SYMCA-B
Roll No-90

Practical 1

Aim: List of Commands (mkdir, touchz, copy from local/put, copy to local/get move from local, cp, rmr, du, dus, stat)

1) hadoop fs

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System

File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 3 items
-rw-r--r--   1 cloudera cloudera          0 2024-09-12 00:18 Demo.txt
drwxr-xr-x   - cloudera cloudera          0 2024-07-18 18:46 Demodirector
drwxr-xr-x   - cloudera cloudera          0 2024-07-18 18:46 Demodirectory
```

2) touchz: It creates an empty file.

```
[cloudera@quickstart ~]$ hadoop fs -touchz Demo.txt
[cloudera@quickstart ~]$ hadoop fs -ls
Found 3 items
-rw-r--r--   1 cloudera cloudera          0 2024-09-12 00:20 Demo.txt
drwxr-xr-x   - cloudera cloudera          0 2024-07-18 18:46 Demodirector
drwxr-xr-x   - cloudera cloudera          0 2024-07-18 18:46 Demodirectory
```

3) copyFromLocal (or) put:

```
[cloudera@quickstart ~]$ hadoop fs -copyFromLocal test.txt Demo.txt
copyFromLocal: `Demo.txt': File exists
```

4) test

5) mkdir

```
[cloudera@quickstart ~]$ hadoop fs -mkdir Demodirectory1
```

6) appendToFile

```
[cloudera@quickstart ~]$ gedit test1.txt
[cloudera@quickstart ~]$ gedit test2.txt
[cloudera@quickstart ~]$ hadoop fs -touchz Demo.txt
[cloudera@quickstart ~]$ hadoop fs -appendToFile test1.txt test2.txt Demo.txt
[cloudera@quickstart ~]$ hadoop fs -cat Demo.txt
test 1
test 2
```

7) usage

```
[cloudera@quickstart ~]$ hadoop fs -usage test
Usage: hadoop fs [generic options] -test -[defsz] <path>
```

8) Count

```
[cloudera@quickstart ~]$ hadoop fs -count -v/
-count: Illegal option -v/
Usage: hadoop fs [generic options] -count [-q] [-h] [-v] [-x] <path> ...
```

9) find

```
[cloudera@quickstart ~]$ hadoop fs -find / -name Demod
```

10) help

```
[cloudera@quickstart ~]$ hadoop fs -help count
-count [-q] [-h] [-v] [-x] <path> ... :
  Count the number of directories, files and bytes under the paths
  that match the specified file pattern.  The output columns are:
  DIR_COUNT FILE_COUNT CONTENT_SIZE PATHNAME
  or, with the -q option:
  QUOTA REM_QUOTA SPACE_QUOTA REM_SPACE_QUOTA
        DIR_COUNT FILE_COUNT CONTENT_SIZE PATHNAME
  The -h option shows file sizes in human readable format.
  The -v option displays a header line.
  The -x option excludes snapshots from being calculated.
[cloudera@quickstart ~]$
```

cloudera@quickstart:~

PRACTICAL 2

Aim: write a Program in Map Reduce for Word Count operation.

WordCountDriver.java

```
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.conf.Configuration;
public class WordCountDriver {
	public static void main(String[] args) throws Exception{
		Job j1=Job.getInstance(new Configuration());
		j1.setJarByClass(WordCountDriver.class);
		j1.setJobName("Average Word Count");

		FileInputFormat.addInputPath(j1,new Path(args[0]));
		FileOutputFormat.setOutputPath(j1, new Path(args[1]));

		j1.setMapperClass(WordCountMapper.class);
		j1.setReducerClass(WordCountReducer.class);
		j1.setOutputKeyClass(Text.class);
		j1.setOutputValueClass(IntWritable.class);
		System.exit(j1.waitForCompletion(true)? 0:1);
	}
}
```

WordCountMapper.java

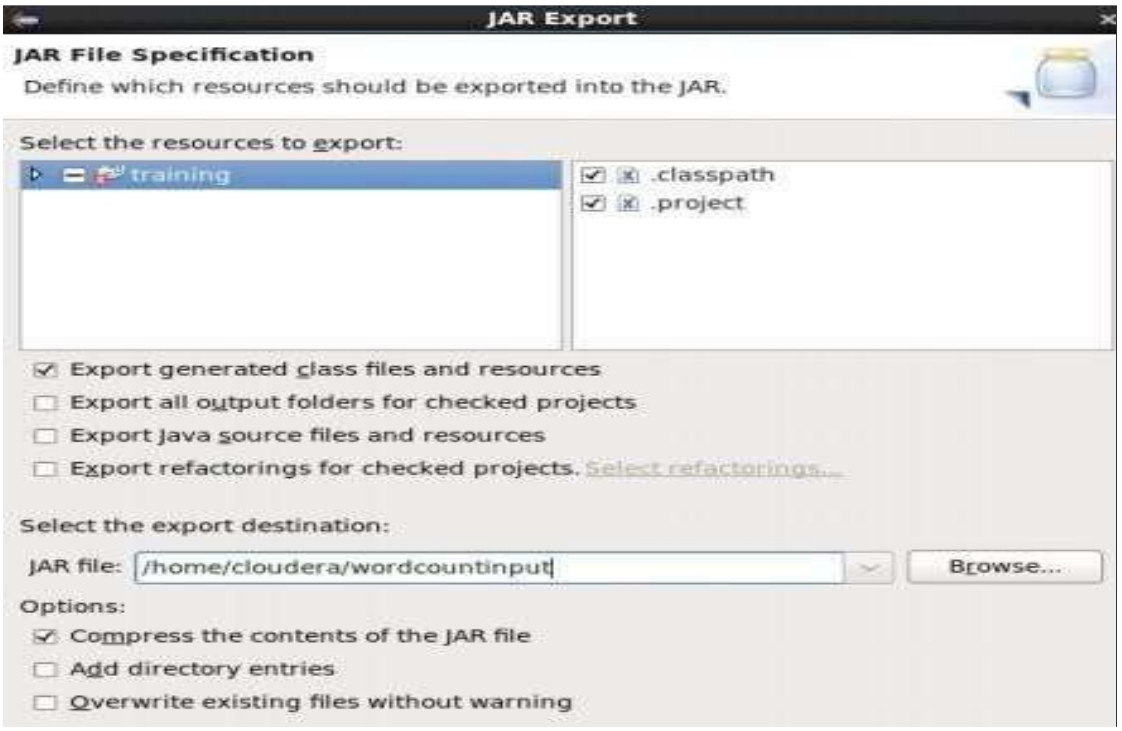
```
import java.io.IOException;
import java.util.StringTokenizer;0
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.Reducer.Context;

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 {
```

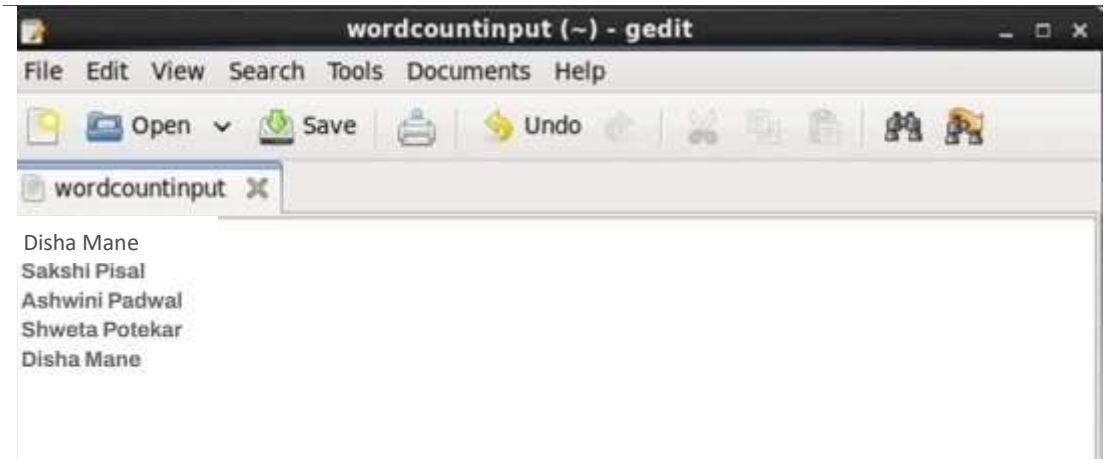
```
            String line=value.toString();
            StringTokenizer ltr=new StringTokenizer(line);
            while(ltr.hasMoreTokens()){
                word.set(ltr.nextToken());
                context.write(word ,one);
            }
        }
}
```

WordCountReducer.java

```
import java.io.IOException;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.IntWritable;
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 value:values)
            {
                sum+=value.get();
            }
            context.write(key,  new IntWritable(sum));
        }
}
```

Export .jar file. Right click on training and select export.






```
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -copyFromLocal wordcountinput /wordcountinputhdfss
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hadoop fs cat /wordcountinputhdfss
cat: Unknown command
Did you mean -cat?  This command begins with a dash.
[cloudera@quickstart ~]$ hadoop fs -cat /wordcountinputhdfss
```

Disha Mane
Saskshi Pisal
Ashwini Padwal
Shweta Potekar
Disha Mane

Executing the jar file using hadoop command:

```
[cloudera@quickstart ~]$ sudo -u hdfs hadoop jar wordcountinput.jar WordCountDriver /wordcountinputhdfss /wordcountoutputdir3
22/11/13 03:27:57 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/10.0.2.15:8032
22/11/13 03:27:58 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool inte
d execute your application with ToolRunner to remedy this.
22/11/13 03:27:58 INFO input.FileInputFormat: Total input paths to process : 1
22/11/13 03:27:58 INFO mapreduce.JobSubmitter: number of splits:1
22/11/13 03:27:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1668337644684_0005
22/11/13 03:27:58 INFO impl.YarnClientImpl: Submitted application application_1668337644684_0005
22/11/13 03:27:58 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1668337644684_
22/11/13 03:27:58 INFO mapreduce.Job: Running job: job_1668337644684_0005
22/11/13 03:28:06 INFO mapreduce.Job: Job job_1668337644684_0005 running in uber mode : false
22/11/13 03:28:06 INFO mapreduce.Job:  map 0% reduce 0%
22/11/13 03:28:12 INFO mapreduce.Job:  map 100% reduce 0%
22/11/13 03:28:18 INFO mapreduce.Job:  map 100% reduce 100%
22/11/13 03:28:19 INFO mapreduce.Job: Job job_1668337644684_0005 completed successfully
22/11/13 03:28:19 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=146
                FILE: Number of bytes written=256611
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=224
                HDFS: Number of bytes written=88
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
```

To output the file

```
[cloudera@quickstart ~]$ hadoop fs -cat /wordcountoutputdir3/part-r-00000
Ashish  3
B       2
Komal   1
Poonam  2
Rujul   2
Shagun  2
Yash    2
hello   1
kirtee  1
komal   1
poonam  1
[cloudera@quickstart ~]$
```

Aim: write a Program in Map Reduce for Matrix Multiplication.

MatrixMultiplication.java

```
import java.io.DataInput;
import java.io.DataOutput;
import java.io.IOException;
import java.util.ArrayList;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Writable;
import org.apache.hadoop.io.WritableComparable;
import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.*;
import org.apache.hadoop.mapreduce.lib.output.*;
import org.apache.hadoop.util.ReflectionUtils;

class Element implements Writable {
	int tag;
	int index;
	double value;
	Element() {
		tag = 0;
		index = 0;
		value = 0.0;
	}
	Element(int tag, int index, double value) {
		this.tag = tag;
		this.index = index;
		this.value = value;
	}
	@Override
	public void readFields(DataInput input) throws IOException {
		tag = input.readInt();
		index = input.readInt();
		value = input.readDouble();
	}
	@Override
	public void write(DataOutput output) throws IOException {
		output.writeInt(tag);
		output.writeInt(index);
		output.writeDouble(value);
	}
```

```
}
class Pair implements WritableComparable<Pair> {
	int i;
	int j;
	Pair() {
		i = 0;
		j = 0;
	}
	Pair(int i, int j) {
		this.i = i;
		this.j = j;
	}
	@Override
	public void readFields(DataInput input) throws IOException {
		i = input.readInt();
		j = input.readInt();
	}
	@Override
	public void write(DataOutput output) throws IOException {

		output.writeInt(i);
		output.writeInt(j);
	}
	@Override
	public int compareTo(Pair compare) {
		if (i > compare.i) {
			return 1;
		}
		else if (i < compare.i) {
			return -1;
		}
		else {
			if (j > compare.j) {
				return 1;
			}
			else if (j < compare.j) {
				return -1;
			}
		}
		return 0;
	}
	public String toString() {
		return i + " " + j + " ";
	}
}
public class MatrixMultiply {
	public static class MatrixMapperM extends
			Mapper<Object, Text, IntWritable, Element> {
		@Override
		public void map(Object key, Text value, Context context)
```

```
                throws IOException, InterruptedException {
            String readLine = value.toString();
            String[] tokens = readLine.split(",");
            int index = Integer.parseInt(tokens[0]);
            double elementVal = Double.parseDouble(tokens[2]);
            Element e = new Element(0, index, elementVal);
            IntWritable keyval = new IntWritable(Integer.parseInt(tokens[1]));
            context.write(keyval, e);
        }
    }
    public static class MatrixMapperN extends
            Mapper<Object, Text, IntWritable, Element> {

        @Override
        public void map(Object key, Text value, Context context)
                throws IOException, InterruptedException {
            String readLine = value.toString();
            String[] tokens = readLine.split(",");
            int index = Integer.parseInt(tokens[1]);
            double elementVal = Double.parseDouble(tokens[2]);
            Element e = new Element(1, index, elementVal);
            IntWritable keyval = new IntWritable(Integer.parseInt(tokens[0]));
            context.write(keyval, e);
        }
    }

    public static class ReducerMN extends
            Reducer<IntWritable, Element, Pair, DoubleWritable> {

        @Override
        public void reduce(IntWritable key, Iterable<Element> values,
                Context context) throws IOException, InterruptedException {
            ArrayList<Element> M = new ArrayList<Element>();
            ArrayList<Element> N = new ArrayList<Element>();
            Configuration conf = context.getConfiguration();
            for (Element element : values) {
                Element temp = ReflectionUtils.newInstance(Element.class, conf);
                ReflectionUtils.copy(conf, element, temp);
                if (temp.tag == 0)
                {
                    M.add(temp);
                }
                else if (temp.tag == 1)
                {
                    N.add(temp);
                }
            }
            for (int i = 0; i < M.size(); i++) {
                for (int j = 0; j < N.size(); j++) {
                    Pair p = new Pair(M.get(i).index, N.get(j).index);
```

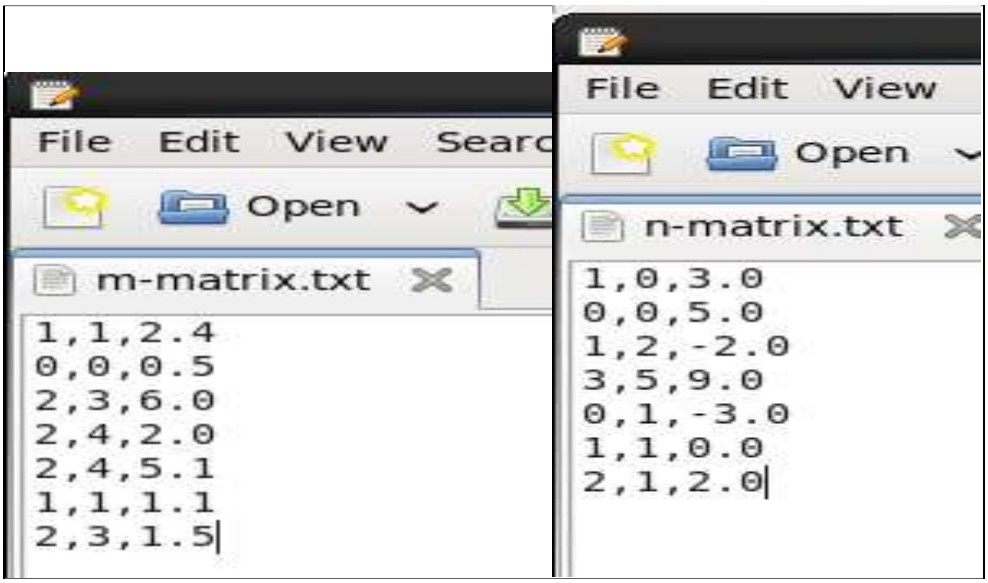
```
                    double mul = M.get(i).value * N.get(j).value;

                    context.write(p, new DoubleWritable(mul));
                }
            }
        }

    }
    public static class MapMN extends
            Mapper<Object, Text, Pair, DoubleWritable> {
        @Override
        public void map(Object key, Text value, Context context)
                throws IOException, InterruptedException {
            String readLine = value.toString();
            String[] pairValue = readLine.split(" ");
            Pair p = new Pair(Integer.parseInt(pairValue[0]),
                    Integer.parseInt(pairValue[1]));
            DoubleWritable val = new DoubleWritable(
                    Double.parseDouble(pairValue[2]));
            context.write(p, val);
        }
    }
    public static class ReduceMN extends
            Reducer<Pair, DoubleWritable, Pair, DoubleWritable> {
        @Override
        public void reduce(Pair key, Iterable<DoubleWritable> values,
                Context context) throws IOException, InterruptedException {
            double sum = 0.0;
            for (DoubleWritable value : values) {
                sum += value.get();
            }
            context.write(key, new DoubleWritable(sum));
        }
    }
    public static void main(String[] args) throws Exception {
        Path MPath = new Path("/expt4/input/M");
        Path NPath = new Path("/expt4/input/N");
        Path intermediatePath = new Path("/expt4/interim");
        Path outputPath = new Path("/expt4/output");
        Job job1 = Job.getInstance();
        job1.setJobName("Map Intermediate");
        job1.setJarByClass(MatrixMultiply.class);
        MultipleInputs.addInputPath(job1, MPath, TextInputFormat.class,
                MatrixMapperM.class);

        MultipleInputs.addInputPath(job1, NPath, TextInputFormat.class,
                MatrixMapperN.class);
        job1.setReducerClass(ReducerMN.class);
        job1.setMapOutputKeyClass(IntWritable.class);
        job1.setMapOutputValueClass(Element.class);
```

```
        job1.setOutputKeyClass(Pair.class);
        job1.setOutputValueClass(DoubleWritable.class);
        job1.setOutputFormatClass(TextOutputFormat.class);
        FileOutputFormat.setOutputPath(job1, intermediatePath);
        job1.waitForCompletion(true);
        Job job2 = Job.getInstance();
        job2.setJobName("Map Final Output");
        job2.setJarByClass(MatrixMultiply.class);
        job2.setMapperClass(MapMN.class);
        job2.setReducerClass(ReduceMN.class);
        job2.setOutputKeyClass(Pair.class);
        job2.setOutputValueClass(DoubleWritable.class);
        job2.setInputFormatClass(TextInputFormat.class);
        job2.setOutputFormatClass(TextOutputFormat.class);
        FileInputFormat.addInputPath(job2, intermediatePath);
        FileOutputFormat.setOutputPath(job2, outputPath);
        job2.waitForCompletion(true);
    }

}
```

Prerequisites create the input directories to store the input matrices M and N

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ hdfs dfs -mkdir /expt4
mkdir: Permission denied: user=cloudera, access=WRITE, inode="/":hdfs:supergroup:drwxr-xr-x
[cloudera@quickstart ~]$ sudo -u hdfs hdfs dfs -mkdir /expt4
[cloudera@quickstart ~]$ sudo -u hdfs hdfs dfs -mkdir /expt4/input
[cloudera@quickstart ~]$ sudo -u hdfs hdfs dfs -mkdir /expt4/input/M
[cloudera@quickstart ~]$ sudo -u hdfs hdfs dfs -mkdir /expt4/input/N
[cloudera@quickstart ~]$ hdfs dfs -ls /expt4/input
Found 2 items
drwxr-xr-x   - hdfs supergroup          0 2022-11-12 10:04 /expt4/input/M
drwxr-xr-x   - hdfs supergroup          0 2022-11-12 10:04 /expt4/input/N
```

The figure below shows the matrix data used for this implementation:



Copy the matrix data from the local system to HDFS

```
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ gedit m-matrix
[cloudera@quickstart ~]$ gedit m-matrix.txt
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ gedit n-matrix
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ gedit n-matrix.txt
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hdfs dfs -copyFromLocal m-matrix.txt /expt4/input/M
copyFromLocal: Permission denied: user=cloudera, access=WRITE, inode="/expt4/input/M":hdfs:supergroup:drwxr-xr-x
[cloudera@quickstart ~]$ sudo -u hdfs hdfs dfs -copyFromLocal m-matrix.txt /expt4/input/M
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo -u hdfs hdfs dfs -copyFromLocal n-matrix.txt /expt4/input/N
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hdfs dfs -ls /expt4/input/M
Found 1 items
-rw-r--r--   1 hdfs supergroup         56 2022-11-12 10:12 /expt4/input/M/m-matrix.txt
[cloudera@quickstart ~]$ hdfs dfs -ls /expt4/input/N
Found 1 items
-rw-r--r--   1 hdfs supergroup         58 2022-11-12 10:12 /expt4/input/N/n-matrix.txt
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ javac MatrixMultiplication.java -cp $(hadoop classpath)
javac: file not found: MatrixMultiplication.java
Usage: javac <options> <source files>
use -help for a list of possible options
[cloudera@quickstart ~]$
```

Compile the code to create the classes

```
cloudera@quickstart:~/workspace/training/src
File Edit View Search Terminal Help
[cloudera@quickstart src]$ javac MatrixMultiplication.java -cp $(hadoop classpath)
[cloudera@quickstart src]$
[cloudera@quickstart src]$
[cloudera@quickstart src]$ gedit Manifest.txt
[cloudera@quickstart src]$
[cloudera@quickstart src]$
```

To indicate the main class file, create a Manifest file to point out to the main driver class.

```
*Manifest.txt (~/workspace/training/src) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
*Manifest.txt
Main-class: MatrixMultiplication
```

Compile and create the Jar file required to run the MapReduce Task

```
cloudera@quickstart:~/workspace/training/src
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart src]$ jar -cfm MatrixMultiplication.jar Manifest.txt *.class
[cloudera@quickstart src]$
```

Run the jar file on the Hadoop ecosystem to trigger all the MapReduce classes.

```
[cloudera@quickstart src]$ sudo -u hdfs hadoop jar MatrixMultiplication.jar
22/11/12 10:29:23 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/127.0.0.1:8032
22/11/12 10:29:24 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface an
d execute your application with ToolRunner to remedy this.
22/11/12 10:29:24 INFO input.FileInputFormat: Total input paths to process : 1
22/11/12 10:29:24 INFO input.FileInputFormat: Total input paths to process : 1
22/11/12 10:29:24 INFO mapreduce.JobSubmitter: number of splits:2
22/11/12 10:29:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1668275009795_0006
22/11/12 10:29:24 INFO impl.YarnClientImpl: Submitted application application_1668275009795_0006
22/11/12 10:29:24 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1668275009795_0006/
22/11/12 10:29:24 INFO mapreduce.Job: Running job: job_1668275009795_0006
22/11/12 10:29:31 INFO mapreduce.Job: Job job_1668275009795_0006 running in uber mode : false
22/11/12 10:29:31 INFO mapreduce.Job:  map 0% reduce 0%
22/11/12 10:29:41 INFO mapreduce.Job:  map 100% reduce 0%
22/11/12 10:29:48 INFO mapreduce.Job:  map 100% reduce 100%
22/11/12 10:29:48 INFO mapreduce.Job: Job job_1668275009795_0006 completed successfully
22/11/12 10:29:48 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=175
                FILE: Number of bytes written=387639
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=640
                HDFS: Number of bytes written=124
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
```

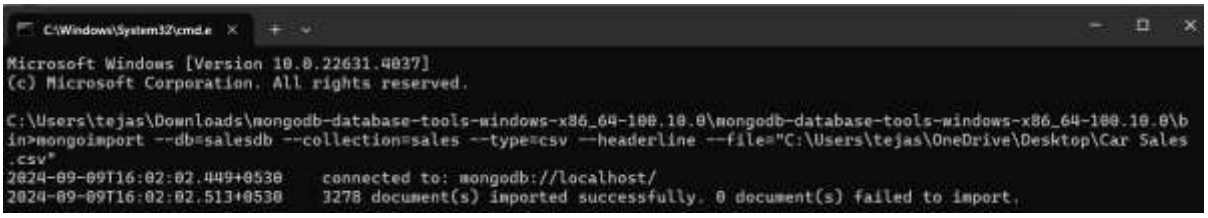
Output of the file

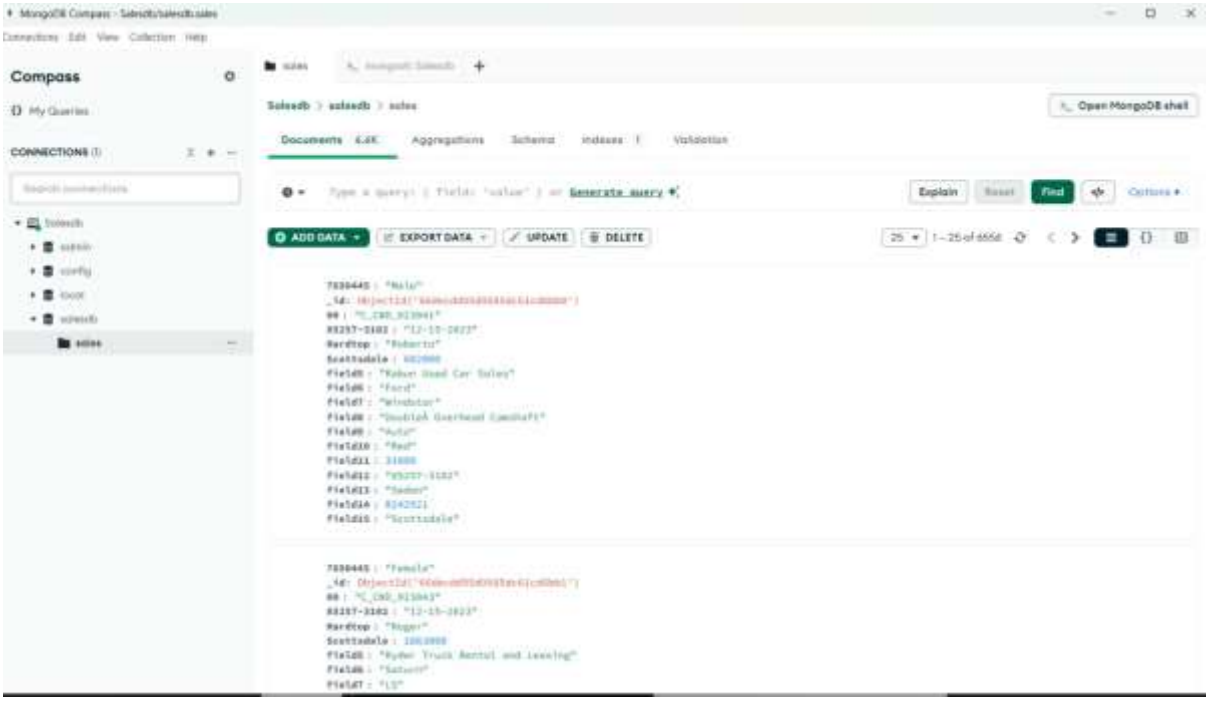
```
cloudera@quickstart:~/workspace/training
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart training]$ hdfs dfs -cat /expt4/interim/part-r-00000
0 1     -1.5
0 0     2.5
1 1     0.0
1 2     -2.2
1 0     3.3000000000000003
1 1     0.0
1 2     -4.8
1 0     7.199999999999999
2 5     13.5
2 5     54.0
[cloudera@quickstart training]$ hdfs dfs -cat /expt4/output/part-r-00000
0 0     2.5
0 1     -1.5
1 0     10.5
1 1     0.0
1 2     -7.0
2 5     67.5
[cloudera@quickstart training]$
```

PRACTICAL 3

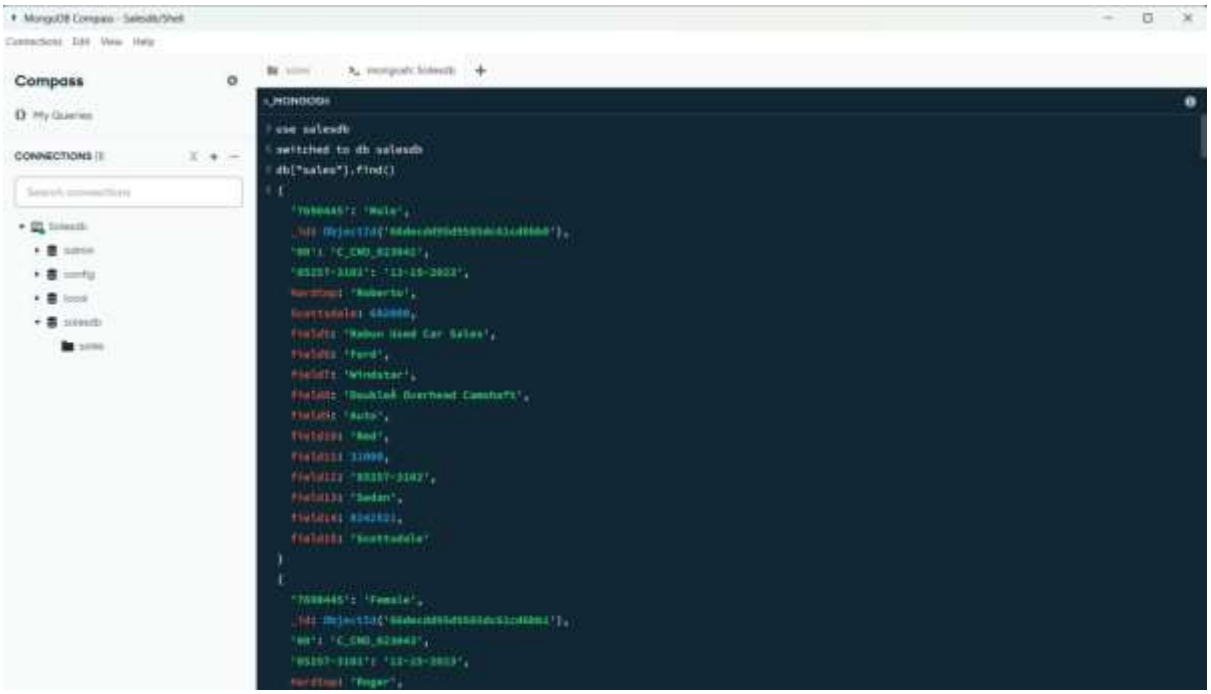
Aim: Query the Sample Database using MongoDB querying commands.

Query :mongoimport --db=salesdb --collection=sales --type=csv --headerline --file="C:\Users\tejas\Desktop\Car Sales.csv"

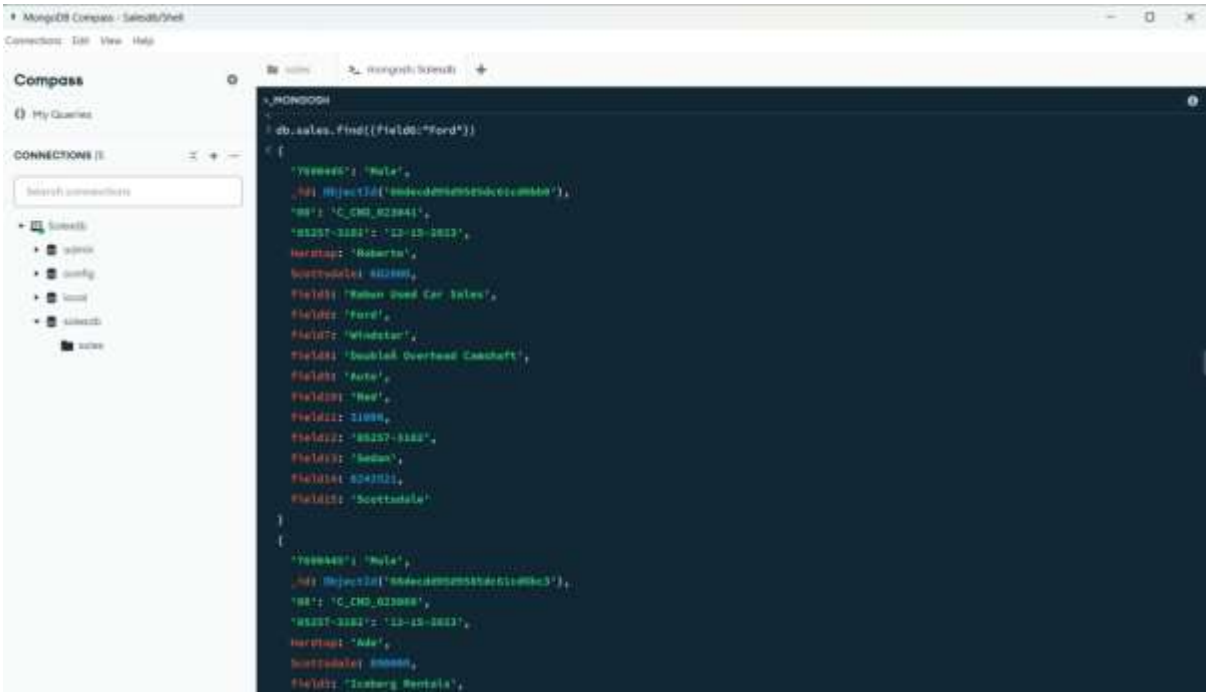




Find data query
Query : db.sales.find()



Db.sales.find({field6:"Ford"})



Update document

db.sales.updateMany({ field6 : "Toyota"}, { $set :{ field:"BMW" }})

```
db.sales.updateMany( { field6:"Toyota"}, { $set :{ field6:"BMW" }})
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 278,
  modifiedCount: 278,
  upsertedCount: 0
```

Delete document
db.sales.deleteOne({field6:”Ford” })

```
> db.sales.deleteOne({ field6: "Ford" })
< {
    acknowledged: true,
    deletedCount: 1
  }
salesdb>
```

Aggregate function

db.sales.aggregate([{$group: {_id: "$CarMake", totalSales: { $sum: 1 }}, {$sort: { totalSales: -1 } }])

```
> db.sales.aggregate([ { $group: {_id: "$CarMake", totalSales: { $sum: 1 }} },{$sort: { totalSales: -1 } }])
< {
    _id: null,
    totalSales: 6555
  }
salesdb>
```

Extract the salesdb database in json

mongoexport --collection=sales --db=salesdb --out=sales.json

```
C:\Windows\System32\cmd.e
Microsoft Windows [Version 10.0.22631.4037]
(c) Microsoft Corporation. All rights reserved.

C:\Users\tejas\Downloads\mongodb-database-tools-windows-x86_64-100.10.0\mongodb-database-tools-windows-x86_64-100.10.0\b
in>mongoexport --collection=sales --db=salesdb --out=sales.json
2024-09-09T16:27:24.326+0530    connected to: mongodb://localhost/
2024-09-09T16:27:24.520+0530    exported 6555 records

C:\Users\tejas\Downloads\mongodb-database-tools-windows-x86_64-100.10.0\mongodb-database-tools-windows-x86_64-100.10.0\b
in>
```

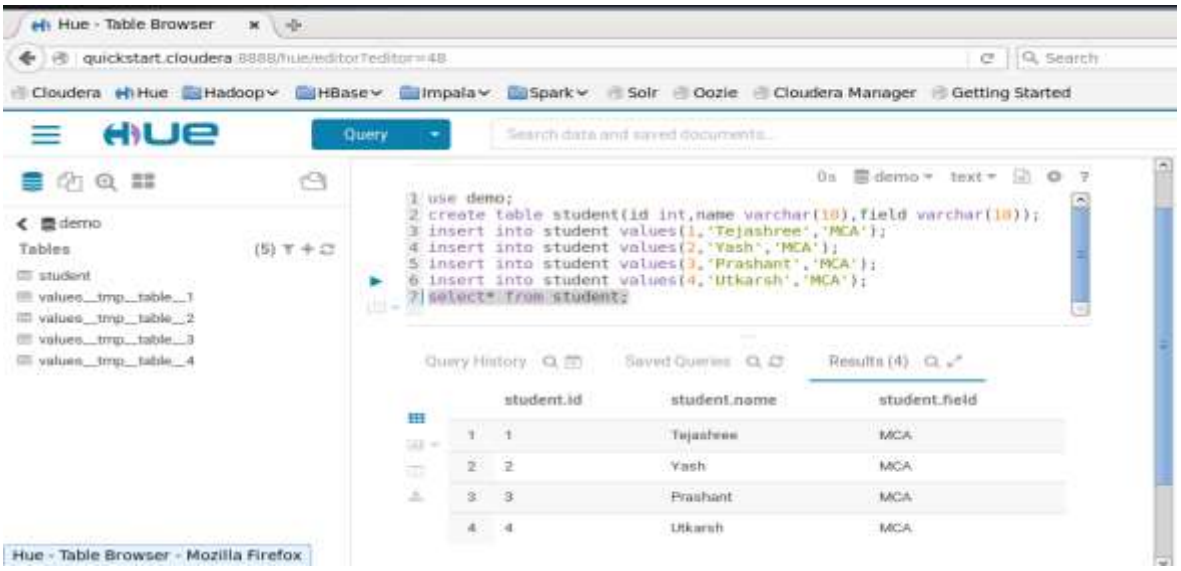
Practical 4

Aim: Basic Hive Commands

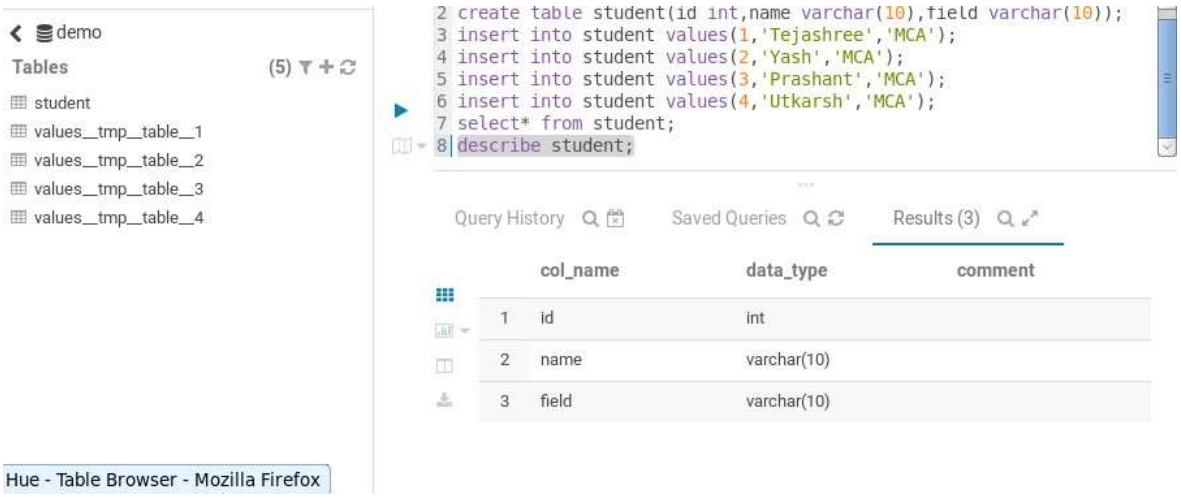
```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ sudo -uhivehive
sudo: unknown user: hivehive
sudo: unable to initialize policy plugin
[cloudera@quickstart ~]$ sudo -uhive hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> show databases;
OK
default
demo
Time taken: 0.321 seconds, Fetched: 2 row(s)
hive> use demo;
OK
Time taken: 0.019 seconds
hive> select * from student;
OK
1       Disha Mane MCA
2       Anket Kohak  MCA
3       Hemraj Gupta MCA
4       Siddhi Rane MCA
Time taken: 0.332 seconds, Fetched: 4 row(s)
hive> show tables;
```

```
hive> show tables;
OK
student
Time taken: 0.022 seconds, Fetched: 1 row(s)
hive> describe student;
OK
id                      int
name                    varchar(10)
field                   varchar(10)
Time taken: 0.037 seconds, Fetched: 3 row(s)
hive>
```

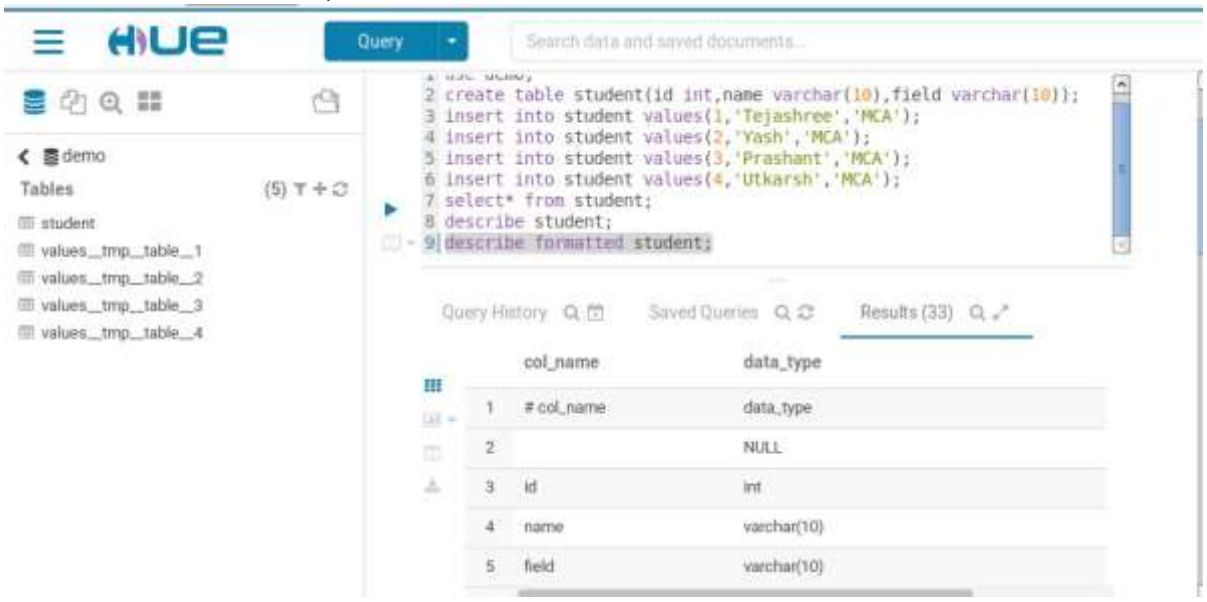
```
use demo;
create table student(id int,name varchar(10),field varchar(10));
insert into student values(1,'Tejashree','MCA');
insert into student values(2,'Yash','MCA');
insert into student values(3,'Prashant','MCA');
insert into student values(4,'Utkarsh','MCA');
select* from student:
```



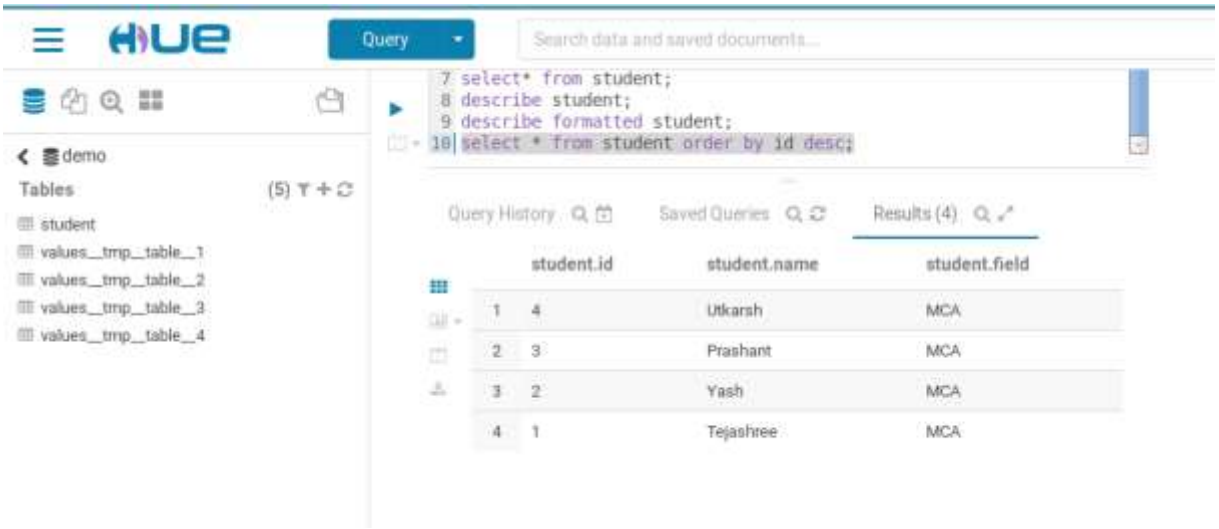
Describe student;



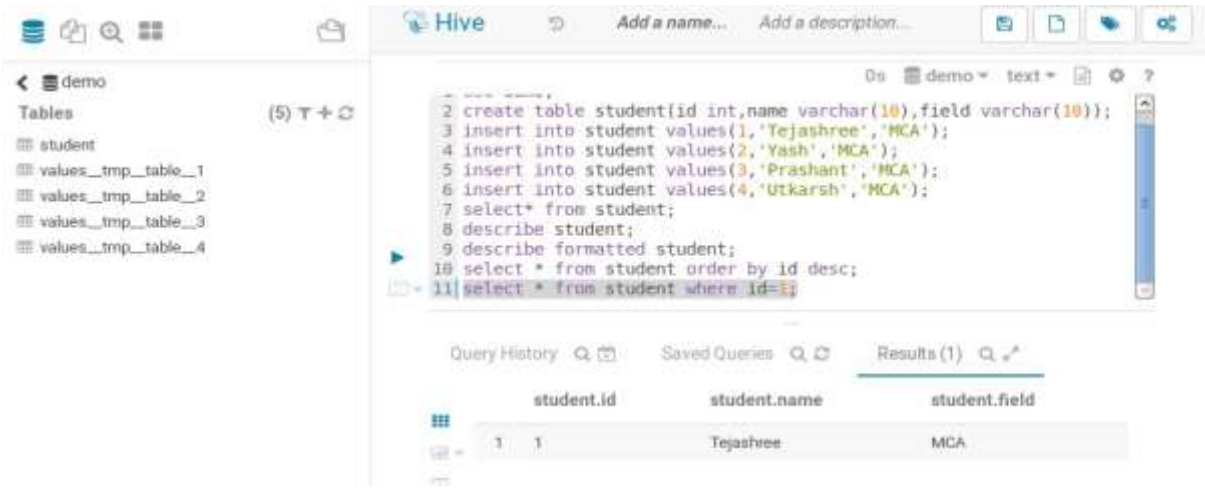
describe formatted student;



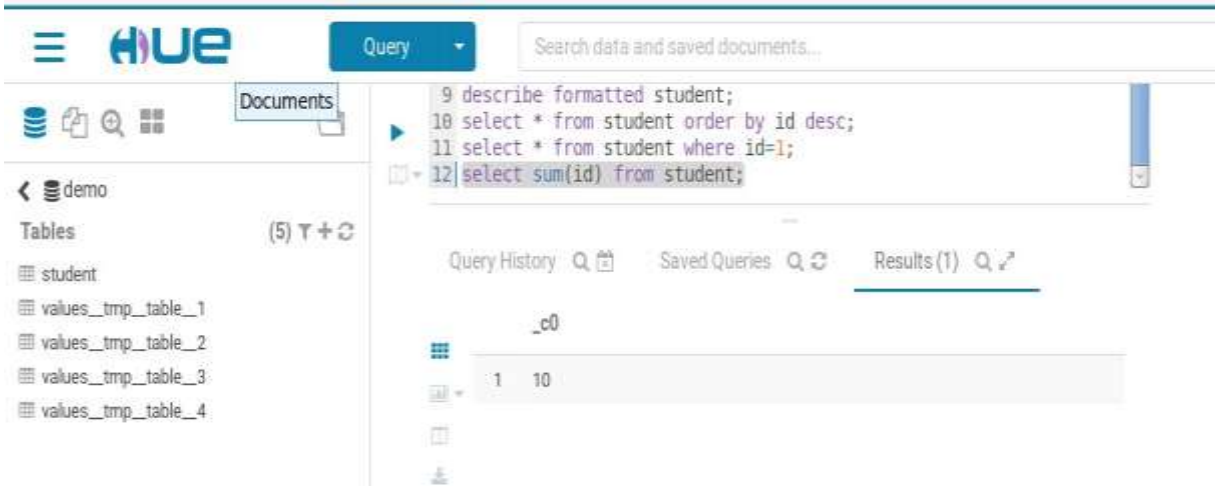
student order by id desc;



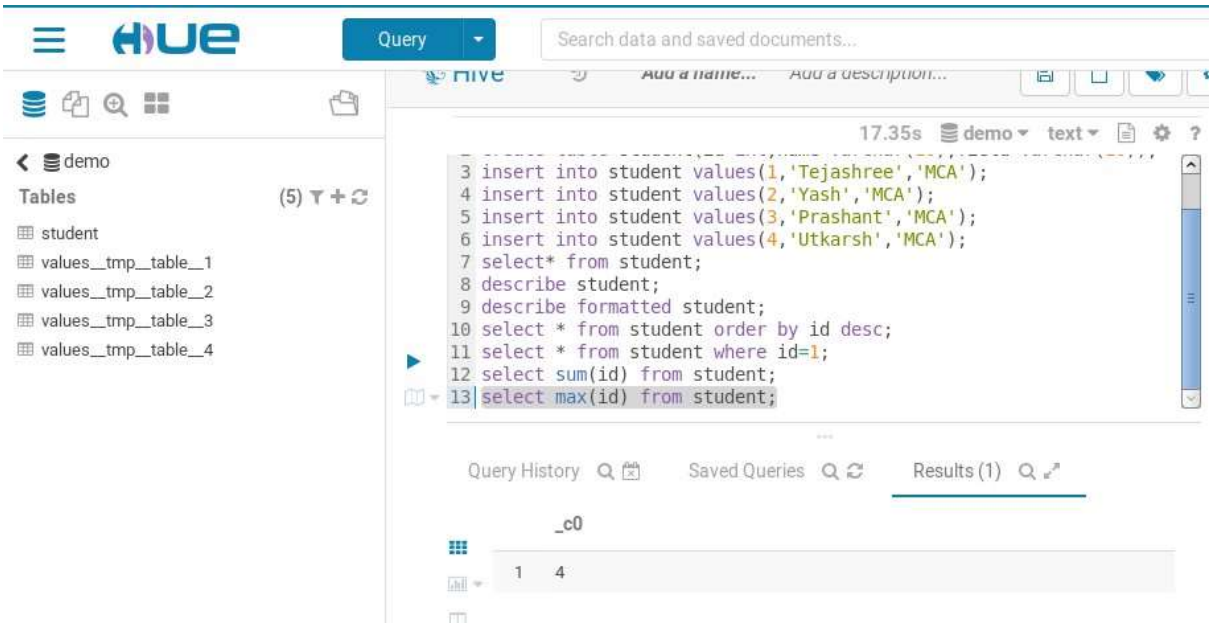
select * from student where id=2;



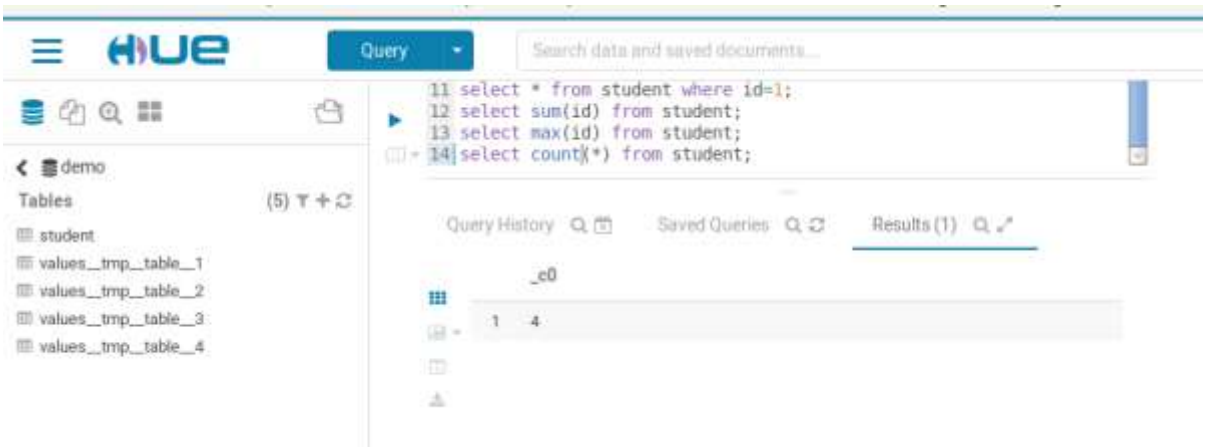
select sum(id) from student;



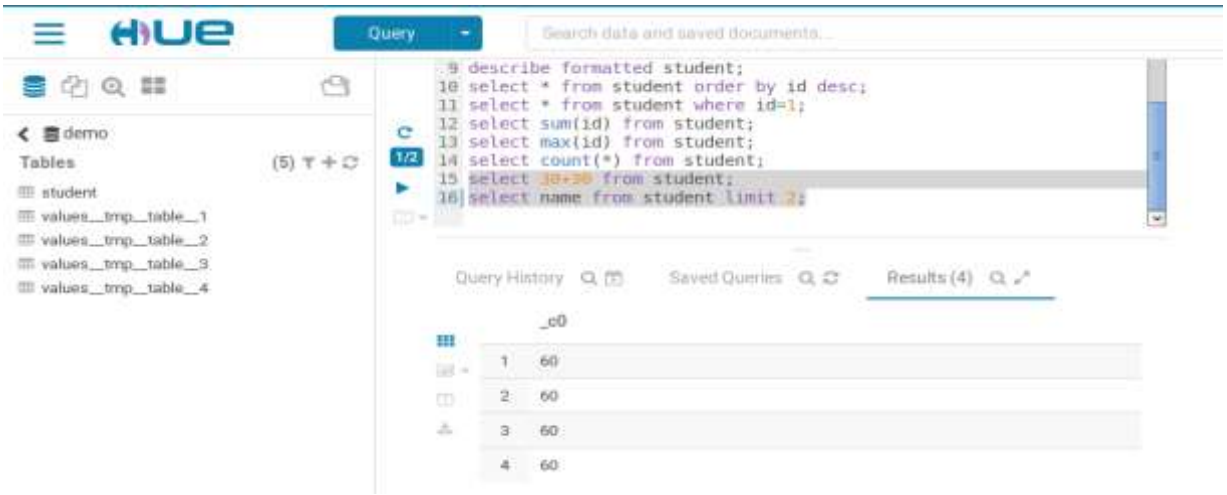
select max(id) from student;



select count(*) from student



select 30+30 from student;
select name from student limit 2;

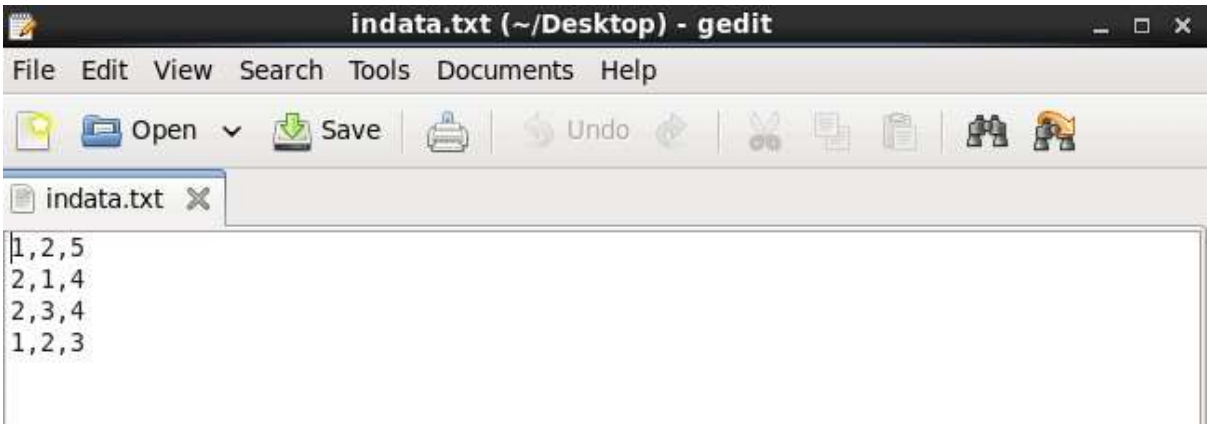


SYMCA-B
Roll No-90

Practical 5

Aim: PIG List of Commands
Student1.txt

Indata.txt



Pig –x local

```
grunt> data = load '/home/cloudera/Desktop/indata.txt'using PigStorage(',') as(A1:int,A2:int,A3:int);
grunt> dump data;
2024-09-18 23:09:02,630 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-09-18 23:09:02,654 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachCol
umnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownFor
achFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-18 23:09:02,720 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-18 23:09:02,732 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-18 23:09:02,733 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-18 23:09:02,747 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - session.id is deprecated. Instead, use dfs.metrics.session-id
2024-09-18 23:09:02,747 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Initializing JVM Metrics with processName=JobTracker, sessionId=
2024-09-18 23:09:02,761 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-18 23:09:02,801 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduc
.reduce.markreset.buffer.percent
2024-09-18 23:09:02,801 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is no
 set, set to default 0.3
2024-09-18 23:09:02,801 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputf
rmat.compress
2024-09-18 23:09:02,820 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-09-18 23:09:02,831 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-18 23:09:02,831 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2024-09-18 23:09:02,831 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.lo
al.dir] with code temp directory: /tmp/1726726142831-0
```

Load data and dump data

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help

Input(s):
Successfully read records from: "/home/cloudera/Desktop/indata.txt"

Output(s):
Successfully stored records in: "file:/tmp/temp2037139829/tmp-61834499"

Job DAG:
job_local486859717_0001


2024-09-18 23:09:21,371 [main] INFO  org.apache.pig.backend.hadoop.executionengi
2024-09-18 23:09:21,376 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
2024-09-18 23:09:21,376 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
2024-09-18 23:09:21,376 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
2024-09-18 23:09:21,376 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
2024-09-18 23:09:21,387 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
2024-09-18 23:09:21,387 [main] INFO  org.apache.pig.backend.hadoop.executionengi
(1,2,5)
(2,1,4)
(2,3,4)
(1,2,3)
(,,)
grunt>
```

Filter

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
grunt> fidata= FILTER data by A3==4;
grunt> dump fidata;
2024-09-18 23:19:43,673 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: FILTER
2024-09-18 23:19:43,673 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-18 23:19:43,675 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2024-09-18 23:19:43,676 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-18 23:19:43,676 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-18 23:19:43,676 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2024-09-18 23:19:43,677 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig script settings are added to the job
2024-09-18 23:19:43,680 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percen
```

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
Successfully stored records in: "file:/tmp/temp2037139829/tmp-1877556993"

Job DAG:
job_local54881604_0007


2024-09-18 23:19:56,196 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2024-09-18 23:19:56,197 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-18 23:19:56,197 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-09-18 23:19:56,197 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-18 23:19:56,198 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2024-09-18 23:19:56,206 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2024-09-18 23:19:56,206 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(2,1,4)
(2,3,4)
grunt>
```

Distinct :

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
grunt> result = DISTINCT data;
grunt> dump result;
2024-09-18 23:21:22,429 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: DISTINCT
2024-09-18 23:21:22,429 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-18 23:21:22,433 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2024-09-18 23:21:22,436 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-18 23:21:22,436 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-18 23:21:22,436 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2024-09-18 23:21:22,436 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig script settings are added to the job
2024-09-18 23:21:22,438 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percen
```

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
job_local1113651125_0008


2024-09-18 23:21:35,017 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2024-09-18 23:21:35,018 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-18 23:21:35,018 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-09-18 23:21:35,019 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-18 23:21:35,019 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2024-09-18 23:21:35,029 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2024-09-18 23:21:35,029 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1,2,3)
(1,2,5)
(2,1,4)
(2,3,4)
(,,)
grunt>
```

Foreach:

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
(2,3,4)
(,,)
grunt> fodata= foreach data generate A1, A2;
grunt> dump fodata;
2024-09-18 23:22:38,771 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNKNOWN
2024-09-18 23:22:38,772 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-18 23:22:38,774 [main] INFO  org.apache.pig.newplan.logical.rules.Column
PruneVisitor - Columns pruned for data: $2
2024-09-18 23:22:38,775 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2024-09-18 23:22:38,776 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-18 23:22:38,776 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-18 23:22:38,776 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
```

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
job_local1526172101_0009


2024-09-18 23:22:51,294 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2024-09-18 23:22:51,295 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-18 23:22:51,295 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-09-18 23:22:51,295 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-18 23:22:51,296 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2024-09-18 23:22:51,306 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2024-09-18 23:22:51,306 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1,2)
(2,1)
(2,3)
(1,2)
(,)
grunt>
```

Load:

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
grunt> stud = load '/home/cloudera/Desktop/student.txt' using PigStorage(',') as
(name:chararray,rollno:int,field:chararray);
grunt> dump stud;
2024-09-18 23:27:38,211 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNKNOWN
2024-09-18 23:27:38,211 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-18 23:27:38,213 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2024-09-18 23:27:38,213 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-18 23:27:38,213 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-18 23:27:38,214 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
```

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
ne.mapReduceLayer.MapReduceLauncher - Success!
2024-09-18 23:27:50,743 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-18 23:27:50,743 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-09-18 23:27:50,745 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-18 23:27:50,745 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2024-09-18 23:27:50,753 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2024-09-18 23:27:50,753 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Tejashree 74 SYMCA,,)
(Yash      82 SYMCA,,)
(Prashant  109 SYMCA,,)
(Utkarsh   78 SYMCA,,)
grunt>
```

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
grunt> grpstud= Group stud by name;
grunt> dump grpstud;
2024-09-18 23:30:22,405 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: GROUP_BY
2024-09-18 23:30:22,405 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-18 23:30:22,407 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2024-09-18 23:30:22,407 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-18 23:30:22,407 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-18 23:30:22,408 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
```

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
ne.mapReduceLayer.MapReduceLauncher - Success!
2024-09-18 23:30:34,944 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-18 23:30:34,944 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-09-18 23:30:34,945 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-18 23:30:34,945 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2024-09-18 23:30:34,954 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2024-09-18 23:30:34,954 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Tejashree 74 SYMCA,{(Tejashree 74 SYMCA,,)})
(Utkarsh   78 SYMCA,{(Utkarsh   78 SYMCA,,)})
(Yash      82 SYMCA,{(Yash      82 SYMCA,,)})
(Prashant  109 SYMCA,{(Prashant  109 SYMCA,,)})
grunt>
```

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help

2024-09-18 23:33:57,634 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2024-09-18 23:33:57,634 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-18 23:33:57,634 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-09-18 23:33:57,635 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-18 23:33:57,635 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2024-09-18 23:33:57,642 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2024-09-18 23:33:57,642 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Tejashree 74 SYMCA,,)
(Yash      82 SYMCA,,)
grunt>
```

Practical 6

Aim: Basic Spark Commands

Move the data into hadoop file system

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 4 items
-rw-r--r--   1 cloudera cloudera          0 2024-07-19 00:46 Demo.txt
drwxr-xr-x   - cloudera cloudera          0 2024-07-31 23:32 inputfolderbr
drwxr-xr-x   - cloudera cloudera          0 2024-07-19 00:42 pract
-rw-r--r--   1 cloudera cloudera          0 2024-09-25 22:51 web-Google.txt
```

```
[cloudera@quickstart ~]$ hdfs dfs -put Downloads/web-Google.txt
put: `web-Google.txt': File exists
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 4 items
-rw-r--r--   1 cloudera cloudera          0 2024-07-19 00:46 Demo.txt
drwxr-xr-x   - cloudera cloudera          0 2024-07-31 23:32 inputfolderbr
drwxr-xr-x   - cloudera cloudera          0 2024-07-19 00:42 pract
-rw-r--r--   1 cloudera cloudera          0 2024-09-25 22:51 web-Google.txt
```

Start pyspark in terminal

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
-rw-r--r--   1 cloudera cloudera          0 2024-09-25 22:51 web-Google.txt
[cloudera@quickstart ~]$ pyspark
Python 2.6.6 (r266:84292, Jul 23 2015, 15:22:56)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-11)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.
jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.j
ar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/parquet/lib/slf4j-log4j12-1.7.5.ja
r!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/avro/avro-tools-1.7.6-cdh5.12.0.ja
r!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanati
on.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/09/25 23:01:50 WARN util.NativeCodeLoader: Unable to load native-hadoop li
brary for your platform... using builtin-java classes where applicable
```

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
r!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanati
on.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/09/25 23:01:50 WARN util.NativeCodeLoader: Unable to load native-hadoop li
brary for your platform... using builtin-java classes where applicable
24/09/25 23:01:51 WARN util.Utils: Your hostname, quickstart.cloudera resolve
s to a loopback address: 127.0.0.1; using 10.0.2.15 instead (on interface eth
0)
24/09/25 23:01:51 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to
another address
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.6.0
      /_/

Using Python version 2.6.6 (r266:84292, Jul 23 2015 15:22:56)
SparkContext available as sc, HiveContext available as sqlContext.
>>> def computeContribs(neighbors, rank):
```

Writing compute contrib function

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
Using Python version 2.6.6 (r266:84292, Jul 23 2015 15:22:56)
SparkContext available as sc, HiveContext available as sqlContext.
>>> def computeContribs(neighbors, rank):
...     for neighbor in neighbors: yield(neighbor, rank/len(neighbors))
...
```

Create a RDD named links with following command

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
KeyboardInterrupt
>>> links = sc.textFile('web-Google.txt')\
...        .map(lambda line: line.split())\
...        .map(lambda pages: (pages[0], pages[1]))\
...        .distinct()\
...        .groupByKey()\
...        .persist()
```

Create a ranks rdd storing the ranks data

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
...        .persist()
>>> ranks=links.map(lambda (page,neighbors): (page,1.0))
```

Create a loop in order to calculate contribs and ranks

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
SyntaxError: invalid syntax
>>> for x in xrange(10):
...     contribs=links\
...     .join(ranks)\
...     .flatMap(lambda(page,(beighbors, rank)):computeContribs(neighbors, ran
k))
...     ranks=contribs\
...     .reduceByKey(lambda v1,v2: v1+v2)\
...     .map(lambda (page,contrib): (page,contrib* 0.85 + 0.15))
```

Code to collect all ranks

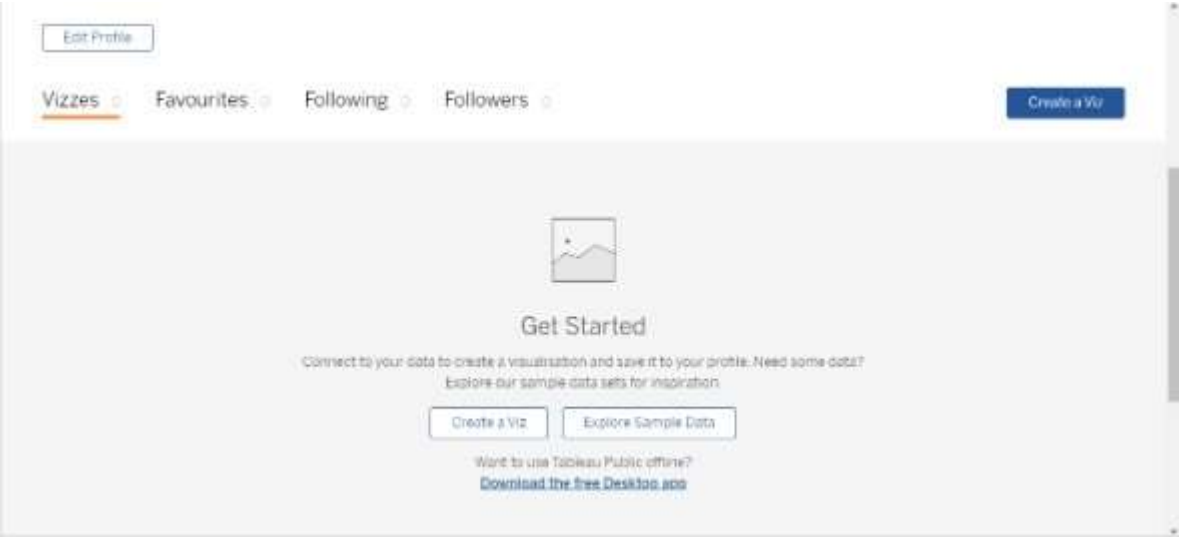
```
...
>>> for rank in ranks.collect(): print rank
...
[Stage 13:>                                                         (0 + 0) /
[Stage 14:==================================================>       (7 + 1) /
[Stage 16:=========================================>                (6 + 1) /
[Stage 18:=========================================>               (7 + 1) /
[Stage 18:=====================================================>    (9 + 1) /
[Stage 20:==========================>                               (5 + 1) /
[Stage 20:=========================================>                (8 + 1) /
[Stage 20:=====================================================>    (10 + 1) /

>>>
```

SYMCA-B
Roll No-90

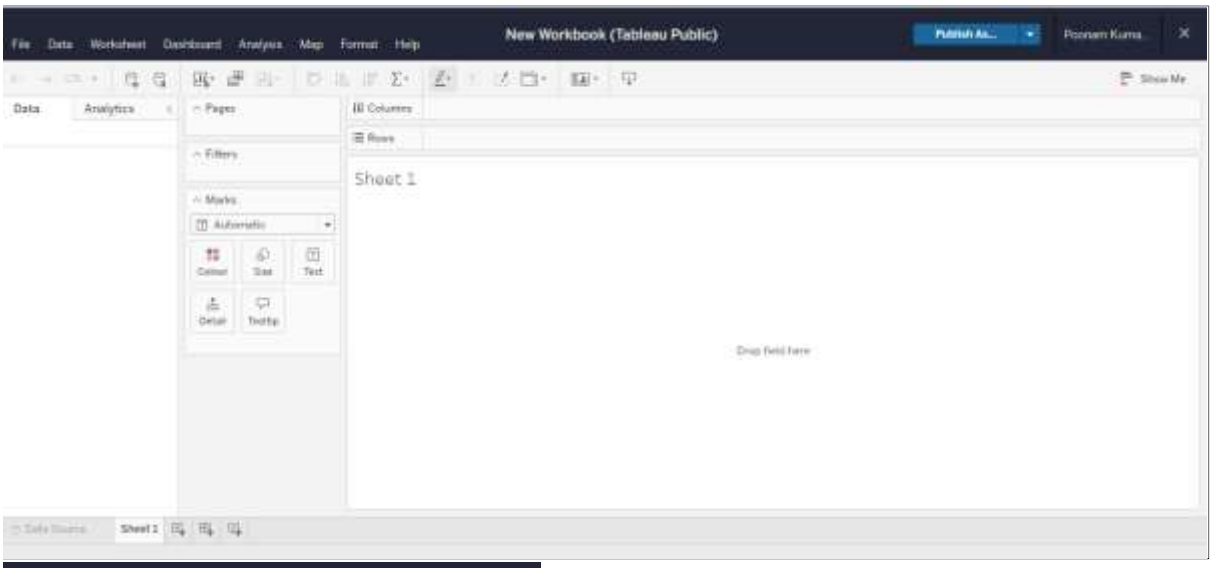
Practical 7

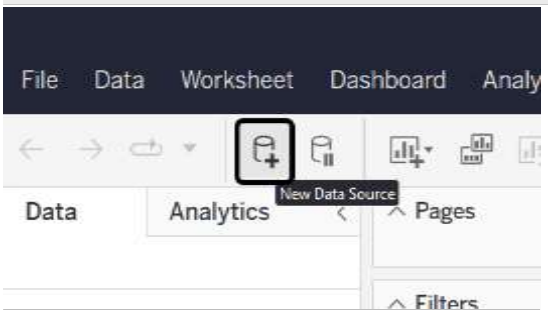
Aim: To Demonstrate Visualization using Tableau.

Create an Account on Tableau Public
Create viz



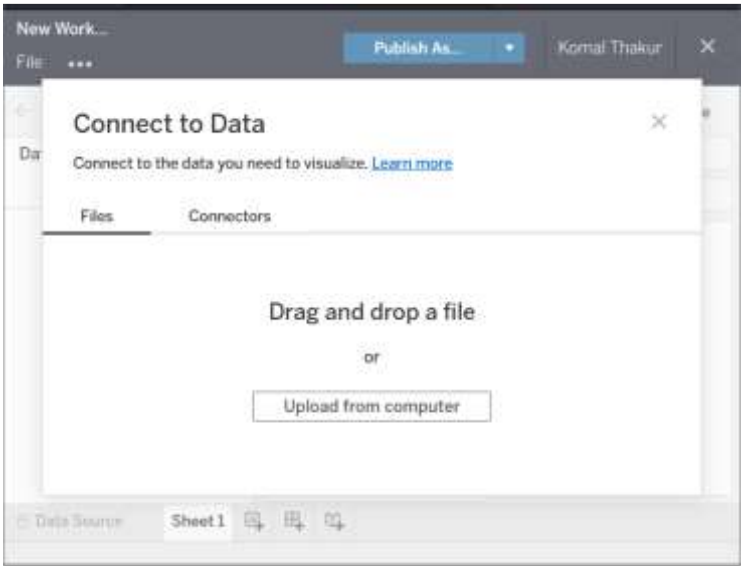
Importing Data



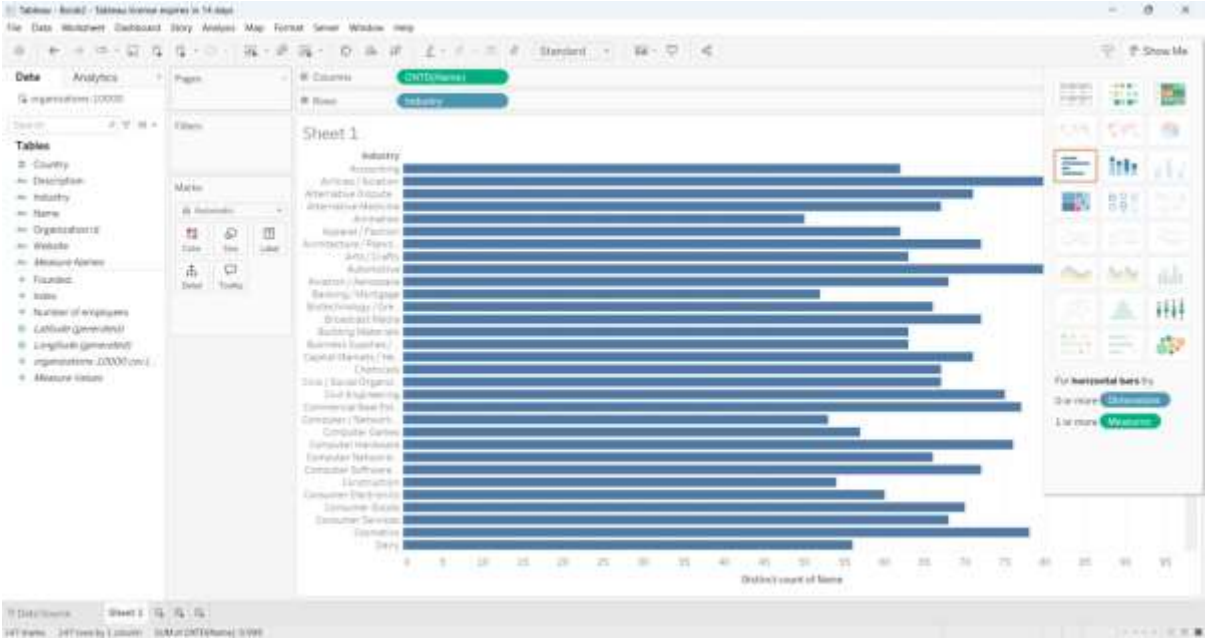


Click on the icon shown below.

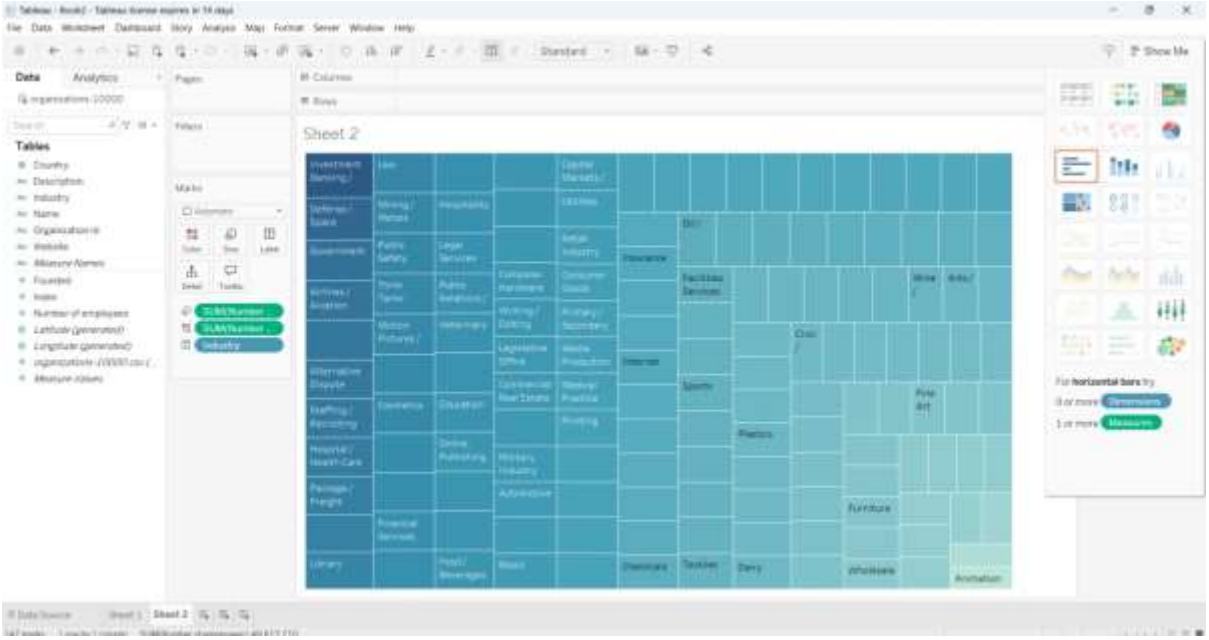
Now drag and drop file you want to use.



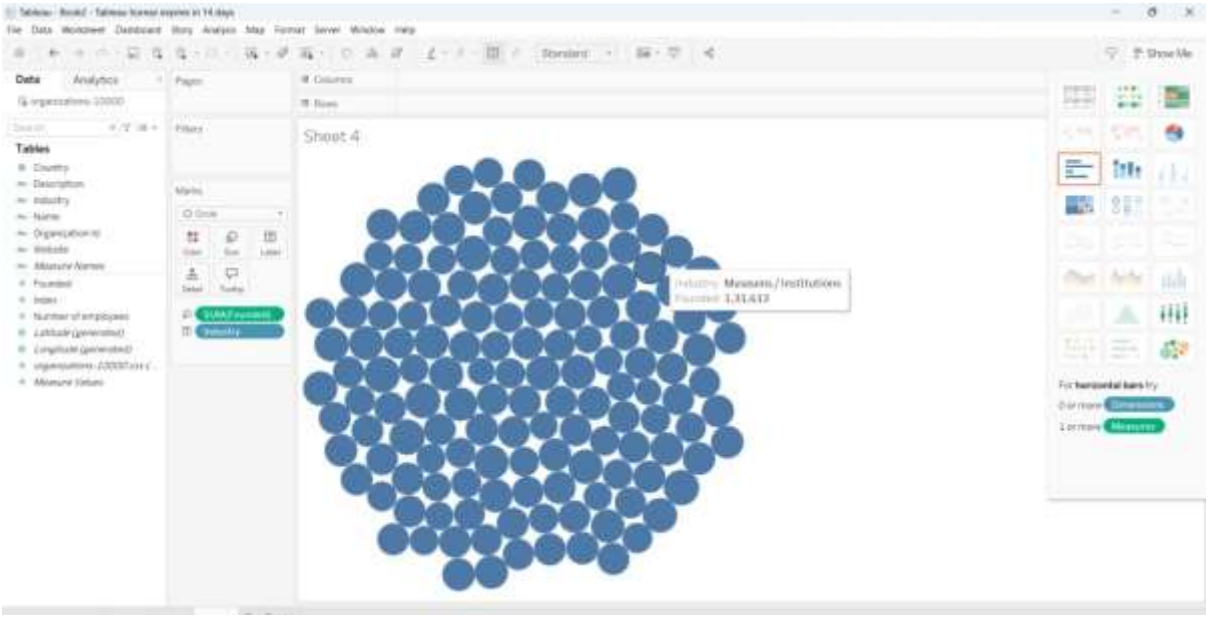
Analysing using charts.

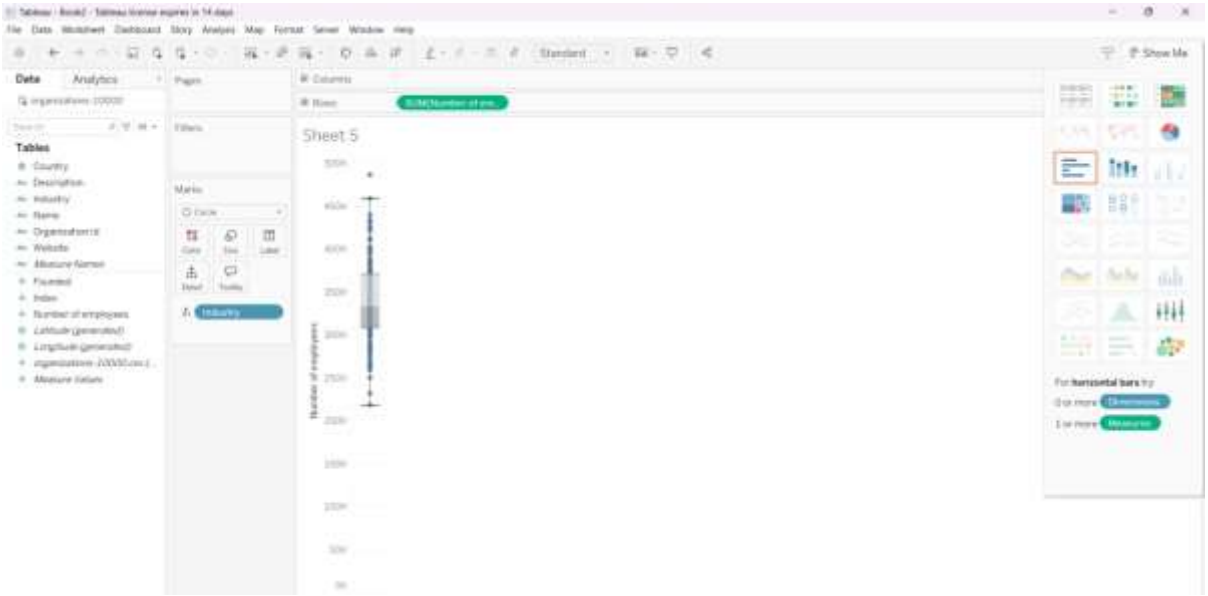


Treemaps



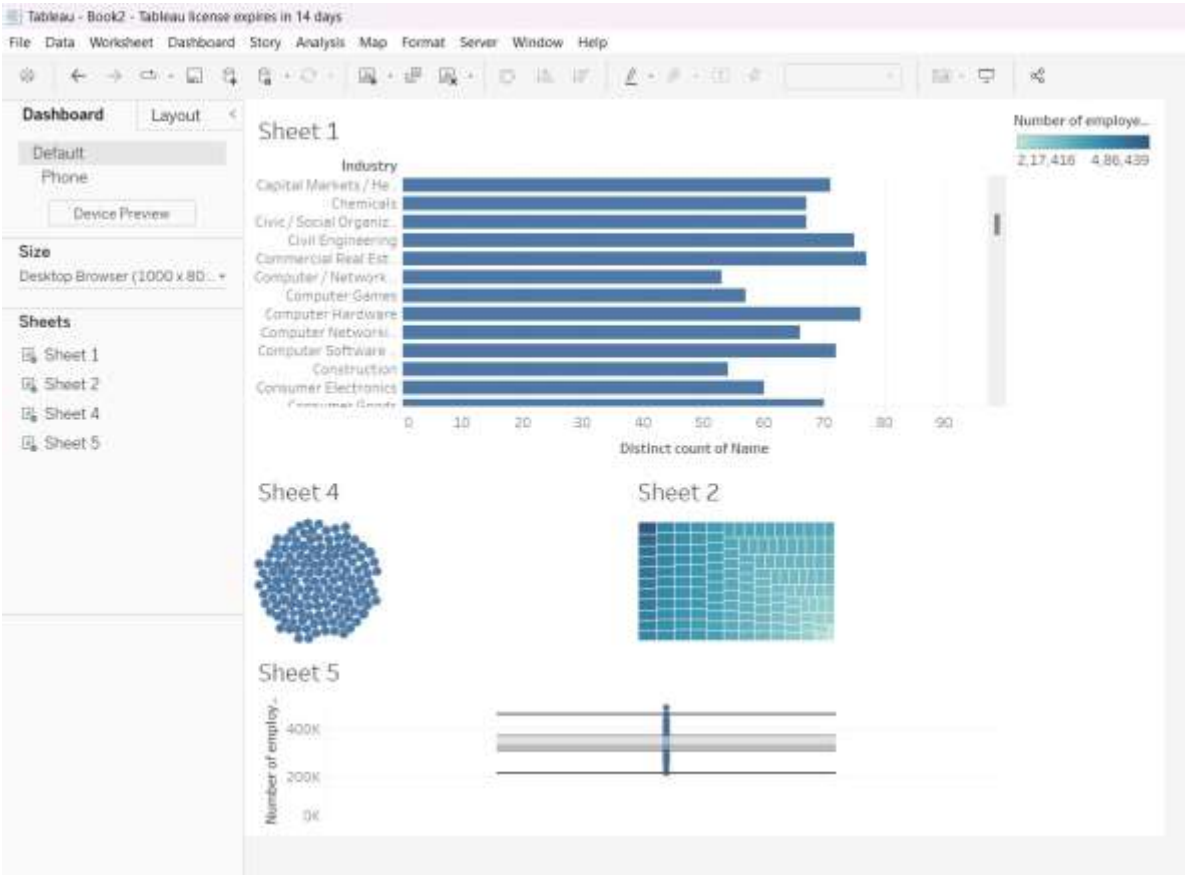
packed bubles





Creating Dashboards

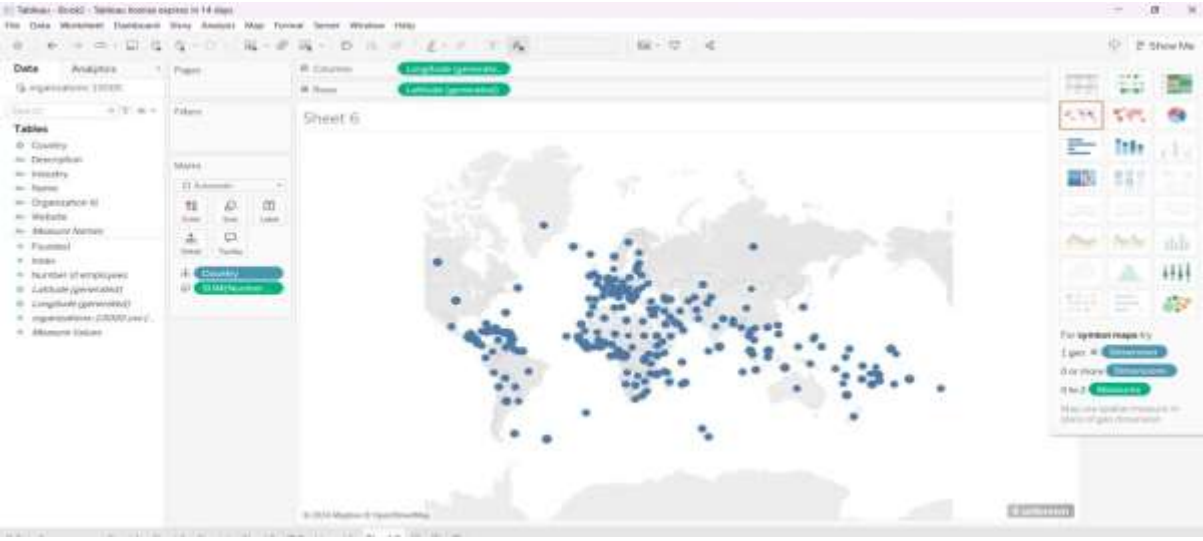
First drag and drop your sheet from Sheets at left side to Main working space
Add other Sheets and arrange them in similar way then your dashboard is ready

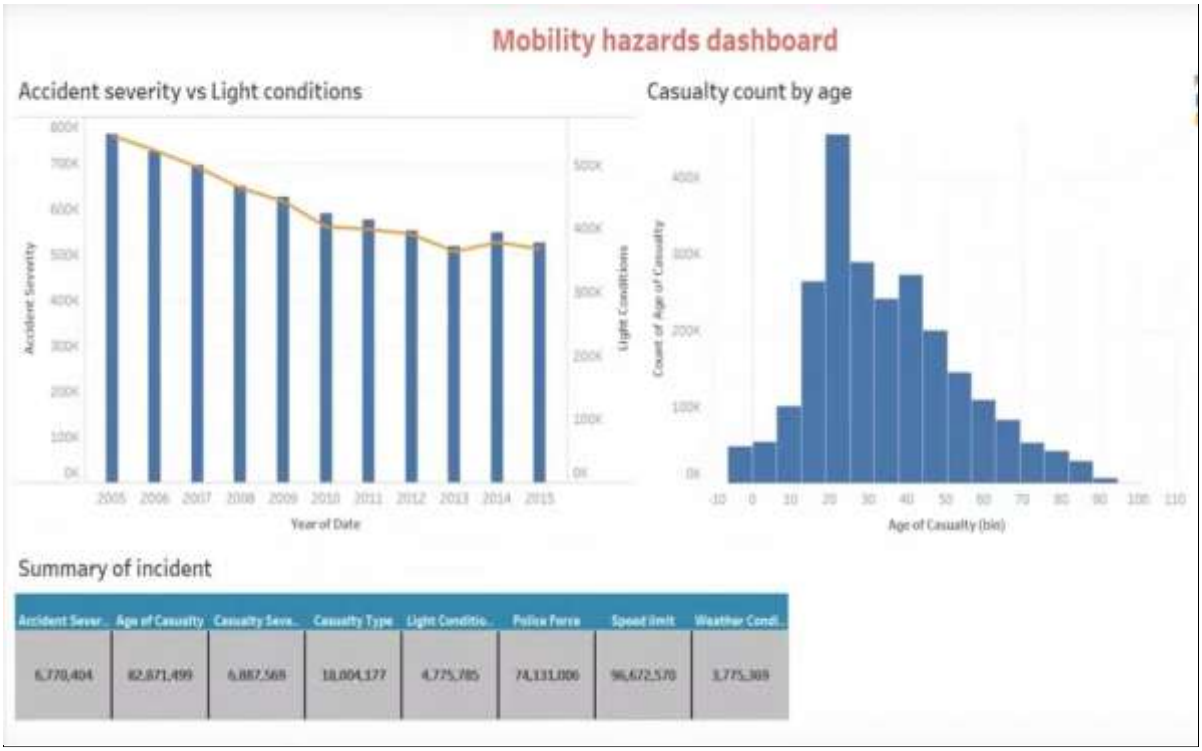


Working with maps

1. Import data and add data as shown below.

2. In this dataset i am using "Country" attribute to be display displayed on given Map.





Telling Stories with Tableau


Mobility hazards data
ampaign
Year of Date


Mobility hazards data collection
The number of casualties specified according to age
Count of Age of Casualty
Age of Casualty (bin)