Data Intake Report

Name: Week 6: File ingestion and schema validation.
Report date: 12/06/2024
Internship Batch: LISUM33
Version: 1.0
Data intake by: Rakshith Nagaraju
Data intake reviewer: -
Data storage location: https://github.com/Rakshith-611/Data-Glacier/tree/Week-6

Tabular data details:

Total number of observations (rows)	42448764
Total number of files	1
Total number of features (columns)	9
Base format of the file	.csv
Size of the data	5.67 GB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- Try out different methods to read the data (pandas, ray, dask, modin).
- Create utility file.
- Create YAML configuration file.
- Perform column validation for reading the data into the pipeline.
- If the file is valid then compress the data into a pipe "|" separated .gz file.