PDP Assignment 2 - Pig

**Naam:** Chris van Loon  
**Studentnummer:** 584694  
**Github url:**

# Code

/\*Load the orders.csv\*/

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);

/\*Remove the extra "" from the csv, code was explained in the lesson\*/

ordersCSV\_Clean = FOREACH ordersCSV GENERATE

REPLACE($0, '"', '') AS (game\_id:chararray),

REPLACE($1, '"', '') AS (unit\_id:chararray),

REPLACE($2, '"', '') AS (unit\_order:chararray),

REPLACE($3, '"', '') AS (location:chararray),

REPLACE($4, '"', '') AS (target:chararray),

REPLACE($5, '"', '') AS (target\_dest:chararray),

REPLACE($6, '"', '') AS (success:chararray),

REPLACE($7, '"', '') AS (reason:chararray),

REPLACE($8, '"', '') AS (turn\_num:chararray);

/\*Filter the csv to only get rows with location Holland\*/

ordersCSV\_Filtered = FILTER ordersCSV\_Clean BY target == 'Holland';

/\*Group the data by location with target being 'Holland'\*/

ordersCSV\_Grouped = GROUP ordersCSV\_Filtered BY (location, target);

/\*Count how many times Holland was target from the location\*/

ordersCSV\_Counted = FOREACH ordersCSV\_Grouped GENERATE FLATTEN(group) as (location, target), COUNT($1);

/\*Order the csv to an alphabetic list by location\*/

ordersCSV\_Ordered = ORDER ordersCSV\_Counted BY location;

DUMP ordersCSV\_Ordered;

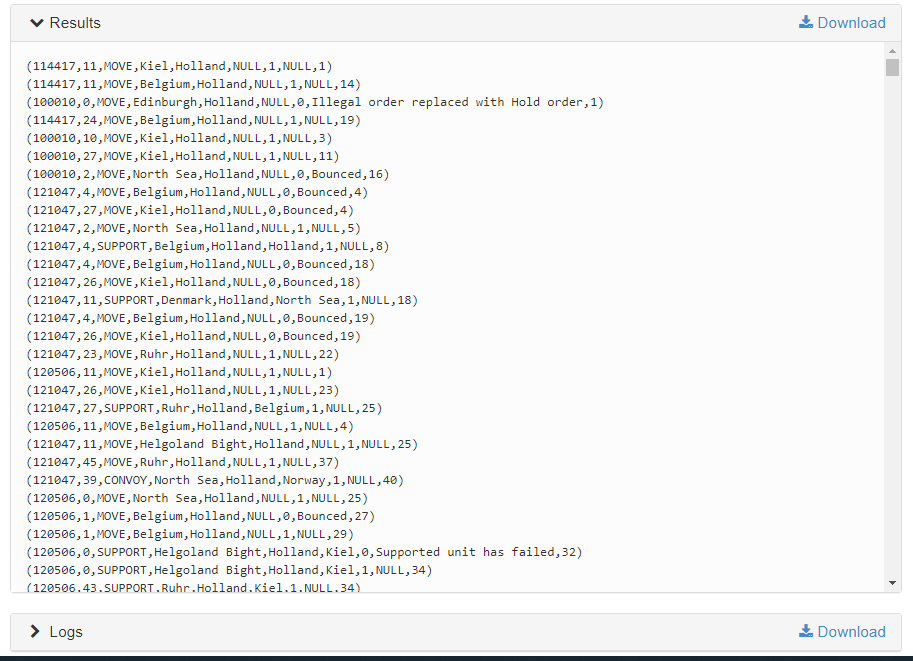
# Explanation how to run the code

There are a couple steps that have to be taken to execute the code. First off the csv files have to be downloaded. Then the user boots up ambari and goes to the files view. After that the user goes to the user/maria\_dev/ directory and creates a folder named ‘diplomacy’. In this the person uploads the needed csv files.

The code starts off by loading the csv from the directory the user just created by using: ordersCSV = LOAD '/user/maria\_dev/diplomacy/orders.csv' USING PigStorage(',').

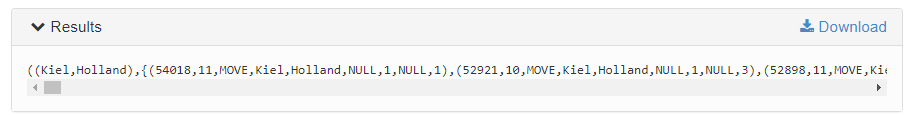
In that line it shows that we load the file from the user/maria\_dev/diplomacy directory. After this the wanted columns are typed and we specify them as chararray.

Next off we make a new variable called ‘ordersCSV\_Clean’ where foreach column we change the text to remove the extra “” ‘s. This is so that when later on we want to get the target of ‘holland’, we don’t get “holland” instead of holland without the “”.

By using the previous cleaned data we can make a variable ‘ordersCSV\_Filtered’ where we can now specify which target to filter by. We do this by using: ordersCSV\_Filtered = FILTER ordersCSV\_Clean BY target == 'Holland';. As you can see, we can now just type: target == ‘holland’;. This will now only show rows where target is holland.

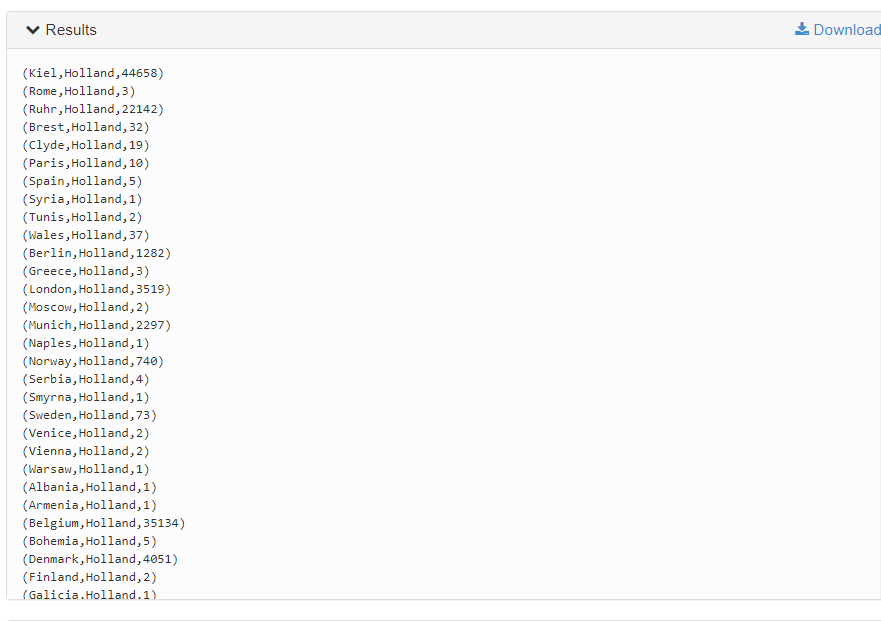
Next off we group the data by location and target using: ordersCSV\_Grouped = GROUP ordersCSV\_Filtered BY (location, target);

which results in:



When we have these groups made we can continue to counting how many times something occurs.We do this by using: ordersCSV\_Counted = FOREACH ordersCSV\_Grouped GENERATE FLATTEN(group) as (location, target), COUNT($1);.

In this line of code we made a foreach in which we go through all groups from ordersCSV\_Grouped where after we use FLATTEN(group) as (location, target), this makes sure that we un-nest the tuple of the ordersCSV\_Grouped so that we can display the location and target in the endresult. This is combined with COUNT($1) which we use to count the amount of times the id occurs in the rows. The result is:



We will now have to order this alphabetically so we use: ordersCSV\_Ordered = ORDER ordersCSV\_Counted BY location;.

Which makes the endresult:

