VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka.

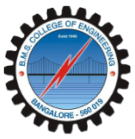
LAB REPORT
on

COURSE TITLE

Submitted by

MANYA VAID (1BM22CS150)

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



B.M.S. COLLEGE OF ENGINEERING
(Autonomous Institution under VTU)
BENGALURU-560019
Feb-2024 to July-2024

B. M. S. College of Engineering,

Bull Temple Road, Bangalore 560019

(Affiliated To Visvesvaraya Technological University, Belgaum)

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the Lab work entitled "Big Data Analytics" carried out by **MANYA VAID (1BM22CS150),** who is bonafide student of **B. M. S. College of Engineering.** It is in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visvesvaraya Technological University, Belgaum during the year 2024. The Lab report has been approved as it satisfies the academic requirements in respect of a **Big Data Analytics-(23CS6PCBDA)** work prescribed for the said degree.

Amrutha
Assistant Professor
Department of CSE
BMSCE, Bengaluru

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

`

Index Sheet

Course Outcome

CO1	Apply the concepts of NoSQL , Hadoop , Spark for a given task.
CO2	Analyse data analytic techniques for a given problem .
CO3	Conduct experiments using data analytics mechanisms for a given problem.

LAB PROGRAM - 01

MongoDB - CRUD Operations Demonstration

Code with Output:

```
C:\Users\HP>mongosh "mongodb+srv://cluster0.j87hg.mongodb.net/" --apiVersion 1 --username manyacs22 --password Manyav20
Current Mongosh Log ID: 67c7290d5703edf7edfa4213
Connecting to:          mongodb+srv://<credentials>@cluster0.j87hg.mongodb.net/?appName=mongosh+2.4.0
Using MongoDB:          8.0.5 (API Version 1)
Using Mongosh:          2.4.0

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


To help improve our products, anonymous usage data is collected and sent to MongoDB periodically (https://www.mongodb.com/legal/privacy-policy).
You can opt-out by running the disableTelemetry() command.

Atlas atlas-ieg1nq-shard-0 [primary] test> db.createCollection("Student");
{ ok: 1 }
Atlas atlas-ieg1nq-shard-0 [primary] test> db.Student.insert({RoRollNo:1,Age:21,Cont:9876,email:"antara.de9@gmail.com"});
DeprecationWarning: Collection.insert() is deprecated. Use insertOne, insertMany, or bulkWrite.
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c729435703edf7edfa4214') }
}
Atlas atlas-ieg1nq-shard-0 [primary] test> db.Student.insert({RoRollNo:2,Age:22,Cont:9976,email:"anushka.de9@gmail.com"});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c7294a5703edf7edfa4215') }
}
Atlas atlas-ieg1nq-shard-0 [primary] test> db.Student.insert({RoRollNo:3,Age:21,Cont:5576,email:"anubhav.de9@gmail.com"});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c729505703edf7edfa4216') }
}
```

```
Atlas atlas-ieg1nq-shard-0 [primary] test> db.Student.insert({RoRollNo:4,Age:20,Cont:4476,email:"pani.de9@gmail.com"});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c729565703edf7edfa4217') }
}
Atlas atlas-ieg1nq-shard-0 [primary] test> db.Student.insert({RoRollNo:10,Age:23,Cont:2276,email:"rekha.de9@gmail.com"});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c7295c5703edf7edfa4218') }
}
Atlas atlas-ieg1nq-shard-0 [primary] test>  db.Student.find()
[
  {
    _id: ObjectId('67c729435703edf7edfa4214'),
    RollNo: 1,
    Age: 21,
    Cont: 9876,
    email: 'antara.de9@gmail.com'
  },
  {
    _id: ObjectId('67c7294a5703edf7edfa4215'),
    RollNo: 2,
    Age: 22,
    Cont: 9976,
    email: 'anushka.de9@gmail.com'
  },
  {
    _id: ObjectId('67c729505703edf7edfa4216'),
    RollNo: 3,
    Age: 21,
    Cont: 5576,
    email: 'anubhav.de9@gmail.com'
  },
  {
```

```
  {
    _id: ObjectId('67c729565703edf7edfa4217'),
    RollNo: 4,
    Age: 20,
    Cont: 4476,
    email: 'pani.de9@gmail.com'
  },
  {
    _id: ObjectId('67c7295c5703edf7edfa4218'),
    RollNo: 10,
    Age: 23,
    Cont: 2276,
    email: 'rekha.de9@gmail.com'
  }
]
Atlas atlas-ieg1nq-shard-0 [primary] test>  db.Student.find()
[
  {
    _id: ObjectId('67c729435703edf7edfa4214'),
    RollNo: 1,
    Age: 21,
    Cont: 9876,
    email: 'antara.de9@gmail.com'
  },
  {
    _id: ObjectId('67c7294a5703edf7edfa4215'),
    RollNo: 2,
    Age: 22,
    Cont: 9976,
    email: 'anushka.de9@gmail.com'
  },
  {
    _id: ObjectId('67c729505703edf7edfa4216'),
    RollNo: 3,
    Age: 21,
```

```
    Cont: 5576,
    email: 'anubhav.de9@gmail.com'
  },
  {
    _id: ObjectId('67c729565703edf7edfa4217'),
    RollNo: 4,
    Age: 20,
    Cont: 4476,
    email: 'pani.de9@gmail.com'
  },
  {
    _id: ObjectId('67c7295c5703edf7edfa4218'),
    RollNo: 10,
    Age: 23,
    Cont: 2276,
    email: 'rekha.de9@gmail.com'
  }
]
Atlas atlas-ieg1nq-shard-0 [primary] test> db.Student.insert({RoAtlas atlas-ieg1nq-shard-0 [primary] test> db.Student.insert({RollNo:
11,Age:22,Name: "ABC",Cont:2276,email:"rea.de9@gmail.com"});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c729a35703edf7edfa4219') }
}
Atlas atlas-ieg1nq-shard-0 [primary] test> db.Student.update({RoRollNo:11,Name:"ABC"},{$set:{Name:"FEM"}})
DeprecationWarning: Collection.update() is deprecated. Use updateOne, updateMany, or bulkWrite.
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
```

LAB PROGRAM - 02

Perform the following DB operations using Cassandra:

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

Code with Output:

```
bmscecse@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.8 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> CREATE KEYSPACE Employee2
   ... WITH replication = {
   ...   'class': 'SimpleStrategy',
   ...   'replication_factor': 1
   ... };
cqlsh>
cqlsh> USE Employee2 ;
cqlsh:employee2> CREATE TABLE Employee_Info (
             ...   Emp_Id int PRIMARY KEY,
             ...   Emp_Name text,
             ...   Designation text,
             ...   Date_of_Joining date,
             ...   Salary decimal,
             ...   Dept_Name text
             ... );
cqlsh:employee2> BEGIN BATCH
             ... INSERT INTO Employee_Info (Emp_Id, Emp_Name, Designation, Date_of_Joining, S
alary, Dept_Name)
             ... VALUES (121, 'Alice', 'Engineer', '2020-03-01', 60000.00, 'IT');
             ...
             ... INSERT INTO Employee_Info (Emp_Id, Emp_Name, Designation, Date_of_Joining, S
alary, Dept_Name)
             ... VALUES (122, 'Bob', 'Manager', '2019-06-15', 75000.00, 'HR');
             ...
             ... INSERT INTO Employee_Info (Emp_Id, Emp_Name, Designation, Date_of_Joining, S
alary, Dept_Name)
             ... VALUES (123, 'Charlie', 'Analyst', '2021-01-10', 50000.00, 'Finance');
             ... APPLY BATCH;
cqlsh:employee2> UPDATE Employee_Info
             ... SET Emp_Name = 'Alicia', Dept_Name = 'Research'
             ... WHERE Emp_Id = 121;
cqlsh:employee2> CREATE TABLE Employee_Salary_Sorted (
             ...   Dept_Name text,
             ...   Salary decimal,
             ...   Emp_Id int,
             ...   Emp_Name text,
             ...   Designation text,
             ...   Date_of_Joining date,
             ...   PRIMARY KEY (Dept_Name, Salary)
             ... ) WITH CLUSTERING ORDER BY (Salary DESC);
cqlsh:employee2> ALTER TABLE Employee_Info
             ... ADD Projects set<text>;
cqlsh:employee2> UPDATE Employee_Info
             ... SET Projects = {'Project A', 'Project B'}
             ... WHERE Emp_Id = 121;
cqlsh:employee2> UPDATE Employee_Info
             ... SET Projects = {'Project C'}
             ... WHERE Emp_Id = 122;
cqlsh:employee2> UPDATE Employee_Info
```

```
cqlsh:employee2> UPDATE Employee_Info
             ... USING TTL 15
             ... SET Emp_Name = 'Temporary', Dept_Name = 'Temp'
             ... WHERE Emp_Id = 124;
cqlsh:employee2> INSERT INTO Employee_Info (Emp_Id, Emp_Name, Designation, Date_of_Joining, S
alary, Dept_Name)
             ... VALUES (124, 'TempUser', 'Intern', '2025-04-08', 30000.00, 'TempDept')
             ... USING TTL 15;
cqlsh:employee2> SELECT * FROM Employee_Info;

 emp_id | date_of_joining | dept_name | designation | emp_name | projects                    |
 salary
--------+-----------------+-----------+-------------+----------+-----------------------------+
----------
    123 |      2021-01-10 |   Finance |     Analyst |  Charlie |                        null |
 50000.00
    122 |      2019-06-15 |        HR |     Manager |      Bob |               {'Project C'} |
 75000.00
    121 |      2020-03-01 |  Research |    Engineer |   Alicia | {'Project A', 'Project B'} |
 60000.00

(3 rows)
cqlsh:employee2>
```

LAB PROGRAM - 03

Perform the following DB operations using Cassandra:

● Create a keyspace by name Library

● 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

● Insert the values into the table in batch

● Display the details of the table created and increase the value of the counter

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

● Export the created column to a csv file

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

Code with Output:

```
bmscecse@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.8 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> CREATE KEYSPACE Library
   ... WITH replication = {
   ...   'class': 'SimpleStrategy',
   ...   'replication_factor': 1
   ... };
cqlsh> USE Library;
cqlsh:library> CREATE TABLE Book_Counter (
           ...   Stud_Id int,
           ...   Book_Name text,
           ...   Counter_Value counter,
           ...   PRIMARY KEY (Stud_Id, Book_Name)
           ... );
cqlsh:library> CREATE TABLE Library_Info (
           ...   Stud_Id int PRIMARY KEY,
           ...   Stud_Name text,
           ...   Book_Name text,
           ...   Book_Id text,
           ...   Date_of_Issue date
           ... );
cqlsh:library> BEGIN BATCH
           ...
           ... INSERT INTO Library_Info (Stud_Id, Stud_Name, Book_Name, Book_Id, Date_of_Issu
e)
           ... VALUES (112, 'Ravi', 'BDA', 'B001', '2025-04-07');
           ...
           ... UPDATE Book_Counter
           ... SET Counter_Value = Counter_Value + 1
           ... WHERE Stud_Id = 112 AND Book_Name = 'BDA';
           ...
           ... APPLY BATCH;
InvalidRequest: Error from server: code=2200 [Invalid query] message="Counter and non-counter
 mutations cannot exist in the same batch"
cqlsh:library> INSERT INTO Library_Info (
           ...     Stud_Id,
           ...     Stud_Name,
           ...     Book_Name,
           ...     Book_Id,
           ...     Date_of_Issue
           ... )
           ... VALUES (
           ...     112,
           ...     'Ravi',
           ...     'BDA',
           ...     'B001',
           ...     '2025-04-07'
           ... );
cqlsh:library> UPDATE Book_Counter
           ... SET Counter_Value = Counter_Value + 1
           ... WHERE Stud_Id = 112 AND Book_Name = 'BDA';
```

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

 stud_id | book_id | book_name | date_of_issue | stud_name
---------+---------+-----------+---------------+-----------
     112 |    B001 |       BDA |    2025-04-07 |      Ravi

(1 rows)
cqlsh:library> UPDATE Book_Counter
           ... SET Counter_Value = Counter_Value + 1
           ... WHERE Stud_Id = 112 AND Book_Name = 'BDA';
cqlsh:library> SELECT * FROM Book_Counter WHERE Stud_Id = 112 AND Book_Name = 'BDA';

 stud_id | book_name | counter_value
---------+-----------+---------------
     112 |       BDA |             2

(1 rows)
cqlsh:library> COPY Library_Info TO 'library_info.csv' WITH HEADER = TRUE;
Using 16 child processes

Starting copy of library.library_info with columns [stud_id, book_id, book_name, date_of_issu
e, stud_name].
Processed: 1 rows; Rate:      12 rows/s; Avg. rate:      12 rows/s
1 rows exported to 1 files in 0.121 seconds.
```

```
0 rows imported from 1 files in 0.365 seconds (0 skipped).
cqlsh:library> COPY Library_Info (Stud_Id, Book_Id, Book_Name, Date_of_Issue, Stud_Name)
           ... FROM '/home/bmscecse/library_info.csv'
           ... WITH HEADER = TRUE;
Using 16 child processes

Starting copy of library.library_info with columns [stud_id, book_id, book_name, date_of_issu
e, stud_name].
Processed: 3 rows; Rate:       6 rows/s; Avg. rate:       8 rows/s
3 rows imported from 1 files in 0.365 seconds (0 skipped).
cqlsh:library> COPY Library_Info (Stud_Id, Book_Id, Book_Name, Date_of_Issue, Stud_Name)
           ... TO '/home/bmscecse/library_info_export.csv'
           ... WITH HEADER = TRUE;
Using 16 child processes

Starting copy of library.library_info with columns [stud_id, book_id, book_name, date_of_issu
e, stud_name].
Processed: 3 rows; Rate:      69 rows/s; Avg. rate:      69 rows/s
3 rows exported to 1 files in 0.076 seconds.
cqlsh:library>
cqlsh:library>
```

LAB PROGRAM - 04

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

Code with Output:

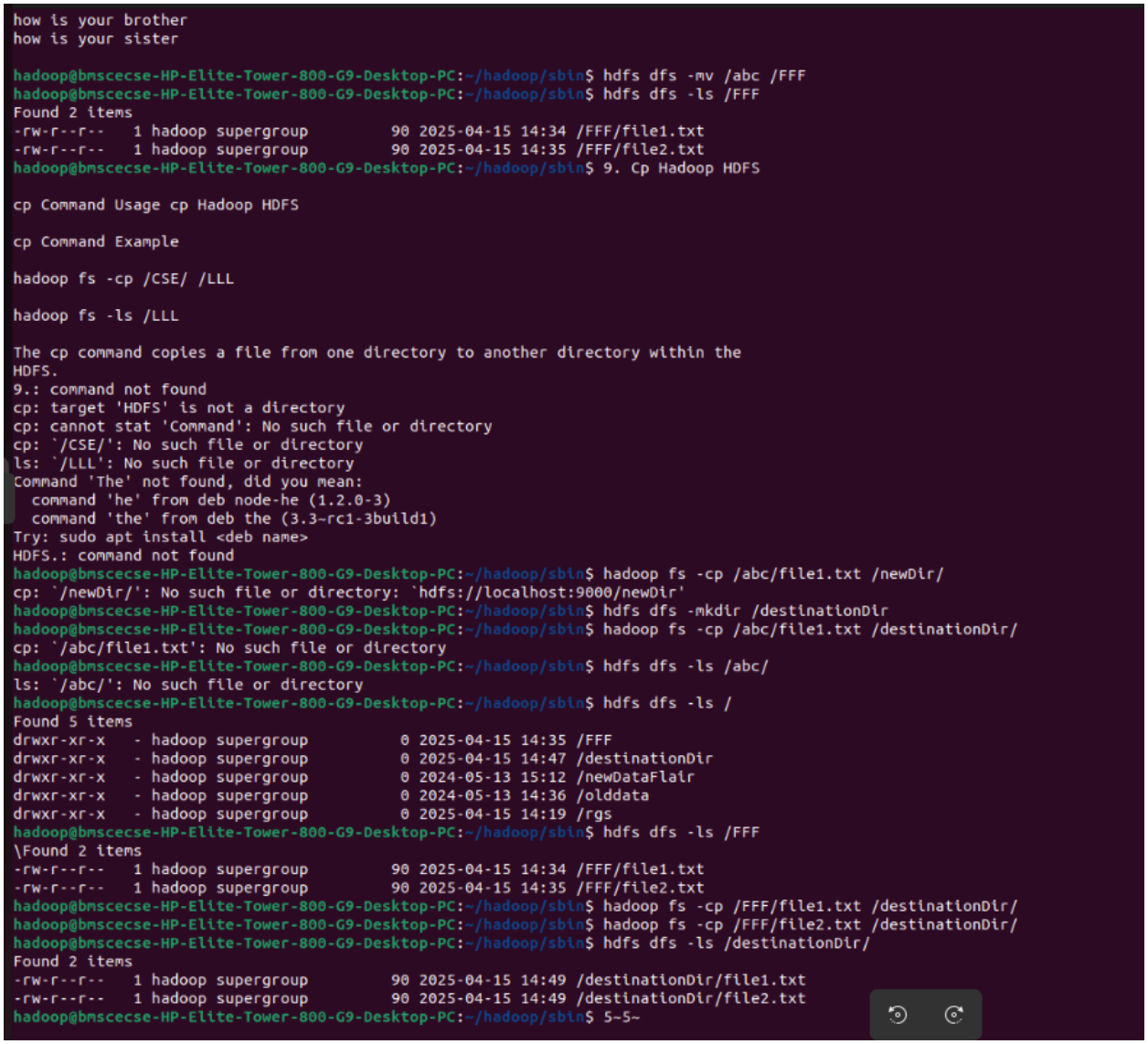
```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -mkdir /abc
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-all sh
start-all: command not found
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ ./start-all.sh
bash: ./start-all.sh: No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd /usr/local/hadoop/sbin
bash: cd: /usr/local/hadoop/sbin: No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
 or    hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
  where CLASSNAME is a user-provided Java class

  OPTIONS is none or any of:

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

  SUBCOMMAND is one of:


    Admin Commands:

daemonlog     get/set the log level for each daemon

    Client Commands:

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

```
version       print the version

    Daemon Commands:

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

SUBCOMMAND may print help when invoked w/o parameters or with -h.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ which hadoop
/home/hadoop/hadoop/bin/hadoop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd /home/hadoop/hadoop/sbin
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 9900.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 10056.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10334.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 10644.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10807.  Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ jps
10644 ResourceManager
10807 NodeManager
10056 DataNode
9900 NameNode
10334 SecondaryNameNode
16111 Jps
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -mkdir /abc
mkdir: `/abc': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -rm -r /abc
Deleted /abc
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -mkdir /abc
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /
Found 4 items
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:32 /abc
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 15:12 /newDataFlair
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:36 /olddata
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:19 /rgs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -put /home/hadoop/Desktop/file1.txt /abc/file1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -cat /abc/file1.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyFromLocal /home/hadoop/Desktop/file1.txt /abc/file1.txt
copyFromLocal: `/abc/file1.txt': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyFromLocal /home/hadoop/Desktop/file1.txt /abc/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /abc
Found 2 items
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:34 /abc/file1.txt
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:35 /abc/file2.txt
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyFromLocal /home/hadoop/Desktop/file1.txt /abc/file1.txt
copyFromLocal: `/abc/file1.txt': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyFromLocal /home/hadoop/Desktop/file1.txt /abc/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /abc
Found 2 items
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:34 /abc/file1.txt
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:35 /abc/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -cat /abc/file2.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/WC.txt /home/hduser/Downloads/WWC.txt
get: `/home/hduser/Downloads/WWC.txt': No such file or directory: `file:///home/hduser/Downloads/WWC.txt'
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hduser/Downloads/
ls: cannot access '/home/hduser/Downloads/': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ mkdir -p /home/hduser/Downloads/
mkdir: cannot create directory ‘/home/hduser’: Permission denied
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo mkdir -p /home/hduser/Downloads/
[sudo] password for hadoop:
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hduser/
Downloads
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/WC.txt /home/hduser/Downloads/WWC.txt
get: `/abc/WC.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/WC.txt /home/hduser/Downloads/WWC.txt
get: `/abc/WC.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /abc/
Found 2 items
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:34 /abc/file1.txt
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:35 /abc/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/file1.txt /home/hduser/Downloads/file1.txt
get: /home/hduser/Downloads/file1.txt._COPYING_ (Permission denied)
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls -ld /home/hduser/Downloads/
drwxr-xr-x 2 root root 4096 Apr 15 14:37 /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo chown -R hadoop:hadoop /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls -ld /home/hduser/Downloads/
drwxr-xr-x 2 hadoop hadoop 4096 Apr 15 14:37 /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls -ld /home/hduser/Downloads/
drwxr-xr-x 2 hadoop hadoop 4096 Apr 15 14:37 /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/file1.txt /home/hduser/Downloads/file1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -getmerge /abc/WC.txt /abc/WC2.txt /home/hduser/Desktop/Merge.txt
getmerge: `/abc/WC.txt': No such file or directory
getmerge: `/abc/WC2.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -getmerge /abc/file1.txt /abc/file2.txt /home/hduser/Desktop/Merge.txt
getmerge: Mkdirs failed to create file:/home/hduser/Desktop (exists=false, cwd=file:/home/hadoop/hadoop/sbin)
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo mkdir -p /home/hduser/Desktop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo chown -R hadoop:hadoop /home/hduser/Desktop
sudo chmod -R 775 /home/hduser/Desktop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -getmerge /abc/file1.txt /abc/file2.txt /home/hduser/Desktop/Merge.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ cat /home/hduser/Desktop/Merge.txt
hi how are you
how is your job
how is your family
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyToLocal /abc/WC.txt /home/hduser/Desktop
copyToLocal: `/abc/WC.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyToLocal /abc/file1.txt /home/hduser/Desktop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ cat /home/hduser/Desktop/file1.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -cat /abc/file1.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -cat /abc/file2.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -mv /abc /FFF
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /FFF
Found 2 items
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:34 /FFF/file1.txt
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:35 /FFF/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ 9. Cp Hadoop HDFS

cp Command Usage cp Hadoop HDFS

cp Command Example

hadoop fs -cp /CSE/ /LLL

hadoop fs -ls /LLL

The cp command copies a file from one directory to another directory within the
HDFS.
9.: command not found
cp: target 'HDFS' is not a directory
cp: cannot stat 'Command': No such file or directory
cp: `/CSE/': No such file or directory
ls: `/LLL': No such file or directory
Command 'The' not found, did you mean:
  command 'he' from deb node-he (1.2.0-3)
  command 'the' from deb the (3.3~rc1-3build1)
Try: sudo apt install <deb name>
HDFS.: command not found
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -cp /abc/file1.txt /newDir/
cp: `/newDir/': No such file or directory: `hdfs://localhost:9000/newDir'
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -mkdir /destinationD
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -cp /abc/file1.txt                Dir/
```


how is your brother
how is your sister
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -mv /abc /FFF
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /FFF
Found 2 items
-rw-r--r-- 1 hadoop supergroup 90 2025-04-15 14:34 /FFF/file1.txt
-rw-r--r-- 1 hadoop supergroup 90 2025-04-15 14:35 /FFF/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ 9. Cp Hadoop HDFS
cp Command Usage cp Hadoop HDFS
cp Command Example
hadoop fs -cp /CSE/ /LLL
hadoop fs -ls /LLL
The cp command copies a file from one directory to another directory within the
HDFS.
9.: command not found
cp: target 'HDFS' is not a directory
cp: cannot stat 'Command': No such file or directory
cp: `/CSE/': No such file or directory
ls: `/LLL': No such file or directory
Command 'The' not found, did you mean:
command 'he' from deb node-he (1.2.0-3)
command 'the' from deb the (3.3~rc1-3build1)
Try: sudo apt install <deb name>
HDFS.: command not found
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -cp /abc/file1.txt /newDir/
cp: `/newDir/': No such file or directory: `hdfs://localhost:9000/newDir'
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -mkdir /destinationDir
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -cp /abc/file1.txt /destinationDir/
cp: `/abc/file1.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /abc/
ls: `/abc/': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /
Found 5 items
drwxr-xr-x - hadoop supergroup 0 2025-04-15 14:35 /FFF
drwxr-xr-x - hadoop supergroup 0 2025-04-15 14:47 /destinationDir
drwxr-xr-x - hadoop supergroup 0 2024-05-13 15:12 /newDataFlair
drwxr-xr-x - hadoop supergroup 0 2024-05-13 14:36 /olddata
drwxr-xr-x - hadoop supergroup 0 2025-04-15 14:19 /rgs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /FFF
\Found 2 items
-rw-r--r-- 1 hadoop supergroup 90 2025-04-15 14:34 /FFF/file1.txt
-rw-r--r-- 1 hadoop supergroup 90 2025-04-15 14:35 /FFF/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -cp /FFF/file1.txt /destinationDir/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -cp /FFF/file2.txt /destinationDir/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /destinationDir/
Found 2 items
-rw-r--r-- 1 hadoop supergroup 90 2025-04-15 14:49 /destinationDir/file1.txt
-rw-r--r-- 1 hadoop supergroup 90 2025-04-15 14:49 /destinationDir/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ 5~5~

```
SUBCOMMAND may print help when invoked w/o parameters or with -h.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: $ cd /home/hadoop/hadoop/sbin
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 9900.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 10056.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10334.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 10644.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10807.  Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ jps
12226 Jps
10644 ResourceManager
10807 NodeManager
10056 DataNode
9900 NameNode
10334 SecondaryNameNode
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 9900.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 10056.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10334.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 10644.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10807.  Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ jps
13012 Jps
10644 ResourceManager
10807 NodeManager
10056 DataNode
9900 NameNode
10334 SecondaryNameNode
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 15:12 /newDataFlair
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:36 /olddata
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/rekhags/Desktop/file1.txt /rgs/test.txt
copyFromLocal: `/rgs/test.txt': No such file or directory: `hdfs://localhost:9000/rgs/test.txt'
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -mkdir /rgs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/rekhags/Desktop/file1.txt /rgs/test.txt
copyFromLocal: `/home/rekhags/Desktop/file1.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hadoop/Desktop/
file1.txt  first.txt  Untitled.jar  WC.jar
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ nano /home/hadoop/Desktop/file1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /hom                ktop/file1.txt /rgs/test.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -mkdir /rgs
mkdir: `/rgs': File exists
```

```
10644 ResourceManager
10807 NodeManager
10056 DataNode
9900 NameNode
10334 SecondaryNameNode
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 9900.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 10056.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10334.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 10644.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10807.  Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ jps
13012 Jps
10644 ResourceManager
10807 NodeManager
10056 DataNode
9900 NameNode
10334 SecondaryNameNode
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 15:12 /newDataFlair
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:36 /olddata
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/rekhags/Desktop/file1.txt /rgs/test.txt
copyFromLocal: `/rgs/test.txt': No such file or directory: `hdfs://localhost:9000/rgs/test.txt'
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -mkdir /rgs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/rekhags/Desktop/file1.txt /rgs/test.txt
copyFromLocal: `/home/rekhags/Desktop/file1.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hadoop/Desktop/
file1.txt  first.txt  Untitled.jar  WC.jar
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ nano /home/hadoop/Desktop/file1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/hadoop/Desktop/file1.txt /rgs/test.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -mkdir /rgs
mkdir: `/rgs': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -ls /
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 15:12 /newDataFlair
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:36 /olddata
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:19 /rgs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/hadoop/Desktop/file1.txt /rgs/test.txt
copyFromLocal: `/rgs/test.txt': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop jar /home/hadoop/Desktop/WordCount.jar wordcount.WordCount /rgs/test.txt /output/
JAR does not exist or is not a normal file: /home/hadoop/Desktop/WordCount.jar
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hadoop/Desktop/WordCount.jar
ls: cannot access '/home/hadoop/Desktop/WordCount.jar': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ START ALL sh
Command 'START' not found, did you mean:
  command 'STAR' from deb rna-star (2.7.10a+dfsg-1)
Try: sudo apt install <deb name>
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$
```

LAB PROGRAM - 05

Implement WordCount Program on Hadoop framework.

Code with Output:

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfs -mkdir /abc
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-all sh
start-all: command not found
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ ./start-all.sh
bash: ./start-all.sh: No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd /usr/local/hadoop/sbin
bash: cd: /usr/local/hadoop/sbin: No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
 or    hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
  where CLASSNAME is a user-provided Java class

  OPTIONS is none or any of:

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

  SUBCOMMAND is one of:


    Admin Commands:

daemonlog     get/set the log level for each daemon

    Client Commands:

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

```
version       print the version

    Daemon Commands:

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

SUBCOMMAND may print help when invoked w/o parameters or with -h.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ which hadoop
/home/hadoop/hadoop/bin/hadoop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd /home/hadoop/hadoop/sbin
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 9900.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 10056.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10334.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 10644.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10807.  Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ jps
10644 ResourceManager
10807 NodeManager
10056 DataNode
9900 NameNode
10334 SecondaryNameNode
16111 Jps
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -mkdir /abc
mkdir: `/abc': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -rm -r /abc
Deleted /abc
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -mkdir /abc
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /
Found 4 items
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:32 /abc
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 15:12 /newDataFlair
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:36 /olddata
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:19 /rgs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -put /home/hadoop/Desktop/file1.txt /abc/file1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -cat /abc/file1.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyFromLocal /home/hadoop/Desktop/file1.txt /abc/file1.txt
copyFromLocal: `/abc/file1.txt': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyFromLocal /home/hadoop/Desktop/file1.txt /abc/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /abc
Found 2 items
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:34 /abc/file1.txt
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:35 /abc/file2.txt
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyFromLocal /home/hadoop/Desktop/file1.txt /abc/file1.txt
copyFromLocal: `/abc/file1.txt': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyFromLocal /home/hadoop/Desktop/file1.txt /abc/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /abc
Found 2 items
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:34 /abc/file1.txt
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:35 /abc/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -cat /abc/file2.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/WC.txt /home/hduser/Downloads/WWC.txt
get: `/home/hduser/Downloads/WWC.txt': No such file or directory: `file:///home/hduser/Downloads/WWC.txt'
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hduser/Downloads/
ls: cannot access '/home/hduser/Downloads/': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ mkdir -p /home/hduser/Downloads/
mkdir: cannot create directory ‘/home/hduser’: Permission denied
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo mkdir -p /home/hduser/Downloads/
[sudo] password for hadoop:
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hduser/
Downloads
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/WC.txt /home/hduser/Downloads/WWC.txt
get: `/abc/WC.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/WC.txt /home/hduser/Downloads/WWC.txt
get: `/abc/WC.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /abc/
Found 2 items
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:34 /abc/file1.txt
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:35 /abc/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/file1.txt /home/hduser/Downloads/file1.txt
get: /home/hduser/Downloads/file1.txt._COPYING_ (Permission denied)
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls -ld /home/hduser/Downloads/
drwxr-xr-x 2 root root 4096 Apr 15 14:37 /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo chown -R hadoop:hadoop /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls -ld /home/hduser/Downloads/
drwxr-xr-x 2 hadoop hadoop 4096 Apr 15 14:37 /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls -ld /home/hduser/Downloads/
drwxr-xr-x 2 hadoop hadoop 4096 Apr 15 14:37 /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/file1.txt /home/hduser/Downloads/file1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -getmerge /abc/WC.txt /abc/WC2.txt /home/hduser/Desktop/Merge.txt
getmerge: `/abc/WC.txt': No such file or directory
getmerge: `/abc/WC2.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -getmerge /abc/file1.txt /abc/file2.txt /home/hduser/Desktop/Merge.txt
getmerge: Mkdirs failed to create file:/home/hduser/Desktop (exists=false, cwd=file:/home/hadoop/hadoop/sbin)
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo mkdir -p /home/hduser/Desktop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo chown -R hadoop:hadoop /home/hduser/Desktop
sudo chmod -R 775 /home/hduser/Desktop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -getmerge /abc/file1.txt /abc/file2.txt /home/hduser/Desktop/Merge.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ cat /home/hduser/Desktop/Merge.txt
hi how are you
how is your job
how is your family
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -cat /abc/file2.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/WC.txt /home/hduser/Downloads/WWC.txt
get: `/home/hduser/Downloads/WWC.txt': No such file or directory: `file:///home/hduser/Downloads/WWC.txt'
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hduser/Downloads/
ls: cannot access '/home/hduser/Downloads/': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ mkdir -p /home/hduser/Downloads/
mkdir: cannot create directory ‘/home/hduser’: Permission denied
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo mkdir -p /home/hduser/Downloads/
[sudo] password for hadoop:
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hduser/
Downloads
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/WC.txt /home/hduser/Downloads/WWC.txt
get: `/abc/WC.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/WC.txt /home/hduser/Downloads/WWC.txt
get: `/abc/WC.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /abc/
Found 2 items
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:34 /abc/file1.txt
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:35 /abc/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/file1.txt /home/hduser/Downloads/file1.txt
get: /home/hduser/Downloads/file1.txt._COPYING_ (Permission denied)
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls -ld /home/hduser/Downloads/
drwxr-xr-x 2 root root 4096 Apr 15 14:37 /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo chown -R hadoop:hadoop /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls -ld /home/hduser/Downloads/
drwxr-xr-x 2 hadoop hadoop 4096 Apr 15 14:37 /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls -ld /home/hduser/Downloads/
drwxr-xr-x 2 hadoop hadoop 4096 Apr 15 14:37 /home/hduser/Downloads/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -get /abc/file1.txt /home/hduser/Downloads/file1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -getmerge /abc/WC.txt /abc/WC2.txt /home/hduser/Desktop/Merge.txt
getmerge: `/abc/WC.txt': No such file or directory
getmerge: `/abc/WC2.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -getmerge /abc/file1.txt /abc/file2.txt /home/hduser/Desktop/Merge.txt
getmerge: Mkdirs failed to create file:/home/hduser/Desktop (exists=false, cwd=file:/home/hadoop/hadoop/sbin)
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo mkdir -p /home/hduser/Desktop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ sudo chown -R hadoop:hadoop /home/hduser/Desktop
sudo chmod -R 775 /home/hduser/Desktop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -getmerge /abc/file1.txt /abc/file2.txt /home/hduser/Desktop/Merge.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ cat /home/hduser/Desktop/Merge.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hi how are you
how is your job
how is your family
how is your brother
how is your sister
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyToLocal /abc/WC.txt /home/hduser/Desktop
copyToLocal: `/abc/WC.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -copyToLocal /abc/file1.txt /home/hduser/Desktop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ cat /home/hduser/Desktop/file1.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -cat /abc/file1.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -cat /abc/file2.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -mv /abc /FFF
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /FFF
Found 2 items
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:34 /FFF/file1.txt
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:35 /FFF/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ 9. Cp Hadoop HDFS

cp Command Usage cp Hadoop HDFS
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -cp /abc/file1.txt /newDir/
cp: `/newDir/': No such file or directory: `hdfs://localhost:9000/newDir'
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -mkdir /destinationDir
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -cp /abc/file1.txt /destinationDir/
cp: `/abc/file1.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /abc/
ls: `/abc/': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /
Found 5 items
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:35 /FFF
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:47 /destinationDir
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 15:12 /newDataFlair
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:36 /olddata
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:19 /rgs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /FFF
\Found 2 items
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:34 /FFF/file1.txt
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:35 /FFF/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -cp /FFF/file1.txt /destinationDir/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -cp /FFF/file2.txt /destinationDir/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hdfs dfs -ls /destinationDir/
Found 2 items
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:49 /destinationDir/file1.txt
-rw-r--r--   1 hadoop supergroup         90 2025-04-15 14:49 /destinationDir/file2.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ 5~5~
```

```
SUBCOMMAND may print help when invoked w/o parameters or with -h.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd /home/hadoop/hadoop/sbin
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 9900.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 10056.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10334.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 10644.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10807.  Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ jps
12226 Jps
10644 ResourceManager
10807 NodeManager
10056 DataNode
9900 NameNode
10334 SecondaryNameNode
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 9900.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 10056.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10334.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 10644.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10807.  Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ jps
13012 Jps
10644 ResourceManager
10807 NodeManager
10056 DataNode
9900 NameNode
10334 SecondaryNameNode
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 15:12 /newDataFlair
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:36 /olddata
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/rekhags/Desktop/file1.txt /rgs/test.txt
copyFromLocal: `/rgs/test.txt': No such file or directory: `hdfs://localhost:9000/rgs/test.txt'
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -mkdir /rgs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/rekhags/Desktop/file1.txt /rgs/test.txt
copyFromLocal: `/home/rekhags/Desktop/file1.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hadoop/Desktop/
file1.txt  first.txt  Untitled.jar  WC.jar
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ nano /home/hadoop/Desktop/file1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/hadoop/Desktop/file1.txt /rgs/test.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -mkdir /rgs
```

```
10644 ResourceManager
10807 NodeManager
10056 DataNode
9900 NameNode
10334 SecondaryNameNode
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 9900.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 10056.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10334.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ./start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 10644.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10807.  Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ jps
13012 Jps
10644 ResourceManager
10807 NodeManager
10056 DataNode
9900 NameNode
10334 SecondaryNameNode
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 15:12 /newDataFlair
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:36 /olddata
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/rekhags/Desktop/file1.txt /rgs/test.txt
copyFromLocal: `/rgs/test.txt': No such file or directory: `hdfs://localhost:9000/rgs/test.txt'
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -mkdir /rgs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/rekhags/Desktop/file1.txt /rgs/test.txt
copyFromLocal: `/home/rekhags/Desktop/file1.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hadoop/Desktop/
file1.txt  first.txt  Untitled.jar  WC.jar
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ nano /home/hadoop/Desktop/file1.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/hadoop/Desktop/file1.txt /rgs/test.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -mkdir /rgs
mkdir: `/rgs': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -ls /
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 15:12 /newDataFlair
drwxr-xr-x   - hadoop supergroup          0 2024-05-13 14:36 /olddata
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:19 /rgs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop fs -copyFromLocal /home/hadoop/Desktop/file1.txt /rgs/test.txt
copyFromLocal: `/rgs/test.txt': File exists
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ hadoop jar /home/hadoop/Desktop/WordCount.jar wordcount.WordCount /rgs/test.txt /output/
JAR does not exist or is not a normal file: /home/hadoop/Desktop/WordCount.jar
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ ls /home/hadoop/Desktop/WordCount.jar
ls: cannot access '/home/hadoop/Desktop/WordCount.jar': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ START ALL sh
Command 'START' not found, did you mean:
  command 'STAR' from deb rna-star (2.7.10a+dfsg-1)
Try: sudo apt install <deb name>
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop_wordcount$ nano WordCount.java
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop_wordcount$
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop_wordcount$ javac -classpath $(hadoop classpath) -d /home/hadoop/wordcount_classes WordCount.java
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop_wordcount$ jar -cvf /home/hadoop/Desktop/WordCount.jar -C /home/hadoop/wordcount_classes/ .
added manifest
adding: wordcount/(in = 0) (out= 0)(stored 0%)
adding: wordcount/WordCount$IntSumReducer.class(in = 1775) (out= 756)(deflated 57%)
adding: wordcount/WordCount$TokenizerMapper.class(in = 1858) (out= 803)(deflated 56%)
adding: wordcount/WordCount.class(in = 1501) (out= 807)(deflated 46%)
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop_wordcount$ hadoop fs -cat /output/part-00000
cat: `/output/part-00000': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop_wordcount$
```

LAB PROGRAM - 06

From the following link extract the weather data:

https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all

Create a Map Reduce program to:

a) Find average temperature for each year from NCDC data set.

b) Find the mean max temperature for every month.

Code with Output:

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -mkdir /ncdc_input
mkdir: Call From bmscecse-HP-Elite-Tower-600-G9-Desktop-PC/127.0.1.1 to localhost:9000 failed on connection exception: java.net.ConnectException: Connection refused; For more details see:  http://wiki.apache.org/hadoop/ConnectionRefused
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ start-dfs.sh
start-yarn.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bmscecse-HP-Elite-Tower-600-G9-Desktop-PC]
Starting resourcemanager
Starting nodemanagers
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ jps
16032 SecondaryNameNode
15745 DataNode
15556 NameNode
7093 org.eclipse.equinox.launcher_1.6.1000.v20250227-1734.jar
16504 NodeManager
16937 Jps
16318 ResourceManager
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -mkdir /ncdc_input
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ git clone https://github.com/tomwhite/hadoop-book.git
fatal: destination path 'hadoop-book' already exists and is not an empty directory.
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cd hadoop-book/input/ncdc/all/
hdfs dfs -put * /ncdc_input
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop-book/input/ncdc/all$ hdfs dfs -ls /ncdc_input
Found 2 items
-rw-r--r--   1 hadoop supergroup      73867 2025-05-06 15:11 /ncdc_input/1901.gz
-rw-r--r--   1 hadoop supergroup      74105 2025-05-06 15:11 /ncdc_input/1902.gz
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop-book/input/ncdc/all$ hadoop jar AverageTemp.jar temp.AverageDriver /ncdc_input /output_avg
JAR does not exist or is not a normal file: /home/hadoop/hadoop-book/input/ncdc/all/AverageTemp.jar
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop-book/input/ncdc/all$ cd ~/Desktop/
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop jar AverageTemp.jar temp.AverageDriver /ncdc_input /output_avg
JAR does not exist or is not a normal file: /home/hadoop/Desktop/AverageTemp.jar
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ ls
ACFrOgC1t8B0DXZ9jzYbibLxQB_J-Myit4QJozqp_tewZO5d859eh0iEqf_um0mI-d7nXMSb849Mx4MgJzFKzS6Ak8WTKMyxCYNBKewr7yOGFiLI7diBJC8Kkdjx9GAsRZJrJpzGQXyTrpgSUCSU.pdf
file1.txt
NCDCWeather
NCDCWeather.jar
sample.txt
Tejaswini.jar
WordCount170.jar
WordCount.jar
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop jar NCDCWeather.jar temp.AverageDriver /ncdc_input /output_avg
2025-05-06 15:14:57,391 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-06 15:14:57,428 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-06 15:14:57,429 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-06 15:14:57,485 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-06 15:14:57,537 INFO input.FileInputFormat: Total input files to process : 2
2025-05-06 15:14:57,553 INFO mapreduce.JobSubmitter: number of splits:2
2025-05-06 15:14:57,607 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1944593155_0001
2025-05-06 15:14:57,607 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-06 15:14:57,656 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-06 15:14:57,657 INFO mapreduce.Job: Running job: job_local1944593155_0001
2025-05-06 15:14:57,657 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-06 15:14:57,661 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:14:57,661 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
```

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop jar NCDCWeather.jar temp.AverageDriver /ncdc_input /output_avg
2025-05-06 15:14:57,391 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-06 15:14:57,428 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-06 15:14:57,429 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-06 15:14:57,485 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-06 15:14:57,537 INFO input.FileInputFormat: Total input files to process : 2
2025-05-06 15:14:57,553 INFO mapreduce.JobSubmitter: number of splits:2
2025-05-06 15:14:57,607 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1944593155_0001
2025-05-06 15:14:57,607 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-06 15:14:57,656 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-06 15:14:57,657 INFO mapreduce.Job: Running job: job_local1944593155_0001
2025-05-06 15:14:57,657 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-06 15:14:57,661 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:14:57,661 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-06 15:14:57,661 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-05-06 15:14:57,692 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-06 15:14:57,693 INFO mapred.LocalJobRunner: Starting task: attempt_local1944593155_0001_m_000000_0
2025-05-06 15:14:57,701 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:14:57,701 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-06 15:14:57,707 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-06 15:14:57,708 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/ncdc_input/1902.gz:0+74105
2025-05-06 15:14:57,739 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-06 15:14:57,739 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-06 15:14:57,739 INFO mapred.MapTask: soft limit at 83886080
2025-05-06 15:14:57,739 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-06 15:14:57,739 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-06 15:14:57,741 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-06 15:14:57,754 INFO zlib.ZlibFactory: Successfully loaded & initialized native-zlib library
2025-05-06 15:14:57,754 INFO compress.CodecPool: Got brand-new decompressor [.gz]
2025-05-06 15:14:57,815 INFO mapred.LocalJobRunner:
2025-05-06 15:14:57,816 INFO mapred.MapTask: Starting flush of map output
2025-05-06 15:14:57,816 INFO mapred.MapTask: Spilling map output
2025-05-06 15:14:57,816 INFO mapred.MapTask: bufstart = 0; bufend = 59085; bufvoid = 104857600
2025-05-06 15:14:57,816 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188140(104752560); length = 26257/6553600
2025-05-06 15:14:57,822 INFO mapred.MapTask: Finished spill 0
2025-05-06 15:14:57,826 INFO mapred.Task: Task:attempt_local1944593155_0001_m_000000_0 is done. And is in the process of committing
2025-05-06 15:14:57,828 INFO mapred.LocalJobRunner: map
2025-05-06 15:14:57,829 INFO mapred.Task: Task 'attempt_local1944593155_0001_m_000000_0' done.
2025-05-06 15:14:57,831 INFO mapred.Task: Final Counters for attempt_local1944593155_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=7873
                FILE: Number of bytes written=719826
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=74105
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=6565
                Map output records=6565
```

```
	Map-Reduce Framework
		Map input records=6565
		Map output records=6565
		Map output bytes=59085
		Map output materialized bytes=72221
		Input split bytes=105
		Combine input records=0
		Spilled Records=6565
		Failed Shuffles=0
		Merged Map outputs=0
		GC time elapsed (ms)=0
		Total committed heap usage (bytes)=526385152
	File Input Format Counters
		Bytes Read=74105
025-05-06 15:14:57,831 INFO mapred.LocalJobRunner: Finishing task: attempt_local1944593155_0001_m_000000_0
025-05-06 15:14:57,831 INFO mapred.LocalJobRunner: Starting task: attempt_local1944593155_0001_m_000001_0
025-05-06 15:14:57,832 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
025-05-06 15:14:57,832 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
025-05-06 15:14:57,832 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
025-05-06 15:14:57,833 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/ncdc_input/1901.gz:0+73867
025-05-06 15:14:57,859 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
025-05-06 15:14:57,859 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
025-05-06 15:14:57,859 INFO mapred.MapTask: soft limit at 83886080
025-05-06 15:14:57,859 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
025-05-06 15:14:57,859 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
 25-05-06 15:14:57,860 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
025-05-06 15:14:57,887 INFO mapred.LocalJobRunner:
025-05-06 15:14:57,887 INFO mapred.MapTask: Starting flush of map output
025-05-06 15:14:57,887 INFO mapred.MapTask: Spilling map output
025-05-06 15:14:57,887 INFO mapred.MapTask: bufstart = 0; bufend = 59076; bufvoid = 104857600
025-05-06 15:14:57,887 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188144(104752576); length = 26253/6553600
025-05-06 15:14:57,890 INFO mapred.MapTask: Finished spill 0
025-05-06 15:14:57,892 INFO mapred.Task: Task:attempt_local1944593155_0001_m_000001_0 is done. And is in the process of committing
025-05-06 15:14:57,894 INFO mapred.LocalJobRunner: map
025-05-06 15:14:57,894 INFO mapred.Task: Task 'attempt_local1944593155_0001_m_000001_0' done.
025-05-06 15:14:57,895 INFO mapred.Task: Final Counters for attempt_local1944593155_0001_m_000001_0: Counters: 23
	File System Counters
		FILE: Number of bytes read=8102
		FILE: Number of bytes written=792068
		FILE: Number of read operations=0
		FILE: Number of large read operations=0
		FILE: Number of write operations=0
		HDFS: Number of bytes read=147972
		HDFS: Number of bytes written=0
		HDFS: Number of read operations=7
		HDFS: Number of large read operations=0
		HDFS: Number of write operations=1
		HDFS: Number of bytes read erasure-coded=0
	Map-Reduce Framework
		Map input records=6565
		Map output records=6564
		Map output bytes=59076
		Map output materialized bytes=72210
		Input split bytes=105
		Combine input records=0
```

```
		Input split bytes=105
		Combine input records=0
		Spilled Records=6564
		Failed Shuffles=0
		Merged Map outputs=0
		GC time elapsed (ms)=0
		Total committed heap usage (bytes)=526385152
	File Input Format Counters
		Bytes Read=73867
2025-05-06 15:14:57,895 INFO mapred.LocalJobRunner: Finishing task: attempt_local1944593155_0001_m_000001_0
2025-05-06 15:14:57,895 INFO mapred.LocalJobRunner: map task executor complete.
2025-05-06 15:14:57,897 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-05-06 15:14:57,897 INFO mapred.LocalJobRunner: Starting task: attempt_local1944593155_0001_r_000000_0
2025-05-06 15:14:57,904 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:14:57,904 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-06 15:14:57,904 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-06 15:14:57,905 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@877f0e5
2025-05-06 15:14:57,906 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-06 15:14:57,915 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5827985408, maxSingleShuffleLimit=1456996352, mergeThreshold=3846470400, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-06 15:14:57,916 INFO reduce.EventFetcher: attempt_local1944593155_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-05-06 15:14:57,927 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1944593155_0001_m_000001_0 decomp: 72206 len: 72210 to MEMORY
2025-05-06 15:14:57,928 INFO reduce.InMemoryMapOutput: Read 72206 bytes from map-output for attempt_local1944593155_0001_m_000001_0
2025-05-06 15:14:57,928 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 72206, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->72206
2025-05-06 15:14:57,929 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1944593155_0001_m_000000_0 decomp: 72217 len: 72221 to MEMORY
`025-05-06 15:14:57,930 INFO reduce.InMemoryMapOutput: Read 72217 bytes from map-output for attempt_local1944593155_0001_m_000000_0
 25-05-06 15:14:57,930 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 72217, inMemoryMapOutputs.size() -> 2, commitMemory -> 72206, usedMemory ->144423
 25-05-06 15:14:57,930 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
_025-05-06 15:14:57,930 WARN io.ReadaheadPool: Failed readahead on ifile
EBADF: Bad file descriptor
	at org.apache.hadoop.io.nativeio.NativeIOSPOSIX.posix_fadvise(Native Method)
	at org.apache.hadoop.io.nativeio.NativeIOSPOSIX.posixFadviseIfPossible(NativeIO.java:419)
	at org.apache.hadoop.io.nativeio.NativeIOSPOSIX$CacheManipulator.posixFadviseIfPossible(NativeIO.java:296)
	at org.apache.hadoop.io.ReadaheadPool$ReadaheadRequestImpl.run(ReadaheadPool.java:220)
	at java.base/java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1128)
	at java.base/java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:628)
	at java.base/java.lang.Thread.run(Thread.java:829)
2025-05-06 15:14:57,930 INFO mapred.LocalJobRunner: 2 / 2 copied.
2025-05-06 15:14:57,931 INFO reduce.MergeManagerImpl: finalMerge called with 2 in-memory map-outputs and 0 on-disk map-outputs
2025-05-06 15:14:57,933 INFO mapred.Merger: Merging 2 sorted segments
2025-05-06 15:14:57,933 INFO mapred.Merger: Down to the last merge-pass, with 2 segments left of total size: 144409 bytes
2025-05-06 15:14:57,938 INFO reduce.MergeManagerImpl: Merged 2 segments, 144423 bytes to disk to satisfy reduce memory limit
2025-05-06 15:14:57,938 INFO reduce.MergeManagerImpl: Merging 1 files, 144425 bytes from disk
2025-05-06 15:14:57,938 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2025-05-06 15:14:57,938 INFO mapred.Merger: Merging 1 sorted segments
2025-05-06 15:14:57,939 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 144414 bytes
2025-05-06 15:14:57,939 INFO mapred.LocalJobRunner: 2 / 2 copied.
2025-05-06 15:14:57,968 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2025-05-06 15:14:58,022 INFO mapred.Task: Task:attempt_local1944593155_0001_r_000000_0 is done. And is in the process of committing
2025-05-06 15:14:58,023 INFO mapred.LocalJobRunner: 2 / 2 copied.
2025-05-06 15:14:58,023 INFO mapred.Task: Task attempt_local1944593155_0001_r_000000_0 is allowed to commit now
2025-05-06 15:14:58,038 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1944593155_0001_r_000000_0' to hdfs://localhost:9000/output_avg
2025-05-06 15:14:58,038 INFO mapred.LocalJobRunner: reduce > reduce
2025-05-06 15:14:58,038 INFO mapred.Task: Task 'attempt_local1944593155_0001_r_000000_0' done.
2025-05-06 15:14:58,038 INFO mapred.Task: Final Counters for attempt_local1944593155_0001_r_000000_0: Counters: 30
	File System Counters
```

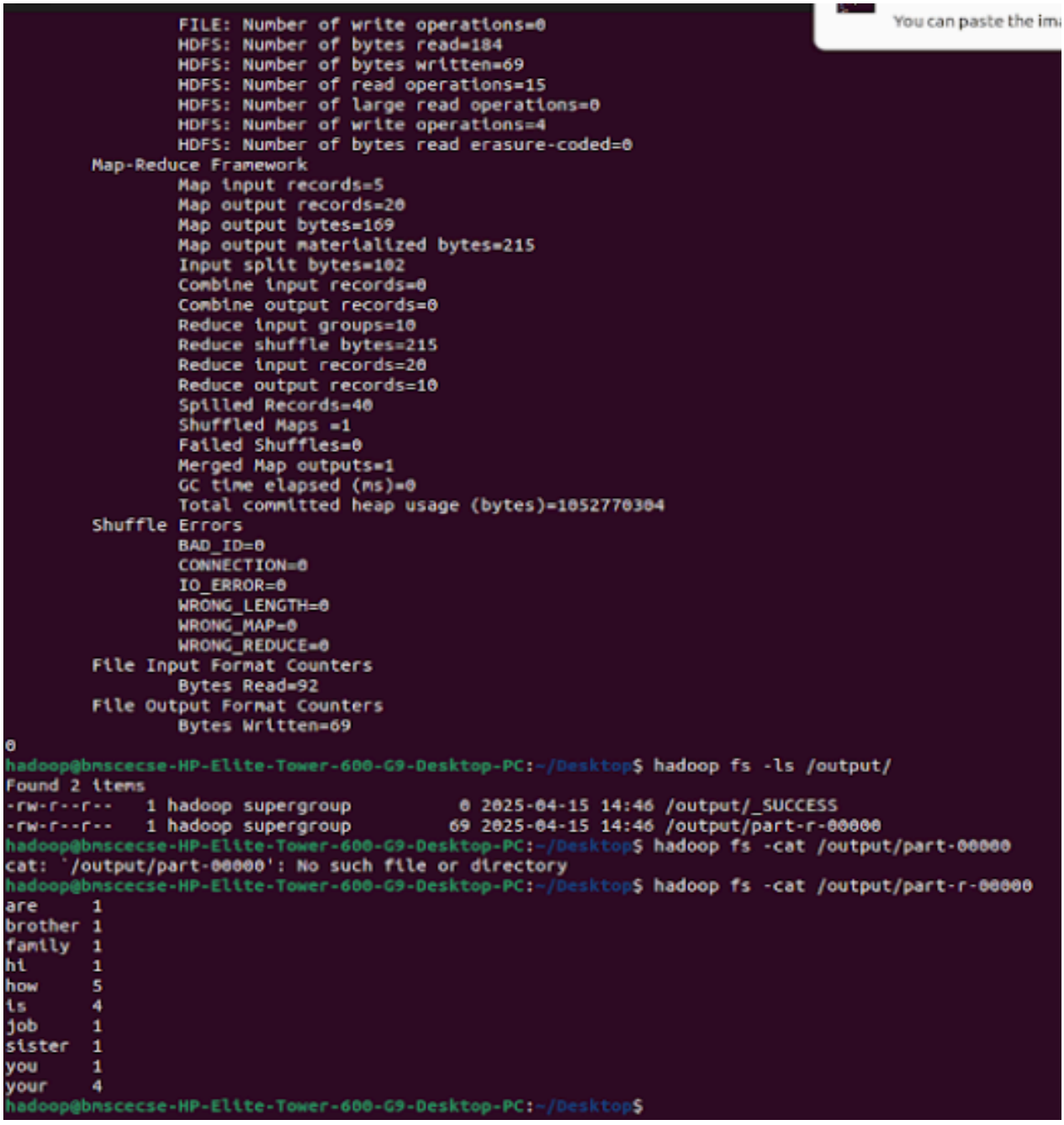
```
2025-05-06 15:14:58,038 INFO mapred.Task: Final Counters for attempt_local1944593155_0001_r_000000_0: Counters: 3
        File System Counters
                FILE: Number of bytes read=297022
                FILE: Number of bytes written=936493
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=147972
                HDFS: Number of bytes written=16
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=2
                Reduce shuffle bytes=144431
                Reduce input records=13129
                Reduce output records=2
                Spilled Records=13129
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=2
                Total committed heap usage (bytes)=633339904
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=16
2025-05-06 15:14:58,038 INFO mapred.LocalJobRunner: Finishing task: attempt_local1944593155_0001_r_000000_0
2025-05-06 15:14:58,039 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-06 15:14:58,659 INFO mapreduce.Job: Job job_local1944593155_0001 running in uber mode : false
2025-05-06 15:14:58,659 INFO mapreduce.Job:  map 100% reduce 100%
2025-05-06 15:14:58,660 INFO mapreduce.Job: Job job_local1944593155_0001 completed successfully
2025-05-06 15:14:58,664 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=312997
                FILE: Number of bytes written=2448387
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=370049
                HDFS: Number of bytes written=16
                HDFS: Number of read operations=24
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=5
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=13130
```

```
                Reduce output records=12
                Spilled Records=26258
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=3
                Total committed heap usage (bytes)=1686110208
        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=147972
        File Output Format Counters
                Bytes Written=72
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -ls /output_avg
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2025-05-06 15:14 /output_avg/_SUCCESS
-rw-r--r--   1 hadoop supergroup         16 2025-05-06 15:14 /output_avg/part-r-00000
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -cat /output_avg/part-r-00000
1901    46
1902    21
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -ls /output_avg
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2025-05-06 15:14 /output_avg/_SUCCESS
-rw-r--r--   1 hadoop supergroup         16 2025-05-06 15:14 /output_avg/part-r-00000
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -cat /output_avg/part-r-00000
1901    46
1902    21
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -cat /output_avg/part-r-00000
1901    46
1902    21
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -cat /output_mean/part-r-00000
01      4
02      1
03      6
04      34
05      89
06      143
07      182
08      172
09      123
10      73
11      21
12      3
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -cat /output_avg/part-r-00000
1901    46
1902    21
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -cat /output_avg/part-r-00000
1901    46
1902    21
```

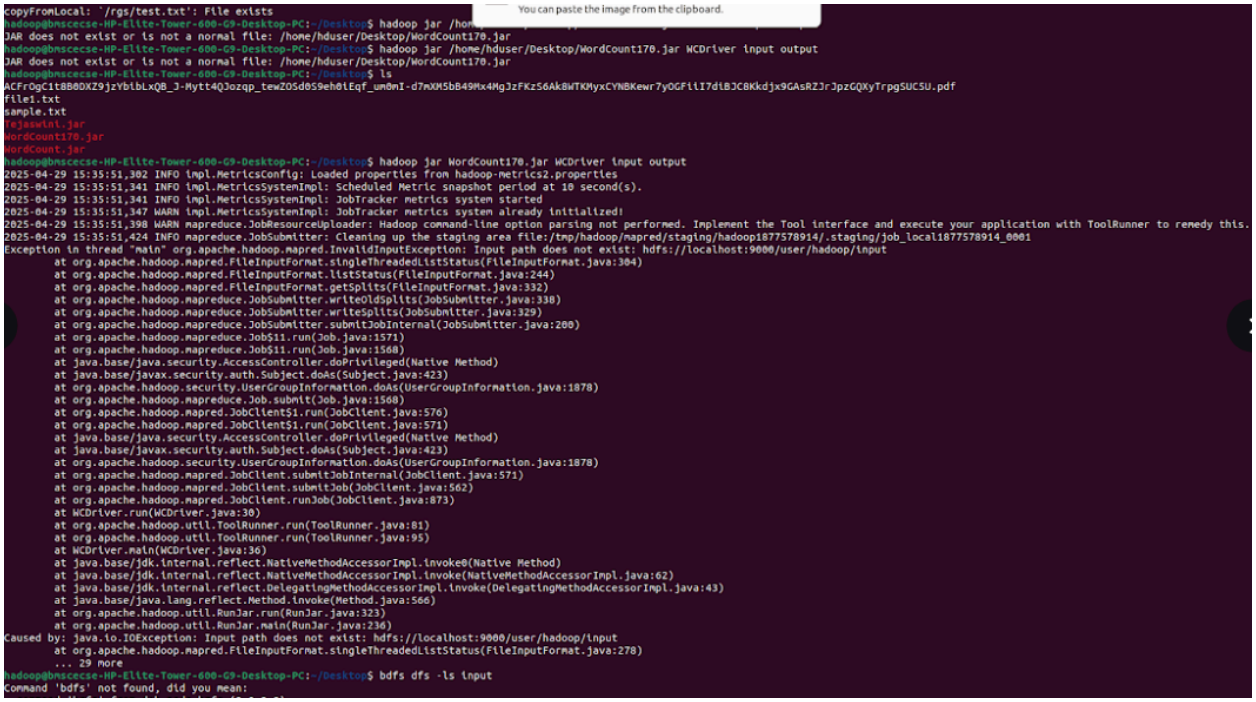
LAB PROGRAM - 07

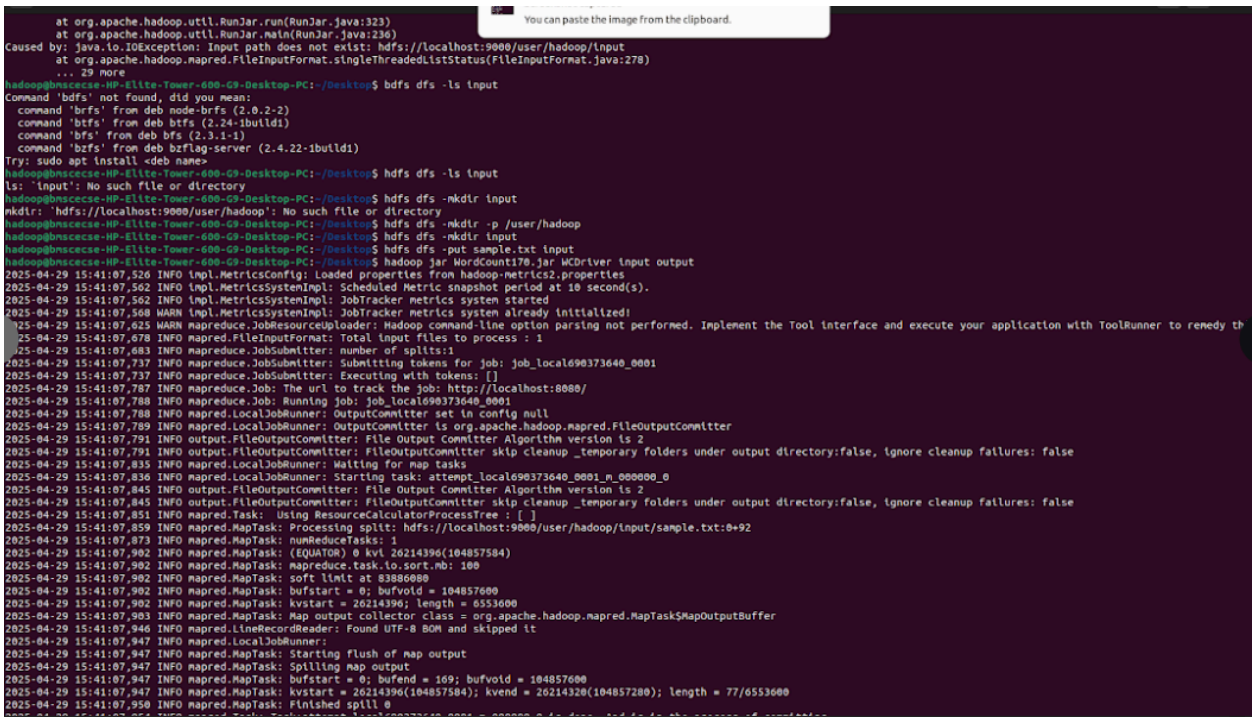
For a given Text file, Create a Map Reduce program to sort the content in an alphabetic order listing only top 10 maximum occurrences of words.

Code with Output:



```
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=10
                Reduce shuffle bytes=215
                Reduce input records=20
                Reduce output records=10
                Spilled Records=20
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=526385152
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=69
25-04-29 15:41:08,074 INFO mapred.LocalJobRunner: Finishing task: attempt_local690373640_0001_r_000000_0
025-04-29 15:41:08,074 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-04-29 15:41:08,791 INFO mapreduce.Job: Job job_local690373640_0001 running in uber mode : false
2025-04-29 15:41:08,791 INFO mapreduce.Job:  map 100% reduce 100%
2025-04-29 15:41:08,792 INFO mapreduce.Job: Job job_local690373640_0001 completed successfully
2025-04-29 15:41:08,796 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=8836
                FILE: Number of bytes written=1283959
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=184
                HDFS: Number of bytes written=69
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=5
                Map output records=20
                Map output bytes=169
                Map output materialized bytes=215
                Input split bytes=102
                Combine input records=0
                Combine output records=0
                Reduce input groups=10
                Reduce shuffle bytes=215
                Reduce input records=20
                Reduce output records=10
```


You can paste the image from the clipboard.
copyFromLocal: `/rgs/test.txt': File exists
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop jar /hom
JAR does not exist or is not a normal file: /home/hduser/Desktop/WordCount170.jar
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop jar /home/hduser/Desktop/WordCount170.jar WCDriver input output
JAR does not exist or is not a normal file: /home/hduser/Desktop/WordCount170.jar
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ ls
ACFrOgC1t8B0DXZ9jzYbtbLxQB_J-Mytt4QJozqp_tewZOSd059eh8tEqf_um0mI-d7mXM5bB49Mx4Mg3zFKzS6Ak8WTKMyxCYNBKewr7yOGFliI7diBJCBKkdjx9GAsRZJrJpzGQXyTrpgSUC5U.pdf
file1.txt
sample.txt
Tejaswini.jar
WordCount170.jar
WordCount.jar
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop jar WordCount170.jar WCDriver input output
2025-04-29 15:35:51,302 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-04-29 15:35:51,341 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-04-29 15:35:51,341 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-04-29 15:35:51,347 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-29 15:35:51,398 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-04-29 15:35:51,424 INFO mapreduce.JobSubmitter: Cleaning up the staging area file:/tmp/hadoop/mapred/staging/hadoop1877578914/.staging/job_local1877578914_0001
Exception in thread "main" org.apache.hadoop.mapred.InvalidInputException: Input path does not exist: hdfs://localhost:9000/user/hadoop/input
at org.apache.hadoop.mapred.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:304)
at org.apache.hadoop.mapred.FileInputFormat.listStatus(FileInputFormat.java:244)
at org.apache.hadoop.mapred.FileInputFormat.getSplits(FileInputFormat.java:332)
at org.apache.hadoop.mapreduce.JobSubmitter.writeOldSplits(JobSubmitter.java:338)
at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java:329)
at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:200)
at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1571)
at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1568)
at java.base/java.security.AccessController.doPrivileged(Native Method)
at java.base/javax.security.auth.Subject.doAs(Subject.java:423)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1878)
at org.apache.hadoop.mapreduce.Job.submit(Job.java:1568)
at org.apache.hadoop.mapred.JobClient$1.run(JobClient.java:576)
at org.apache.hadoop.mapred.JobClient$1.run(JobClient.java:571)
at java.base/java.security.AccessController.doPrivileged(Native Method)
at java.base/javax.security.auth.Subject.doAs(Subject.java:423)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1878)
at org.apache.hadoop.mapred.JobClient.submitJobInternal(JobClient.java:571)
at org.apache.hadoop.mapred.JobClient.submitJob(JobClient.java:562)
at org.apache.hadoop.mapred.JobClient.runJob(JobClient.java:873)
at WCDriver.run(WCDriver.java:30)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:81)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:95)
at WCDriver.main(WCDriver.java:36)
at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at java.base/jdk.internal.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.base/java.lang.reflect.Method.invoke(Method.java:566)
at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
Caused by: java.io.IOException: Input path does not exist: hdfs://localhost:9000/user/hadoop/input
at org.apache.hadoop.mapred.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:278)
... 29 more
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ bdfs dfs -ls input
Command 'bdfs' not found, did you mean:


You can paste the image from the clipboard.
at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
Caused by: java.io.IOException: Input path does not exist: hdfs://localhost:9000/user/hadoop/input
at org.apache.hadoop.mapred.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:278)
... 29 more
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ bdfs dfs -ls input
Command 'bdfs' not found, did you mean:
command 'brfs' from deb node-brfs (2.0.2-2)
command 'btfs' from deb btfs (2.24-1build1)
command 'bfs' from deb bfs (2.3.1-1)
command 'bzfs' from deb bzflag-server (2.4.22-1build1)
Try: sudo apt install <deb name>
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -ls input
ls: `input': No such file or directory
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -mkdir input
mkdir: `hdfs://localhost:9000/user/hadoop': No such file or directory
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -mkdir -p /user/hadoop
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -mkdir input
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hdfs dfs -put sample.txt input
hadoop@bnscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop jar WordCount170.jar WCDriver input output
2025-04-29 15:41:07,526 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-04-29 15:41:07,562 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-04-29 15:41:07,562 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-04-29 15:41:07,568 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-29 15:41:07,625 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy th
2025-04-29 15:41:07,678 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-29 15:41:07,683 INFO mapreduce.JobSubmitter: number of splits:1
2025-04-29 15:41:07,737 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local690373640_0001
2025-04-29 15:41:07,737 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-29 15:41:07,787 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-04-29 15:41:07,788 INFO mapreduce.Job: Running job: job_local690373640_0001
2025-04-29 15:41:07,788 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-04-29 15:41:07,789 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2025-04-29 15:41:07,791 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-04-29 15:41:07,791 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-04-29 15:41:07,835 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-04-29 15:41:07,836 INFO mapred.LocalJobRunner: Starting task: attempt_local690373640_0001_m_000000_0
2025-04-29 15:41:07,845 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-04-29 15:41:07,845 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-04-29 15:41:07,851 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-04-29 15:41:07,859 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hadoop/input/sample.txt:0+92
2025-04-29 15:41:07,873 INFO mapred.MapTask: numReduceTasks: 1
2025-04-29 15:41:07,902 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-04-29 15:41:07,902 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-04-29 15:41:07,902 INFO mapred.MapTask: soft limit at 83886080
2025-04-29 15:41:07,902 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-04-29 15:41:07,902 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-04-29 15:41:07,903 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-04-29 15:41:07,946 INFO mapred.LineRecordReader: Found UTF-8 BOM and skipped it
2025-04-29 15:41:07,947 INFO mapred.LocalJobRunner:
2025-04-29 15:41:07,947 INFO mapred.MapTask: Starting flush of map output
2025-04-29 15:41:07,947 INFO mapred.MapTask: Spilling map output
2025-04-29 15:41:07,947 INFO mapred.MapTask: bufstart = 0; bufend = 169; bufvoid = 104857600
2025-04-29 15:41:07,947 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214320(104857280); length = 77/6553600
2025-04-29 15:41:07,950 INFO mapred.MapTask: Finished spill 0

```
Starting nodemanagers
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop/etc/hadoop$ cd ~
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 15897.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 16080.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-600-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: secondarynamenode is running as process 16370.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
resourcemanager is running as process 16661.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 16829.  Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ jps
8256 org.eclipse.equinox.launcher_1.6.1000.v20250227-1734.jar
16080 DataNode
16370 SecondaryNameNode
16661 ResourceManager
15897 NameNode
16829 NodeManager
18399 Jps
 \doop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ ls
  sktop              hadoopdata          Public       wc1.jar
 ocuments            hs_err_pid4959.log  salpoorvika  wc.jar
ownloads             hs_err_pid6247.log  shara        WordCountClasses
clipse-workspace     hs_err_pid7049.log  snap         WordCountProject
hadoop               Music               Templates
hadoop-3.3.4.tar.gz  Pictures            Videos
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cd Desktop
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ nano sample.txt
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hadoop supergroup          0 2024-05-14 15:15 /FF
drwxr-xr-x   - hadoop supergroup          0 2024-05-14 14:58 /FFF
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:46 /output
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 14:38 /rgs
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop fs -mkdir /rgs
mkdir: `/rgs': File exists
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop fs -copyFromLocal D:/sample.txt /rgs/test.txt
copyFromLocal: `/rgs/test.txt': File exists
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop jar /home/hduser/Desktop/WordCount170.jar WCDriver input output
JAR does not exist or is not a normal file: /home/hduser/Desktop/WordCount170.jar
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop jar /home/hduser/Desktop/WordCount170.jar WCDriver input output
JAR does not exist or is not a normal file: /home/hduser/Desktop/WordCount170.jar
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ ls
ACFrOgC1t8B0DXZ9jzYbibLxQB_J-Mytt4QJozqp_tewZOSd0S9eh0iEqf_um0mI-d7mXM5bB49Mx4MgJzFKzS6Ak8WTKMyxCYNBKewr7yOGFiiI7diBJC8Kkdjx9GAsRZJrJpzGQXyTrpgSUC5U.pdf
file1.txt
sample.txt
Tejaswini.jar
WordCount170.jar
WordCount.jar
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/Desktop$ hadoop jar WordCount170.jar WCDriver input output
```

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ ls
Desktop              hadoopdata          Public       wc1.jar
Documents            hs_err_pid4959.log  salpoorvika  wc.jar
Downloads            hs_err_pid6247.log  shara        WordCountClasses
eclipse-workspace    hs_err_pid7049.log  snap         WordCountProject
hadoop               Music               Templates
hadoop-3.3.4.tar.gz  Pictures            Videos
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cd hadoop
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop$ ls
bin      lib          licenses-binary  Merge.txt     README.txt  WC.txt
etc      libexec      LICENSE.txt      NOTICE-binary  sbin       WWC.txt
include  LICENSE-binary  logs         NOTICE.txt    share
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop$ cd etc
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop/etc$ ls
hadoop
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop/etc$ cd hadoop
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop/etc/hadoop$ ls
capacity-scheduler.xml            kms-log4j.properties
configuration.xsl                 kms-site.xml
container-executor.cfg            log4j.properties
core-site.xml                     mapred-env.cmd
hadoop-env.cmd                    mapred-env.sh
hadoop-env.sh                     mapred-queues.xml.template
hadoop-metrics2.properties        mapred-site.xml
 doop-policy.xml                  shellprofile.d
  doop-user-functions.sh.example  ssl-client.xml.example
 fs-rbf-site.xml                  ssl-server.xml.example
ndfs-site.xml                     user_ec_policies.xml.template
httpfs-env.sh                     workers
httpfs-log4j.properties           yarn-env.cmd
httpfs-site.xml                   yarn-env.sh
kms-acls.xml                      yarnservice-log4j.properties
kms-env.sh                        yarn-site.xml
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop/etc/hadoop$ nano mapred-site.xml
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop/etc/hadoop$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bmscecse-HP-Elite-Tower-600-G9-Desktop-PC]
Starting resourcemanager
Starting nodemanagers
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/hadoop/etc/hadoop$ cd ~
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 15897.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 16080.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-600-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: secondarynamenode is running as process 16370.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
```

LAB PROGRAM - 08

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

Code with Output:

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

scala> object PrintNumbers {
     |   def main(args: Array[String]): Unit = {
     |     for (i <- 1 to 100) {
     |       println(i)
     |     }
     |   }
     | }
defined object PrintNumbers

scala> PrintNumbers
res0: PrintNumbers.type = PrintNumbers$@7157413e

scala> PrintNumbers.main(Array())
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

LAB PROGRAM - 09

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

Code with Output:

```
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ echo -e "hello world hello\nspark is awesome\nthis is a test spark world\nhello spark w
orld\nhello" > sample.txt
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ nano word_count.py
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ spark-submit word_count.py
25/05/20 11:39:52 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicabl
e
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
25/05/20 11:39:52 INFO SparkContext: Running Spark version 3.0.3
25/05/20 11:39:52 INFO ResourceUtils: ==============================================================
25/05/20 11:39:52 INFO ResourceUtils: Resources for spark.driver:

25/05/20 11:39:52 INFO ResourceUtils: ==============================================================
25/05/20 11:39:52 INFO SparkContext: Submitted application: SimpleWordCount
25/05/20 11:39:53 INFO SecurityManager: Changing view acls to: bmscecse
25/05/20 11:39:53 INFO SecurityManager: Changing modify acls to: bmscecse
25/05/20 11:39:53 INFO SecurityManager: Changing view acls groups to:
25/05/20 11:39:53 INFO SecurityManager: Changing modify acls groups to:
25/05/20 11:39:53 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(bmscecs
e); groups with view permissions: Set(); users  with modify permissions: Set(bmscecse); groups with modify permissions: Set()
25/05/20 11:39:53 INFO Utils: Successfully started service 'sparkDriver' on port 34445.
25/05/20 11:39:53 INFO SparkEnv: Registering MapOutputTracker
25/05/20 11:39:53 INFO SparkEnv: Registering BlockManagerMaster
25/05/20 11:39:53 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
25/05/20 11:39:53 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
25/05/20 11:39:53 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
25/05/20 11:39:53 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-ede53b67-2dce-41cc-ae20-7cc9c45cde11
25/05/20 11:39:53 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
25/05/20 11:39:53 INFO SparkEnv: Registering OutputCommitCoordinator
25/05/20 11:39:53 INFO Utils: Successfully started service 'SparkUI' on port 4040.
25/05/20 11:39:53 INFO SparkUI: Bound SparkUI to 127.0.0.1, and started at http://localhost:4040
25/05/20 11:39:53 INFO Executor: Starting executor ID driver on host localhost
25/05/20 11:39:53 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 34187.
25/05/20 11:39:53 INFO NettyBlockTransferService: Server created on localhost:34187
25/05/20 11:39:53 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
25/05/20 11:39:53 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, localhost, 34187, None)
25/05/20 11:39:53 INFO BlockManagerMasterEndpoint: Registering block manager localhost:34187 with 366.3 MiB RAM, BlockManagerId(driver, local
host, 34187, None)
```

```
25/05/20 11:39:53 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, localhost, 34187, None)
25/05/20 11:39:53 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, localhost, 34187, None)
25/05/20 11:39:53 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 241.7 KiB, free 366.1 MiB)
25/05/20 11:39:53 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 23.4 KiB, free 366.0 MiB)
25/05/20 11:39:53 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:34187 (size: 23.4 KiB, free: 366.3 MiB)
25/05/20 11:39:53 INFO SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:0
25/05/20 11:39:54 INFO FileInputFormat: Total input paths to process : 1
25/05/20 11:39:54 INFO SparkContext: Starting job: collect at /home/bmscecse/word_count.py:16
25/05/20 11:39:54 INFO DAGScheduler: Registering RDD 3 (reduceByKey at /home/bmscecse/word_count.py:12) as input to shuffle 0
25/05/20 11:39:54 INFO DAGScheduler: Got job 0 (collect at /home/bmscecse/word_count.py:16) with 1 output partitions
25/05/20 11:39:54 INFO DAGScheduler: Final stage: ResultStage 1 (collect at /home/bmscecse/word_count.py:16)
25/05/20 11:39:54 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)
25/05/20 11:39:54 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 0)
25/05/20 11:39:54 INFO DAGScheduler: Submitting ShuffleMapStage 0 (PairwiseRDD[3] at reduceByKey at /home/bmscecse/word_count.py:12), which h
as no missing parents
25/05/20 11:39:54 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 11.5 KiB, free 366.0 MiB)
25/05/20 11:39:54 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 7.0 KiB, free 366.0 MiB)
25/05/20 11:39:54 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:34187 (size: 7.0 KiB, free: 366.3 MiB)
25/05/20 11:39:54 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1223
25/05/20 11:39:54 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (PairwiseRDD[3] at reduceByKey at /home/bmscecse/word_
count.py:12) (first 15 tasks are for partitions Vector(0))
25/05/20 11:39:54 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
25/05/20 11:39:54 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, executor driver, partition 0, PROCESS_LOCAL, 7360 by
tes)
25/05/20 11:39:54 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
25/05/20 11:39:54 INFO HadoopRDD: Input split: file:/home/bmscecse/sample.txt:0+86
25/05/20 11:39:54 INFO PythonRunner: Times: total = 212, boot = 166, init = 46, finish = 0
25/05/20 11:39:55 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1803 bytes result sent to driver
25/05/20 11:39:55 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 837 ms on localhost (executor driver) (1/1)
25/05/20 11:39:55 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
25/05/20 11:39:55 INFO PythonAccumulatorV2: Connected to AccumulatorServer at host: 127.0.0.1 port: 58787
25/05/20 11:39:55 INFO DAGScheduler: ShuffleMapStage 0 (reduceByKey at /home/bmscecse/word_count.py:12) finished in 0.905 s
25/05/20 11:39:55 INFO DAGScheduler: looking for newly runnable stages
25/05/20 11:39:55 INFO DAGScheduler: running: Set()
25/05/20 11:39:55 INFO DAGScheduler: waiting: Set(ResultStage 1)
25/05/20 11:39:55 INFO DAGScheduler: failed: Set()
25/05/20 11:39:55 INFO DAGScheduler: Submitting ResultStage 1 (PythonRDD[6] at collect at /home/bmscecse/word_count.py:16), which has no miss
```

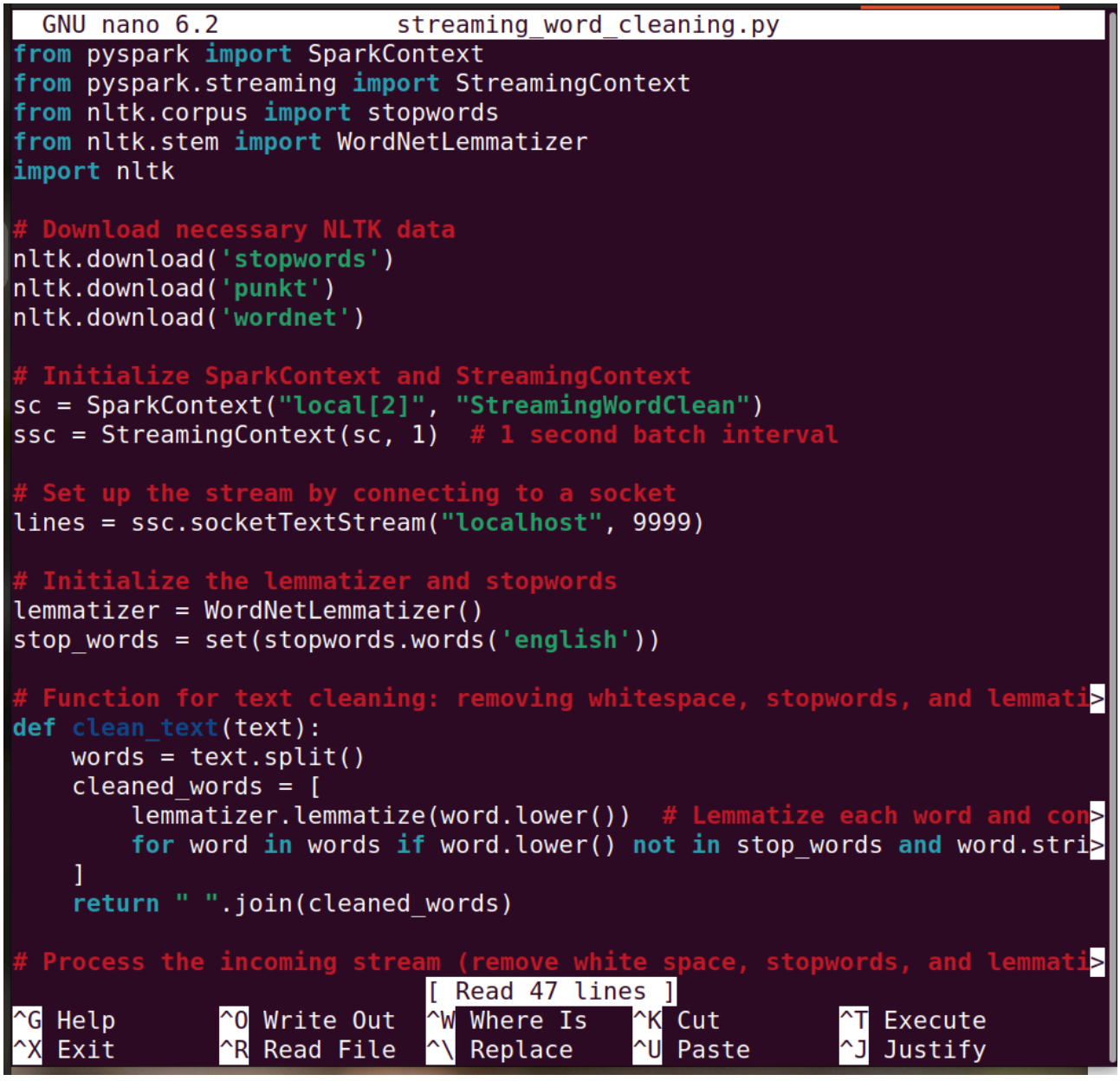
```
25/05/20 11:39:55 INFO DAGScheduler: Submitting ResultStage 1 (PythonRDD[6] at collect at /home/bmscecse/word_count.py:16), which has no miss
ing parents
25/05/20 11:39:55 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 8.7 KiB, free 366.0 MiB)
25/05/20 11:39:55 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 5.2 KiB, free 366.0 MiB)
25/05/20 11:39:55 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:34187 (size: 5.2 KiB, free: 366.3 MiB)
25/05/20 11:39:55 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1223
25/05/20 11:39:55 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (PythonRDD[6] at collect at /home/bmscecse/word_count.py:1
6) (first 15 tasks are for partitions Vector(0))
25/05/20 11:39:55 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
25/05/20 11:39:55 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, NODE_LOCAL, 7143 bytes
)
25/05/20 11:39:55 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
25/05/20 11:39:55 INFO ShuffleBlockFetcherIterator: Getting 1 (156.0 B) non-empty blocks including 1 (156.0 B) local and 0 (0.0 B) host-local
 and 0 (0.0 B) remote blocks
25/05/20 11:39:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 3 ms
25/05/20 11:39:55 INFO PythonRunner: Times: total = 42, boot = -518, init = 560, finish = 0
25/05/20 11:39:55 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1767 bytes result sent to driver
25/05/20 11:39:55 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 67 ms on localhost (executor driver) (1/1)
25/05/20 11:39:55 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
25/05/20 11:39:55 INFO DAGScheduler: ResultStage 1 (collect at /home/bmscecse/word_count.py:16) finished in 0.073 s
25/05/20 11:39:55 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
25/05/20 11:39:55 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
25/05/20 11:39:55 INFO DAGScheduler: Job 0 finished: collect at /home/bmscecse/word_count.py:16, took 1.006372 s
25/05/20 11:39:55 INFO SparkUI: Stopped Spark web UI at http://localhost:4040
25/05/20 11:39:55 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/05/20 11:39:55 INFO MemoryStore: MemoryStore cleared
25/05/20 11:39:55 INFO BlockManager: BlockManager stopped
25/05/20 11:39:55 INFO BlockManagerMaster: BlockManagerMaster stopped
25/05/20 11:39:55 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/05/20 11:39:55 INFO SparkContext: Successfully stopped SparkContext
25/05/20 11:39:56 INFO ShutdownHookManager: Shutdown hook called
25/05/20 11:39:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-aa3fe98f-0602-4d46-9aa7-736cc067dbc4
25/05/20 11:39:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-46383acd-daf9-43ba-9b6c-337b1f545cc0
25/05/20 11:39:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-46383acd-daf9-43ba-9b6c-337b1f545cc0/pyspark-bab4954c-8903-4094-94e
c-b517c5aef7db
```

```
  GNU nano 6.2                        word_count.py
from pyspark import SparkContext

# Initialize SparkContext
sc = SparkContext("local", "SimpleWordCount")

# Read the file (use the temporary sample.txt)
rdd = sc.textFile("sample.txt")  # Use the sample file created above

# Count words
counts = (rdd.flatMap(lambda line: line.split())          # Split each li>
             .map(lambda word: (word, 1))                # Map each word >
             .reduceByKey(lambda a, b: a + b)           # Reduce by key t>
             .filter(lambda x: x[1] > 4))               # Filter words wi>

# Show result
for word, count in counts.collect():
    print(word, count)

# Stop the SparkContext
sc.stop()









                          [ Read 20 lines ]
^G Help       ^O Write Out  ^W Where Is   ^K Cut        ^T Execute
^X Exit       ^R Read File  ^\ Replace    ^U Paste      ^J Justify
```

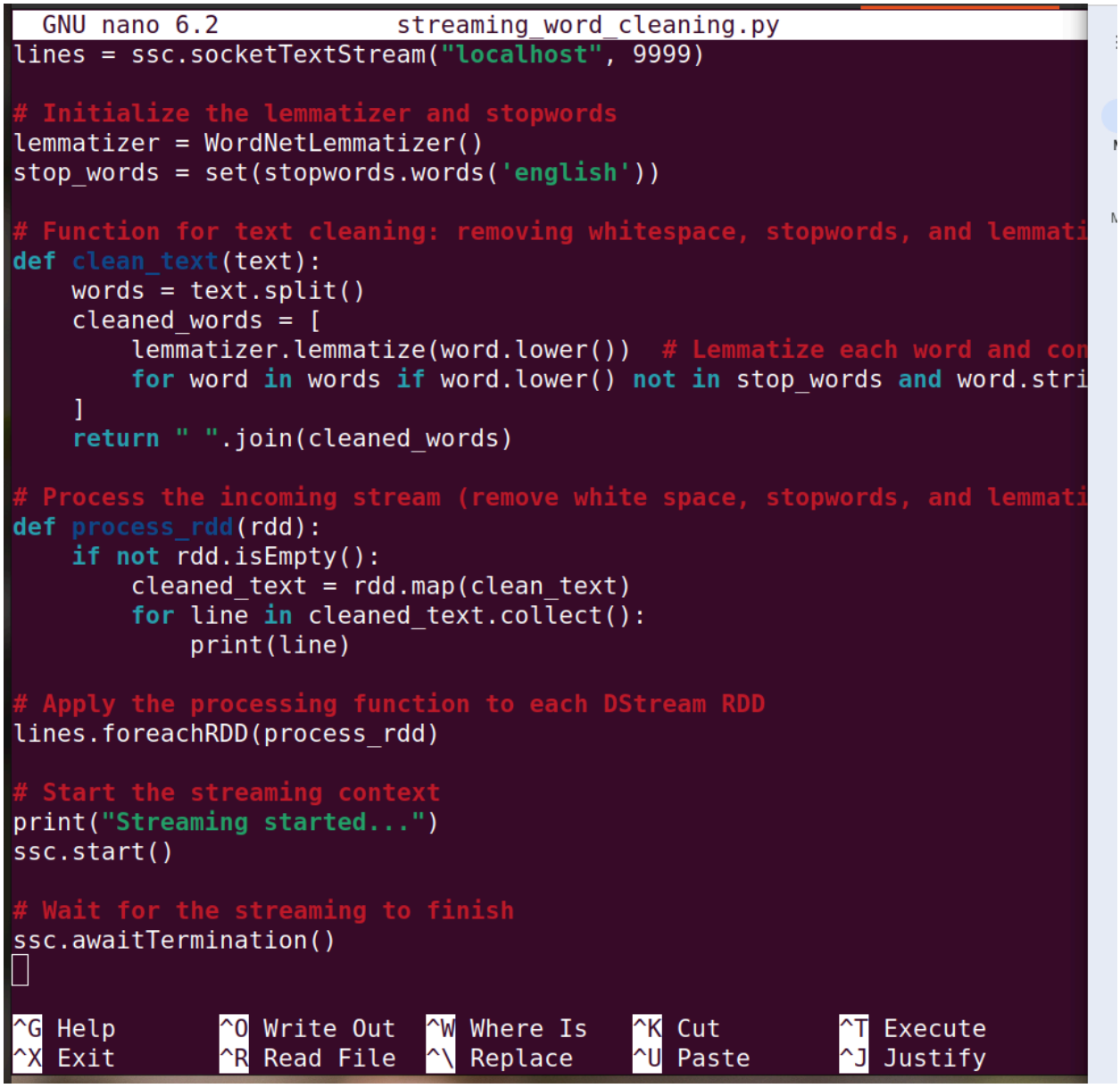
LAB PROGRAM - 10

Write a simple streaming program in Spark to receive text data streams on a particular port, perform basic text cleaning (like white space removal, stop words removal, lemmatization, etc.), and print the cleaned text on the screen. (Open Ended Question).

Code with Output:



```
  GNU nano 6.2              streaming_word_cleaning.py
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
import nltk

# Download necessary NLTK data
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')

# Initialize SparkContext and StreamingContext
sc = SparkContext("local[2]", "StreamingWordClean")
ssc = StreamingContext(sc, 1)  # 1 second batch interval

# Set up the stream by connecting to a socket
lines = ssc.socketTextStream("localhost", 9999)

# Initialize the lemmatizer and stopwords
lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))

# Function for text cleaning: removing whitespace, stopwords, and lemmati>
def clean_text(text):
    words = text.split()
    cleaned_words = [
        lemmatizer.lemmatize(word.lower())  # Lemmatize each word and con>
        for word in words if word.lower() not in stop_words and word.stri>
    ]
    return " ".join(cleaned_words)

# Process the incoming stream (remove white space, stopwords, and lemmati>
                        [ Read 47 lines ]
^G Help       ^O Write Out  ^W Where Is   ^K Cut        ^T Execute
^X Exit       ^R Read File  ^\ Replace    ^U Paste      ^J Justify
```


GNU nano 6.2 streaming_word_cleaning.py
lines = ssc.socketTextStream("localhost", 9999)
# Initialize the lemmatizer and stopwords
lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))
# Function for text cleaning: removing whitespace, stopwords, and lemmati
def clean_text(text):
words = text.split()
cleaned_words = [
lemmatizer.lemmatize(word.lower()) # Lemmatize each word and con
for word in words if word.lower() not in stop_words and word.stri
]
return " ".join(cleaned_words)
# Process the incoming stream (remove white space, stopwords, and lemmati
def process_rdd(rdd):
if not rdd.isEmpty():
cleaned_text = rdd.map(clean_text)
for line in cleaned_text.collect():
print(line)
# Apply the processing function to each DStream RDD
lines.foreachRDD(process_rdd)
# Start the streaming context
print("Streaming started...")
ssc.start()
# Wait for the streaming to finish
ssc.awaitTermination()
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify