

# VISVESVARAYA TECHNOLOGICAL UNIVERSITY

“JnanaSangama”, Belgaum -590014, Karnataka.



**LAB REPORT**

**on**

## **BIG DATA ANALYTICS (22CS6PEBDA)**

*Submitted by*

**VYLERI KEZHEKE SIDDHARTH (1BM21CS247)**

*in partial fulfillment for the award of the degree of*

**BACHELOR OF ENGINEERING**

*in*

**COMPUTER SCIENCE AND ENGINEERING**



**B.M.S. COLLEGE OF ENGINEERING**

(Autonomous Institution under VTU)

**BENGALURU-560019**

**Feb-2024 to July-2024**

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



**CERTIFICATE**

This is to certify that the Lab work entitled "BIG DATA ANALYTICS" carried out by **VYLERI KEZHEKE SIDDHARTH (1BM21CS247)**, who is bonafide student of **B. M. S. College of Engineering**. It is in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visvesvaraya Technological University, Belgaum during the year 2024. The Lab report has been approved as it satisfies the academic requirements in respect of a **Big Data Analytics – (22CS6PEBDA)** work prescribed for the said degree.

**Saritha A.N**  
Assistant Professor  
Department of CSE  
BMSCE, Bengaluru

**Dr. Jyothi S Nayak**  
Professor and Head  
Department of CSE  
BMSCE, Bengaluru

## Index Sheet

Sl. No.	Experiment Title	Page No.
1	Perform the following DB operations using MongoDB	1
2	Perform the following DB operations using Cassandra	4
3	Hadoop installation	11
4	Implement Word Count Program on Hadoop framework	13
5	Create a Map Reduce program for the given dataset	17

## **BDA LAB-2**

**DATE:01-04-2024**

I Perform the following DB operations using MongoDB.

1. Create a database “Student” with the following attributes Rollno, Age, ContactNo, Email-Id.
2. Insert appropriate values
3. Write a query to update the Email-Id of a student with roll no 10.
4. . Replace the student name from “ABC” to “FEM” of roll no 11

```
Atlas atlas-xnuql-shard-0 [primary] test> db.Student.insert({_id:1,roll_no:1,stud_name:"ABC",age:20,contact_no:9988776655,email:"abc@gmail.com"});
{ acknowledged: true, insertedIds: { '0': 1 } }
Atlas atlas-xnuql-shard-0 [primary] test> db.Student.update({roll_no:10},{ $set:{email:'abcd@gmail.com'}});
Uncaught:
SyntaxError: Unexpected token, expected ",", (1:61)

> 1 | db.Student.update({roll_no:10},{ $set:{email:'abcd@gmail.com'}});
    |                                     ^
    2 |

Atlas atlas-xnuql-shard-0 [primary] test> db.Student.update({roll_no:10},{ $set:{email:'abcd@gmail.com'}},{upsert:true});
{
  acknowledged: true,
  insertedId: ObjectId("660a84f713da6f733017258d"),
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 1
}
Atlas atlas-xnuql-shard-0 [primary] test> db.Student.update({roll_no:1},{ $set:{stud_name:'FEM'}},{upsert:true});
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
Atlas atlas-xnuql-shard-0 [primary] test> db.Student.find({});
```

```
Atlas atlas-xnu1gl-shard-0 [primary] test> db.Student.find({});
[
  {
    _id: 1,
    roll_no: 1,
    stud_name: 'FEM',
    age: 20,
    contact_no: 9988776655,
    email: 'abc@gmail.com'
  },
  {
    _id: ObjectId("660a84f713da6f733017258d"),
    roll_no: 10,
    email: 'abcd@gmail.com'
  }
]
```

II. Perform the following DB operations using MongoDB.

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

3. Write a query to display those records whose total account balance is greater than 1200 of account type 'Z' for each customer\_id.

4. Determine Minimum and Maximum account balance for each customer\_id

```

Atlas atlas-xnulgl-shard-0 [primary] test> db.createCollection('customer');
{ ok: 1 }
Atlas atlas-xnulgl-shard-0 [primary] test> db.customer.insert({cust_id:100,acc_bal:1500,acc_type:'z'});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId("660a85c23be552442cee58a4") }
}
Atlas atlas-xnulgl-shard-0 [primary] test> db.customer.insert({cust_id:101,acc_bal:1300,acc_type:'a'});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId("660a85d63be552442cee58a5") }
}
Atlas atlas-xnulgl-shard-0 [primary] test> db.customer.insert({cust_id:102,acc_bal:1200,acc_type:'x'});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId("660a85e63be552442cee58a6") }
}
Atlas atlas-xnulgl-shard-0 [primary] test> db.customer.insert({cust_id:101,acc_bal:1210,acc_type:'z'});

```

```

{
  acknowledged: true,
  insertedIds: { '0': ObjectId("660a85f83be552442cee58a7") }
}
Atlas atlas-xnulgl-shard-0 [primary] test> db.customer.insert({cust_id:103,acc_bal:1210,acc_type:'a'});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId("660a869b3be552442cee58a8") }
}
Atlas atlas-xnulgl-shard-0 [primary] test> db.customer.aggregate({$match:{acc_type:'x'}},{ $group:{_id:'cust_id',total_acc_bal:{$sum:'$acc_bal'}}},{ $match:{total_acc_bal:{$gt:1200}}});
[ { _id: 'cust_id', total_acc_bal: 2710 } ]
Atlas atlas-xnulgl-shard-0 [primary] test> db.customer.aggregate({$match:{acc_type:'x'}},{ $group:{_id:'$cust_id',total_acc_bal:{$sum:'$acc_bal'}}},{ $match:{total_acc_bal:{$gt:1200}}});
[
  { _id: 101, total_acc_bal: 1210 },
  { _id: 100, total_acc_bal: 1500 }
]
Atlas atlas-xnulgl-shard-0 [primary] test> db.customer.aggregate({$group:{_id:'$cust_id',min_bal:{$min:'$acc_bal'},max_bal:{$max:'$acc_bal'}}});
[
  { _id: 101, min_bal: 1210, max_bal: 'acc_type' },
  { _id: 100, min_bal: 1500, max_bal: 'acc_type' },
  { _id: 102, min_bal: 1200, max_bal: 'acc_type' },
]
Atlas atlas-xnulgl-shard-0 [primary] test>

```

## BDA LAB-3 06-05-2024

### Cassandra

```
bmscscse@bmscscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.4 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> CREATE KEYSPACE Students WITH REPLICATION={
... 'class':'SimpleStrategy','replication_factor':1};
cqlsh> DESCRIBE KEYSPACES

students      system_auth      system_schema      system_views
system        system_distributed system_traces       system_virtual_schema

cqlsh> SELECT * FROM system.schema_keyspaces;
InvalidRequest: Error from server: code=2200 [Invalid query] message="table schema_keyspaces does not exist"
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;
Table 'students' not found in keyspace 'students'
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 mentable = '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 mentable_flush_period_in_ms = 0
AND min_index_interval = 128
AND read_repair = 'BLOCKING'
AND speculative_retry = '99p';

cqlsh:students> Begin batch insert into Students_info(Roll_no, StudName,DateOfJoining, last_exam_Percent) values(1,'Sadhana','2023-10-09', 98) insert into Students_info(Roll_no, StudName,DateOfJoining, last_exam_Percent) values(2,'Rutu','2023-10-10', 97) insert into Students_info(Roll_no, StudName,DateOfJoining, last_exam_Percent) values(3,'Rachana','2023-10-10', 97.5) insert into Students_info(Roll_no, StudName,DateOfJoining, last_exam_Percent) values(4,'Charu','2023-10-06', 96.5) apply batch;
cqlsh:students> select * from students_info;

roll_no | dateofjoining | last_exam_percent | studname
-----|-----|-----|-----
1 | 2023-10-08 18:30:00.000000+0000 | 98 | Sadhana
2 | 2023-10-09 18:30:00.000000+0000 | 97 | Rutu
4 | 2023-10-05 18:30:00.000000+0000 | 96.5 | Charu
3 | 2023-10-09 18:30:00.000000+0000 | 97.5 | Rachana

(4 rows)
cqlsh:students> select * from students_info where roll_no in (1,2,3);

roll_no | dateofjoining | last_exam_percent | studname
-----|-----|-----|-----
1 | 2023-10-08 18:30:00.000000+0000 | 98 | Sadhana
2 | 2023-10-09 18:30:00.000000+0000 | 97 | Rutu
3 | 2023-10-09 18:30:00.000000+0000 | 97.5 | Rachana

(3 rows)
cqlsh:students> select * from students_info where Studname='Charu';
InvalidRequest: Error from server: code=2200 [Invalid query] message="Cannot execute this query as it might involve data filtering and thus may have unpredictable performance. If you want to execute this query despite the performance unpredictability, use ALLOW FILTERING"
cqlsh:students> create index on Students_info(StudName);
cqlsh:students> select * from students_info where Studname='Charu';

roll_no | dateofjoining | last_exam_percent | studname
-----|-----|-----|-----
4 | 2023-10-05 18:30:00.000000+0000 | 96.5 | Charu

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

```

(4 rows)
cqlsh:students> select * from students_info where roll_no in (1,2,3);

roll_no | dateofjoining | last_exam_percent | studname
-----|-----|-----|-----
1 | 2023-10-08 18:30:00.000000+0000 | 98 | Sadhana
2 | 2023-10-09 18:30:00.000000+0000 | 97 | Rutu
3 | 2023-10-09 18:30:00.000000+0000 | 97.5 | Rachana

(3 rows)
cqlsh:students> select * from students_info where Studname='Charu';
InvalidRequest: Error from server: code=2200 [Invalid query] message="Cannot execute this query as it might involve data filtering and thus may have unpredictable performance. If you want to execute this query despite the performance unpredictability, use ALLOW FILTERING"
cqlsh:students> create index on Students_info(StudName);
cqlsh:students> select * from students_info where Studname='Charu';

roll_no | dateofjoining | last_exam_percent | studname
-----|-----|-----|-----
4 | 2023-10-05 18:30:00.000000+0000 | 96.5 | Charu

(1 rows)
cqlsh:students> select Roll_no,StudName from students_info LIMIT 2;

roll_no | studname
-----|-----
1 | Sadhana
2 | Rutu

(2 rows)
cqlsh:students> SELECT Roll_no as "USN" from Students_info;

USN
----
1
2
4
3

```

bmscecse@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~\$ cqlsh  
 Connected to Test Cluster at 127.0.0.1:9042  
 [cqlsh 6.1.0 | Cassandra 4.1.4 | CQL spec 3.4.6 | Native protocol v5]  
 Use HELP for help.

```

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

```

students system\_auth system\_schema system\_views  
 systemsystem\_distributed system\_traces system\_virtual\_schema

```

cqlsh> SELECT * FROM system.schema_keyspaces;
InvalidRequest: Error from server: code=2200 [Invalid query] message="table
schema_keyspaces does not exist"
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;
Table 'students' not found in keyspace 'students'
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> Begin batch insert into Students_info(Roll_no, StudName,DateOfJoining,
last_exam_Percent) values(1,'Sadhana','2023-10-09', 98)
insert into Students_info(Roll_no, StudName,DateOfJoining, last_exam_Percent)
values(2,'Rutu','2023-10-10', 97)
insert into Students_info(Roll_no, StudName,DateOfJoining, last_exam_Percent)
values(3,'Rachana','2023-10-10', 97.5)
insert into Students_info(Roll_no, StudName,DateOfJoining, last_exam_Percent)
values(4,'Charu','2023-10-06', 96.5) apply batch;
cqlsh:students> select * from students_info;

```

roll_no	dateofjoining	last_exam_percent	studname
1	2023-10-08 18:30:00.000000+0000	98	Sadhana
2	2023-10-09 18:30:00.000000+0000	97	Rutu
4	2023-10-05 18:30:00.000000+0000	96.5	Charu
3	2023-10-09 18:30:00.000000+0000	97.5	Rachana

(4 rows)

```

cqlsh:students> select * from students_info where roll_no in (1,2,3);

```

roll_no	dateofjoining	last_exam_percent	studname
1	2023-10-08 18:30:00.000000+0000	98	Sadhana
2	2023-10-09 18:30:00.000000+0000	97	Rutu
3	2023-10-09 18:30:00.000000+0000	97.5	Rachana

cqlsh:students> select \* from students\_info where Studname='Charu';

InvalidRequest: Error from server: code=2200 [Invalid query] message="Cannot execute this query as it might involve data filtering and thus may have unpredictable performance. If you want to execute this query despite the performance unpredictability, use ALLOW FILTERING"

cqlsh:students> create index on Students\_info(StudName);

cqlsh:students> select \* from students\_info where Studname='Charu';

roll_no	dateofjoining	last_exam_percent	studname
4	2023-10-05 18:30:00.000000+0000	96.5	Charu

(1 rows)

cqlsh:students> select Roll\_no,StudName from students\_info LIMIT 2;

roll_no	studname
1	Sadhana
2	Rutu

(2 rows)

cqlsh:students> SELECT Roll\_no as "USN" from Students\_info;

USN
1
2
4
3

(4 rows)

cqlsh:students> update students\_info set StudName='Shreya' where Roll\_no=3;

cqlsh:students> select \* from students\_info;

roll_no	dateofjoining	last_exam_percent	studname
1	2023-10-08 18:30:00.000000+0000	98	Sadhana
2	2023-10-09 18:30:00.000000+0000	97	Rutu
4	2023-10-05 18:30:00.000000+0000	96.5	Charu
3	2023-10-09 18:30:00.000000+0000	97.5	Shreya

(4 rows)

```
cqlsh:students> update students_info set roll_no=8 where Roll_no=3;
```

InvalidRequest: Error from server: code=2200 [Invalid query] message="PRIMARY KEY part roll\_no found in SET part"

```
cqlsh:students> delete last_exam_percent from students_info where roll_no=2;
```

```
cqlsh:students> select * from students_info;
```

roll_no	dateofjoining	last_exam_percent	studname
1	2023-10-08 18:30:00.000000+0000	98	Sadhana
2	2023-10-09 18:30:00.000000+0000	null	Rutu
4	2023-10-05 18:30:00.000000+0000	96.5	Charu
3	2023-10-09 18:30:00.000000+0000	97.5	Shreya

(4 rows)

```
cqlsh:students> delete from students_info where roll_no=2;
```

```
cqlsh:students> select * from students_info;
```

roll_no	dateofjoining	last_exam_percent	studname
1	2023-10-08 18:30:00.000000+0000	98	Sadhana
4	2023-10-05 18:30:00.000000+0000	96.5	Charu
3	2023-10-09 18:30:00.000000+0000	97.5	Shreya

(3 rows)

Cassandra : Employee

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

```

hmsccsc@hmsccsc-HP-Elite-Tower-880-G9-Desktop-PC: $ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.4 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> create keyspace Employee with replication = {'class': 'SimpleStrategy', 'replication_factor': 1};
SyntaxException: Line 1:85 mismatched input ';' expecting '}' (...with replication = {'class': 'SimpleStrategy', 'replication_factor': 1}...)
cqlsh> create keyspace Employee WITH replication={'class': 'SimpleStrategy', 'replication_factor': 1};
SyntaxException: unrecognized strategy option (replication_factor) passed to SimpleStrategy for keyspace Employee
cqlsh> create keyspace Employee WITH replication={'class': 'SimpleStrategy', 'replication_factor': 1};
cqlsh> DESCRIBE KEYSPACES

employee      system_auth      system_schema    system_views
system        system_distributed system_traces     system_virtual_schema

cqlsh> CREATE TABLE IF NOT EXISTS Employee_Info(
... Emp_Id INT PRIMARY KEY,
... Emp_name TEXT,
... designation TEXT,
... date_of_joining DATE,
... Salary FLOAT,
... Dep_name TEXT,
... Projects SET<TEXT>);
InvalidRequest: Error from server: code=2200 [Invalid query] message="No keyspace has been specified. USE a keyspace, or explicitly specify keyspace.tableName"
cqlsh> USE EMPLOYEE
cqlsh> USE Employee
...
cqlsh> USE Employee;
cqlsh:employee> CREATE TABLE IF NOT EXISTS Employee_Info( Emp_Id INT PRIMARY KEY, Emp_name TEXT, designation TEXT, date_of_joining DATE, Salary FLOAT, Dep_name TEXT, Projects SET<TEXT>);
cqlsh:employee> describe keyspace Employee

CREATE KEYSPACE employee WITH replication = {'class': 'SimpleStrategy', 'replication_factor': 1} AND durable_writes = true;

CREATE TABLE employee.employee_info (
  emp_id int PRIMARY KEY,
  date_of_joining date,
  dep_name text,
  designation text,
  emp_name text,
  salary float,
  projects set<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 mentable = '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 mentable_flush_period_in_ms = 0
AND min_index_interval = 128

```

```

cqlsh:employee> update employee_info using ttl 15 set salary = 0 where emp_id = 121;
cqlsh:employee> select * from employee_info;

```

emp_id	bonus	date_of_joining	dep_name	designation	emp_name	projects	salary
120	12000	2024-05-06	Engineering	Developer	Priyanka GH	{'Project B', 'ProjectA'}	1e+06
123	null	2024-05-07	Engineering	Engineer	Sadhana	{'Project M', 'Project P'}	1.2e+06
122	null	2024-05-06	Management	HR	Rachana	{'Project C', 'Project M'}	9e+05
121	11000	2024-05-06	Management	Developer	Shreya	{'Project C', 'ProjectA'}	0

(4 rows)

```
cqlsh:employee> select * from employee_info;
```

emp_id	bonus	date_of_joining	dep_name	designation	emp_name	projects	salary
120	12000	2024-05-06	Engineering	Developer	Priyanka GH	{'Project B', 'ProjectA'}	1e+06
123	null	2024-05-07	Engineering	Engineer	Sadhana	{'Project M', 'Project P'}	1.2e+06
122	null	2024-05-06	Management	HR	Rachana	{'Project C', 'Project M'}	9e+05
121	11000	2024-05-06	Management	Developer	Shreya	{'Project C', 'ProjectA'}	null

(4 rows)

```
cqlsh:employee>
```

```

AND speculative_retry = '99p';
cqlsh:employee> select * from employee_info;

emp_id | date_of_joining | dep_name | designation | emp_name | projects | salary
-----|-----|-----|-----|-----|-----|-----
120 | 2024-05-06 | Engineering | Developer | Priyanka GH | ('Project B', 'ProjectA') | 1e+06
123 | 2024-05-07 | Engineering | Engineer | Sadhana | ('Project M', 'Project P') | 1.2e+06
122 | 2024-05-06 | Management | HR | Rachana | ('Project C', 'Project M') | 9e+05
121 | 2024-05-06 | Management | Developer | Shreya | ('Project C', 'ProjectA') | 9e+05
(4 rows)
cqlsh:employee> update employee_info set emp_name = 'Priyanka GH' where emp_id = '120';
InvalidRequest: Error from server: code=2208 [Invalid query] message="Invalid STRING constant (120) for "emp_id" of type int"
cqlsh:employee> update employee_info set emp_name = 'Priyanka GH' where emp_id=120;
cqlsh:employee> select * from employee_info;

emp_id | date_of_joining | dep_name | designation | emp_name | projects | salary
-----|-----|-----|-----|-----|-----|-----
120 | 2024-05-06 | Engineering | Developer | Priyanka GH | ('Project B', 'ProjectA') | 1e+06
123 | 2024-05-07 | Engineering | Engineer | Sadhana | ('Project M', 'Project P') | 1.2e+06
122 | 2024-05-06 | Management | HR | Rachana | ('Project C', 'Project M') | 9e+05
121 | 2024-05-06 | Management | Developer | Shreya | ('Project C', 'ProjectA') | 9e+05
(4 rows)
cqlsh:employee> select * from employee_info order by salary;
InvalidRequest: Error from server: code=2208 [Invalid query] message="ORDER BY is only supported when the partition key is restricted by an EQ or an IN."
cqlsh:employee> alter table employee_info add bonus INT;
cqlsh:employee> select * from employee_info;

emp_id | bonus | date_of_joining | dep_name | designation | emp_name | projects | salary
-----|-----|-----|-----|-----|-----|-----|-----
120 | null | 2024-05-06 | Engineering | Developer | Priyanka GH | ('Project B', 'ProjectA') | 1e+06
123 | null | 2024-05-07 | Engineering | Engineer | Sadhana | ('Project M', 'Project P') | 1.2e+06
122 | null | 2024-05-06 | Management | HR | Rachana | ('Project C', 'Project M') | 9e+05
121 | null | 2024-05-06 | Management | Developer | Shreya | ('Project C', 'ProjectA') | 9e+05
(4 rows)
cqlsh:employee> update employee_info set bonus = 12000 where emp_id = 120;
cqlsh:employee> select * from employee_info;

emp_id | bonus | date_of_joining | dep_name | designation | emp_name | projects | salary
-----|-----|-----|-----|-----|-----|-----|-----
120 | 12000 | 2024-05-06 | Engineering | Developer | Priyanka GH | ('Project B', 'ProjectA') | 1e+06
123 | null | 2024-05-07 | Engineering | Engineer | Sadhana | ('Project M', 'Project P') | 1.2e+06
122 | null | 2024-05-06 | Management | HR | Rachana | ('Project C', 'Project M') | 9e+05
121 | null | 2024-05-06 | Management | Developer | Shreya | ('Project C', 'ProjectA') | 9e+05
(4 rows)
cqlsh:employee> update employee_info set bonus = 11000 where emp_id = 121;
cqlsh:employee> select * from employee_info using ttl 15 where emp_id = 123;
SyntaxException: line 1:44 statement 'using' expecting an 'insert' * from employee_info [using] ttl...)
cqlsh:employee> select * from employee_info where emp_id = 121 using ttl 15;
SyntaxException: line 1:47 no viable alternative at input 'using' (...employee_info where emp_id = 121 [using]...)
cqlsh:employee> update employee_info using ttl 15 set salary = 0 where emp_id = 121;
cqlsh:employee> select * from employee_info;

```

## **HADOOP** 13-05-24

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-all.sh
```

```
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
```

```
WARNING: This is not a recommended production deployment configuration.
```

```
WARNING: Use CTRL-C to abort.
```

```
Starting namenodes on [localhost]
```

```
Starting datanodes
```

```
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
```

```
Starting resourcemanager
```

```
Starting nodemanagers
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop dfs -mkdir /sadh
```

```
WARNING: Use of this script to execute dfs is deprecated.
```

```
WARNING: Attempting to execute replacement "hdfs dfs" instead.
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -mkdir /sadh
```

```
mkdir: `/sadh': File exists
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /
```

```
Found 1 items
```

```
drwxr-xr-x - hadoop supergroup    0 2024-05-13 14:27 /sadh
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /sadh
```

```

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -put
/home/hadoop/Desktop/example/Welcome.txt /sadh/WC.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -cat /sadh/WC.txt
hiii
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -get /sadh/WC.txt
/home/hadoop/Desktop/example/WWC.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -get /sadh/WC.txt
/home/hadoop/Desktop/example/WWC2.txt

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -put
/home/hadoop/Desktop/example/Welcome.txt /sadh/WC2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -getmerge /sadh/WC.txt
/sadh/WC2.txt /home/hadoop/Desktop/example/Merge.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -getfacl /sadh/
# file: /sadh
# owner: hadoop
# group: supergroup
user::rwx
group::r-x
other::r-x

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mv /sadh /WC2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /sadh /WC2.txt
ls: `/sadh': No such file or directory
Found 2 items
-rw-r--r-- 1 hadoop supergroup      6 2024-05-13 14:51 /WC2.txt/WC.txt
-rw-r--r-- 1 hadoop supergroup      6 2024-05-13 15:03 /WC2.txt/WC2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cp /WC2.txt/ /WC.txt

```

## **BDA LAB-5**

**DATE:-27-05-2024**

### **Implement WordCount Program on Hadoop framework**

Mapper Code :

```
import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapred.Mapper;

import org.apache.hadoop.mapred.OutputCollector;

import org.apache.hadoop.mapred.Reporter;

public class WCMapper extends Mapper<LongWritable, Text,
    Text,
    IntWritable> {

    public void map(LongWritable key, Text value, OutputCollector<IntWritable> collector, Reporter r) throws IOException {

        String line = value.toString();

        for (String word : line.split(" ")) {

            word.toLowerCase().getBytes().length > 0)
        }
    }
}
```



```
o"tp"t.collecl(→cw rct(woid), →cw I→tWitablc(1));

||||
```

```
// Importing libraries
```

14

```

{
    IntWritable i = value.next();
    count += i.get();
}

output.collect(key, new IntWritable(count));
}}

```

Driver Code: You have to copy paste this program into the WCDriver Java Class file.

```

// Importing libraries

import java.io.IOException;

import org.apache.hadoop.conf.Configured;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapred.FileInputFormat;

import org.apache.hadoop.mapred.FileOutputFormat;

import org.apache.hadoop.mapred.JobClient;

import org.apache.hadoop.mapred.JobConf;

import org.apache.hadoop.util.Tool;

import org.apache.hadoop.util.ToolRunner;

public class WCDriver extends Configured implements Tool {

    public int run(String args[]) throws IOException

    {

        if (args.length < 2)

        {

```

```

System.out.println("Please give valid inputs");

return -1;

}

JobConf conf = new JobConf(WCDriver.class);

FileInputFormat.setInputPaths(conf, new Path(args[0]));

FileOutputFormat.setOutputPath(conf, new Path(args[1]));

conf.setMapperClass(WCMapper.class);

conf.setReducerClass(WCReducer.class);

conf.setMapOutputKeyClass(Text.class);

conf.setMapOutputValueClass(IntWritable.class);

conf.setOutputKeyClass(Text.class);

conf.setOutputValueClass(IntWritable.class);

JobClient.runJob(conf);

return 0;

}

// Main Method

public static void main(String args[]) throws Exception

{

int exitCode = ToolRunner.run(new WCDriver(), args);

System.out.println(exitCode);

}

}

```

**From the following link extract the weather**

**data** <https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all>

**Cícatc a Map Rcducc píogíam to**

**a) fi→d a:cíagc tcmpcíatuíc foí cack Qcaí fíom NCDC data sct.**

### **AverageDriver**

```
package temp;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class AverageDriver {

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

        if (args.length != 2) {

            System.err.println("Please Enter the input and output parameters");

            System.exit(-1);

        }

        Job job = new Job();

        job.setJarByClass(AverageDriver.class);

        job.setJobName("Max temperature");

        FileInputFormat.addInputPath(job, new Path(args[0]));

        FileOutputFormat.setOutputPath(job, new Path(args[1]));
```

```

job.setMapperClass(AverageMapper.class);
job.setReducerClass(AverageReducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}

```

### **AverageMapper**

```

package temp;

import java.io.IOException;

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

public class AverageMapper extends Mapper<LongWritable, Text, Text, IntWritable> {

    public static final int MISSING = 9999;

    public void map(LongWritable key, Text value, Mapper<LongWritable, Text, Text,
IntWritable>.Context context) throws IOException, InterruptedException {

        int temperature;

        String line = value.toString();

        String year = line.substring(15, 19);

        if (line.charAt(87) == '+') {

            temperature = Integer.parseInt(line.substring(88, 92));

        } else {

```

```

temperature = Integer.parseInt(line.substring(87, 92));
}

String quality = line.substring(92, 93);

if (temperature != 9999 && quality.matches("[01459]"))
context.write(new Text(year), new IntWritable(temperature));
}
}

AverageReducer

package temp;

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));

}}

```

```

C:\hadoop-3.3.0\bin>hadoop jar C:\avgtemp.jar temp.AverageDriver /input_dir/temp.txt /avgtemp_outputdir
2021-05-15 14:52:50,635 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-05-15 14:52:51,805 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-05-15 14:52:51,111 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Anusree/.staging/job_1621060230696_0005
2021-05-15 14:52:51,735 INFO input.FileInputFormat: Total input files to process : 1
2021-05-15 14:52:52,751 INFO mapreduce.JobSubmitter: number of splits:1
2021-05-15 14:52:53,073 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1621060230696_0005
2021-05-15 14:52:53,073 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-05-15 14:52:53,237 INFO conf.Configuration: resource-types.xml not found
2021-05-15 14:52:53,238 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-05-15 14:52:53,312 INFO impl.YarnClientImpl: Submitted application application_1621060230696_0005
2021-05-15 14:52:53,352 INFO mapreduce.Job: The url to track the job: http://LAPTOP-JG329ESD:8088/proxy/application_1621060230696_0005/
2021-05-15 14:52:53,353 INFO mapreduce.Job: Running job: job_1621060230696_0005
2021-05-15 14:53:06,648 INFO mapreduce.Job: Job job_1621060230696_0005 running in uber mode : false
2021-05-15 14:53:06,643 INFO mapreduce.Job:  map 0% reduce 0%
2021-05-15 14:53:12,758 INFO mapreduce.Job:  map 100% reduce 0%
2021-05-15 14:53:19,868 INFO mapreduce.Job:  map 100% reduce 100%
2021-05-15 14:53:25,967 INFO mapreduce.Job: Job job_1621060230696_0005 completed successfully
2021-05-15 14:53:26,096 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=72210
        FILE: Number of bytes written=674341
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=894860
        HDFS: Number of bytes written=8
        HDFS: Number of read operations=8
        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=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=3782

```

```

C:\hadoop-3.3.0\bin>hdfs dfs -ls /avgtemp_outputdir
Found 2 items
-rw-r--r-- 1 Anusree supergroup          0 2021-05-15 14:53 /avgtemp_outputdir/_SUCCESS
-rw-r--r-- 1 Anusree supergroup      8 2021-05-15 14:53 /avgtemp_outputdir/part-r-00000

C:\hadoop-3.3.0\bin>hdfs dfs -cat /avgtemp_outputdir/part-r-00000
1901    46

C:\hadoop-3.3.0\bin>

```

b) find the mean max temperature for every month

### MeanMaxDriver.class

```
package meanmax;
```

```
import org.apache.hadoop.fs.Path;
```

```
import org.apache.hadoop.io.IntWritable;
```

```
import org.apache.hadoop.io.Text;
```

```
import org.apache.hadoop.mapreduce.Job;
```

```
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
```

```

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class MeanMaxDriver {

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

        if (args.length != 2) {

            System.err.println("Please Enter the input and output parameters");

            System.exit(-1);

        }

        Job job = new Job();

        job.setJarByClass(MeanMaxDriver.class);

        job.setJobName("Max temperature");

        FileInputFormat.addInputPath(job, new Path(args[0]));

        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        job.setMapperClass(MeanMaxMapper.class);

        job.setReducerClass(MeanMaxReducer.class);

        job.setOutputKeyClass(Text.class);

        job.setOutputValueClass(IntWritable.class);

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

    }

}

```

### **MeanMaxMapper.class**

```

package meanmax;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.LongWritable;

```



```

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Mapper;

public class MeanMaxMapper extends Mapper<LongWritable, Text, Text, IntWritable> {

    public static final int MISSING = 9999;

    public void map(LongWritable key, Text value, Mapper<LongWritable, Text, Text,
    IntWritable>.Context context) throws IOException, InterruptedException {

        int temperature;

        String line = value.toString();

        String month = line.substring(19, 21);

        if (line.charAt(87) == '+') {

            temperature = Integer.parseInt(line.substring(88, 92));

        } else {

            temperature = Integer.parseInt(line.substring(87, 92));

        }

        String quality = line.substring(92, 93);

        if (temperature != 9999 && quality.matches("[01459]"))

            context.write(new Text(month), new IntWritable(temperature));

    }

}

```

### **MeanMaxReducer.class**

```

package meanmax;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

```

```

import org.apache.hadoop.mapreduce.Reducer;

public class MeanMaxReducer 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 total_temp = 0;

        int count = 0;

        int days = 0;

        for (IntWritable value : values) {

            int temp = value.get();

            if (temp > max_temp)

                max_temp = temp;

            count++;

            if (count == 3) {

                total_temp += max_temp;

                max_temp = 0;

                count = 0;

                days++;

            }

        }

        context.write(key, new IntWritable(total_temp / days));

    }

}

```

```

C:\hadoop-3.3.0\sbin>hadoop jar C:\meanmax.jar meanmax.MeanMaxDriver /input_dir/temp.txt /meanmax_output
2021-05-21 20:28:05,250 INFO client.DefaultHARFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-05-21 20:28:06,662 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-05-21 20:28:06,916 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ksusree/staging/job_1621608943095_0001
2021-05-21 20:28:08,426 INFO input.FileInputFormat: Total input files to process : 1
2021-05-21 20:28:09,187 INFO mapreduce.JobSubmitter: number of splits:1
2021-05-21 20:28:09,741 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1621608943095_0001
2021-05-21 20:28:09,741 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-05-21 20:28:10,029 INFO conf.Configuration: resource-types.xml not found
2021-05-21 20:28:10,030 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-05-21 20:28:10,676 INFO impl.YarnClientImpl: Submitted application application_1621608943095_0001
2021-05-21 20:28:11,005 INFO mapreduce.Job: The url to track the job: http://LAPTOP-36329ESD-8008/proxy/application_1621608943095_0001/
2021-05-21 20:28:11,006 INFO mapreduce.Job: Running job: job_1621608943095_0001
2021-05-21 20:28:29,385 INFO mapreduce.Job: Job job_1621608943095_0001 running in uber mode : false
2021-05-21 20:28:29,389 INFO mapreduce.Job:  map 0% reduce 0%
2021-05-21 20:28:40,664 INFO mapreduce.Job:  map 100% reduce 0%
2021-05-21 20:28:50,832 INFO mapreduce.Job:  map 100% reduce 100%
2021-05-21 20:28:58,965 INFO mapreduce.Job: Job job_1621608943095_0001 completed successfully
2021-05-21 20:28:59,178 INFO mapreduce.Job: Counters: 54
    File System Counters
      FILE: Number of bytes read=59882
      FILE: Number of bytes written=648091
      FILE: Number of read operations=0
      FILE: Number of large read operations=0
      FILE: Number of write operations=0
      HDFS: Number of bytes read=894860
      HDFS: Number of bytes written=74
      HDFS: Number of read operations=0
      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=1
      Launched reduce tasks=1
      Data-local map tasks=1
      Total time spent by all maps in occupied slots (ms)=8877
      Total time spent by all reduces in occupied slots (ms)=7511
      Total time spent by all map tasks (ms)=8877
      Total time spent by all reduce tasks (ms)=7511
      Total vcore-milliseconds taken by all map tasks=8877
      Total vcore-milliseconds taken by all reduce tasks=7511
      Total megabyte-milliseconds taken by all map tasks=8270848
      Total megabyte-milliseconds taken by all reduce tasks=7091264

```

```

C:\hadoop-3.3.0\sbin>hdfs dfs -cat /meanmax_output/*
01      4
02      0
03      7
04     44
05    100
06    168
07    219
08    198
09    141
10    100
11     19
12     3

C:\hadoop-3.3.0\sbin>

```

**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.**

#### **Driver-TopN.class**

```
package samples.topn;

import java.io.IOException;

import java.util.StringTokenizer;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import org.apache.hadoop.util.GenericOptionsParser;

public class TopN {

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

        Configuration conf = new Configuration();

        String[] otherArgs = (new GenericOptionsParser(conf, args)).getRemainingArgs();

        if (otherArgs.length != 2) {

            System.err.println("Usage: TopN <in> <out>");

            System.exit(2);

        }

        Job job = Job.getInstance(conf);

        job.setJobName("Top N");
```

```

job.setJarByClass(TopN.class);

job.setMapperClass(TopNMapper.class);

job.setReducerClass(TopNReducer.class);

job.setOutputKeyClass(Text.class);

job.setOutputValueClass(IntWritable.class);

FileInputFormat.addInputPath(job, new Path(otherArgs[0]));

FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));

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

}

public static class TopNMapper extends Mapper<Object, Text, Text, IntWritable> {

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

    private Text word = new Text();

    private String tokens = "[_!$#<>\\^=\\[\\]\\*^\\\\\\\\,;,.\\-:()?!\"']";

    public void map(Object key, Text value, Mapper<Object, Text, Text, IntWritable>.Context
context) throws IOException, InterruptedException {

        String cleanLine = value.toString().toLowerCase().replaceAll(this.tokens, " ");

        StringTokenizer itr = new StringTokenizer(cleanLine);

        while (itr.hasMoreTokens()) {

            this.word.set(itr.nextToken().trim());

            context.write(this.word, one);

        }

    }

}

```

### **TopNCombiner.class**

```
package samples.topn;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Reducer;

public class TopNCombiner 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 sum = 0;

        for (IntWritable val : values)

            sum += val.get();

        context.write(key, new IntWritable(sum));

    }

}
```

### **TopNMapper.class**

```
package samples.topn;

import java.io.IOException;

import java.util.StringTokenizer;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Mapper;

public class TopNMapper extends Mapper<Object, Text, Text, IntWritable> {

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

```

private Text word = new Text();

private String tokens = "[_!$#<>\\^=\\[\\]\\*^\\\\\\,;\\.\\-:()?!\"']";

public void map(Object key, Text value, Mapper<Object, Text, Text, IntWritable>.Context
context) throws IOException, InterruptedException {

String cleanLine = value.toString().toLowerCase().replaceAll(this.tokens, " ");

StringTokenizer itr = new StringTokenizer(cleanLine);

while (itr.hasMoreTokens()) {

this.word.set(itr.nextToken().trim());

context.write(this.word, one);

}

}

}

```

### **TopNReducer.class**

```

package samples.topn;

import java.io.IOException;

import java.util.HashMap;

import java.util.Map;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Reducer;

import utils.MiscUtils;

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

private Map<Text, IntWritable> countMap = new HashMap<>();

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

```

```

Text, IntWritable>.Context context) throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable val : values)
        sum += val.get();
    this.countMap.put(new Text(key), new IntWritable(sum));
}

protected void cleanup(Reducer<Text, IntWritable, Text, IntWritable>.Context context)
    throws IOException, InterruptedException {
    Map<Text, IntWritable> sortedMap = MiscUtils.sortByValues(this.countMap);
    int counter = 0;
    for (Text key : sortedMap.keySet()) {
        if (counter++ == 20)
            break;
        context.write(key, sortedMap.get(key));
    }
}
}

```



```

C:\hadoop-3.3.0\sbin>jps
11072 DataNode
20528 Jps
5620 ResourceManager
15532 NodeManager
6140 NameNode

C:\hadoop-3.3.0\sbin>hdfs dfs -mkdir /input_dir

C:\hadoop-3.3.0\sbin>hdfs dfs -ls /
Found 1 items
drwxr-xr-x - Anusree supergroup 0 2021-05-08 19:46 /input_dir

C:\hadoop-3.3.0\sbin>hdfs dfs -copyFromLocal C:\input.txt /input_dir

C:\hadoop-3.3.0\sbin>hdfs dfs -ls /input_dir
Found 1 items
-rw-r--r-- 1 Anusree supergroup 36 2021-05-08 19:48 /input_dir/input.txt

C:\hadoop-3.3.0\sbin>hdfs dfs -cat /input_dir/input.txt
hello
world
hello
hadoop
bye

```

```

C:\hadoop-3.3.0\sbin>hadoop jar C:\sort.jar samples.topn.TopN /input_dir/input.txt /output_dir
2021-05-08 19:54:54,582 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-05-08 19:54:55,291 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Anusree/.staging/job_1620483374279_0001
2021-05-08 19:54:55,821 INFO input.FileInputFormat: Total input files to process : 1
2021-05-08 19:54:56,261 INFO mapreduce.JobSubmitter: number of splits:1
2021-05-08 19:54:56,552 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1620483374279_0001
2021-05-08 19:54:56,552 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-05-08 19:54:56,843 INFO conf.Configuration: resource-types.xml not found
2021-05-08 19:54:56,843 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-05-08 19:54:57,387 INFO impl.YarnClientImpl: Submitted application application_1620483374279_0001
2021-05-08 19:54:57,507 INFO mapreduce.Job: The url to track the job: http://LAPTOP-JG329E50:8088/proxy/application_1620483374279_0001/
2021-05-08 19:54:57,508 INFO mapreduce.Job: Running job: job_1620483374279_0001
2021-05-08 19:55:13,792 INFO mapreduce.Job: Job job_1620483374279_0001 running in uber mode : false
2021-05-08 19:55:13,794 INFO mapreduce.Job: map 0% reduce 0%
2021-05-08 19:55:20,020 INFO mapreduce.Job: map 100% reduce 0%
2021-05-08 19:55:27,116 INFO mapreduce.Job: map 100% reduce 100%
2021-05-08 19:55:33,199 INFO mapreduce.Job: Job job_1620483374279_0001 completed successfully
2021-05-08 19:55:33,334 INFO mapreduce.Job: Counters: 54
File System Counters
  FILE: Number of bytes read=65
  FILE: Number of bytes written=530397
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=142
  HDFS: Number of bytes written=31
  HDFS: Number of read operations=8
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
  HDFS: Number of bytes read erasure-coded=0

```

```
C:\hadoop-3.3.0\sbin>hdfs dfs -cat /output_dir/*  
hello 2  
hadoop 1  
world 1  
bye 1  
  
C:\hadoop-3.3.0\sbin>
```