PDP - Assignment 2 - Pig

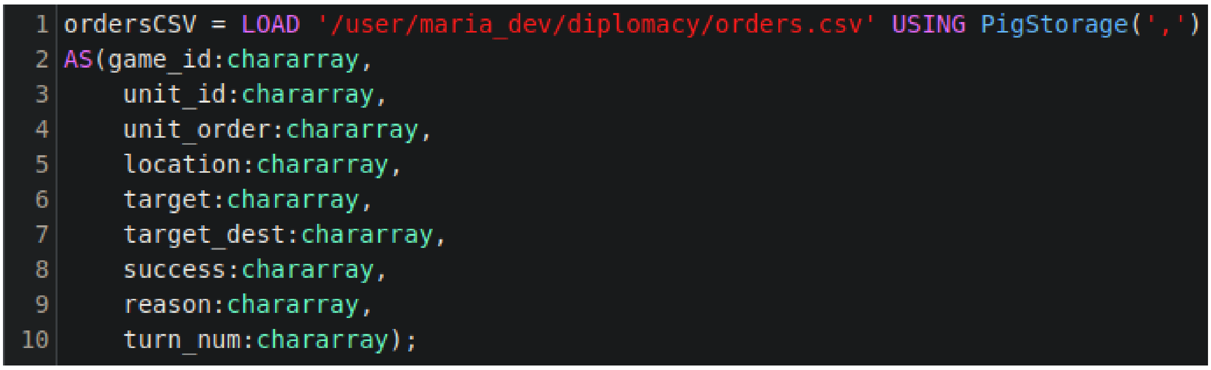
Student: Caitlyn Smith
Number: 644844
Date: 25-06-2022
Version 1
Github url:
https://github.com/C-Smith-InHolland/HadoopAssignmentsYear3/tree/main/assignment2

Part 1: Counting the locations

file: sort_locations

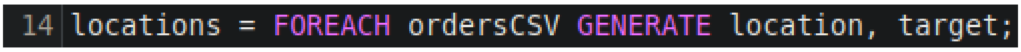
The code:

Load in all the data and columns in the ordersCSV variable.



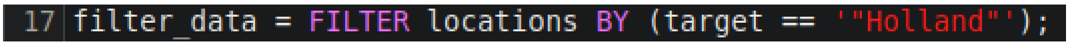
```
ordersCSV = LOAD '/user/maria_dev/diplomacy/orders.csv' USING PigStorage(',')
AS(game_id:chararray,
    unit_id:chararray,
    unit_order:chararray,
    location:chararray,
    target:chararray,
    target_dest:chararray,
    success:chararray,
    reason:chararray,
    turn_num:chararray);
```

From the above data we only need the location and target columns. With the below code we make a new variable only containing these columns.



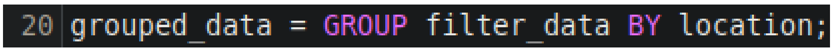
```
locations = FOREACH ordersCSV GENERATE location, target;
```

In this assignment we only want to show the location which has Holland as its target. So for this we filter the data so the target column matches with Holland.



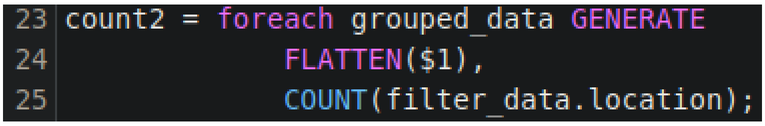
```
filter_data = FILTER locations BY (target == '"Holland"');
```

We then want to group the locations together. For each location we will get a tuple of all the occurrences. For this we group the data on location.



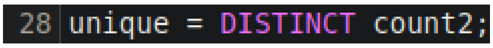
```
grouped_data = GROUP filter_data BY location;
```

From this grouped data we want to go over each location, flatten the data so we get each value on its own line and append the number of tuples that the location has to the end of the line.



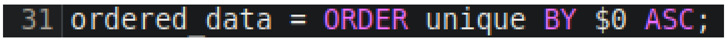
```
count2 = foreach grouped_data GENERATE
             FLATTEN($1),
             COUNT(filter_data.location);
```

Since we now have a lot of duplicate values (1 for each occurence of the location and target combo) we need to only retrieve the unique values using DISTINCT.



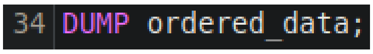
```
unique = DISTINCT count2;
```

Lastly we order the data on alphabetical order based on the location starting at the letter A.



```
ordered_data = ORDER unique BY $0 ASC;
```

And display the result:



```
DUMP ordered_data;
```

To run it:

Load the orders.csv file into hadoop.

Open the pig view.

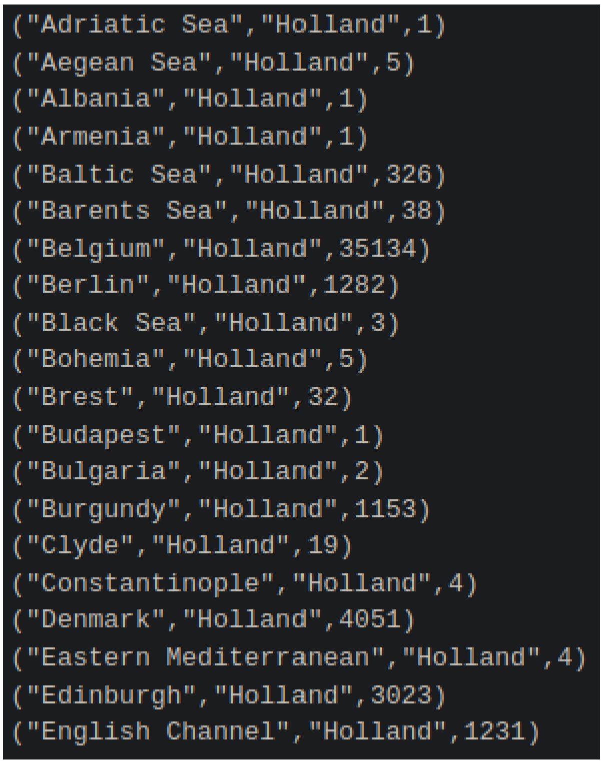
Make a new script.

Paste the code into the script view.

Execute the code using fez.

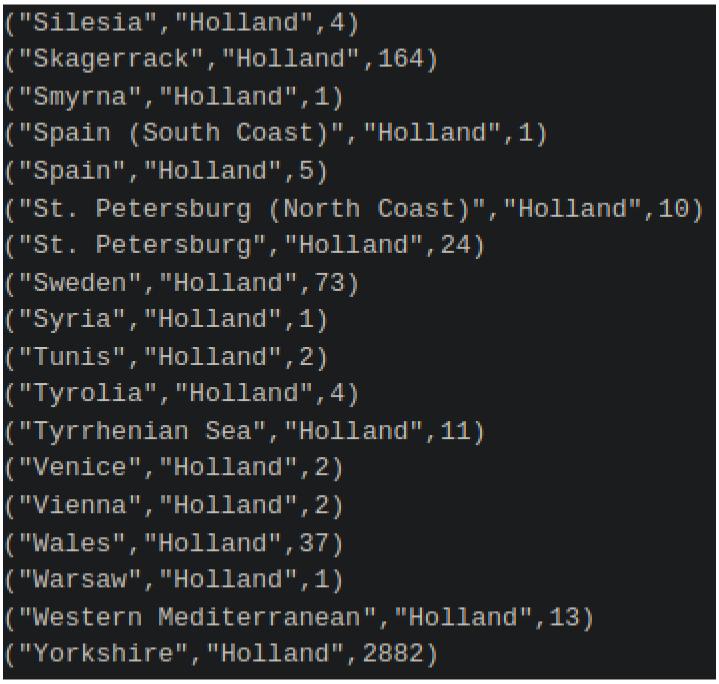
The result:

The top results:



```
("Adriatic Sea","Holland",1)
("Aegean Sea","Holland",5)
("Albania","Holland",1)
("Armenia","Holland",1)
("Baltic Sea","Holland",326)
("Barents Sea","Holland",38)
("Belgium","Holland",35134)
("Berlin","Holland",1282)
("Black Sea","Holland",3)
("Bohemia","Holland",5)
("Brest","Holland",32)
("Budapest","Holland",1)
("Bulgaria","Holland",2)
("Burgundy","Holland",1153)
("Clyde","Holland",19)
("Constantinople","Holland",4)
("Denmark","Holland",4051)
("Eastern Mediterranean","Holland",4)
("Edinburgh","Holland",3023)
("English Channel","Holland",1231)
```

The bottom results:



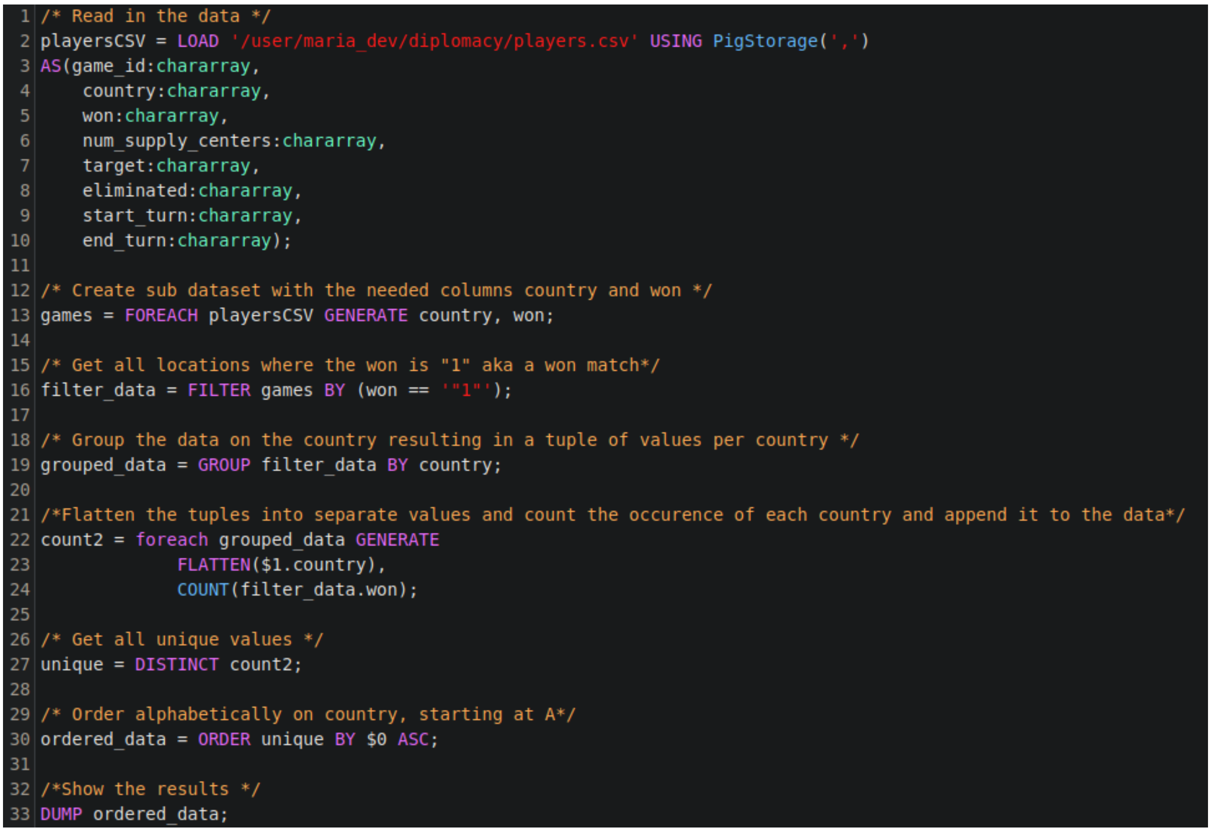
```
("Silesia","Holland",4)
("Skagerrack","Holland",164)
("Smyrna","Holland",1)
("Spain (South Coast)","Holland",1)
("Spain","Holland",5)
("St. Petersburg (North Coast)","Holland",10)
("St. Petersburg","Holland",24)
("Sweden","Holland",73)
("Syria","Holland",1)
("Tunis","Holland",2)
("Tyrolia","Holland",4)
("Tyrrhenian Sea","Holland",11)
("Venice","Holland",2)
("Vienna","Holland",2)
("Wales","Holland",37)
("Warsaw","Holland",1)
("Western Mediterranean","Holland",13)
("Yorkshire","Holland",2882)
```

Part 2: Wins per country

File: winning_countries

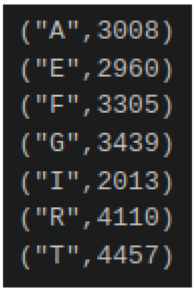
The code:

The code works very similar to the code before but altered to load the players data and to group on the wins.



```
/* Read in the data */
playersCSV = LOAD '/user/maria_dev/diplomacy/players.csv' USING PigStorage(',')
AS(game_id:chararray,
    country:chararray,
    won:chararray,
    num_supply_centers:chararray,
    target:chararray,
    eliminated:chararray,
    start_turn:chararray,
    end_turn:chararray);

/* Create sub dataset with the needed columns country and won */
games = FOREACH playersCSV GENERATE country, won;

/* Get all locations where the won is "1" aka a won match*/
filter_data = FILTER games BY (won == '"1"');

/* Group the data on the country resulting in a tuple of values per country */
grouped_data = GROUP filter_data BY country;

/*Flatten the tuples into separate values and count the occurence of each country and append it to the data*/
count2 = foreach grouped_data GENERATE
             FLATTEN($1.country),
             COUNT(filter_data.won);

/* Get all unique values */
unique = DISTINCT count2;

/* Order alphabetically on country, starting at A*/
ordered_data = ORDER unique BY $0 ASC;

/*Show the results */
DUMP ordered_data;
```

The results:



```
("A",3008)
("E",2960)
("F",3305)
("G",3439)
("I",2013)
("R",4110)
("T",4457)
```