PDP Assignment 2 – Pig

Name: Sjors Grooff

Student number: 634293

Email: 634293@student.inholland.nl

Github url: https://github.com/groofy98/PDP/tree/main/Assignment%202

Date: 9-9-2021

The assignment

I ended up doing the first two exercises of assignment 2. The first one counting the times holland was targeted by location. And secondly the number of wins per country. Both in their own scripts

The script for 2A

```
/* @Author: Sjors Grooff
Date: 09-09-2021

Load the data from the CSV files with the correct datatypes.
I used the apache CSVExcelstorage loader because the csv was formatted as such
 and omitted the quote removal step.
*/
orders = LOAD '/user/maria_dev/orders.csv' USING org.apache.pig.piggybank.stor
age.CSVExcelStorage() AS
    (game_id:int,
    unit_id:int,
    unit_order:chararray,
    location:chararray,
    target:chararray,
    target_dest:chararray,
    success:chararray,
    reason:chararray,
    turn_num:int);

-- We only need moves that target holland so we do a filter operation
targetsHolland = FILTER orders BY (target == 'Holland');

-- Next we need to group all moves by location
targetsHollandByLocation = GROUP targetsHolland BY (location, target);

-
- We need to add a count to the touple and we do this by firt flattening the e
xisting tuple and creating a new tuple with count added
```

```
targetsHollandByLocationWithCount = foreach targetsHollandByLocation generate
flatten(group) as (location, target), COUNT($1);

-- Lastly we order the data by location
result = ORDER targetsHollandByLocationWithCount BY location;

-- Finally show the result
DUMP result;
```

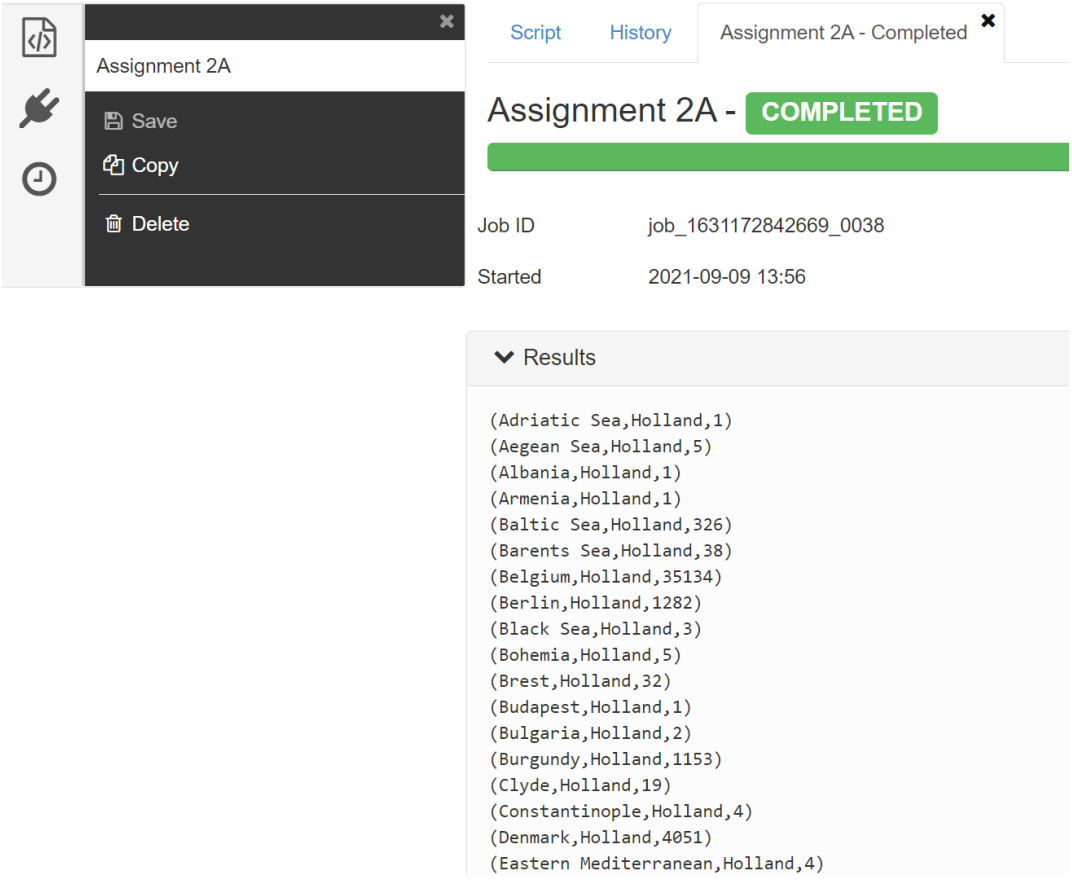
How it works

We start by loading the orders CSV from hdfs using PIG storage. I used the CSVExcelstorage from Apache because the csv was formatted as such. Then we start handling the loaded data in the following order:

1. Filter out the rows that don't target Holland
2. Next we group by location and target (I don't think target is necessary but it looks cleaner)
3. Then we flatten the created groups and add a count to the tuples.
4. The list is then ordered alphabetically by location
5. Finally the list is dumped to output the list

The result of 2A

The picture below shows the result in PIG VIEW:



The script for 2B

```
/* @Author: Sjors Grooff
Date: 09-09-2021

Load the data from the CVS files with the correct datatypes.
I used the CSVExcelstorage loader because the csv was formatted as such and om
itted the quote removal step.
*/
players = LOAD '/user/maria_dev/players.csv' USING org.apache.pig.piggybank.st
orage.CSVExcelStorage() AS
    (game_id:int,
    country:chararray,
    won:int,
    num_supply_centers:int,
    eliminated:int,
    start_turn:int,
    end_turn:int
    );

-- We downselect to only won games
playerWon = FILTER players BY (won == 1);

-- Next we need to group by country
playerWonByCountry = GROUP playerWon BY (country);

-
- We need to add a count to the touple and we do this by firt flattening the e
xisting tuple and creating a new tuple with count added
playerWonByCountryWithCount = foreach playerWonByCountry generate flatten(grou
p) as (country), COUNT($1);

-- Lastly we order the data by location
result = ORDER playerWonByCountryWithCount BY country;

DUMP result;
```

How it works

This time we load the players CSV file the same way as assignment 2A. Then we execute the following operations:

1. Filter and keep the players that won.
2. Next we group by country
3. Then like last time we flatten the tuples and count the amount that each country won
4. To keep the scripts similar I also ordered by country which wasn't requested
5. Finally we dump so the result shows in PIG view

The result for 2B

The results that can be seen in PIG view:

ob ID job_1631172842669_0046

tarted 2021-09-09 20:58

Results

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

How to run it

To get the scripts working the following steps need to be taken:

1. Upload the orders.csv and players.csv to virtualbox via sftp
2. Copy the files from local to hdfs with the ‘hadoop fs -put’ command
3. Create a new script in the ambari Pig view
4. Copy and paste the script of your choice into the editor
5. Execute with TEZ