

Name: Mattis Meeuwesse

Student number: 622948

URL Github: <https://github.com/Maestro1334/Hadoop>

## PDP Assignment 2

Grade	Description	Example result
6	<ul style="list-style-type: none"><li>• Make a alphabetic list from all locations from the orders.csv.</li><li>• Group by "location" with target "Holland"</li><li>• Count how many times Holland was the target from that location</li><li>• Code is executed from Pig View</li><li>• Upload a document on Moodle with the following information:<ul style="list-style-type: none"><li>• Name and student number</li><li>• URL from GitHub with your source code</li><li>• Explain in the document what steps have to be taken to execute the code</li><li>• Explain always in your own words every step included source code.</li><li>• Make a screenshot from the result and include it in your document</li></ul></li></ul>	Adriatic Sea , Holland, 6 Albania, Holland, 5 etc...
7	Extra on grade 6 : <ul style="list-style-type: none"><li>• Make a list how many times every country won</li></ul>	(T,6000) (R,5000) Etc..
8	Make a list from the top 10 countries that had the most turns and won the game.  Names had to be the full name and not the index.	(Russia, 600) (France, 500) Etc..
9	Write UDF with Python that take care replacing "I" with "Italy" etc. explain the code in your win words.	(Russia, 600) (France, 500) Etc..
10	Analyze, what is strange when reading the rules from the original boardgame <a href="https://media.wizards.com/2015/downloads/ah/diplomacy_rules.pdf">https://media.wizards.com/2015/downloads/ah/diplomacy_rules.pdf</a> and comparing the rules with the data? Explain your findings.	

## Steps

1. Have Hadoop configured and up and running according to slides of lesson 3
2. In Hadoop Files view, upload the required CSVs: Moodle -> diplomacy dataset: Players.csv & orders.csv
3. In Pig view, create new script & paste script provided in files into the field OR upload script
4. Execute / Run script

## Output

Script

History

Assignment 2 - Completed ✕

Assignment 2 - **COMPLETED**

Job ID

job\_1632818448565\_0057

Started

2021-09-28 19:17

▼ Results

Download

```
result: {location: chararray,target: chararray,total: long}
("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)
("Finland","Holland",2)
("Galicia","Holland",1)
("Gascony","Holland",6)
("Greece","Holland",3)
("Gulf of Bothnia","Holland",2)
("Gulf of Lyons","Holland",3)
("Helgoland Bight","Holland",9107)
("Holland","Holland",325)
```

▼ Results

Download

```
("Rumania","Holland",2)
("Serbia","Holland",4)
("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)
ordered_result: {country: chararray,wins: long}
("T",4457)
("R",4110)
("G",3439)
("F",3305)
("A",3008)
("E",2960)
("I",2013)
```

## Script part 1

### # Load orders data

```
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);
```

### # Filter data on target: Holland

```
filtered_data = FILTER ordersCSV BY (target == "Holland");
```

### # Group by location (country)

```
group_by_location = GROUP filtered_data BY location;
```

### # Count the number of records for move with "Holland" as target

```
count_specified_location = FOREACH group_by_location GENERATE group as location, "Holland" as  
target, COUNT($1) as total;
```

### # Order the results alphabetically by Location

```
result = ORDER count_specified_location BY location;
```

### # Describe the schema and DUMP the results

```
describe result;
```

```
DUMP result;
```

## Script part 2

### # Load players data

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

### # Filter players by value "1" in won, meaning all lost games/players are filtered out

```
filtered_players = FILTER playersCSV BY (won == "1");
```

### # Group data by country

```
group_by_country = GROUP filtered_players BY country;
```

### # Count the number of won games for each country

```
count_wins = FOREACH group_by_country GENERATE group as country, COUNT($1) as wins;
```

### # Order results by number of wins descending

```
ordered_result = ORDER count_wins BY wins DESC;
```

### # Describe the schema and DUMP the results

```
describe ordered_result;  
DUMP ordered_result;
```