Assignment 2 – Parallel Distributed Processing

Student: Jesse van Duijne

Student number: 552779

Date of submission: 14-06-2020

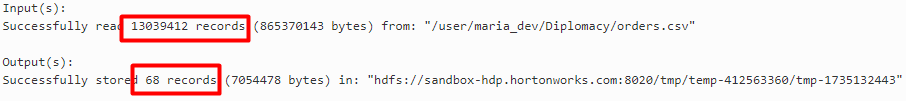
Github link: <https://github.com/Jessevanduijne/4.2_PDP_Assignments>

# Locations

1. **Define:** After several tests of executing a query using PigStorage, it turned out that there was always no result. The logs made it clear that the Excel file was too big to be read by PigStorage. It was for this reason that CSVExcelStorage is currently used. It’s defined at the beginning of the query with *DEFINE CSVExcelStorage org.apache.pig.piggybank.storage.CSVExcelStorage;*

The default delimiter of CSVExcelStorage is a comma (,) so this doesn’t have to be defined.

1. **Load:** The file needs to be loaded. This is simply done by loading the HDFS-file orders.csv and by defining all of its properties as defined in the Diplomacy Codebook
2. **Filter:** Filtering is a logical first step, since grouping all of the unfiltered data would take up more time and resources. Only the rows with *target\_dest == 'Holland’* are being used
3. **Group:** The 13039412 records get grouped to 68 records with *GROUP orders\_filtered BY (location, target).* The records left all start with a location after which all corresponding records are shown, e.g. *(Kiel, {(record), (record), (record), ..})*



‘Target’ will of course always be ‘Holland’, because its filtered on it. It does however make the end result look better cosmetically, it’ll be more clear what the count number means if the target is described as well as the location.

1. **Count:** In each group, a count takes place for the amount of records in that group. This is done with *FOREACH orders\_grouped GENERATE group, COUNT(orders\_filtered).* For every grouped location, the amount of ordered\_filters containing that location is being counted. The schema which is used as input is shown below.



This results in the following unordered list:



1. **Order:** The list still needs to be alphabetically ordered. This is done with *ORDER orders\_unordered BY $0 ASC.* The dollar-symbol stands for its positional notation, which means that the list will be filtered by the first (first is always zero) position which is the location. ‘ASC’ stands for ascending, first alphabet-letter first.
2. **Dump:** print result. Looks as follows:

