Data Intake Report

Name: Data Ingestion Project

Report date: 18th of July 2021

Internship Batch: LISUM01

Version:<1.0>

Data intake by: Enias Vontas

Data intake reviewer:

Data storage location: https://www.kaggle.com/zanjibar/100-million-data-csv?select=custom\_1988\_2020.csv

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 113607322 |
| **Total number of files** | 1, **custom\_1988\_2020.csv** |
| **Total number of features** | 8 |
| **Base format of the file** | .csv |
| **Size of the data** | 4.22 GB |

No authorization was required to access the datasets and no missing values were found. The dataset has no header for the column names, so we had to assume an extra row and then name the columns ourselves.

**Proposed Approaches of reading the dataset:**

* First approach would be to read the dataset with ‘read\_csv’ function from ‘pandas’, which would take about 72.0 seconds.
* Second approach would be to use ‘Dask’ library and execute the ‘read\_csv’ of that, which would take about 0.0135 seconds. A very considerable difference in reading the file, which is expected as the dataset is not stored in memory and thus can read files that exceed our computer’s RAM.
* Third approach would be through ‘modin.pandas’

Presented below is a table with various functions and the time (in seconds) needed to excecute them:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Function** | **Pandas** | **Modin** | **Dask** | **Vaex** |
| read\_csv | 74.0 | 52.7 | 0.016 | 87 |
| groupby(exp\_imp).count() | 2.23 | 8.31 | 22.9 | 0.387 |
| yr.max( ) | 0.111 | 1.63 | 19.4 | 0.111 |
| value.mean( ) | 0.111 | 3.03 | 19.7 | 0.245 |
| LinearRegression( ) | 5.23 | 4.17 | 22.5 | 5.16 |
| describe( ) | 29.7 | MemoryError | 35.1 | 7.47 |

It is clear that Pandas is slower than Modin and Dask when first reading the file but not from Vaex. Pandas makes up for it in the computations such as ‘groupby’ and some statistical functions. If our file was not in ‘.csv’ format, but in ‘.hdf5’, then Vaex library would have been faster in every computation as well as when first reading the data file.

File Summary

After we zip the file in gz format, we have:

|  |  |
| --- | --- |
| **Total number of observations** | 113607322 |
| **Total number of files** | 1, **custom\_1988\_2020.txt.gz** |
| **Total number of features** | 8 |
| **Base format of the file** | .txt.gz |
| **Size of the data** | 1.01 GB |