Data Intake Report

Name: Week 06 Data Ingestion and Schema Validation

Report date: 2022/12/11

Internship Batch: LISUM15

Version: 1.0

Data intake by: Flores, Richard

Data intake reviewer: N/A

Data storage location: Kaggle URL: https://www.kaggle.com/datasets/najzeko/steam-reviews-2021

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 21,700,000 |
| **Total number of files** | 01 |
| **Total number of features** | 23 |
| **Base format of the file** | CSV |
| **Size of the data** | 8.17 GB |

**Proposed Approach:**

* Read the file using Modin and Pandas
* Compare file read speed using Modin and Pandas libraries in term of computational efficiency
* Perform basic validation on data columns: e.g.: remove special character, white spaces from the col name
* Define separator of read and write file, column name in YAML
* Validate number of columns and column name of ingested file with YAML.
* Write the file in pipe separated text file (|) in gz format.
* Create a summary of the file:
* Total number of rows,
* total number of columns
* file size