Data Intake Report

Name: File ingestion and schema validation
Report date: 12/12/2024
Internship Batch: LISUM39

Version: 1.0
Data intake by: Madoka Fujii
Data intake reviewer: Data Glacier
Data storage location:

Tabular data details: train.csv

Total number of observations	100000
Total number of files	1
Total number of features	5
Base format of the file	.csv
Size of the data	2GB

Proposed Approach:

-Find a large size of CSV
-Try to read with Pandas, Pandas with chunks, Dask, Dask with chunks, Ray and check the speed to complete
-Conduct basic Validations (check missing data, white spaces from the col name, and data type)
-Generate YAML and util.py
-Validate with YAML file
-Create a gz format of CSV with pipe separated text file (|)
-Summary the total number of rows, total number of columns, and the file size