Data Intake Report

Name: Data Ingestion
Report date: 2023.1.6
Internship Batch: LISUM16
Version:<1.0>
Data intake by: Zeru Zhou
Data intake reviewer: NA
Data storage location: Kaggle dataset. I use Google Colab to scrape the data and didn't store it locally. https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store?select=2019-Oct.csv

Tabular data details:

Total number of observations	42448764
Total number of files	1
Total number of features	9
Base format of the file	.csv
Size of the data	5406 Mb

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

Proposed Approach:

- Read the data in several ways including pandas, dask, ray, and modin
- Designed utility script that allow user to validate the columns and extracted general information about the dataset.
- Wrote YAML file that store the configuration information about the dataset.
- Implement column validation using the script and YAML file.
- No assumptions from data.