Name: Mattis Meeuwesse

Student number: 622948

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

Graphical user interface, application

Description automatically generated

# 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

1. In Pig view, create new script & paste script provided in files into the field OR upload script
2. Execute / Run script

# Output

Graphical user interface, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

# 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;

# Descibe 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;

# Descibe the schema and DUMP the results

describe ordered\_result;

DUMP ordered\_result;