**Naam** Stefan de Jong  
**Student-nr** 625583

**Git URL** https://github.com/Stefan-de-jong/pdp\_opdrachten

**Uitleg**

Opdracht 2 bestaat uit twee pig scripts; alphabetic\_holland\_target.pig (voor grade 6) en wins\_per\_country.pig (voor grade 7).

Met de scripting taal ‘Pig Latin’ kun je .pig files schrijven, waar een makkelijkere syntax gebruikt kan worden dan MapReduce (vergelijkbaar met SQL queries). Binnen beiden scripts worden queries als het ware ‘gechained’ of aan elkaar geknoopt. De uitkomtst van iedere query wordt in een variabele opgeslagen, en gebruikt in de volgende query. Dit maakt het makkelijk om complexe taken op te splitsen in kleinere stappen.

Om de scripts te gebruiken dienen de bijgevoegde .csv files (players.csv en orders.csv) geupload te worden (ambari file explorer) naar de maria\_dev folder. Vervolgens kunnen in pig nieuwe scripts gemaakt worden, met daarin de code zoals in alphabetic\_holland\_target.pig en wins\_per\_country.pig

Het uitvoeren kan het beste gedaan worden met ‘Execute on tez’, aangezien dat vele malen sneller gaat.

alphabetic\_holland\_target.pig

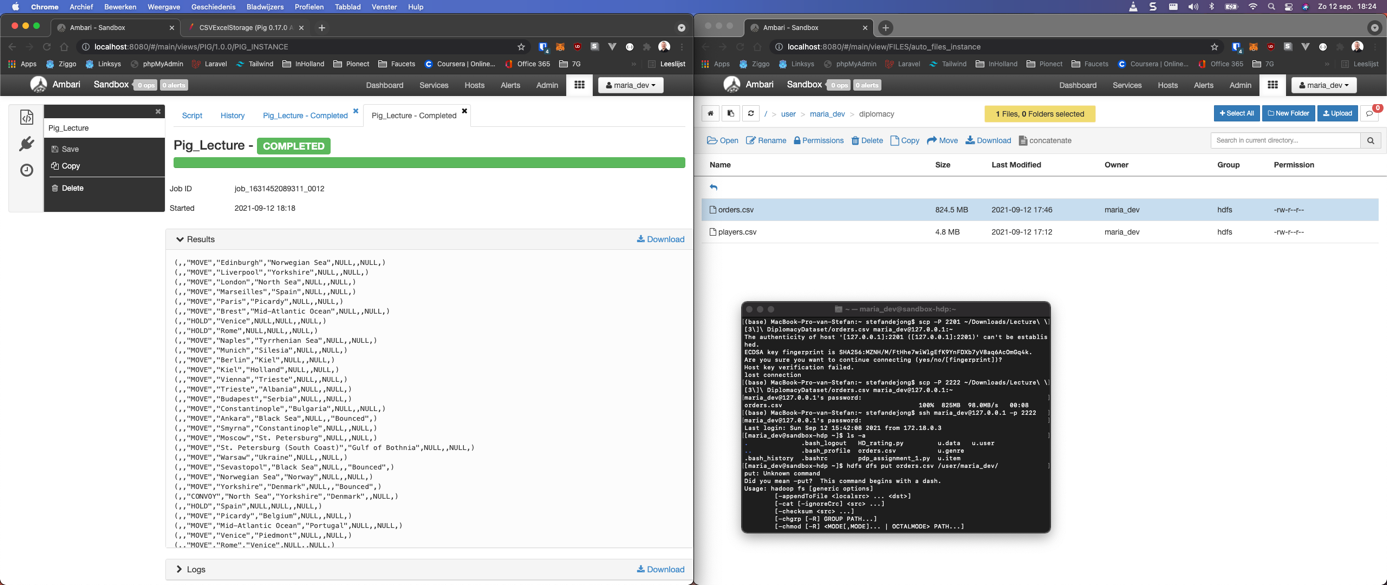
1. Allereerst wordt het orders.csv bestand ingelezen
2. Daarna worden van iedere regel de locatie en doelwit opgeslagen in ‘start\_locations’
3. Start\_locations wordt gefilterd op enkel rijen waar het doelwit ‘Holland’ was, en opgeslagen in ‘targeted\_holland’
4. Targeted\_holland wordt gegroepeerd op locatie, opgeslagen als ‘grouped’
5. Voor iedere regel in Grouped wordt een regel gemaakt met ‘<locatie>, Holland, <aantal keer dat Holland doelwit was>’ als syntax, opgeslagen in ‘counts’
6. Counts wordt gesorteerd op locatie en opgeslagen als ‘order\_by’
7. Order\_by wordt geprint als output

wins\_per\_country.pig

1. Allereerst wordt het players.csv bestand ingelezen
2. Players wordt gefilterd op enkel de rijen waar won ‘true’ is, en opgeslagen in ‘wins’
3. Wins wordt gegroepeerd op country, opgeslagen in ‘grouped\_by’
4. Voor iedere rij in grouped\_by, wordt een regel gemaakt met ‘<land>, <aantal wins>’ als syntax. Opgeslagen in ‘counts’
5. Counts wordt geprints als output

**Screenshots**

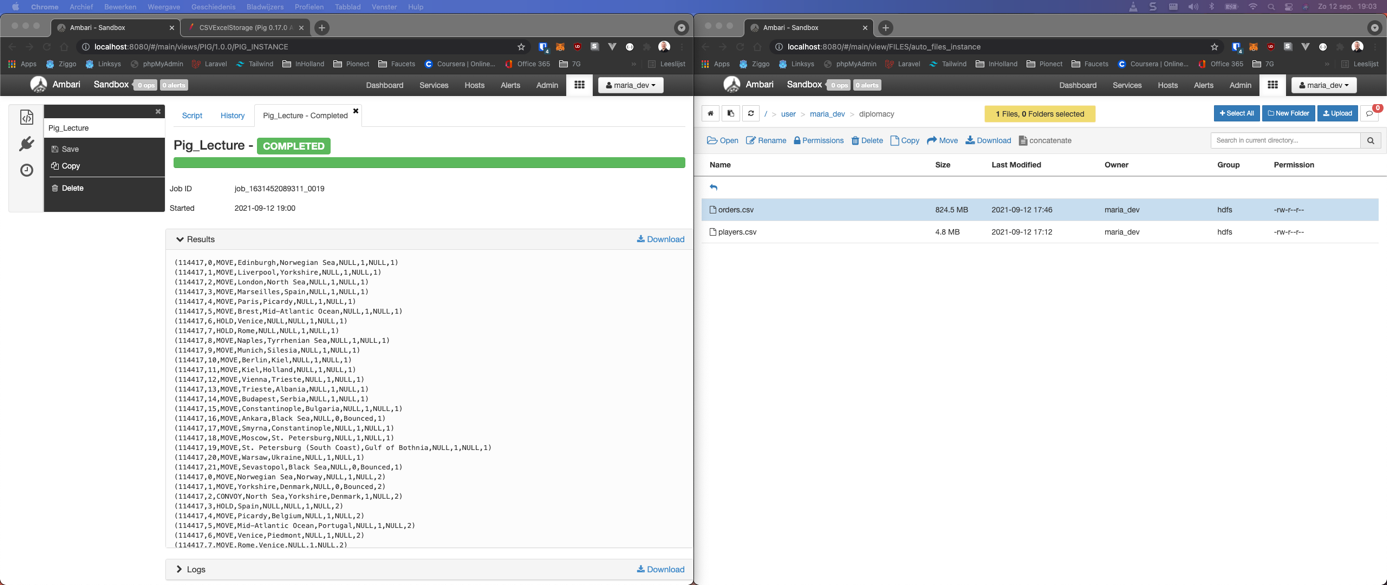
De syntax voor het importeren van de CSV ( USING PigStorage(‘,’) ) gaf verkeerde resultaten (lege velden of NULL waardes):



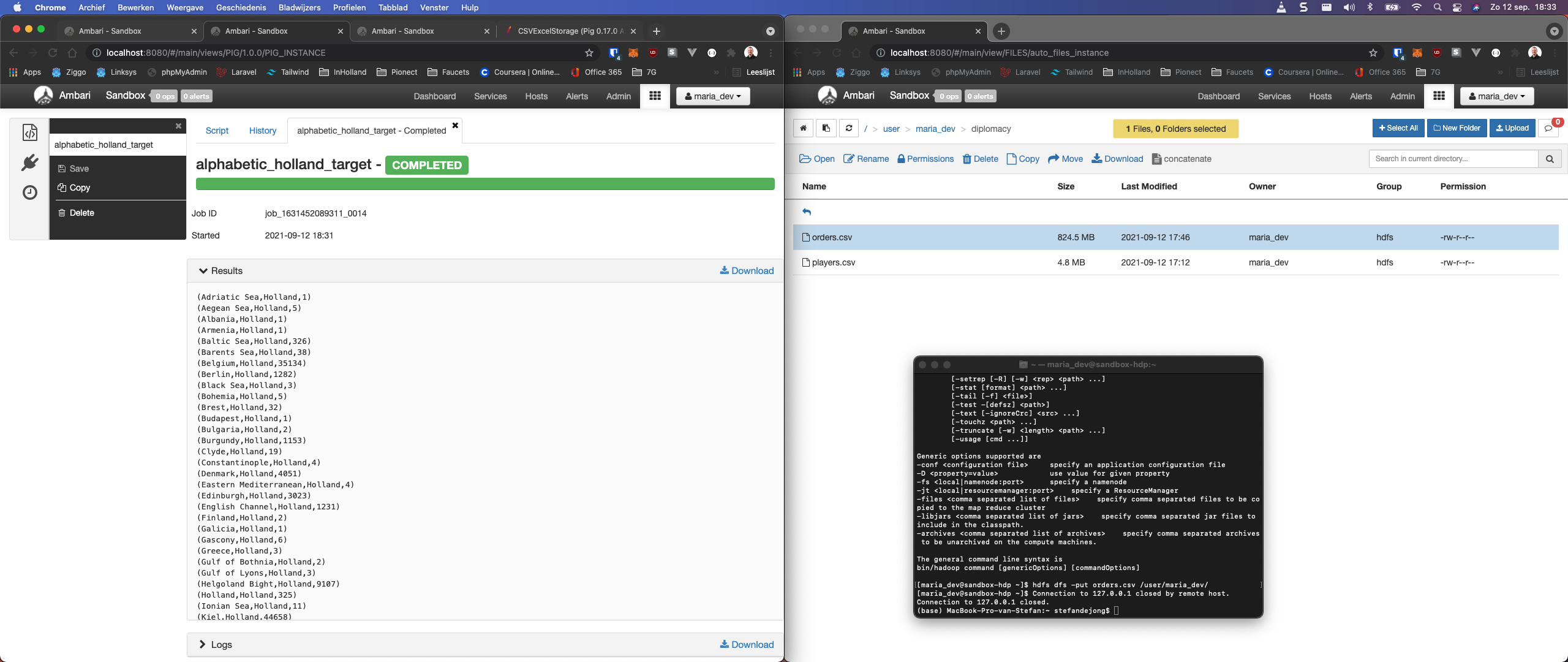
Daarom daarna gebruik gemaakt van deze functie:

<https://pig.apache.org/docs/latest/api/org/apache/pig/piggybank/storage/CSVExcelStorage.html>

Daarna werkte het importeren wel naar behoren:



**Alfabetische lijst van locaties met Holland als doelwit:**



**Overwinningen per land:**

