VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka.

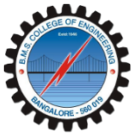
LAB REPORT
on

BIG DATA ANALYTICS
(20CS6PEBDA)

Submitted by

Kusum M R (1BM19CS077)

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
May-2022 to July-2022

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 **Kusum M R(1BM19CS077),** who is a bonafide student of **B. M. S. College of Engineering.** It is in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visvesvaraya Technological University, Belgaum during the year 2022. The Lab report has been approved as it satisfies the academic requirements in respect of a **BIG DATA ANALYTICS - (20CS6PEBDA)** work prescribed for the said degree.

Dr.Pallavi G B
Assistant Professor
Department of CSE
BMSCE, Bengaluru

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

`

Index Sheet

Course Outcome

CO1	Apply the concept of NoSQL, Hadoop or Spark for a given task
CO2	Analyze the Big Data and obtain insight using data analytics mechanisms.
CO3	Design and implement Big data applications by applying NoSQL, Hadoop or Spark

BDA LAB 1

Program 1. Perform the following DB operations using Cassandra.

1. Create a key space by name Employee

```
cqlsh> CREATE KEYSPACE Employee WITH replication = {'class': 'SimpleStrategy', 'replication_factor': 1};
cqlsh> describe keyspace
No keyspace specified and no current keyspace
cqlsh> describe 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

```
cqlsh> create table Employee.Employee_Info(Emp_Id int Primary Key,Emp_Name text,Designation text,Date_of_Joining timestamp,Salary dou
ble,Dept_Name text);
```

```
cqlsh> select * from Employee.Employee_Info;

 emp_id | date_of_joining | dept_name | designation | emp_name | salary
--------+-----------------+-----------+-------------+----------+--------

(0 rows)
```

3. Insert the values into the table in batch

```
cqlsh> begin batch insert into Employee.Employee_Info(emp_id,date_of_joining,dept_name,designation,emp_name,salary)values(1,'2021-06-
03','Deployment','Manager','Kusum',1500000.50); apply batch;
cqlsh> select * from Employee.Employee_Info;

 emp_id | date_of_joining                 | dept_name  | designation | emp_name | salary
--------+---------------------------------+------------+-------------+----------+---------
      1 | 2021-06-03 00:00:00.000000+0000 | Deployment |     Manager |    Kusum | 1.5e+06

(1 rows)
```

```
cqlsh> begin batch
   ... insert into Employee.Employee_Info(emp_id,date_of_joining,dept_name,designation,emp_name,salary)values(2,'2020-09-03','Develop
ment','Web developer','Karan',1700000.50);
   ... insert into Employee.Employee_Info(emp_id,date_of_joining,dept_name,designation,emp_name,salary)values(121,'2019-05-03','R&D',
'Intern','Kia',2000000.50);
   ... apply batch;
cqlsh> select * from Employee.Employee_Info;

 emp_id | date_of_joining                 | dept_name   | designation   | emp_name | salary
--------+---------------------------------+-------------+---------------+----------+---------
      1 | 2021-06-03 00:00:00.000000+0000 |  Deployment |       Manager |    Kusum | 1.5e+06
      2 | 2020-09-03 00:00:00.000000+0000 | Development | Web developer |    Karan | 1.7e+06
    121 | 2019-05-03 00:00:00.000000+0000 |         R&D |        Intern |      Kia |   2e+06

(3 rows)
```

4. Update Employee name and Department of Emp-Id 121

```
cqlsh> update Employee.Employee_Info SET emp_name='Kushi',dept_name='Testing' where emp_id=121;
```

```
cqlsh> select * from Employee.Employee_Info;

 emp_id | date_of_joining                 | dept_name   | designation    | emp_name | salary
--------+---------------------------------+-------------+----------------+----------+---------
      1 | 2021-06-03 00:00:00.000000+0000 |  Deployment |        Manager |    Kusum | 1.5e+06
      2 | 2020-09-03 00:00:00.000000+0000 | Development | Web developer |    Karan | 1.7e+06
    121 | 2019-05-03 00:00:00.000000+0000 |     Testing |         Intern |    Kushi |   2e+06

(3 rows)
```

5. Sort the details of Employee records based on salary

```
cqlsh> create table Employee.emp(Emp_Id int,Emp_name text,Designation text,Date_Of_Joining timestamp,Salary double,Dept_Name text,primary ke
y(Emp_Id,Salary));
```

```
cqlsh> begin batch
   ... insert into Employee.emp(emp_id,salary,date_of_joining,dept_name,designation,emp_name)values(1,1500000.50,'2021-06-03','Deployment','
Manager','Kusum');
   ... insert into Employee.emp(emp_id,salary,date_of_joining,dept_name,designation,emp_name)values(2,1100000.50,'2022-05-03','Development',
'Web Developer','Karan');
   ... insert into Employee.emp(emp_id,salary,date_of_joining,dept_name,designation,emp_name)values(121,1900000.50,'2022-05-03','R&D','Inter
n','Kia');
   ... apply batch;
cqlsh> select * from Employee.emp;

 emp_id | salary  | date_of_joining                 | dept_name   | designation    | emp_name
--------+---------+---------------------------------+-------------+----------------+----------
      1 | 1.5e+06 | 2021-06-03 00:00:00.000000+0000 |  Deployment |        Manager |    Kusum
      2 | 1.1e+06 | 2022-05-03 00:00:00.000000+0000 | Development | Web Developer |    Karan
    121 | 1.9e+06 | 2022-05-03 00:00:00.000000+0000 |         R&D |         Intern |      Kia

(3 rows)
cqlsh> paging off;
Disabled Query paging.
cqlsh> select * from Employee.emp where emp_id in (1,2,121) order by salary;

 emp_id | salary  | date_of_joining                 | dept_name   | designation    | emp_name
--------+---------+---------------------------------+-------------+----------------+----------
      2 | 1.1e+06 | 2022-05-03 00:00:00.000000+0000 | Development | Web Developer |    Karan
      1 | 1.5e+06 | 2021-06-03 00:00:00.000000+0000 |  Deployment |        Manager |    Kusum
    121 | 1.9e+06 | 2022-05-03 00:00:00.000000+0000 |         R&D |         Intern |      Kia

(3 rows)
cqlsh>
```

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.

```
cqlsh> alter table Employee.Employee_Info add Projects set<text>;
cqlsh> select * from Employee.Employee_Info;

 emp_id | date_of_joining                 | dept_name   | designation    | emp_name | projects | salary
--------+---------------------------------+-------------+----------------+----------+----------+---------
      1 | 2021-06-03 00:00:00.000000+0000 |  Deployment |        Manager |    Kusum |     null | 1.5e+06
      2 | 2020-09-03 00:00:00.000000+0000 | Development | Web developer |    Karan |     null | 1.7e+06
    121 | 2019-05-03 00:00:00.000000+0000 |     Testing |         Intern |    Kushi |     null |   2e+06

(3 rows)
```

7. Update the altered table to add project names.

```
cqlsh> update Employee.Employee_Info set projects=projects+{'abc','xyz'} where emp_id=1;
cqlsh> select * from Employee.Employee_Info;

 emp_id | date_of_joining                 | dept_name   | designation    | emp_name | projects       | salary
--------+---------------------------------+-------------+----------------+----------+----------------+---------
      1 | 2021-06-03 00:00:00.000000+0000 |  Deployment |        Manager |    Kusum | {'abc', 'xyz'} | 1.5e+06
      2 | 2020-09-03 00:00:00.000000+0000 | Development | Web developer |    Karan |           null | 1.7e+06
    121 | 2019-05-03 00:00:00.000000+0000 |     Testing |         Intern |    Kushi |           null |   2e+06

(3 rows)
cqlsh> update Employee.Employee_Info set projects=projects+{'pqr','lmn'} where emp_id=2;
cqlsh> update Employee.Employee_Info set projects=projects+{'tuv','def'} where emp_id=2;
cqlsh> select * from Employee.Employee_Info;

 emp_id | date_of_joining                 | dept_name   | designation    | emp_name | projects                       | salary
--------+---------------------------------+-------------+----------------+----------+--------------------------------+---------
      1 | 2021-06-03 00:00:00.000000+0000 |  Deployment |        Manager |    Kusum |                {'abc', 'xyz'} | 1.5e+06
      2 | 2020-09-03 00:00:00.000000+0000 | Development | Web developer |    Karan | {'def', 'lmn', 'pqr', 'tuv'} | 1.7e+06
    121 | 2019-05-03 00:00:00.000000+0000 |     Testing |         Intern |    Kushi |                           null |   2e+06

(3 rows)
```

```
cqlsh> update Employee.Employee_Info set projects=projects+{'lab','jkl'} where emp_id=121;
cqlsh> select * from Employee.Employee_Info;

 emp_id | date_of_joining                 | dept_name   | designation    | emp_name | projects                       | salary
--------+---------------------------------+-------------+----------------+----------+--------------------------------+---------
      1 | 2021-06-03 00:00:00.000000+0000 |  Deployment |        Manager |    Kusum |                {'abc', 'xyz'} | 1.5e+06
      2 | 2020-09-03 00:00:00.000000+0000 | Development | Web developer |    Karan | {'def', 'lmn', 'pqr', 'tuv'} | 1.7e+06
    121 | 2019-05-03 00:00:00.000000+0000 |     Testing |         Intern |    Kushi |                {'lab', 'jkl'} |   2e+06

(3 rows)
```

8 Create a TTL of 15 seconds to display the values of Employees.

```
cqlsh> insert into Employee.Employee_Info(emp_id,date_of_joining,dept_name,designation,emp_name,salary)values(11,'2019-05-05','R&D','
Intern','Kajal',1000000.50) using TTL 15;
cqlsh> select * from Employee.Employee_Info;

 emp_id | date_of_joining                 | dept_name   | designation    | emp_name | projects                       | salary
--------+---------------------------------+-------------+----------------+----------+--------------------------------+---------
     11 | 2019-05-05 00:00:00.000000+0000 |         R&D |         Intern |    Kajal |                           null |   1e+06
      1 | 2021-06-03 00:00:00.000000+0000 |  Deployment |        Manager |    Kusum |                {'abc', 'xyz'} | 1.5e+06
      2 | 2020-09-03 00:00:00.000000+0000 | Development | Web developer |    Karan | {'def', 'lmn', 'pqr', 'tuv'} | 1.7e+06
    121 | 2019-05-03 00:00:00.000000+0000 |     Testing |         Intern |    Kushi |                {'lab', 'jkl'} |   2e+06

(4 rows)
cqlsh> select * from Employee.Employee_Info;

 emp_id | date_of_joining                 | dept_name   | designation    | emp_name | projects                       | salary
--------+---------------------------------+-------------+----------------+----------+--------------------------------+---------
      1 | 2021-06-03 00:00:00.000000+0000 |  Deployment |        Manager |    Kusum |                {'abc', 'xyz'} | 1.5e+06
      2 | 2020-09-03 00:00:00.000000+0000 | Development | Web developer |    Karan | {'def', 'lmn', 'pqr', 'tuv'} | 1.7e+06
    121 | 2019-05-03 00:00:00.000000+0000 |     Testing |         Intern |    Kushi |                {'lab', 'jkl'} |   2e+06

(3 rows)
cqlsh>
```

BDA LAB 2

Perform the following DB operations using Cassandra:

1 Create a key space by name Library

```
cqlsh> CREATE KEYSPACE LIBRARY WITH replication = {'class':'SimpleStrategy','replication_factor':3};
cqlsh> Use LIBRARY;
cqlsh:library>
```

2. Create a column family by name Library-Info with attributes Stud_Id Primary Key, Counter_value of type Counter,Stud_Name, Book-Name, Book-Id, Date_of_issue.

```
cqlsh:library> create table library_info(stud_id int, counter_value Counter, stud_name text,book_name text, date_of_issue timestamp, book_id int, PRIMARY KEY(stud_id,stud_name,book_name,date_of_issue,book_id));
```

```
cqlsh:library> select * from library.library_info;

 stud_id | stud_name | book_name | date_of_issue | book_id | counter_value
---------+-----------+-----------+---------------+---------+---------------

(0 rows)
```

3. Insert the values into the table in batch

```
cqlsh:library> UPDATE library_info SET counter_value = counter_value + 1 WHERE stud_id = 111 and stud_name = 'SAM' and book_name = 'ML' and date_of_issue = '2020-10-11' and book_id = 200;
cqlsh:library> UPDATE library_info SET counter_value = counter_value + 1 WHERE stud_id = 112 and stud_name = 'SHAAN' and book_name = 'BDA' and date_of_issue = '2020-09-21' and book_id = 300;
cqlsh:library> UPDATE library_info SET counter_value = counter_value + 1 WHERE stud_id = 113 and stud_name = 'AYMAN' and book_name = 'OOMD' and date_of_issue = '2020-04-01' and book_id = 400;
```

```
cqlsh:library> select * from library.library_info;

 stud_id | stud_name | book_name | date_of_issue                   | book_id | counter_value
---------+-----------+-----------+---------------------------------+---------+---------------
     111 |       SAM |        ML | 2020-10-10 18:30:00.000000+0000 |     200 |             1
     113 |     AYMAN |      OOMD | 2020-03-31 18:30:00.000000+0000 |     400 |             1
     112 |     SHAAN |       BDA | 2020-09-20 18:30:00.000000+0000 |     300 |             1

(3 rows)
```

4. Display the details of the table created and increase the value of the counter

```
cqlsh:library> UPDATE library_info SET counter_value = counter_value + 1 WHERE stud_id = 112 and stud_name = 'SHAAN' and book_name = 'BDA' and date_of_issue = '2020-09-21' and book_id = 300;
```

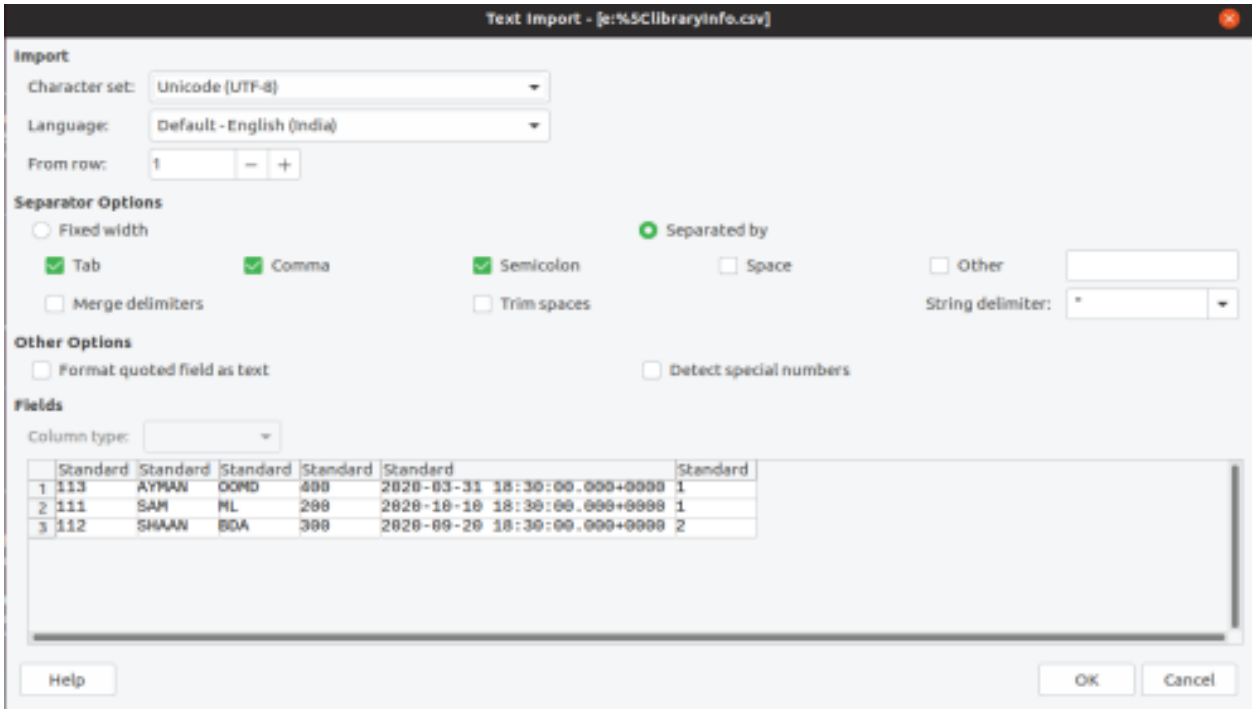
```
cqlsh:library> select * from library.library_info;

 stud_id | stud_name | book_name | date_of_issue                   | book_id | counter_value
---------+-----------+-----------+---------------------------------+---------+---------------
     111 |       SAM |        ML | 2020-10-10 18:30:00.000000+0000 |     200 |             1
     113 |     AYMAN |      OOMD | 2020-03-31 18:30:00.000000+0000 |     400 |             1
     112 |     SHAAN |       BDA | 2020-09-20 18:30:00.000000+0000 |     300 |             2

(3 rows)
```

5. Write a query to show that a student with id 112 has taken a book “BDA” 2 times.

```
cqlsh:library> SELECT * FROM library_info WHERE stud_id = 112;

 stud_id | stud_name | book_name | date_of_issue                   | book_id | counter_value
---------+-----------+-----------+---------------------------------+---------+---------------
     112 |     SHAAN |       BDA | 2020-09-20 18:30:00.000000+0000 |     300 |             2

(1 rows)
```

6. Export the created column to a csv file

```
cqlsh:library> COPY Library_Info(Stud_Id,Stud_Name,Book_Name,Book_Id,Date_Of_Issue,Counter_value) TO 'e:\libraryInfo.csv';
Using 11 child processes

Starting copy of library.library_info with columns [stud_id, stud_name, book_name, book_id, date_of_issue, counter_value].
Processed: 3 rows; Rate:      17 rows/s; Avg. rate:      17 rows/s
3 rows exported to 1 files in 0.204 seconds.
```



7. Import a given csv dataset from local file system into Cassandra column family

```
cqlsh:library> create table library_info2(stud_id int, counter_value Counter, stud_name text,book_name text, date_of_issue timestamp, book_id int, PRIMARY KEY(stud_id,stud_name,book_name,date_of_issue,
book_id));
```

```
cqlsh:library> SELECT * FROM library_info2;

 stud_id | stud_name | book_name | date_of_issue | book_id | counter_value
---------+-----------+-----------+---------------+---------+---------------

(0 rows)
```

```
cqlsh:library> COPY library_info2(stud_id,stud_name,book_name,book_id,date_of_issue,counter_value) FROM 'e:\libraryInfo.csv';
Using 11 child processes

Starting copy of library.library_info2 with columns [stud_id, stud_name, book_name, book_id, date_of_issue, counter_value].
Processed: 3 rows; Rate:       5 rows/s; Avg. rate:       7 rows/s
3 rows imported from 1 files in 0.416 seconds (0 skipped).
```

BDA LAB 3

Mongo (CRUD)

```
bmsce@bmsce-OptiPlex-3060:~$ mongo
MongoDB shell version v4.0.28
connecting to: mongodb://127.0.0.1:27017/?gssapiServiceName=mongodb
Implicit session: session { "id" : UUID("c2e3109b-0341-483b-ba3a-f9fb3b1aed87") }
MongoDB server version: 4.0.28
Server has startup warnings:
2022-04-11T14:03:08.254+0530 I STORAGE  [initandlisten]
2022-04-11T14:03:08.254+0530 I STORAGE  [initandlisten] ** WARNING: Using the XFS filesystem is strongly recommended with the WiredTiger storage engine
2022-04-11T14:03:08.254+0530 I STORAGE  [initandlisten] **          See http://dochub.mongodb.org/core/prodnotes-filesystem
2022-04-11T14:03:10.024+0530 I CONTROL  [initandlisten]
2022-04-11T14:03:10.024+0530 I CONTROL  [initandlisten] ** WARNING: Access control is not enabled for the database.
2022-04-11T14:03:10.024+0530 I CONTROL  [initandlisten] **          Read and write access to data and configuration is unrestricted.
2022-04-11T14:03:10.024+0530 I CONTROL  [initandlisten]
---
Enable MongoDB's free cloud-based monitoring service, which will then receive and display
metrics about your deployment (disk utilization, CPU, operation statistics, etc).

The monitoring data will be available on a MongoDB website with a unique URL accessible to you
and anyone you share the URL with. MongoDB may use this information to make product
improvements and to suggest MongoDB products and deployment options to you.

To enable free monitoring, run the following command: db.enableFreeMonitoring()
To permanently disable this reminder, run the following command: db.disableFreeMonitoring()
---

> db;
test
> use lab1DB;
switched to db lab1DB
> db;
lab1DB
```

```
> db.Student.update({_id:5, StudName:"Kusum", Grade:"VI"}, {$set:{Hobbies: ["Golf", "Sea Shell Collection"]}}, {upsert:true});
WriteResult({ "nMatched" : 0, "nUpserted" : 1, "nModified" : 0, "_id" : 5 })
> db.Student.find({}).pretty();
{
        "_id" : 1,
        "StudName" : "Jeevan",
        "Grade" : "VI",
        "Hobbies" : "InternetSurfing"
}
{
        "_id" : 2,
        "StudName" : "Vamsi",
        "Grade" : "VI",
        "Hobbies" : [
                "Watching Movies",
                "Reading Novels",
                "Drugs"
        ]
}
{ "_id" : 3, "StudName" : "Sharat", "Grade" : "V", "Hobbies" : "Reading" }
{
        "_id" : 5,
        "Grade" : "VI",
        "StudName" : "Kusum",
        "Hobbies" : [
                "Golf",
                "Sea Shell Collection"
        ]
}
> db.Student.find({StudName:"Kusum"});
{ "_id" : 5, "Grade" : "VI", "StudName" : "Kusum", "Hobbies" : [ "Golf", "Sea Shell Collection" ] }
```

```
> db.Student.find({StudName:"Kusum"}, {StudName:1, Grade:1}).pretty();
{ "_id" : 5, "Grade" : "VI", "StudName" : "Kusum" }
> db.Student.count();
4
> db.Student.find({}).sort({StudName:1});
{ "_id" : 1, "StudName" : "Jeevan", "Grade" : "VI", "Hobbies" : "InternetSurfing" }
{ "_id" : 5, "Grade" : "VI", "StudName" : "Kusum", "Hobbies" : [ "Golf", "Sea Shell Collection" ] }
{ "_id" : 3, "StudName" : "Sharat", "Grade" : "V", "Hobbies" : "Reading" }
{ "_id" : 2, "StudName" : "Vamsi", "Grade" : "VI", "Hobbies" : [ "Watching Movies", "Reading Novels", "Drugs" ] }
> db.Student.find({}).sort({StudName:-1});
{ "_id" : 2, "StudName" : "Vamsi", "Grade" : "VI", "Hobbies" : [ "Watching Movies", "Reading Novels", "Drugs" ] }
{ "_id" : 3, "StudName" : "Sharat", "Grade" : "V", "Hobbies" : "Reading" }
{ "_id" : 5, "Grade" : "VI", "StudName" : "Kusum", "Hobbies" : [ "Golf", "Sea Shell Collection" ] }
{ "_id" : 1, "StudName" : "Jeevan", "Grade" : "VI", "Hobbies" : "InternetSurfing" }
> db.Student.update({_id:5}, {$set:{Location:"NIHMANS"}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.Student.find({StudName:"Kusum"}).pretty();
{
        "_id" : 5,
        "Grade" : "VI",
        "StudName" : "Kusum",
        "Hobbies" : [
                "Golf",
                "Sea Shell Collection"
        ],
        "Location" : "NIHMANS"
}
```

```
> db.Student.find({}).skip(2).pretty();
{ "_id" : 3, "StudName" : "Sharat", "Grade" : "V", "Hobbies" : "Reading" }
{
        "_id" : 5,
        "Grade" : "VI",
        "StudName" : "Kusum",
        "Hobbies" : [
                "Golf",
                "Sea Shell Collection"
        ],
        "Location" : "NIHMANS"
}
> db.Student.find().skip(2).pretty();
{ "_id" : 3, "StudName" : "Sharat", "Grade" : "V", "Hobbies" : "Reading" }
{
        "_id" : 5,
        "Grade" : "VI",
        "StudName" : "Kusum",
        "Hobbies" : [
                "Golf",
                "Sea Shell Collection"
        ],
        "Location" : "NIHMANS"
}
> db.createCollection("food")
{ "ok" : 1 }
> db.food.insert({_id:1,fruits:['avacado','dragon fruit']})
WriteResult({ "nInserted" : 1 })
```

```
> db.food.insert({_id:1,fruits:['avacado','dragon fruit']})
WriteResult({ "nInserted" : 1 })
> db.food.insert({_id:2,fruits:['strawberry','dragon fruit']})
WriteResult({ "nInserted" : 1 })
> db.food.find({'fruits.1':'avacado'}).pretty()
> db.food.find().pretty()
{ "_id" : 1, "fruits" : [ "avacado", "dragon fruit" ] }
{ "_id" : 2, "fruits" : [ "strawberry", "dragon fruit" ] }
> db.food.find({'fruits.1':"avacado"}).pretty()
> db.food.find({'fruits.1':"avacado"})
> db.food.find({'fruits.0':"avacado"})
{ "_id" : 1, "fruits" : [ "avacado", "dragon fruit" ] }
> db.food.find({'fruits.0':"avacado"}).pretty()
{ "_id" : 1, "fruits" : [ "avacado", "dragon fruit" ] }
> db.food.find({'fruits.0':"avacado"}).pretty();
{ "_id" : 1, "fruits" : [ "avacado", "dragon fruit" ] }
> db.food.find({'fruits.0':{$size:2}}).pretty();
> db.food.find({'fruits':{$size:2}})
{ "_id" : 1, "fruits" : [ "avacado", "dragon fruit" ] }
{ "_id" : 2, "fruits" : [ "strawberry", "dragon fruit" ] }
> db.food.find({_id:2},{'fruits':{$slice:2}});
{ "_id" : 2, "fruits" : [ "strawberry", "dragon fruit" ] }
> db.food.find({_id:2},{'fruits':{$slice:1}});
{ "_id" : 2, "fruits" : [ "strawberry" ] }
> db.food.find({fruits:{$all:["avacado"]}})
{ "_id" : 1, "fruits" : [ "avacado", "dragon fruit" ] }
> db.food.find({fruits:{$all:["avacado","dragon fruit"]}})
{ "_id" : 1, "fruits" : [ "avacado", "dragon fruit" ] }
> db.food.find({fruits:{$all:["dragon fruit"]}})
{ "_id" : 1, "fruits" : [ "avacado", "dragon fruit" ] }
{ "_id" : 2, "fruits" : [ "strawberry", "dragon fruit" ] }
```

```
> show collections;
Student
customer
food
> db.customer.aggregate({$match:{AcctType:"FD"}},{$group:{_id:"$custID",TotalAccBal:{$sum:"$AcctBal"}}})
{ "_id" : 2, "TotalAccBal" : 20000000 }
{ "_id" : 1, "TotalAccBal" : 10000000 }
> db.customer.find()
{ "_id" : ObjectId("6253f945d7ce1043c6d5c8cc"), "custID" : 1, "AcctBal" : 10000000, "AcctType" : "FD" }
{ "_id" : ObjectId("6253f963d7ce1043c6d5c8cd"), "custID" : 2, "AcctBal" : 20000000, "AcctType" : "FD" }
{ "_id" : ObjectId("6253f973d7ce1043c6d5c8ce"), "custID" : 3, "AcctBal" : 30000000, "AcctType" : "RD" }
> db.customer.aggregate({$match:{AcctType:"FD"}},{$group:{_id:"$custID",TotalAccBal:{$sum:"$AcctBal"}}},{$match:{TotAccBal:{$gt:10000000}}});
> db.customer.aggregate({$match:{AcctType:"FD"}},{$group:{_id:"$custID",TotalAccBal:{$sum:"$AcctBal"}}},{$match:{TotalAccBal:{$gt:10000000}}});
{ "_id" : 2, "TotalAccBal" : 20000000 }
> quit()
```

1) Using MongoDB
i) Create a database for Students and Create a Student Collection (_id,Name, USN, Semester, Dept_Name, CGPA, Hobbies(Set)).
ii) Insert required documents to the collection.
iii) First Filter on “Dept_Name:CSE” and then group it on “Semester” and compute the Average CPGA for that semester and filter those documents where the “Avg_CPGA” is greater than 7.5.
iv) Command used to export MongoDB JSON documents from “Student” Collection into the “Students” database into a CSV file “Output.txt”.

```
> db.Student.insert({_id:1,Name:"Aravind",USN:"1BM19CS001",Sem:6,Dept_name:"CSE",CGPA:"9.6",Hobbies:"Badminton"});
WriteResult({ "nInserted" : 1 })
> db.Student.insert({_id:2,Name:"Aman",USN:"1BM19EC002",Sem:7,Dept_name:"ECE",CGPA:"9.1",Hobbies:"Swimming"});
WriteResult({ "nInserted" : 1 })
> db.Student.insert({_id:3,Name:"Latha",USN:"1BM19CS003",Sem:6,Dept_name:"CSE",CGPA:"8.1",Hobbies:"Reading"});
WriteResult({ "nInserted" : 1 })
> db.Student.insert({_id:4,Name:"Sam",USN:"1BM19CS004",Sem:6,Dept_name:"CSE",CGPA:"6.5",Hobbies:"Cycling"});
WriteResult({ "nInserted" : 1 })
> db.Student.insert({_id:5,Name:"Suman",USN:"1BM19CS005",Sem:5,Dept_name:"CSE",CGPA:"7.6",Hobbies:"Cycling"});
WriteResult({ "nInserted" : 1 })
```

```
> db.Student.update({_id:1},{$set:{CGPA:9.0}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.Student.update({_id:2},{$set:{CGPA:9.1}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.Student.update({_id:3},{$set:{CGPA:8.1}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.Student.update({_id:4},{$set:{CGPA:6.5}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.Student.update({_id:5},{$set:{CGPA:8.6}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.Student.aggregate({$match:{Dept_name:"CSE"}},{$group:{_id:"$Sem",AvgCGPA:{$avg:"$CGPA"}}},{$match:{AvgCGPA:{$gt:7.5}}});
{ "_id" : 5, "AvgCGPA" : 8.6 }
{ "_id" : 6, "AvgCGPA" : 7.866666666666667 }
```

```
bmsce@bmsce-Precision-T1700:~$ mongoexport --host localhost --db Niharika_db --collection Student --csv --out /home/bmsce/Desktop/output.txt --fields "_id","Name","USN","Sem","Dept-name","CGPA","Hobbies";
2022-04-20T15:04:30.836+0530    csv flag is deprecated; please use --type=csv instead
2022-04-20T15:04:30.836+0530    connected to: localhost
2022-04-20T15:04:30.836+0530    exported 5 records
```

Open

output.txt
~/Desktop

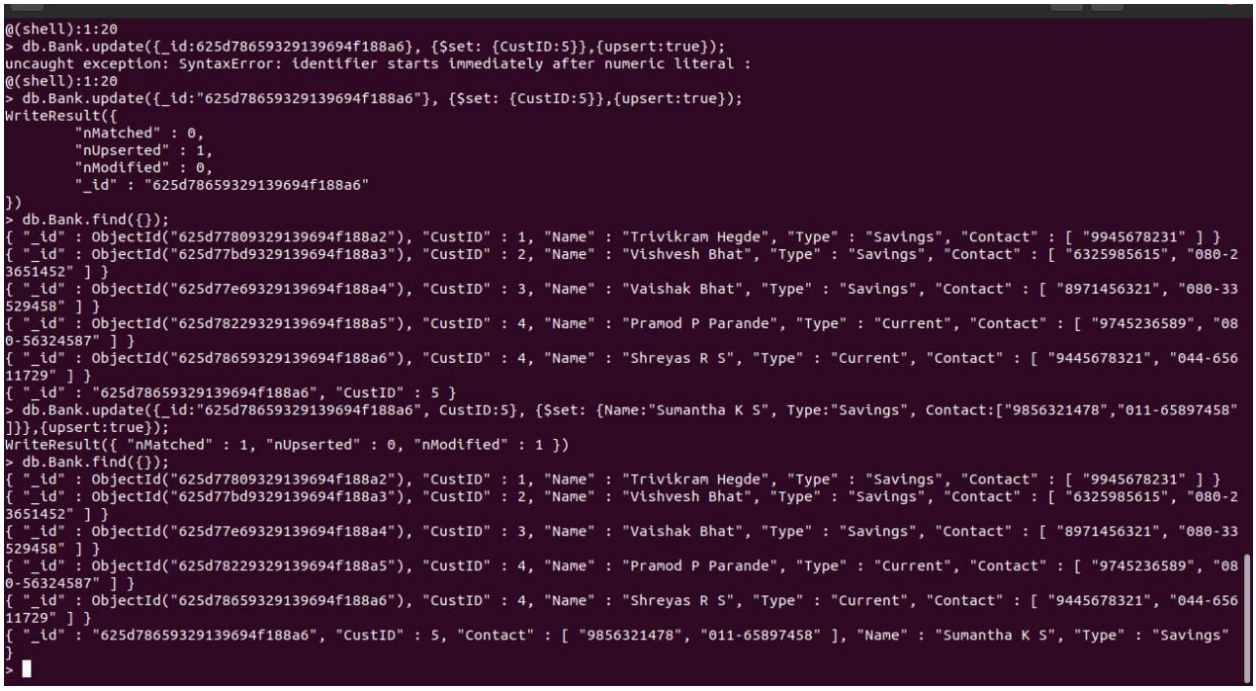
```
_id,Name,USN,Sem,Dept-name,CGPA,Hobbies
1,Aravind,1BM19CS001,6,,9,Badminton
2,Aman,1BM19EC002,7,,9.1,Swimming
3,Latha,1BM19CS003,6,,8.1,Reading
4,Sam,1BM19CS004,6,,6.5,Cycling
5,Suman,1BM19CS005,5,,8.6,Cycling
```

2)Create a mongodb collection Bank. Demonstrate the following by choosing fields of your choice.

1. Insert three documents
2. Use Arrays(Use Pull and Pop operation)
3. Use Index
4. Use Cursors
5. Updation

```
> db.createCollection("Bank");
{ "ok" : 1 }
> db.insert({CustID:1, Name:"Trivikram Hegde", Type:"Savings", Contact:["9945678231", "080-22364587"]});
uncaught exception: TypeError: db.insert is not a function :
@(shell):1:1
> db.Bank.insert({CustID:1, Name:"Trivikram Hegde", Type:"Savings", Contact:["9945678231", "080-22364587"]});
WriteResult({ "nInserted" : 1 })
> db.Bank.insert({CustID:2, Name:"Vishvesh Bhat", Type:"Savings", Contact:["6325985615", "080-23651452"]});
WriteResult({ "nInserted" : 1 })
> db.Bank.insert({CustID:3, Name:"Vaishak Bhat", Type:"Savings", Contact:["8971456321", "080-33529458"]});
WriteResult({ "nInserted" : 1 })
> db.Bank.insert({CustID:4, Name:"Pramod P Parande", Type:"Current", Contact:["9745236589", "080-56324587"]});
WriteResult({ "nInserted" : 1 })
> db.Bank.insert({CustID:4, Name:"Shreyas R S", Type:"Current", Contact:["9445678321","044-65611729", "080-25639856"]});
WriteResult({ "nInserted" : 1 })
> db.Bank.find({});
{ "_id" : ObjectId("625d77809329139694f188a2"), "CustID" : 1, "Name" : "Trivikram Hegde", "Type" : "Savings", "Contact" : [ "9945678231", "080
-22364587" ] }
{ "_id" : ObjectId("625d77bd9329139694f188a3"), "CustID" : 2, "Name" : "Vishvesh Bhat", "Type" : "Savings", "Contact" : [ "6325985615", "080-2
3651452" ] }
{ "_id" : ObjectId("625d77e69329139694f188a4"), "CustID" : 3, "Name" : "Vaishak Bhat", "Type" : "Savings", "Contact" : [ "8971456321", "080-33
529458" ] }
{ "_id" : ObjectId("625d78229329139694f188a5"), "CustID" : 4, "Name" : "Pramod P Parande", "Type" : "Current", "Contact" : [ "9745236589", "08
0-56324587" ] }
{ "_id" : ObjectId("625d78659329139694f188a6"), "CustID" : 4, "Name" : "Shreyas R S", "Type" : "Current", "Contact" : [ "9445678321", "044-656
11729", "080-25639856" ] }
> db.Bank.updateMany({CustID:1},{$pop:{Contact:1}} );
{ "acknowledged" : true, "matchedCount" : 1, "modifiedCount" : 1 }
> db.Bank.find({});
{ "_id" : ObjectId("625d77809329139694f188a2"), "CustID" : 1, "Name" : "Trivikram Hegde", "Type" : "Savings", "Contact" : [ "9945678231" ] }
{ "_id" : ObjectId("625d77bd9329139694f188a3"), "CustID" : 2, "Name" : "Vishvesh Bhat", "Type" : "Savings", "Contact" : [ "6325985615", "080-2
3651452" ] }
{ "_id" : ObjectId("625d77e69329139694f188a4"), "CustID" : 3, "Name" : "Vaishak Bhat", "Type" : "Savings", "Contact" : [ "8971456321", "080-33
529458" ] }
{ "_id" : ObjectId("625d78229329139694f188a5"), "CustID" : 4, "Name" : "Pramod P Parande", "Type" : "Current", "Contact" : [ "9745236589", "08
0-56324587" ] }
{ "_id" : ObjectId("625d78659329139694f188a6"), "CustID" : 4, "Name" : "Shreyas R S", "Type" : "Current", "Contact" : [ "9445678321", "044-656
11729", "080-25639856" ] }
```

```
{ "_id" : ObjectId("625d78659329139694f188a6"), "CustID" : 4, "Name" : "Shreyas R S", "Type" : "Current", "Contact" : [ "9445678321", "044-656
11729", "080-25639856" ] }
> db.Bank.updateMany({},{$pull:{Contact:"080-25639856"}} );
{ "acknowledged" : true, "matchedCount" : 5, "modifiedCount" : 1 }
> db.Bank.find({});
{ "_id" : ObjectId("625d77809329139694f188a2"), "CustID" : 1, "Name" : "Trivikram Hegde", "Type" : "Savings", "Contact" : [ "9945678231" ] }
{ "_id" : ObjectId("625d77bd9329139694f188a3"), "CustID" : 2, "Name" : "Vishvesh Bhat", "Type" : "Savings", "Contact" : [ "6325985615", "080-2
3651452" ] }
{ "_id" : ObjectId("625d77e69329139694f188a4"), "CustID" : 3, "Name" : "Vaishak Bhat", "Type" : "Savings", "Contact" : [ "8971456321", "080-33
529458" ] }
{ "_id" : ObjectId("625d78229329139694f188a5"), "CustID" : 4, "Name" : "Pramod P Parande", "Type" : "Current", "Contact" : [ "9745236589", "08
0-56324587" ] }
{ "_id" : ObjectId("625d78659329139694f188a6"), "CustID" : 4, "Name" : "Shreyas R S", "Type" : "Current", "Contact" : [ "9445678321", "044-656
11729" ] }
> db.Bank.createIndex({Name:1, Type:1},{name:});
uncaught exception: SyntaxError: expected expression, got '}' :
@(shell):1:43
> db.Bank.createIndex({Name:1, Type:1},{name:"Find current account holders"});
{
        "createdCollectionAutomatically" : false,
        "numIndexesBefore" : 1,
        "numIndexesAfter" : 2,
        "ok" : 1
}
> db.Bank.find({});
{ "_id" : ObjectId("625d77809329139694f188a2"), "CustID" : 1, "Name" : "Trivikram Hegde", "Type" : "Savings", "Contact" : [ "9945678231" ] }
{ "_id" : ObjectId("625d77bd9329139694f188a3"), "CustID" : 2, "Name" : "Vishvesh Bhat", "Type" : "Savings", "Contact" : [ "6325985615", "080-2
3651452" ] }
{ "_id" : ObjectId("625d77e69329139694f188a4"), "CustID" : 3, "Name" : "Vaishak Bhat", "Type" : "Savings", "Contact" : [ "8971456321", "080-33
529458" ] }
{ "_id" : ObjectId("625d78229329139694f188a5"), "CustID" : 4, "Name" : "Pramod P Parande", "Type" : "Current", "Contact" : [ "9745236589", "08
0-56324587" ] }
{ "_id" : ObjectId("625d78659329139694f188a6"), "CustID" : 4, "Name" : "Shreyas R S", "Type" : "Current", "Contact" : [ "9445678321", "044-656
11729" ] }
> db.Bank.getIndexes()
[
        {
                "v" : 2,
```



1) Using MongoDB,
i) Create a database for Faculty and Create a Faculty Collection(Faculty_id, Name, Designation ,Department, Age, Salary, Specialization(Set)).
ii) Insert required documents to the collection.
iii) First Filter on “Dept_Name:MECH” and then group it on “Designation” and compute the Average Salary for that Designation and filter those documents where the “Avg_Sal” is greater than 650000.
iv) Demonstrate usage of import and export commands

Write MongoDB queries for the following:
1)To display only the product name from all the documents of the product collection.
2)To display only the Product ID, ExpiryDate as well as the quantity from the document of the product collection where the _id column is 1.
3)To find those documents where the price is not set to 15000.
4)To find those documents from the Product collection where the quantity is set to 9 and the product name is set to ‘monitor’.
5)To find documents from the Product collection where the Product name ends in ‘d’.

```
}
> db.createCollection("faculty");
{ "ok" : 1 }
> db.faculty.insert({_id:1,name:"Dr. Balaraman Ravindran",designation:"Professor",department:"CSE",age:45,salary:100000,specialization:['pytho
n','mysql','sklearn', 'tensorflow']});
WriteResult({ "nInserted" : 1 })
> db.faculty.insert({_id:2,name:"Dr. Mahadev Ghorki",designation:"Assistant Professor",department:"CSE",age:35,salary:80000,specialization:['p
ython','numpy','sklearn', 'tensorflow', 'java']});
WriteResult({ "nInserted" : 1 })
> db.faculty.insert({_id:3,name:"Dr. Praveen Borade",designation:"Associate Professor",department:"ME",age:40,salary:75000,specialization:['au
tocad', 'aerodynamics', 'thermal physics']});
WriteResult({ "nInserted" : 1 })
> db.faculty.insert({_id:4,name:"Dr. Madhav Nayak",designation:"Assistant Professor",department:"ME",age:37,salary:95000,specialization:['auto
cad', 'flight-dynamics', 'Finite Element Analysis']});
WriteResult({ "nInserted" : 1 })
> db.faculty.aggregate ( {$match:{department:"ME"}}, {$group : {_id : "$designation", AverageSal :{$avg:"$salary"} } },  {$match:{AverageSal:{
$gt:50000}}});
{ "_id" : "Associate Professor", "AverageSal" : 75000 }
{ "_id" : "Assistant Professor", "AverageSal" : 95000 }
> db.createCollection("product");
{ "ok" : 1 }
> db.product.insert({pid:1,pname:"keyboard",mdate:2001,price:1800,quantity:2});
WriteResult({ "nInserted" : 1 })
> db.product.insert({pid:2,pname:"mouse",mdate:2005,price:1500,quantity:5});
WriteResult({ "nInserted" : 1 })
> db.product.insert({pid:3,pname:"monitor",mdate:2015,price:10000,quantity:9});
WriteResult({ "nInserted" : 1 })
> db.product.insert({pid:4,pname:"motherboard",mdate:2021,price:15000,quantity:4});
WriteResult({ "nInserted" : 1 })
> db.product.find({},{pname:1,_id:0})
{ "pname" : "keyboard" }
{ "pname" : "mouse" }
{ "pname" : "monitor" }
{ "pname" : "motherboard" }
> db.product.find({pid:1},{pid:1,_id:0,mdate:1,quantity:1});
{ "pid" : 1, "mdate" : 2001, "quantity" : 2 }
>  db.product.find({price:{$ne:15000}},{pname:1,_id:0});
{ "pname" : "keyboard" }
```

3)Create a mongodb collection Hospital. Demonstrate the following by choosing fields of your choice.

1. Insert three documents
2. Use Arrays(Use Pull and Pop operation)
3. Use Index
4. Use Cursors
5. Updation

```
{ "pname" : "motherboard" }
> db.product.find({pid:1},{pid:1,_id:0,mdate:1,quantity:1});
{ "pid" : 1, "mdate" : 2001, "quantity" : 2 }
>  db.product.find({price:{$ne:15000}},{pname:1,_id:0});
{ "pname" : "keyboard" }
{ "pname" : "mouse" }
{ "pname" : "monitor" }
> db.product.find({$and:[{quantity:{$eq:9}},{pname:{$eq:"monitor"}}]},{pname:1,_id:0})
{ "pname" : "monitor" }
> db.product.find({pname:/d$/},{pname:1,quantity:1,_id:0})
{ "pname" : "keyboard", "quantity" : 2 }
{ "pname" : "motherboard", "quantity" : 4 }
> db.createCollection("hospital");
{ "ok" : 1 }
> db.hospital.insert({_id:1, Name: "Anshuman Agarwal", age:23, diseases:["fever", "diarrhoea", "wheezing", "gastritis"]});
WriteResult({ "nInserted" : 1 })
> db.hospital.insert({_id:2, Name: "Pinky Chaubey", age:35, diseases:["fever","nausea", "food infection", "indigestion", "kidney stones"]});
WriteResult({ "nInserted" : 1 })
> db.hospital.insert({_id:3, Name: "Amresh Chowpati", age:63, diseases:["hyperglycemia", "diabetes mellitus", "food poisoning", "cold"]});
WriteResult({ "nInserted" : 1 })
> db.hospital.updateMany({},{$pull:{diseases:"fever"}});
{ "acknowledged" : true, "matchedCount" : 3, "modifiedCount" : 2 }
> db.hospital.updateOne({_id:1},{$pop:{diseases:-1}});
{ "acknowledged" : true, "matchedCount" : 1, "modifiedCount" : 1 }
> db.hospital.find({"diseases.2":"nausea"});
> db.hospital.find({"diseases.1":"nausea"});
> d.hospital.find({});
uncaught exception: ReferenceError: d is not defined :
@(shell):1:1
> db.hospital.find({});
{ "_id" : 1, "Name" : "Anshuman Agarwal", "age" : 23, "diseases" : [ "wheezing", "gastritis" ] }
{ "_id" : 2, "Name" : "Pinky Chaubey", "age" : 35, "diseases" : [ "nausea", "food infection", "indigestion", "kidney stones" ] }
{ "_id" : 3, "Name" : "Amresh Chowpati", "age" : 63, "diseases" : [ "hyperglycemia", "diabetes mellitus", "food poisoning", "cold" ] }
> db.hospital.find({"diseases.0":"nausea"});
{ "_id" : 2, "Name" : "Pinky Chaubey", "age" : 35, "diseases" : [ "nausea", "food infection", "indigestion", "kidney stones" ] }
> db.hospital.update({_id:3},{$set:{'diseases.1':'sarscov'}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

BDA LAB 4

4. Hadoop Installation

```
shashi@Shashis-MacBook-Air-2 ~ % hadoop -version
ERROR: -version is not COMMAND nor fully qualified CLASSNAME.
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
  or    hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
    where CLASSNAME is a user-provided Java class

  OPTIONS is none or any of:

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

  SUBCOMMAND is one of:


    Admin Commands:

daemonlog     get/set the log level for each daemon

    Client Commands:

archive       create a Hadoop archive
checknative   check native Hadoop and compression libraries availability
classpath     prints the class path needed to get the Hadoop jar and the
              required libraries
conftest      validate configuration XML files
credential    interact with credential providers
distch        distributed metadata changer
distcp        copy file or directories recursively
dtutil        operations related to delegation tokens
envvars       display computed Hadoop environment variables
fs            run a generic filesystem user client
gridmix       submit a mix of synthetic job, modeling a profiled from
              production load
jar <jar>     run a jar file. NOTE: please use "yarn jar" to launch YARN
              applications, not this command.
jnipath       prints the java.library.path
kdiag         Diagnose Kerberos Problems
kerbname      show auth_to_local principal conversion
key           manage keys via the KeyProvider
rumenfolder   scale a rumen input trace
rumentrace    convert logs into a rumen trace
s3guard       manage metadata on S3
trace         view and modify Hadoop tracing settings
version       print the version

    Daemon Commands:

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

SUBCOMMAND may print help when invoked w/o parameters or with -h.
```

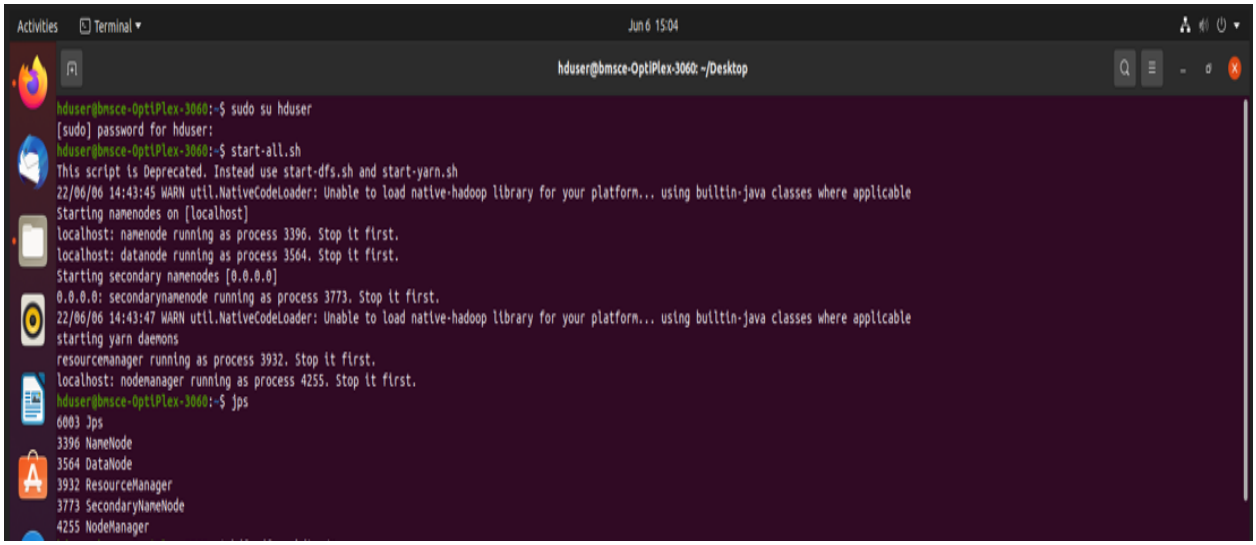
```
shashi@Shashis-MacBook-Air-2 ~ % hadoop fs -mkdir hadoopDir
```

```
shashi@Shashis-MacBook-Air-2 ~ % hadoop fs -ls
2022-07-10 17:34:47,585 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 32 items
-r--------   1 shashi staff          7 2022-04-19 02:59 .CFUserTextEncoding
-rw-r--r--   1 shashi staff      12292 2022-07-10 17:33 .DS_Store
drwx------   - shashi staff       3008 2022-07-05 09:24 .Trash
drwxr-xr-x   - shashi staff         64 2022-06-09 21:16 .cassandra
drwx------   - shashi staff        128 2022-06-01 22:11 .config
drwxr-xr-x   - shashi staff         96 2022-06-02 21:47 .ipython
drwxr-xr-x   - shashi staff         96 2022-06-11 11:15 .jupyter
drwxr-xr-x   - shashi staff        160 2022-06-01 20:34 .npm
-rw-------   1 shashi staff         65 2022-07-04 19:00 .python_history
-rw-------   1 shashi staff       1583 2022-06-22 13:17 .viminfo
drwxr-xr-x   - shashi staff        128 2022-05-22 16:35 .vscode
-rw-r--r--   1 shashi staff        208 2022-06-01 20:26 .zprofile
-rw-------   1 shashi staff       6437 2022-07-10 17:30 .zsh_history
drwx------   - shashi staff        192 2022-07-10 17:30 .zsh_sessions
-rw-r--r--   1 shashi staff         36 2022-05-22 16:29 .zshrc
drwx------   - shashi staff         96 2022-05-23 17:59 Applications
drwx------   - shashi staff        448 2022-07-04 18:54 Desktop
drwx------   - shashi staff        448 2022-07-05 09:25 Documents
drwx------   - shashi staff       3616 2022-07-09 19:39 Downloads
drwxr-xr-x   - shashi staff        448 2022-05-22 18:01 DragonMS
drwxr-xr-x   - shashi staff        128 2022-06-26 15:58 Graphs
drwxr-xr-x   - shashi staff        128 2022-06-30 16:47 Kusum
drwx------   - shashi staff       2560 2022-06-02 21:47 Library
drwxr-xr-x   - shashi staff         64 2022-07-10 17:29 Makedir
drwx------   - shashi staff        192 2022-05-23 17:56 Movies
drwx------   - shashi staff        128 2022-05-06 04:01 Music
drwx------   - shashi staff        160 2022-05-06 20:14 Pictures
drwxr-xr-x   - shashi staff        160 2022-06-05 14:04 Public
drwxr-xr-x   - shashi staff        320 2022-06-01 22:20 eth-todo-list
-rw-r--r--   1 shashi staff    2680203 2022-05-22 16:14 get-pip.py
drwxr-xr-x   - shashi staff        544 2022-05-22 18:15 hacking-with-python
drwxr-xr-x   - shashi staff         64 2022-07-10 17:31 hadoopDir
```

BDA LAB 5

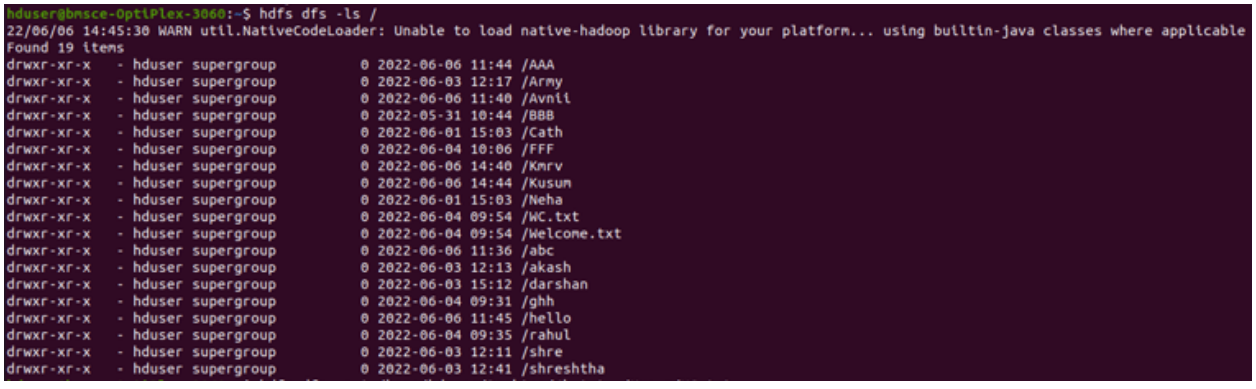
5.Execution of HDFS Commands for interaction with Hadoop Environment. (Minimum 10 commands to be executed)

1. jps



2. mkdir

3. ls



4. put

```
hduser@bmsce-OptiPlex-3060:~$ hdfs dfs -put /home/hduser/Desktop/6b.txt  /Kusum/WC.txt
22/06/06 14:46:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@bmsce-OptiPlex-3060:~$ hdfs dfs -cat /Kusum/WC.txt
22/06/06 14:47:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hello fron 6B

D
B

B
```

5. copyFromLocal

```
hduser@bmsce-OptiPlex-3060:~$ hdfs dfs -put /home/hduser/Desktop/6b.txt  /Kusum/newWC.txt
22/06/06 14:50:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@bmsce-OptiPlex-3060:~$ hdfs dfs -cat /Kusum/newWC.txt
22/06/06 14:50:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hello fron 6B

D
B

B
```

6. Get

i)

```
hduser@bmsce-OptiPlex-3060:~$ hdfs dfs -get /Kusum/WC.txt /home/hduser/Downloads/newWWC.txt
22/06/06 14:51:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@bmsce-OptiPlex-3060:~$ cd Downloads
hduser@bmsce-OptiPlex-3060:~/Downloads$ cat newWWC.txt
hello fron 6B

D
B

B
```

ii)

```
hduser@bmsce-OptiPlex-3060:~$ hdfs dfs -ls /Kusum/
22/06/06 14:54:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 hduser supergroup         23 2022-06-06 14:46 /Kusum/WC.txt
-rw-r--r--   1 hduser supergroup         23 2022-06-06 14:50 /Kusum/newWC.txt
hduser@bmsce-OptiPlex-3060:~$ hdfs dfs -getmerge /Kusum/WC.txt /Kusum/newWC.txt /home/hduser/Desktop/newmerge.txt
22/06/06 14:55:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@bmsce-OptiPlex-3060:~$ cd Desktop
hduser@bmsce-OptiPlex-3060:~/Desktop$ cat newmerge.txt
hello fron 6B

D
B

B

hello fron 6B

D
B

B
```

iii)

```
hduser@bmsce-OptiPlex-3060:~/Desktop$ hadoop fs -getfacl /Kusum/
22/06/06 14:56:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
# file: /Kusum
# owner: hduser
# group: supergroup
user::rwx
group::r-x
other::r-x
```

7. copyToLocal

```
hduser@bmsce-OptiPlex-3060:~/Desktop$ hdfs dfs -copyToLocal /Kusum/WC.txt /home/hduser/Desktop
22/06/06 14:58:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@bmsce-OptiPlex-3060:~/Desktop$ cat WC.txt
hello fron 6B

D
B

B
```

8. cat

```
hduser@bmsce-OptiPlex-3060:~/Desktop$ hdfs dfs -cat /Kusum/WC.txt
22/06/06 14:58:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hello fron 6B

D
B

B
```

9. Mv

```
hduser@bmsce-OptiPlex-3060:~/Desktop$ hadoop fs -mv /Kusum /FFF
22/06/06 14:59:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@bmsce-OptiPlex-3060:~/Desktop$ hadoop fs -ls /FFF
22/06/06 15:00:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - hduser supergroup          0 2022-06-06 14:50 /FFF/Kusun
-rw-r--r--   1 hduser supergroup         17 2022-06-04 10:06 /FFF/WC.txt
```

10. cp

```
hduser@bmsce-OptiPlex-3060:~/Desktop$ hadoop fs -cp /FFF/ /LLL
22/06/06 15:09:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@bmsce-OptiPlex-3060:~/Desktop$ hadoop fs -ls /LLL
22/06/06 15:10:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - hduser supergroup          0 2022-06-06 15:09 /LLL/Kusun
-rw-r--r--   1 hduser supergroup         17 2022-06-06 15:09 /LLL/WC.txt
hduser@bmsce-OptiPlex-3060:~/Desktop$
```

BDA LAB 6

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

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));
  }
}
```

b) find the mean max temperature for every month

MeanMax
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));
  }
}
```

```
hduser@bmsce-Precision-T1700:~/Desktop/temperature$  start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-
yarn.shStarting namenodes on [localhost]
hduser@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser-namenode-bmsce-
Precision-T1700.out
hduser@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser-datanode-
bmsce-Precision-T1700.out
Starting secondary namenodes
[0.0.0.0]hduser@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-
hduser-secondarynamenode-bmsce-Precision-T1700.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resourcemanager-bmsce-
Precision-T1700.out
hduser@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduser-nodemanager-bmsce-
Precision-T1700.out
hduser@bmsce-Precision-T1700:~/Desktop/temperature$
jps6832 NodeManager
6498 ResourceManager
6339 SecondaryNameNode
4887 org.eclipse.equinox.launcher_1.5.600.v20191014-2022.jar
6954 Jps
6123 DataNode
5951 NameNode
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -le /
-le: Unknown command
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs
-ls /Found 31 items
 drwxr-xr-          -   supergroup         0   2022-06-   12:   /CSE
 x             hduser                              06    35

 drwxr-xr-          -   supergroup         0   2022-06-   12:   /FFF
 x             hduser                              06    23

 drwxr-xr-          -   supergroup         0   2022-06-   12:   /LLL
 x             hduser                              06    36

 drwxr-xr-          -   supergroup         0   2022-06-   12:   /amit_bda
 x             hduser                              20    06

 drwxr-xr-          -   supergroup         0   2022-06-   11:   /amit_lab
 x             hduser                              27    42

 drwxr-xr-          -   supergroup         0   2022-06-   14:   /bharath
 x             hduser                              03    52

 drwxr-xr-          -   supergroup         0   2022-06-   14:   /bharath035
 x             hduser                              03    43

 drwxr-xr-          -   supergroup         0   2022-06-   14:   /chi
 x             hduser                              24    54

 drwxr-xr-          -   supergroup         0   2022-05-   10:   /example
 x             hduser                              31    21

 drwxr-xr-          -   supergroup         0   2022-06-   15:   /foldernew
 x             hduser                              01    13

 drwxr-xr-          -   supergroup         0   2022-06-   15:   /hemang061
 x             hduser                              06    04

 drwxr-xr-          -   supergroup         0   2022-06-   15:   /input_kusum
 x             hduser                              20    16
```

drwxr-xr-x	-	hduser	supergroup	0	2022-06-03	12:27 /irfan
drwxr-xr-x	-	hduser	supergroup	0	2022-06-22	10:44 /lwde
drwxr-xr-x	-	hduser	supergroup	0	2022-06-27	13:03 /mapreducejoin_amit
drwxr-xr-x	-	hduser	supergroup	0	2022-06-22	15:32 /muskan
drwxr-xr-x	-	hduser	supergroup	0	2022-06-22	15:06 /muskan_op
drwxr-xr-x	-	hduser	supergroup	0	2022-06-22	15:35 /muskan_output
drwxr-xr-x	-	hduser	supergroup	0	2022-06-06	15:04 /new_folder
drwxr-xr-x	-	hduser	supergroup	0	2022-05-31	10:26 /one
drwxr-xr-x	-	hduser	supergroup	0	2022-06-24	15:30 /out55
drwxr-xr-x	-	hduser	supergroup	0	2022-06-20	12:17 /output
drwxr-xr-x	-	hduser	supergroup	0	2022-06-27	13:04 /output_TOPn
drwxr-xr-x	-	hduser	supergroup	0	2022-06-27	12:14 /output_Topn
drwxr-xr-x	-	hduser	supergroup	0	2022-06-24	12:42 /r1
drwxr-xr-x	-	hduser	supergroup	0	2022-06-24	12:24 /rgs

```
drwxr-xr-x   - hduser supergroup          0 2022-06-03 12:08 /kusum
drwxrwxr-x   - hduser supergroup          0 2019-08-01 16:19 /tmp
drwxr-xr-x   - hduser supergroup          0 2019-08-01 16:03 /user
drwxr-xr-x   - hduser supergroup          0 2022-06-01 09:46 /user1
-rw-r--r--   1 hduser supergroup       2436 2022-06-24 12:17 /wc.jar
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -mkdir /kusum_temperature
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -put ./1901 /kusum_temperature
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -put ./1902 /kusum_temperature
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -ls /kusum_temperature
Found 2 items
-rw-r--r--   1 hduser supergroup     888190 2022-06-27 14:47 /kusum_temperature/1901
-rw-r--r--   1 hduser supergroup     888978 2022-06-27 14:47 /kusum_temperature/1902
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hadoop jar ./avgtemp.jar AverageDriver
/kusum_temperature/1901 /kusum_temperature/output/
Exception in thread "main" java.lang.ClassNotFoundException:
 AverageDriverat
 java.net.URLClassLoader.findClass(URLClassLoader.java:382)
 a                                                        t
 java.lang.ClassLoader.loadClass(ClassLoader.java:418)at
 java.lang.ClassLoader.loadClass(ClassLoader.java:351) at
 java.lang.Class.forName0(Native Method)
 at java.lang.Class.forName(Class.java:348)
 at  org.apache.hadoop.util.RunJar.run(RunJar.java:214)
 at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hadoop jar ./avgtemp.jar
temperature.AverageDriver /kusum_temperature/1901 /kusum_temperature/output/ 22/06/27
14:53:27 INFO Configuration.deprecation: session.id is deprecated.  Instead, use
dfs.metrics.session-id
22/06/27 14:53:27 INFO jvm.JvmMetrics: Initializing JVM Metrics with
processName=JobTracker,sessionId=
22/06/27 14:53:27 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not
performed. Implement the Tool interface and execute your application with ToolRunner to remedy
this.
22/06/27 14:53:27 INFO input.FileInputFormat: Total input paths to process :
122/06/27 14:53:27 INFO mapreduce.JobSubmitter: number of splits:1
22/06/27 14:53:28 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local254968295_000122/06/27 14:53:28 INFO mapreduce.Job: The url to track the job: http://
localhost:8080/
22/06/27 14:53:28 INFO mapreduce.Job: Running job: job_local254968295_0001
22/06/27 14:53:28 INFO mapred.LocalJobRunner: OutputCommitter set in config
null22/06/27 14:53:28 INFO mapred.LocalJobRunner: OutputCommitter is
org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
22/06/27 14:53:28 INFO mapred.LocalJobRunner: Waiting for map tasks
22/06/27 14:53:28 INFO mapred.LocalJobRunner: Starting task:
attempt_local254968295_0001_m_000000_022/06/27 14:53:28 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
22/06/27 14:53:28 INFO mapred.MapTask: Processing
split:hdfs://localhost:54310/kusum_temperature/
1901:0+888190
22/06/27 14:53:28 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
22/06/27 14:53:28 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
10022/06/27 14:53:28 INFO mapred.MapTask: soft limit at 83886080
22/06/27 14:53:28 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
22/06/27 14:53:28 INFO mapred.MapTask: kvstart = 26214396; length =
655360022/06/27 14:53:28 INFO mapred.MapTask: Map output collector class =
org.apache.hadoop.mapred.MapTask$MapOutputBuffer
22/06/27 14:53:28 INFO mapred.LocalJobRunner:
22/06/27 14:53:28 INFO mapred.MapTask: Starting flush of map
output22/06/27 14:53:28 INFO mapred.MapTask: Spilling map output
22/06/27 14:53:28 INFO mapred.MapTask: bufstart = 0; bufend = 59076; bufvoid = 104857600
22/06/27 14:53:28 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188144(104752576);
length = 26253/6553600
22/06/27 14:53:28 INFO mapred.MapTask: Finished spill 0
```

```
22/06/27 14:53:28 INFO mapred.Task: Task:attempt_local254968295_0001_m_000000_0 is done. And is
inthe process of committing
22/06/27 14:53:28 INFO mapred.LocalJobRunner: map
22/06/27 14:53:28 INFO mapred.Task: Task 'attempt_local254968295_0001_m_000000_0' done.
22/06/27 14:53:28 INFO mapred.LocalJobRunner: Finishing task:
attempt_local254968295_0001_m_000000_022/06/27 14:53:28 INFO mapred.LocalJobRunner: map task
executor complete.
22/06/27 14:53:28 INFO mapred.LocalJobRunner: Waiting for reduce tasks
22/06/27 14:53:28 INFO mapred.LocalJobRunner: Starting task:
attempt_local254968295_0001_r_000000_022/06/27 14:53:28 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
22/06/27 14:53:28 INFO mapred.ReduceTask: Using
ShuffleConsumerPlugin:
org.apache.hadoop.mapreduce.task.reduce.Shuffle@262cb2a9
22/06/27 14:53:28 INFO reduce.MergeManagerImpl: MergerManager:
memoryLimit=349752512,maxSingleShuffleLimit=87438128, mergeThreshold=230836672,
ioSortFactor=10, memToMemMergeOutputsThreshold=10
22/06/27 14:53:28 INFO reduce.EventFetcher: attempt_local254968295_0001_r_000000_0 Thread
started:EventFetcher for fetching Map Completion Events
22/06/27 14:53:28 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of
mapattempt_local254968295_0001_m_000000_0 decomp: 72206 len: 72210 to MEMORY
22/06/27 14:53:28 INFO reduce.InMemoryMapOutput: Read 72206 bytes from map-output
forattempt_local254968295_0001_m_000000_0
22/06/27 14:53:28 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size:
72206,inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->72206
22/06/27 14:53:28 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning22/06/27 14:53:28 INFO mapred.LocalJobRunner: 1 / 1 copied.
22/06/27 14:53:28 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and
0on-disk map-outputs
22/06/27 14:53:28 INFO mapred.Merger: Merging 1 sorted segments
22/06/27 14:53:28 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
totalsize: 72199 bytes
22/06/27 14:53:28 INFO reduce.MergeManagerImpl: Merged 1 segments, 72206 bytes to disk to
satisfyreduce memory limit
22/06/27 14:53:28 INFO reduce.MergeManagerImpl: Merging 1 files, 72210 bytes from disk
22/06/27 14:53:28 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into
reduce22/06/27 14:53:28 INFO mapred.Merger: Merging 1 sorted segments
22/06/27 14:53:28 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
totalsize: 72199 bytes
22/06/27 14:53:28 INFO mapred.LocalJobRunner: 1 / 1 copied.
22/06/27 14:53:28 INFO Configuration.deprecation: mapred.skip.on is deprecated.  Instead,
usemapreduce.job.skiprecords
22/06/27 14:53:28 INFO mapred.Task: Task:attempt_local254968295_0001_r_000000_0 is done. And is
inthe process of committing
22/06/27 14:53:28 INFO mapred.LocalJobRunner: 1 / 1 copied.
22/06/27 14:53:28 INFO mapred.Task: Task attempt_local254968295_0001_r_000000_0 is allowed to
commitnow
22/06/27 14:53:28 INFO output.FileOutputCommitter: Saved output of task
'attempt_local254968295_0001_r_000000_0' to hdfs://localhost:54310/kusum_temperature/output/
_temporary/0/task_local254968295_0001_r_00000022/06/27 14:53:28 INFO mapred.LocalJobRunner: reduce >
reduce
22/06/27 14:53:28 INFO mapred.Task: Task 'attempt_local254968295_0001_r_000000_0' done.
22/06/27 14:53:28 INFO mapred.LocalJobRunner: Finishing task:
attempt_local254968295_0001_r_000000_022/06/27 14:53:28 INFO mapred.LocalJobRunner: reduce task
executor complete.
22/06/27 14:53:29 INFO mapreduce.Job: Job job_local254968295_0001 running in uber mode :
false22/06/27 14:53:29 INFO mapreduce.Job:  map 100% reduce 100%
22/06/27 14:53:29 INFO mapreduce.Job: Job job_local254968295_0001 completed
successfully22/06/27 14:53:29 INFO mapreduce.Job: Counters: 38
 File System Counters
 FILE: Number of bytes read=153102
```

```
FILE: Number of bytes written=723014
FILE: Number of read operations=0
FILE: Number of large read
operations=0FILE: Number of write
operations=0 HDFS: Number of bytes
read=1776380 HDFS: Number of bytes
written=8
HDFS: Number of read operations=13
HDFS: Number of large read
operations=0HDFS: Number of write
operations=4
Map-Reduce Framework
Map input records=6565
Map output
records=6564Map output
bytes=59076
Map output materialized
bytes=72210Input split bytes=112
Combine input records=0
Combine output records=0
Reduce input groups=1
Reduce shuffle
bytes=72210Reduce input
records=6564 Reduce output
records=1 Spilled
Records=13128 Shuffled
Maps =1
Failed Shuffles=0
Merged  Map  outputs=1
GC time elapsed
(ms)=55CPU time spent
(ms)=0
Physical memory (bytes)
snapshot=0Virtual memory (bytes)
snapshot=0
Total committed heap usage
(bytes)=999292928Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=
0WRONG_MAP=0
WRONG_REDUCE=0
File Input Format
CountersBytes Read=888190
File Output Format
CountersBytes Written=8
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -ls /kusum_temperature/output/
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2022-06-27 14:53 /kusum_temperature/output/_SUCCESS
-rw-r--r--   1 hduser supergroup          8 2022-06-27 14:53 /kusum_temperature/output/part-r-
00000
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -cat /kusum_temperature/output/part-
r-00000
1901    46
hduser@bmsce-Precision-T1700:~/Desktop/temperature$
```

BDA LAB 7

7. For a given Text file, Create a Map Reduce program to sort the content in an alphabetic order listing only top n 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 vo```\\id 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));
    }
  }
}
```

```
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -mkdir /kusum_topn
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -put ./input.txt /kusum_topn/
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -ls /kusum_topn/
Found 1 items
-rw-r--r--   1 hduser supergroup        103 2022-06-27 15:43 /kusum_topn/input.txt
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hadoop jar topn.jar TopNDriver
/kusum_topn/input.txt /kusum_topn/output
Exception in thread "main" java.lang.ClassNotFoundException:
 TopNDriverat
 java.net.URLClassLoader.findClass(URLClassLoader.java:382)
 a                                                      t
 java.lang.ClassLoader.loadClass(ClassLoader.java:418)at
 java.lang.ClassLoader.loadClass(ClassLoader.java:351) at
 java.lang.Class.forName0(Native Method)
 at java.lang.Class.forName(Class.java:348)
 at  org.apache.hadoop.util.RunJar.run(RunJar.java:214)
 at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hadoop jar topn.jar topn.TopNDriver
/kusum_topn/input.txt /kusum_topn/output
22/06/27 15:45:22 INFO Configuration.deprecation: session.id is deprecated.  Instead,
usedfs.metrics.session-id
22/06/27 15:45:22 INFO jvm.JvmMetrics: Initializing JVM Metrics with
processName=JobTracker,sessionId=
22/06/27 15:45:22 INFO input.FileInputFormat: Total input paths to process :
122/06/27 15:45:22 INFO mapreduce.JobSubmitter: number of splits:1
22/06/27 15:45:22 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local691635730_000122/06/27 15:45:22 INFO mapreduce.Job: The url to track the job: http://
localhost:8080/
22/06/27 15:45:22 INFO mapreduce.Job: Running job: job_local691635730_0001
22/06/27 15:45:22 INFO mapred.LocalJobRunner: OutputCommitter set in config
null22/06/27 15:45:22 INFO mapred.LocalJobRunner: OutputCommitter is
org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
22/06/27 15:45:22 INFO mapred.LocalJobRunner: Waiting for map tasks
22/06/27 15:45:22 INFO mapred.LocalJobRunner: Starting task:
attempt_local691635730_0001_m_000000_022/06/27 15:45:22 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
22/06/27 15:45:22 INFO mapred.MapTask: Processing
split:hdfs://localhost:54310/kusum_topn/input.txt:0+103
22/06/27 15:45:22 INFO mapred.MapTask:  (EQUATOR) 0 kvi 26214396(104857584)
22/06/27 15:45:22 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
10022/06/27 15:45:22 INFO mapred.MapTask: soft limit at 83886080
22/06/27 15:45:22 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
22/06/27 15:45:22 INFO mapred.MapTask: kvstart = 26214396; length =
655360022/06/27 15:45:22 INFO mapred.MapTask: Map output collector class =
org.apache.hadoop.mapred.MapTask$MapOutputBuffer
22/06/27 15:45:22 INFO mapred.LocalJobRunner:
22/06/27 15:45:22 INFO mapred.MapTask: Starting flush of map
output22/06/27 15:45:22 INFO mapred.MapTask: Spilling map output
22/06/27 15:45:22 INFO mapred.MapTask: bufstart = 0; bufend = 187; bufvoid = 104857600
22/06/27 15:45:22 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214316(104857264);
length = 81/6553600
22/06/27 15:45:22 INFO mapred.MapTask: Finished spill 0
22/06/27 15:45:22 INFO mapred.Task: Task:attempt_local691635730_0001_m_000000_0 is done. And is
inthe process of committing
22/06/27 15:45:22 INFO mapred.LocalJobRunner: map
22/06/27 15:45:22 INFO mapred.Task: Task 'attempt_local691635730_0001_m_000000_0' done.
22/06/27 15:45:22 INFO mapred.LocalJobRunner: Finishing task:
attempt_local691635730_0001_m_000000_022/06/27 15:45:22 INFO mapred.LocalJobRunner: map task
executor complete.
22/06/27 15:45:22 INFO mapred.LocalJobRunner: Waiting for reduce tasks
22/06/27 15:45:22 INFO mapred.LocalJobRunner: Starting task:
attempt_local691635730_0001_r_000000_022/06/27 15:45:22 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
```

```
22/06/27 15:45:22 INFO mapred.ReduceTask: Using
ShuffleConsumerPlugin:
org.apache.hadoop.mapreduce.task.reduce.Shuffle@40a5e65a
22/06/27 15:45:22 INFO reduce.MergeManagerImpl: MergerManager:
memoryLimit=334338464,maxSingleShuffleLimit=83584616, mergeThreshold=220663392,
ioSortFactor=10, memToMemMergeOutputsThreshold=10
22/06/27 15:45:22 INFO reduce.EventFetcher: attempt_local691635730_0001_r_000000_0 Thread
started:EventFetcher for fetching Map Completion Events
22/06/27 15:45:22 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of
mapattempt_local691635730_0001_m_000000_0 decomp: 231 len: 235 to MEMORY
22/06/27 15:45:22 INFO reduce.InMemoryMapOutput: Read 231 bytes from map-output
forattempt_local691635730_0001_m_000000_0
22/06/27 15:45:22 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size:
231, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->231
22/06/27 15:45:22 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning22/06/27 15:45:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
22/06/27 15:45:22 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and
0on-disk map-outputs
22/06/27 15:45:22 INFO mapred.Merger: Merging 1 sorted segments
22/06/27 15:45:22 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
totalsize: 226 bytes
22/06/27 15:45:22 INFO reduce.MergeManagerImpl: Merged 1 segments, 231 bytes to disk to
satisfyreduce memory limit
22/06/27 15:45:22 INFO reduce.MergeManagerImpl: Merging 1 files, 235 bytes from disk
22/06/27 15:45:22 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into
reduce22/06/27 15:45:22 INFO mapred.Merger: Merging 1 sorted segments
22/06/27 15:45:22 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
totalsize: 226 bytes
22/06/27 15:45:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
22/06/27 15:45:22 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead,
usemapreduce.job.skiprecords
22/06/27 15:45:23 INFO mapred.Task: Task:attempt_local691635730_0001_r_000000_0 is done. And is
inthe process of committing
22/06/27 15:45:23 INFO mapred.LocalJobRunner: 1 / 1 copied.
22/06/27 15:45:23 INFO mapred.Task: Task attempt_local691635730_0001_r_000000_0 is allowed to
commitnow
22/06/27 15:45:23 INFO output.FileOutputCommitter: Saved output of task
'attempt_local691635730_0001_r_000000_0' to hdfs://localhost:54310/kusum_topn/output/
_temporary/0/task_local691635730_0001_r_00000022/06/27 15:45:23 INFO mapred.LocalJobRunner:
reduce > reduce
22/06/27 15:45:23 INFO mapred.Task: Task 'attempt_local691635730_0001_r_000000_0' done.
22/06/27 15:45:23 INFO mapred.LocalJobRunner: Finishing task:
attempt_local691635730_0001_r_000000_022/06/27 15:45:23 INFO mapred.LocalJobRunner: reduce task
executor complete.
22/06/27 15:45:23 INFO mapreduce.Job: Job job_local691635730_0001 running in uber mode :
false22/06/27 15:45:23 INFO mapreduce.Job:  map 100% reduce 100%
22/06/27 15:45:23 INFO mapreduce.Job: Job job_local691635730_0001 completed
successfully22/06/27 15:45:23 INFO mapreduce.Job: Counters: 38
 File System Counters
 FILE: Number of bytes read=18078
 FILE: Number of bytes
 written=516697FILE: Number of read
 operations=0
 FILE: Number of large read
 operations=0FILE: Number of write
 operations=0 HDFS: Number of bytes
 read=206
 HDFS: Number of bytes written=105
 HDFS: Number of read operations=13
 HDFS: Number of large read
 operations=0HDFS: Number of write
 operations=4
 Map-Reduce Framework
```

```
    Map input records=6
    Map output
    records=21Map output
    bytes=187
    Map output materialized
    bytes=235Input split bytes=110
    Combine input records=0
    Combine output
    records=0Reduce input
    groups=15 Reduce shuffle
    bytes=235Reduce input
    records=21 Reduce output
    records=15Spilled
    Records=42 Shuffled Maps
    =1
    Failed Shuffles=0
    Merged  Map  outputs=1
    GC time elapsed
    (ms)=42CPU time spent
    (ms)=0
    Physical memory (bytes)
    snapshot=0Virtual memory (bytes)
    snapshot=0
    Total committed heap usage
    (bytes)=578289664Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=
    OWRONG_MAP=0
    WRONG_REDUCE=0
    File Input Format
    CountersBytes Read=103
    File Output Format
    CountersBytes Written=105
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -ls /kusum_topn/output/
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2022-06-27 15:45 /kusum_topn/output/_SUCCESS
-rw-r--r--   1 hduser supergroup        105 2022-06-27 15:45 /kusum_topn/output/part-r-00000
hduser@bmsce-Precision-T1700:~/Desktop/temperature$ hdfs dfs -cat /kusum_topn/output/part-r-00000
 hadoo  4
 p

 i3

 am     2

 hi     1

 im     1

 is     1

 there  1

 bye    1

learing 1
awesome 1
love    1
kusum 1
cool    1
and     1
using   1
hduser@bmsce-Precision-T1700:~/Desktop/temperature$
```

BDA LAB 8

8. Create a Map Reduce program to demonstrating join operation

```
// JoinDriver.java
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.mapred.lib.MultipleInputs;
import org.apache.hadoop.util.*;

public class JoinDriver extends Configured implements Tool {

public static class KeyPartitioner implements Partitioner<TextPair, Text> {
@Override
public void configure(JobConf job) {}

@Override
public int getPartition(TextPair key, Text value, int numPartitions) {
return (key.getFirst().hashCode() & Integer.MAX_VALUE) %
numPartitions;
}
}

@Override

public int run(String[] args) throws Exception {

if (args.length != 3) {
System.out.println("Usage: <Department Emp Strength input>

<Department Name input> <output>");
return -1;
}

JobConf conf = new JobConf(getConf(), getClass());

conf.setJobName("Join 'Department Emp Strength input' with 'Department Name
input'");

Path AInputPath = new Path(args[0]);
Path BInputPath = new Path(args[1]);
Path outputPath = new Path(args[2]);

MultipleInputs.addInputPath(conf, AInputPath, TextInputFormat.class,

Posts.class);

MultipleInputs.addInputPath(conf, BInputPath, TextInputFormat.class,

User.class);
```

```
FileOutputFormat.setOutputPath(conf, outputPath);

conf.setPartitionerClass(KeyPartitioner.class);

conf.setOutputValueGroupingComparator(TextPair.FirstComparator.class);

conf.setMapOutputKeyClass(TextPair.class);

conf.setReducerClass(JoinReducer.class);

conf.setOutputKeyClass(Text.class);

JobClient.runJob(conf);

return 0;
}

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

int exitCode = ToolRunner.run(new JoinDriver(), args);
System.exit(exitCode);
}
}

// JoinReducer.java
import java.io.IOException;
import java.util.Iterator;

import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;

public class JoinReducer extends MapReduceBase implements Reducer<TextPair, Text, Text,
Text> {

@Override
public void reduce (TextPair key, Iterator<Text> values, OutputCollector<Text, Text>
output, Reporter reporter)

throws IOException
{

Text nodeId = new Text(values.next());
while (values.hasNext()) {

Text node = values.next();
Text outValue = new Text(nodeId.toString() + "\t\t" + node.toString());
output.collect(key.getFirst(), outValue);
}
}
}

// User.java
import java.io.IOException;
```

```
import java.util.Iterator;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FSDataInputStream;
import org.apache.hadoop.fs.FSDataOutputStream;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;

import org.apache.hadoop.io.IntWritable;

public class User extends MapReduceBase implements Mapper<LongWritable, Text, TextPair,
Text> {

@Override
public void map(LongWritable key, Text value, OutputCollector<TextPair, Text> output,
Reporter reporter)

throws IOException

{

String valueString = value.toString();

String[] SingleNodeData = valueString.split("\t");
output.collect(new TextPair(SingleNodeData[0], "1"), new

Text(SingleNodeData[1]));
}
}

//Posts.java
import java.io.IOException;

import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;

public class Posts extends MapReduceBase implements Mapper<LongWritable, Text, TextPair,
Text> {

@Override
public void map(LongWritable key, Text value, OutputCollector<TextPair, Text> output,
Reporter reporter)
throws IOException
{
String valueString = value.toString();
String[] SingleNodeData = valueString.split("\t");
output.collect(new TextPair(SingleNodeData[3], "0"), new

Text(SingleNodeData[9]));
}
}
```

```
// TextPair.java
import java.io.*;

import org.apache.hadoop.io.*;

public class TextPair implements WritableComparable<TextPair> {

private Text first;
private Text second;

public TextPair() {
set(new Text(), new Text());
}

public TextPair(String first, String second) {
set(new Text(first), new Text(second));
}

public TextPair(Text first, Text second) {
set(first, second);
}

public void set(Text first, Text second) {
this.first = first;
this.second = second;
}

public Text getFirst() {
return first;
}

public Text getSecond() {
return second;
}

@Override
public void write(DataOutput out) throws IOException {
first.write(out);
second.write(out);
}

@Override
public void readFields(DataInput in) throws IOException {
first.readFields(in);
second.readFields(in);
}

@Override
public int hashCode() {
return first.hashCode() * 163 + second.hashCode();
}
```

```
@Override
public boolean equals(Object o) {
if (o instanceof TextPair) {
TextPair tp = (TextPair) o;
return first.equals(tp.first) && second.equals(tp.second);
}
return false;
}

@Override
public String toString() {
return first + "\t" + second;
}

@Override
public int compareTo(TextPair tp) {
int cmp = first.compareTo(tp.first);
if (cmp != 0) {
return cmp;
}
return second.compareTo(tp.second);
}
// ^^ TextPair

// vv TextPairComparator
public static class Comparator extends WritableComparator {
private static final Text.Comparator TEXT_COMPARATOR = new Text.Comparator();
public Comparator() {
super(TextPair.class);
}
@Override
public int compare(byte[] b1, int s1, int l1,
byte[] b2, int s2, int l2) {
try {
int firstL1 = WritableUtils.decodeVIntSize(b1[s1]) + readVInt(b1, s1);
int firstL2 = WritableUtils.decodeVIntSize(b2[s2]) + readVInt(b2, s2);
int cmp = TEXT_COMPARATOR.compare(b1, s1, firstL1, b2, s2, firstL2);
if (cmp != 0) {
return cmp;
}
return TEXT_COMPARATOR.compare(b1, s1 + firstL1, l1 - firstL1,

b2, s2 + firstL2, l2 - firstL2);
} catch (IOException e) {
throw new IllegalArgumentException(e);
}
}
}
static {
WritableComparator.define(TextPair.class, new Comparator());
}
public static class FirstComparator extends WritableComparator {
private static final Text.Comparator TEXT_COMPARATOR = new Text.Comparator();
```

```
public FirstComparator() {
super(TextPair.class);
}

@Override
public int compare(byte[] b1, int s1, int l1,
byte[] b2, int s2, int l2) {
try {
int firstL1 = WritableUtils.decodeVIntSize(b1[s1]) + readVInt(b1, s1);
int firstL2 = WritableUtils.decodeVIntSize(b2[s2]) + readVInt(b2, s2);
return TEXT_COMPARATOR.compare(b1, s1, firstL1, b2, s2, firstL2);
} catch (IOException e) {
throw new IllegalArgumentException(e);
}
}
@Override
public int compare(WritableComparable a, WritableComparable b) {
if (a instanceof TextPair && b instanceof TextPair) {
return ((TextPair) a).first.compareTo(((TextPair) b).first);
}
return super.compare(a, b);
}
} }
```

```
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$ hdfs dfs -ls /kusum_joinls:
`/kusum_join': No such file or directory
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$ hdfs dfs -mkdir /kusum_join
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$ hdfs dfs -ls /kusum_join
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$ hdfs dfs -put ./DeptName.txt
/kusum_join/
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$ hdfs dfs -put ./DeptStrength.txt
/kusum_join/
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$ hadoop jar MapReduceJoin.jar
/kusum_join/DeptName.txt  /kusum_join/DeptStrength.txt  /kusum_join/output/ 22/06/27
15:12:24 INFO Configuration.deprecation: session.id is deprecated.  Instead, use
dfs.metrics.session-id
22/06/27 15:12:24 INFO jvm.JvmMetrics: Initializing JVM Metrics with
processName=JobTracker,sessionId=
22/06/27 15:12:24 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker,sessionId= - already initialized
22/06/27 15:12:24 INFO mapred.FileInputFormat: Total input paths to process :
122/06/27 15:12:24 INFO mapred.FileInputFormat: Total input paths to process :
122/06/27 15:12:24 INFO mapreduce.JobSubmitter: number of splits:2
22/06/27 15:12:24 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local1238804660_000122/06/27 15:12:24 INFO mapreduce.Job: The url to track the job: http://
localhost:8080/
22/06/27 15:12:24 INFO mapred.LocalJobRunner: OutputCommitter set in config
null22/06/27 15:12:24 INFO mapreduce.Job: Running job: job_local1238804660_0001
22/06/27 15:12:24 INFO mapred.LocalJobRunner: OutputCommitter is
org.apache.hadoop.mapred.FileOutputCommitter
22/06/27 15:12:24 INFO mapred.LocalJobRunner: Waiting for map tasks
22/06/27 15:12:24 INFO mapred.LocalJobRunner: Starting task:
attempt_local1238804660_0001_m_000000_022/06/27 15:12:24 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
22/06/27 15:12:24 INFO mapred.MapTask: Processing
split:hdfs://localhost:54310/kusum_join/
DeptName.txt:0+59 22/06/27 15:12:24 INFO mapred.MapTask:
numReduceTasks: 1
22/06/27 15:12:24 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
22/06/27 15:12:24 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
10022/06/27 15:12:24 INFO mapred.MapTask: soft limit at 83886080
22/06/27 15:12:24 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
22/06/27 15:12:24 INFO mapred.MapTask: kvstart = 26214396; length =
6553600 22/06/27 15:12:24 INFO mapred.MapTask: Map output collector class =
org.apache.hadoop.mapred.MapTask$MapOutputBuffer
22/06/27 15:12:24 INFO mapred.LocalJobRunner:
22/06/27 15:12:24 INFO mapred.MapTask: Starting flush of map
output22/06/27 15:12:24 INFO mapred.MapTask: Spilling map output
22/06/27 15:12:24 INFO mapred.MapTask: bufstart = 0; bufend = 63; bufvoid = 104857600
22/06/27 15:12:24 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214384(104857536);
length = 13/6553600
22/06/27 15:12:24 INFO mapred.MapTask: Finished spill 0
22/06/27 15:12:24 INFO mapred.Task: Task:attempt_local1238804660_0001_m_000000_0 is done.  And is
inthe process of committing
22/06/27 15:12:24 INFO mapred.LocalJobRunner: hdfs://localhost:54310/kusum_join/DeptName.txt:0+59
22/06/27 15:12:24 INFO mapred.Task: Task 'attempt_local1238804660_0001_m_000000_0' done.
22/06/27 15:12:24 INFO mapred.LocalJobRunner: Finishing
task:attempt_local1238804660_0001_m_000000_0
22/06/27 15:12:24 INFO mapred.LocalJobRunner: Starting task:
attempt_local1238804660_0001_m_000001_022/06/27 15:12:24 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
22/06/27 15:12:24 INFO mapred.MapTask: Processing split:
hdfs://localhost:54310/kusum_join/DeptStrength.txt:0+50
22/06/27 15:12:24 INFO mapred.MapTask: numReduceTasks: 1
22/06/27 15:12:24 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
22/06/27 15:12:24 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
```

```
22/06/27 15:12:24 INFO mapred.MapTask: soft limit at 83886080
22/06/27 15:12:24 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
22/06/27 15:12:24 INFO mapred.MapTask: kvstart = 26214396; length =
655360022/06/27 15:12:24 INFO mapred.MapTask: Map output collector class =
org.apache.hadoop.mapred.MapTask$MapOutputBuffer
22/06/27 15:12:24 INFO mapred.LocalJobRunner:
22/06/27 15:12:24 INFO mapred.MapTask: Starting flush of map
output22/06/27 15:12:24 INFO mapred.MapTask: Spilling map output
22/06/27 15:12:24 INFO mapred.MapTask: bufstart = 0; bufend = 54; bufvoid = 104857600
22/06/27 15:12:24 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214384(104857536);
length = 13/6553600
22/06/27 15:12:24 INFO mapred.MapTask: Finished spill 0
22/06/27 15:12:24 INFO mapred.Task: Task:attempt_local1238804660_0001_m_000001_0 is done. And is
inthe process of committing
22/06/27 15:12:24 INFO mapred.LocalJobRunner: hdfs://
localhost:54310/kusum_join/DeptStrength.txt:0+50
22/06/27 15:12:24 INFO mapred.Task: Task 'attempt_local1238804660_0001_m_000001_0'
done.22/06/27 15:12:24 INFO mapred.LocalJobRunner: Finishing task:
attempt_local1238804660_0001_m_000001_0
22/06/27 15:12:24 INFO mapred.LocalJobRunner: map task executor
complete.22/06/27 15:12:24 INFO mapred.LocalJobRunner: Waiting for reduce
tasks
22/06/27 15:12:24 INFO mapred.LocalJobRunner: Starting task:
attempt_local1238804660_0001_r_000000_022/06/27 15:12:24 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
22/06/27 15:12:24 INFO mapred.ReduceTask: Using
ShuffleConsumerPlugin:
org.apache.hadoop.mapreduce.task.reduce.Shuffle@45cb1c
22/06/27 15:12:24 INFO reduce.MergeManagerImpl: MergerManager:
memoryLimit=334338464,maxSingleShuffleLimit=83584616, mergeThreshold=220663392,
ioSortFactor=10, memToMemMergeOutputsThreshold=10
22/06/27 15:12:24 INFO reduce.EventFetcher: attempt_local1238804660_0001_r_000000_0 Thread
started:EventFetcher for fetching Map Completion Events
22/06/27 15:12:24 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of
mapattempt_local1238804660_0001_m_000001_0 decomp: 64 len: 68 to MEMORY
22/06/27 15:12:24 INFO reduce.InMemoryMapOutput: Read 64 bytes from map-output
forattempt_local1238804660_0001_m_000001_0
22/06/27 15:12:24 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size:
64,inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->64
22/06/27 15:12:24 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of
mapattempt_local1238804660_0001_m_000000_0 decomp: 73 len: 77 to MEMORY
22/06/27 15:12:24 INFO reduce.InMemoryMapOutput: Read 73 bytes from map-output
forattempt_local1238804660_0001_m_000000_0
22/06/27 15:12:24 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size:
73,inMemoryMapOutputs.size() -> 2, commitMemory -> 64, usedMemory ->137
22/06/27 15:12:24 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning22/06/27 15:12:24 INFO mapred.LocalJobRunner: 2 / 2 copied.
22/06/27 15:12:24 INFO reduce.MergeManagerImpl: finalMerge called with 2 in-memory map-outputs and
0on-disk map-outputs
22/06/27 15:12:24 INFO mapred.Merger: Merging 2 sorted segments
22/06/27 15:12:24 INFO mapred.Merger: Down to the last merge-pass, with 2 segments left of
totalsize: 121 bytes
22/06/27 15:12:24 INFO reduce.MergeManagerImpl: Merged 2 segments, 137 bytes to disk to
satisfyreduce memory limit
22/06/27 15:12:24 INFO reduce.MergeManagerImpl: Merging 1 files, 139 bytes from disk
22/06/27 15:12:24 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into
reduce22/06/27 15:12:24 INFO mapred.Merger: Merging 1 sorted segments
22/06/27 15:12:24 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
totalsize: 127 bytes
22/06/27 15:12:24 INFO mapred.LocalJobRunner: 2 / 2 copied.
```

```
22/06/27 15:12:24 INFO mapred.Task: Task:attempt_local1238804660_0001_r_000000_0 is done. And is
inthe process of committing
22/06/27 15:12:24 INFO mapred.LocalJobRunner: 2 / 2 copied.
22/06/27 15:12:24 INFO mapred.Task: Task attempt_local1238804660_0001_r_000000_0 is allowed
tocommit now
22/06/27 15:12:24 INFO output.FileOutputCommitter: Saved output of task
'attempt_local1238804660_0001_r_000000_0' to hdfs://localhost:54310/kusum_join/output/
_temporary/0/task_local1238804660_0001_r_00000022/06/27 15:12:24 INFO mapred.LocalJobRunner:
reduce > reduce
22/06/27 15:12:24 INFO mapred.Task: Task 'attempt_local1238804660_0001_r_000000_0'
done.22/06/27 15:12:24 INFO mapred.LocalJobRunner: Finishing task:
attempt_local1238804660_0001_r_000000_0
22/06/27 15:12:24 INFO mapred.LocalJobRunner: reduce task executor complete.
22/06/27 15:12:25 INFO mapreduce.Job: Job job_local1238804660_0001 running in uber mode :
false22/06/27 15:12:25 INFO mapreduce.Job:  map 100% reduce 100%
22/06/27 15:12:25 INFO mapreduce.Job: Job job_local1238804660_0001 completed
successfully22/06/27 15:12:25 INFO mapreduce.Job: Counters: 38
 File System Counters
 FILE: Number of bytes read=26370
 FILE: Number of bytes
 written=782865FILE: Number of read
 operations=0
 FILE: Number of large read
 operations=0FILE: Number of write
 operations=0 HDFS: Number of bytes
 read=277
 HDFS: Number of bytes written=85
 HDFS: Number of read operations=28
 HDFS: Number of large read
 operations=0HDFS: Number of write
 operations=5
 Map-Reduce
 FrameworkMap input
 records=8 Map output
 records=8Map output
 bytes=117
 Map output materialized
 bytes=145Input split bytes=443
 Combine input records=0
 Combine output
 records=0Reduce input
 groups=4 Reduce shuffle
 bytes=145Reduce input
 records=8 Reduce output
 records=4 Spilled
 Records=16 Shuffled Maps
 =2
 Failed Shuffles=0
 Merged Map outputs=2
 GC time elapsed
 (ms)=2CPU time spent
 (ms)=0
 Physical memory (bytes)
 snapshot=0Virtual memory (bytes)
 snapshot=0
 Total committed heap usage
 (bytes)=916979712Shuffle Errors
 BAD_ID=0
 CONNECTION=0
 IO_ERROR=0
 WRONG_LENGTH=
 0WRONG_MAP=0
 WRONG_REDUCE=0
 File Input Format
 CountersBytes Read=0
```

```
  File Output Format
  CountersBytes Written=85
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$ hdfs dfs -ls /kusum_join/output/
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2022-06-27 15:12 /kusum_join/output/_SUCCESS
-rw-r--r--   1 hduser supergroup         85 2022-06-27 15:12 /kusum_join/output/part-00000
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$ hdfs dfs -cat /kusum_join/output/part-
00000
A11     Finance             50
B12     HR          100
C13     Manufacturing       250
Dept_ID Dept_Name           Total_Employee
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$ hadoop jar MapReduceJoin.jar
/kusum_join/DeptStrength.txt  /kusum_join/DeptName.txt  /kusum_join/output/ 22/06/27
15:15:17 INFO Configuration.deprecation: session.id is deprecated. Instead, use
dfs.metrics.session-id
22/06/27 15:15:17 INFO jvm.JvmMetrics: Initializing JVM Metrics with
processName=JobTracker,sessionId=
22/06/27 15:15:17 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker,sessionId= - already initialized
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException: Output
directoryhdfs://localhost:54310/kusum_join/output already exists
 at
 org.apache.hadoop.mapred.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:132)at
 org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:564)
 at
 org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:432)at
 org.apache.hadoop.mapreduce.Job$10.run(Job.java:1296)
 at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1293)
 at java.security.AccessController.doPrivileged(Native
 Method)at javax.security.auth.Subject.doAs(Subject.java:422)
 at
 org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1628)at
 org.apache.hadoop.mapreduce.Job.submit(Job.java:1293)
 at
 org.apache.hadoop.mapred.JobClient$1.run(JobClient.java:562)at
 org.apache.hadoop.mapred.JobClient$1.run(JobClient.java:557)at
 java.security.AccessController.doPrivileged(Native Method) at
 javax.security.auth.Subject.doAs(Subject.java:422)
 at
 org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1628)at
 org.apache.hadoop.mapred.JobClient.submitJobInternal(JobClient.java:557)
 at
 org.apache.hadoop.mapred.JobClient.submitJob(JobClient.java:548)at
 org.apache.hadoop.mapred.JobClient.runJob(JobClient.java:833) at
 MapReduceJoin.JoinDriver.run(JoinDriver.java:53)
 a                                                    t
 org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)at
 org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84) at
 MapReduceJoin.JoinDriver.main(JoinDriver.java:60)
 at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
 at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
 at
 sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)at
 java.lang.reflect.Method.invoke(Method.java:498)
 at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
 at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$ hadoop jar MapReduceJoin.jar
/kusum_join/DeptStrength.txt /kusum_join/DeptName.txt /kusum_join/output2/ 22/06/27
15:15:26 INFO Configuration.deprecation: session.id is deprecated. Instead, use
dfs.metrics.session-id
22/06/27 15:15:26 INFO jvm.JvmMetrics: Initializing JVM Metrics with
processName=JobTracker,sessionId=
22/06/27 15:15:26 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker,sessionId= - already initialized
22/06/27 15:15:26 INFO mapred.FileInputFormat: Total input paths to process : 1
```

```
22/06/27 15:15:26 INFO mapred.FileInputFormat: Total input paths to process :
122/06/27 15:15:26 INFO mapreduce.JobSubmitter: number of splits:2
22/06/27 15:15:26 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local1698947086_000122/06/27 15:15:26 INFO mapreduce.Job: The url to track the job: http://
localhost:8080/
22/06/27 15:15:26 INFO mapred.LocalJobRunner: OutputCommitter set in config
null22/06/27 15:15:26 INFO mapreduce.Job: Running job: job_local1698947086_0001
22/06/27 15:15:26 INFO mapred.LocalJobRunner: OutputCommitter is
org.apache.hadoop.mapred.FileOutputCommitter
22/06/27 15:15:26 INFO mapred.LocalJobRunner: Waiting for map tasks
22/06/27 15:15:26 INFO mapred.LocalJobRunner: Starting task:
attempt_local1698947086_0001_m_000000_022/06/27 15:15:26 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
22/06/27 15:15:26 INFO mapred.MapTask: Processing
split:hdfs://localhost:54310/kusum_join/
DeptName.txt:0+59 22/06/27 15:15:26 INFO mapred.MapTask:
numReduceTasks: 1
22/06/27 15:15:26 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
22/06/27 15:15:26 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
10022/06/27 15:15:26 INFO mapred.MapTask: soft limit at 83886080
22/06/27 15:15:26 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
22/06/27 15:15:26 INFO mapred.MapTask: kvstart = 26214396; length =
655360022/06/27 15:15:26 INFO mapred.MapTask: Map output collector class =
org.apache.hadoop.mapred.MapTask$MapOutputBuffer
22/06/27 15:15:26 INFO mapred.LocalJobRunner:
22/06/27 15:15:26 INFO mapred.MapTask: Starting flush of map
output22/06/27 15:15:26 INFO mapred.MapTask: Spilling map output
22/06/27 15:15:26 INFO mapred.MapTask: bufstart = 0; bufend = 63; bufvoid = 104857600
22/06/27 15:15:26 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214384(104857536);
length = 13/6553600
22/06/27 15:15:26 INFO mapred.MapTask: Finished spill 0
22/06/27 15:15:26 INFO mapred.Task: Task:attempt_local1698947086_0001_m_000000_0 is done. And is
inthe process of committing
22/06/27 15:15:26 INFO mapred.LocalJobRunner: hdfs://localhost:54310/kusum_join/DeptName.txt:0+59
22/06/27 15:15:26 INFO mapred.Task: Task 'attempt_local1698947086_0001_m_000000_0' done.
22/06/27 15:15:26 INFO mapred.LocalJobRunner: Finishing
task:attempt_local1698947086_0001_m_000000_0
22/06/27 15:15:26 INFO mapred.LocalJobRunner: Starting task:
attempt_local1698947086_0001_m_000001_022/06/27 15:15:26 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
22/06/27 15:15:26 INFO mapred.MapTask: Processing split:
hdfs://localhost:54310/kusum_join/DeptStrength.txt:0+50
22/06/27 15:15:26 INFO mapred.MapTask: numReduceTasks: 1
22/06/27 15:15:27 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
22/06/27 15:15:27 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
10022/06/27 15:15:27 INFO mapred.MapTask: soft limit at 83886080
22/06/27 15:15:27 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
22/06/27 15:15:27 INFO mapred.MapTask: kvstart = 26214396; length =
655360022/06/27 15:15:27 INFO mapred.MapTask: Map output collector class =
org.apache.hadoop.mapred.MapTask$MapOutputBuffer
22/06/27 15:15:27 INFO mapred.LocalJobRunner:
22/06/27 15:15:27 INFO mapred.MapTask: Starting flush of map
output22/06/27 15:15:27 INFO mapred.MapTask: Spilling map output
22/06/27 15:15:27 INFO mapred.MapTask: bufstart = 0; bufend = 54; bufvoid = 104857600
22/06/27 15:15:27 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214384(104857536);
length = 13/6553600
22/06/27 15:15:27 INFO mapred.MapTask: Finished spill 0
22/06/27 15:15:27 INFO mapred.Task: Task:attempt_local1698947086_0001_m_000001_0 is done. And is
inthe process of committing
22/06/27 15:15:27 INFO mapred.LocalJobRunner: hdfs://
localhost:54310/kusum_join/DeptStrength.txt:0+50
22/06/27 15:15:27 INFO mapred.Task: Task 'attempt_local1698947086_0001_m_000001_0' done.
```

```
22/06/27 15:15:27 INFO mapred.LocalJobRunner: Finishing
task:attempt_local1698947086_0001_m_000001_0
22/06/27 15:15:27 INFO mapred.LocalJobRunner: map task executor
complete.22/06/27 15:15:27 INFO mapred.LocalJobRunner: Waiting for reduce
tasks
22/06/27 15:15:27 INFO mapred.LocalJobRunner: Starting task:
attempt_local1698947086_0001_r_000000_022/06/27 15:15:27 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
22/06/27 15:15:27 INFO mapred.ReduceTask: Using
ShuffleConsumerPlugin:
org.apache.hadoop.mapreduce.task.reduce.Shuffle@477ec0d7
22/06/27 15:15:27 INFO reduce.MergeManagerImpl: MergerManager:
memoryLimit=334338464,maxSingleShuffleLimit=83584616, mergeThreshold=220663392,
ioSortFactor=10, memToMemMergeOutputsThreshold=10
22/06/27 15:15:27 INFO reduce.EventFetcher: attempt_local1698947086_0001_r_000000_0 Thread
started:EventFetcher for fetching Map Completion Events
22/06/27 15:15:27 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of
mapattempt_local1698947086_0001_m_000000_0 decomp: 73 len: 77 to MEMORY
22/06/27 15:15:27 INFO reduce.InMemoryMapOutput: Read 73 bytes from map-output
forattempt_local1698947086_0001_m_000000_0
22/06/27 15:15:27 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size:
73,inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->73
22/06/27 15:15:27 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of
mapattempt_local1698947086_0001_m_000001_0 decomp: 64 len: 68 to MEMORY
22/06/27 15:15:27 INFO reduce.InMemoryMapOutput: Read 64 bytes from map-output
forattempt_local1698947086_0001_m_000001_0
22/06/27 15:15:27 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size:
64,inMemoryMapOutputs.size() -> 2, commitMemory -> 73, usedMemory ->137
22/06/27 15:15:27 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning22/06/27 15:15:27 INFO mapred.LocalJobRunner: 2 / 2 copied.
22/06/27 15:15:27 INFO reduce.MergeManagerImpl: finalMerge called with 2 in-memory map-outputs and
0on-disk map-outputs
22/06/27 15:15:27 INFO mapred.Merger: Merging 2 sorted segments
22/06/27 15:15:27 INFO mapred.Merger: Down to the last merge-pass, with 2 segments left of
totalsize: 121 bytes
22/06/27 15:15:27 INFO reduce.MergeManagerImpl: Merged 2 segments, 137 bytes to disk to
satisfyreduce memory limit
22/06/27 15:15:27 INFO reduce.MergeManagerImpl: Merging 1 files, 139 bytes from disk
22/06/27 15:15:27 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into
reduce22/06/27 15:15:27 INFO mapred.Merger: Merging 1 sorted segments
22/06/27 15:15:27 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
totalsize: 127 bytes
22/06/27 15:15:27 INFO mapred.LocalJobRunner: 2 / 2 copied.
22/06/27 15:15:27 INFO mapred.Task: Task:attempt_local1698947086_0001_r_000000_0 is done. And is
inthe process of committing
22/06/27 15:15:27 INFO mapred.LocalJobRunner: 2 / 2 copied.
22/06/27 15:15:27 INFO mapred.Task: Task attempt_local1698947086_0001_r_000000_0 is allowed
tocommit now
22/06/27 15:15:27 INFO output.FileOutputCommitter: Saved output of task
'attempt_local1698947086_0001_r_000000_0' to hdfs://localhost:54310/kusum_join/output2/
_temporary/0/task_local1698947086_0001_r_00000022/06/27 15:15:27 INFO mapred.LocalJobRunner:
reduce > reduce
22/06/27 15:15:27 INFO mapred.Task: Task 'attempt_local1698947086_0001_r_000000_0'
done.22/06/27 15:15:27 INFO mapred.LocalJobRunner: Finishing task:
attempt_local1698947086_0001_r_000000_0
22/06/27 15:15:27 INFO mapred.LocalJobRunner: reduce task executor complete.
22/06/27 15:15:27 INFO mapreduce.Job: Job job_local1698947086_0001 running in uber mode :
false22/06/27 15:15:27 INFO mapreduce.Job:  map 100% reduce 100%
22/06/27 15:15:27 INFO mapreduce.Job: Job job_local1698947086_0001 completed
successfully22/06/27 15:15:27 INFO mapreduce.Job: Counters: 38
 File System Counters
```

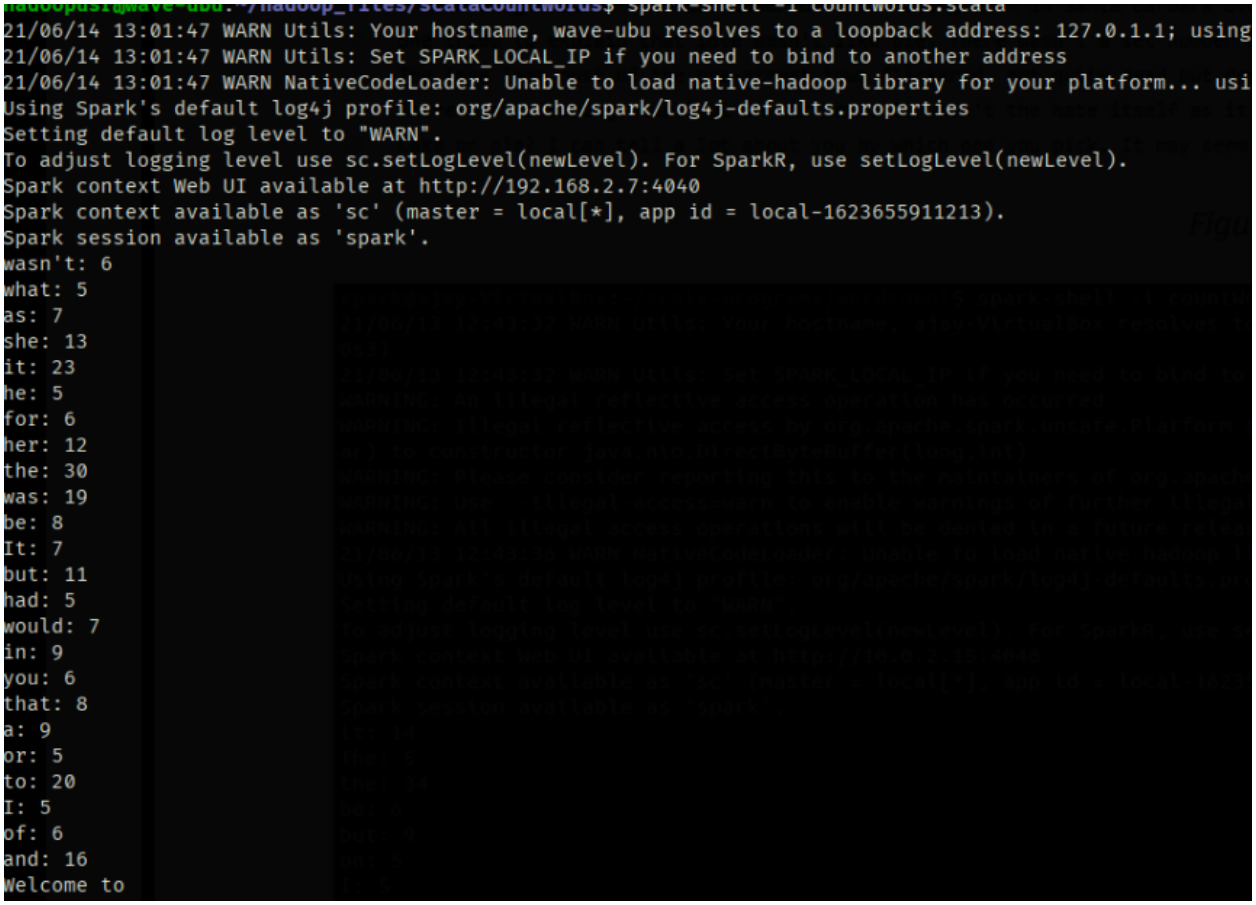
```
        FILE: Number of bytes read=26370
        FILE: Number of bytes
        written=782871FILE: Number of read
        operations=0
        FILE: Number of large read
        operations=0FILE: Number of write
        operations=0 HDFS: Number of bytes
        read=277
        HDFS: Number of bytes written=85
        HDFS: Number of read operations=28
        HDFS: Number of large read
        operations=0HDFS: Number of write
        operations=5
        Map-Reduce
        FrameworkMap input
        records=8 Map output
        records=8Map output
        bytes=117
        Map output materialized
        bytes=145Input split bytes=443
        Combine input records=0
        Combine output
        records=0Reduce input
        groups=4 Reduce shuffle
        bytes=145Reduce input
        records=8 Reduce output
        records=4 Spilled
        Records=16 Shuffled Maps
        =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed
        (ms)=2CPU time spent
        (ms)=0
        Physical memory (bytes)
        snapshot=0Virtual memory (bytes)
        snapshot=0
        Total committed heap usage
        (bytes)=913833984Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=
        0WRONG_MAP=0
        WRONG_REDUCE=0
        File Input Format
        CountersBytes Read=0
        File Output Format
        CountersBytes Written=85
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$ hdfs dfs -cat /kusum_join/output2/part-
00000
A11     50              Finance
B12     100             HR
C13     250             Manufacturing
Dept_ID Total_Employee          Dept_Name
hduser@bmsce-Precision-T1700:~/kusum/join/MapReduceJoin$
```

BDA LAB 9

Program to print word count on scala shell and print “Hello world” on scala IDE

```
scala> println("Hello World!");
Hello World!
```

```
val data=sc.textFile("sparkdata.txt")
data.collect;
val splitdata = data.flatMap(line => line.split(" "));
splitdata.collect;
val mapdata = splitdata.map(word => (word,1));
mapdata.collect;
val reducedata = mapdata.reduceByKey(_+_);
reducedata.collect;
```



BDA LAB 10

Using RDD and Flat Map count how many times each word appears in a file and write out a list of words whose count is strictly greater than n using Spark

```
scala> val textfile = sc.textFile("/home/sam/Desktop/abc.txt")
textfile: org.apache.spark.rdd.RDD[String] = /home/sam/Desktop/abc.txt MapPartitionsRDD[8] at textFile at <conso
le>:25

scala> val counts = textfile.flatMap(line => line.split(" ")).map(word => (word,1)).reduceByKey(_+_)
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[11] at reduceByKey at <console>:26

scala> import scala.collection.immutable.ListMap
import scala.collection.immutable.ListMap

scala> val sorted = ListMap(counts.collect.sortWith(_._2>_._2):_*)
sorted: scala.collection.immutable.ListMap[String,Int] = ListMap(hello -> 3, apple -> 2, unicorn -> 1, world ->
1)

scala> println(sorted)
ListMap(hello -> 3, apple -> 2, unicorn -> 1, world -> 1)
```