

# VISVESVARAYA TECHNOLOGICAL UNIVERSITY

“JnanaSangama”, Belgaum -590014, Karnataka.



## LAB REPORT

on

## BIG DATA ANALYTICS

*Submitted by*

**VARUN URS M S (1BM20CS182)**

*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**

**April-2023 to July-2023**

**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 “LAB COURSE **BIG DATA ANALYTICS**” carried out by **VARUN URS M S (1BM20CS182)**, who is bonafide student of **B. M. S. College of Engineering**. It is in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visvesvaraya Technological University, Belgaum during the year 2023. 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. Shyamala G

Assistant Professor  
Department of CSE  
BMSCE, Bengaluru

.

**Dr. Jyothi S Nayak**

Professor and Head  
Department of CSE  
BMSCE, Bengaluru

## Index Sheet

<b>Sl. No.</b>	<b>Experiment Title</b>	<b>Page No.</b>
1	Cassandra Lab Program 1:- Employee Database	
2	Cassandra Lab Program 2:- Library Database	
3	MongoDB- CRUD Demonstration	
4	Hadoop installation	
5	Hadoop Commands	
6	Hadoop Program: Average Temperature	
7	Hadoop Program: Word Count,TopN	
8	Hadoop program: Join operation	
9	Scala Program	
10	Scala Program: Word Count	

## **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

## LAB 1 Perform the following DB operations using Cassandra.

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

```

Activities Terminal Jun 18 21:26 • ise@ise-VirtualBox: ~/Desktop
(S rows)
cqish:employee> update Employee_info set project = project+'AI','Data
... warehouse') where Emp_id = 111 and salary = 75000;
InvalidRequest: Error from server: code=200 [Invalid query] message="Undefined column name project in table employee.employee_info"
cqish:employee> alter employee_info add project set+text;
cqish:employee> update Employee_info set project = project+'AI','Data warehouse' where Emp_id = 111 and salary = 75000;
cqish:employee> update Employee_info set project = project+'IOT','Data
... warehouse') where Emp_id = 121 and salary = 85000;
cqish:employee> select * from employee_info;
cqish:employee> select * from employee_info;

emp_id | salary | dept_name | designation | doj | emp_name | project
-----+-----+-----+-----+-----+-----+-----+
111 | 75000 | CSE | Assistant professor | 2022-05-10 10:30:00.000000+0000 | John | ('AI', 'Data warehouse')
151 | 95000 | ISE | Associate professor | 2022-05-10 10:30:00.000000+0000 | Yelena | null
123 | 85000 | ECE | Assistant professor | 2022-05-10 10:30:00.000000+0000 | Josh | ('Data\warehouse', 'IOT')
141 | 1.05e+05 | ISE | Associate professor | 2022-05-10 10:30:00.000000+0000 | Jane | null
131 | 95000 | ECE | Associate professor | 2022-05-10 10:30:00.000000+0000 | Mary | null
(5 rows)

cqish:employee> update Employee_info set project = project+'IOT','AI' where
... Emp_id = 131 and salary = 95000;
cqish:employee> update Employee_info set project = project+'IOT','machine
... learning') where Emp_id = 141 and salary = 95000;
cqish:employee> update Employee_info set project = project+'IOT','data science'
... where Emp_id = 141 and salary = 105000;
cqish:employee> select * from Employee_info;
SyntaxException: line 10 no viable alternative at input 'cqish' ([cqish]...)
cqish:employee> select * from Employee_info;

emp_id | salary | dept_name | designation | doj | emp_name | project
-----+-----+-----+-----+-----+-----+-----+
111 | 75000 | CSE | Assistant professor | 2022-05-10 10:30:00.000000+0000 | John | ('AI', 'Data warehouse')
151 | 95000 | ISE | Associate professor | 2022-05-10 10:30:00.000000+0000 | Yelena | null
123 | 85000 | ECE | Assistant professor | 2022-05-10 10:30:00.000000+0000 | Josh | ('Data\warehouse', 'IOT')
141 | null | null | null | null | null | ('IOT', 'machine\learning')
141 | 1.05e+05 | ISE | Associate professor | 2022-05-10 10:30:00.000000+0000 | Jane | ('IOT', 'data science')
131 | 95000 | ECE | Associate professor | 2022-05-10 10:30:00.000000+0000 | Mary | ('AI', 'IOT')
(6 rows)

cqish:employee> update Employee_info set project = project+'IOT','AI' where
... Emp_id = 131 and salary = 95000;
cqish:employee> update Employee_info set project = project+'IOT','machine
... learning') where Emp_id = 141 and salary = 95000;
cqish:employee> update Employee_info set project = project+'IOT','data science'
... where Emp_id = 141 and salary = 105000;
cqish:employee> select * from Employee_info;

```

```

Activities Terminal Jun 18 21:26 • ise@ise-VirtualBox: ~/Desktop
(S rows)
cqish:employee> update Employee_info set project = project+'IOT','AI' where
... Emp_id = 131 and salary = 95000;
cqish:employee> update Employee_info set project = project+'IOT','machine
... learning') where Emp_id = 141 and salary = 95000;
cqish:employee> update Employee_info set project = project+'IOT','data science'
... where Emp_id = 141 and salary = 105000;
cqish:employee> select * from Employee_info;

emp_id | salary | dept_name | designation | doj | emp_name | project
-----+-----+-----+-----+-----+-----+-----+
141 | 1.05e+05 | ISE | Associate professor | 2022-05-10 10:30:00.000000+0000 | Jane | ('IOT', 'data science')
131 | 95000 | ECE | Associate professor | 2022-05-10 10:30:00.000000+0000 | Mary | ('AI', 'IOT')
(6 rows)

cqish:employee> update Employee_info set project = project+'IOT','AI' where
... Emp_id = 131 and salary = 95000;
cqish:employee> update Employee_info set project = project+'IOT','machine
... learning') where Emp_id = 141 and salary = 95000;
cqish:employee> update Employee_info set project = project+'IOT','data science'
... where Emp_id = 141 and salary = 105000;
cqish:employee> select * from Employee_info;
cqish:employee> select * from Employee_info;

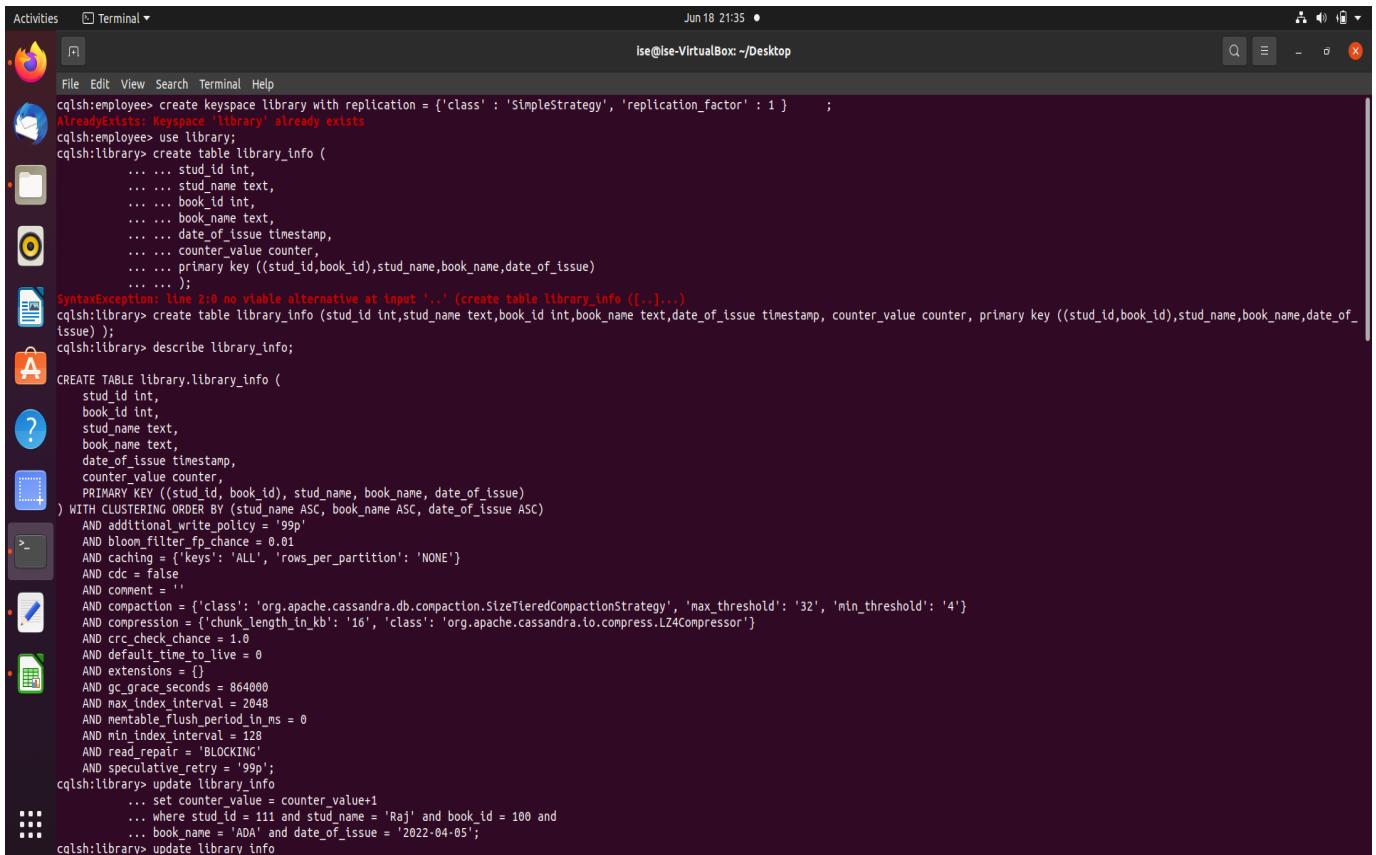
emp_id | salary | dept_name | designation | doj | emp_name | project
-----+-----+-----+-----+-----+-----+-----+
111 | 75000 | CSE | Assistant professor | 2022-05-10 10:30:00.000000+0000 | John | ('AI', 'Data warehouse')
151 | 95000 | ISE | Associate professor | 2022-05-10 10:30:00.000000+0000 | Yelena | null
123 | 85000 | ECE | Assistant professor | 2022-05-10 10:30:00.000000+0000 | Josh | ('Data\warehouse', 'IOT')
141 | 95000 | null | null | null | null | ('IOT', 'machine\learning')
141 | 1.05e+05 | ISE | Associate professor | 2022-05-10 10:30:00.000000+0000 | Jane | ('IOT', 'data science')
131 | 95000 | ECE | Associate professor | 2022-05-10 10:30:00.000000+0000 | Mary | ('AI', 'IOT')
(6 rows)

cqish:employee> insert into
... Employee_info(Emp_id,Emp_name,Designation,DOJ,salary,Dept_name) values
... (161,'Ryan','Associate professor','2022-05-11',95000,'ISE') using ttl 60;
cqish:employee> select ttl(Emp_name) from Employee_info where Emp_id = 161
... and salary = 95000;
ttl(emp_name)
52
(1 rows)
cqish:employee>

```

## LAB 2 Perform the following DB operations using Cassandra.

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



The screenshot shows a terminal window titled "Terminal" running on a Linux desktop. The terminal session is as follows:

```
Activities Terminal Jun 18 21:35 ● ise@ise-VirtualBox: ~/Desktop
ise@ise-VirtualBox: ~$ cqlsh:employee> create keyspace library with replication = {'class' : 'SimpleStrategy', 'replication_factor' : 1} ;
AlreadyExists: Keyspace 'library' already exists
ise@ise-VirtualBox: ~$ cqlsh:employee> use library;
cqlsh:library> create table library_info (
    ... stud_id int,
    ... stud_name text,
    ... book_id int,
    ... book_name text,
    ... date_of_issue timestamp,
    ... counter_value counter,
    ... primary key ((stud_id,book_id),stud_name,book_name,date_of_issue)
    ... );
SyntaxException: line 2:0 no viable alternative at input '... (create table library_info ([..]...
cqlsh:library> create table library_info (stud_id int,stud_name text,book_id int,book_name text,date_of_issue timestamp, counter_value counter, primary key ((stud_id,book_id),stud_name,book_name,date_of_issue));
cqlsh:library> describe library_info;
CREATE TABLE library.library_info (
    stud_id int,
    book_id int,
    stud_name text,
    book_name text,
    date_of_issue timestamp,
    counter_value counter,
    PRIMARY KEY ((stud_id,book_id), stud_name, book_name, date_of_issue)
) WITH CLUSTERING ORDER BY (stud_name ASC, book_name ASC, date_of_issue ASC)
AND additional_write_policy = '99p'
AND bloom_filter_fp_chance = 0.01
AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
AND cdc = false
AND comment = ''
AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactionStrategy', 'max_threshold': '32', 'min_threshold': '4'}
AND compression = {'chunk_length_in_kb': '16', 'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}
AND crc_check_chance = 1.0
AND default_time_to_live = 0
AND extensions = {}
AND gc_grace_seconds = 864000
AND max_index_interval = 2048
AND memtable_flush_period_in_ms = 0
AND min_index_interval = 128
AND read_repair = 'BLOCKING'
AND speculative_retry = '99p';
cqlsh:library> update library_info
    ... set counter_value = counter_value+1
    ... where stud_id = 111 and stud_name = 'Raj' and book_id = 100 and
    ... book_name = 'ADA' and date_of_issue = '2022-04-05';
cqlsh:library> update library info
```

```

Activities Terminal Jun 18 21:35 • ise@ise-VirtualBox: ~/Desktop
File Edit View Search Terminal Help
AND read_repair = 'BLOCKING'
AND speculative_retry = '99';
cqqlsh:library> update library.info
... set counter_value = counter_value+1
... where stud_id = 111 and stud_name = 'Raj' and book_id = 100 and
... book_name = 'ADA' and date_of_issue = '2022-04-05';
cqqlsh:library> update library.info
... set counter_value = counter_value+1
... where stud_id = 112 and stud_name = 'Ram' and book_id = 200 and
... book_name = 'DSA' and date_of_issue = '2022-04-06';
cqqlsh:library> update library.info
... set counter_value = counter_value+1
... where stud_id = 113 and stud_name = 'sohan' and book_id = 300 and
... book_name = 'JAVA' and date_of_issue = '2022-04-07';
cqqlsh:library> update library.info
... set counter_value = counter_value+1
... where stud_id = 114 and stud_name = 'rohan' and book_id = 400 and
... book_name = 'UNIX' and date_of_issue = '2022-04-07';
cqqlsh:library> select * from library.info;
stud_id | book_id | stud_name | book_name | date_of_issue | counter_value
-----+-----+-----+-----+-----+-----
114 | 400 | rohan | UNIX | 2022-04-06 18:30:00.000000+0000 | 1
111 | 100 | Raj | ADA | 2022-04-04 18:30:00.000000+0000 | 1
112 | 200 | Ram | DSA | 2022-04-05 18:30:00.000000+0000 | 1
113 | 300 | sohan | JAVA | 2022-04-06 18:30:00.000000+0000 | 1
(4 rows)
cqqlsh:library> update library.info
... set counter_value = counter_value+1
... where stud_id = 114 and stud_name = 'rohan' and book_id = 400 and
... book_name = 'UNIX' and date_of_issue = '2022-04-07';
cqqlsh:library> select * from library.info;
stud_id | book_id | stud_name | book_name | date_of_issue | counter_value
-----+-----+-----+-----+-----+-----
114 | 400 | rohan | UNIX | 2022-04-06 18:30:00.000000+0000 | 2
111 | 100 | Raj | ADA | 2022-04-04 18:30:00.000000+0000 | 1
112 | 200 | Ram | DSA | 2022-04-05 18:30:00.000000+0000 | 1
113 | 300 | sohan | JAVA | 2022-04-06 18:30:00.000000+0000 | 1
(4 rows)
cqqlsh:library> select stud_id from library.info where book_name = 'UNIX' and
... counter_value = 2 allow filtering;
stud_id
-----
114
(1 rows)

```

```

Activities Terminal Jun 18 21:35 • ise@ise-VirtualBox: ~/Desktop
File Edit View Search Terminal Help
stud_id | book_id | stud_name | book_name | date_of_issue | counter_value
-----+-----+-----+-----+-----+-----
114 | 400 | rohan | UNIX | 2022-04-06 18:30:00.000000+0000 | 2
111 | 100 | Raj | ADA | 2022-04-04 18:30:00.000000+0000 | 1
112 | 200 | Ram | DSA | 2022-04-05 18:30:00.000000+0000 | 1
113 | 300 | sohan | JAVA | 2022-04-06 18:30:00.000000+0000 | 1
(4 rows)
cqqlsh:library> select stud_id from library.info where book_name = 'UNIX' and
... counter_value = 2 allow filtering;
stud_id
-----
114
(1 rows)
cqqlsh:library> copy
... library.info(stud_id,stud_name,book_id,book_name,date_of_issue,counter_value
... ) to '/home/ise/Desktop/library_info.csv';
Using 1 child processes
Starting copy of library.library.info with columns [stud_id, stud_name, book_id, book_name, date_of_issue, counter_value].
Processed: 4 rows; Rate: 31 rows/s; Avg. rate: 12 rows/s
4 rows exported to 1 files in 0.334 seconds.
cqqlsh:library> truncate library.info;
cqqlsh:library> select * from library.info;
stud_id | book_id | stud_name | book_name | date_of_issue | counter_value
-----+-----+-----+-----+-----+-----
(0 rows)
cqqlsh:library> copy library.info(stud_id,book_id,stud_name,book_name,date_of_issue,counter_value
... ) from 'd:\library_info.csv' with header = true;
Using 1 child processes
Starting copy of library.library.info with columns [stud_id, book_id, stud_name, book_name, date_of_issue, counter_value].
Failed to import 0 rows: OSError - Can't open 'd:\library_info.csv' for reading: no matching file found, given up after 1 attempts
Processed: 0 rows; Rate: 0 rows/s; Avg. rate: 0 rows/s
0 rows imported from 0 files in 0.354 seconds (0 skipped).
cqqlsh:library> copy library.info(stud_id,book_id,stud_name,book_name,date_of_issue,counter_value ) from 'd:\library_info.csv' with header = true;
Using 1 child processes
Starting copy of library.library.info with columns [stud_id, book_id, stud_name, book_name, date_of_issue, counter_value].
Failed to import 0 rows: OSError - Can't open 'd:\library_info.csv' for reading: no matching file found, given up after 1 attempts
Processed: 0 rows; Rate: 0 rows/s; Avg. rate: 0 rows/s
0 rows imported from 0 files in 0.505 seconds (0 skipped).
cqqlsh:library> copy library.info(stud_id,book_id,stud_name,book_name,date_of_issue,counter.value ) from '/home/ise/Desktop/library_info.csv' with header = true;

```

## LAB 3 MongoDB- CRUD Demonstration

```
mongosh mongoDB://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
mscsece@mscsece-HP-EliteTower-600-C9:Desktop-PC: $ mongosh
Current Mongosh Log ID: 6427a8a279fc2f07a1c5a5d1
Connecting to:          mongosh://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000&appName=mongosh+1.8.0
Using MongoDB:          6.0.5
Using Mongosh:          1.8.0
For mongosh info see: https://docs.mongodb.com/mongodb-shell/
-----
The server generated these startup warnings when booting
2023-04-01T09:06:25.676+05:30: Using the XFS filesystem is strongly recommended with the WiredTiger storage engine. See http://dochub.mongodb.org/core/prodnotes-filesystem
2023-04-01T09:06:28.365+05:30: Access control is not enabled for the database. Read and write access to data and configuration is unrestricted
-----
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()
-----
test> show dbs
admin   40.00 KB
config  72.00 KB
local   72.00 KB
varunurldb
switched to db varunurldb
varunurldb> db.createCollection("FirstCollection")
{
  "ok": 1
}
varunurldb> db.FirstCollection.insert({_id:1,Student_Name:'Varun Urs M S',grade:'6 Sem'})
DeprecationWarning: Collection.insert() is deprecated. Use insertOne, insertMany, or bulkWrite.
{
  "acknowledged": true,
  "insertedId": "63755a2a5a5a5a5a5a5a5a5a"
}
varunurldb> show dbs
admin   40.00 KB
config  108.00 KB
local   72.00 KB
varunurldb  8.00 KB
varunurldb> show collections
FirstCollection
varunurldb> db.Students.update({_id:2,Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"Skating"}},{upsert:true})
{
  "acknowledged": false,
  "error": {
    "code": 11000,
    "errmsg": "MongoDB::Error: SyntaxError: Unexpected token (1:50)"
  }
}
varunurldb> db.Students.update({_id:2,Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"Skating"}},{upsert:true})
{
  "acknowledged": true,
  "insertedId": null,
  "matchedCount": 1,
  "modifiedCount": 1,
  "upsertedCount": 0
}
varunurldb> db.Students.update({_id:2,Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"Skating"}},{upsert:true})
DeprecationWarning: Collection.update() is deprecated. Use updateOne, updateMany, or bulkWrite.
{
  "acknowledged": true,
  "insertedId": null,
  "matchedCount": 1,
  "modifiedCount": 1,
  "upsertedCount": 0
}
varunurldb> db.Students.update({_id:2,Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"Chess"}},{upsert:true})
{
  "acknowledged": true,
  "insertedId": null,
  "matchedCount": 1,
  "modifiedCount": 1,
  "upsertedCount": 0
}
varunurldb> db.Students.update({_id:2,Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"Chess"}},{upsert:true})
{
  "acknowledged": true,
  "insertedId": null,
  "matchedCount": 1,
  "modifiedCount": 1,
  "upsertedCount": 0
}
varunurldb> db.Students.update({_id:2,Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"Chess"}},{upsert:true})
{
  "acknowledged": true,
  "insertedId": null,
  "matchedCount": 1,
  "modifiedCount": 1,
  "upsertedCount": 0
}
varunurldb> db.Students.update({_id:2,Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"Kung fu"}},{upsert:true})
{
  "acknowledged": true,
  "insertedId": null,
  "matchedCount": 1,
  "modifiedCount": 1,
  "upsertedCount": 0
}
varunurldb> db.Students.update({_id:2,Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"Kung fu"}},{upsert:true})
{
  "acknowledged": true,
  "insertedId": null,
  "matchedCount": 1,
  "modifiedCount": 1,
  "upsertedCount": 0
}
varunurldb> db.Students.update({_id:2,Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"Chess"}},{upsert:true})
{
  "acknowledged": true,
  "insertedId": null,
  "matchedCount": 1,
  "modifiedCount": 1,
  "upsertedCount": 0
}
varunurldb> db.Students.update({_id:2,Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"Chess"}},{upsert:true})
{
  "acknowledged": true,
  "insertedId": null,
  "matchedCount": 1,
  "modifiedCount": 1,
  "upsertedCount": 0
}
varunurldb> db.Students.find()
[ { "_id": 2, "Name": "Aryan David", "grade": "7 Sem", "Hobbies": "Kung fu" } ]
varunurldb> db.Students.find().pretty()
[ { "_id": 2, "Name": "Aryan David", "grade": "7 Sem", "Hobbies": "Kung fu" } ]
varunurldb> db.Students.update({Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"Chess"}},{upsert:true})
{
  "acknowledged": true,
  "insertedId": null,
  "matchedCount": 1,
  "modifiedCount": 1,
  "upsertedCount": 0
}
varunurldb> db.Students.find().pretty()
```

```

mongosh mongoDB://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
varunursdb> db.Students.find()
[ { _id: 2, Name: 'Aryan David', grade: '7 Sem', Hobbies: 'Kung fu' } ]
varunursdb> db.Students.findOne().pretty()
{ "_id": 2, "Name": "Aryan David", "grade": "7 Sem", "Hobbies": "Kung fu" }
varunursdb> db.Students.update({Name: 'Aryan David',grade:'7 Sem"},{$set:{Hobbies:"Chess"}},{upsert:true})
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
varunursdb> db.Students.find().pretty()
[ { _id: 2, Name: 'Aryan David', grade: '7 Sem', Hobbies: 'Chess' } ]
varunursdb> db.Students.update({Name: 'Ramesh',grade:'6 Sem"},{$set:{Hobbies:'Cricket'}},{upsert:true})
{
  acknowledged: true,
  insertedId: ObjectId("6427ac30aee986d1b52df926"),
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 1
}
varunursdb> db.Students.find().pretty()
[ { _id: 2, Name: 'Aryan David', grade: '7 Sem', Hobbies: 'Chess' },
  { _id: ObjectId("6427ac30aee986d1b52df926"),
    Name: 'Ramesh',
    grade: '6 Sem',
    Hobbies: 'Cricket'
  }
]
varunursdb> db.Students.find({Name:'Ramesh'}).pretty()
[ {
  _id: ObjectId("6427ac30aee986d1b52df926"),
  Name: 'Ramesh',
  grade: '6 Sem',
  Hobbies: 'Cricket'
}
]
varunursdb> db.Students.find({Name:'Aryan David'}).pretty()
[ { _id: 2, Name: 'Aryan David', grade: '7 Sem', Hobbies: 'Chess' } ]
varunursdb> db.Students.find({name:'Aryan Davl'}).pretty()
{
  acknowledged: true,
  insertedId: db.Students.find({name:'Aryan Davl'}).find(),
  [SyntaxError: db.Students.find({name:'Aryan Davl'}).find is not a function]
  matchedCount: 0
}
varunursdb> db.Students.find({name:'Aryan Davl'}).pretty()
[ { _id: 2, Name: 'Aryan David', grade: '7 Sem', Hobbies: 'Chess' } ]
varunursdb> db.Students.find({}, {Name:1,Grade:1,_id:0})
[ { Name: 'Aryan David' }, { Name: 'Ramesh' } ]
varunursdb> db.Students.find({}, {Name:2,Grade:1,_id:0})
[ { Name: 'Aryan David' }, { Name: 'Ramesh' } ]
varunursdb> db.Students.update({Name: 'Varun Urs',grade:'6 Sem"},{$set:{Hobbies:"Designing"}},{upsert:true})
[

mongosh mongoDB://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
varunursdb> db.Students.find({name:'Aryan David'}).pretty()
[ { _id: 2, Name: 'Aryan David', grade: '7 Sem', Hobbies: 'Chess' } ]
varunursdb> db.Students.find({name:'Aryan Davl'}).pretty()
{
  acknowledged: true,
  insertedId: db.Students.find({name:'Aryan Davl'}).find(),
  [TypeError: db.Students.find({name:'Aryan Davl'}).find is not a function]
  matchedCount: 0
}
varunursdb> db.Students.find({name:'Aryan Davl'}).pretty()
[ { _id: 2, Name: 'Aryan David', grade: '7 Sem', Hobbies: 'Chess' } ]
varunursdb> db.Students.find({name:'Arun'}, {Name:1,Grade:1,_id:0})
[ { Name: 'Arun' }, { Name: 'Ramesh' } ]
varunursdb> db.Students.find({name:'Arun'}, {Name:2,Grade:1,_id:0})
[ { Name: 'Arun' }, { Name: 'Ramesh' } ]
varunursdb> db.Students.update({Name: 'Varun Urs',grade:'6 Sem"},{$set:{Hobbies:"Designing"}},{upsert:true})
{
  acknowledged: true,
  insertedId: ObjectId("6427ad6aee986d1b52df971"),
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 1
}
varunursdb> db.Students.update({Name: 'Arun',grade:'7 Sem"},{$set:{Hobbies:'Anime'}},{upsert:true})
{
  acknowledged: true,
  insertedId: ObjectId("6427ad7aaee986d1b52df977"),
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 1
}
varunursdb> db.Students.update({Name: 'Chirag',grade:'7 Sem"},{$set:{Hobbies:'Cycling'}},{upsert:true})
{
  acknowledged: true,
  insertedId: ObjectId("6427ad89aee986d1b52df97d"),
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 1
}
varunursdb> db.Students.find().pretty()
[ { _id: 2, Name: 'Aryan David', grade: '7 Sem', Hobbies: 'Chess' },
  { _id: ObjectId("6427ac30aee986d1b52df926"),
    Name: 'Ramesh',
    grade: '6 Sem',
    Hobbies: 'Cricket'
  },
  {
    _id: ObjectId("6427ad6aee986d1b52df971"),
    Name: 'Varun Urs',
    grade: '6 Sem',
    Hobbies: 'Designing'
  },
  {
    _id: ObjectId("6427ad7aaee986d1b52df977"),
    Name: 'Arun',
    grade: '7 Sem',
    Hobbies: 'Anime'
  }
]

```

```

mongosh mongoDB://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
varunursdb> db.Students.find().pretty()
[
  {
    "_id": 2,
    "Name": "Aryan David",
    "grade": "7 Sem",
    "Hobbies": "Skating"
  },
  {
    "_id": ObjectId("6427ac30aee986d1b52df926"),
    "Name": "Ramesh",
    "grade": "6 Sem",
    "Hobbies": "Cricket"
  },
  {
    "_id": ObjectId("6427ad6aaee986d1b52df971"),
    "Name": "Varun Urs",
    "grade": "6 Sem",
    "Hobbies": "Designing"
  },
  {
    "_id": ObjectId("6427ad7aaee986d1b52df977"),
    "Name": "Arun",
    "grade": "7 Sem",
    "Hobbies": "Anime"
  },
  {
    "_id": ObjectId("6427ad89aee986d1b52df97d"),
    "Name": "Chirag",
    "grade": "7 Sem",
    "Hobbies": "Cycling"
  }
]
varunursdb> db.Students.remove({Name:"Aryan David"})
DeprecationWarning: Collection.remove() is deprecated. Use deleteOne, deleteMany, findOneAndDelete, or bulkWrite.
{
  acknowledged: true,
  deletedCount: 0
}
varunursdb> db.Students.deleteOne({Name:"Aryan David"})
{
  acknowledged: true,
  deletedCount: 0
}
varunursdb> db.Students.find().pretty()
[
  {
    "_id": 2,
    "Name": "Aryan David",
    "grade": "7 Sem",
    "Hobbies": "Skating"
  },
  {
    "_id": ObjectId("6427ac30aee986d1b52df926"),
    "Name": "Ramesh",
    "grade": "6 Sem",
    "Hobbies": "Cricket"
  },
  {
    "_id": ObjectId("6427ad6aaee986d1b52df971"),
    "Name": "Varun Urs",
    "grade": "6 Sem",
    "Hobbies": "Designing"
  },
  {
    "_id": ObjectId("6427ad7aaee986d1b52df977")
  }
]
mongosh mongoDB://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
varunursdb> db.Students.update({Name:"Aryan David",grade:"7 Sem"},{$set:{Hobbies:"skating"}},{upsert:true})
{
  acknowledged: true,
  insertedId: ObjectId("6427af3cae986d1b52df9e8"),
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 1
}
varunursdb> db.Students.find().pretty()
[
  {
    "_id": ObjectId("6427ac30aee986d1b52df926"),
    "Name": "Ramesh",
    "grade": "6 Sem",
    "Hobbies": "Cricket"
  },
  {
    "_id": ObjectId("6427ad6aaee986d1b52df971"),
    "Name": "Varun Urs",
    "grade": "6 Sem",
    "Hobbies": "Designing"
  },
  {
    "_id": ObjectId("6427ad7aaee986d1b52df977"),
    "Name": "Arun",
    "grade": "7 Sem",
    "Hobbies": "Anime"
  },
  {
    "_id": ObjectId("6427ad89aee986d1b52df97d"),
    "Name": "Chirag",
    "grade": "7 Sem",
    "Hobbies": "Cycling"
  },
  {
    "_id": ObjectId("6427af3cae986d1b52df9e8"),
    "Name": "Aryan David",
    "grade": "7 Sem",
    "Hobbies": "skating"
  }
]
varunursdb> db.Students.find(),{Name:1}
[
  {
    "_id": ObjectId("6427ac30aee986d1b52df926"), Name: "Ramesh" },
  {
    "_id": ObjectId("6427ad6aaee986d1b52df971"), Name: "Varun Urs" },
  {
    "_id": ObjectId("6427ad7aaee986d1b52df977"), Name: "Arun" },
  {
    "_id": ObjectId("6427ad89aee986d1b52df97d"), Name: "Chirag" },
  {
    "_id": ObjectId("6427af3cae986d1b52df9e8"), Name: "Aryan David" }
]
varunursdb> db.Students.find(),{Name:1}, {_id:0}
Uncaught:
syntaxError: Unexpected token, expected "," (1:32)
< 1 | db.Students.find({}, {Name:1}, {_id:0})

```

```

mongosh mongoDB://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
1 | db.Students.find({}, {Name:1}, {_id:0})
| |
2 |
varunursdb> db.Students.find({}, {Name:1}, {_id:0})
[
  { Name: 'Ramesh' },
  { Name: 'Varun Urs' },
  { Name: 'Arun' },
  { Name: 'Chirag' },
  { Name: 'Aryan David' }
]
varunursdb> db.Students.find({}, {Name:2}, {_id:0})
[
  { Name: 'Ramesh' },
  { Name: 'Varun Urs' },
  { Name: 'Arun' },
  { Name: 'Chirag' },
  { Name: 'Aryan David' }
]
varunursdb> db.Students.find({}, {Grade:1}, {_id:0})
[ {}, {}, {}, {} ]
varunursdb> db.Students.find({}, {grade:1}, {_id:0})
[
  { grade: '6 Sem' },
  { grade: '6 Sem' },
  { grade: '7 Sem' },
  { grade: '7 Sem' },
  { grade: '7 Sem' }
]
varunursdb> db.Students.find({grade:1}, {_id:0})
varunursdb> db.Students.find({}, {grade:1}, {_id:0})
[
  { grade: '6 Sem' },
  { grade: '6 Sem' },
  { grade: '7 Sem' },
  { grade: '7 Sem' },
  { grade: '7 Sem' }
]
varunursdb> db.Students.find({grade:1}, {_id:0})
[
  { _id: ObjectId("6427ac30acee986d1b52df926"), grade: '6 Sem' },
  { _id: ObjectId("6427ad6aaee986d1b52df971"), grade: '6 Sem' },
  { _id: ObjectId("6427ad7aaee986d1b52df977"), grade: '7 Sem' },
  { _id: ObjectId("6427ad89aee986d1b52df97d"), grade: '7 Sem' },
  { _id: ObjectId("6427af3cae986d1b52df9e8"), grade: '7 Sem' }
]
varunursdb> db.Students.find({grade:1}, {_id:1})
varunursdb> db.Students.find({grade:1}, {_id:0})
[
  { grade: '6 Sem' },
  { grade: '6 Sem' },
  { grade: '7 Sem' }
]
varunursdb> db.Students.find({grade:1}, {_id:1})
{
  "_id": ObjectId("6427af3cae986d1b52df9e8"),
  "Name": "Aryan David",
  "grade": "7 Sem",
  "Hobbies": "Skating"
}
varunursdb> db.Students.find({grade:{Seq:"6 Sem"}})
[
  {
    "_id": ObjectId("6427ac30acee986d1b52df926"),
    "Name": "Ramesh",
    "grade": '6 Sem',
    "Hobbies": 'Cricket'
  },
  {
    "_id": ObjectId("6427ad6aaee986d1b52df971"),
    "Name": "Varun Urs",
    "grade": '6 Sem',
    "Hobbies": 'Designing'
  }
]
varunursdb> db.Students.find({grade:{Seq:"7 Sem"}})
[
  {
    "_id": ObjectId("6427ad7aaee986d1b52df977"),
    "Name": 'Arun',
    "grade": '7 Sem',
    "Hobbies": 'Antne'
  },
  {
    "_id": ObjectId("6427ad89aee986d1b52df97d"),
    "Name": 'Chirag',
    "grade": '7 Sem',
    "Hobbies": 'Cycling'
  },
  {
    "_id": ObjectId("6427af3cae986d1b52df9e8"),
    "Name": 'Aryan David',
    "grade": '7 Sem',
    "Hobbies": 'Skating'
  }
]
varunursdb> db.Students.find({hobbies:{$in:[ "Chess", "Cricket"]}})
[
  {
    "_id": ObjectId("6427ac30acee986d1b52df926"),
    "Name": "Ramesh",
    "grade": '6 Sem',
    "Hobbies": 'Cricket'
  }
]
varunursdb> db.Students.find({hobbies:{$in:[ "Cycling", "Cricket"]}})
[
  {

```

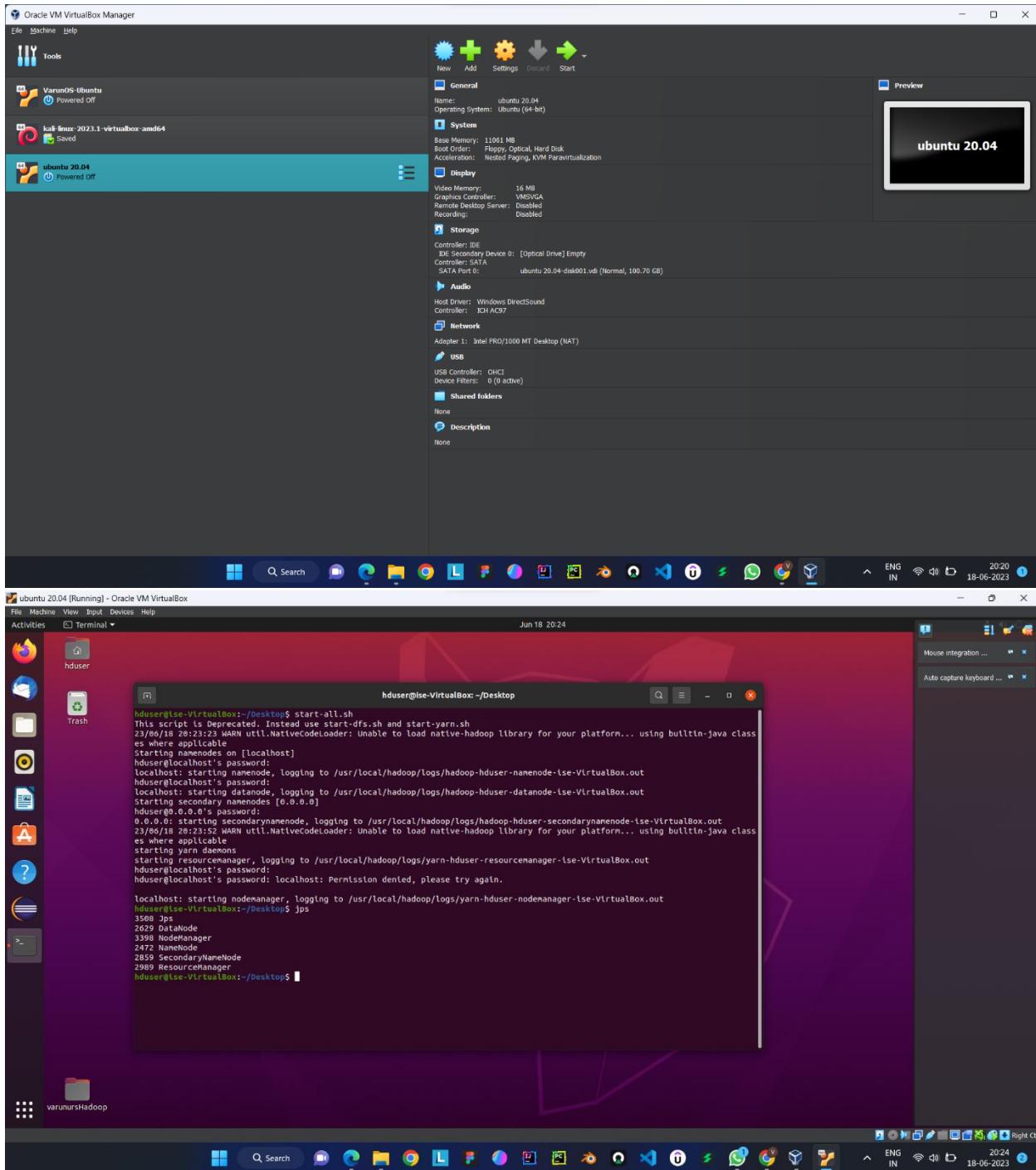
```

mongosh mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
varunursdb> db.Students.countDocuments()
varunursdb> db.Students.find().sort([{"name": -1})
[
  {
    "_id": ObjectId("6427ad6aaee986d1b52df971"),
    "Name": "Varun Urs",
    "grade": "6 Sem",
    "Hobbies": "Designing"
  },
  {
    "_id": ObjectId("6427ac30aee986d1b52df926"),
    "Name": "Ramesh",
    "grade": "6 Sem",
    "Hobbies": "Cricket"
  },
  {
    "_id": ObjectId("6427ad89aee986d1b52df97d"),
    "Name": "Chirag",
    "grade": "7 Sem",
    "Hobbies": "Cycling"
  },
  {
    "_id": ObjectId("6427af3caee986d1b52df9e8"),
    "Name": "Aryan David",
    "grade": "7 Sem",
    "Hobbies": "Skating"
  },
  {
    "_id": ObjectId("6427ad7aaee986d1b52df977"),
    "Name": "Arun",
    "grade": "7 Sem",
    "Hobbies": "Anime"
  }
]
varunursdb> db.Students.find()
[
  {
    "_id": ObjectId("6427ac30aee986d1b52df926"),
    "Name": "Ramesh",
    "grade": "6 Sem",
    "Hobbies": "Cricket"
  },
  {
    "_id": ObjectId("6427ad6aaee986d1b52df971"),
    "Name": "Varun Urs",
    "grade": "6 Sem",
    "Hobbies": "Designing"
  },
  {
    "_id": ObjectId("6427ad7aaee986d1b52df977"),
    "Name": "Arun",
    "grade": "7 Sem",
    "Hobbies": "Anime"
  }
]

mongosh mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
varunursdb> db.Students.find({name:/nS/})
[
  {
    "_id": ObjectId("6427ad7aaee986d1b52df977"),
    "Name": "Arun",
    "grade": "7 Sem",
    "Hobbies": "Anime"
  }
]
varunursdb> mongoimport --db Student --collection airlines --type csv --headerline --file /home/hduser/Desktop/airline.csv
Uncaught:
SyntaxError: Missing semicolon. (1:14)
> 1 | mongoimport -db Student --collection airlines --type csv --headerline --file /home/hduser/Desktop/airline.csv
| |
2 |
varunursdb> mongoimport --db Students --collection airlines --type txt --headerline --file /home/hduser/Desktop/file.txt
Uncaught:
SyntaxError: Missing semicolon. (1:14)
> 1 | mongoimport -db Students --collection airlines --type txt --headerline --file /home/hduser/Desktop/file.txt
| |
2 |
varunursdb> db.Students.save({name: "Tushar"})
SyntaxError: db.Students.save is not a function
varunursdb> db.Students.save({_id: ObjectId('6427ad7aaee986d1b52df978'), Name: "Tushar"})
SyntaxError: db.Students.save is not a function
varunursdb> help

```

## LAB 4. Screenshot of Hadoop installed



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

```
Activities Terminal Jun 18 20:56
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -mkdir /newdir
23/06/18 20:49:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
mkdir: '/newdir': File exists
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -mkdir /abc
23/06/18 20:50:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
mkdir: '/abc': File exists
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /
23/06/18 20:50:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 8 items
drwxr-xr-x - hduser supergroup 0 2023-06-18 20:50 /abc
drwxr-xr-x - hduser supergroup 0 2023-06-16 23:37 /final
drwxr-xr-x - hduser supergroup 0 2022-05-13 20:51 /home
drwxr-xr-x - hduser supergroup 0 2023-06-16 23:23 /newdir
drwxr-xr-x - hduser supergroup 0 2022-05-13 11:04 /tmp
drwxr-xr-x - hduser supergroup 0 2023-06-16 20:19 /user
drwxr-xr-x - hduser supergroup 0 2023-06-18 20:46 /varunurs
drwxr-xr-x - hduser supergroup 0 2023-06-16 21:59 /varunurs18
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -put /home/hduser/Desktop/varunursHadoop/sample.txt /new
23/06/18 20:51:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
copyFromLocal: Parent path is not a directory: /new
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -copyFromLocal /home/hduser/Desktop/varunursHadoop/sample.txt /new/test.txt
23/06/18 20:51:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
copyToLocal: /home/hduser/Desktop/varunursHadoop/sample.txt does not exist
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -get /new/sample.txt /home/hduser/Desktop
23/06/18 20:52:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
get: '/new/sample.txt': No such file or directory
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -get /abc/sample.txt /home/hduser/Desktop
23/06/18 20:52:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
get: '/abc/sample.txt': No such file or directory
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /abc/sample.txt
:ld: Unknown command
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /abc/sample.txt
23/06/18 20:53:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ls: '/abc/sample.txt': No such file or directory
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /abc
23/06/18 20:53:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /new
23/06/18 20:53:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
-rw-r--r-- 1 hduser supergroup 105 2023-06-18 20:51 /new
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -copyFromLocal /home/hduser/Desktop/varunursHadoop/sample.txt /abc
23/06/18 20:53:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /abc
23/06/18 20:53:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
-rw-r--r-- 1 hduser supergroup 105 2023-06-18 20:53 /abc/sample.txt
Found 1 items
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -copyToLocal /abc/sample.txt /home/hduser/Desktop
23/06/18 20:54:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Activities Terminal Jun 18 20:56
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /abc/sample.txt
23/06/18 20:53:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ls: '/abc/sample.txt': No such file or directory
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /abc
23/06/18 20:53:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /new
23/06/18 20:53:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
-rw-r--r-- 1 hduser supergroup 105 2023-06-18 20:51 /new
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -copyFromLocal /home/hduser/Desktop/varunursHadoop/sample.txt /abc
23/06/18 20:53:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /abc
23/06/18 20:53:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 hduser supergroup 105 2023-06-18 20:53 /abc/sample.txt
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -copyToLocal /abc/sample.txt /home/hduser/Desktop
23/06/18 20:53:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -cat /abc/sample.txt
23/06/18 20:54:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hit how are you
how is your job
how is your family
how is your brother
how is your sister
save your file
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -mv /abc/sample.txt /abc
23/06/18 20:55:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
/mv: '/abc/sample.txt': File exists
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /
23/06/18 20:55:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 9 items
drwxr-xr-x - hduser supergroup 0 2023-06-18 20:53 /abc
drwxr-xr-x - hduser supergroup 0 2023-06-16 23:37 /final
drwxr-xr-x - hduser supergroup 0 2022-05-13 20:51 /home
drwxr-xr-x - hduser supergroup 105 2023-06-18 20:51 /new
drwxr-xr-x - hduser supergroup 0 2023-06-16 23:23 /newdir
drwxrwxr-x - hduser supergroup 0 2022-05-13 11:04 /tmp
drwxr-xr-x - hduser supergroup 0 2023-06-16 20:19 /user
drwxr-xr-x - hduser supergroup 0 2023-06-18 20:46 /varunurs
drwxr-xr-x - hduser supergroup 0 2023-06-16 21:59 /varunurs18
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -cp /abc/sample.txt /new
23/06/18 20:55:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
cp: '/new': File exists
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /new
23/06/18 20:55:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
-rw-r--r-- 1 hduser supergroup 105 2023-06-18 20:51 /new
```

**LAB 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 NCDC data set.**
- b) **find the mean max temperature for every month**

### Average Temperature

#### AverageDriver

```
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

```
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

```
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));
}
}

```

Activities Terminal Jun 18 20:38 hduser@lse-VirtualBox: ~/Desktop

```

hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /varunurs
23/06/18 20:35:40 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 0 2023-06-18 20:33 /varunurs/output-wc
-rw-r--r-- 1 hduser supergroup 105 2023-06-18 20:32 /varunurs/sample.txt
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -copyFromLocal /home/hduser/Desktop/varunursHadoop/1901 /varunurs/
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /varunurs
23/06/18 20:36:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 item
-rw-r--r-- 1 hduser supergroup 888190 2023-06-18 20:35 /varunurs/1901
drwxr-xr-x 0 hduser supergroup 0 2023-06-18 20:33 /varunurs/output-wc
-rw-r--r-- 1 hduser supergroup 105 2023-06-18 20:32 /varunurs/sample.txt
hduser@lse-VirtualBox:~/Desktop$ hadoop jar /home/hduser/Desktop/varunurshadoop/AvgTemp.jar AverageDriver /varunurs/1901 /varunurs/output-temp
23/06/18 20:37:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/06/18 20:37:48 INFO configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
23/06/18 20:37:49 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
23/06/18 20:37:49 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
23/06/18 20:37:49 INFO mapreduce.JobSubmitter: Total number of tasks to process: 1
23/06/18 20:37:50 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1705332837_0001
23/06/18 20:37:50 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
23/06/18 20:37:50 INFO mapreduce.Job: Runner job: job_local1705332837_0001
23/06/18 20:37:50 INFO mapred.LocalJobRunner: OutputCommitter is null
23/06/18 20:37:50 INFO mapred.LocalJobRunner: JobConf艰辛的org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/06/18 20:37:51 INFO mapred.LocalJobRunner: Waiting for map tasks
23/06/18 20:37:51 INFO mapred.LocalJobRunner: Starting task: attempt_local1705332837_0001_m_000000_0
23/06/18 20:37:50 INFO mapred.Task: Using ResourceCalculatorProcessTree : []
23/06/18 20:37:50 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/varunurs/1901:0+888190
23/06/18 20:37:51 INFO mapred.MapTask: (EQUATOR) 0 kv=26253/6553600
23/06/18 20:37:51 INFO mapred.MapTask: Input split file length: 100
23/06/18 20:37:51 INFO mapred.MapTask: soft limit at 83884800
23/06/18 20:37:51 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
23/06/18 20:37:51 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
23/06/18 20:37:51 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTaskMapOutputBuffer
23/06/18 20:37:51 INFO mapreduce.Job: Job job_local1705332837_0001 running in uber mode : false
23/06/18 20:37:51 INFO mapred.LocalJobRunner: Reduce 0
23/06/18 20:37:51 INFO mapred.LocalJobRunner:
23/06/18 20:37:51 INFO mapred.MapTask: Starting flush of map output
23/06/18 20:37:51 INFO mapred.MapTask: Spilling map output
23/06/18 20:37:51 INFO mapred.MapTask: bufstart = 0; bufvoid = 59676; bufvoid = 104857600
23/06/18 20:37:51 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188144(104752576); length = 26253/6553600
23/06/18 20:37:51 INFO mapred.Task: Task attempt_local1705332837_0001_m_000000_t is done. And is in the process of committing
23/06/18 20:37:51 INFO mapred.Task: map
23/06/18 20:37:51 INFO mapred.Task: Task 'attempt_local1705332837_0001_m_000000_0' done.

```

Activities Terminal Jun 18 20:38 hduser@lse-VirtualBox: ~/Desktop

```

FILE: Number of large read operations=0
FILE: Number of small read 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=0
HDFS: Number of write operations=4
Map-Reduce metrics:
  Map input records=6565
  Map output records=564
  Map output bytes=59676
  Map output materialized bytes=72210
  Input split bytes=101
  Combiner output records=0
  Combiner output bytes=0
  Reduce input groups=1
  Reduce shuffle bytes=72210
  Reduce input records=6564
  Reduce output records=1
  Shuffled records=19128
  Shuffled Maps=0
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=36
  CPU time spent (ms)=0
  Physical memory (bytes) snapshot=0
  Virtual memory (bytes) snapshot=0
  Total committed heap usage (bytes)=344981504
Shuffle Errors:
  BAD_ID=0
  CONNECTION=0
  ID=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
  File Input Format Counters
    Bytes Read=888190
  File Output Format Counters
    BytesWritten=8
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /varunurs/output-temp
23/06/18 20:38:02 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 0 2023-06-18 20:37 /varunurs/output-temp/_SUCCESS
-rw-r--r-- 1 hduser supergroup 5 2023-06-18 20:37 /varunurs/output-temp/part-r-00000
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -cat /varunurs/output-temp/part-r-00000
23/06/18 20:38:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
1901 46
hduser@lse-VirtualBox:~/Desktop$ 

```

## MeanMax

### MeanMaxDriver

```
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

```
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

```

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;
        }
        if (count != 0)
            total_temp = max_temp;
        context.write(key, new IntWritable(total_temp));
    }
}

```

```

count++;
if (count == 3) {
    total_temp += max_temp;
    max_temp = 0;
    count = 0;
    days++;
}
} context.write(key, new IntWritable(total_temp / days));
}

```

The screenshot displays two terminal windows on a Linux desktop environment.

**Top Terminal Window:**

```

Activities Terminal Jun 18 20:40
hduser@lse-VirtualBox: ~/Desktop
hduser@lse-VirtualBox:~/Desktop$ hadoop jar /home/hduser/Desktop/varunursHadoop/meanMax.jar MeanMaxDriver /varunurs/1901 /varunurs/output-meanmax
23/06/18 20:39:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/06/18 20:39:35 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
23/06/18 20:39:35 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
23/06/18 20:39:35 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
23/06/18 20:39:35 INFO mapreduce.JobSubmitter: Total number of splits: 1
23/06/18 20:39:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1216575569_0001
23/06/18 20:39:35 INFO mapreduce.Job: The url to track the job: http://localhost:8088/
23/06/18 20:39:35 INFO mapreduce.Job: Running job: job_local1216575569_0001
23/06/18 20:39:35 INFO mapred.LocalJobRunner: OutputCommitter set in config null
23/06/18 20:39:35 INFO mapred.LocalJobRunner: Waiting for map tasks
23/06/18 20:39:36 INFO mapred.Task: Using ResourceCalculatorProcessTree : []
23/06/18 20:39:36 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/varunurs/1901:0+888190
23/06/18 20:39:36 INFO mapred.MapTask: (EQUATOR) 0 kv=26214396(10485784)
23/06/18 20:39:36 INFO mapred.MapTask: soft LLInt=83886080
23/06/18 20:39:36 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
23/06/18 20:39:36 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
23/06/18 20:39:36 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
23/06/18 20:39:36 INFO mapred.MapTask: Spilling map output
23/06/18 20:39:36 INFO mapred.MapTask: Spilling map output
23/06/18 20:39:36 INFO mapred.MapTask: bufvoid = 45948; bufvoid = 104857600
23/06/18 20:39:36 INFO mapred.MapTask: kvstart = 26214396(10485784); kvend = 26188144(104752576); length = 26253/6553600
23/06/18 20:39:36 INFO mapred.MapTask: Finished spill 0
23/06/18 20:39:36 INFO mapred.Task:attempt_local1216575569_0001_m_000000_0 is done. And is in the process of committing
23/06/18 20:39:36 INFO mapred.Task: Task attempt_local1216575569_0001_m_000000_0 done.
23/06/18 20:39:36 INFO mapred.LocalJobRunner: Finishing task: attempt_local1216575569_0001_m_000000_0
23/06/18 20:39:36 INFO mapred.LocalJobRunner: map task executor complete.
23/06/18 20:39:36 INFO mapred.LocalJobRunner: Waiting for reduce tasks
23/06/18 20:39:36 INFO mapred.LocalJobRunner: Starting task: attempt_local1216575569_0001_r_000000_0
23/06/18 20:39:36 INFO mapred.ReduceTask: Using ShuffleConsumerPlanner: org.apache.hadoop.mapreduce.task.reduce.Shuffle@2b426f54
23/06/18 20:39:36 INFO reduce.MergeManagerImpl: MergerManager: memoryLimits=363285696, maxSingleShuffleLimit=90821424, mergeThreshold=239768576, ioSortFactor=10, memToMemMergeOutputsThreshold=10
23/06/18 20:39:36 INFO reduce.EventFetcher: attempt_local1216575569_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
23/06/18 20:39:36 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1216575569_0001_m_000000_0
23/06/18 20:39:36 INFO reduce.InMemoryMapOutput: read 59078 bytes from map-output for attempt_local1216575569_0001_m_000000_0
23/06/18 20:39:36 INFO reduce.InMemoryMapOutput: close map-output file --> map-output of size: 59078, InMemoryMapOutputs.size() -> 1, commitInMemory -> 0, usedMemory -> 59078
23/06/18 20:39:36 INFO reduce.EventFetcher: EventFetcher is interrupted. Returning
23/06/18 20:39:36 INFO mapred.LocalJobRunner: I / 1 copied.
23/06/18 20:39:36 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
23/06/18 20:39:36 INFO mapred.Merger: Merged 1 sorted segments
23/06/18 20:39:36 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 59073 bytes
23/06/18 20:39:36 INFO reduce.MergeManagerImpl: Merged 1 segments, 59078 bytes to disk to satisfy reduce memory limit

```

**Bottom Terminal Window:**

```

Activities Terminal Jun 18 20:40
hduser@lse-VirtualBox: ~/Desktop
Map output materialized bytes=59082
Input split bytes<1
Combine input records=0
Combine output records=0
Reduce input groups=12
Reduce output bytes=59082
Reduce input records=564
Reduce output records=12
Spilled Records=13128
Shuffled Maps =1
Failed Shuffles=0
Merged Map Outputs=1
GC time elapsed (ms)=28
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=344981504
Shuffle Errors:
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_SOURCE=0
File Input Format Counters
  Bytes Read=888190
File Output Format Counters
  Bytes Written=74
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /varunurs/output-meanmax
23/06/18 20:40:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
-rw-r--r-- 1 hduser supergroup 0 2023-06-18 20:39 /varunurs/output-meanmax/ SUCCESS
-rw-r--r-- 1 hduser supergroup 74 2023-06-18 20:39 /varunurs/output-meanmax/part-r-00000
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -cat /varunurs/output-meanmax/part-r-00000
01
02
03
04
04 44
05
05 100
06
06 168
07
07 219
08 198
09 141
10 100
11 19
12 3
hduser@lse-VirtualBox:~/Desktop$ 

```

**LAB 7. For a given Text file, Create a Map Reduce program to sort the content in an alphabetic order listing only top 10 maximum occurrences of words.**

**TopNDriver**

```
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

```

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

```
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class TopNMapper extends Mapper<Object, Text, Text, IntWritable> {
    private static final IntWritable one = new IntWritable(1);

    private Text word = new Text();

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

    public void map(Object key, Text value, Mapper<Object, Text, Text, IntWritable>.Context context) throws IOException, InterruptedException {
        String cleanLine = value.toString().toLowerCase().replaceAll(this.tokens, " ");
        StringTokenizer itr = new StringTokenizer(cleanLine);
        while (itr.hasMoreTokens()) {
            this.word.set(itr.nextToken().trim());
            context.write(this.word, one);
        }
    }
}
```

## TopNReducer

```
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));
        }
    }
}

```

## MiscUtils

```

package utils;
import java.util.*;

public class MiscUtils {
    public static <K extends Comparable, V extends Comparable> Map<K, V> sortByValues(Map<K, V> map) {
        List<Map.Entry<K, V>> entries = new LinkedList<Map.Entry<K, V>>(map.entrySet());
        Collections.sort(entries, new Comparator<Map.Entry<K, V>>() {
            @Override
            public int compare(Map.Entry<K, V> o1, Map.Entry<K, V> o2) {
                return o2.getValue().compareTo(o1.getValue());
            }
        });
    }
}

```

```

    }
};

Map<K, V> sortedMap = new LinkedHashMap<K, V>();
for (Map.Entry<K, V> entry : entries) {
sortedMap.put(entry.getKey(), entry.getValue());
}
return sortedMap;
}
}

```

```

Activities Terminal Jun 18 20:34
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -mkdir /varunurs
23/06/18 20:31:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /varunurs
23/06/18 20:31:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -copyFromLocal /home/hduser/Desktop/varunurs/Hadoop/sample.txt /varunurs/
23/06/18 20:31:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /varunurs
23/06/18 20:32:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 hduser supergroup 105 2023-06-18 20:32 /varunurs/sample.txt
hduser@lse-VirtualBox:~/Desktop$ hadoop jar /home/hduser/Desktop/varunurs/Hadoop/WordCount12.jar WordCount /varunurs/sample.txt /varunurs/output-wc
23/06/18 20:32:30 INFO util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/06/18 20:32:59 INFO Configuration: Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
23/06/18 20:32:59 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId= - already initialized
23/06/18 20:33:08 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
23/06/18 20:33:08 INFO mapred.FileInputFormat: Total input paths to process : 1
A 23/06/18 20:33:08 INFO mapreduce.Job: Submitting tokens for job: job_local1668463559_0001
23/06/18 20:33:08 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
23/06/18 20:33:08 INFO mapred.LocalJobRunner: OutputCommitter set in config null
23/06/18 20:33:08 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
23/06/18 20:33:08 INFO mapred.LocalJobRunner: InputFormat is org.apache.hadoop.mapred.TextInputFormat
23/06/18 20:33:08 INFO mapred.LocalJobRunner: Starting task: attempt_local1668463559_0001_m_000000_0
23/06/18 20:33:08 INFO mapred.MapTask: Using ResourceCalculatorProcessTree : []
23/06/18 20:33:08 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/varunurs/sample.txt:0+105
23/06/18 20:33:08 INFO mapred.MapTask: numReduceTasks: 1
23/06/18 20:33:08 INFO mapreduce.Job: Job job_local1668463559_0001 running in uber mode : false
23/06/18 20:33:08 INFO mapred.MapTask: Starting task: attempt_local1668463559_0001_m_000000_0
23/06/18 20:33:08 INFO mapred.MapTask: (EQUATOR) 0 kv 26214396(104857584)
23/06/18 20:33:08 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
23/06/18 20:33:08 INFO mapred.MapTask: soft limit at 83886080
23/06/18 20:33:08 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
23/06/18 20:33:08 INFO mapred.MapTask: kvstart = 20214396; length = 6553600
23/06/18 20:33:08 INFO mapred.MapTask: Map input collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
23/06/18 20:33:08 INFO mapred.LocalJobRunner:
23/06/18 20:33:08 INFO mapred.MapTask: Starting flush of map output
23/06/18 20:33:08 INFO mapred.MapTask: Spilling map output
23/06/18 20:33:08 INFO mapred.MapTask: bufstart = 0; bufend = 196; bufvoid = 104857600
23/06/18 20:33:08 INFO mapred.MapTask: kvstart = 20214396(104857584); kvend = 20214398(104857232); length = 89/6553600
23/06/18 20:33:08 INFO mapred.Task: Finishing task
23/06/18 20:33:08 INFO mapred.Task: Task attempt_local1668463559_0001_m_000000_0 is done. And is in the process of committing
23/06/18 20:33:08 INFO mapred.LocalJobRunner: hdfs://localhost:54310/varunurs/sample.txt:0+105
23/06/18 20:33:08 INFO mapred.Task: Task attempt_local1668463559_0001_m_000000_0 done.
23/06/18 20:33:08 INFO mapred.LocalJobRunner: Finishing task: attempt_local1668463559_0001_m_000000_0
23/06/18 20:33:08 INFO mapred.LocalJobRunner: Map task executor complete.
23/06/18 20:33:08 INFO mapred.LocalJobRunner: Waiting for reduce tasks

```

```

Activities Terminal Jun 18 20:42
hduser@lse-VirtualBox:~/Desktop$ 
HDFS: Number of write operations=4
Map-Reduce Framework
  Map input records=7
  Map output records=23
  Map output bytes=196
  Map output materialized bytes=248
  Input split bytes=107
  Combine output records=0
  Grouping output records=0
  Reduce input groups=12
  Reduce shuffle bytes=248
  Reduce input records=23
  Reduce output records=12
  Shuffled Maps=40
  Shuffled Records=40
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=0
  CPU time spent (ms)=0
  Physical Memory (bytes) snapshot=0
  Virtual memory (bytes) snapshot=0
  Total committed heap usage (bytes)=344981504
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=105
File Output Format Counters
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -cat /varunurs/output-topn/part-r-00000
23/06/18 20:42:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
your 5
how 5
ts 4
brother 1
are 1
ht 1
file 1
sister 1
Family 1
your 1
save 1
job 1
hduser@lse-VirtualBox:~/Desktop$ 

```

## LAB 8. Create a Map Reduce program to demonstrating join operation

### JoinDriver

```
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,
DeptNameMapper.class);
        MultipleInputs.addInputPath(conf, BInputPath, TextInputFormat.class,
DeptEmpStrengthMapper.class);

        FileOutputFormat.setOutputPath(conf, outputPath);

        conf.setPartitionerClass(KeyPartitioner.class);
        conf.setOutputValueGroupingComparator(TextPair.FirstComparator.class);

        conf.setMapOutputKeyClass(TextPair.class);

        conf.setReducerClass(JoinReduce.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);
    }
}

```

### JoinReduce

```

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

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

```

```
public class JoinReduce 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);
        }
    }
}
```

## TextPair

```
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 Text getFirst() {
        return first;
    }

    public Text getSecond() {
        return second;
    }

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

    public int compareTo(TextPair other) {
        if (first.equals(other.getFirst()) && second.equals(other.getSecond()))
            return 0;
        else if (first.equals(other.getFirst()))
            return 1;
        else if (second.equals(other.getSecond()))
            return -1;
        else
            return 1;
    }

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

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

    public String toString() {
        return first + "\t\t" + 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());
}
// ^^ TextPairComparator

// vv TextPairFirstComparator
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);
    }
}
// ^^ TextPairFirstComparator

// vv TextPair
}
```

## Posts

```
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]));
}
}
```

## User

```
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]));
}
}
```

### DeptEmpStrengthMapper

```
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 DeptEmpStrengthMapper extends MapReduceBase implements
Mapper<LongWritable, Text, TextPair, Text> {

    @Override
    public void map(LongWritable key, Text value, OutputCollector<TextPair, Text> output,
Reportер reporter)
        throws IOException
    {

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

```

### DeptNameMapper

```

import java.io.IOException;

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

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

    @Override
    public void map(LongWritable key, Text value, OutputCollector<TextPair, Text> output,
Reportер reporter)
        throws IOException
    {

```

```

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

```

Activities Terminal Jun 18 20:47

```

hduser@lse-VirtualBox:~/Desktop$ hadoop fs -copyFromLocal /home/hduser/Desktop/varunurshadoop/DeptName.txt /home/hduser/Desktop/varunurshadoop/DeptStrength.txt /varunurus/
23/06/18 20:44:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /varunurus/
23/06/18 20:44:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 8 items
-rw-r--r-- 1 hduser supergroup 888190 2023-06-18 20:35 /varunurus/_root0
-rw-r--r-- 1 hduser supergroup 59 2023-06-18 20:45 /varunurus/DeptName.txt
-rw-r--r-- 1 hduser supergroup 50 2023-06-18 20:45 /varunurus/DeptStrength.txt
drwxr-xr-x 1 hduser supergroup 0 2023-06-18 20:37 /varunurus/_output-mnmax
drwxr-xr-x 1 hduser supergroup 0 2023-06-18 20:37 /varunurus/_output-temp
drwxr-xr-x 1 hduser supergroup 0 2023-06-18 20:42 /varunurus/_output-topn
drwxr-xr-x 1 hduser supergroup 0 2023-06-18 20:33 /varunurus/_output-wc
-rw-r--r-- 1 hduser supergroup 105 2023-06-18 20:32 /varunurus/sample.txt

```

hduser@lse-VirtualBox:~/Desktop\$ hadoop jar /home/hduser/Desktop/varunurshadoop/Join.jar JoinDriver /varunurus/DeptName.txt /varunurus/DeptStrength.txt /varunurus/\_output-join
Usage: hadoop jar /home/hduser/Desktop/varunurshadoop/Join.jar /varunurus/DeptName.txt /varunurus/DeptStrength.txt /varunurus/\_output-join
23/06/18 20:46:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/06/18 20:46:04 INFO configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
23/06/18 20:46:04 INFO jvm.JvmMetrics: Initializing JVM Metrics with processname=JobTracker, sessionId=
23/06/18 20:46:04 INFO mapred.JobClient: Cannot initialize HDFS Metrics with processname=JobTracker, sessionId= - already initialized
23/06/18 20:46:04 INFO mapred.FileInputFormat: Total input paths to process : 1
23/06/18 20:46:04 INFO mapred.FileInputFormat: Total output paths to process : 1
23/06/18 20:46:04 INFO mapreduce.JobSubmitter: number of splits:2
23/06/18 20:46:04 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_local1277420989\_0001
23/06/18 20:46:05 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
23/06/18 20:46:05 INFO mapreduce.Job: Job tracking job: job\_local1277420989\_0001
23/06/18 20:46:05 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
23/06/18 20:46:05 INFO mapred.LocalJobRunner: config null
23/06/18 20:46:05 INFO mapred.LocalJobRunner: Waiting for map tasks
23/06/18 20:46:05 INFO mapred.LocalJobRunner: Starting task: attempt\_local1277420989\_0001\_m\_000000
23/06/18 20:46:05 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
23/06/18 20:46:05 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/varunurus/DeptName.txt:0+59
23/06/18 20:46:05 INFO mapred.MapTask: Map Task
23/06/18 20:46:05 INFO mapred.MapTask: (EQUATOR) 0 kvi:26114396((104857584)
23/06/18 20:46:05 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
23/06/18 20:46:05 INFO mapred.MapTask: soft limit at 8388e000
23/06/18 20:46:05 INFO mapred.MapTask: bufstart = 0; bufvold = 104857600
23/06/18 20:46:05 INFO mapred.MapTask: kvstart = 20234396; length = 6553600
23/06/18 20:46:05 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask\$MapOutputBuffer
23/06/18 20:46:05 INFO mapred.LocalJobRunner: 
23/06/18 20:46:05 INFO mapred.MapTask: Starting flush of map output
23/06/18 20:46:05 INFO mapred.MapTask: Spilling map output
23/06/18 20:46:05 INFO mapred.MapTask: bufstart = 0; bufvold = 63; bufvold = 104857600
23/06/18 20:46:05 INFO mapred.MapTask: kvstart = 20234396((104857584)); kvend = 20234384((104857536)); length = 13/6553600
23/06/18 20:46:05 INFO mapred.MapTask: Map output flushed
23/06/18 20:46:05 INFO mapred.Task: Task attempt\_local1277420989\_0001\_m\_000000.0 is done. And is in the process of committing
23/06/18 20:46:06 INFO mapred.LocalJobRunner: hdfs://localhost:54310/varunurus/DeptName.txt:0+59
23/06/18 20:46:06 INFO mapred.Task: Task 'attempt\_local1277420989\_0001\_m\_000000.0' done.

Activities Terminal Jun 18 20:47

```

hduser@lse-VirtualBox:~/Desktop$ hadoop fs -cat /varunurus/_output-join/part-r-00000
Reduce output records=4
Split Input Records=10
Shuffled Maps=1
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=62
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=517472256
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=0
  File Output Format Counters
  Bytes Written=5
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -cat /varunurus/_output-join/part-r-00000
23/06/18 20:46:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
cat: /varunurus/_output-join/part-r-00000: No such file or directory
hduser@lse-VirtualBox:~/Desktop$ hadoop fs -ls /varunurus
23/06/18 20:46:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 9 items
-rw-r--r-- 1 hduser supergroup 888190 2023-06-18 20:35 /varunurus/_root0
-rw-r--r-- 1 hduser supergroup 59 2023-06-18 20:45 /varunurus/DeptName.txt
-rw-r--r-- 1 hduser supergroup 50 2023-06-18 20:45 /varunurus/DeptStrength.txt
drwxr-xr-x 1 hduser supergroup 0 2023-06-18 20:37 /varunurus/_output-mnmax
drwxr-xr-x 1 hduser supergroup 0 2023-06-18 20:37 /varunurus/_output-temp
drwxr-xr-x 1 hduser supergroup 0 2023-06-18 20:42 /varunurus/_output-topn
drwxr-xr-x 1 hduser supergroup 0 2023-06-18 20:33 /varunurus/_output-wc
-rw-r--r-- 1 hduser supergroup 105 2023-06-18 20:32 /varunurus/sample.txt

```

hduser@lse-VirtualBox:~/Desktop\$ hadoop fs -ls /varunurus/\_output-join
Found 2 items
-rw-r--r-- 1 hduser supergroup 0 2023-06-18 20:46 /varunurus/\_output-join/\_SUCCESS
-rw-r--r-- 1 hduser supergroup 85 2023-06-18 20:46 /varunurus/\_output-join/part-r-00000
hduser@lse-VirtualBox:~/Desktop\$ hadoop fs -cat /varunurus/\_output-join/part-r-00000
23/06/18 20:47:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
A1 50
B2 100
C13 Manufacturing 250
Dept ID Dept Name Total\_Employee
\*\*\* hduser@lse-VirtualBox:~/Desktop\$

## LAB 9. Program to print word count on scala shell and print “Hello world” on scala IDE

### Word Count

```
scala> val data = sc.textFile("sparkdata.txt")
data: org.apache.spark.rdd.RDD[String] = sparkdata.txt MapPartitionsRDD[6] at textFile at <console>:2
4

scala> data.collect;
res5: Array[String] = Array(welcome to spark by varunurs and also welcome to the user by akash;)

scala> val splitdata1 = data.flatMap(line=>line.split(" "));
splitdata1: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[7] at flatMap at <console>:25

scala> splitdata1.collect;
res6: Array[String] = Array(welcome, to, spark, by, varunurs, and, also, welcome, to, the, user, by,
akash;)

scala> val mapdata1 = splitdata1.map(word=>(word,1));
mapdata1: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[8] at map at <console>:25

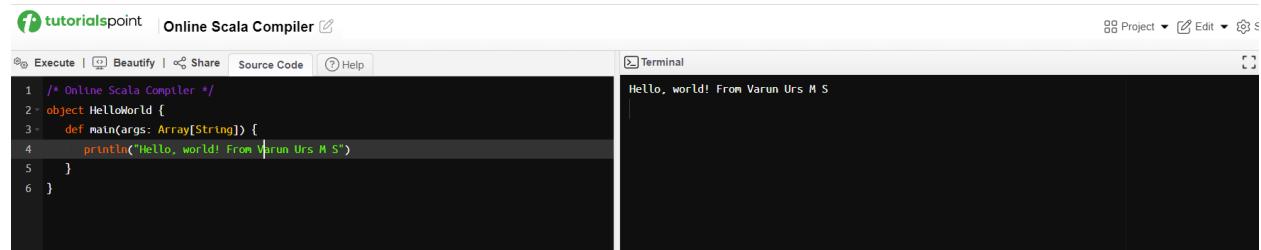
scala> mapdata1.collect;
res7: Array[(String, Int)] = Array((welcome,1), (to,1), (spark,1), (by,1), (varunurs,1), (and,1), (al
so,1), (welcome,1), (to,1), (the,1), (user,1), (by,1), (akash,1))

scala> val reddata = mapdata1.reduceByKey(_+_);
reddata: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[9] at reduceByKey at <console>:25

scala> reddata.collect;
res8: Array[(String, Int)] = Array((welcome,2), (varunurs,1), (spark,1), (akash,1), (to,2), (by,2),
(also,1), (user,1), (and,1), (the,1))

scala>
```

### Hello World



The screenshot shows the tutorialspoint Online Scala Compiler interface. On the left, the code editor contains the following Scala code:

```
/* Online Scala Compiler */
object HelloWorld {
  def main(args: Array[String]) {
    println("Hello, world! From Varun Urs M S")
  }
}
```

On the right, the terminal window displays the output: "Hello, world! From Varun Urs M S".

**LAB 10. Using RDD and FlaMap count how many times each word appears in a file and write out a list of words whose count is strictly greater than 4 using Spark**

```
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~/Desktop
```

```
textFile: org.apache.spark.rdd.RDD[String] = /home/bmscecse/Desktop/wc182.txt MapPartitionsRDD[19] at textFile at <console>:24
```

```
scala> val counts = textFile.flatMap(line=>line.split(" ")).map(word=>(word,1)).reduceByKey(_+_)
```

```
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[22] at reduceByKey at <console>:25
```

```
scala> counts.collect;
```

```
res9: Array[(String, Int)] = Array((hills,1), (Georgia,1), (have,3), (one,1), (will,1), (today!,1), (day,1), ([...],2), (my,1), (little,1), (four,1), (that,2), (a,3), (on,1), (dream,3), (I,3), (of,1), (children,1), (red,1), (the,1))
```

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

```
import scala.collection.immutable.ListMap
```

```
scala> val sorted = ListMap(counts.collect.sortWith(_.2>_.2):_*)
```

```
<console>:1: error: ')' expected but double literal found.
```

```
    val sorted = ListMap(counts.collect.sortWith(_.2>_.2):_*)
```

```
                                ^
```

```
scala> val sorted = ListMap(counts.collect.sortWith(_.2>_.2):_*)
```

```
<console>:1: error: ')' expected but double literal found.
```

```
    val sorted = ListMap(counts.collect.sortWith(_.2>_.2):_*)
```

```
                                ^
```

```
scala> val sorted = ListMap(counts.collect.sortWith(_.2>_.2):_*)
```

```
sorted: scala.collection.immutable.ListMap[String,Int] = ListMap(have -> 3, a -> 3, dream -> 3, I -> 3, [...] -> 2, that -> 2, hills -> 1, Georgia -> 1, one -> 1, will -> 1, today! -> 1, day -> 1, my -> 1, little -> 1, four -> 1, on -> 1, of -> 1, children -> 1, red -> 1, the -> 1)
```

```
scala> println(sorted)
```

```
ListMap(have -> 3, a -> 3, dream -> 3, I -> 3, [...] -> 2, that -> 2, hills -> 1, Georgia -> 1, one -> 1, will -> 1, today! -> 1, day -> 1, my -> 1, little -> 1, four -> 1, on -> 1, of -> 1, children -> 1, red -> 1, the -> 1)
```

```
scala> for((k,v)<-sorted){if(v>4){print(k + ",") print(v) println()}}
```

```
<console>:27: error: value print is not a member of Unit
```

```
    for((k,v)<-sorted){if(v>4){print(k + ",") print(v) println()}}
```

```
                                ^
```

```
scala> for((k,v)<-sorted){if(v>4){print(k + ",") print(v) println()}}
```

```
<console>:27: error: value print is not a member of Unit
```

```
    for((k,v)<-sorted){if(v>4){print(k + ",") print(v) println()}}
```

```
                                ^
```

```
scala> for((k,v)<-sorted){if(v>4){print(k + ","); print(v); println()}}
```

```
scala> for((k,v)<-sorted){if(v>4){print(k + ","); print(v); println()}}
```

```
scala> for((k,v)<-sorted){if(v>2){print(k + ","); print(v); println()}}
```

```
have,3
```

```
a,3
```

```
dream,3
```

```
I,3
```

```
scala> []
```