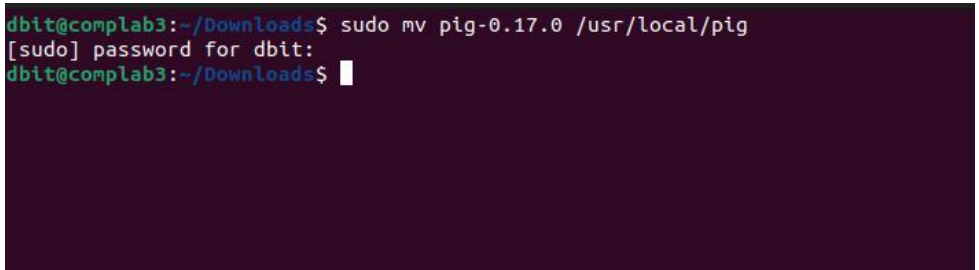
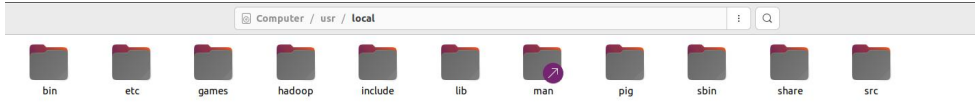
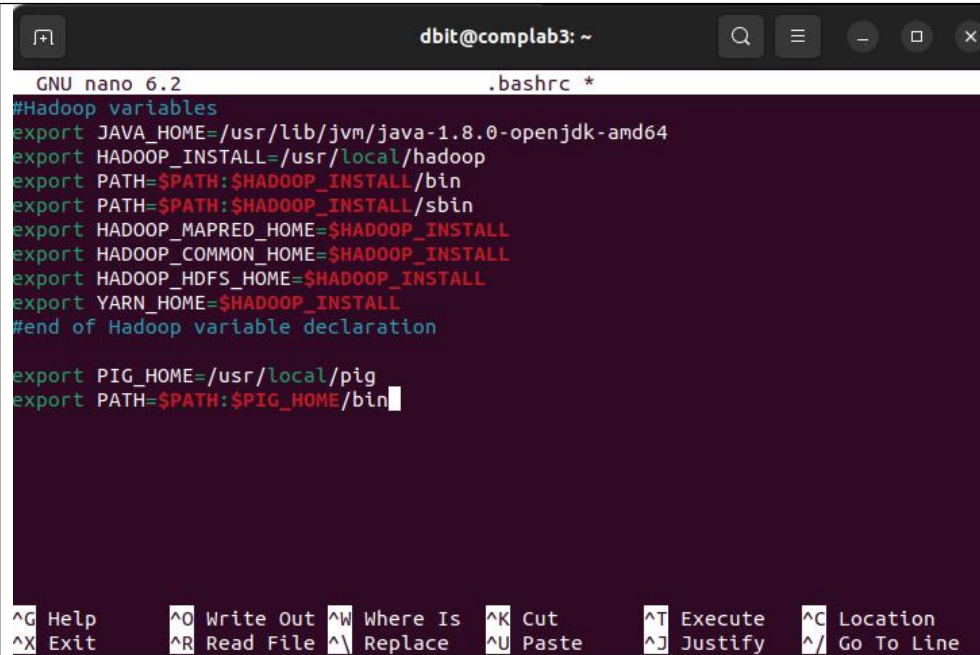


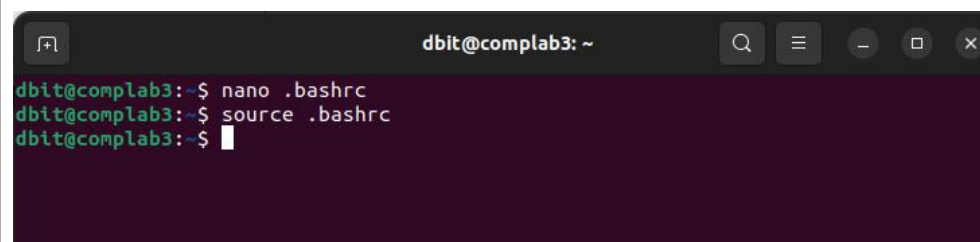
Experiment No: 6**Name:** Alston Fernandes**Roll No:** 19**Batch:** A

Topic:	Install, configure and execute Apache PIG Latin commands
Mapping With COs:	CSL702.3
Objective:	To write queries in PIG Latin language to filter data.
Outcomes:	Students will be able to install and manage big data using Apache PIG.
Instructions:	This experiment is a compulsory experiment. All the students are required to perform this experiment individually.
Deliverables:	<div>Installing PIG  </div> Updating bashrc file



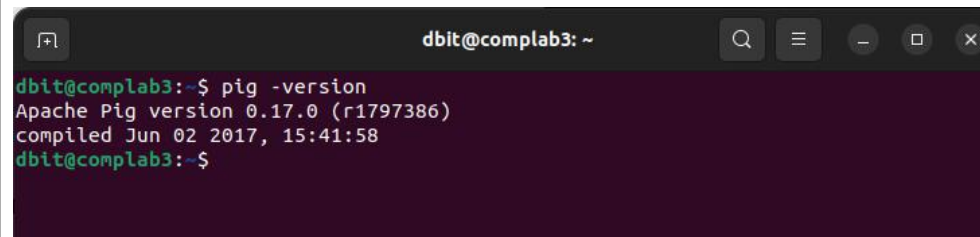
```
dbit@complab3: ~  
GNU nano 6.2 .bashrc *  
#Hadoop variables  
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64  
export HADOOP_INSTALL=/usr/local/hadoop  
export PATH=$PATH:$HADOOP_INSTALL/bin  
export PATH=$PATH:$HADOOP_INSTALL/sbin  
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL  
export HADOOP_COMMON_HOME=$HADOOP_INSTALL  
export HADOOP_HDFS_HOME=$HADOOP_INSTALL  
export YARN_HOME=$HADOOP_INSTALL  
#end of Hadoop variable declaration  
  
export PIG_HOME=/usr/local/pig  
export PATH=$PATH:$PIG_HOME/bin
```

Help Write Out Where Is Cut Execute Location
Exit Read File Replace Paste Justify Go To Line



```
dbit@complab3: ~  
dbit@complab3:~$ nano .bashrc  
dbit@complab3:~$ source .bashrc  
dbit@complab3:~$
```

Checking PIG is installed correctly



```
dbit@complab3: ~  
dbit@complab3:~$ pig -version  
Apache Pig version 0.17.0 (r1797386)  
compiled Jun 02 2017, 15:41:58  
dbit@complab3:~$
```

```
dbit@complab3: ~  
dbit@complab3:~$ pig  
23/09/23 18:33:49 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL  
23/09/23 18:33:49 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE  
23/09/23 18:33:49 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType  
2023-09-23 18:33:49,635 [main] INFO org.apache.pig.Main - Apache Pig version 0.  
17.0 (r1797386) compiled Jun 02 2017, 15:41:58  
2023-09-23 18:33:49,635 [main] INFO org.apache.pig.Main - Logging error message  
s to: /home/dbit/pig_1695474229633.log  
2023-09-23 18:33:49,767 [main] INFO org.apache.pig.impl.util.Utils - Default bo  
otup file /home/dbit/.pigbootup not found  
2023-09-23 18:33:51,037 [main] INFO org.apache.hadoop.conf.Configuration.deprec  
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr  
ess  
2023-09-23 18:33:51,037 [main] INFO org.apache.pig.backend.hadoop.executionengi  
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000  
2023-09-23 18:33:52,622 [main] INFO org.apache.pig.PigServer - Pig Script ID fo  
r the session: PIG-default-3d85f299-a1da-45bb-9f6a-5fa7bc5fb011  
2023-09-23 18:33:52,622 [main] WARN org.apache.pig.PigServer - ATS is disabled  
since yarn.timeline-service.enabled set to false  
grunt>
```

Dataset 1

```
dbit@complab3:~$ cat pig.txt  
1991, 33, Mumbai  
1991, 38, Pune  
1991, 38, Pune  
1991, 36, Delhi  
1990, 55, Pune  
1990, 55, Delhi  
1996, 65, Mumbai  
1996, 36, Pune  
2020, 66, Mumbai  
2020, 52, Pune  
2025, 55, Mumbai  
2025, 62, Delhi  
2025, 62, Pune  
2025, 62, Pune  
dbit@complab3:~$
```

LOAD keyword with column name and datatype

```
grunt> my_bag = LOAD '/home/dbit/pig.txt' as (year:int, temp:int, city:chararray);  
2023-09-23 18:46:24,241 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.  
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
grunt> describe my_bag;  
my_bag: {year: int,temp: int,city: chararray}  
grunt>
```

Dump Keyword

```
2023-09-24 11:52:22,149 [main] INFO org.apache.hadoop.conf.Configuration - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
grunt> dump my_bag;
```

```
SchemaTupleBackend has already been initialized
2023-09-24 11:53:02,166 [main] INFO org.apache.hadoop.mapreduce.lib.in
putFormat - Total input paths to process : 1
2023-09-24 11:53:02,166 [main] INFO org.apache.pig.backend.hadoop.exec
ne.util.MapRedUtil - Total input paths to process : 1
(1991, 33, Mumbai)
(1991, 38, Pune)
(1991, 38, Pune)
(1991, 36, Delhi)
(1990, 55, Pune)
(1990, 55, Delhi)
(1996, 65, Mumbai)
(1996, 36, Pune)
(2020, 66, Mumbai)
(2020, 52, Pune)
(2025, 55, Mumbai)
(2025, 62, Delhi)
(2025, 62, Pune)
(2025, 62, Pune)
grunt>
```

LOAD keyword with column name

```
grunt> my_bag1 = LOAD '/home/dbit/pig.txt' as (year, temp, city);
2023-09-23 19:23:54,441 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> describe my_bag1;
my_bag1: {year: bytearray,temp: bytearray,city: bytearray}
grunt>
```

LOAD keyword without parameter

```
grunt> my_bag2 = LOAD '/home/dbit/pig.txt';
2023-09-23 19:24:36,894 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> describe my_bag2;
Schema for my_bag2 unknown.
grunt>
```

Dataset 2

```
dbit@complab3:~$ cat data.txt
Sam, Mumbai, 66
Jim, Pune, 77
Tom, Pune, 55
Herry, Mumbai, 65
Sony, Delhi, 51
Sia, Pune, 65
Sia, Pune, 65
Sia, Pune, 65
Tina, Mumbai, 53
Jia, Delhi, 76
Sara, Delhi, 65
dbit@complab3:~$
```

Q1. Remove duplicate tuples and put in a bag say "result1" and show its content.

```
grunt> Q1 = LOAD '/home/dbit/data.txt';
2023-09-23 19:28:03,678 [main] INFO org.apache.hadoop.conf.Configuration.deprecation -
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> result1 = DISTINCT Q1;
grunt> dump result1;
2023-09-23 19:29:04,365 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig fe
d in the script: DISTINCT
2023-09-23 19:29:04,389 [main] INFO org.apache.hadoop.conf.Configuration.deprecation -
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-23 19:29:04,389 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTup
has already been initialized
2023-09-23 19:29:04,390 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPl
```



```
2023-09-24 11:55:04,645 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-09-24 11:55:04,645 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Jim, Pune, 77)
(Sia, Pune, 65)
(Tom, Pune, 55)
(Jia, Delhi, 76)
(Sam, Mumbai, 66)
(Sara, Delhi, 65)
(Sony, Delhi, 51)
(Tina, Mumbai, 53)
(Herry, Mumbai, 65)
grunt> █
```

```
grunt> store result1 into '/home/dbit/result1' using PigStorage(';');
2023-09-24 11:56:39,360 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-24 11:56:39,364 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2023-09-24 11:56:39,400 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: DISTINCT
2023-09-24 11:56:39,416 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-24 11:56:39,416 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-09-24 11:56:39,416 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeF
```

```
Job DAG:
job_local1902519006_0003
```

```
2023-09-23 19:31:06,568 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-09-23 19:31:06,568 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-09-23 19:31:06,569 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-09-23 19:31:06,570 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> █
```

Home / result1

part-r-00000 _SUCCESS

part-r-00000
~/result1

1 Jim, Pune, 77
2 Sia, Pune, 65
3 Tom, Pune, 55
4 Jia, Delhi, 76
5 Sam, Mumbai, 66
6 Sara, Delhi, 65
7 Sony, Delhi, 51
8 Tina, Mumbai, 53
9 Herry, Mumbai, 65

Q2. Sort the data in ascending and descending both and put the sorted data in the bag “result2” and “result3” respectively.

```
grunt> Q2 = LOAD '/home/dbit/data.txt' as (Name, Location, marks);
2023-09-23 19:37:02,550 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.byte
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> describe Q2;
Q2: {Name: bytearray,Location: bytearray,marks: bytearray}
grunt> result2 = ORDER Q2 by Name;
grunt> result3 = ORDER Q2 by Name DESC;
grunt> store result2 into '/home/dbit/result2' using PigStorage(';');
2023-09-23 19:38:12,021 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.byte
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-23 19:38:12,062 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features
d in the script: ORDER_BY
2023-09-23 19:38:12,077 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.byte
```

Home / result2

part-r-00000 _SUCCESS

Open [icon] part-r-00000
~/result2

```
1 Herry, Mumbai, 65
2 Jia, Delhi, 76
3 Jim, Pune, 77
4 Sam, Mumbai, 66
5 Sara, Delhi, 65
6 Sia, Pune, 65
7 Sia, Pune, 65
8 Sia, Pune, 65
9 Sony, Delhi, 51
10 Tina, Mumbai, 53
11 Tom, Pune, 55
```

```
grunt> store result3 into '/home/dbit/result3' using PigStorage(';');
2023-09-23 19:39:21,781 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.byte
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-23 19:39:21,815 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.byte
```

Home / result3

part-r-00000 _SUCCESS

Open [icon] part-r-00000
~/result3

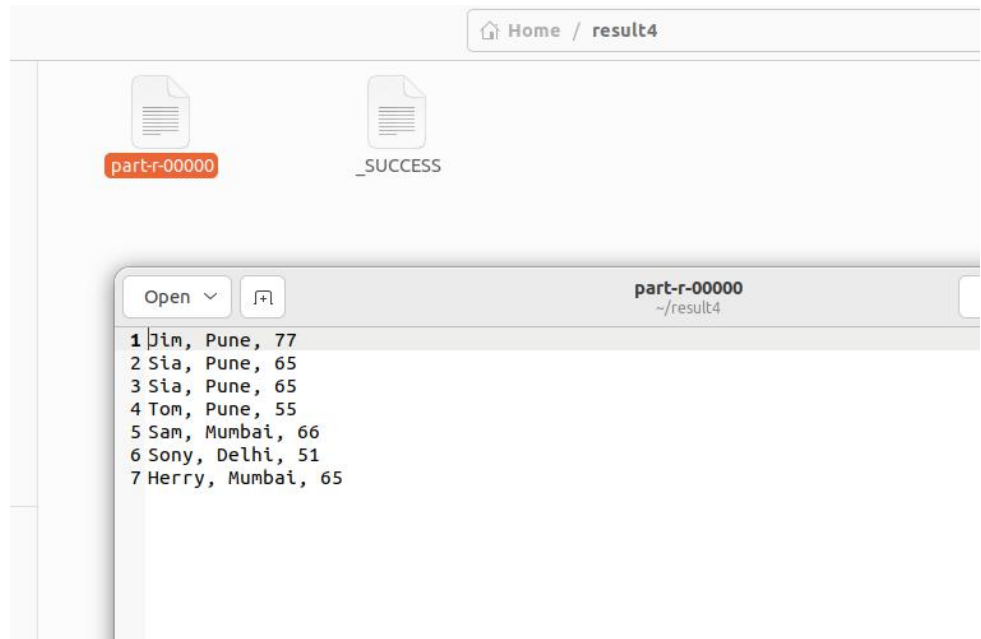
```
1 Tom, Pune, 55
2 Tina, Mumbai, 53
3 Sony, Delhi, 51
4 Sia, Pune, 65
5 Sia, Pune, 65
6 Sia, Pune, 65
7 Sara, Delhi, 65
8 Sam, Mumbai, 66
9 Jim, Pune, 77
10 Jia, Delhi, 76
11 Herry, Mumbai, 65
```

Q3. Put the first seven tuples in the bag “result4” and show its content.

```
grunt> Q3 = LOAD '/home/dbit/data.txt';
2023-09-23 21:02:03,992 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-23 21:02:04,249 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> result4 = LIMIT Q3 7;
grunt> dump result4;
```

```
hemaTupleBackend has already been initialized
2023-09-24 12:07:21,069 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2023-09-24 12:07:21,069 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Sam, Mumbai, 66)
(Jim, Pune, 77)
(Tom, Pune, 55)
(Herry, Mumbai, 65)
(Sony, Delhi, 51)
(Sia, Pune, 65)
(Sia, Pune, 65)
grunt>
```

```
grunt> store result4 into '/home/dbit/result4' using PigStorage(';');
2023-09-23 21:03:32,695 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-23 21:03:33,075 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features use
d in the script: LIMIT
```



Q4. Show all the details of the students who scored more than 60 marks in “result5”.

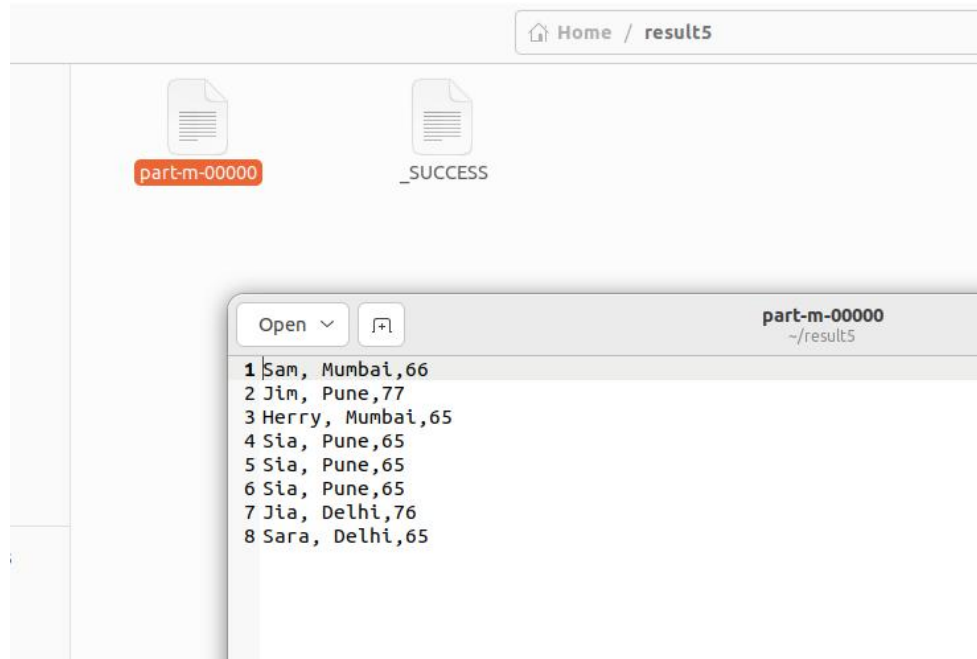
```
grunt> Q4 = LOAD '/home/dbit/data.txt' using PigStorage(';') as (Name:chararray,
Location:chararray, Marks:int);
2023-09-24 12:21:38,527 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> dump Q4;
2023-09-24 12:21:40,086 [main] INFO org.apache.pig.tools.pigstats.ScriptState -
Pig features used in the script: UNKNOWN
2023-09-24 12:21:40,097 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
```



```
2023-09-24 12:21:40,412 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Sam, Mumbai,66)
(Jim, Pune,77)
(Tom, Pune,55)
(Herry, Mumbai,65)
(Sony, Delhi,51)
(Sia, Pune,65)
(Sia, Pune,65)
(Sia, Pune,65)
(Tina, Mumbai,53)
(Jia, Delhi,76)
(Sara, Delhi,65)
grunt> result5 = FILTER Q4 by Marks>60;
grunt> dump result5;
```

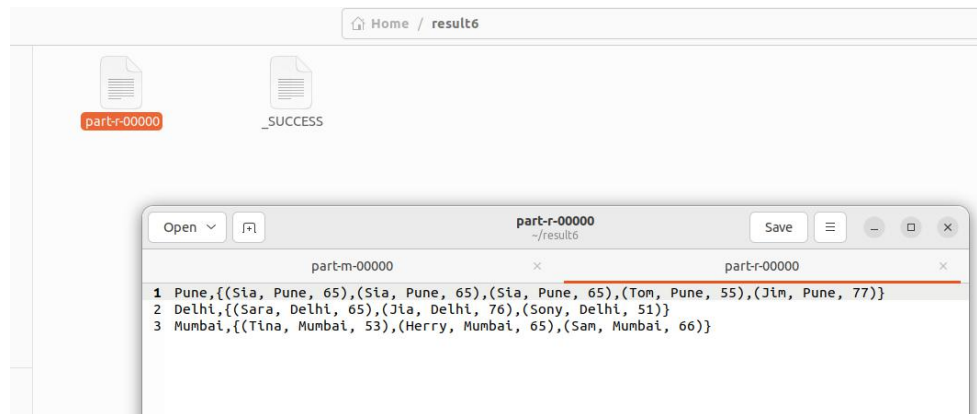
```
InputFormat - Total input paths to process : 1
2023-09-24 12:23:31,876 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Sam, Mumbai,66)
(Jim, Pune,77)
(Herry, Mumbai,65)
(Sia, Pune,65)
(Sia, Pune,65)
(Sia, Pune,65)
(Jia, Delhi,76)
(Sara, Delhi,65)
grunt>
```

```
grunt> store result5 into '/home/dbit/result5' using PigStorage(',');
2023-09-24 12:24:51,066 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-24 12:24:51,081 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
```



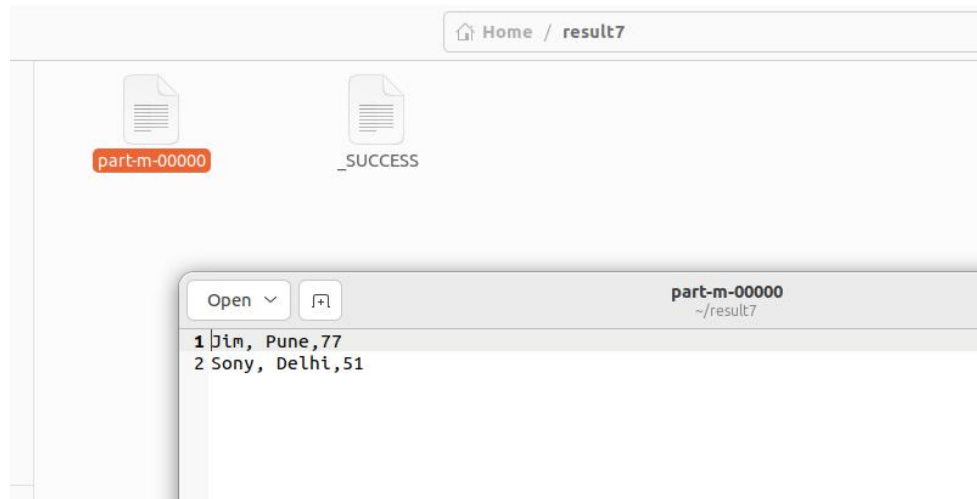
Q5. Group the students according to their city and store it in the bag saying “result6”.


```
grunt> Q5 = LOAD '/home/dbit/data.txt' using PigStorage(',') as (Name, Location:
chararray, Marks);
2023-09-24 12:26:58,773 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-24 12:26:58,788 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> result6 = GROUP Q5 by Location;
grunt> store result6 into '/home/dbit/result6' using PigStorage(',');
```



Q6. Show which student scored maximum and which student scored minimum marks in “result7”

```
grunt> Q6 = LOAD '/home/dbit/data.txt' USING PigStorage(',')
>> AS (Name:chararray, Location:chararray, Marks:int);
2023-09-24 13:38:01,415 [main] INFO org.apache.hadoop.conf.Configuration.deprecation -
io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> min_score = FOREACH (GROUP Q6 ALL) GENERATE MIN(Q6.Marks) AS (Marks:int);
grunt> max_score = FOREACH (GROUP Q6 ALL) GENERATE MAX(Q6.Marks) AS (Marks:int);
grunt> result7 = FILTER Q6 by (Marks==min_score.Marks OR Marks==max_score.Marks);
grunt> dump result7;
2023-09-24 13:38:41,370 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig fe
atures used in the script: GROUP_BY,FILTER
2023-09-24 13:38:41,381 [main] INFO org.apache.hadoop.conf.Configuration.deprecation -
io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-24 13:38:41,382 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [oig
2023-09-24 13:38:42,118 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTup
leBackend has already been initialized
2023-09-24 13:38:42,124 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFor
mat - Total input paths to process : 1
2023-09-24 13:38:42,124 [main] INFO org.apache.pig.backend.hadoop.executionengine.util
..MapRedUtil - Total input paths to process : 1
(Jim, Pune,77)
(Sony, Delhi,51)
grunt>
grunt>
```



Conclusion:	Apache Pig offers a high-level scripting language that abstracts complex Hadoop MapReduce operations. This simplicity makes it an excellent tool for teaching beginners or students who are new to big data processing and distributed computing. Pig seamlessly integrates with other components of the Hadoop ecosystem, such as HDFS (Hadoop Distributed File System) and Hive. This integration exposes students to a broader ecosystem of tools and technologies used for big data processing and analytics.
References:	https://pig.apache.org/docs/latest/basic.html#schemas https://www.tutorialspoint.com/apache_pig/pig_latin_basics.htm https://www.tutorialspoint.com/apache_pig/apache_pig_min.htm https://data-flair.training/blogs/apache-pig-built-in-functions/

Don Bosco Institute of Technology
Department of Computer Engineering

Academic year – 2022-2023

Big Data Analytics

Assessment Rubric for Experiment No.: 06

Performance Date :
Submission Date :

Title of Experiment :

Year and Semester : IVth Year and VIIth Semester

Batch :

Name of Student :

Roll No. :

Performance	Poor	Satisfactory	Good	Excellent	
	1 point	2 points	3 points	4 points	
Results and Documentations	Poor	Satisfactory	Good	Excellent	
	1 point	2 points	3 points	4 points	
Viva	Poor	Satisfactory	Good	Excellent	
	1 point	2 points	3 points	4 points	
Timely Submission	Submission beyond 7 days of the deadline	Late submission till 7 days	Submission on time		
	1 points	2 points	3 points		

Shaikh

Faculty Incharge : Ms. Sana