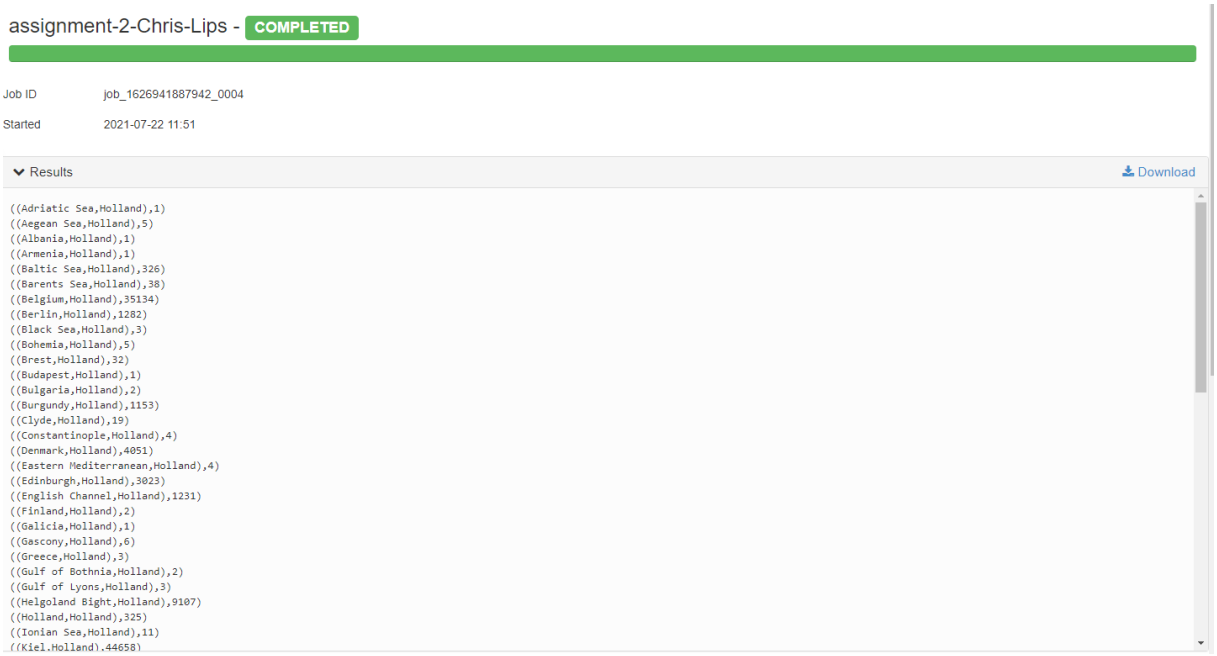
PDP assignment 2 – Chris Lips

Name: Christiaan Lips
Student number: 623434
Github: https://github.com/Lipsinator/Hadoop-fundamentals-Chris-Lips

Steps:

1. To be able to run this assignment the same steps as mentioned in assignment one as far as Hadoop counts. Meaning the Hadoop cluster needs to be created.
2. Secondly when the cluster is up and running we need to go to files view from the ambari windows and select upload file. We want to upload the dimplomacy dataset (from this link: https://moodle.inholland.nl/mod/resource/view.php?id=863082) to the files view so we can use it for the pig file.
3. Go to the pig view and select create new script.
4. Execute the assignment that is included in the file and you're all good to go.

Results:



Code explanation:

1. Load the csv file orders to the user directory.
2. Filter all the entries that contain 'Holland'
3. Group these entries by the location and target being Holland in this case.
4. Count the grouped rows.
5. Order the final locations list to a ascending list for better readability.