Academic year 2019-2020

Department of Computer Science and Engineering

KARNATAKA LAW SOCIETY’S

GOGTE INSTITUTE OF TECHNOLOGY

UDYAMBAG, BELAGAVI-590008

Course Project Report

“Execution of Apache Pig MAX Function“

Sem: 7

1. Abhishek Tadkod	2GI17CS002
2. Akshay Raichur	2GI17CS012
3. Chetana Bhat	2GI17CS031
4. Jayanth Apagundi	2GI17CS182

Guide
Prof. Arati Shapurkar
Gogte Institute of Technology
Belagavi.

Contents

Definition:

The Pig Latin **MAX()** function is used to calculate the highest value for a column (numeric values or chararrays) in a single-column bag. While calculating the maximum value, the **Max()** function ignores the NULL values.

Note −

- To get the global maximum value, we need to perform a **Group All** operation, and calculate the maximum value using the MAX() function.
- To get the maximum value of a group, we need to group it using the **Group By** operator and proceed with the maximum function.

Syntax:

```
grunt> Max(expression)
```

Example:

Assume that we have a file named **student_details.txt** in the HDFS directory **/pig_data/** as shown below.

student_details.txt

```
001,Rajiv,Reddy,21,9848022337,Hyderabad,89
002,siddarth,Battacharya,22,9848022338,Kolkata,78
003,Rajesh,Khanna,22,9848022339,Delhi,90
004,Preethi,Agarwal,21,9848022330,Pune,93
005,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar,75
006,Archana,Mishra,23,9848022335,Chennai,87
007,Komal,Nayak,24,9848022334,trivendram,83
008,Bharathi,Nambiayar,24,9848022333,Chennai,72
```

And the file is loaded into Pig with the relation name **student_details** as shown below.

```
grunt> student_details = LOAD 'pig_data/student_details.txt' USING
PigStorage(',')
as (id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,
city:chararray, gp
```

Calculating the Maximum GPA:

We can use the built-in function **MAX()** to calculate the maximum value from a set of given numerical values. Let us group the relation **student_details** using the **Group All** operator, and store the result in the relation named **student_group_all** as shown below.

```
grunt> student_group_all = Group student_details All;
```

This will produce a relation as shown below.

```
grunt> Dump student_group_all;
```

```
(all,{(8,Bharathi,Nambiayar,24,9848022333,Chennai,72),
(7,Komal,Nayak,24,9848022 334,trivendram,83),
(6,Archana,Mishra,23,9848022335,Chennai,87),
(5,Trupthi,Mohan thy,23,9848022336,Bhuwaneshwar,75),
(4,Preethi,Agarwal,21,9848022330,Pune,93),
(3,Rajesh,Khanna,22,9848022339,Delhi,90),
(2,siddarth,Battacharya,22,9848022338,Ko lkata,78),
(1,Rajiv,Reddy,21,9848022337,Hyderabad,89)})
```

The global maximum of GPA, i.e., maximum among the GPA values of all the students using the **MAX()** function as shown below.

```
grunt> student_gpa_max = foreach student_group_all  Generate
   (student_details.firstname, student_details.gpa), MAX(student_details.gpa);
```

Verification:

Verify the relation **student_gpa_max** using the **DUMP** operator as shown below.

```
grunt > Dump student_gpa_max;
```

Output:

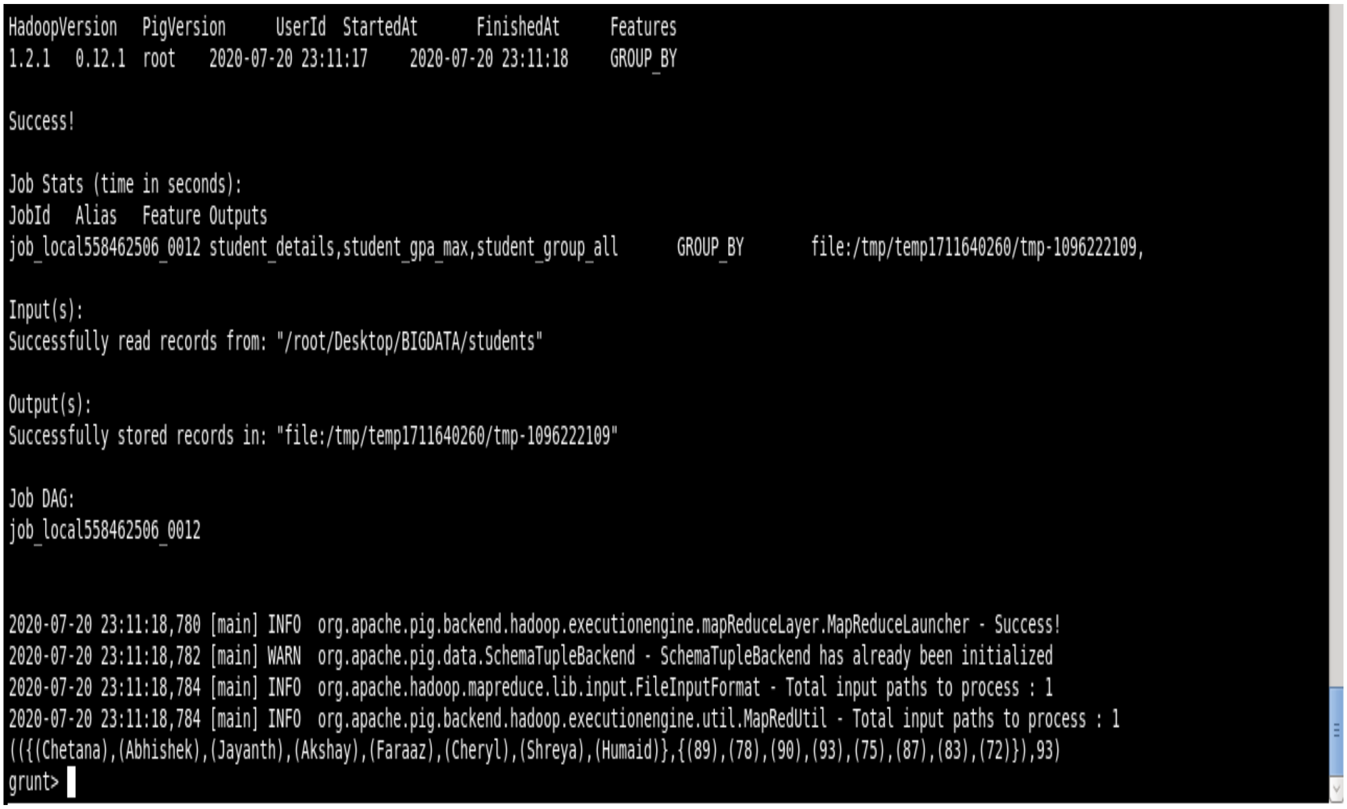
It will produce the as shown, displaying the contents of the relation **student_gpa_max**.

```
(({(Bharathi),(Komal),(Archana),(Trupthi),(Preethi),(Rajesh),(siddarth),(Rajiv) } ,
   {   (72)   ,  (83)  ,   (87)  ,  (75)  ,  (93)  ,  (90)  ,   (78)  , (89)   }) ,93)
```

Program:

```
student_details = LOAD '/root/Desktop/BIGDATA/students' USING
PigStorage(',')

as (id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,
city:chararray, gpa:int);

student_group_all = Group student_details All;

student_gpa_max = foreach student_group_all  Generate

(student_details.firstname, student_details.gpa), MAX(student_details.gpa);

Dump student_gpa_max;
```

Output:



```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
1.2.1   0.12.1  root    2020-07-20 23:11:17     2020-07-20 23:11:18     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local558462506_0012 student_details,student_gpa_max,student_group_all       GROUP_BY        file:/tmp/temp1711640260/tmp-1096222109,

Input(s):
Successfully read records from: "/root/Desktop/BIGDATA/students"

Output(s):
Successfully stored records in: "file:/tmp/temp1711640260/tmp-1096222109"

Job DAG:
job_local558462506_0012


2020-07-20 23:11:18,780 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-07-20 23:11:18,782 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-07-20 23:11:18,784 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2020-07-20 23:11:18,784 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(({(Chetana),(Abhishek),(Jayanth),(Akshay),(Faraaz),(Cheryl),(Shreya),(Humaid)},{(89),(78),(90),(93),(75),(87),(83),(72)}),93)
grunt> 
```