BIG DATA ANALYTICS

LAB MANUAL

VI Semester AI&DS

Designed By,

1. **Prof. Chetan Patil**
2. **Prof. Vaibhav Chavan**

Institute Vision

To become premier institute committed to **academic excellence** and **global competence** for the **holistic development** of students.

Key words: academic excellence, global competence, holistic development.

Institute Mission

M1: Develop competent human resources, adopt outcome based education (OBE) and implement cognitive assessment of students.

M2: Inculcate the traits of global competencies amongst the students.

M3: Nurture and train our students to have domain knowledge, develop the qualities of global professionals and to have social consciousness for holistic development.

Department Vision

To deliver a **quality and responsive** education in the field of artificial intelligence and data science emphasizing **professional skills** to face **global challenges** in the evolving IT paradigm.

Key words: quality and responsive, professional skills, global challenges.

Department Mission

M1: Leverage multiple pedagogical approaches to impart knowledge on the current and emerging AI technologies.

M2: Develop an inclusive and holistic ambiance that bolsters problem solving, cognitive abilities and critical thinking.

M3: Enable students to develop trust worthiness, team spirit, understanding law-of-the-land, social behavior to be a global stake holder.

Program Specific Outcomes (PSOs):

PSO1: To apply core knowledge of Artificial Intelligence, Machine Learning, Deep Learning, Data Science, Big Data Analytics and Statistical Learning to develop effective solutions for real-world problems.

PSO2: To demonstrate proficiency in specialized and emerging technologies such as Natural Language Processing, Cloud Computing, Robotic Process Automation, Storage Area Networks and the Internet of Things to meet the stringent and diverse professional challenges.

PSO3: To imbibe managerial skills, social responsibility, ethical and moral values through courses in Management and Entrepreneurship, Software Engineering Principles, Universal Human Values and Ability Enhancement Programs to meet the industry and societal expectations.

Program Educational Objectives (PEOs)

PEO 1: Build a strong foundation in mathematics, core programming, artificial intelligence, machine learning, and data science to enable graduates to analyze, design, and implement intelligent systems for solving complex real-world problems.

PEO 2: Foster creativity, cognitive and research skills to analyze the requirements and technical specifications of software to articulate novel engineering solutions for an efficient product design.

PEO 3: Prepare graduates for dynamic career opportunities in AI and Data Science by equipping them with interdisciplinary knowledge, adaptability, and practical exposure to tools and techniques required for industry and research.

PEO 4: Instill a strong sense of ethics, professional responsibility, and human values, empowering graduates to contribute positively to society and lead with integrity in their professional domains.

PEO 5: Encourage graduates to pursue higher education, certification program, entrepreneurial ventures, etc. by nurturing a mindset of continuous learning and awareness of global trends and challenges.

INDEX

Experiment-1

Aim: Install Hadoop and Implement the following file management tasks in Hadoop: Adding files and directories Retrieving files Deleting files and directories. Hint: A typical Hadoop workflow creates data files (such as log files) elsewhere and copies them into HDFS using one of the above command line utilities.

Sol:

1. Software Requirements:

- Operating System: Ubuntu/Linux (preferred)
- Java JDK 8 or later
- Hadoop 3.x

Step 1: Install Java

- Download JDK from any website (Oracle)
- Install the JDK in System
- Edit Environment Variable
- System variable>>New System Variable>>set as **JAVA_HOME** with path
- System variable>>Path>>New>>set **%JAVA_HOME%\bin**
- User variable>>Path>>New>>set **%JAVA_HOME%\bin**

Step 2: Install Hadoop

- Download Hadoop 3.x
- hadoop-3.3.0.tar.gz (Extract the file)
- Edit Environment Variable
- System variable>>New System Variable>>set as **HADOOP_HOME** with path
- System variable>>Path>>New>>set **%HADOOP_HOME%\bin**
- System variable>>Path>>New>>set **%HADOOP_HOME%\sbin**
- User variable>>Path>>New>>set **%HADOOP_HOME%\bin**
- User variable>>Path>>New>>set **%HADOOP_HOME%\sbin**

Note : Create the folder in Hadoop name Data>>datanode>>namenode

Step 3: Verification of Installation

- java -version
- hadoop -version

Step 4: Configure Hadoop:

Edit some files in Hadoop folder—hadoop>>etc>>Hadoop

1.core-site.xml:

```
<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://0.0.0.0:19000</value>
  </property>
</configuration>
```

2.hdfs-site.xml:

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/hadoop/hadoop-3.3.0/data/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/hadoop/hadoop-3.3.0/data/datanode</value>
  </property>
</configuration>
```

3.mapred-site.xml:

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>mapreduce.application.classpath</name>

<value>%HADOOP_HOME%/share/hadoop/mapreduce/*,%HADOOP_HOME%/share/hadoop/mapreduce/lib/*,%HADOOP_HOME%/share/hadoop/common/*,%HADOOP_HOME%/share/hadoop/common/lib/*,%HADOOP_HOME%/share/hadoop/yarn/*,%HADOOP_HOME%/share/hadoop/yarn/lib/*,%HADOOP_HOME%/share/hadoop/hdfs/*,%HADOOP_HOME%/share/hadoop/hdfs/lib/*</value>
  </property>
</configuration>
```

4.yarn-site.xml:

```
    <configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.env-whitelist</name>

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,
CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value
>
  </property>
</configuration>
```

Step 5: Replace Hadoop bin file from winutils.master bin as required

Step 6: Format the NameNode: cmd>>Run as administrator>>Type>>
"hdfs namenode -format"

Step 7: Start HDFS daemons:

"%HADOOP_HOME%\sbin\start-dfs.cmd"

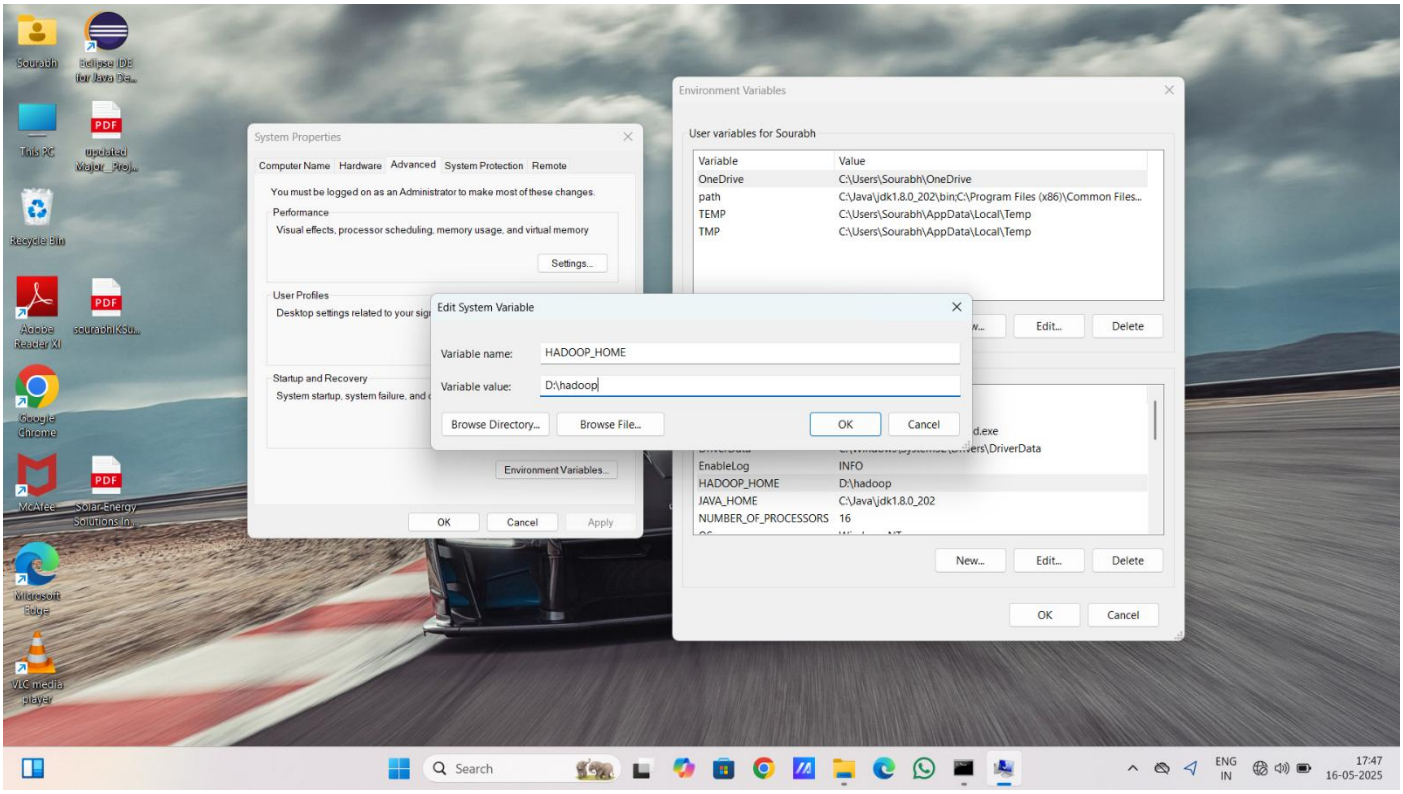
Step 8: Start YARN daemons:

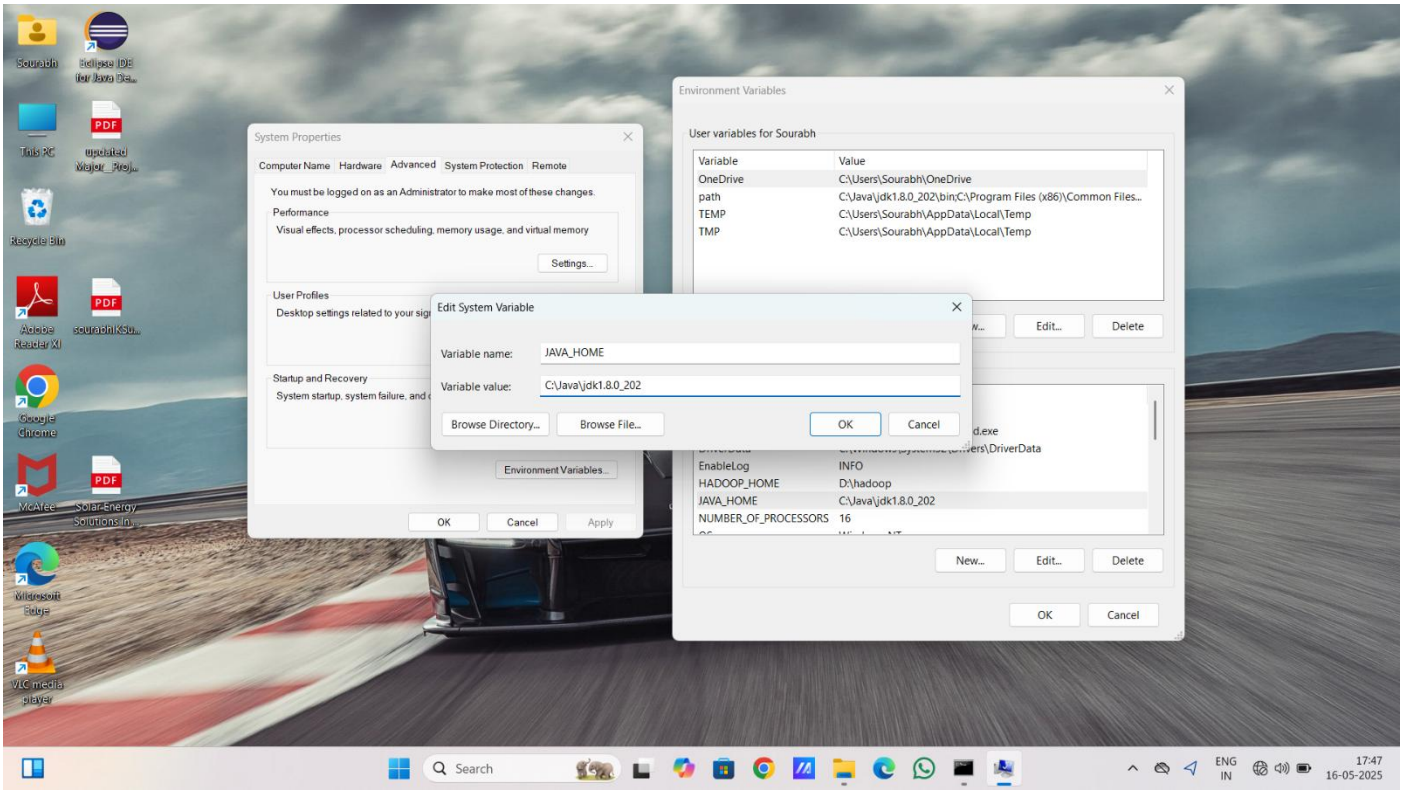
"%HADOOP_HOME%\sbin\start-yarn.cmd"

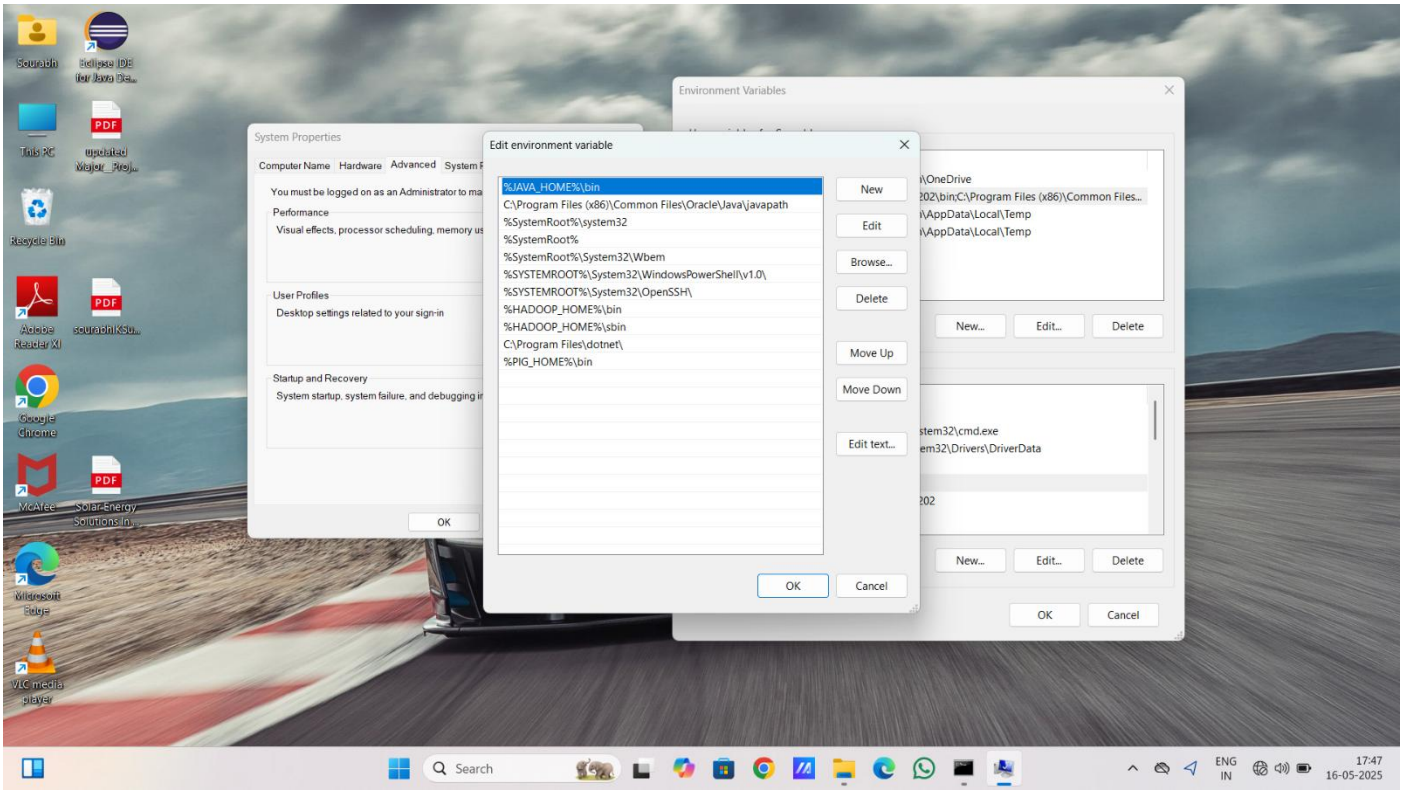
Step 9: You can verify YARN resource manager UI when all services are started successfully.

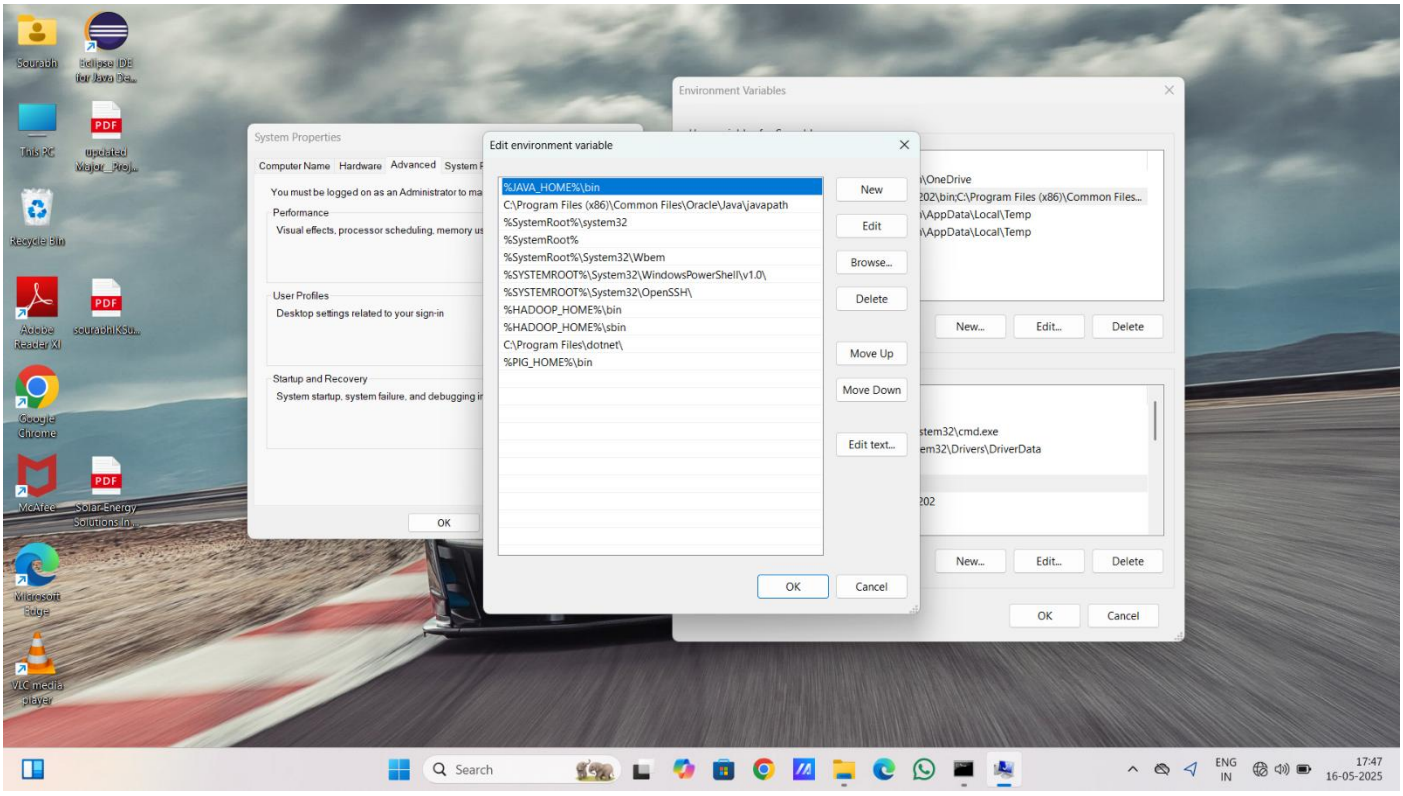
Type localhost 8088 in Address Bar or URL Bar

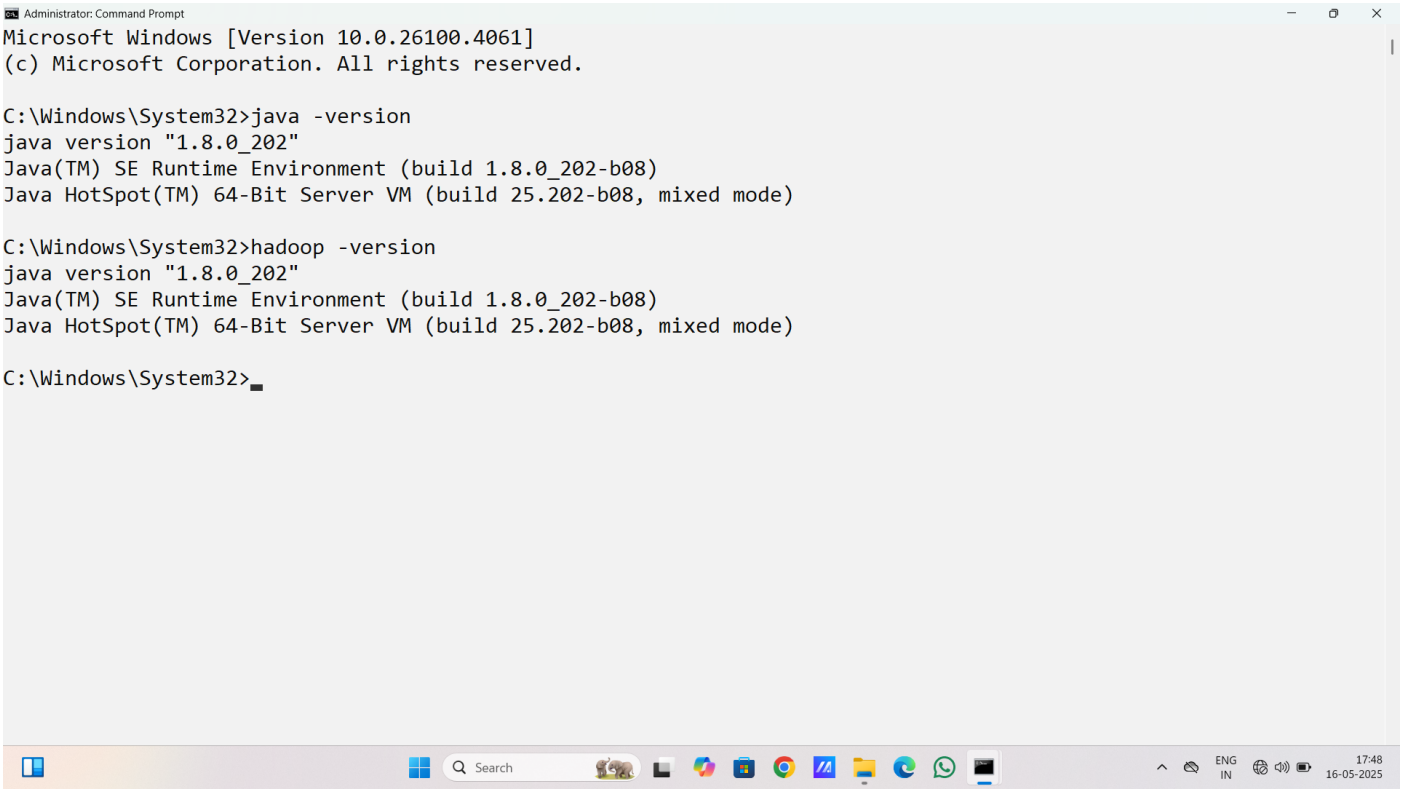
Output:

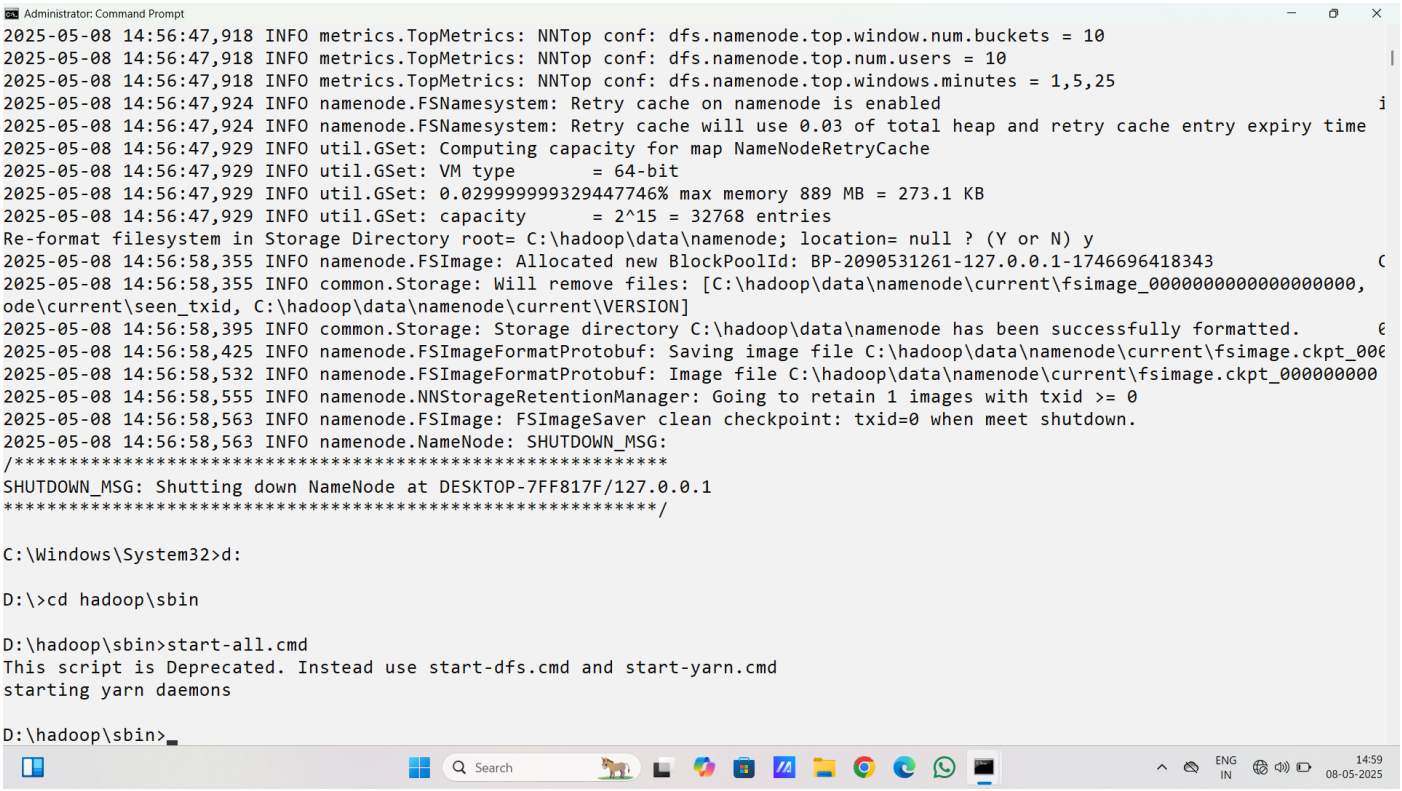

System Properties
Computer Name Hardware Advanced System Protection Remote
You must be logged on as an Administrator to make most of these changes.
Performance
Visual effects, processor scheduling, memory usage, and virtual memory
Settings...
User Profiles
Desktop settings related to your sign-in
Startup and Recovery
System startup, system failure, and debugging information
Environment Variables...
OK
Cancel
Apply
Edit System Variable
Variable name: HADOOP_HOME
Variable value: D:\hadoop
Browse Directory...
Browse File...
OK
Cancel
Environment Variables
User variables for Sourabh
Variable
Value
OneDrive
C:\Users\Sourabh\OneDrive
path
C:\Java\jdk1.8.0_202\bin;C:\Program Files (x86)\Common Files...
TEMP
C:\Users\Sourabh\AppData\Local\Temp
TMP
C:\Users\Sourabh\AppData\Local\Temp
New...
Edit...
Delete
EnableLog
INFO
HADOOP_HOME
D:\hadoop
JAVA_HOME
C:\Java\jdk1.8.0_202
NUMBER_OF_PROCESSORS
16
OK
Cancel
Search
ENG
IN
17:47
16-05-2025

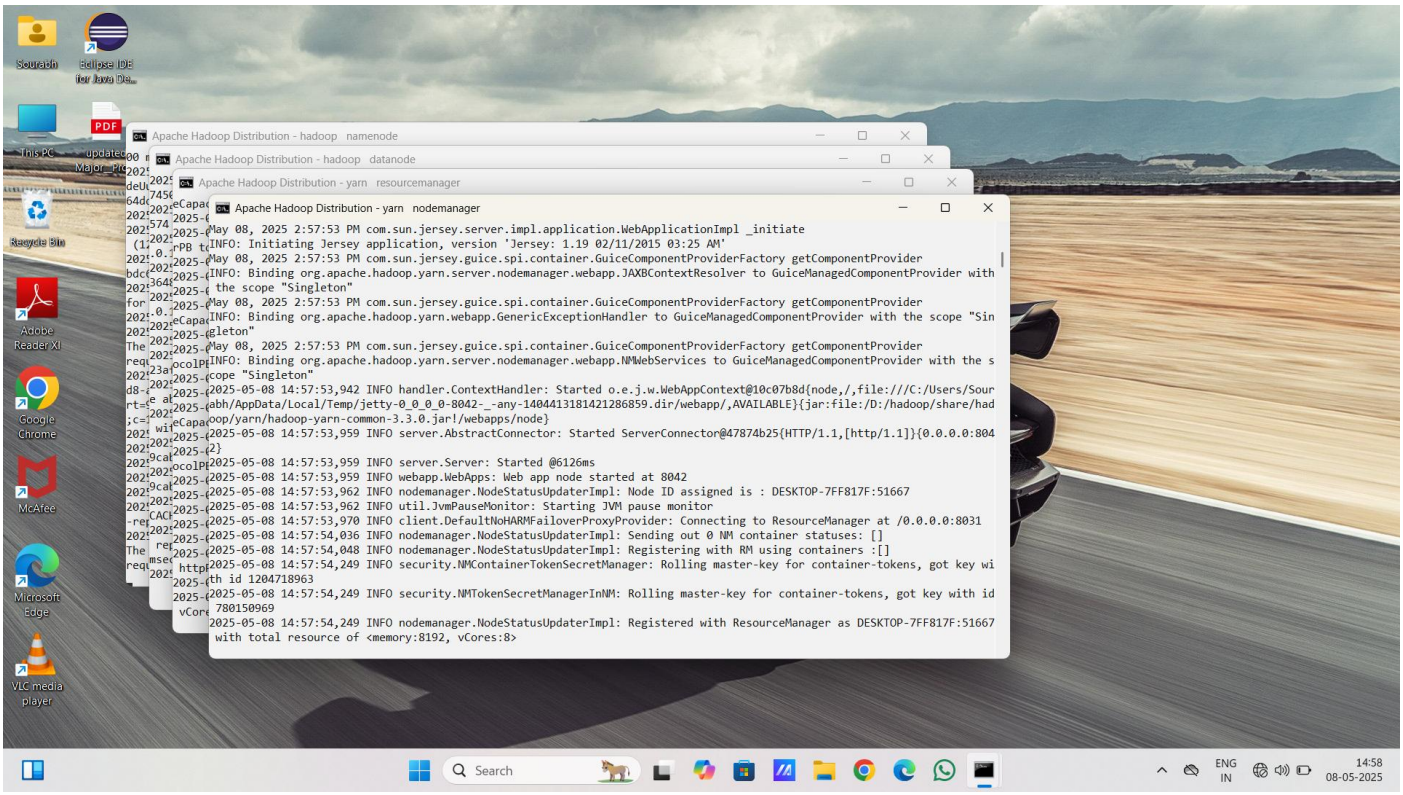

System Properties
Computer Name Hardware Advanced System Protection Remote
You must be logged on as an Administrator to make most of these changes.
Performance
Visual effects, processor scheduling, memory usage, and virtual memory
Settings...
User Profiles
Desktop settings related to your sign-in
Startup and Recovery
System startup, system failure, and debugging information
Environment Variables...
OK
Cancel
Apply
Edit System Variable
Variable name: JAVA_HOME
Variable value: C:\Java\jdk1.8.0_202
Browse Directory...
Browse File...
OK
Cancel
Environment Variables
User variables for Sourabh
Variable
Value
OneDrive
C:\Users\Sourabh\OneDrive
path
C:\Java\jdk1.8.0_202\bin;C:\Program Files (x86)\Common Files...
TEMP
C:\Users\Sourabh\AppData\Local\Temp
TMP
C:\Users\Sourabh\AppData\Local\Temp
New...
Edit...
Delete
EnableLog
INFO
HADOOP_HOME
D:\hadoop
JAVA_HOME
C:\Java\jdk1.8.0_202
NUMBER_OF_PROCESSORS
16
OK
Cancel
Search
ENG
IN
17:47
16-05-2025

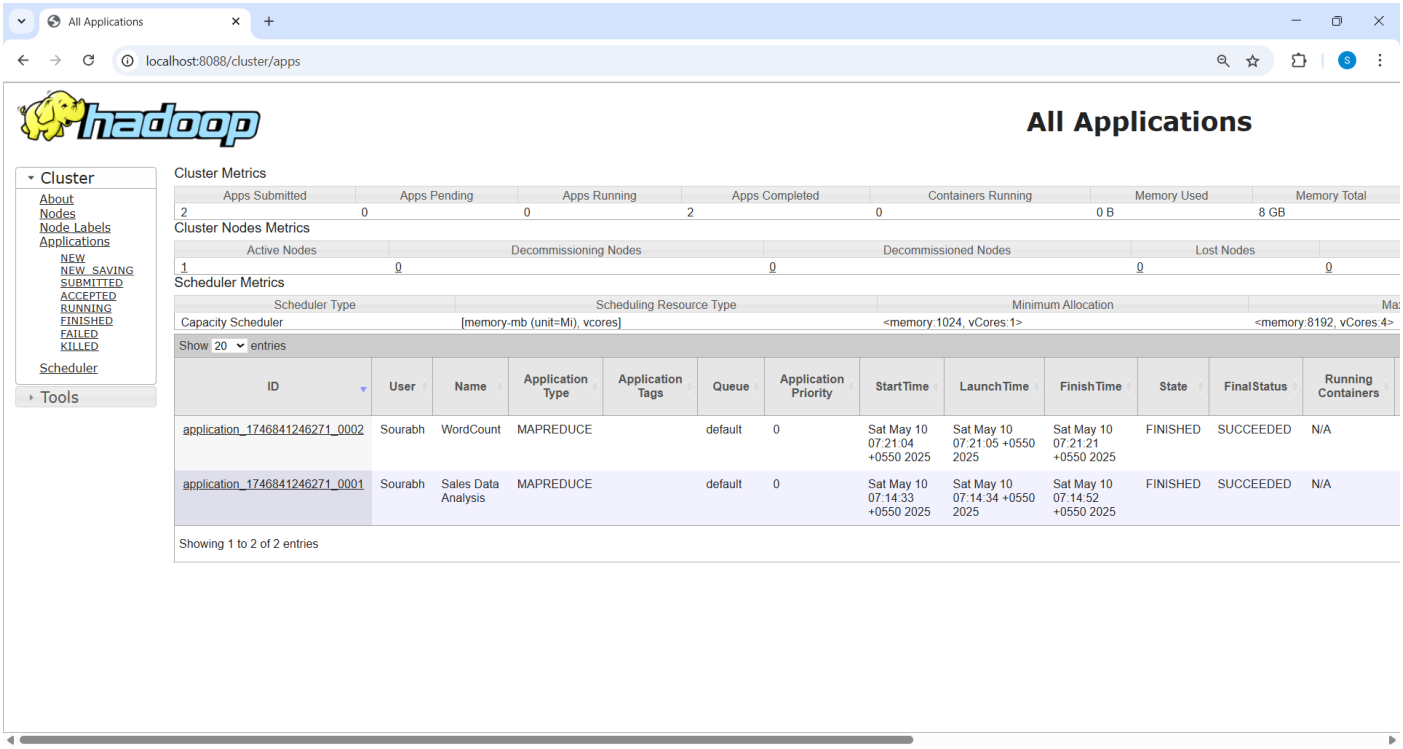

Edit environment variable
%JAVA_HOME%\bin
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
%HADOOP_HOME%\bin
%HADOOP_HOME%\sbin
C:\Program Files\dotnet\
%PIG_HOME%\bin
New
Edit
Browse...
Delete
Move Up
Move Down
Edit text...
OK
Cancel
Environment Variables
System Properties
Search
ENG
IN
17:47
16-05-2025


Edit environment variable
%JAVA_HOME%\bin
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
%HADOOP_HOME%\bin
%HADOOP_HOME%\sbin
C:\Program Files\dotnet\
%PIG_HOME%\bin
New
Edit
Browse...
Delete
Move Up
Move Down
Edit text...
OK
Cancel
Environment Variables
System Properties
Search
ENG
IN
17:47
16-05-2025


Administrator: Command Prompt
Microsoft Windows [Version 10.0.26100.4061]
(c) Microsoft Corporation. All rights reserved.
C:\Windows\System32>java -version
java version "1.8.0_202"
Java(TM) SE Runtime Environment (build 1.8.0_202-b08)
Java HotSpot(TM) 64-Bit Server VM (build 25.202-b08, mixed mode)
C:\Windows\System32>hadoop -version
java version "1.8.0_202"
Java(TM) SE Runtime Environment (build 1.8.0_202-b08)
Java HotSpot(TM) 64-Bit Server VM (build 25.202-b08, mixed mode)
C:\Windows\System32>
Search
ENG
IN
17:48
16-05-2025


Administrator: Command Prompt
2025-05-08 14:56:47,918 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2025-05-08 14:56:47,918 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2025-05-08 14:56:47,918 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2025-05-08 14:56:47,924 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2025-05-08 14:56:47,924 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time
2025-05-08 14:56:47,929 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2025-05-08 14:56:47,929 INFO util.GSet: VM type = 64-bit
2025-05-08 14:56:47,929 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
2025-05-08 14:56:47,929 INFO util.GSet: capacity = 2^15 = 32768 entries
Re-format filesystem in Storage Directory root= C:\hadoop\data\namenode; location= null ? (Y or N) y
2025-05-08 14:56:58,355 INFO namenode.FSImage: Allocated new BlockPoolId: BP-2090531261-127.0.0.1-1746696418343
2025-05-08 14:56:58,355 INFO common.Storage: Will remove files: [C:\hadoop\data\namenode\current\fsimage_0000000000000000000,
ode\current\seen_txid, C:\hadoop\data\namenode\current\VERSION]
2025-05-08 14:56:58,395 INFO common.Storage: Storage directory C:\hadoop\data\namenode has been successfully formatted.
2025-05-08 14:56:58,425 INFO namenode.FSImageFormatProtobuf: Saving image file C:\hadoop\data\namenode\current\fsimage.ckpt_000
2025-05-08 14:56:58,532 INFO namenode.FSImageFormatProtobuf: Image file C:\hadoop\data\namenode\current\fsimage.ckpt_000000000
2025-05-08 14:56:58,555 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2025-05-08 14:56:58,563 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2025-05-08 14:56:58,563 INFO namenode.NameNode: SHUTDOWN_MSG:
/**
SHUTDOWN_MSG: Shutting down NameNode at DESKTOP-7FF817F/127.0.0.1
**/
C:\Windows\System32>d:
D:\>cd hadoop\sbin
D:\hadoop\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
D:\hadoop\sbin>
Search
ENG
IN
14:59
08-05-2025


Apache Hadoop Distribution - hadoop namenode
Apache Hadoop Distribution - hadoop datanode
Apache Hadoop Distribution - yarn resourcemanager
Apache Hadoop Distribution - yarn nodemanager
May 08, 2025 2:57:53 PM com.sun.jersey.server.impl.application.WebApplicationImpl _initiate
INFO: Initiating Jersey application, version 'Jersey: 1.19 02/11/2015 03:25 AM'
May 08, 2025 2:57:53 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.JAXBContextResolver to GuiceManagedComponentProvider with the scope "Singleton"
May 08, 2025 2:57:53 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.webapp.GenericExceptionHandler to GuiceManagedComponentProvider with the scope "Singleton"
May 08, 2025 2:57:53 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.NMWebServices to GuiceManagedComponentProvider with the scope "Singleton"
2025-05-08 14:57:53,942 INFO handler.ContextHandler: Started o.e.j.w.WebAppContext@10c07b8d{node,/,file:///C:/Users/Sourabh/AppData/Local/Temp/jetty-0_0_0_0-8042-_-any-14044131814212868 59.dir/webapp/,AVAILABLE}{jar:file:/D:/hadoop/share/hadoop/yarn/hadoop-yarn-common-3.3.0.jar!/webapps/node}
2025-05-08 14:57:53,959 INFO server.AbstractConnector: Started ServerConnector@47874b25{HTTP/1.1,[http/1.1]}{0.0.0.0:8042}
2025-05-08 14:57:53,959 INFO server.Server: Started @6126ms
2025-05-08 14:57:53,959 INFO webapp.WebApps: Web app node started at 8042
2025-05-08 14:57:53,962 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : DESKTOP-7FF817F:51667
2025-05-08 14:57:53,962 INFO util.JvmPauseMonitor: Starting JVM pause monitor
2025-05-08 14:57:53,970 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8031
2025-05-08 14:57:54,036 INFO nodemanager.NodeStatusUpdaterImpl: Sending out 0 NM container statuses: []
2025-05-08 14:57:54,048 INFO nodemanager.NodeStatusUpdaterImpl: Registering with RM using containers :[]
2025-05-08 14:57:54,249 INFO security.NMContainerTokenSecretManager: Rolling master-key for container-tokens, got key with id 1204718963
2025-05-08 14:57:54,249 INFO security.NMTokenSecretManagerInNM: Rolling master-key for container-tokens, got key with id 780150969
2025-05-08 14:57:54,249 INFO nodemanager.NodeStatusUpdaterImpl: Registered with ResourceManager as DESKTOP-7FF817F:51667 with total resource of <memory:8192, vCores:8>
Search
ENG IN
14:58
08-05-2025


All Applications
localhost:8088/cluster/apps
All Applications
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Cluster Metrics
Apps Submitted 2
Apps Pending 0
Apps Running 0
Apps Completed 2
Containers Running 0
Memory Used 0 B
Memory Total 8 GB
Cluster Nodes Metrics
Active Nodes 1
Decommissioning Nodes 0
Decommissioned Nodes 0
Lost Nodes 0
Scheduler Metrics
Scheduler Type: Capacity Scheduler
Scheduling Resource Type: [memory-mb (unit=Mi), vcores]
Minimum Allocation: <memory:1024, vCores:1>
Ma: <memory:8192, vCores:4>
Show 20 entries
ID | User | Name | Application Type | Application Tags | Queue | Application Priority | StartTime | LaunchTime | FinishTime | State | FinalStatus | Running Containers
application_1746841246271_0002 | Sourabh | WordCount | MAPREDUCE | | default | 0 | Sat May 10 07:21:04 +0550 2025 | Sat May 10 07:21:05 +0550 2025 | Sat May 10 07:21:21 +0550 2025 | FINISHED | SUCCEEDED | N/A
application_1746841246271_0001 | Sourabh | Sales Data Analysis | MAPREDUCE | | default | 0 | Sat May 10 07:14:33 +0550 2025 | Sat May 10 07:14:34 +0550 2025 | Sat May 10 07:14:52 +0550 2025 | FINISHED | SUCCEEDED | N/A
Showing 1 to 2 of 2 entries

Experiment 2

Aim: Develop a MapReduce program to implement Matrix Multiplication

Code:

Map:

```
package com.mapreduce.wc;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
//import org.apache.hadoop.mapreduce.Mapper;
import java.io.IOException;
public class Map extends org.apache.hadoop.mapreduce.Mapper<LongWritable, Text, Text, Text>
{
        @Override
public void map(LongWritable key, Text value, Context context)
throws IOException, InterruptedException {
                Configuration conf = context.getConfiguration();
                int m = Integer.parseInt(conf.get("m"));
                int p = Integer.parseInt(conf.get("p"));
String line = value.toString();
// (M, i, j, Mij);
String[] indicesAndValue = line.split(",");
Text outputKey = new Text();
Text outputValue = new Text();
if (indicesAndValue[0].equals("M")) {
for (int k = 0; k < p; k++) {
outputKey.set(indicesAndValue[1] + "," + k);
// outputKey.set(i,k);
outputValue.set(indicesAndValue[0] + "," + indicesAndValue[2]
```

```
+ "," + indicesAndValue[3]);
// outputValue.set(M,j,Mij);
context.write(outputKey, outputValue);
}
} else {
        // (N, j, k, Njk);
        for (int i = 0; i < m; i++) {
        outputKey.set(i + "," + indicesAndValue[2]); outputValue.set("N," + indicesAndValue[1] + ","
        + indicesAndValue[3]); context.write(outputKey, outputValue);
        }
        }
        }
}
```

MatrixMultiply:

```
package com.mapreduce.wc;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
public class MatrixMultiply {
public static void main(String[] args) throws Exception { if (args.length != 2) {
System.err.println("Usage: MatrixMultiply <in_dir> <out_dir>");
System.exit(2);
}
Configuration conf = new Configuration();
```

```
conf.set("m", "1000");
conf.set("n", "100");
conf.set("p", "1000");
@SuppressWarnings("deprecation")
Job job = new Job(conf, "MatrixMultiply");
job.setJarByClass(MatrixMultiply.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class);
job.setMapperClass(Map.class);
job.setReducerClass(Reduce.class);
job.setInputFormatClass(TextInputFormat.class);
job.setOutputFormatClass(TextOutputFormat.class);
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new Path(args[1]));
job.waitForCompletion(true);
}
}
```

Reduce:

```
package com.mapreduce.wc;
import org.apache.hadoop.io.Text;
// import org.apache.hadoop.mapreduce.Reducer;
import java.io.IOException;
import java.util.HashMap;
public class Reduce
extends org.apache.hadoop.mapreduce.Reducer<Text, Text, Text, Text> { @Override
public void reduce(Text key, Iterable<Text> values, Context context)
throws IOException, InterruptedException {
String[] value;
//key=(i,k),
```

```
//Values = [(M/N,j,V/W),..]

HashMap<Integer, Float> hashA = new HashMap<Integer, Float>(); HashMap<Integer, Float> hashB =
new HashMap<Integer, Float>(); for (Text val : values) {

value = val.toString().split(",");

if (value[0].equals("M")) {

hashA.put(Integer.parseInt(value[1]), Float.parseFloat(value[2])); } else {

hashB.put(Integer.parseInt(value[1]), Float.parseFloat(value[2]));

}

}

int n = Integer.parseInt(context.getConfiguration().get("n"));

float result = 0.0f;

float m_ij;

float n_jk;

for (int j = 0; j < n; j++) {

m_ij = hashA.containsKey(j) ? hashA.get(j) : 0.0f; n_jk = hashB.containsKey(j) ? hashB.get(j) : 0.0f; result
+= m_ij * n_jk;

}

if (result != 0.0f) {

context.write(null,

new Text(key.toString() + "," + Float.toString(result)));

}

}

}
```

Commands for Execution

1. D:\hadoop\sbin>start-all.cmd

#Create Directory under Hadoop

2. D:\hadoop\sbin>hadoop fs -mkdir /pgm2
3. D:\hadoop\sbin>hadoop fs -put D:\hadoop_experiments\MatrixMultiply_M.txt /pgm2

#upload a file on to the hadoop Directory

4. D:\hadoop\sbin>hadoop fs -put D:\hadoop_experiments\MatrixMultiply_N.txt /pgm2

#To check file is uploaded or not

5. D:\hadoop\sbin>hadoop fs -ls /pgm2

#TO run jar

6. D:\hadoop\sbin>hadoop jar D:\hadoop_experiments\Multiplication.jar com.mapreduce.wc/MatrixMultiply /pgm2/* /pgm2OP
7. D:\hadoop\sbin>hadoop fs -cat /pgm2OP/part-r-00000

Matrix input file:

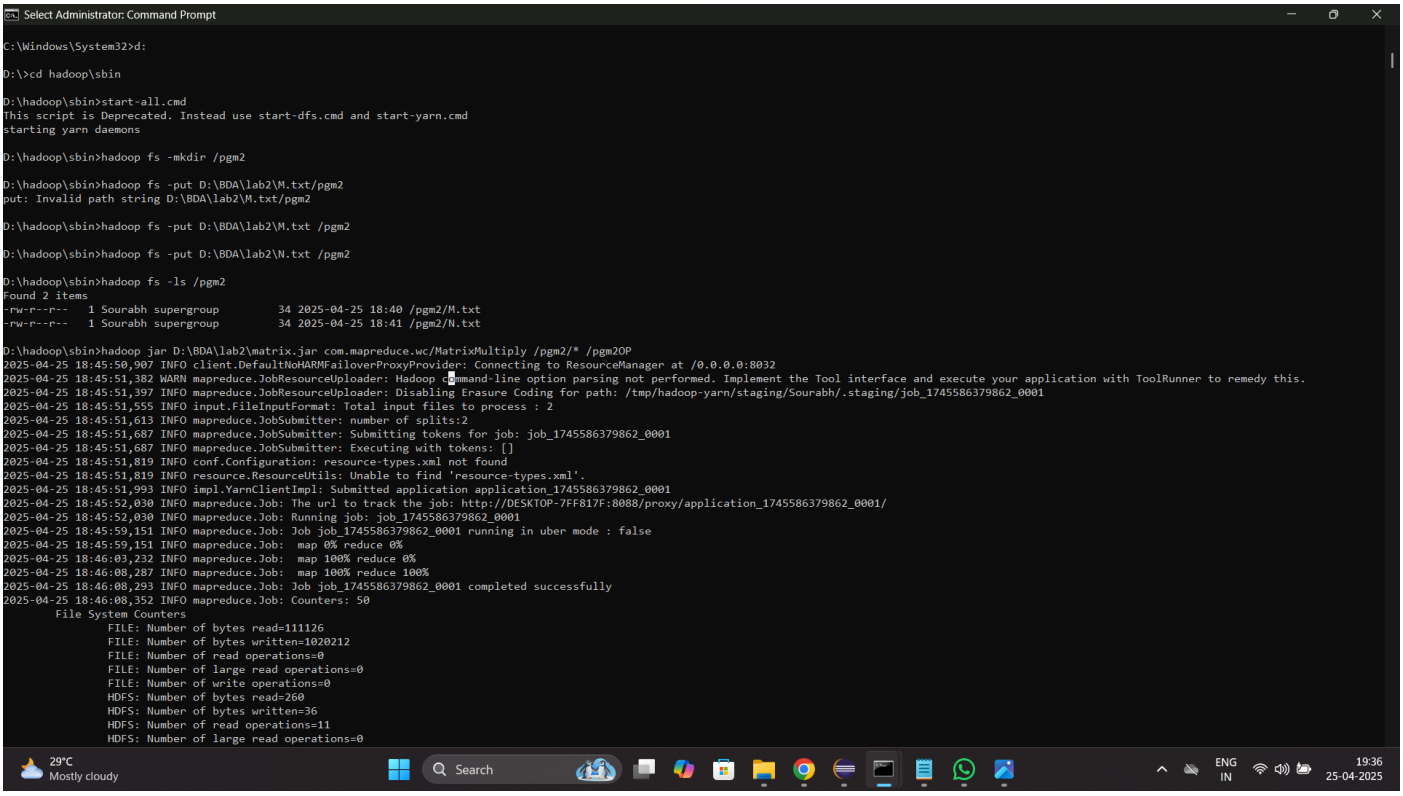
M.txt:

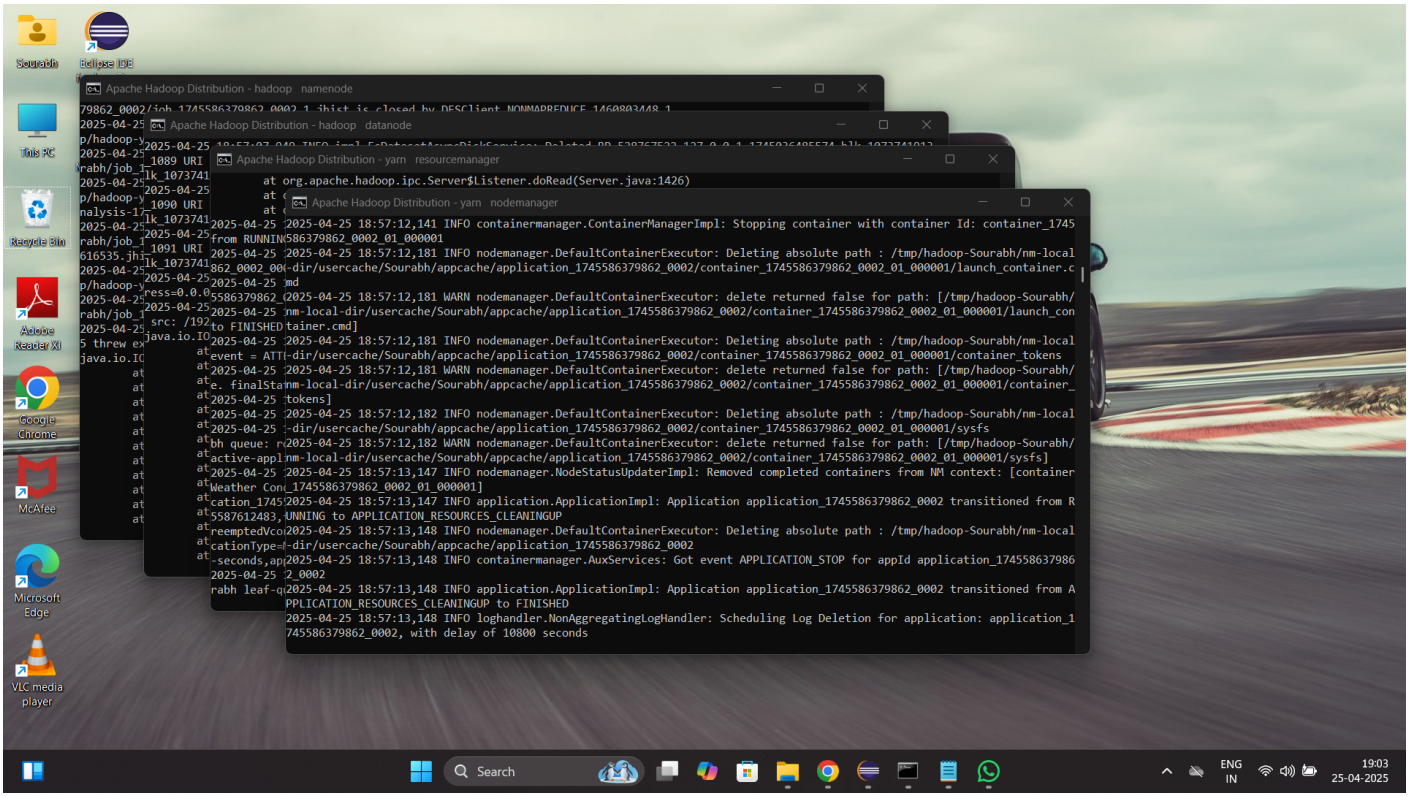
M,0,0,1
M,0,1,2
M,1,0,3
M,1,1,4

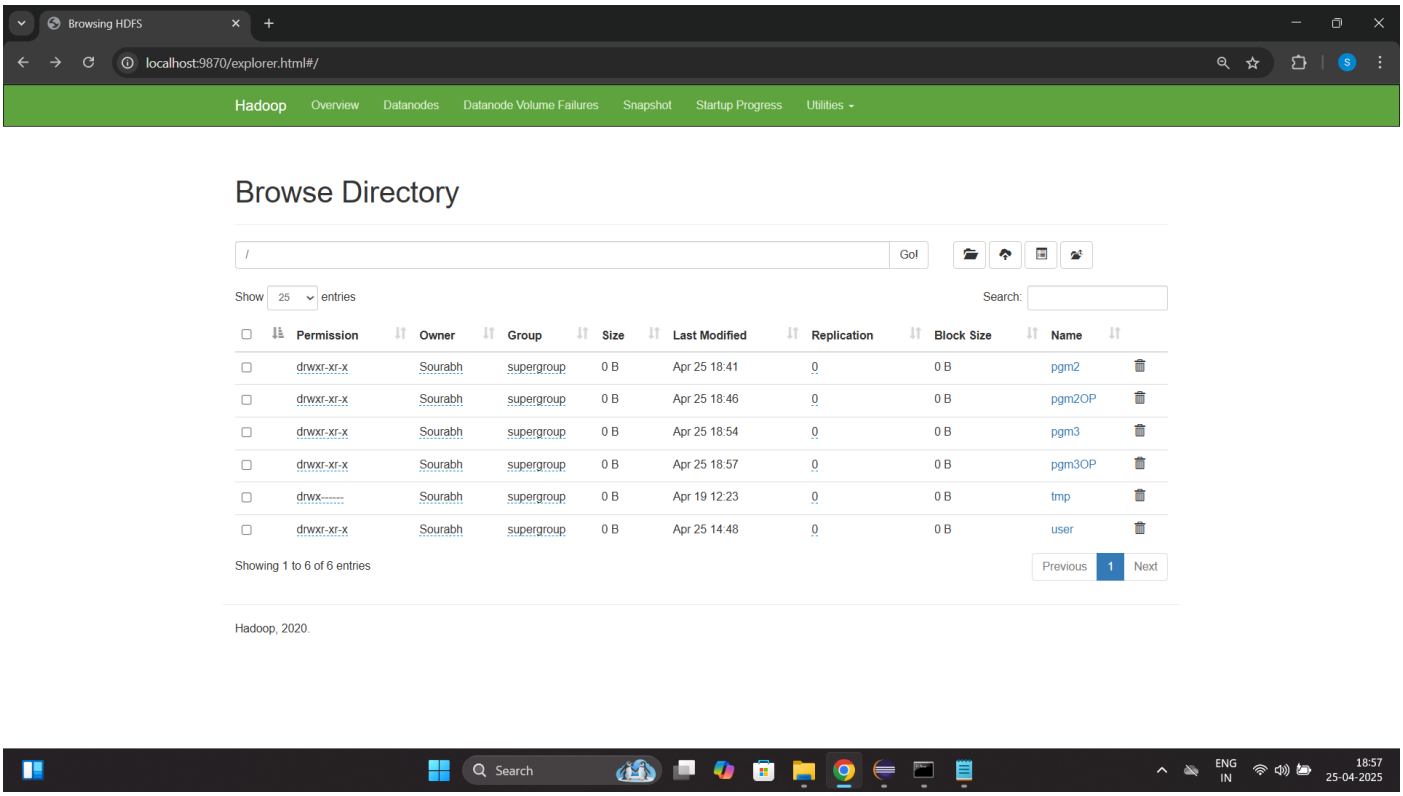
N.txt:

N,0,0,5
N,0,1,6
N,1,0,7
N,1,1,8

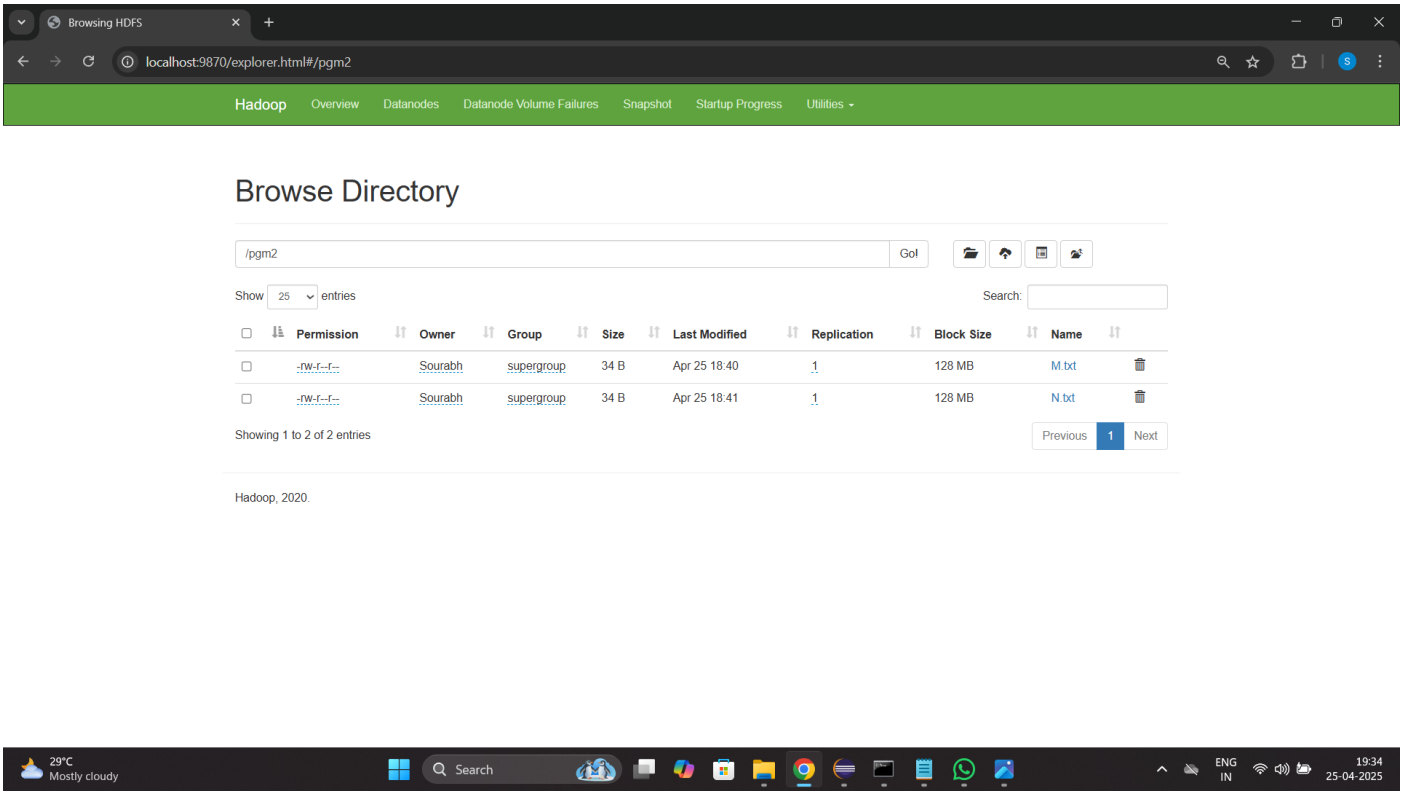
Output:


Select Administrator: Command Prompt
C:\Windows\System32>d:
D:\>cd hadoop\sbin
D:\hadoop\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
D:\hadoop\sbin>hadoop fs -mkdir /pgm2
D:\hadoop\sbin>hadoop fs -put D:\BDA\lab2\M.txt/pgm2
put: Invalid path string D:\BDA\lab2\M.txt/pgm2
D:\hadoop\sbin>hadoop fs -put D:\BDA\lab2\M.txt /pgm2
D:\hadoop\sbin>hadoop fs -put D:\BDA\lab2\N.txt /pgm2
D:\hadoop\sbin>hadoop fs -ls /pgm2
Found 2 items
-rw-r--r-- 1 Sourabh supergroup 34 2025-04-25 18:40 /pgm2/M.txt
-rw-r--r-- 1 Sourabh supergroup 34 2025-04-25 18:41 /pgm2/N.txt
D:\hadoop\sbin>hadoop jar D:\BDA\lab2\matrix.jar com.mapreduce.wc/MatrixMultiply /pgm2/* /pgm2OP
2025-04-25 18:45:50,907 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-25 18:45:51,382 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-04-25 18:45:51,397 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Sourabh/.staging/job_1745586379862_0001
2025-04-25 18:45:51,555 INFO input.FileInputFormat: Total input files to process : 2
2025-04-25 18:45:51,613 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-25 18:45:51,687 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745586379862_0001
2025-04-25 18:45:51,687 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-25 18:45:51,819 INFO conf.Configuration: resource-types.xml not found
2025-04-25 18:45:51,819 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-25 18:45:51,993 INFO impl.YarnClientImpl: Submitted application application_1745586379862_0001
2025-04-25 18:45:52,030 INFO mapreduce.Job: The url to track the job: http://DESKTOP-7FF817F:8088/proxy/application_1745586379862_0001/
2025-04-25 18:45:52,030 INFO mapreduce.Job: Running job: job_1745586379862_0001
2025-04-25 18:45:59,151 INFO mapreduce.Job: Job job_1745586379862_0001 running in uber mode : false
2025-04-25 18:45:59,151 INFO mapreduce.Job: map 0% reduce 0%
2025-04-25 18:46:03,232 INFO mapreduce.Job: map 100% reduce 0%
2025-04-25 18:46:08,287 INFO mapreduce.Job: map 100% reduce 100%
2025-04-25 18:46:08,293 INFO mapreduce.Job: Job job_1745586379862_0001 completed successfully
2025-04-25 18:46:08,352 INFO mapreduce.Job: Counters: 50
File System Counters
FILE: Number of bytes read=111126
FILE: Number of bytes written=1020212
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=260
HDFS: Number of bytes written=36
HDFS: Number of read operations=11
HDFS: Number of large read operations=0
29°C
Mostly cloudy
Search
ENG
IN
19:36
25-04-2025

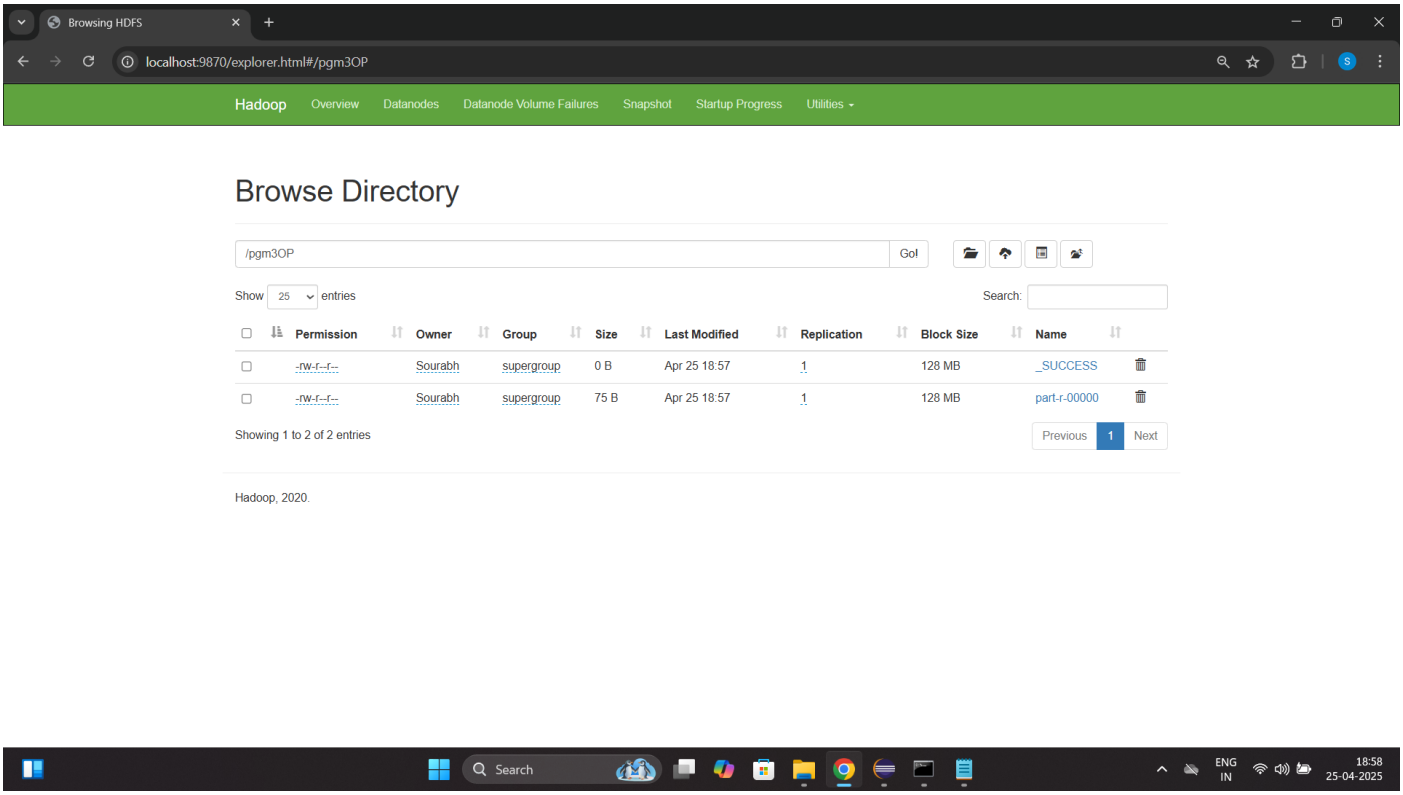

Sourabh
Eclipse IDE
This PC
Recycle Bin
Adobe Reader XI
Google Chrome
McAfee
Microsoft Edge
VLC media player
Apache Hadoop Distribution - hadoop namenode
Apache Hadoop Distribution - hadoop datanode
Apache Hadoop Distribution - yarn resourcemanager
Apache Hadoop Distribution - yarn nodemanager
at org.apache.hadoop.ipc.Server$Listener.doRead(Server.java:1426)
2025-04-25 18:57:12,141 INFO containermanager.ContainerManagerImpl: Stopping container with container Id: container_1745
2025-04-25 18:57:12,181 INFO nodemanager.DefaultContainerExecutor: Deleting absolute path : /tmp/hadoop-Sourabh/nm-local
2025-04-25 18:57:12,181 WARN nodemanager.DefaultContainerExecutor: delete returned false for path: [/tmp/hadoop-Sourabh/
2025-04-25 18:57:12,181 INFO nodemanager.DefaultContainerExecutor: Deleting absolute path : /tmp/hadoop-Sourabh/nm-local
2025-04-25 18:57:12,181 WARN nodemanager.DefaultContainerExecutor: delete returned false for path: [/tmp/hadoop-Sourabh/
2025-04-25 18:57:12,182 INFO nodemanager.DefaultContainerExecutor: Deleting absolute path : /tmp/hadoop-Sourabh/nm-local
2025-04-25 18:57:12,182 WARN nodemanager.DefaultContainerExecutor: delete returned false for path: [/tmp/hadoop-Sourabh/
2025-04-25 18:57:13,147 INFO nodemanager.NodeStatusUpdaterImpl: Removed completed containers from NM context: [container
2025-04-25 18:57:13,147 INFO application.ApplicationImpl: Application application_1745586379862_0002 transitioned from R
2025-04-25 18:57:13,148 INFO nodemanager.DefaultContainerExecutor: Deleting absolute path : /tmp/hadoop-Sourabh/nm-local
2025-04-25 18:57:13,148 INFO containermanager.AuxServices: Got event APPLICATION_STOP for appId application_174558637986
2025-04-25 18:57:13,148 INFO application.ApplicationImpl: Application application_1745586379862_0002 transitioned from A
2025-04-25 18:57:13,148 INFO loghandler.NonAggregatingLogHandler: Scheduling Log Deletion for application: application_1
745586379862_0002, with delay of 10800 seconds
Search
ENG
IN
19:03
25-04-2025

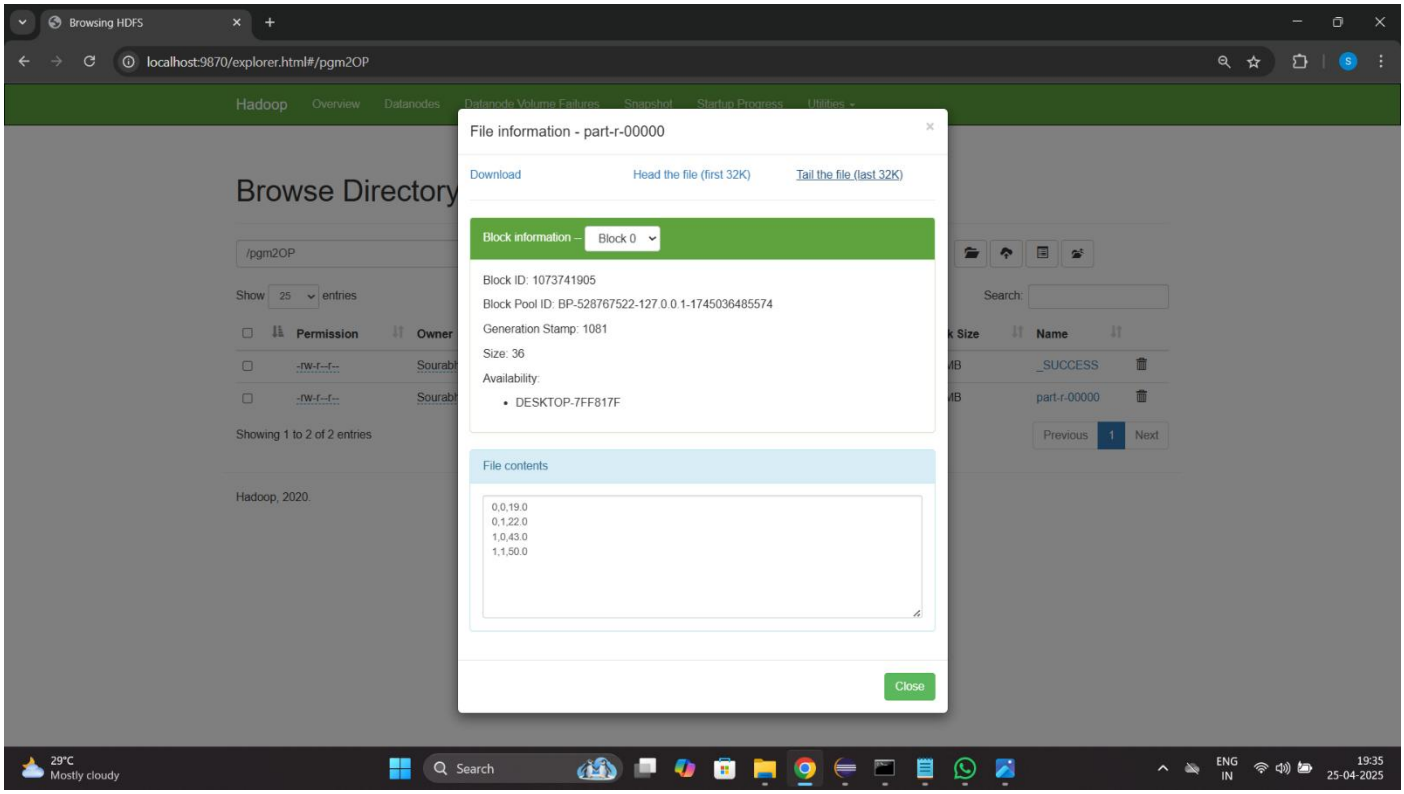

Browsing HDFS
localhost:9870/explorer.html#/
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:41 0 0 B pgm2
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:46 0 0 B pgm2OP
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:54 0 0 B pgm3
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:57 0 0 B pgm3OP
drwx------ Sourabh supergroup 0 B Apr 19 12:23 0 0 B tmp
drwxr-xr-x Sourabh supergroup 0 B Apr 25 14:48 0 0 B user
Showing 1 to 6 of 6 entries
Previous
1
Next
Hadoop, 2020.

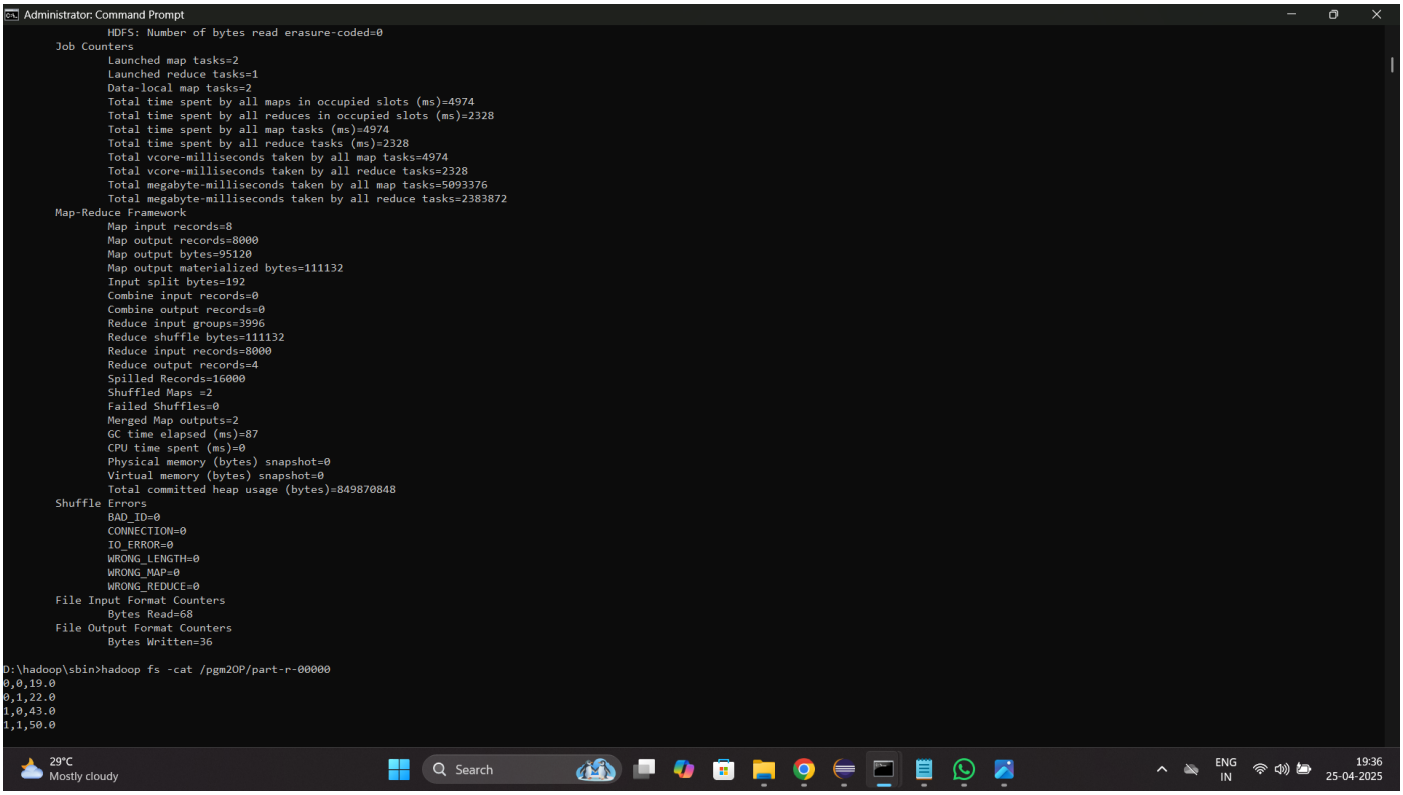
Search
ENG
IN
18:57
25-04-2025


Browsing HDFS
localhost:9870/explorer.html#/pgm2
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm2
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r-- Sourabh supergroup 34 B Apr 25 18:40 1 128 MB M.txt
-rw-r--r-- Sourabh supergroup 34 B Apr 25 18:41 1 128 MB N.txt
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.

29°C
Mostly cloudy
Search
ENG
IN
19:34
25-04-2025


Browsing HDFS
localhost:9870/explorer.html#/pgm3OP
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm3OP
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
Sourabh
supergroup
0 B
Apr 25 18:57
1
128 MB
_SUCCESS
-rw-r--r--
Sourabh
supergroup
75 B
Apr 25 18:57
1
128 MB
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.
Search
ENG
IN
18:58
25-04-2025


Browsing HDFS
localhost:9870/explorer.html#/pgm2OP
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm2OP
Show 25 entries
Search:
Permission
Owner
Name
-rw-r--r--
_SUCCESS
-rw-r--r--
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.
File information - part-r-00000
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information -- Block 0
Block ID: 1073741905
Block Pool ID: BP-528767522-127.0.0.1-1745036485574
Generation Stamp: 1081
Size: 36
Availability:
DESKTOP-7FF817F
File contents
0,0,19.0
0,1,22.0
1,0,43.0
1,1,50.0
Close
29°C
Mostly cloudy
Search
ENG
IN
19:35
25-04-2025


Administrator: Command Prompt
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=4974
Total time spent by all reduces in occupied slots (ms)=2328
Total time spent by all map tasks (ms)=4974
Total time spent by all reduce tasks (ms)=2328
Total vcore-milliseconds taken by all map tasks=4974
Total vcore-milliseconds taken by all reduce tasks=2328
Total megabyte-milliseconds taken by all map tasks=5093376
Total megabyte-milliseconds taken by all reduce tasks=2383872
Map-Reduce Framework
Map input records=8
Map output records=8000
Map output bytes=95120
Map output materialized bytes=111132
Input split bytes=192
Combine input records=0
Combine output records=0
Reduce input groups=3996
Reduce shuffle bytes=111132
Reduce input records=8000
Reduce output records=4
Spilled Records=16000
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=87
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=849870848
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=68
File Output Format Counters
Bytes Written=36
D:\hadoop\sbin>hadoop fs -cat /pgm2OP/part-r-00000
0,0,19.0
0,1,22.0
1,0,43.0
1,1,50.0
29°C
Mostly cloudy
Search
ENG
IN
19:36
25-04-2025

Experiment 3

Aim: Develop a Map Reduce program that mines weather data and displays appropriate messages indicating the weather conditions of the day.

Code:

WeatherDriver:

```
package weather;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
public class WeatherDriver {
   public static void main(String[] args) throws Exception {
      if (args.length != 2) {
         System.err.println("Usage: WeatherDriver <input path> <output path>");
         System.exit(-1);
      }
      Configuration conf = new Configuration();
      Job job = Job.getInstance(conf, "Weather Condition Analysis");
      job.setJarByClass(WeatherDriver.class);
      job.setMapperClass(WeatherMapper.class);
      job.setReducerClass(WeatherReducer.class);
      job.setOutputKeyClass(Text.class);
      job.setOutputValueClass(Text.class);
      FileInputFormat.addInputPath(job, new Path(args[0]));
      FileOutputFormat.setOutputPath(job, new Path(args[1]));
      System.exit(job.waitForCompletion(true) ? 0 : 1);
   }}
```

WeatherMapper:

```
package weather;

import java.io.IOException;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.mapreduce.Mapper;

public class WeatherMapper extends Mapper<LongWritable, Text, Text, Text> {

   public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException {

      // Skip header line

      if (key.get() == 0 && value.toString().contains("City")) {

         return;

      }

     String[] fields = value.toString().split(",");

      if (fields.length == 3) {

         String city = fields[0];

         String temperatureStr = fields[2];

         try {

            int temp = Integer.parseInt(temperatureStr);

            String condition;

            if (temp >= 35) {

               condition = "Hot Day";

            } else if (temp <= 15) {

               condition = "Cold Day";

            } else if (temp >= 16 && temp <= 25) {

               condition = "Pleasant Day";

            } else {

               condition = "Moderate Day";

            }
```

```
            context.write(new Text(city), new Text(condition));
            } catch (NumberFormatException e) {
                // Ignore malformed lines
            }
        }
    }
}
```

WeatherReducer:

```
package weather;
import java.io.IOException;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class WeatherReducer extends Reducer<Text, Text, Text, Text> {
    public void reduce(Text key, Iterable<Text> values, Context context) throws IOException,
InterruptedException {
        // Just write the city and condition (1 value only per city)
        for (Text val : values) {
            context.write(key, val);
        }
    }
}
```

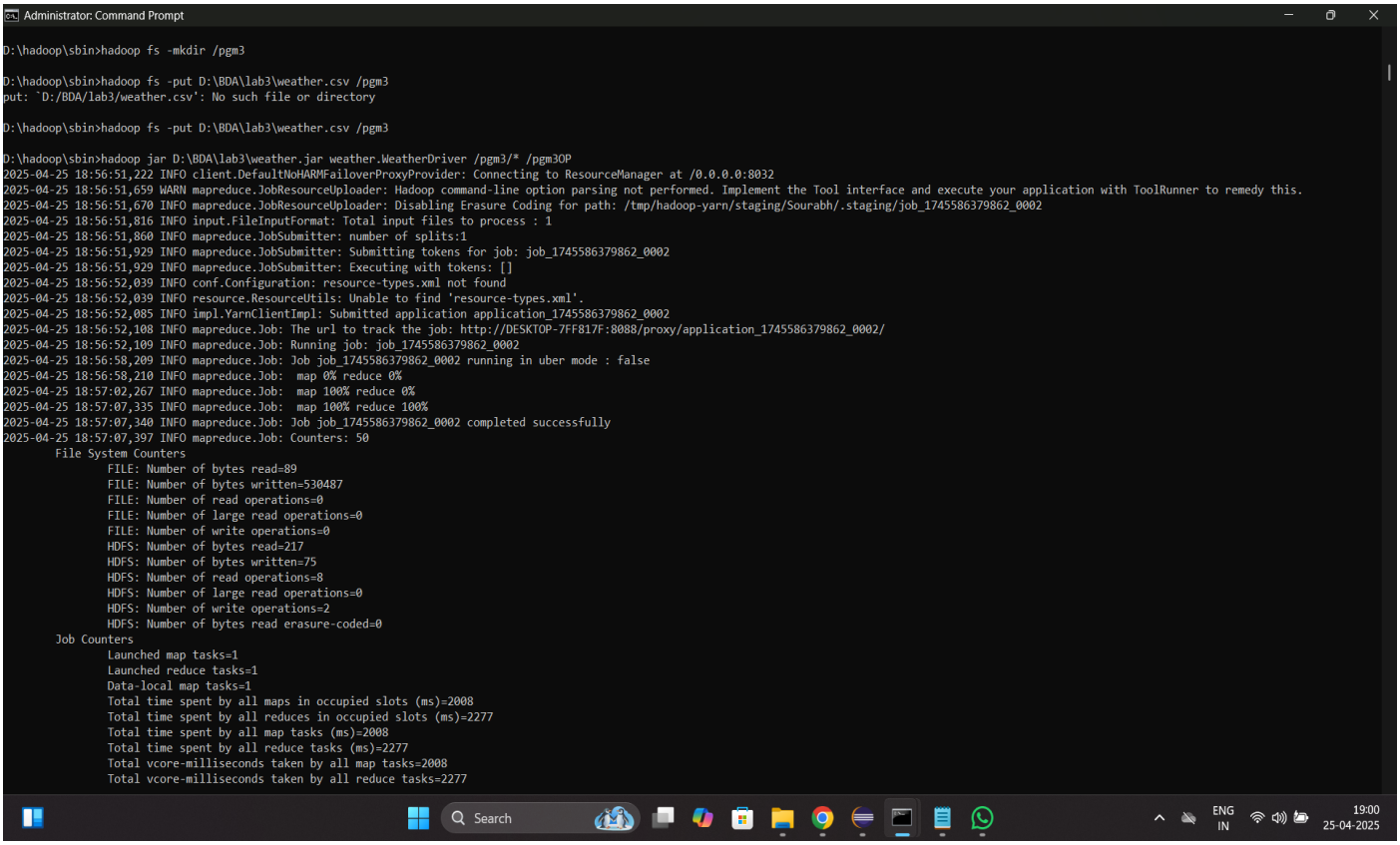
Commands for Execution

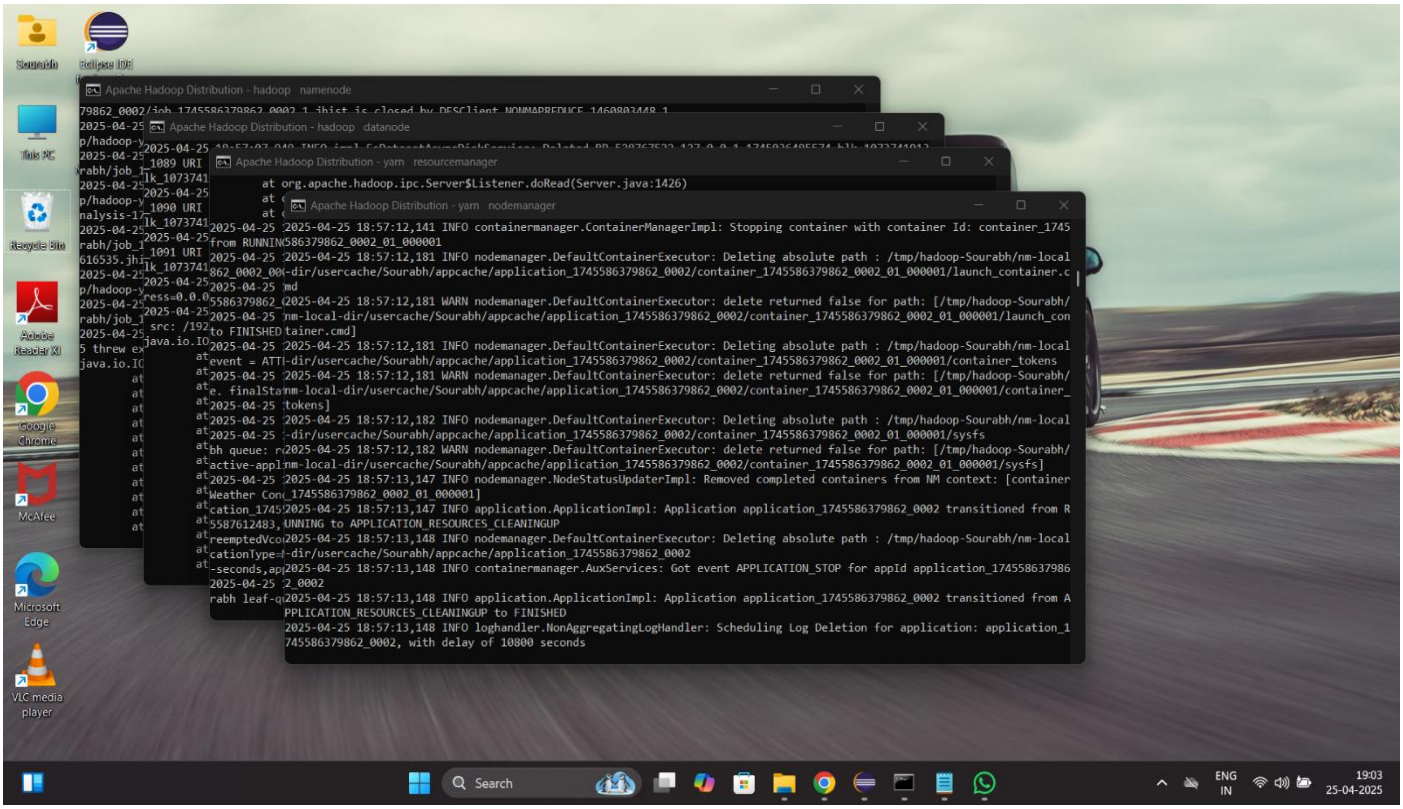
1. D:\hadoop\sbin>start-all.cmd
2. D:\hadoop\sbin>hadoop fs -mkdir /pgm3
3. D:\hadoop\sbin>hadoop fs -put D:\hadoop_experiments\weather.csv /pgm3
4. D:\hadoop\sbin>hadoop jar D:\hadoop_experiments\Weather.jar weather.WeatherDriver /pgm3 */pgm3OP
5. D:\hadoop\sbin>hadoop fs -cat /pgm3OP/part-r-00000

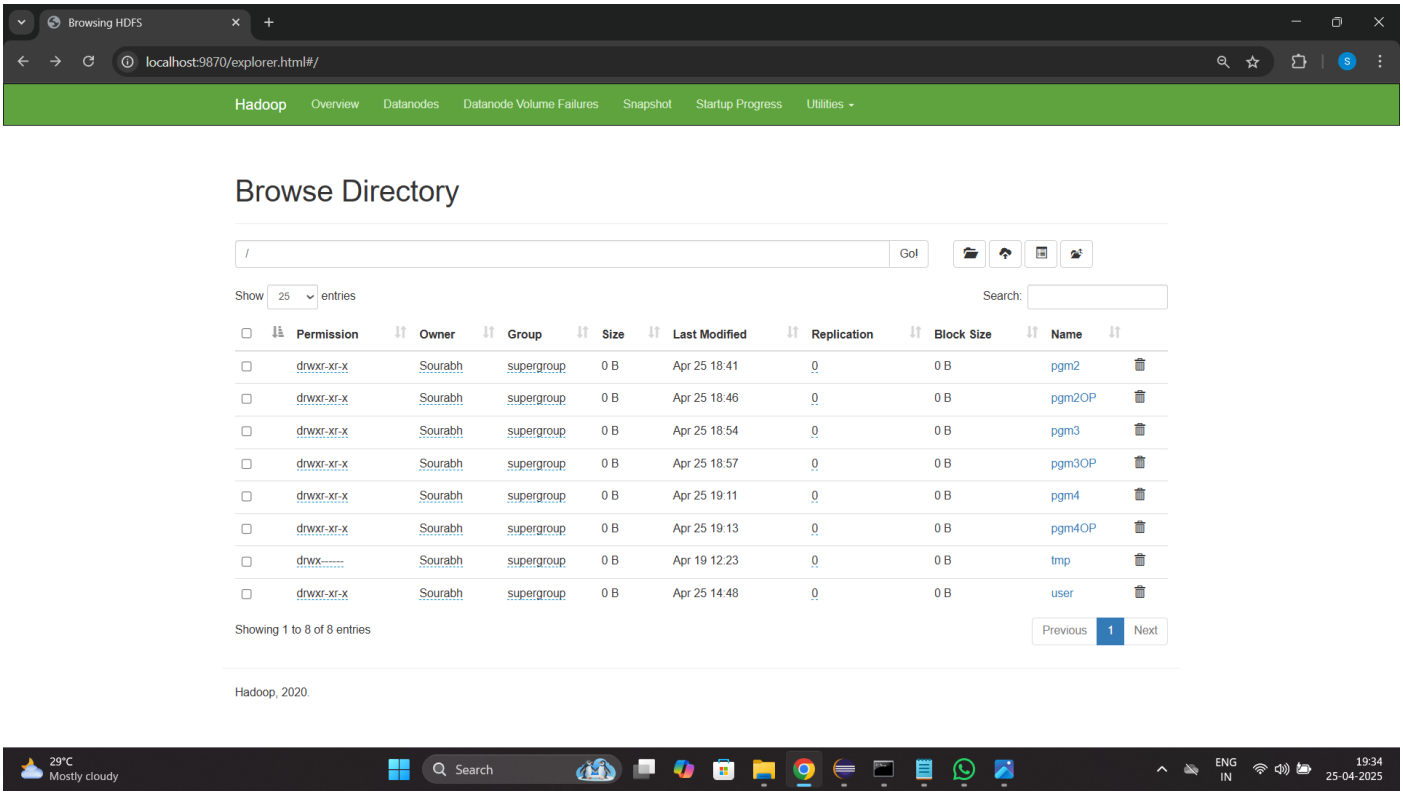
Weather input file

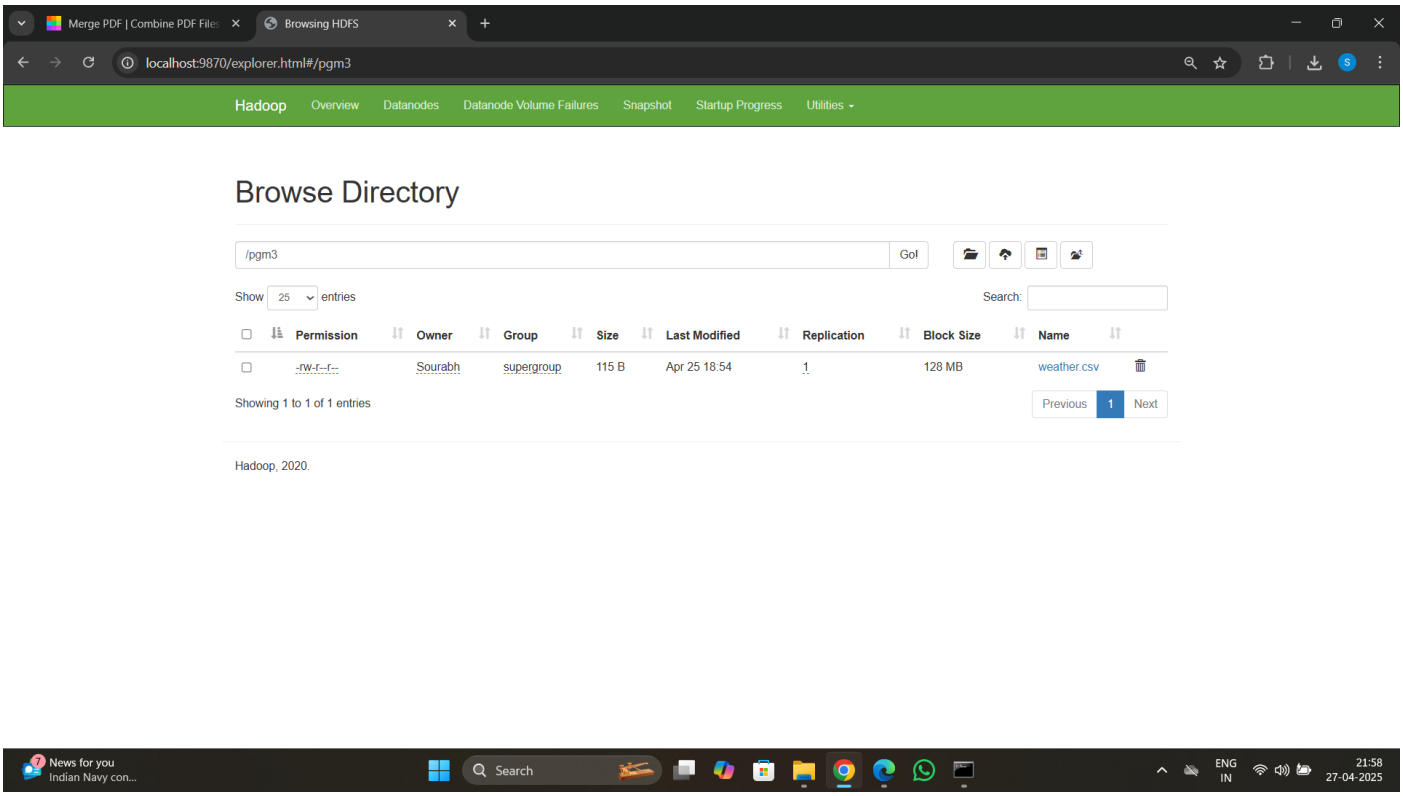
City	Date	Temperatu
Belgaum	16-04-2025	38
Mumbai	16-04-2025	29
Shimla	16-04-2025	12
Bengaluru	16-04-2025	22

Output:

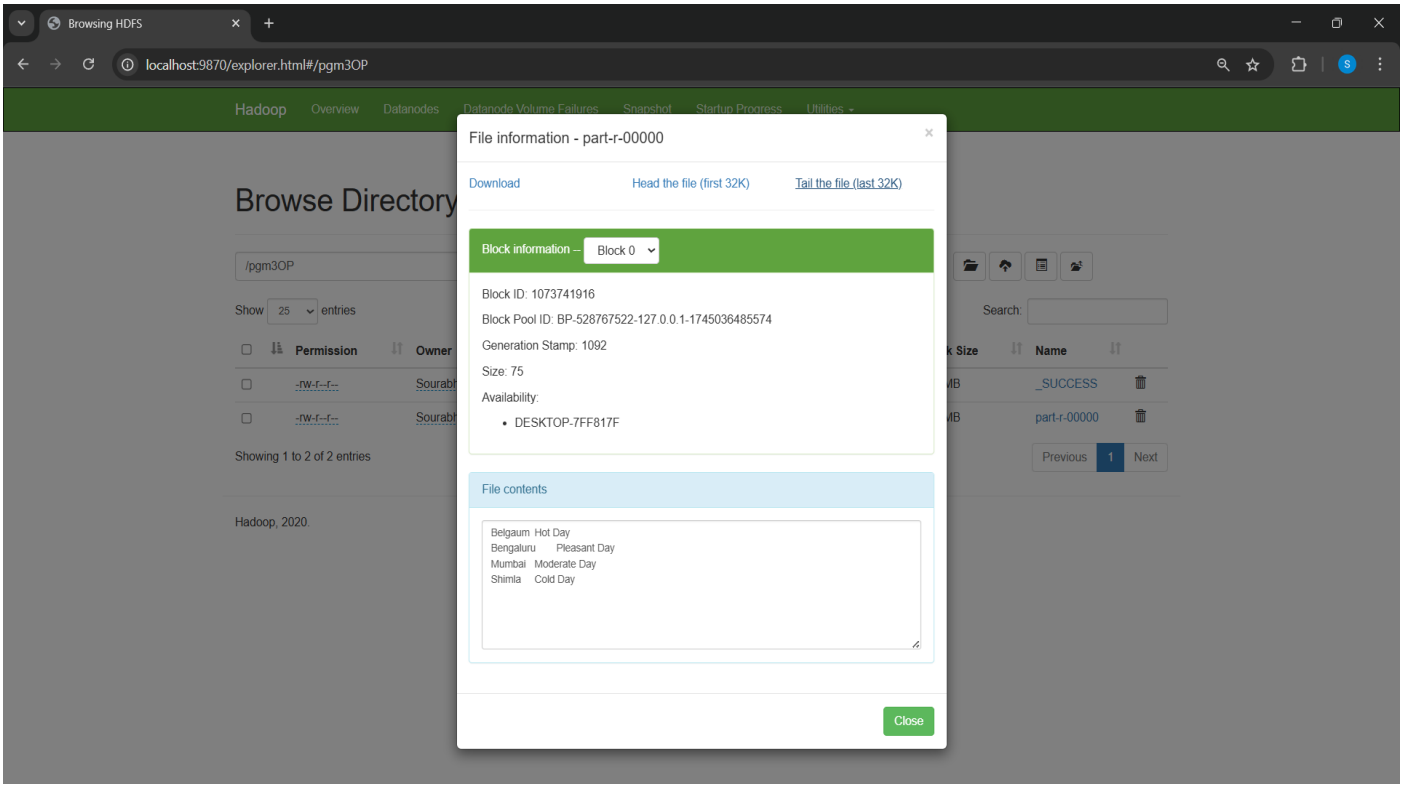


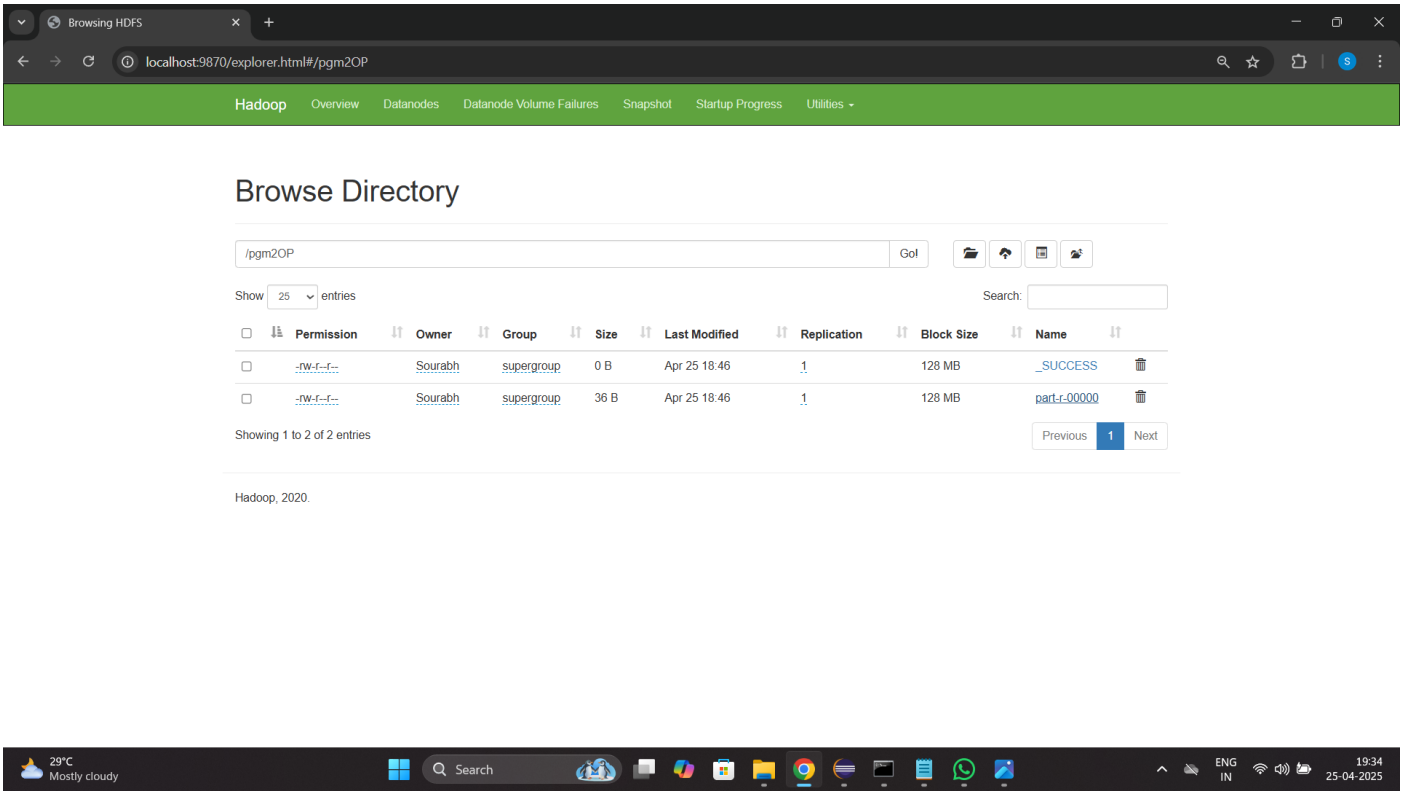

Apache Hadoop Distribution - hadoop namenode
Apache Hadoop Distribution - hadoop datanode
Apache Hadoop Distribution - yarn resourcemanager
Apache Hadoop Distribution - yarn nodemanager
at org.apache.hadoop.ipc.Server$Listener.doRead(Server.java:1426)
2025-04-25 18:57:12,141 INFO containermanager.ContainerManagerImpl: Stopping container with container Id: container_1745
2025-04-25 18:57:12,181 INFO nodemanager.DefaultContainerExecutor: Deleting absolute path : /tmp/hadoop-Sourabh/nm-local
2025-04-25 18:57:12,181 WARN nodemanager.DefaultContainerExecutor: delete returned false for path: [/tmp/hadoop-Sourabh/
2025-04-25 18:57:12,181 INFO nodemanager.DefaultContainerExecutor: Deleting absolute path : /tmp/hadoop-Sourabh/nm-local
2025-04-25 18:57:12,181 WARN nodemanager.DefaultContainerExecutor: delete returned false for path: [/tmp/hadoop-Sourabh/
2025-04-25 18:57:12,182 INFO nodemanager.DefaultContainerExecutor: Deleting absolute path : /tmp/hadoop-Sourabh/nm-local
2025-04-25 18:57:12,182 WARN nodemanager.DefaultContainerExecutor: delete returned false for path: [/tmp/hadoop-Sourabh/
2025-04-25 18:57:13,147 INFO nodemanager.NodeStatusUpdaterImpl: Removed completed containers from NM context: [container
2025-04-25 18:57:13,147 INFO application.ApplicationImpl: Application application_1745586379862_0002 transitioned from R
2025-04-25 18:57:13,148 INFO nodemanager.DefaultContainerExecutor: Deleting absolute path : /tmp/hadoop-Sourabh/nm-local
2025-04-25 18:57:13,148 INFO containermanager.AuxServices: Got event APPLICATION_STOP for appId application_1745586379862
2025-04-25 18:57:13,148 INFO application.ApplicationImpl: Application application_1745586379862_0002 transitioned from A
2025-04-25 18:57:13,148 INFO loghandler.NonAggregatingLogHandler: Scheduling Log Deletion for application: application_1
1745586379862_0002, with delay of 10800 seconds
Search
ENG
IN
19:03
25-04-2025


Browsing HDFS
localhost:9870/explorer.html#/
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:41 0 0 B pgm2
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:46 0 0 B pgm2OP
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:54 0 0 B pgm3
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:57 0 0 B pgm3OP
drwxr-xr-x Sourabh supergroup 0 B Apr 25 19:11 0 0 B pgm4
drwxr-xr-x Sourabh supergroup 0 B Apr 25 19:13 0 0 B pgm4OP
drwx------ Sourabh supergroup 0 B Apr 19 12:23 0 0 B tmp
drwxr-xr-x Sourabh supergroup 0 B Apr 25 14:48 0 0 B user
Showing 1 to 8 of 8 entries
Previous
1
Next
Hadoop, 2020.
29°C
Mostly cloudy
Search
ENG
IN
19:34
25-04-2025

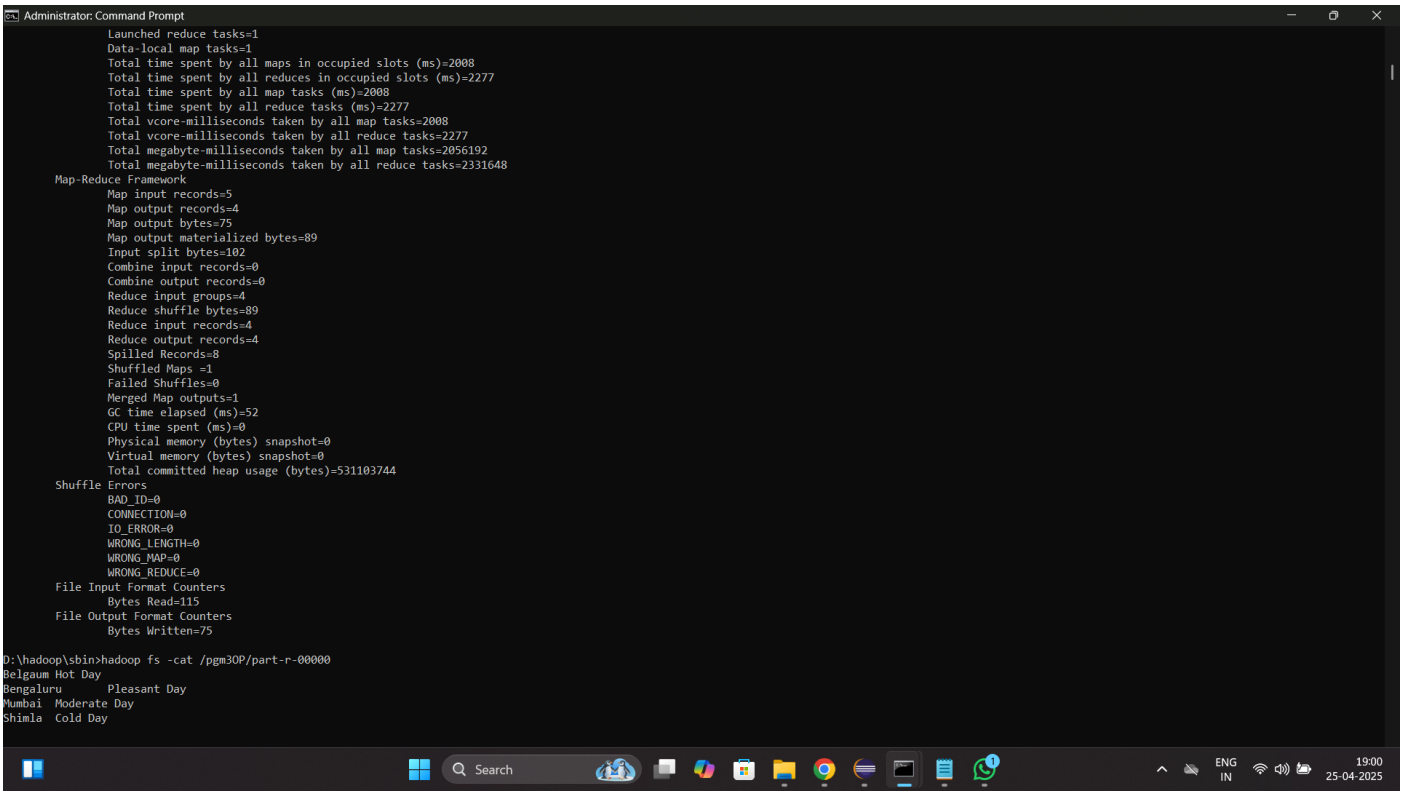

Merge PDF | Combine PDF Files
Browsing HDFS
localhost:9870/explorer.html#/pgm3
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm3
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
Sourabh
supergroup
115 B
Apr 25 18:54
1
128 MB
weather.csv
Showing 1 to 1 of 1 entries
Previous
1
Next
Hadoop, 2020.

News for you
Indian Navy con...
Search
ENG
IN
21:58
27-04-2025


Browsing HDFS
localhost:9870/explorer.html#/pgm3OP
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm3OP
Show 25 entries
Search:
Permission
Owner
Name
-rw-r--r--
Sourabh
_SUCCESS
-rw-r--r--
Sourabh
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.
File information - part-r-00000
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information –
Block 0
Block ID: 1073741916
Block Pool ID: BP-528767522-127.0.0.1-1745036485574
Generation Stamp: 1092
Size: 75
Availability:
DESKTOP-7FF817F
File contents
Belgaum Hot Day
Bengaluru Pleasant Day
Mumbai Moderate Day
Shimla Cold Day
Close


Browsing HDFS
localhost:9870/explorer.html#/pgm2OP
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm2OP
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
Sourabh
supergroup
0 B
Apr 25 18:46
1
128 MB
_SUCCESS
-rw-r--r--
Sourabh
supergroup
36 B
Apr 25 18:46
1
128 MB
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.

29°C
Mostly cloudy
Search
ENG
IN
19:34
25-04-2025


Administrator: Command Prompt
Launched reduce tasks=1
Data-local map tasks=1
Total time spent by all maps in occupied slots (ms)=2008
Total time spent by all reduces in occupied slots (ms)=2277
Total time spent by all map tasks (ms)=2008
Total time spent by all reduce tasks (ms)=2277
Total vcore-milliseconds taken by all map tasks=2008
Total vcore-milliseconds taken by all reduce tasks=2277
Total megabyte-milliseconds taken by all map tasks=2056192
Total megabyte-milliseconds taken by all reduce tasks=2331648
Map-Reduce Framework
Map input records=5
Map output records=4
Map output bytes=75
Map output materialized bytes=89
Input split bytes=102
Combine input records=0
Combine output records=0
Reduce input groups=4
Reduce shuffle bytes=89
Reduce input records=4
Reduce output records=4
Spilled Records=8
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=52
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=531103744
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=115
File Output Format Counters
Bytes Written=75
D:\hadoop\sbin>hadoop fs -cat /pgm3OP/part-r-00000
Belgaum Hot Day
Bengaluru Pleasant Day
Mumbai Moderate Day
Shimla Cold Day
Search
ENG
IN
19:00
25-04-2025

Experiment 4

Aim: Develop a MapReduce program to find the tags associated with each movie by analyzing movie lens data.

Code:

TagDriver:

```
package movie;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
public class TagDriver {
   public static void main(String[] args) throws Exception {
      Configuration conf = new Configuration();
      Job job = Job.getInstance(conf, "Movie Tags");
      job.setJarByClass(TagDriver.class);
      job.setMapperClass(TagMapper.class);
      job.setReducerClass(TagReducer.class);
      job.setOutputKeyClass(Text.class);
      job.setOutputValueClass(Text.class);
      // args[0] = input path, args[1] = output path
      FileInputFormat.addInputPath(job, new Path(args[0]));
      FileOutputFormat.setOutputPath(job, new Path(args[1]));

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

TagMapper:

```
package movie;


import java.io.IOException;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
public class TagMapper extends Mapper<LongWritable, Text, Text, Text> {
   public void map(LongWritable key, Text value, Context context)
         throws IOException, InterruptedException {
      // Skip header
      if (key.get() == 0 && value.toString().contains("userId")) return;
      String[] parts = value.toString().split(",");
      if (parts.length >= 4) {
         String movieId = parts[1].trim();
         String tag = parts[2].trim();
         context.write(new Text(movieId), new Text(tag));
      }
   }
}
```

TagReducer:

```
package movie;


import java.io.IOException;
import java.util.HashSet;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
public class TagReducer extends Reducer<Text, Text, Text, Text> {
   public void reduce(Text key, Iterable<Text> values, Context context)
         throws IOException, InterruptedException {
      HashSet<String> uniqueTags = new HashSet<>();
      for (Text tag : values) {
         uniqueTags.add(tag.toString());
      }
      StringBuilder tagList = new StringBuilder();
      for (String tag : uniqueTags) {
         if (tagList.length() > 0) tagList.append(", ");
         tagList.append(tag);
      }
      context.write(key, new Text(tagList.toString()));
   }
}
```

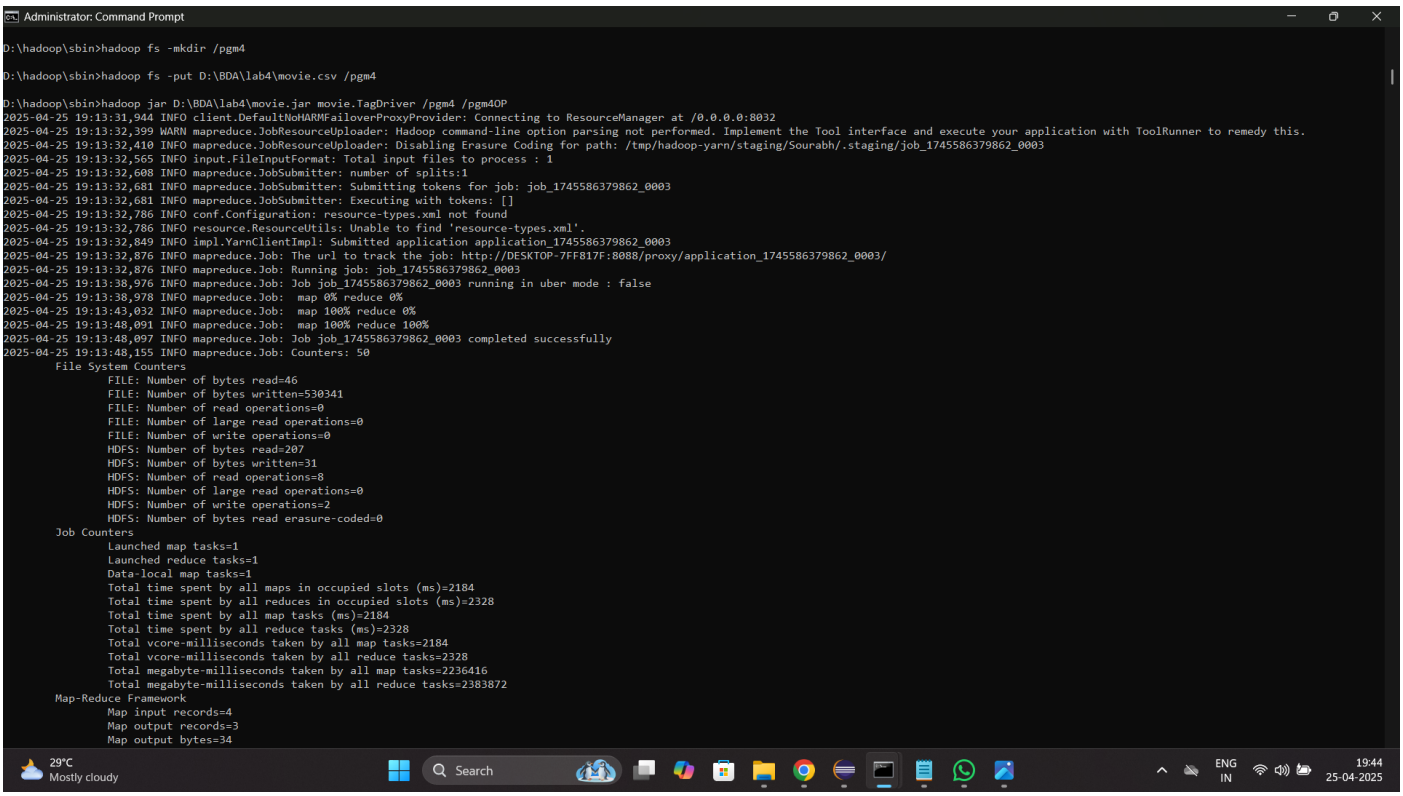
Commands for Execution

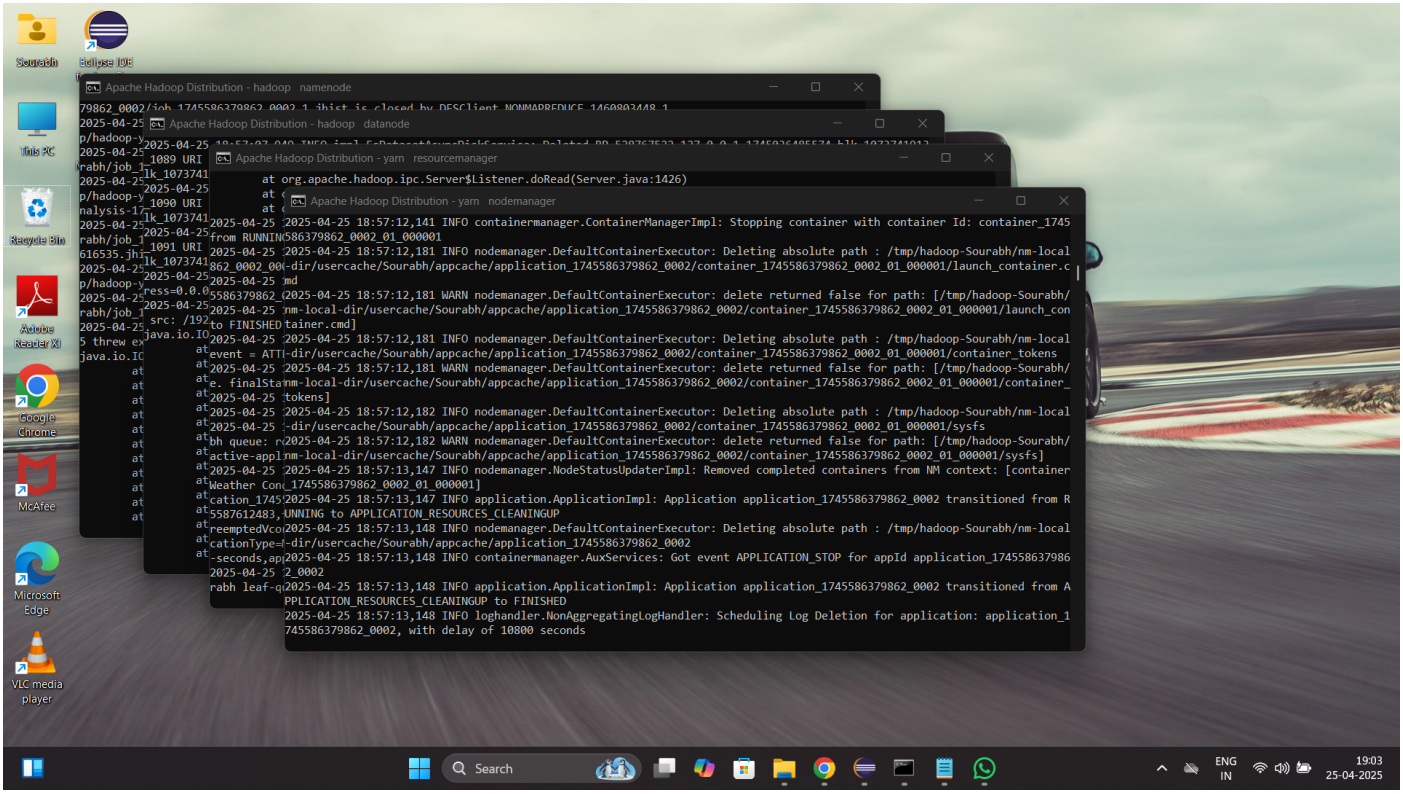
1. D:\hadoop\sbin>start-all.cmd
2. D:\hadoop\sbin>hadoop fs -mkdir /pgm4
3. D:\hadoop\sbin>hadoop fs -put D:\hadoop_experiments\movie.csv /pgm4
4. D:\hadoop\sbin>hadoop jar D:\hadoop_experiments\movietags.jar movie.TagDriver /pgm4 /pgm4OP
5. D:\hadoop\sbin>hadoop fs -cat /pgm4OP/part-r-00000

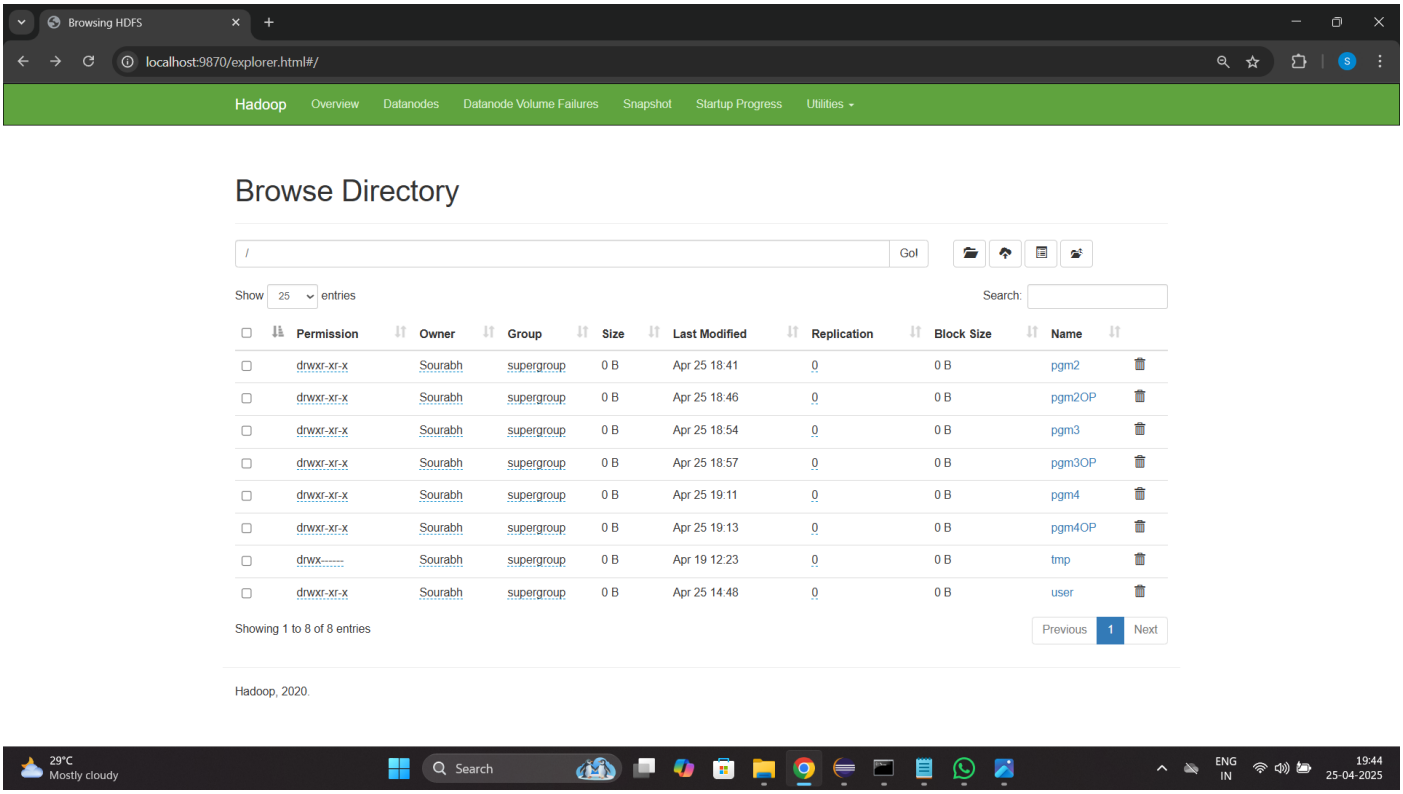
Movie Input file

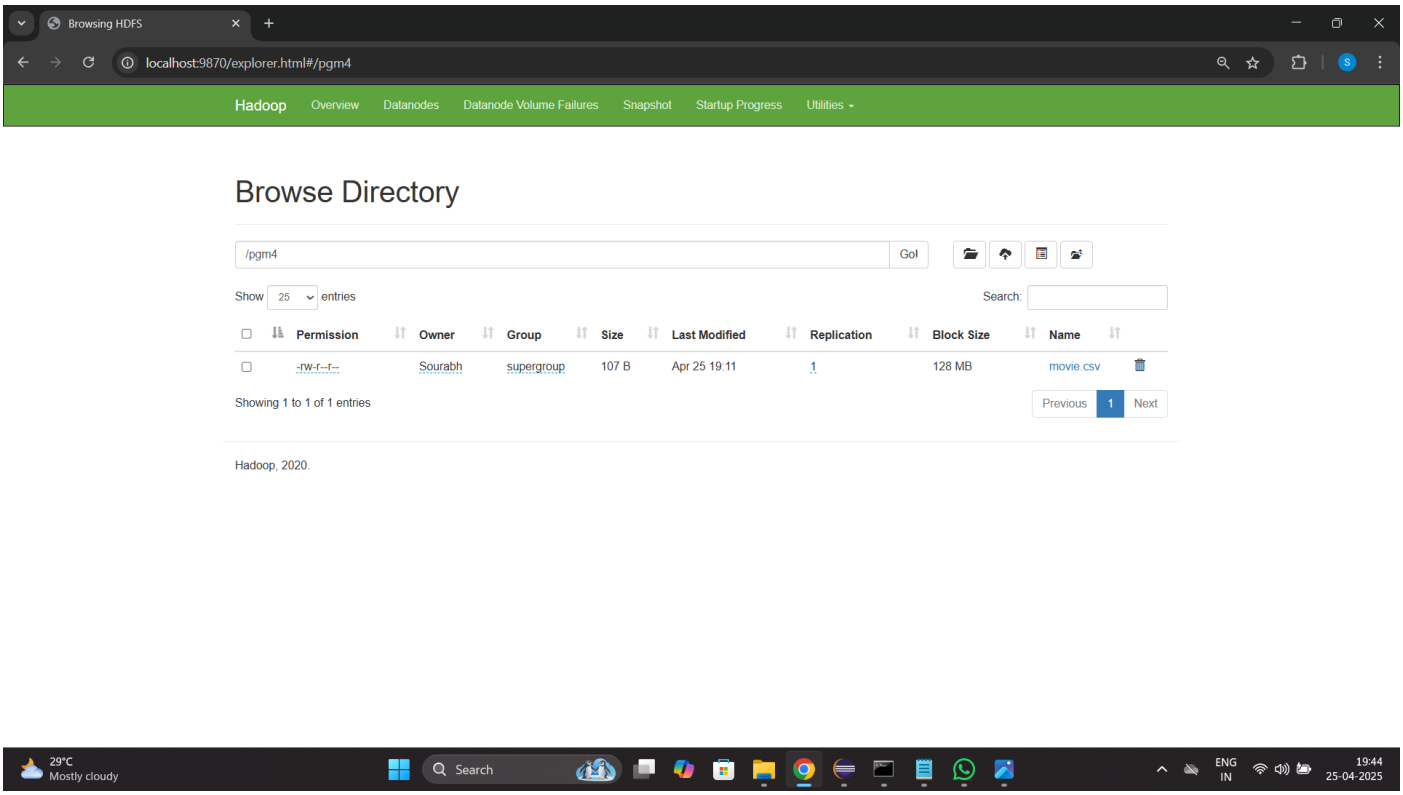
userId	movieId	tag	timestamp
15	339	action	1215184630
20	858	classic	1215208005
20	858	sci-fi	1215208010

Output:

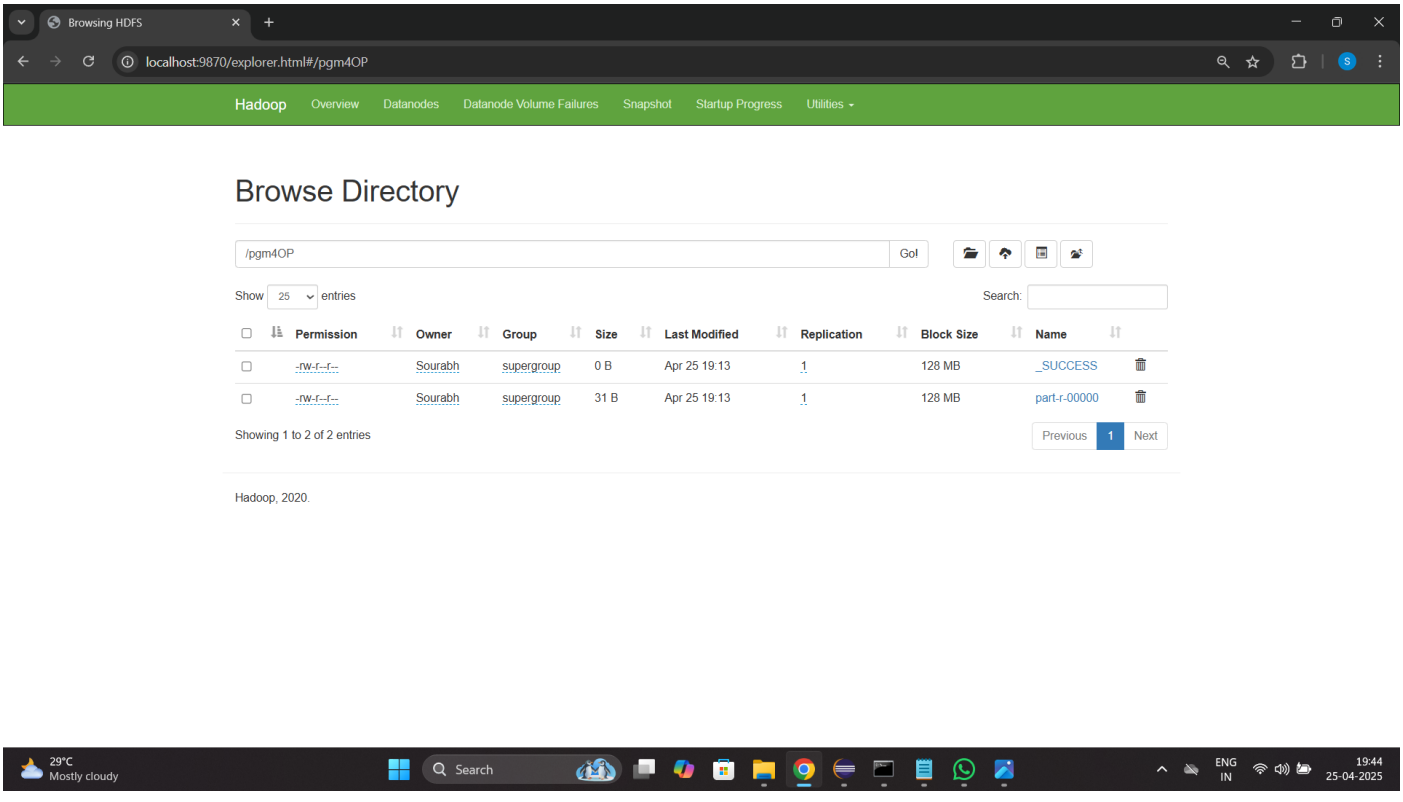



Apache Hadoop Distribution - hadoop namenode
Apache Hadoop Distribution - hadoop datanode
Apache Hadoop Distribution - yarn resourcemanager
Apache Hadoop Distribution - yarn nodemanager
Search
ENG
IN
19:03
25-04-2025

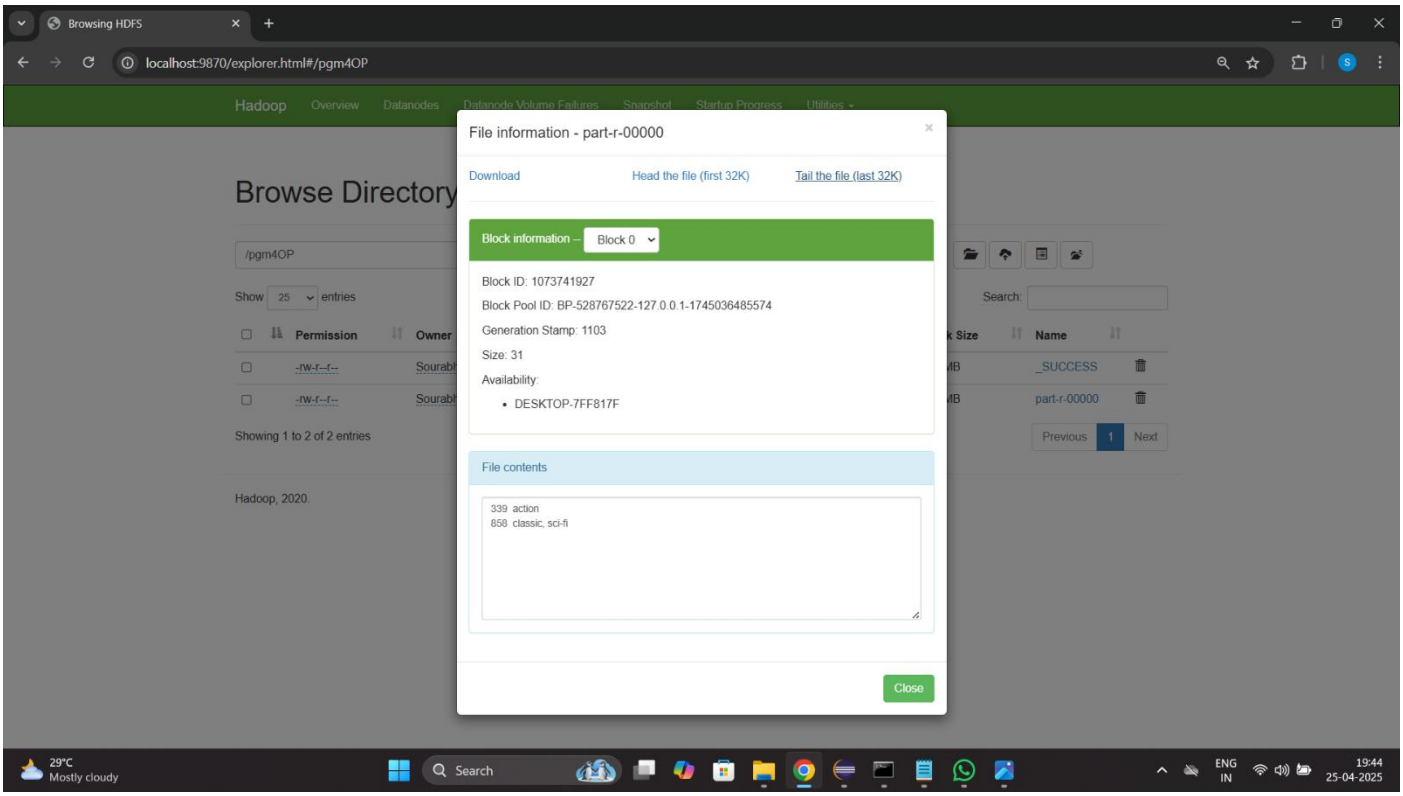

Browsing HDFS
localhost:9870/explorer.html#/
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:41 0 0 B pgm2
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:46 0 0 B pgm2OP
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:54 0 0 B pgm3
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:57 0 0 B pgm3OP
drwxr-xr-x Sourabh supergroup 0 B Apr 25 19:11 0 0 B pgm4
drwxr-xr-x Sourabh supergroup 0 B Apr 25 19:13 0 0 B pgm4OP
drwx------ Sourabh supergroup 0 B Apr 19 12:23 0 0 B tmp
drwxr-xr-x Sourabh supergroup 0 B Apr 25 14:48 0 0 B user
Showing 1 to 8 of 8 entries
Previous
1
Next
Hadoop, 2020.
29°C
Mostly cloudy
Search
ENG
IN
19:44
25-04-2025

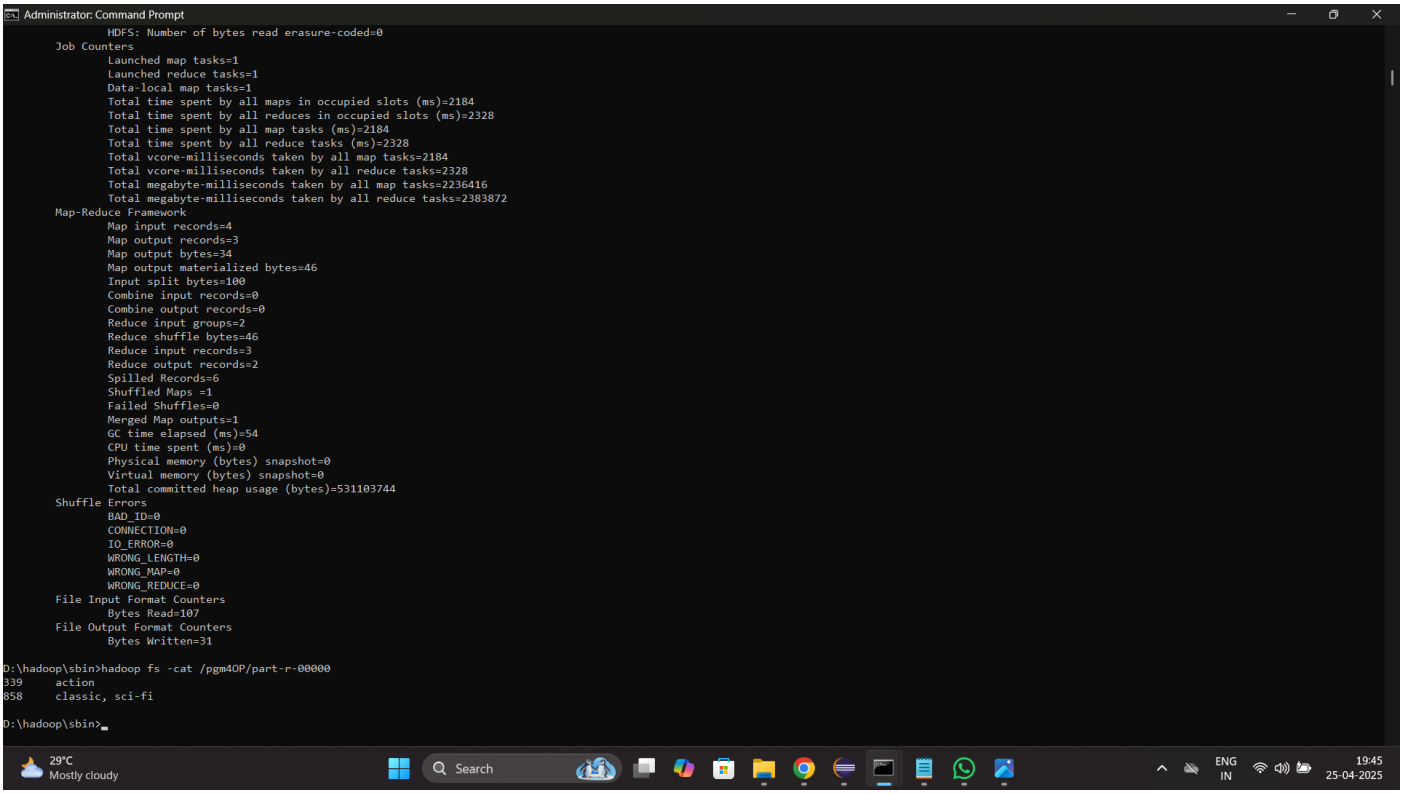

Browsing HDFS
localhost:9870/explorer.html#/pgm4
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm4
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
Sourabh
supergroup
107 B
Apr 25 19:11
1
128 MB
movie.csv
Showing 1 to 1 of 1 entries
Previous
1
Next
Hadoop, 2020.

29°C
Mostly cloudy
Search
ENG
IN
19:44
25-04-2025


Browsing HDFS
localhost:9870/explorer.html#/pgm4OP
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm4OP
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
Sourabh
supergroup
0 B
Apr 25 19:13
1
128 MB
_SUCCESS
-rw-r--r--
Sourabh
supergroup
31 B
Apr 25 19:13
1
128 MB
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.

29°C
Mostly cloudy
Search
ENG
IN
19:44
25-04-2025


Browsing HDFS
localhost:9870/explorer.html#/pgm4OP
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm4OP
Show 25 entries
Search:
Permission
Owner
Name
-rw-r--r--
_SUCCESS
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.
File information - part-r-00000
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information -- Block 0
Block ID: 1073741927
Block Pool ID: BP-528767522-127.0.0.1-1745036485574
Generation Stamp: 1103
Size: 31
Availability:
DESKTOP-7FF817F
File contents
339 action
858 classic, sci-fi
Close
29°C
Mostly cloudy
Search
ENG
IN
19:44
25-04-2025


Administrator: Command Prompt
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=1
Launched reduce tasks=1
Data-local map tasks=1
Total time spent by all maps in occupied slots (ms)=2184
Total time spent by all reduces in occupied slots (ms)=2328
Total time spent by all map tasks (ms)=2184
Total time spent by all reduce tasks (ms)=2328
Total vcore-milliseconds taken by all map tasks=2184
Total vcore-milliseconds taken by all reduce tasks=2328
Total megabyte-milliseconds taken by all map tasks=2236416
Total megabyte-milliseconds taken by all reduce tasks=2383872
Map-Reduce Framework
Map input records=4
Map output records=3
Map output bytes=34
Map output materialized bytes=46
Input split bytes=100
Combine input records=0
Combine output records=0
Reduce input groups=2
Reduce shuffle bytes=46
Reduce input records=3
Reduce output records=2
Spilled Records=6
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=54
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=531103744
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=107
File Output Format Counters
Bytes Written=31
D:\hadoop\sbin>hadoop fs -cat /pgm4OP/part-r-00000
339 action
858 classic, sci-fi
D:\hadoop\sbin>
29°C
Mostly cloudy
Search
ENG
IN
19:45
25-04-2025

Experiment 5

Aim: Implement Functions: Count – Sort – Limit – Skip – Aggregate using MongoDB

Code:

```
db.students.insertMany([
  { name: "Alice", age: 22, marks: 85 },
  { name: "Bob", age: 23, marks: 75 },
  { name: "Charlie", age: 22, marks: 95 },
  { name: "David", age: 24, marks: 65 },
  { name: "Eve", age: 23, marks: 70 },
  { name: "Frank", age: 22, marks: 80 }
])
```

1)count

```
db.students.countDocuments()
db.students.countDocuments({ marks: { $gt: 80 } })
```

2)sort

```
db.students.find().sort({ marks: 1 })  // ascending order
db.students.find().sort({ marks: -1 }) // descending order
```

3)limit // Get only 3 students

```
db.students.find().limit(3)
```

4)skip // Skip the first 2 students

```
db.students.find().skip(2)
```

5)aggregate

```
db.students.aggregate([
  {
    $group: {
      _id: "$age",
      averageMarks: { $avg: "$marks" },
      totalStudents: { $sum: 1 }
    }
  }
])
```

```
db.students.aggregate([
  { $match: { marks: { $gt: 80 } } },  // Filter students
  { $sort: { marks: -1 } }             // Sort by marks descending
])
```

Output:

```
test> db.students.insertMany([
...     { name: "Alice", age: 22, marks: 85 },
...     { name: "Bob", age: 23, marks: 75 },
...     { name: "Charlie", age: 22, marks: 95 },
...     { name: "David", age: 24, marks: 65 },
...     { name: "Eve", age: 23, marks: 70 },
...     { name: "Frank", age: 22, marks: 80 }
... ])
...
{
  acknowledged: true,
  insertedIds: {
    '0': ObjectId('681758fe495939eae5b5f899'),
    '1': ObjectId('681758fe495939eae5b5f89a'),
    '2': ObjectId('681758fe495939eae5b5f89b'),
    '3': ObjectId('681758fe495939eae5b5f89c'),
    '4': ObjectId('681758fe495939eae5b5f89d'),
    '5': ObjectId('681758fe495939eae5b5f89e')
  }
}
```

```
test> db.students.countDocuments()
6
```

```
test> db.students.find().limit(3)
[
  {
    _id: ObjectId('681758fe495939eae5b5f899'),
    name: 'Alice',
    age: 22,
    marks: 85
  },
  {
    _id: ObjectId('681758fe495939eae5b5f89a'),
    name: 'Bob',
    age: 23,
    marks: 75
  },
  {
    _id: ObjectId('681758fe495939eae5b5f89b'),
    name: 'Charlie',
    age: 22,
    marks: 95
  }
]
```

```
test> db.students.countDocuments()
... db.students.countDocuments({ marks: { $gt: 80 } })
2
test> db.students.find().sort({ marks: 1 })
... db.students.find().sort({ marks: -1 })
[
  {
    _id: ObjectId('681758fe495939eae5b5f89b'),
    name: 'Charlie',
    age: 22,
    marks: 95
  },
  {
    _id: ObjectId('681758fe495939eae5b5f899'),
    name: 'Alice',
    age: 22,
    marks: 85
  },
  {
    _id: ObjectId('681758fe495939eae5b5f89e'),
    name: 'Frank',
    age: 22,
    marks: 80
  },
  {
    _id: ObjectId('681758fe495939eae5b5f89a'),
    name: 'Bob',
    age: 23,
    marks: 75
  },
  {
    _id: ObjectId('681758fe495939eae5b5f89d'),
    name: 'Eve',
    age: 23,
    marks: 70
  },
  {
    _id: ObjectId('681758fe495939eae5b5f89c'),
    name: 'David',
    age: 24,
    marks: 65
  }
]
```

```
test> db.students.find().skip(2)
[
  {
    _id: ObjectId('681758fe495939eae5b5f89b'),
    name: 'Charlie',
    age: 22,
    marks: 95
  },
  {
    _id: ObjectId('681758fe495939eae5b5f89c'),
    name: 'David',
    age: 24,
    marks: 65
  },
  {
    _id: ObjectId('681758fe495939eae5b5f89d'),
    name: 'Eve',
    age: 23,
    marks: 70
  },
  {
    _id: ObjectId('681758fe495939eae5b5f89e'),
    name: 'Frank',
    age: 22,
    marks: 80
  }
]
```

```
test> db.students.aggregate([
...   {
...     $group: {
...       _id: "$age",
...       averageMarks: { $avg: "$marks" },
...       totalStudents: { $sum: 1 }
...     }
...   }
... ])
... db.students.aggregate([
...   { $match: { marks: { $gt: 80 } } },  // Filter students
...   { $sort: { marks: -1 } }             // Sort by marks descending
... ])
...
[
  {
    _id: ObjectId('681758fe495939eae5b5f89b'),
    name: 'Charlie',
    age: 22,
    marks: 95
  },
  {
    _id: ObjectId('681758fe495939eae5b5f899'),
    name: 'Alice',
    age: 22,
    marks: 85
  }
]
```

Experiment 6

Aim: Develop Pig Latin scripts to sort, group, join, project, and filter the data.

Apache Pig Installation Steps:

Software Requirements:

- Linux/Ubuntu OS
- Hadoop (already installed and configured)
- Java (JDK 8 or later)

Step 1: Download Apache Pig

- Go to the official Apache website and Download latest version of Apache pig
- Apache pig 0.17.0 (Extract the file)
- System variable>>New System Variable>>set as **PIG_HOME** with path (**D:\pig**)
- System variable>>Path>>New>>set **%PIG_HOME%\bin**
- User variable>>Path>>New>>set **%PIG_HOME%\bin**

Step 2: Verify Pig Installation

"pig -x local"

Students.txt input File data:

1,John,21,101
2,Alice,22,102
3,Bob,20,101
4,David,23,103
5,Eve,22,102
6,Frank,21,104
7,Grace,20,103
8,Hannah,24,101

department.txt input File data:

101,Computer Science
102,Mechanical Engineering
103,Civil Engineering
104,Electrical Engineering

Code:

LOAD:

```
grunt>students = LOAD 'D:/BDA/lab6/students.txt'
>>>USING PigStorage(',')
>> AS (student_id:int, name: chararray, age:int, department_id:int);

grunt> departments = LOAD 'D:/BDA/lab6/departments.txt'
>>> USING PigStorage(',')
>> AS (department_id:int, department_name: chararray);
```

1) SORT

```
grunt>sorted_students = ORDER students BY age ASC;
grunt>DUMP sorted_students;
```

2)FILTER

```
grunt>filtered_students = FILTER students BY age > 20;
grunt>DUMP filtered_students;
```

3)JOIN

```
grunt>joined_data = JOIN students BY department_id, departments BY department_id;
grunt>DUMP  joined_data;
```

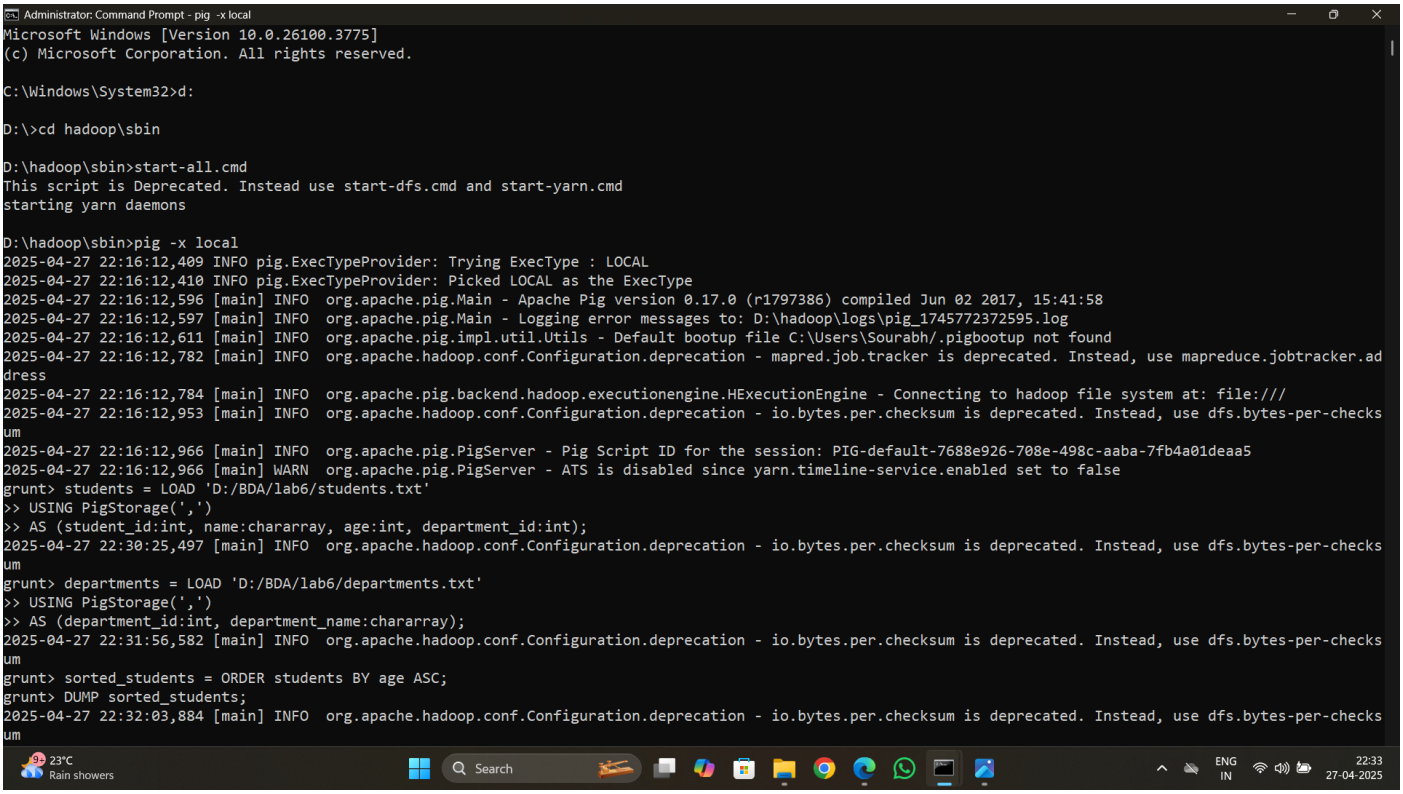
4)PROJECT

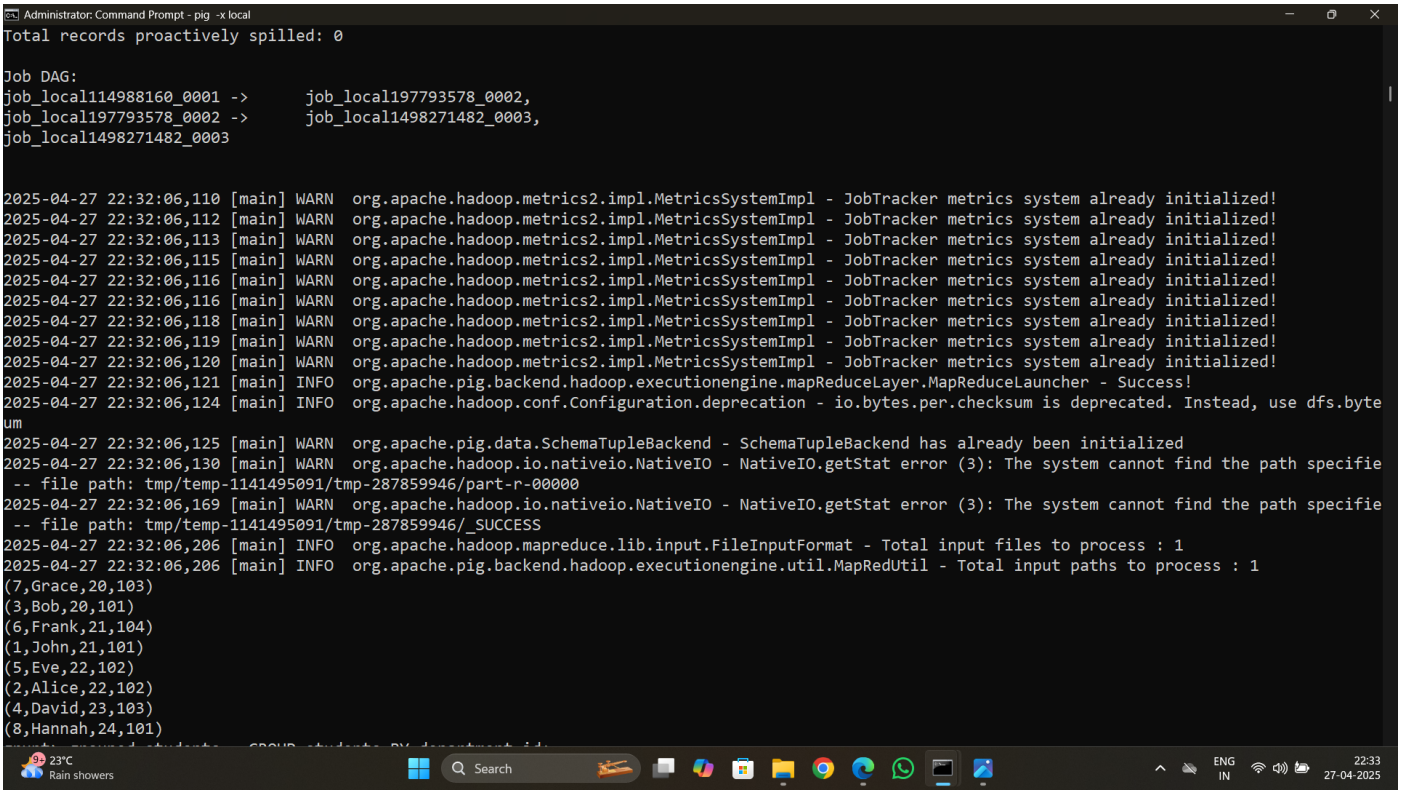
```
grunt>projected_data = FOREACH joined_data GENERATE students::name,
departments::department_name;
grunt>DUMP  projected_data;
```

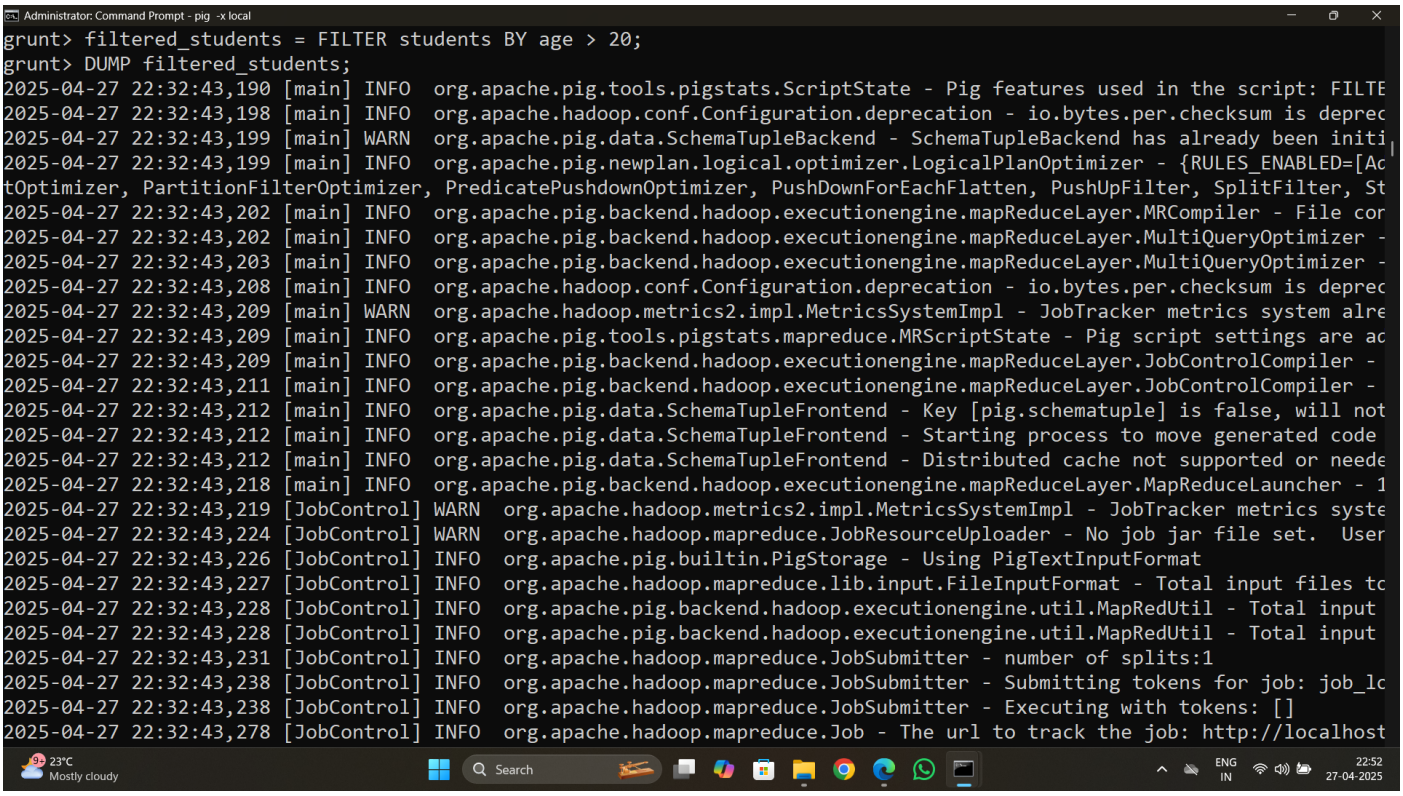
5)GROUP

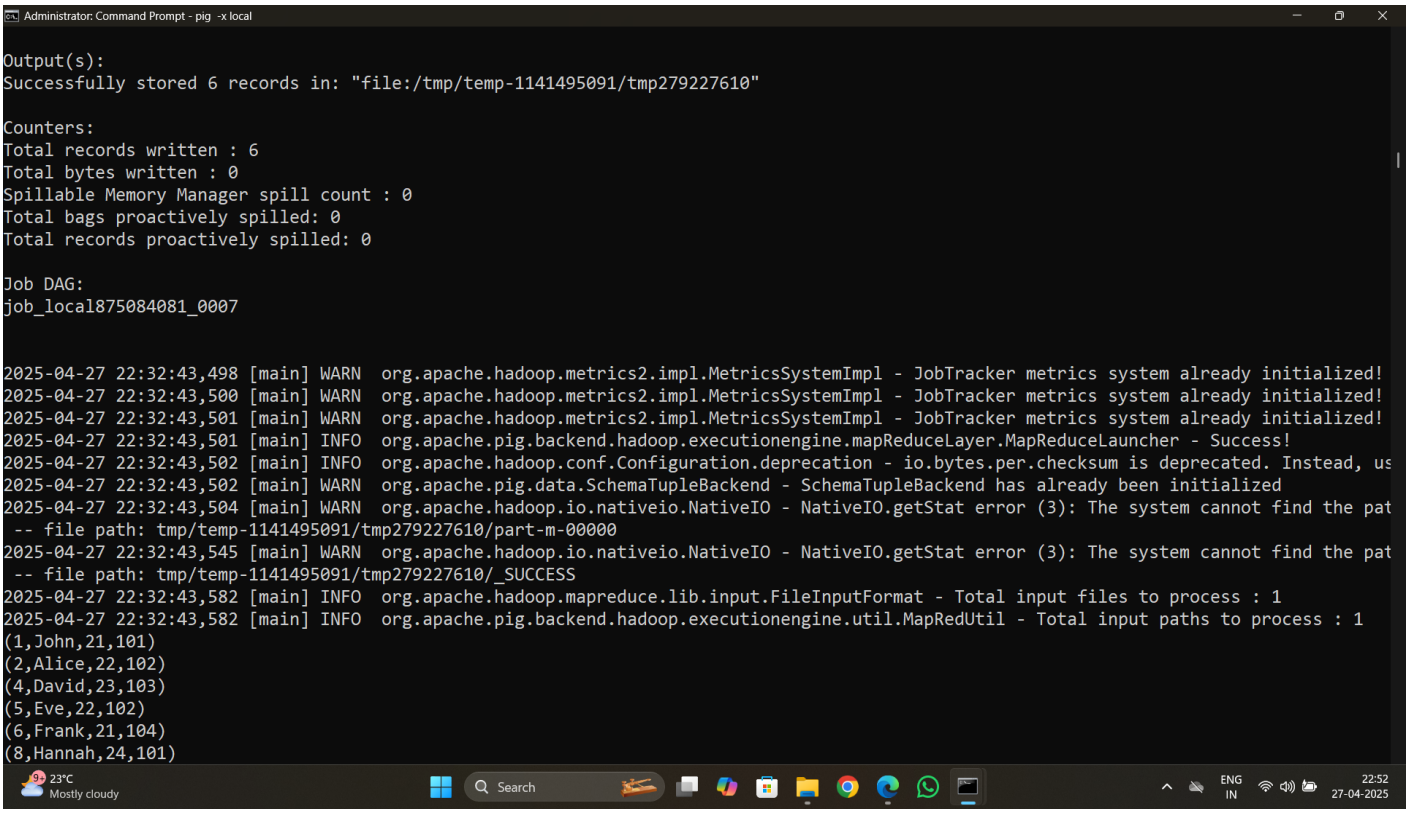
```
grunt>grouped_students = GROUP students BY department_id;
grunt>DUMP grouped_students;
```

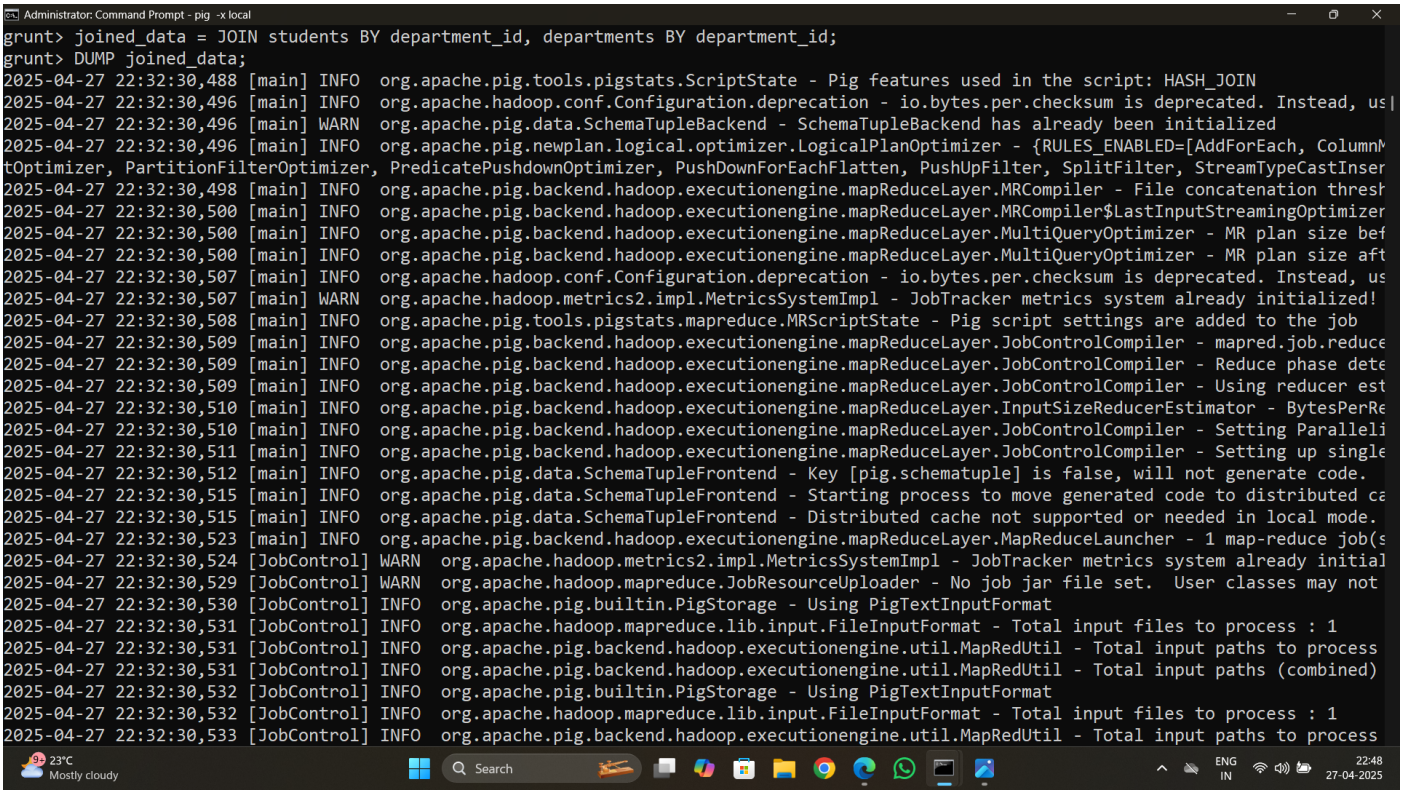
Output:

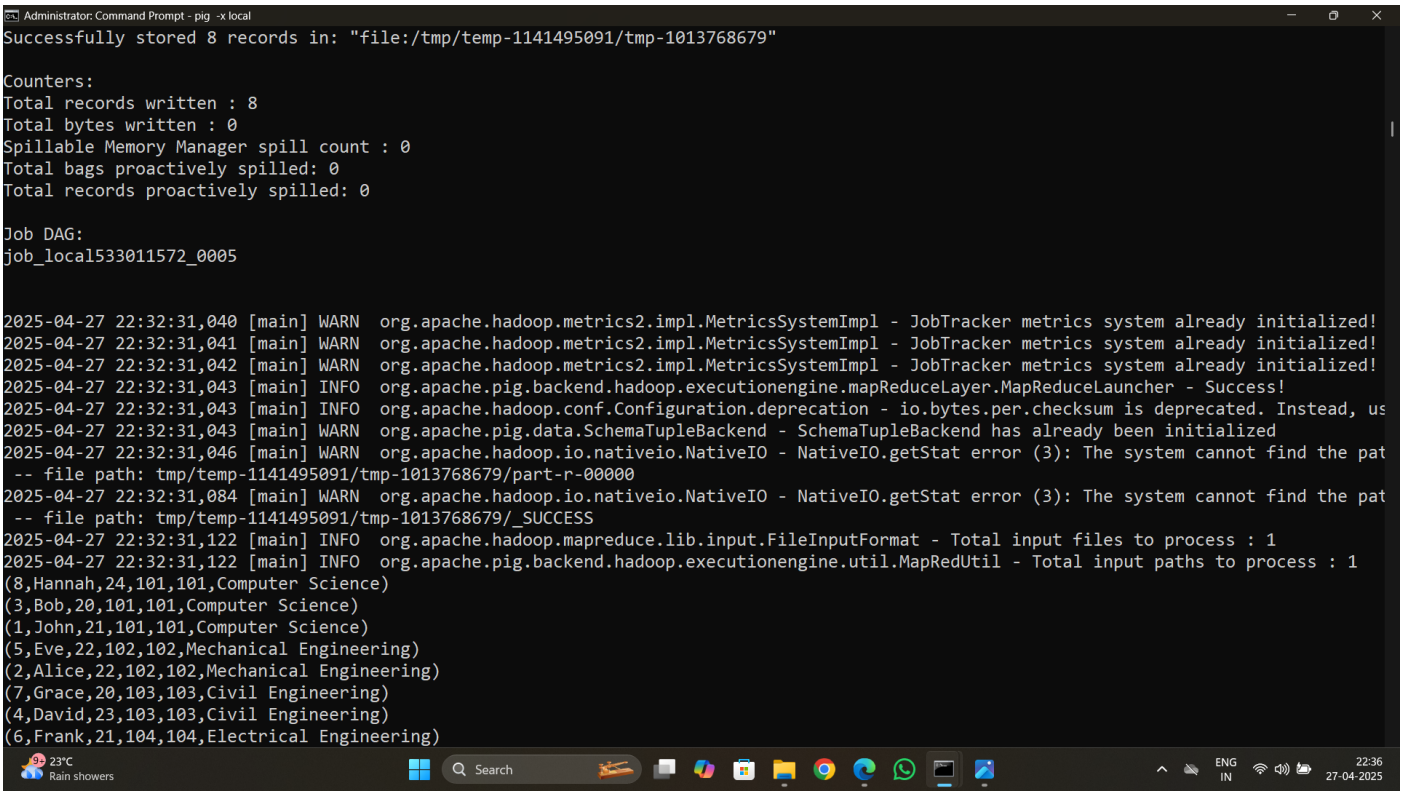

Administrator: Command Prompt - pig -x local
Microsoft Windows [Version 10.0.26100.3775]
(c) Microsoft Corporation. All rights reserved.
C:\Windows\System32>d:
D:\>cd hadoop\sbin
D:\hadoop\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
D:\hadoop\sbin>pig -x local
2025-04-27 22:16:12,409 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2025-04-27 22:16:12,410 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2025-04-27 22:16:12,596 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2025-04-27 22:16:12,597 [main] INFO org.apache.pig.Main - Logging error messages to: D:\hadoop\logs\pig_1745772372595.log
2025-04-27 22:16:12,611 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file C:\Users\Sourabh/.pigbootup not found
2025-04-27 22:16:12,782 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2025-04-27 22:16:12,784 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2025-04-27 22:16:12,953 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-04-27 22:16:12,966 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-7688e926-708e-498c-aaba-7fb4a01deaa5
2025-04-27 22:16:12,966 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> students = LOAD 'D:/BDA/lab6/students.txt'
>> USING PigStorage(',')
>> AS (student_id:int, name:chararray, age:int, department_id:int);
2025-04-27 22:30:25,497 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> departments = LOAD 'D:/BDA/lab6/departments.txt'
>> USING PigStorage(',')
>> AS (department_id:int, department_name:chararray);
2025-04-27 22:31:56,582 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> sorted_students = ORDER students BY age ASC;
grunt> DUMP sorted_students;
2025-04-27 22:32:03,884 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
23°C
Rain showers
Search
ENG
IN
22:33
27-04-2025

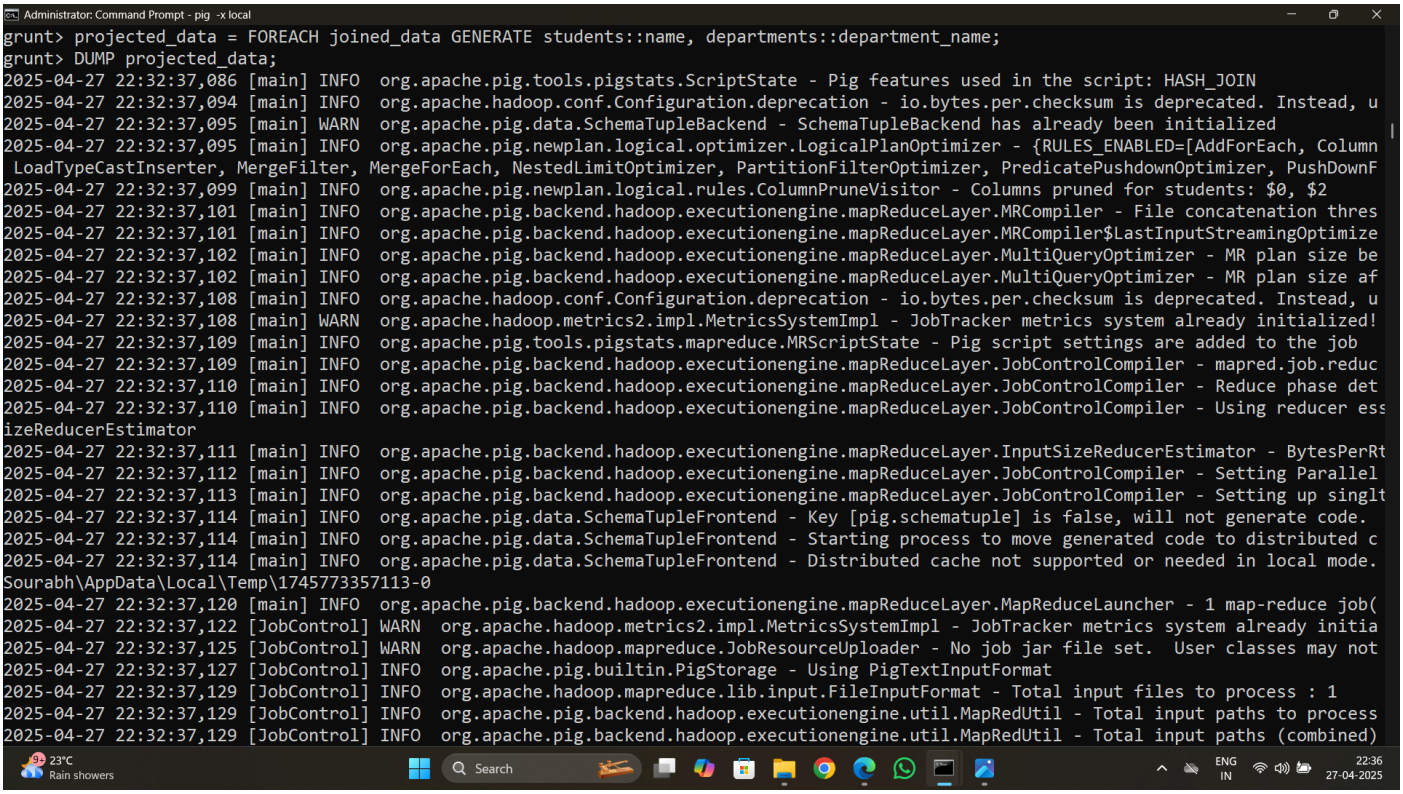

Administrator: Command Prompt - pig -x local
Total records proactively spilled: 0
Job DAG:
job_local114988160_0001 -> job_local197793578_0002,
job_local197793578_0002 -> job_local1498271482_0003,
job_local1498271482_0003
2025-04-27 22:32:06,110 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:06,112 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:06,113 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:06,115 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:06,116 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:06,116 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:06,118 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:06,119 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:06,120 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:06,121 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2025-04-27 22:32:06,124 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-04-27 22:32:06,125 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-04-27 22:32:06,130 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system cannot find the path specified
-- file path: tmp/temp-1141495091/tmp-287859946/part-r-00000
2025-04-27 22:32:06,169 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system cannot find the path specified
-- file path: tmp/temp-1141495091/tmp-287859946/_SUCCESS
2025-04-27 22:32:06,206 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-04-27 22:32:06,206 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(7,Grace,20,103)
(3,Bob,20,101)
(6,Frank,21,104)
(1,John,21,101)
(5,Eve,22,102)
(2,Alice,22,102)
(4,David,23,103)
(8,Hannah,24,101)
23°C
Rain showers
Search
ENG
IN
22:33
27-04-2025

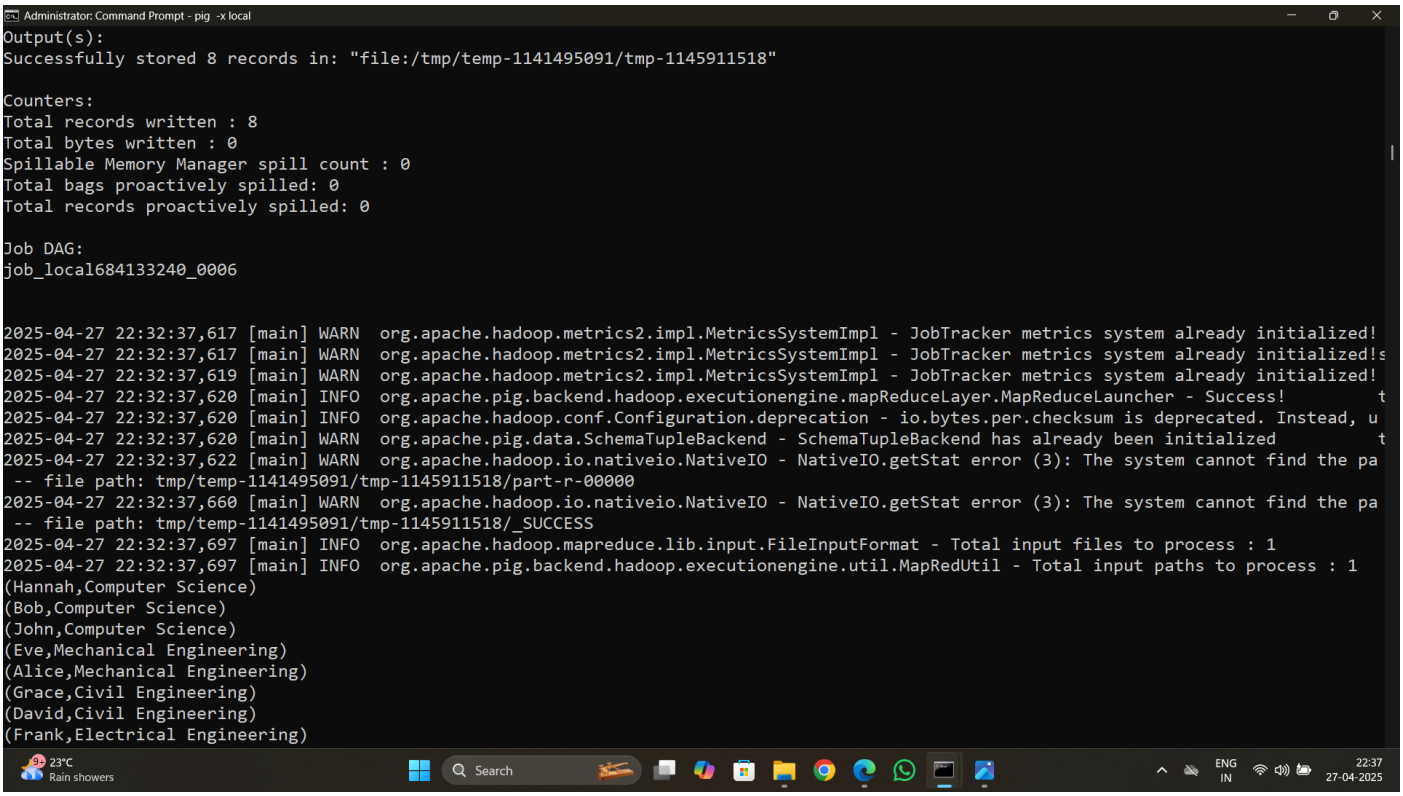

Administrator: Command Prompt - pig -x local
grunt> filtered_students = FILTER students BY age > 20;
grunt> DUMP filtered_students;
2025-04-27 22:32:43,190 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTE
2025-04-27 22:32:43,198 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprec
2025-04-27 22:32:43,199 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initi
2025-04-27 22:32:43,199 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[Ac
tOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, St
2025-04-27 22:32:43,202 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File cor
2025-04-27 22:32:43,202 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer -
2025-04-27 22:32:43,203 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer -
2025-04-27 22:32:43,208 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprec
2025-04-27 22:32:43,209 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system alre
2025-04-27 22:32:43,209 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are ac
2025-04-27 22:32:43,209 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
2025-04-27 22:32:43,211 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
2025-04-27 22:32:43,212 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not
2025-04-27 22:32:43,212 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code
2025-04-27 22:32:43,212 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or neede
2025-04-27 22:32:43,218 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1
2025-04-27 22:32:43,219 [JobControl] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics syste
2025-04-27 22:32:43,224 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User
2025-04-27 22:32:43,226 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2025-04-27 22:32:43,227 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to
2025-04-27 22:32:43,228 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
2025-04-27 22:32:43,228 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
2025-04-27 22:32:43,231 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2025-04-27 22:32:43,238 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_lo
2025-04-27 22:32:43,238 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2025-04-27 22:32:43,278 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost
23°C
Mostly cloudy
Search
ENG
IN
22:52
27-04-2025

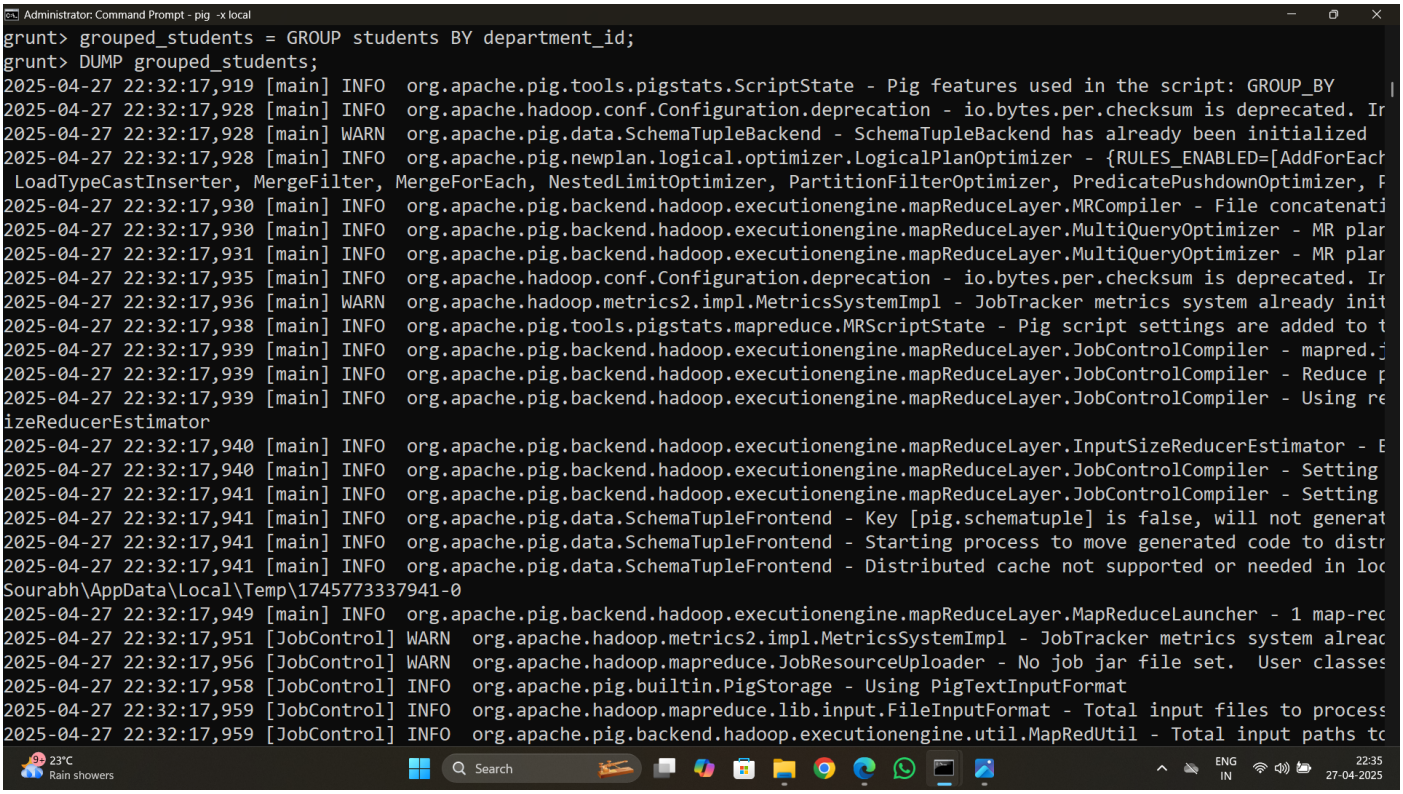

Administrator: Command Prompt - pig -x local
Output(s):
Successfully stored 6 records in: "file:/tmp/temp-1141495091/tmp279227610"
Counters:
Total records written : 6
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_local875084081_0007
2025-04-27 22:32:43,498 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:43,500 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:43,501 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:43,501 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2025-04-27 22:32:43,502 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, us
2025-04-27 22:32:43,502 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-04-27 22:32:43,504 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system cannot find the pat
-- file path: tmp/temp-1141495091/tmp279227610/part-m-00000
2025-04-27 22:32:43,545 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system cannot find the pat
-- file path: tmp/temp-1141495091/tmp279227610/_SUCCESS
2025-04-27 22:32:43,582 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-04-27 22:32:43,582 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,John,21,101)
(2,Alice,22,102)
(4,David,23,103)
(5,Eve,22,102)
(6,Frank,21,104)
(8,Hannah,24,101)
23°C
Mostly cloudy
Search
ENG
IN
22:52
27-04-2025

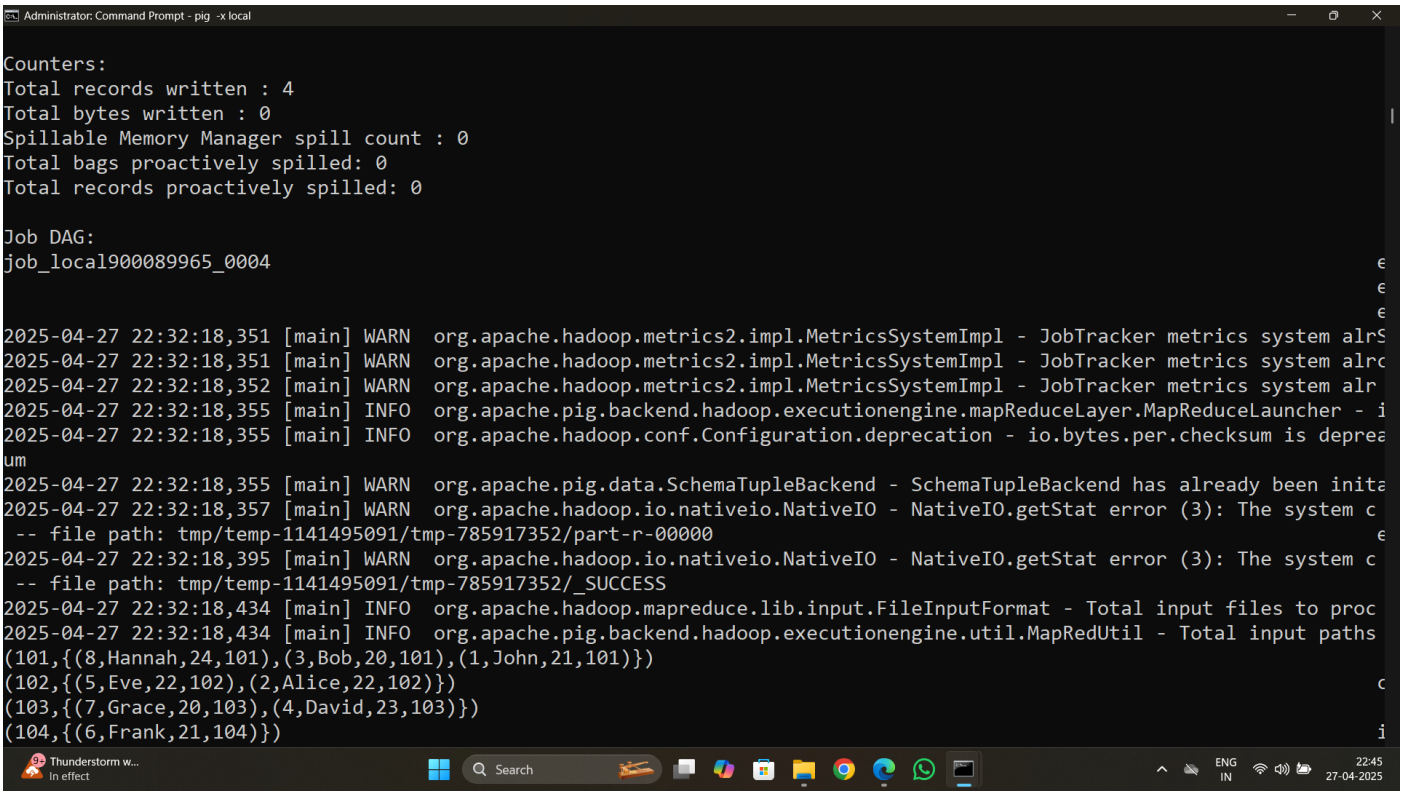

Administrator: Command Prompt - pig -x local
grunt> joined_data = JOIN students BY department_id, departments BY department_id;
grunt> DUMP joined_data;
2025-04-27 22:32:30,488 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN
2025-04-27 22:32:30,496 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, us
2025-04-27 22:32:30,496 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-04-27 22:32:30,496 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnP
tOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInser
2025-04-27 22:32:30,498 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresh
2025-04-27 22:32:30,500 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer
2025-04-27 22:32:30,500 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size bef
2025-04-27 22:32:30,500 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size aft
2025-04-27 22:32:30,507 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, us
2025-04-27 22:32:30,507 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:30,508 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2025-04-27 22:32:30,509 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce
2025-04-27 22:32:30,509 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase dete
2025-04-27 22:32:30,509 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer est
2025-04-27 22:32:30,510 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerRe
2025-04-27 22:32:30,510 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelis
2025-04-27 22:32:30,511 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single
2025-04-27 22:32:30,512 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2025-04-27 22:32:30,515 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed ca
2025-04-27 22:32:30,515 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode.
2025-04-27 22:32:30,523 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s
2025-04-27 22:32:30,524 [JobControl] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initial
2025-04-27 22:32:30,529 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not
2025-04-27 22:32:30,530 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2025-04-27 22:32:30,531 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-04-27 22:32:30,531 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process
2025-04-27 22:32:30,531 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined)
2025-04-27 22:32:30,532 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2025-04-27 22:32:30,532 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-04-27 22:32:30,533 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process
23°C
Mostly cloudy
Search
ENG
IN
22:48
27-04-2025


Administrator: Command Prompt - pig -x local
Successfully stored 8 records in: "file:/tmp/temp-1141495091/tmp-1013768679"
Counters:
Total records written : 8
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_local533011572_0005
2025-04-27 22:32:31,040 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:31,041 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:31,042 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:31,043 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2025-04-27 22:32:31,043 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, us
2025-04-27 22:32:31,043 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-04-27 22:32:31,046 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system cannot find the pat
-- file path: tmp/temp-1141495091/tmp-1013768679/part-r-00000
2025-04-27 22:32:31,084 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system cannot find the pat
-- file path: tmp/temp-1141495091/tmp-1013768679/_SUCCESS
2025-04-27 22:32:31,122 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-04-27 22:32:31,122 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(8,Hannah,24,101,101,Computer Science)
(3,Bob,20,101,101,Computer Science)
(1,John,21,101,101,Computer Science)
(5,Eve,22,102,102,Mechanical Engineering)
(2,Alice,22,102,102,Mechanical Engineering)
(7,Grace,20,103,103,Civil Engineering)
(4,David,23,103,103,Civil Engineering)
(6,Frank,21,104,104,Electrical Engineering)
23°C
Rain showers
Search
ENG
IN
22:36
27-04-2025


Administrator: Command Prompt - pig -x local
grunt> projected_data = FOREACH joined_data GENERATE students::name, departments::department_name;
grunt> DUMP projected_data;
2025-04-27 22:32:37,086 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN
2025-04-27 22:32:37,094 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, u
2025-04-27 22:32:37,095 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-04-27 22:32:37,095 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, Column
LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownF
2025-04-27 22:32:37,099 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for students: $0, $2
2025-04-27 22:32:37,101 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thres
2025-04-27 22:32:37,101 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimize
2025-04-27 22:32:37,102 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size be
2025-04-27 22:32:37,102 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size af
2025-04-27 22:32:37,108 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, u
2025-04-27 22:32:37,108 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:37,109 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2025-04-27 22:32:37,109 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduc
2025-04-27 22:32:37,110 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase det
2025-04-27 22:32:37,110 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer ess
izeReducerEstimator
2025-04-27 22:32:37,111 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerRt
2025-04-27 22:32:37,112 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallel
2025-04-27 22:32:37,113 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up singlt
2025-04-27 22:32:37,114 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2025-04-27 22:32:37,114 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed c
2025-04-27 22:32:37,114 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode.
Sourabh\AppData\Local\Temp\1745773357113-0
2025-04-27 22:32:37,120 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(
2025-04-27 22:32:37,122 [JobControl] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initia
2025-04-27 22:32:37,125 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not
2025-04-27 22:32:37,127 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2025-04-27 22:32:37,129 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-04-27 22:32:37,129 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process
2025-04-27 22:32:37,129 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined)
23°C
Rain showers
Search
ENG
IN
22:36
27-04-2025


Administrator: Command Prompt - pig -x local
Output(s):
Successfully stored 8 records in: "file:/tmp/temp-1141495091/tmp-1145911518"
Counters:
Total records written : 8
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_local684133240_0006
2025-04-27 22:32:37,617 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:37,617 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!s
2025-04-27 22:32:37,619 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-27 22:32:37,620 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success! t
2025-04-27 22:32:37,620 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, u
2025-04-27 22:32:37,620 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized t
2025-04-27 22:32:37,622 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system cannot find the pa
-- file path: tmp/temp-1141495091/tmp-1145911518/part-r-00000
2025-04-27 22:32:37,660 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system cannot find the pa
-- file path: tmp/temp-1141495091/tmp-1145911518/_SUCCESS
2025-04-27 22:32:37,697 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-04-27 22:32:37,697 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Hannah,Computer Science)
(Bob,Computer Science)
(John,Computer Science)
(Eve,Mechanical Engineering)
(Alice,Mechanical Engineering)
(Grace,Civil Engineering)
(David,Civil Engineering)
(Frank,Electrical Engineering)
23°C
Rain showers
Search
ENG
IN
22:37
27-04-2025


Administrator: Command Prompt - pig -x local
grunt> grouped_students = GROUP students BY department_id;
grunt> DUMP grouped_students;
2025-04-27 22:32:17,919 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2025-04-27 22:32:17,928 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Ir
2025-04-27 22:32:17,928 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-04-27 22:32:17,928 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach
LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, F
2025-04-27 22:32:17,930 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenati
2025-04-27 22:32:17,930 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plar
2025-04-27 22:32:17,931 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plar
2025-04-27 22:32:17,935 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Ir
2025-04-27 22:32:17,936 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already init
2025-04-27 22:32:17,938 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to t
2025-04-27 22:32:17,939 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.j
2025-04-27 22:32:17,939 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce p
2025-04-27 22:32:17,939 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using re
izeReducerEstimator
2025-04-27 22:32:17,940 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - E
2025-04-27 22:32:17,940 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting
2025-04-27 22:32:17,941 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting
2025-04-27 22:32:17,941 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generat
2025-04-27 22:32:17,941 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distr
2025-04-27 22:32:17,941 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in loc
Sourabh\AppData\Local\Temp\1745773337941-0
2025-04-27 22:32:17,949 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-rec
2025-04-27 22:32:17,951 [JobControl] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system alreac
2025-04-27 22:32:17,956 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes
2025-04-27 22:32:17,958 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2025-04-27 22:32:17,959 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process
2025-04-27 22:32:17,959 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
23°C
Rain showers
Search
ENG
IN
22:35
27-04-2025


Administrator: Command Prompt - pig -x local
Counters:
Total records written : 4
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_local1900089965_0004
2025-04-27 22:32:18,351 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system alrS
2025-04-27 22:32:18,351 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system alrc
2025-04-27 22:32:18,352 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system alr
2025-04-27 22:32:18,355 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - i
2025-04-27 22:32:18,355 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprea
um
2025-04-27 22:32:18,355 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been inita
2025-04-27 22:32:18,357 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system c
-- file path: tmp/temp-1141495091/tmp-785917352/part-r-00000
2025-04-27 22:32:18,395 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system c
-- file path: tmp/temp-1141495091/tmp-785917352/_SUCCESS
2025-04-27 22:32:18,434 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to proc
2025-04-27 22:32:18,434 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths
(101,{(8,Hannah,24,101),(3,Bob,20,101),(1,John,21,101)})
(102,{(5,Eve,22,102),(2,Alice,22,102)})
(103,{(7,Grace,20,103),(4,David,23,103)})
(104,{(6,Frank,21,104)})
Thunderstorm w...
In effect
Search
ENG
IN
22:45
27-04-2025

Experiment 7

Aim: Use Hive to create, alter, and drop databases, tables, views, functions, and indexes

Apache Hive Installation Steps:

Software Requirements:

- Ubuntu/Linux OS
- Java JDK 8 or higher
- Hadoop (HDFS configured)
- Apache Hive 3.x

Step 1: Download Hive

- Go to the official Apache website and Download latest version of Apache Hive
- Apache Hive 3.x (Extract the file)
- System variable>>New System Variable>>set as **HIVE_HOME** with path
- System variable>>Path>>New>>set **%HIVE_HOME%\bin**

Step 2: Create Hive Directories in HDFS

- hdfs dfs -mkdir /user/hive/warehouse
- hdfs dfs -chmod g+w /user/hive/warehouse

Step 3: Initialize Hive Metastore

- schematool -dbType derby -initSchema

Step 4: Start Hive CLI

- hive

Steps and Commands:

1)Create Database & Table

```
CREATE DATABASE IF NOT EXISTS my_lab_db;

USE my_lab_db;

CREATE TABLE IF NOT EXISTS student (
   roll_no INT,
   name STRING,
   marks INT
)
```

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ',';

Output: OK

Time taken: 0.22 seconds

2) Insert Sample Data

INSERT INTO TABLE student VALUES

(101, 'Alice', 85),

(102, 'Bob', 90),

(103, 'Charlie', 78);

Output: OK

Time taken: 0.35 seconds

3) Add a New Column (ALTER TABLE)

ALTER TABLE student ADD COLUMNS (grade STRING);

Output: OK

Time taken: 0.10 seconds

4) Insert Grade Values

INSERT OVERWRITE TABLE student VALUES
(101, 'Alice', 85, 'A'),
(102, 'Bob', 90, 'A+'),
(103, 'Charlie', 78, 'B');

Output: OK

Time taken: 0.40 seconds

5) Select Query to View Table Data

SELECT * FROM student;

Output: 101 Alice 85 A

102 Bob 90 A+

103 Charlie 78 B

Time taken: 0.2 seconds

6) Create a View

CREATE VIEW student_view AS

SELECT roll_no, name FROM student;

SELECT * FROM student_view;

Output: 101 Alice

102 Bob

103 Charlie

Time taken: 0.12 seconds

7) Create an Index (on roll_no)

CREATE INDEX idx_roll_no

ON TABLE student (roll_no)

AS 'COMPACT'

WITH DEFERRED REBUILD;

Output: OK

Time taken: 0.30 seconds

8) Clean Up (DROP View, Index, Table, Database)

DROP VIEW IF EXISTS student_view;

DROP INDEX IF EXISTS idx_roll_no ON student;

DROP TABLE IF EXISTS student;

DROP DATABASE IF EXISTS my_lab_db CASCADE;

Output: OK

Time taken: 0.15–0.30 seconds each

9) Note on CREATE FUNCTION:

In Hive, user-defined functions (UDFs) allow you to extend Hive with custom logic.
To use a UDF, you must first compile it in Java, create a JAR, and register it in Hive using:

ADD JAR /path/to/your_udf.jar;
CREATE TEMPORARY FUNCTION my_length AS 'com.example.MyLengthFunction';

Experiment 8

Aim: Implement a word count program in Hadoop and Spark.

Code:

WordCountDriver:

```
package WordCount;


import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
public class WordCountDriver {
     public static void main(String[] args) throws Exception {
       Configuration conf = new Configuration();
       Job job = Job.getInstance(conf, "Word Count");


       job.setJarByClass(WordCountDriver.class);
       job.setMapperClass(WordCountMapper.class);
       job.setReducerClass(WordCountReducer.class);
       job.setOutputKeyClass(Text.class);
       job.setOutputValueClass(IntWritable.class);
       FileInputFormat.addInputPath(job, new Path(args[0]));  // input path
       FileOutputFormat.setOutputPath(job, new Path(args[1])); // output path
       System.exit(job.waitForCompletion(true) ? 0 : 1);
   }
}
```

WordCountMapper:

```
package WordCount;


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


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


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


   public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException {


      String line = value.toString();
      String[] words = line.split("\\s+"); // Split by spaces
      for (String str : words) {
         word.set(str);
         context.write(word, one);
      }
   }
}
```

WordCountReducer:

```
package WordCount;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Reducer;


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


   public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException,
InterruptedException {


      int sum = 0;

      for (IntWritable val : values) {

         sum += val.get();

      }

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

   }

}
```

Commands for Execution

1. D:\hadoop\sbin>start-all.cmd
2. D:\hadoop\sbin>hadoop fs -mkdir /pgm8
3. D:\hadoop\sbin>hadoop fs -put D:\BDA\lab8\sample.txt /pgm8
4. D:\hadoop\sbin>hadoop jar D:\BDA\lab8\sample.jar WordCount.WordCountDriver /pgm8 /pgm8OP
5. D:\hadoop\sbin>hadoop fs -cat /pgm8OP/part-r-00000

Sample.txt input file:

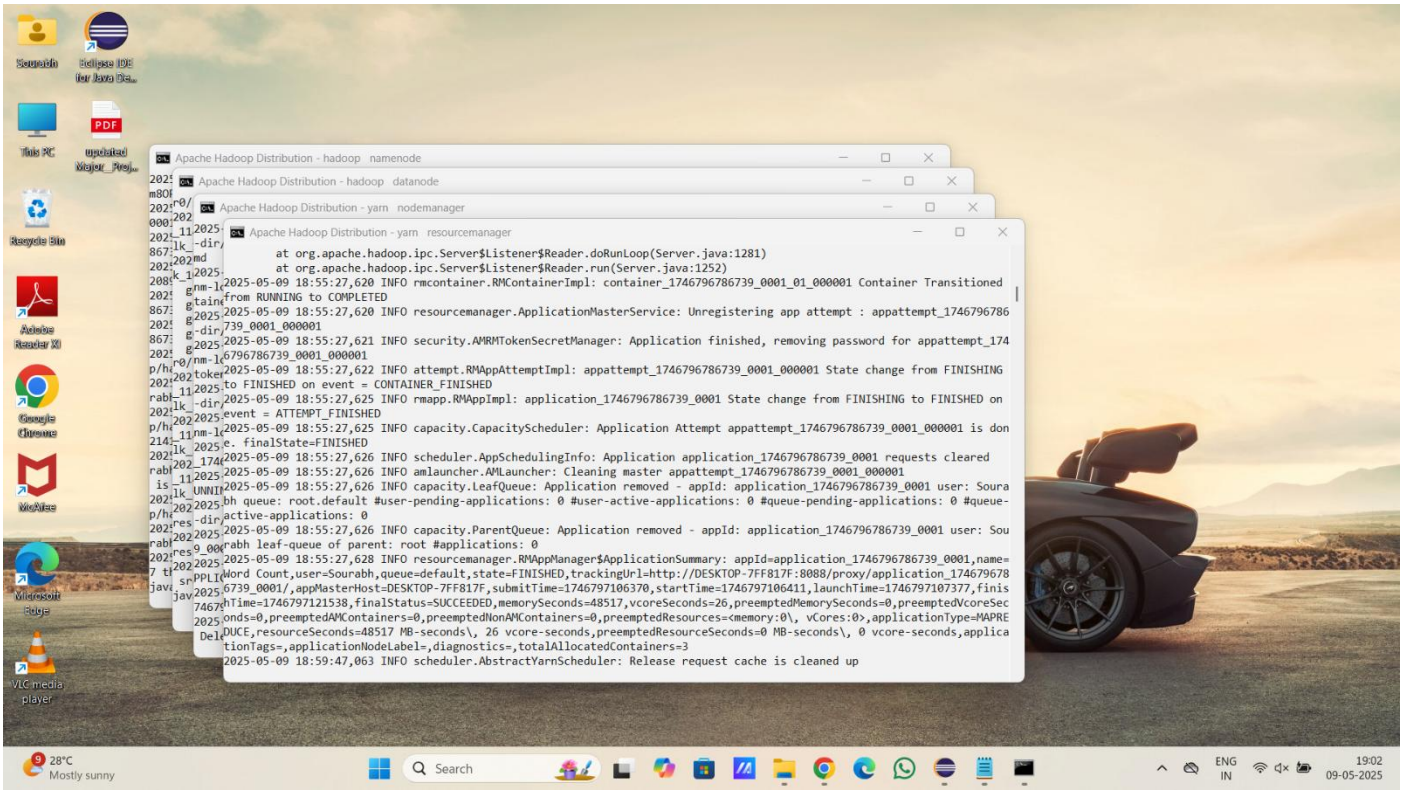
Sachin Virat Rohit Dhoni Sachin

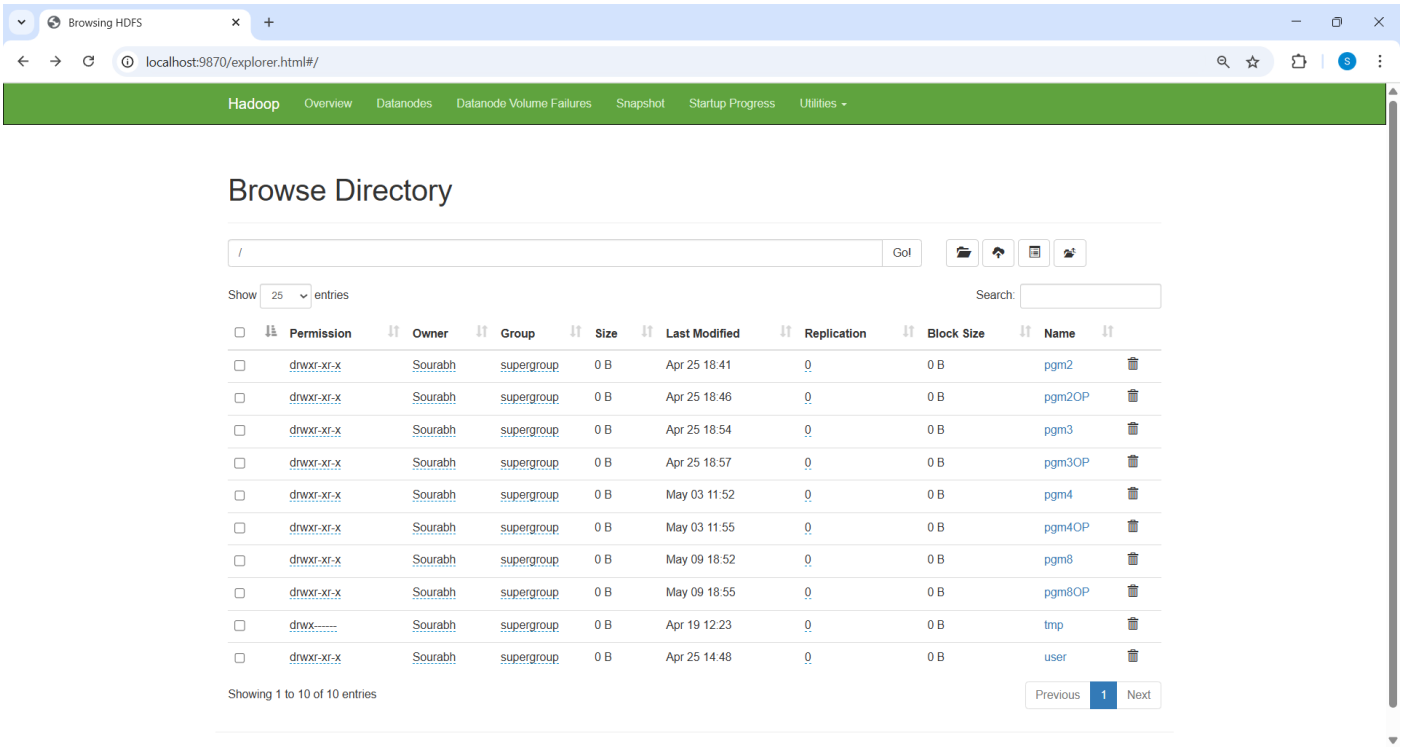
Sourav Rohit Virat Dhoni Virat

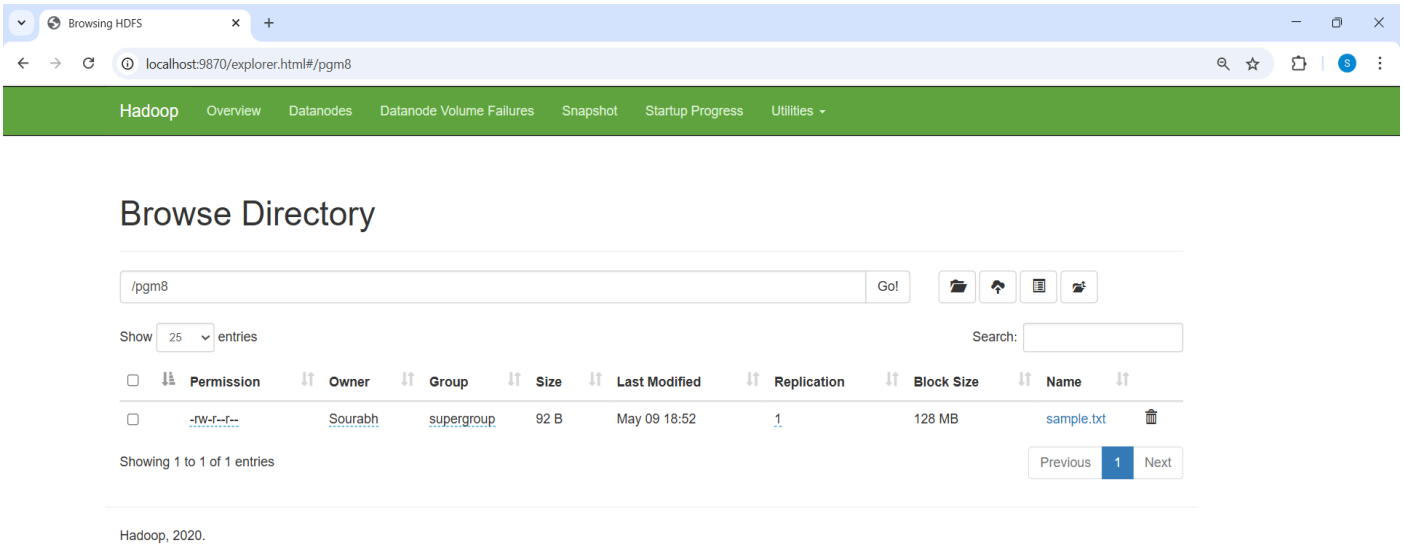
Rohit Sourav Sourav Rohit

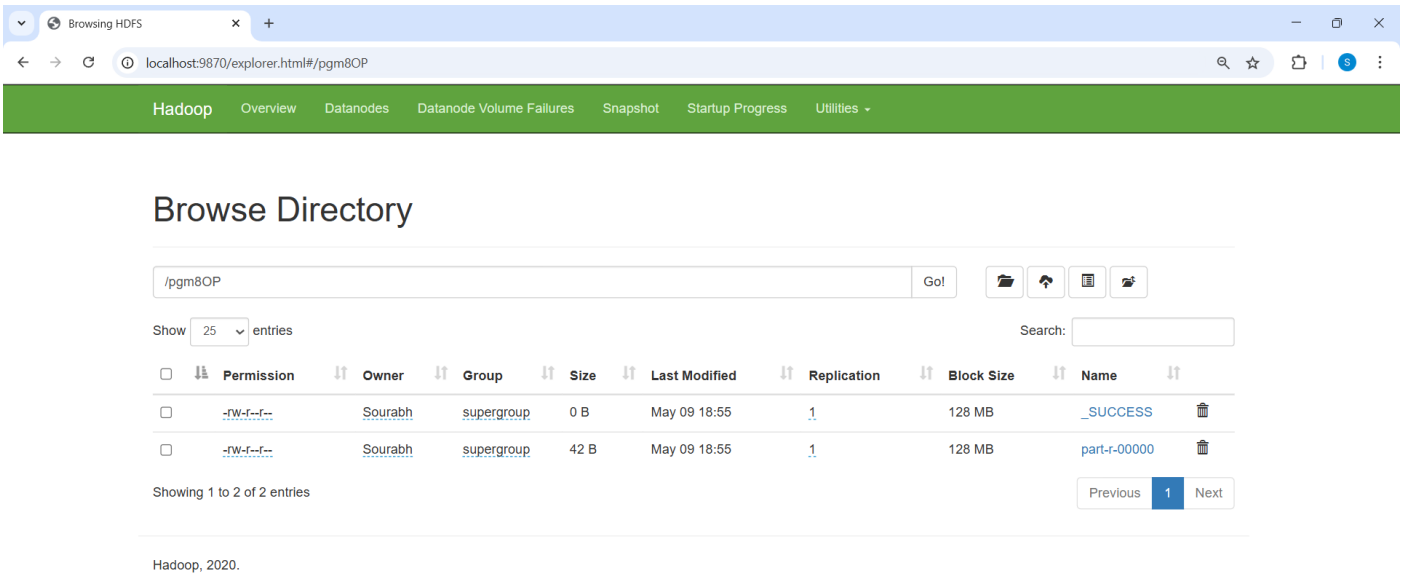
Output:

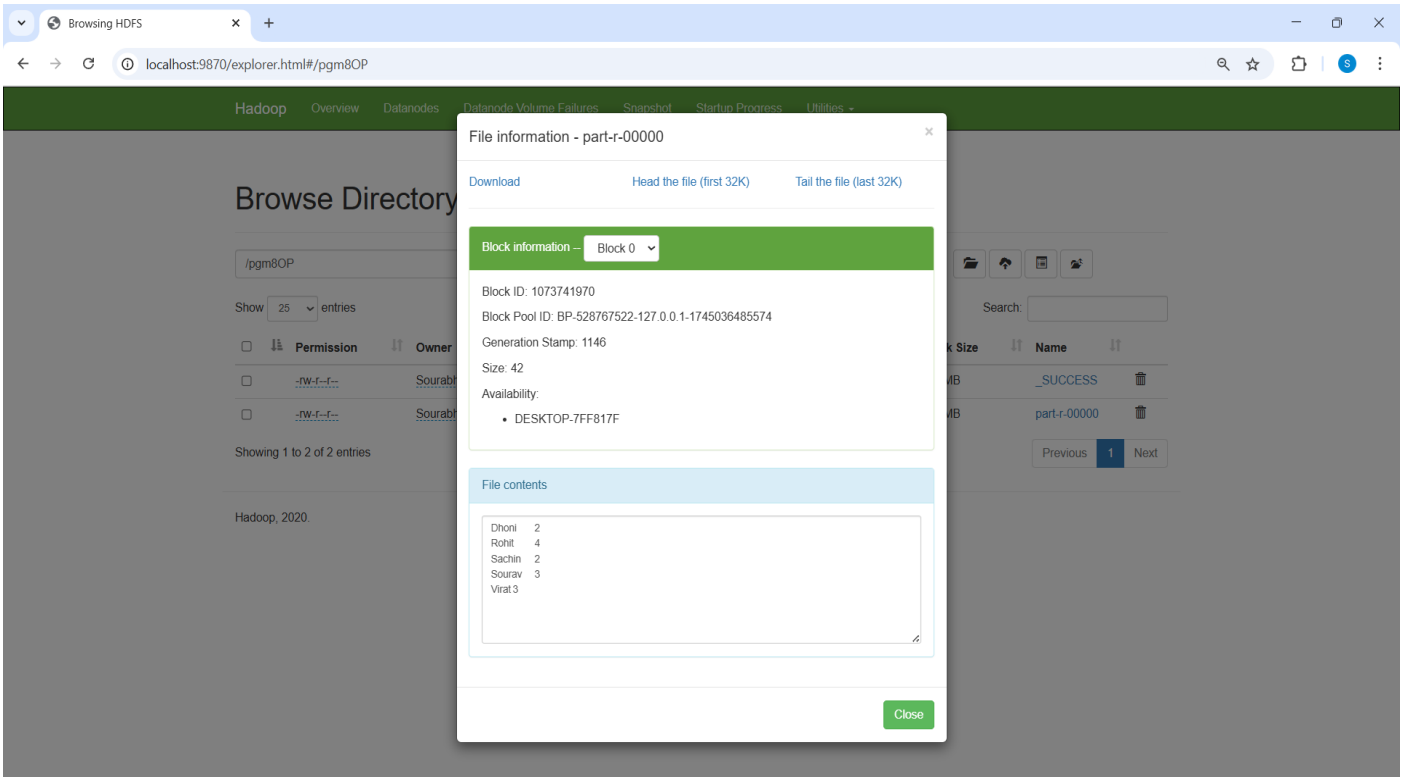
```
Administrator: Command Prompt

C:\Windows\System32>d:

D:\>cd hadoop\sbin

D:\hadoop\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

D:\hadoop\sbin>hadoop fs -mkdir /pgm8

D:\hadoop\sbin>hadoop fs -put D:\BDA\lab8\sample.txt /pgm8

D:\hadoop\sbin>hadoop jar D:\BDA\lab8\sample.jar WordCount.WordCountDriver /pgm8 /pgm8OP
2025-05-09 18:55:05,485 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-05-09 18:55:05,923 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
2025-05-09 18:55:05,938 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
2025-05-09 18:55:06,096 INFO input.FileInputFormat: Total input files to process : 1
2025-05-09 18:55:06,136 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-09 18:55:06,205 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746796786739_0001
2025-05-09 18:55:06,205 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-09 18:55:06,324 INFO conf.Configuration: resource-types.xml not found
2025-05-09 18:55:06,324 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-05-09 18:55:06,482 INFO impl.YarnClientImpl: Submitted application application_1746796786739_0001
2025-05-09 18:55:06,516 INFO mapreduce.Job: The url to track the job: http://DESKTOP-7FF817F:8088/proxy/application_174
2025-05-09 18:55:06,517 INFO mapreduce.Job: Running job: job_1746796786739_0001
2025-05-09 18:55:13,625 INFO mapreduce.Job: Job job_1746796786739_0001 running in uber mode : false
2025-05-09 18:55:13,627 INFO mapreduce.Job:  map 0% reduce 0%
2025-05-09 18:55:17,703 INFO mapreduce.Job:  map 100% reduce 0%
```

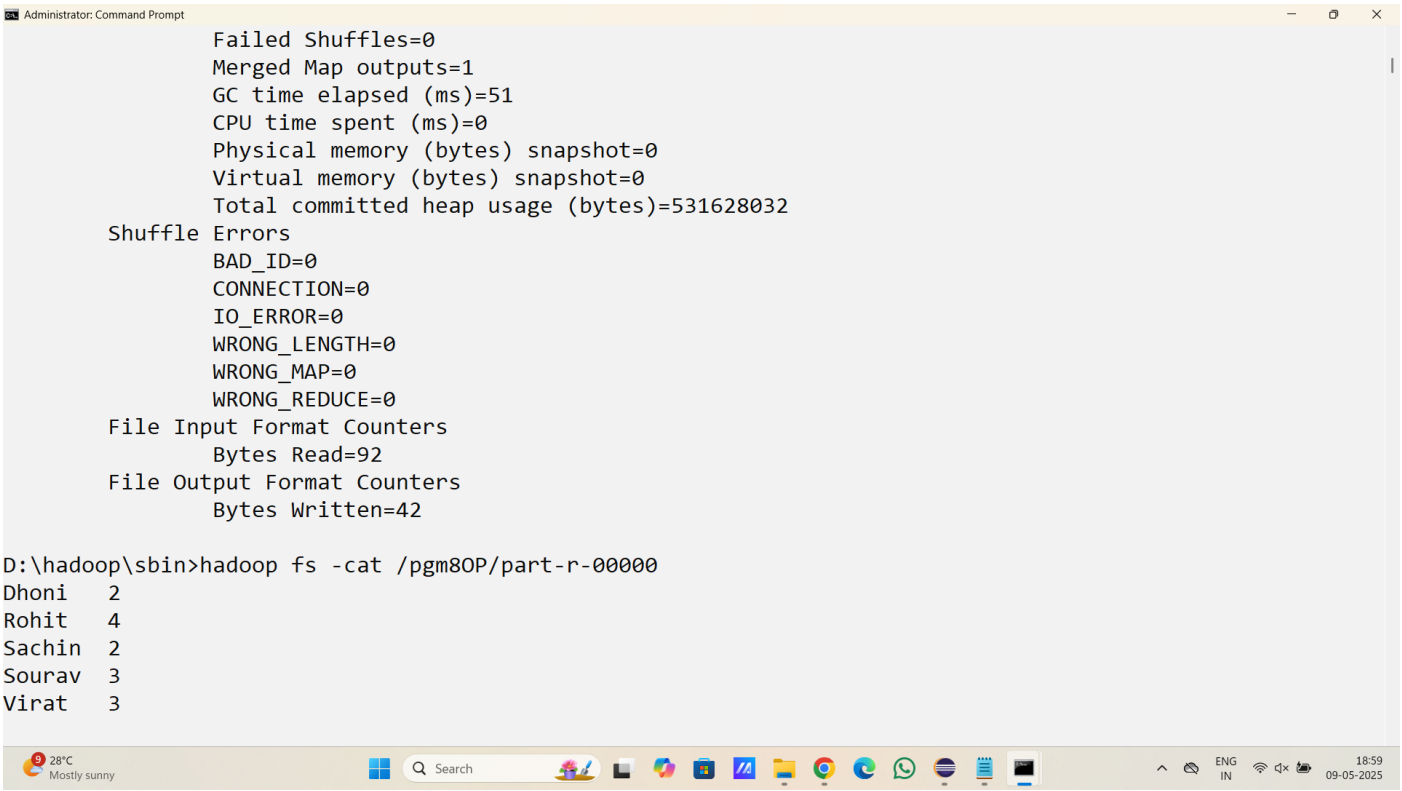



Browsing HDFS
localhost:9870/explorer.html#/
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:41 0 0 B pgm2
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:46 0 0 B pgm2OP
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:54 0 0 B pgm3
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:57 0 0 B pgm3OP
drwxr-xr-x Sourabh supergroup 0 B May 03 11:52 0 0 B pgm4
drwxr-xr-x Sourabh supergroup 0 B May 03 11:55 0 0 B pgm4OP
drwxr-xr-x Sourabh supergroup 0 B May 09 18:52 0 0 B pgm8
drwxr-xr-x Sourabh supergroup 0 B May 09 18:55 0 0 B pgm8OP
drwx------ Sourabh supergroup 0 B Apr 19 12:23 0 0 B tmp
drwxr-xr-x Sourabh supergroup 0 B Apr 25 14:48 0 0 B user
Showing 1 to 10 of 10 entries
Previous
1
Next


Browsing HDFS
localhost:9870/explorer.html#/pgm8
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm8
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r-- Sourabh supergroup 92 B May 09 18:52 1 128 MB sample.txt
Showing 1 to 1 of 1 entries
Previous
1
Next
Hadoop, 2020.


Browsing HDFS
localhost:9870/explorer.html#/pgm8OP
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm8OP
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
Sourabh
supergroup
0 B
May 09 18:55
1
128 MB
_SUCCESS
-rw-r--r--
Sourabh
supergroup
42 B
May 09 18:55
1
128 MB
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.


Browsing HDFS
localhost:9870/explorer.html#/pgm8OP
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm8OP
Show 25 entries
Search:
Permission
Owner
Name
-rw-r--r--
_SUCCESS
-rw-r--r--
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.
File information - part-r-00000
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information --
Block 0
Block ID: 1073741970
Block Pool ID: BP-528767522-127.0.0.1-1745036485574
Generation Stamp: 1146
Size: 42
Availability:
DESKTOP-7FF817F
File contents
Dhoni 2
Rohit 4
Sachin 2
Sourav 3
Virat 3
Close


Administrator: Command Prompt
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=51
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=531628032
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=92
File Output Format Counters
Bytes Written=42
D:\hadoop\sbin>hadoop fs -cat /pgm8OP/part-r-00000
Dhoni 2
Rohit 4
Sachin 2
Sourav 3
Virat 3
28°C
Mostly sunny
Search
ENG
IN
18:59
09-05-2025

Experiment 9

Aim: Use CDH (Cloudera Distribution for Hadoop) and HUE (Hadoop User Interface) to analyze data and generate reports for sample datasets

Code:

SalesDriver:

```
package cdh;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

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

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


public class SalesDriver {
   public static void main(String[] args) throws Exception {
      Configuration conf = new Configuration();
      Job job = Job.getInstance(conf, "Sales Data Analysis");
      job.setJarByClass(SalesDriver.class);
      job.setMapperClass(SalesMapper.class);
      job.setReducerClass(SalesReducer.class);
      job.setOutputKeyClass(Text.class);
      job.setOutputValueClass(IntWritable.class);
      FileInputFormat.addInputPath(job, new Path(args[0]));
      FileOutputFormat.setOutputPath(job, new Path(args[1]));
      System.exit(job.waitForCompletion(true) ? 0 : 1);
   }
}
```

SalesMapper:

```
package cdh;


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


public class SalesMapper extends Mapper<Object, Text, Text, IntWritable> {
   private Text customerId = new Text();
   private IntWritable amount = new IntWritable();


   public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
      String line = value.toString();
      if (!line.startsWith("order_id")) {  // skip header
         String[] fields = line.split(",");
         customerId.set(fields[1]);  // customer_id
         amount.set(Integer.parseInt(fields[3])); // amount
         context.write(customerId, amount);
      }
   }
}
```

SalesReducer:

```
package cdh;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Reducer;

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

   public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException,
InterruptedException {

      int total = 0;

      for (IntWritable val : values) {

         total += val.get();

      }

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

   }

}
```

Commends for Execution

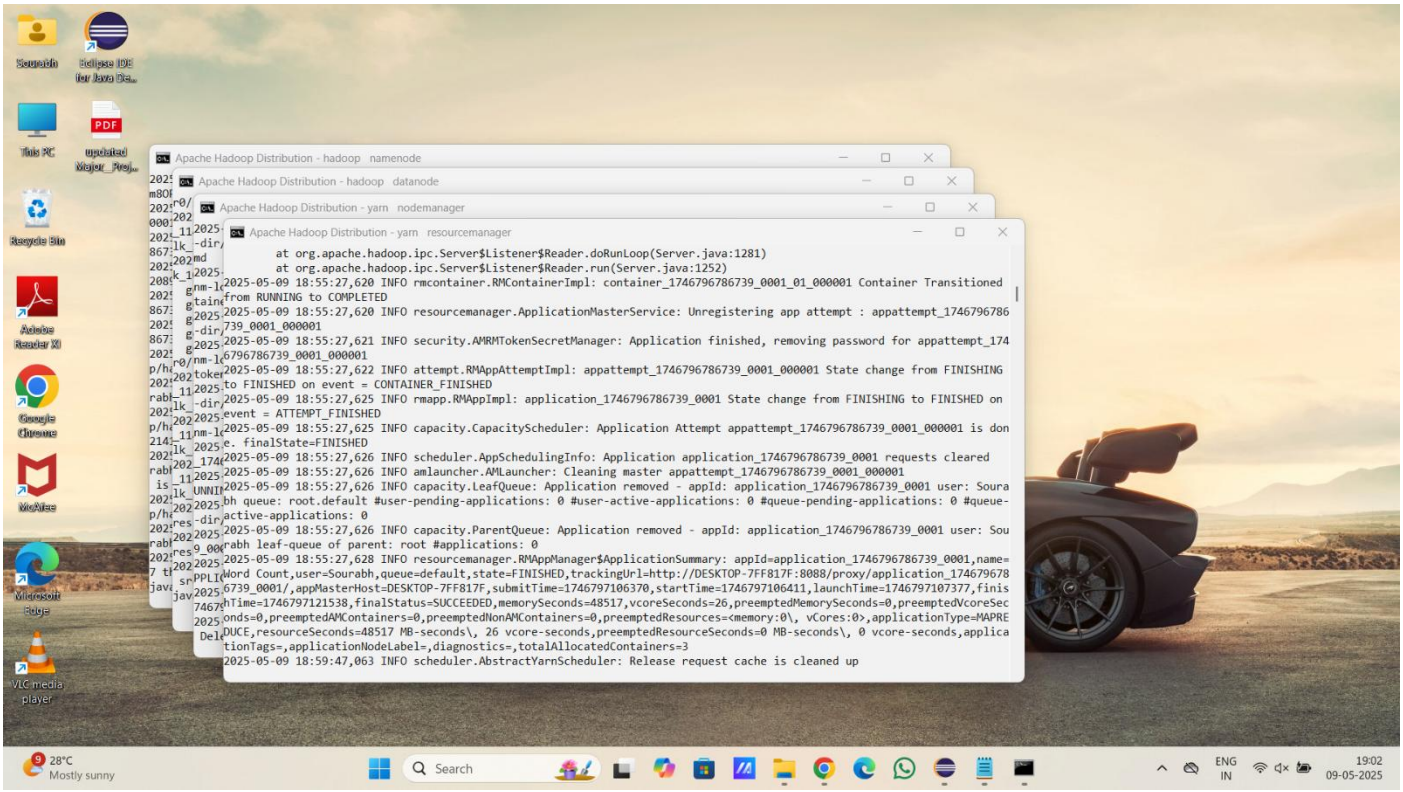
1. D:\hadoop\sbin>start-all.cmd
2. D:\hadoop\sbin>hadoop fs -mkdir /pgm9
3. D:\hadoop\sbin>hadoop fs -put D:\BDA\lab9\sales.csv /pgm9
4. D:\hadoop\sbin>hadoop jar D:\BDA\lab9\sales.jar cdh.SalesDriver /pgm9 /pgm9OP
5. D:\hadoop\sbin>hadoop fs -cat /pgm9OP/part-r-00000

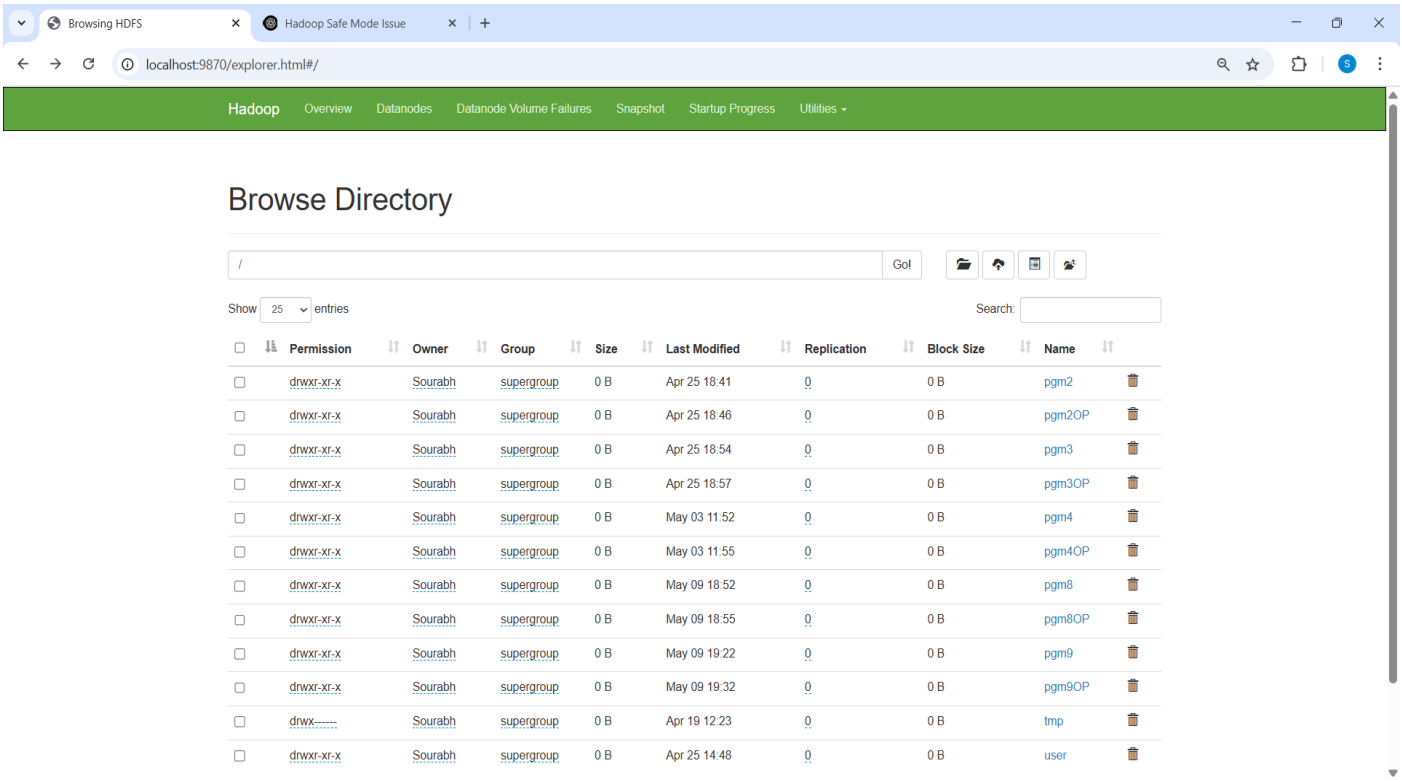
Sales.CSV Input file

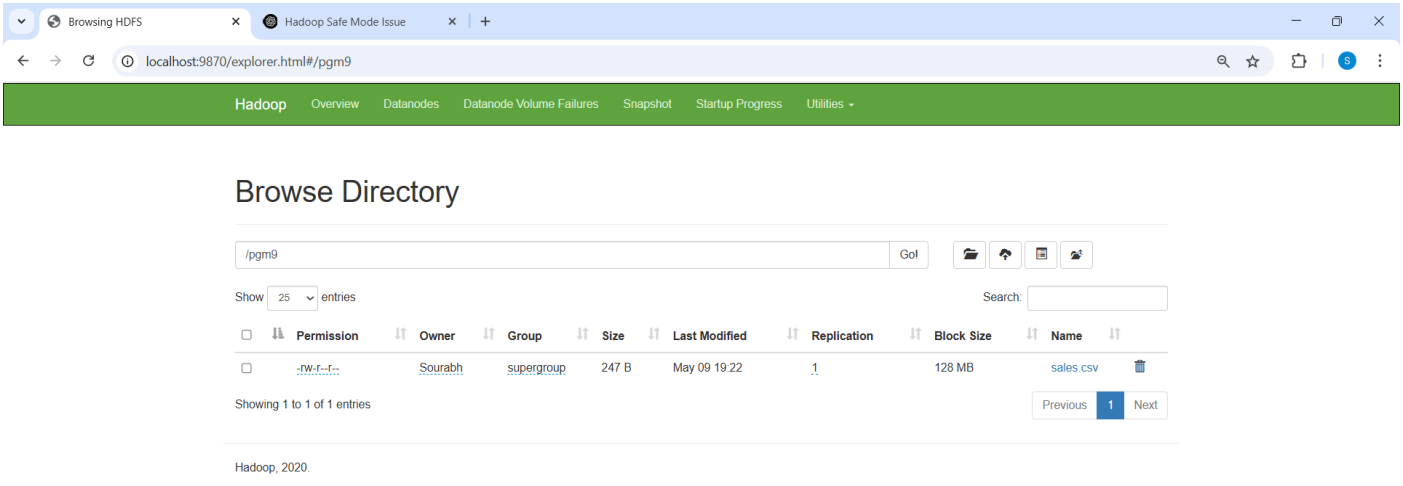
order_id	customer_id	order_date	amount
1	1001	01-01-2024	250
2	1002	01-01-2024	150
3	1001	02-01-2024	400
4	1003	02-01-2024	100
5	1002	03-01-2024	200
6	1005	04-01-2024	300
7	1006	05-01-2024	400
8	1009	06-01-2024	500
9	1007	07-01-2024	600

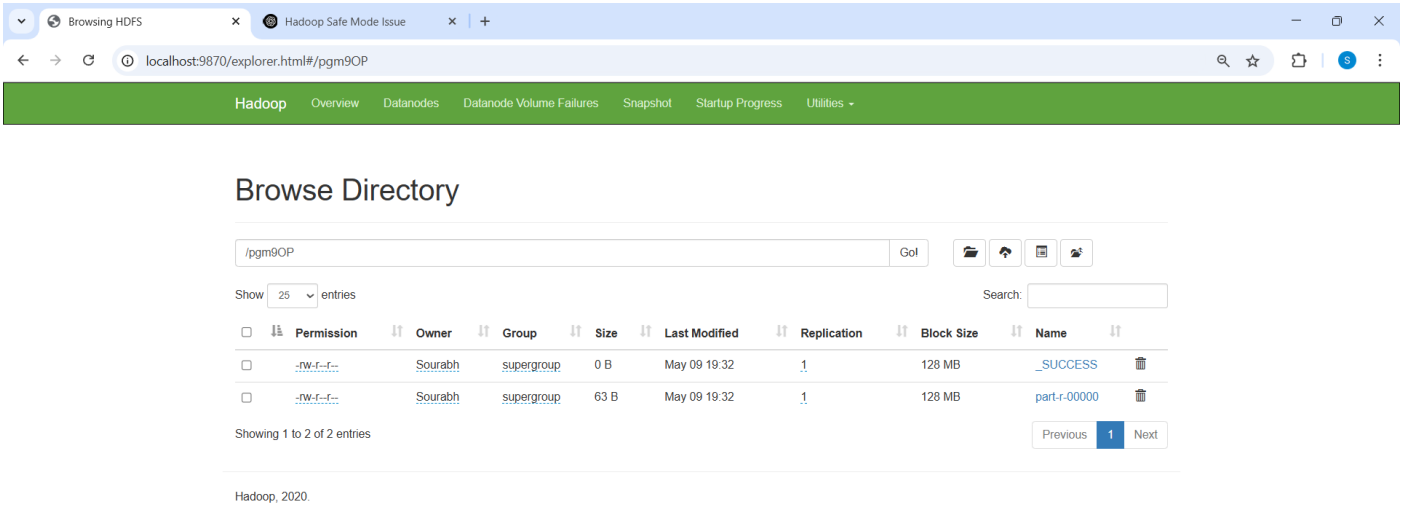
Output:

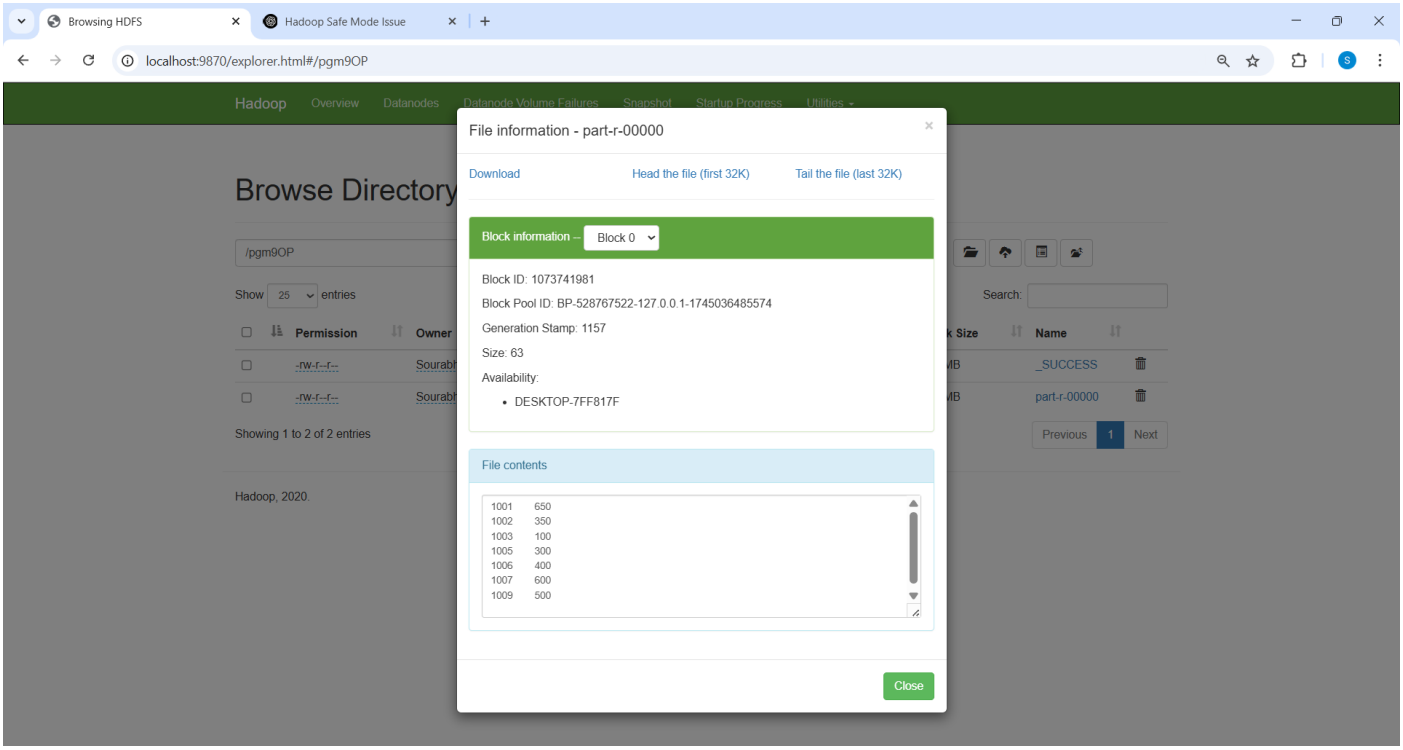
```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.26100.3915]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>d:

D:\>cd hadoop\sbin

D:\hadoop\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

D:\hadoop\sbin>hadoop fs -mkdir /pgm9
mkdir: Cannot create directory /pgm9. Name node is in safe mode.

D:\hadoop\sbin>hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
Safe mode is OFF

D:\hadoop\sbin>hadoop fs -mkdir /pgm9

D:\hadoop\sbin>hadoop fs -put D:\BDA\lab9\sales.csv /pgm9

D:\hadoop\sbin>hadoop jar D:\BDA\lab9\sales.jar cdh.SalesDriver /pgm9 /pgm9OP
2025-05-09 19:32:01,766 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /
2025-05-09 19:32:02,212 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not perform
```

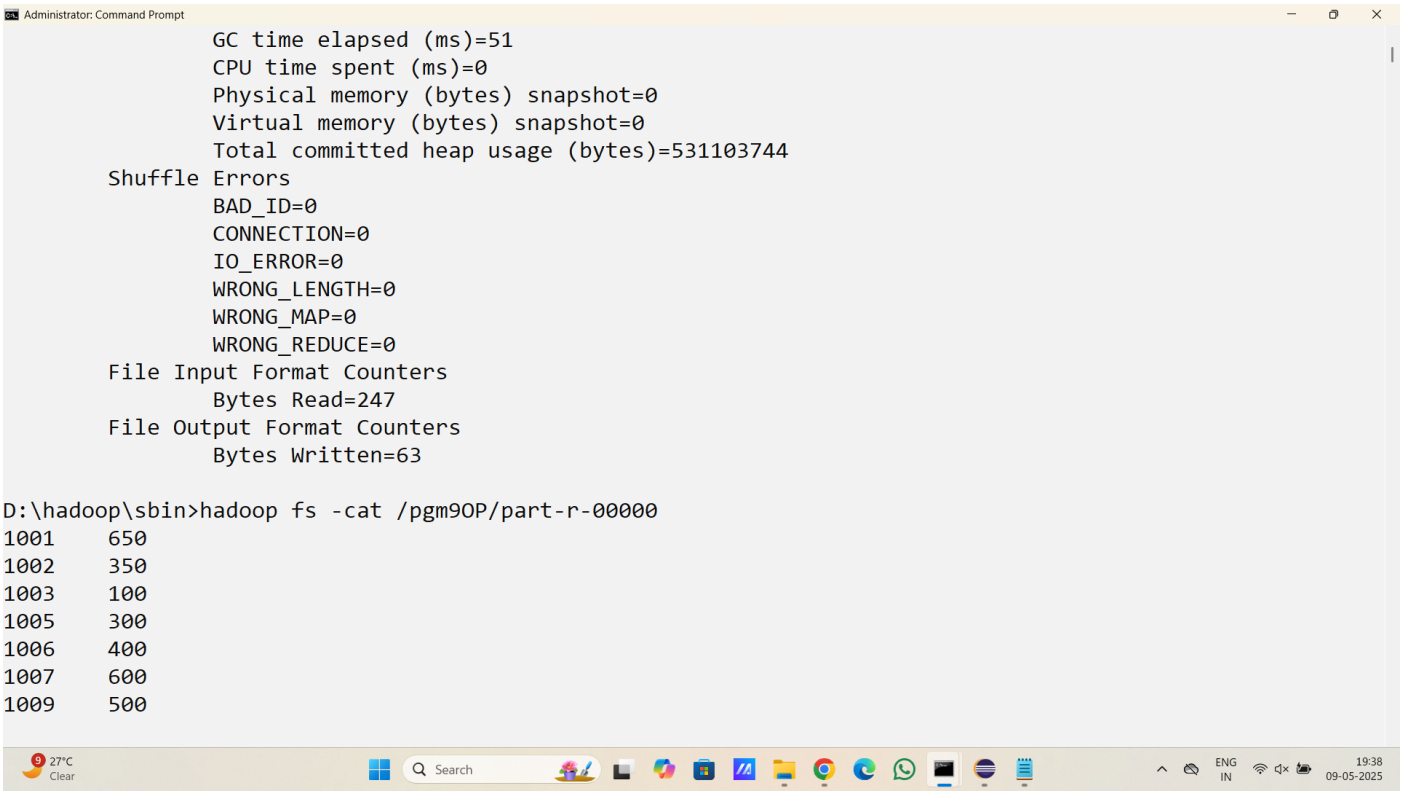


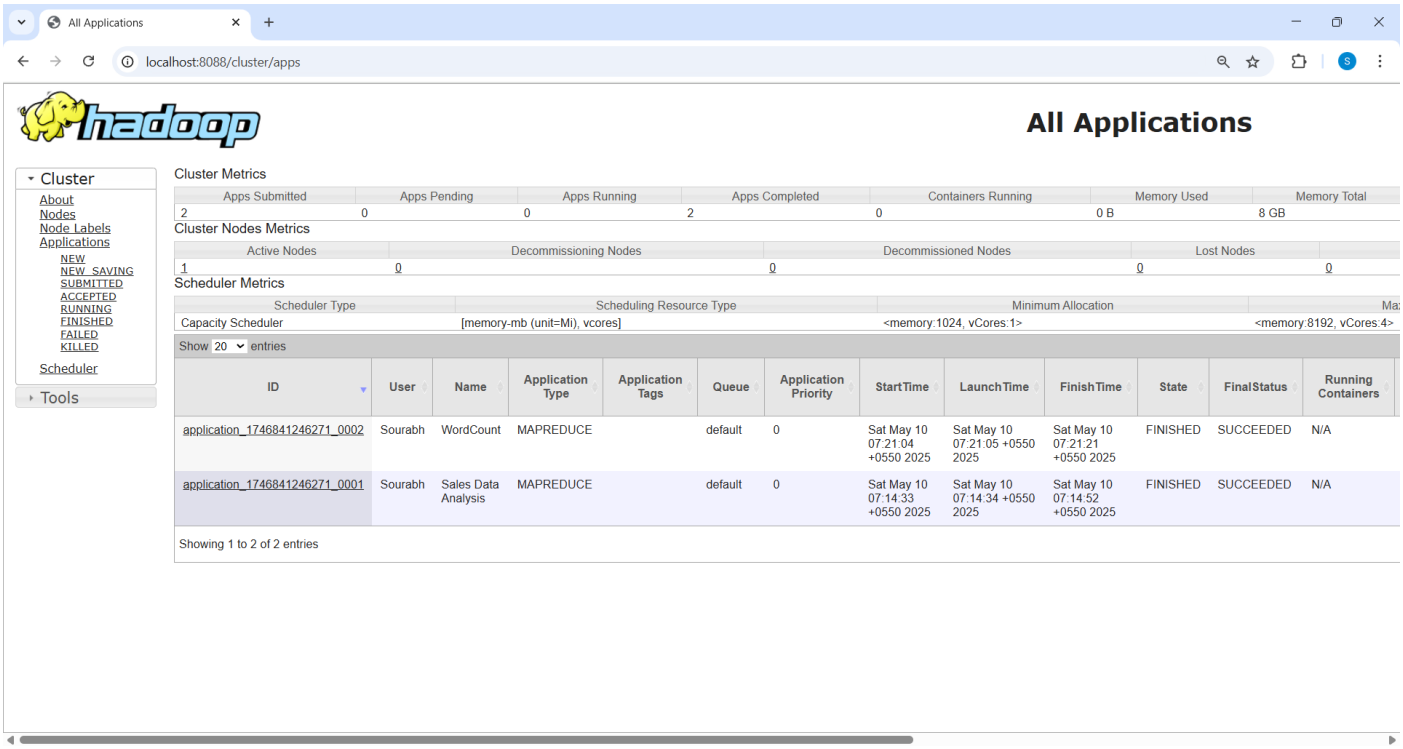

Browsing HDFS
Hadoop Safe Mode Issue
localhost:9870/explorer.html#/
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:41 0 0 B pgm2
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:46 0 0 B pgm2OP
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:54 0 0 B pgm3
drwxr-xr-x Sourabh supergroup 0 B Apr 25 18:57 0 0 B pgm3OP
drwxr-xr-x Sourabh supergroup 0 B May 03 11:52 0 0 B pgm4
drwxr-xr-x Sourabh supergroup 0 B May 03 11:55 0 0 B pgm4OP
drwxr-xr-x Sourabh supergroup 0 B May 09 18:52 0 0 B pgm8
drwxr-xr-x Sourabh supergroup 0 B May 09 18:55 0 0 B pgm8OP
drwxr-xr-x Sourabh supergroup 0 B May 09 19:22 0 0 B pgm9
drwxr-xr-x Sourabh supergroup 0 B May 09 19:32 0 0 B pgm9OP
drwx------ Sourabh supergroup 0 B Apr 19 12:23 0 0 B tmp
drwxr-xr-x Sourabh supergroup 0 B Apr 25 14:48 0 0 B user


Browsing HDFS
Hadoop Safe Mode Issue
localhost:9870/explorer.html#/pgm9
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm9
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r-- Sourabh supergroup 247 B May 09 19:22 1 128 MB sales.csv
Showing 1 to 1 of 1 entries
Previous
1
Next
Hadoop, 2020.


Browsing HDFS
Hadoop Safe Mode Issue
localhost:9870/explorer.html#/pgm9OP
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm9OP
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
Sourabh
supergroup
0 B
May 09 19:32
1
128 MB
_SUCCESS
-rw-r--r--
Sourabh
supergroup
63 B
May 09 19:32
1
128 MB
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.


Browsing HDFS
Hadoop Safe Mode Issue
localhost:9870/explorer.html#/pgm9OP
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/pgm9OP
Show 25 entries
Search:
Permission
Owner
Name
-rw-r--r--
_SUCCESS
-rw-r--r--
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2020.
File information - part-r-00000
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information --
Block 0
Block ID: 1073741981
Block Pool ID: BP-528767522-127.0.0.1-1745036485574
Generation Stamp: 1157
Size: 63
Availability:
DESKTOP-7FF817F
File contents
1001 650
1002 350
1003 100
1005 300
1006 400
1007 600
1009 500
Close


Administrator: Command Prompt
GC time elapsed (ms)=51
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=531103744
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=247
File Output Format Counters
Bytes Written=63
D:\hadoop\sbin>hadoop fs -cat /pgm9OP/part-r-00000
1001 650
1002 350
1003 100
1005 300
1006 400
1007 600
1009 500
27°C
Clear
Search
ENG
IN
19:38
09-05-2025


All Applications
localhost:8088/cluster/apps
hadoop
All Applications
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Cluster Metrics
Apps Submitted 2
Apps Pending 0
Apps Running 0
Apps Completed 2
Containers Running 0
Memory Used 0 B
Memory Total 8 GB
Cluster Nodes Metrics
Active Nodes 1
Decommissioning Nodes 0
Decommissioned Nodes 0
Lost Nodes 0
Scheduler Metrics
Scheduler Type Capacity Scheduler
Scheduling Resource Type [memory-mb (unit=Mi), vcores]
Minimum Allocation <memory:1024, vCores:1>
Show 20 entries
ID User Name Application Type Application Tags Queue Application Priority StartTime LaunchTime FinishTime State FinalStatus Running Containers
application_1746841246271_0002 Sourabh WordCount MAPREDUCE default 0 Sat May 10 07:21:04 +0550 2025 Sat May 10 07:21:05 +0550 2025 Sat May 10 07:21:21 +0550 2025 FINISHED SUCCEEDED N/A
application_1746841246271_0001 Sourabh Sales Data Analysis MAPREDUCE default 0 Sat May 10 07:14:33 +0550 2025 Sat May 10 07:14:34 +0550 2025 Sat May 10 07:14:52 +0550 2025 FINISHED SUCCEEDED N/A
Showing 1 to 2 of 2 entries