Name: Matthew Zhou

Batch code: LISUM20

Submission Date: 5/10/2023

Submitted to: Canvas/Github

Data Ingestion Pipeline

Reading CSV files with:

Pandas - A screen shot of a computer code

Description automatically generated with low confidence

Dask –

A screenshot of a computer

Description automatically generated with medium confidence

Modin –

A screen shot of a computer code

Description automatically generated with low confidence

Ray –

A picture containing text, font, screenshot

Description automatically generated

Dask and Ray were the best at parallelizing the reading of files. Pandas and Modin had to read each file in sequence, which is very inefficient.

Column Validation:

A screen shot of a computer

Description automatically generated with medium confidence

Outputted to single file