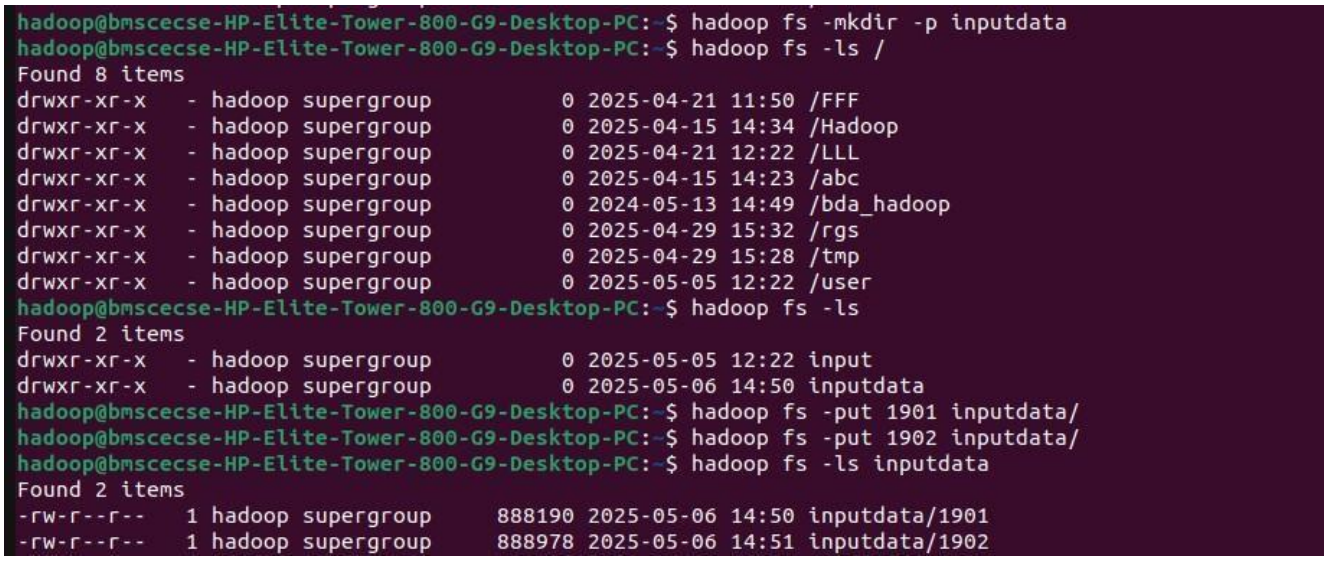

eclipse-workspace-2 - WordCount/src/wordcount/WCDriver.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer
WordCount
JRE System Library [JavaSE-1.8]
src
wordcount
WCDriver.java
WCMapper.java
WCReducer.java
Referenced Libraries
hadoop-common-3.3.6.jar - /home/hadoop/hadoop/share
hadoop-common-3.3.6-tests.jar - /home/hadoop/hadoop/
hadoop-kms-3.3.6.jar - /home/hadoop/hadoop/share/had
hadoop-mapreduce-client-app-3.3.6.jar - /home/hadoop/h
hadoop-mapreduce-client-common-3.3.6.jar - /home/hadc
hadoop-mapreduce-client-core-3.3.6.jar - /home/hadoop/
hadoop-mapreduce-client-hs-3.3.6.jar - /home/hadoop/ha
hadoop-mapreduce-client-hs-plugins-3.3.6.jar - /home/had
hadoop-mapreduce-client-jobclient-3.3.6.jar - /home/hadc
hadoop-mapreduce-client-jobclient-3.3.6-tests.jar - /home
hadoop-mapreduce-client-nativetask-3.3.6.jar - /home/hac
hadoop-mapreduce-client-shuffle-3.3.6.jar - /home/hadoo
hadoop-mapreduce-client-uploader-3.3.6.jar - /home/hadc
hadoop-mapreduce-examples-3.3.6.jar - /home/hadoop/h
hadoop-nfs-3.3.6.jar - /home/hadoop/hadoop/share/hadc
hadoop-registry-3.3.6.jar - /home/hadoop/hadoop/share/
WCDriver.java
WCMapper.java
WCReducer.java
1 package wordcount;
2
3 import java.io.IOException;
15
16
17
18 public class WCDriver extends Configured implements Tool {
19 public int run(String args[]) throws IOException
20 {
21 if (args.length < 2)
22 {
23 System.out.println("Please give valid inputs");
24 return -1;
25 }
26 JobConf conf = new JobConf(WCDriver.class);
27 FileInputFormat.setInputPaths(conf, new Path(args[0]));
28 FileOutputFormat.setOutputPath(conf, new Path(args[1]));
29 conf.setMapperClass(WCMapper.class);
30 conf.setReducerClass(WCReducer.class);
31 conf.setMapOutputKeyClass(Text.class);
32 conf.setMapOutputValueClass(IntWritable.class);
33 conf.setOutputKeyClass(Text.class);
34 conf.setOutputValueClass(IntWritable.class);
35 JobClient.runJob(conf);
36 return 0;
37 }
38 // Main Method
39 public static void main(String args[]) throws Exception
40 {
41 int exitCode = ToolRunner.run(new WCDriver(), args);
42 System.out.println(exitCode);
43 }
44 }
45
Task List
Find
All
Activate...
Connect Mylyn
Connect to your task and ALM tools or create a local task.
Outline
wordcount
WCDriver
run(String[]) : int
main(String[]) : void
Problems
Javadoc
Declaration
Install Java 24 Support
Java 24 has been released.
Would you like to install Eclipse Java Development Tools support for Java 24 into your IDE?
Install Java 24 Support
Remind Me Later
Don't Ask Again
Would you like to download a Java 24 JDK?
Download Temurin JDK 24
Download OpenJDK 24
WordCount

LAB 6 – WEATHER DATA HADOOP

From the following link extract the weather data https://github.com/tomwhite/hadoop book/tree/master/input/ncdc/all a) Create a MapReduce program to find average temperature for each year from NCDC data set. b) find the mean max temperature for every month.

Screenshots:

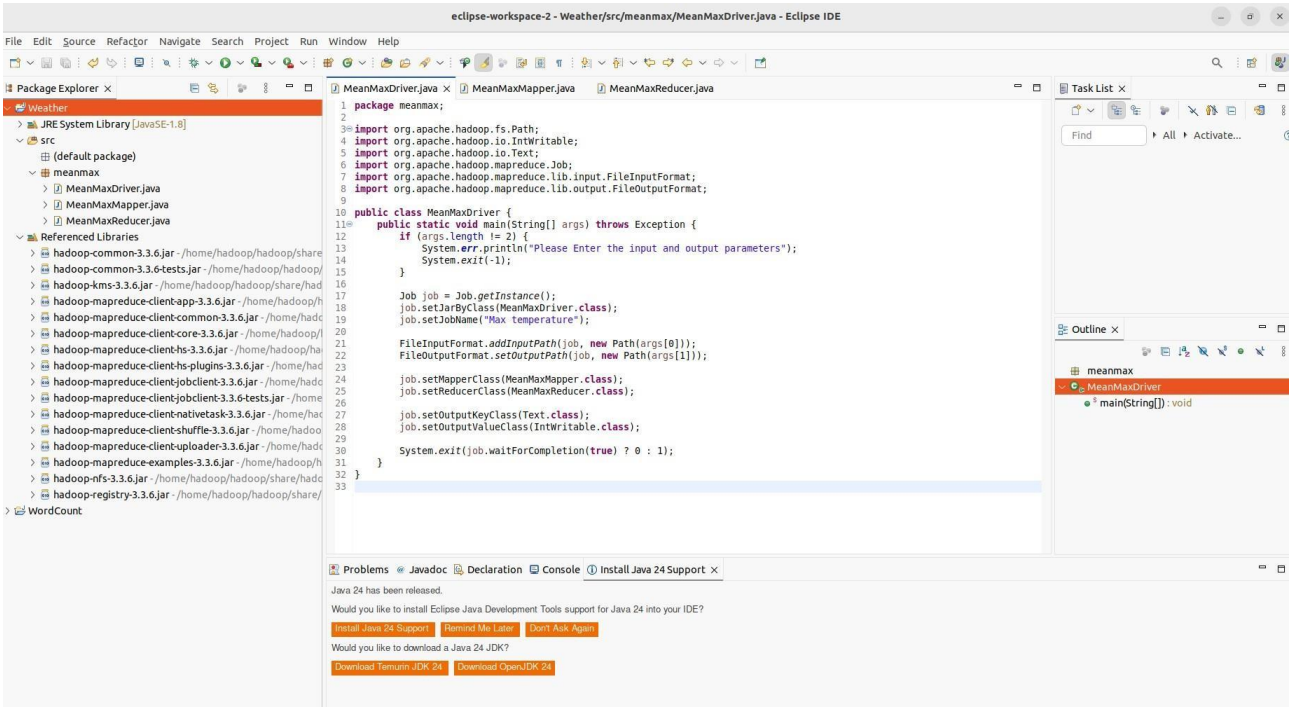




```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: $ hadoop jar Weather.jar meanmax.MeanMaxDriver inputdata outputdata
2025-05-06 14:54:51,797 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-05-06 14:54:52,065 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-06 14:54:52,086 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
hadoop/.staging/job_1746522831414_0001
2025-05-06 14:54:52,277 INFO input.FileInputFormat: Total input files to process : 2
2025-05-06 14:54:52,411 INFO mapreduce.JobSubmitter: number of splits:2
2025-05-06 14:54:52,519 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746522831414_0001
2025-05-06 14:54:52,519 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-06 14:54:52,598 INFO conf.Configuration: resource-types.xml not found
2025-05-06 14:54:52,598 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-05-06 14:54:52,709 INFO impl.YarnClientImpl: Submitted application application_1746522831414_0001
2025-05-06 14:54:52,734 INFO mapreduce.Job: The url to track the job: http://bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:8
088/proxy/application_1746522831414_0001/
2025-05-06 14:54:52,734 INFO mapreduce.Job: Running job: job_1746522831414_0001
2025-05-06 14:54:57,790 INFO mapreduce.Job: Job job_1746522831414_0001 running in uber mode : false
2025-05-06 14:54:57,792 INFO mapreduce.Job:  map 0% reduce 0%
2025-05-06 14:55:01,857 INFO mapreduce.Job:  map 100% reduce 0%
2025-05-06 14:55:05,884 INFO mapreduce.Job:  map 100% reduce 100%
2025-05-06 14:55:05,903 INFO mapreduce.Job: Job job_1746522831414_0001 completed successfully
2025-05-06 14:55:05,968 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=118167
                FILE: Number of bytes written=1064123
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1777394
                HDFS: Number of bytes written=72
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=2666
                Total time spent by all reduces in occupied slots (ms)=1373
                Total time spent by all map tasks (ms)=2666
                Total time spent by all reduce tasks (ms)=1373
                Total vcore-milliseconds taken by all map tasks=2666
                Total vcore-milliseconds taken by all reduce tasks=1373
                Total megabyte-milliseconds taken by all map tasks=2729984
                Total megabyte-milliseconds taken by all reduce tasks=1405952
```

```
        Map-Reduce Framework
                Map input records=13130
                Map output records=13129
                Map output bytes=91903
                Map output materialized bytes=118173
                Input split bytes=226
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=118173
                Reduce input records=13129
                Reduce output records=12
                Spilled Records=26258
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=15
                CPU time spent (ms)=2290
                Physical memory (bytes) snapshot=1042640896
                Virtual memory (bytes) snapshot=8414855168
                Total committed heap usage (bytes)=1572864000
                Peak Map Physical memory (bytes)=379871232
                Peak Map Virtual memory (bytes)=2805063680
                Peak Reduce Physical memory (bytes)=286236672
                Peak Reduce Virtual memory (bytes)=2810429440
        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=1777168
        File Output Format Counters
                Bytes Written=72
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls outputdata
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2025-05-06 14:55 outputdata/_SUCCESS
-rw-r--r--   1 hadoop supergroup         72 2025-05-06 14:55 outputdata/part-r-00000
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat outputdata/part-r-00000
01      4
02      1
03      6
04      34
05      89
06      143
07      182
08      172
09      123
10      73
11      21
12      3
```


eclipse-workspace-2 - Weather/src/meanmax/MeanMaxDriver.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer
Weather
JRE System Library [JavaSE-1.8]
src
(default package)
meanmax
MeanMaxDriver.java
MeanMaxMapper.java
MeanMaxReducer.java
Referenced Libraries
hadoop-common-3.3.6.jar - /home/hadoop/hadoop/share
hadoop-common-3.3.6-tests.jar - /home/hadoop/hadoop/
hadoop-kms-3.3.6.jar - /home/hadoop/hadoop/share/had
hadoop-mapreduce-client-app-3.3.6.jar - /home/hadoop/h
hadoop-mapreduce-client-common-3.3.6.jar - /home/hadc
hadoop-mapreduce-client-core-3.3.6.jar - /home/hadoop/l
hadoop-mapreduce-client-hs-3.3.6.jar - /home/hadoop/ha
hadoop-mapreduce-client-hs-plugins-3.3.6.jar - /home/had
hadoop-mapreduce-client-jobclient-3.3.6.jar - /home/hadc
hadoop-mapreduce-client-jobclient-3.3.6-tests.jar - /home
hadoop-mapreduce-client-nativetask-3.3.6.jar - /home/hac
hadoop-mapreduce-client-shuffle-3.3.6.jar - /home/hadoo
hadoop-mapreduce-client-uploader-3.3.6.jar - /home/hadc
hadoop-mapreduce-examples-3.3.6.jar - /home/hadoop/h
hadoop-nfs-3.3.6.jar - /home/hadoop/hadoop/share/hadc
hadoop-registry-3.3.6.jar - /home/hadoop/hadoop/share/
WordCount
MeanMaxDriver.java
MeanMaxMapper.java
MeanMaxReducer.java
1 package meanmax;
2
3 import org.apache.hadoop.fs.Path;
4 import org.apache.hadoop.io.IntWritable;
5 import org.apache.hadoop.io.Text;
6 import org.apache.hadoop.mapreduce.Job;
7 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
8 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
9
10 public class MeanMaxDriver {
11 public static void main(String[] args) throws Exception {
12 if (args.length != 2) {
13 System.err.println("Please Enter the input and output parameters");
14 System.exit(-1);
15 }
16
17 Job job = Job.getInstance();
18 job.setJarByClass(MeanMaxDriver.class);
19 job.setJobName("Max temperature");
20
21 FileInputFormat.addInputPath(job, new Path(args[0]));
22 FileOutputFormat.setOutputPath(job, new Path(args[1]));
23
24 job.setMapperClass(MeanMaxMapper.class);
25 job.setReducerClass(MeanMaxReducer.class);
26
27 job.setOutputKeyClass(Text.class);
28 job.setOutputValueClass(IntWritable.class);
29
30 System.exit(job.waitForCompletion(true) ? 0 : 1);
31 }
32 }
33
Task List
Find
All
Activate...
Outline
meanmax
MeanMaxDriver
main(String[]) : void
Problems
Javadoc
Declaration
Console
Install Java 24 Support
Java 24 has been released.
Would you like to install Eclipse Java Development Tools support for Java 24 into your IDE?
Install Java 24 Support
Remind Me Later
Don't Ask Again
Would you like to download a Java 24 JDK?
Download Temurin JDK 24
Download OpenJDK 24

LAB 7 – Top N Words Hadoop

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.

```
2025-04-29 15:32:09,761 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager
 at /0.0.0.0:8032
2025-04-29 15:32:09,829 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager
 at /0.0.0.0:8032
2025-04-29 15:32:09,918 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not pe
rformed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-04-29 15:32:09,944 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/h
adoop-yarn/staging/hadoop/.staging/job_1745919848818_0003
2025-04-29 15:32:10,138 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-29 15:32:10,227 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-29 15:32:10,318 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745919848818_000
3
2025-04-29 15:32:10,318 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-29 15:32:10,405 INFO conf.Configuration: resource-types.xml not found
2025-04-29 15:32:10,405 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-29 15:32:10,556 INFO impl.YarnClientImpl: Submitted application application_1745919848818_000
3
2025-04-29 15:32:10,574 INFO mapreduce.Job: The url to track the job: http://bmscecse-HP-Elite-Tower-
800-G9-Desktop-PC:8088/proxy/application_1745919848818_0003/
2025-04-29 15:32:10,575 INFO mapreduce.Job: Running job: job_1745919848818_0003
2025-04-29 15:32:15,652 INFO mapreduce.Job: Job job_1745919848818_0003 running in uber mode : false
2025-04-29 15:32:15,654 INFO mapreduce.Job:  map 0% reduce 0%
2025-04-29 15:32:18,772 INFO mapreduce.Job:  map 100% reduce 0%
2025-04-29 15:32:22,799 INFO mapreduce.Job:  map 100% reduce 100%
2025-04-29 15:32:23,824 INFO mapreduce.Job: Job job_1745919848818_0003 completed successfully
2025-04-29 15:32:23,882 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=215
                FILE: Number of bytes written=829242
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=306
                HDFS: Number of bytes written=69
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=2555
                Total time spent by all reduces in occupied slots (ms)=1281
                Total time spent by all map tasks (ms)=2555
                Total time spent by all reduce tasks (ms)=1281
                Total vcore-milliseconds taken by all map tasks=2555
                Total vcore-milliseconds taken by all reduce tasks=1281
                Total megabyte-milliseconds taken by all map tasks=2616320
                Total megabyte-milliseconds taken by all reduce tasks=1311744
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /rgs/output
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2025-04-29 15:32 /rgs/output/_SUCCESS
-rw-r--r--   1 hadoop supergroup         69 2025-04-29 15:32 /rgs/output/part-00000
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /rgs/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-800-G9-Desktop-PC:~$
```

LAB 8 – SCALA PROGRAM

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

```scala
object ExampleForLoop1 {
  def main(args: Array[String]): Unit = {
    for (counter <- 1 to 100)
      print(counter + " ")
    println()
  }
}
```

```
>_ Console (F3)

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 5
```

LAB 9 – RDD And FlatMap:

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.