Data Intake Report

Name: Data Ingestion Pipeline
Report date: 10 October, 2022
Internship Batch: LISUM13: 30
Data intake by: Jay Panara
Data intake reviewer: Data Glacier
Data storage location: https://github.com/jaypanara/Data_Glacier_Internship/tree/main/Week6

Tabular data details:

Total number of observations	22489348
Total number of files	4
Total number of features	11
Base format of the file	.csv
Size of the data	2.24 GB

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

Proposed Approach:

- **At first a test utility file is written with all the functions used in the code.**
- **Next a YAML file is created.**
- **Next step is the reading of the file using config file. Here two methods are used for reading the csv file, one with pandas and the other with Dask.**
- **Next is the removal of special characters and white spaces.**
- **Followed by validation of the files.**
- **Later on a GZ compressed file is created.**