

Pig Assignment

Ashvin Bachoo | Big Data Query | 25/11/2020

# INTRODUCTION:

MapReduce is a way of writing programs designed for processing vast amounts of data, and a system for running those programs in a distributed and fault-tolerant way. Apache Hadoop is one such system designed primarily to run Java code.

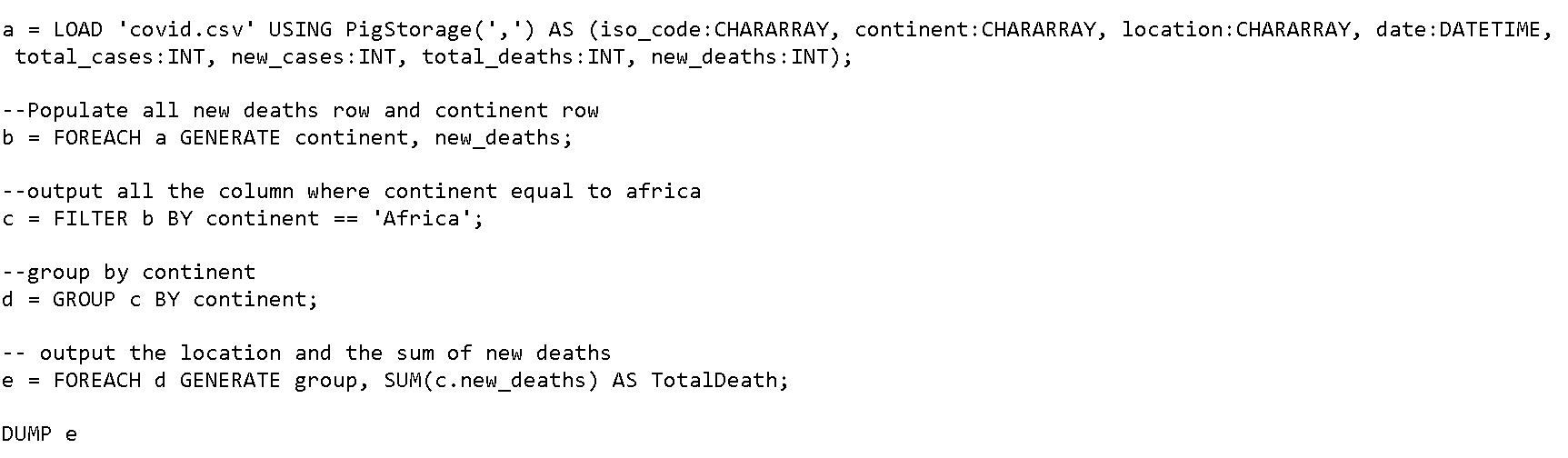
This report consists of five questions.

1. Total number of deaths for Africa.
2. The country with the most number of cases as at now.
3. Which date had the highest number of new cases and in which country?
4. Top ten countries in terms of new cases in the last 2 months.
5. Using the movie dataset, write scripts to output the movie with a rating of 5 but rated by the least number of users.

# Question 1 (Total number of deaths in Africa)

Script:

* Using localhost:4200(maria\_dev) to run/write/execute scripts.
* Using nano filename.pig to write scripts.



Operators:

**LOAD:**

Data is loaded into Apache Pig from the file system (HDFS/ Local) using LOAD operator of Pig Latin(important to mention ‘=’ and where and how to load the data.)

**PigStorage(‘,’):**

We have used the PigStorage() function. It loads and stores data as structured text files. It takes a delimiter using which each entity of a tuple is separated, as a parameter. By default, it takes ‘,’ as a parameter.

**FOREACH/GENERATE**

The FOREACH operator is used to generate specified data transformations based on the column data. In the FOREACH statement, the field in relation B is referred to by positional notation($0).

**Group**:

The GROUP operator is used to group the data in one or more relations. It collects the data having the same key.

**FILTER BY:**

The FILTER operator is used to select the required tuples from a relation based on a condition.

**SUM():**

The SUM() function of Pig Latin is used to get the total of the numeric values of a column in a single-column bag. While computing the total, the SUM() function ignores the NULL values.

**DUMP :**

The Dump operator is used to run the Pig Latin statements and display the results on the screen. It is generally used for debugging Purpose.

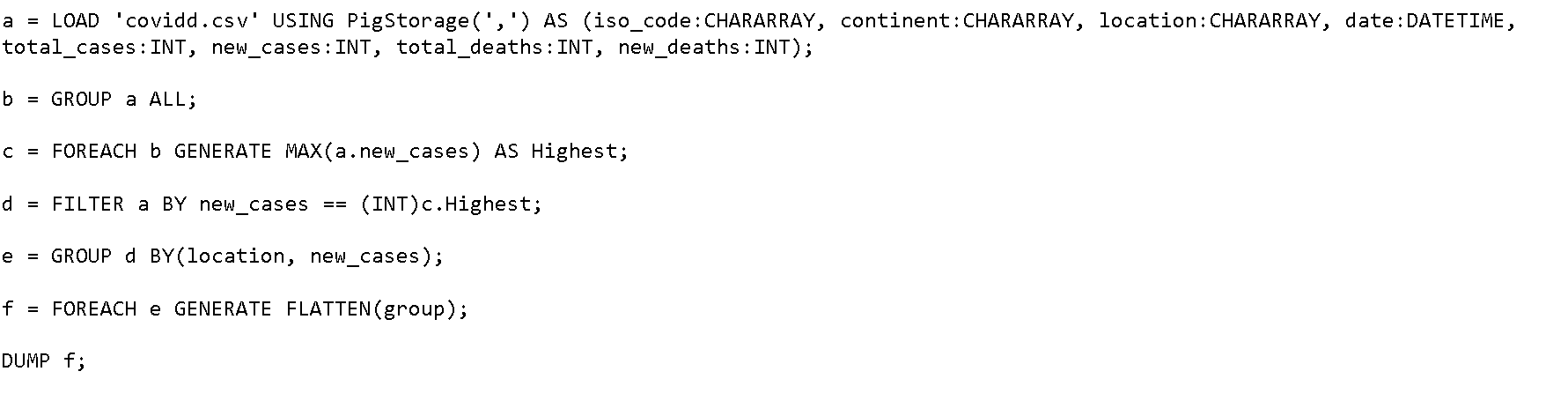
Result:

# C:\Users\Ashvin\Desktop\PIG\QUESTION1_Output.PNG

# QUESTION2: (Country with the greatest number of cases as at now)

Script:

* Using localhost:4200(maria\_dev) to run/write/execute scripts.
* Using nano filename.pig to write scripts



Operators:

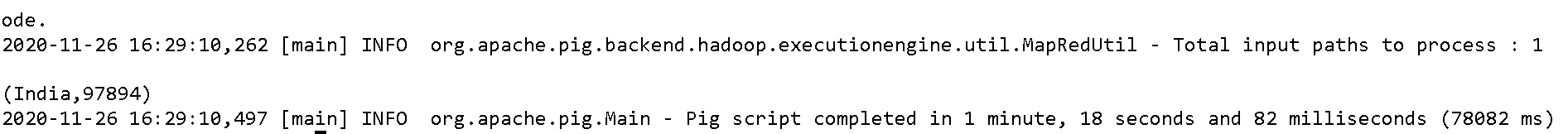
**FLATTEN:**

The FLATTEN operator looks like a UDF syntactically, but it is actually an operator that changes the structure of tuples and bags in a way that a UDF cannot

**MAX():**

This 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.

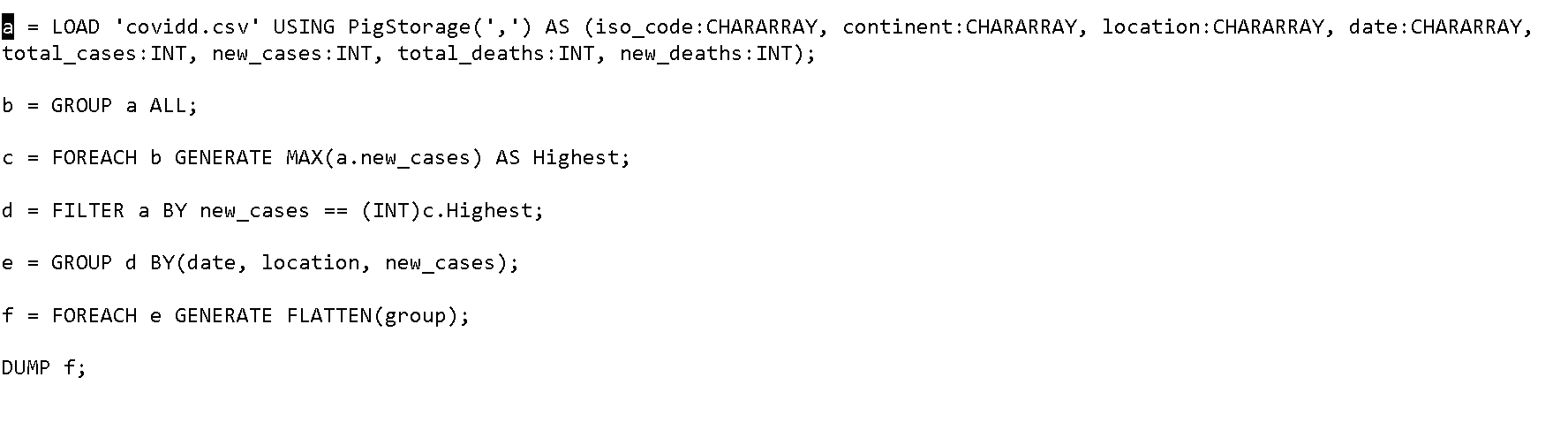
Result:



# QUESTION3: (Which date has the highest number of new cases and which country?)

Script:

* Using localhost:4200(maria\_dev) to run/write/execute scripts.
* Using nano filename.pig to write scripts.



Operators:

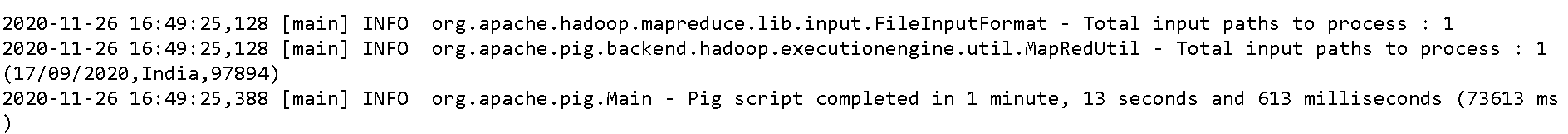
**GROUP filcases BY (dateregis, location, new\_cases):**

Outputs the date, location and new\_cases respectively.

**.HIGHEST:**

Rather using Max() func, (.HIGHEST) func has been used.

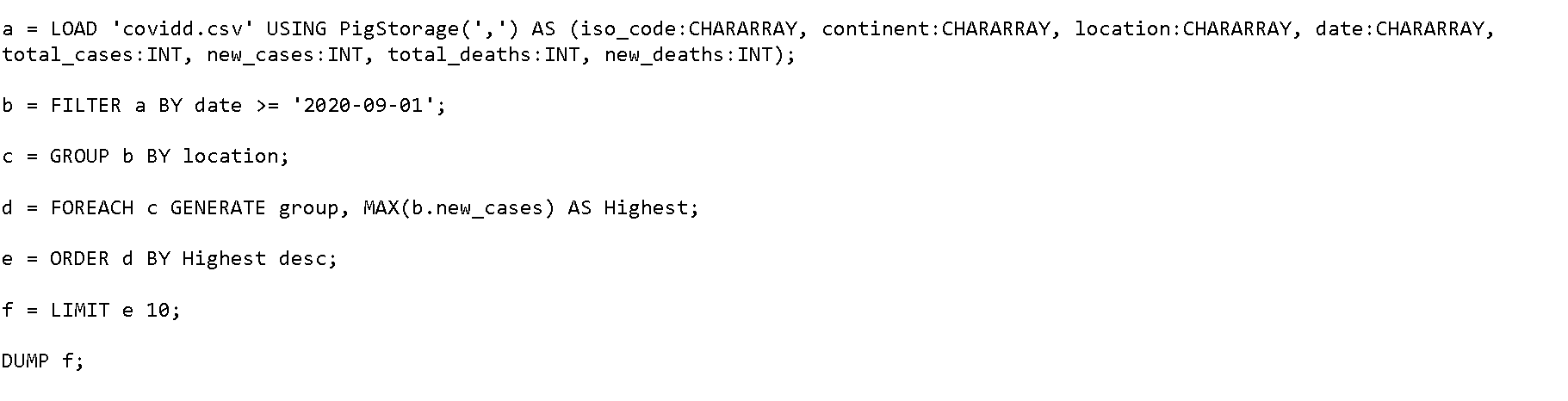
Result:



# QUESTION4: (Top 10 countries in terms of new cases in the last 2 months)

Script:

* Using localhost:4200(maria\_dev) to run/write/execute scripts.
* Using nano filename.pig to write scripts.
* Used datetime to read date with its formatting to further filter date to two last months.



Operators:

**Datetime:**

It is a datatype which read information such as formats of date

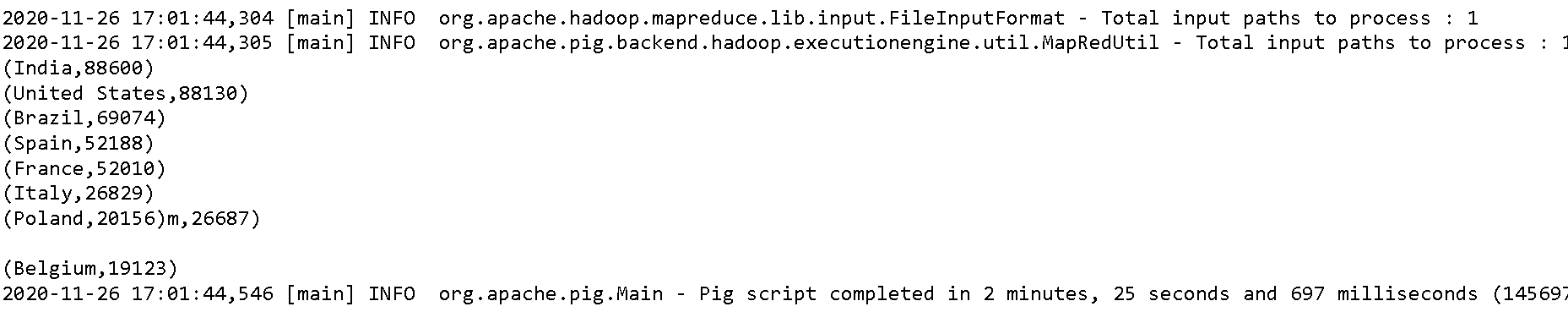
**DESC:**

is a function used along with ORDER BY to read information in descending order

**LIMIT:**

This operator is used to get a limited number of tuples from a relation.

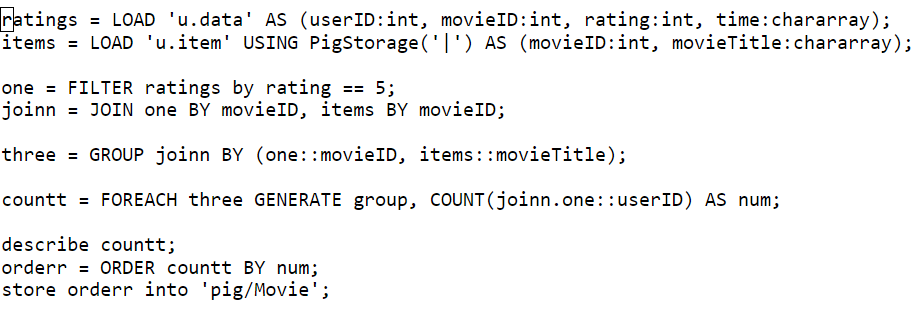
Result:



# Question 5(Using the movie dataset, write scripts to output the movie with a rating of 5 but rated by the least number of users.)

Script:

* Using localhost:4200(maria\_dev) to run/write/execute scripts.
* Using nano filename.pig to write scripts.



Operators:

**JOIN:**

It is used to combine records from two or more relations. While performing a join operation,we declare one (or a group of) tuple(s) from each relation, as keys

**COUNT():**

function of Pig Latin is used to get the number of elements in a bag. While counting the number of tuples in a bag, the COUNT() function ignores (will not count) the tuples having a NULL value in the first field.

**DESCRIBE :**

The describe operator is used to view the schema of a relation.

**STORE INTO:**

Stores the relation in the HDFS directory “/movie/” usually as a file named part-m00000 as shown below.

Result:



Thank you