Data Intake Report

Name: File ingestion and schema validation
Report date: 07th April 2023
Internship Batch: LISUM19
Version: 1.0
Data intake by: B. Harika
Data intake reviewer: Data Glacier
Data storage location:
https://github.com/HarikaReddyB/File_ingestion_and_schema_validation

Tabular data details: Yellow_tripdata_2015

Total number of observations	1048575
Total number of files	1
Total number of features	19
Base format of the file	csv
Size of the data	150.0+ MB

Proposed Approach:

- The data was read using different methods like Pandas, Dask, Ray, Vaex and the vaex has least reading time when compared to others.
- Performed basic validation on the data columns to remove special characters, white space, duplicates and null values.
- Validated the number of columns inorder to avoid errors while processing the data.
- Converted the file into pipe separated text file in gz format, to reduce the file size.
- Created the summary of table to provide the insights of the data, this contains number of rows, columns, and the File size.