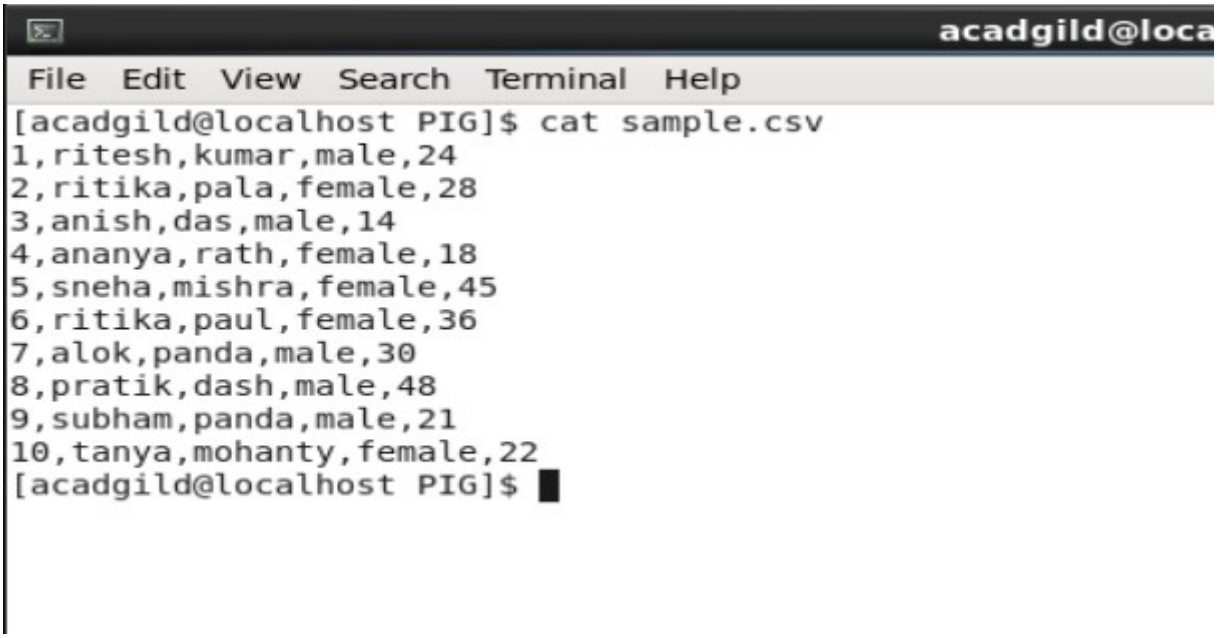
Bigdata Assignment2.2

Create a sample dataset and implement the below Pig commands on the same dataset.
1) Concat
2) Tokenize
3) Sum
4) Min
5) Max
6) Limit
7) Store
8) Distinct
9) Flatten
10) IsEmpty

<u>Created the dataset:-</u>

cat sample.csv;



<u>Entered into grunt shell:-</u>

Command - **pig -x local;**

```
acadgild@localhost:~/INPUT/PIG
File  Edit  View  Search  Terminal  Help
[acadgild@localhost PIG]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/05/28 20:13:43 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/28 20:13:43 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-05-28 20:13:43,911 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
9
2018-05-28 20:13:43,911 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/INPUT/PIG/pig_1527518623
909.log
2018-05-28 20:13:44,011 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-05-28 20:13:44,246 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2018-05-28 20:13:44,246 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-05-28 20:13:44,248 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: file:///
2018-05-28 20:13:44,518 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-28 20:13:44,571 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-f74ec77d-bcca-4fa3
-ba9a-03eb4a8a6ec7
2018-05-28 20:13:44,571 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to fa
lse
grunt>
```

Loaded the data into alias:-

Command - **data = LOAD 'sample.csv' using PigStorage(',') as (id:int,first:charaaray,last:chararray,gender:chararray,age:int);**

```
acadgild@localhost:~/INPUT/PIG
File  Edit  View  Search  Terminal  Help

grunt> data = LOAD 'sample.csv' using PigStorage(',') as (id:int,first:chararray,last:chararray,gender:chararray,age:int);
2018-05-28 20:17:30,714 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-28 20:17:30,714 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> dump data;
```

Displayed the data:-

```
acadgild@localhost:~/INPUT/PIG
File  Edit  View  Search  Terminal  Help
grunt> dump data;
2018-05-28 20:25:06,998 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
```

```
(1,ritesh,kumar,male,24)
(2,ritika,pala,female,28)
(3,anish,das,male,14)
(4,ananya,rath,female,18)
(5,sneha,mishra,female,45)
(6,ritika,paul,female,36)
(7,alok,panda,male,30)
(8,pratik,dash,male,48)
(9,subham,panda,male,21)
(10,tanya,mohanty,female,22)
grunt>
```

1. Concat -

first_last = foreach data generate CONCAT(first,' ',last) as name;
dump first_last;

```
acadgild@localhost:~/INPUT/PIG
File  Edit  View  Search  Terminal  Help

grunt> first_last = foreach data generate CONCAT(first,' ',last) as name;
grunt> dump first_last;
2018-05-28 20:29:49,041 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features
```

```
(ritesh kumar)
(ritika pala)
(anish das)
(ananya rath)
(sneha mishra)
(ritika paul)
(alok panda)
(pratik dash)
(subham panda)
(tanya mohanty)
grunt>
```

O/p – Here using CONCAT we concated the two columns first and last. Hence two columns merged data was the output.

2. Tokenize -

tokenized = foreach first_last generate TOKENIZE(name);

```
acadgild@localhost:~/INPUT/
File  Edit  View  Search  Terminal  Help
grunt> tokenized = foreach first_last generate TOKENIZE(name);
grunt> dump tokenized;

({(ritesh),(kumar)})
({(ritika),(pala)})
({(anish),(das)})
({(ananya),(rath)})
({(sneha),(mishra)})
({(ritika),(paul)})
({(alok),(panda)})
({(pratik),(dash)})
({(subham),(panda)})
({(tanya),(mohanty)})
grunt>
```

O/p- Here using TOKENIZE when a chararray is given as an argument, this method will split the chararray and return a bag with a tuple for each chararray that results from the split.
We slitted the first_last charaaray.

3.SUM:-

data_grp = GROUP data all;
result = foreach data_grp generate SUM(data.age);
dump result;

```
acadgild@localhos
File  Edit  View  Search  Terminal  Help

grunt> data_grp = GROUP data all;
grunt> result =  foreach data_grp generate SUM(data.age);
grunt> dump result;

cess : 1
(286)
grunt>
```

O/P- SUM function is used get the total of the numeric values of a column in a single-column bag. Here we first made the data into a single column bag then we found the sum of the age column.

4. MIN:-

result = foreach data_grp generate MIN(data.age) , data.first;
dump result;

```
acadgild@localhost:~/INPUT/
File  Edit  View  Search  Terminal  Help

grunt> result =  foreach data_grp generate MIN(data.age),data.first;
grunt> dump result;

cess : 1
(14,{(tanya),(subham),(pratik),(alok),(ritika),(sneha),(ananya),(anish),(ritika),(ritesh)})
grunt>
```

O/P- MIN function is used get the minimum value of a certain column in a single-column bag. Here we first made the data into a single column bag then we found the minimum value of the age column.

5.MAX:-

result = foreach data_grp generate MAX(data.age) , data.first;
dump result;

```
acadgild@localhost:~/INPUT/PIG
File  Edit  View  Search  Terminal  Help
grunt> result =  foreach data_grp generate MAX(data.age),data.first;
grunt> dump result;

(48,{(tanya),(subham),(pratik),(alok),(ritika),(sneha),(ananya),(anish),(ritika),(ritesh)})
grunt>
```

O/P- MAX function is used get the minimum value of a certain column in a single-column bag. Here we first made the data into a single column bag then we found the maximum value of the age column.

6.LIMIT:-

first_3 = LIMIT data 3;
dump first_3;

```
acadgild@localhost:~/INPUT/PIG
File  Edit  View  Search  Terminal  Help

grunt> first_3 = LIMIT data 3;
grunt> dump first_3;
2018-05-29 10:07:41,520 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
2018-05-29 10:07:41,559 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2018-05-29 10:07:41,649 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-29 10:07:41,650 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-05-29 10:07:41,708 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEa
ch, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, Mer
geForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTyp
eCastInserter]}
2018-05-29 10:07:41,835 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699
400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2018-05-29 10:07:42,060 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-29 10:07:42,060 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-05-29 10:07:42,061 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-29 10:07:42,064 [main] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2018-05-29 10:07:42,068 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-29 10:07:42,073 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
2018-05-29 10:07:42,136 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attem
pt__0001_m_000001_1' to file:/tmp/temp1614455752/tmp184462639/_temporary/0/task__0001_m_000001
2018-05-29 10:07:42,165 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-29 10:07:42,189 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-29 10:07:42,189 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(1,ritesh,kumar,male,24)
(2,ritika,pala,female,28)
(3,anish,das,male,14)
grunt>
```

O/p – LIMIT is used to get limited no of tuples from a relation.Here we got 3 no of tuples.

7. STORE:-

STORE first_3 into '/home/acadgild/INPUT/PIG/first_3';

To check the content – **cat part-r-00000;**

```
acadgild@localhost:~/INPUT/PIG
File  Edit  View  Search  Terminal  Help

grunt> STORE first_3 into '/home/acadgild/INPUT/PIG/first_3' using PigStorage(',');
```

```
acadgild@localhost:~/INPUT/PIG/first_3
File  Edit  View  Search  Terminal  Help
[acadgild@localhost first_3]$ cat part-r-00000
1,ritesh,kumar,male,24
2,ritika,pala,female,28
3,anish,das,male,14
[acadgild@localhost first_3]$
```

O/P – STORE is used to store the loaded data into the file system . Here we , loaded the data first_3 into a folder in the file system and then checked it.

8.DISTINCT:-

distinct_data = DISTINCT data;

```
File  Edit  View  Search  Terminal  Help

grunt> distinct_data = DISTINCT data;
```

```
(1,ritesh,kumar,male,24)
(2,ritika,pala,female,28)
(3,anish,das,male,14)
(4,ananya,rath,female,18)
(5,sneha,mishra,female,45)
(6,ritika,paul,female,36)
(7,alok,panda,male,30)
(8,pratik,dash,male,48)
(9,subham,panda,male,21)
(10,tanya,mohanty,female,22)
grunt>
```

O/P- DISTINCT is used to remove redundant tuples from a relation. Here , as there were no redundant tuples , the entire data was the output.

9. <u>FLATTEN:-</u>

flatten_data = foreach data_grp generate FLATTEN(data);
dump flatten-data;

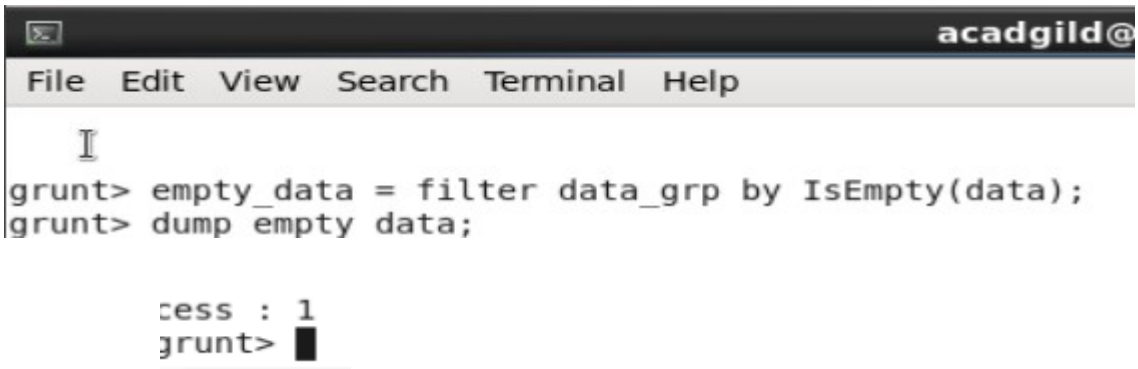
```
acadgild@localhost:~/IN
File  Edit  View  Search  Terminal  Help
grunt> flatten_data = foreach data_grp generate FLATTEN(data);
grunt> dump flatten_data;
```

```
(10,tanya,mohanty,female,22)
(9,subham,panda,male,21)
(8,pratik,dash,male,48)
(7,alok,panda,male,30)
(6,ritika,paul,female,36)
(5,sneha,mishra,female,45)
(4,ananya,rath,female,18)
(3,anish,das,male,14)
(2,ritika,pala,female,28)
(1,ritesh,kumar,male,24)
grunt>
```

O/P – This flattens the nested schema into unnested schema. Here , we converted the data_grp bag into flatten-data.

10. IsEmpty:-

empty_data = filter data_grp by IsEmpty(data);
dump empty_data;



O/P – ISEmpty is used to check if empty map or bags are present. Here , the output is nothing since there are no empty bags.