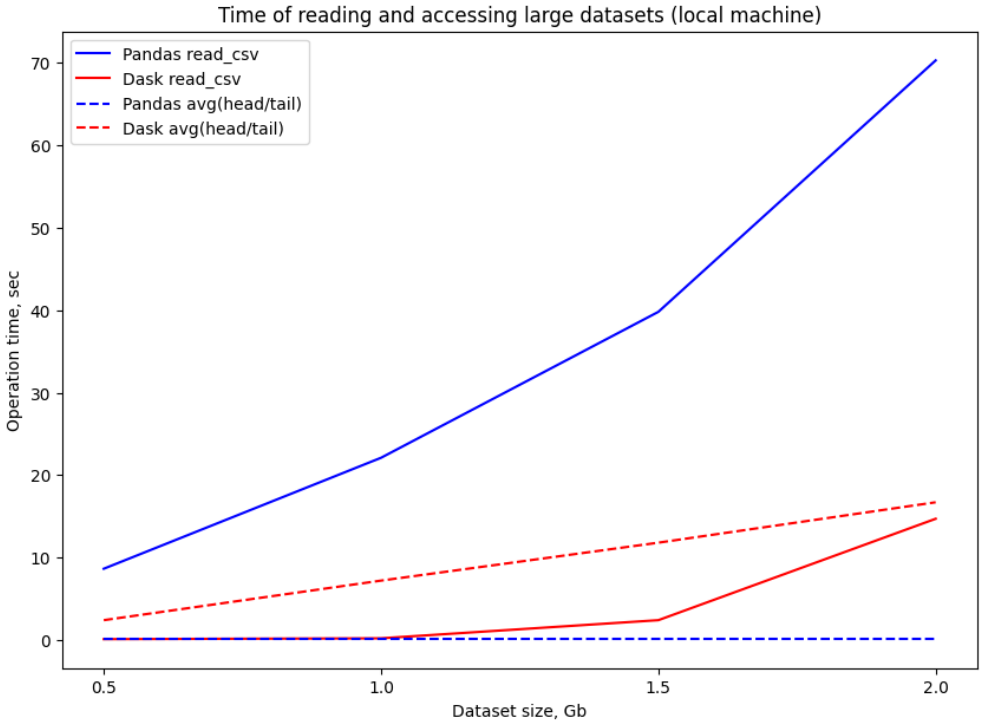
Week 6 Assignment - Week 6: File ingestion and schema validation

Name: Dmitry Sharukhin
E-mail: sharuhinda@gmail.com
Batch code: LISUM15
Submission date: Dec 11, 2022
Submitted to: Data Glacier

1. Read file (read_csv()) and data access (head() / tail()) operations were tested on datasets of the size 0.5, 1.0, 1.5 and 2.0 Gb on local machine using Pandas and Dask libraries



Here we can see the difference between Pandas and Dask packages performing file read and data access operations. Pandas is winning in data access time as long as the dataset fits into RAM. But time of reading large files rises significantly as the dataset size grows. Dask demonstrates less significant dependency between the size of the dataset and data read and accesss time but for the price of increased data access time.

2. Created config file and utils code to perform columns validation (available at https://github.com/sharuhinda/data_glacier_vi/tree/main/week_6). Both initial and processed files were located at local hard drive.

Final processed file contains:

rows: 8 599 212

columns: 7

size (pure .csv format): 0.51 Gb

size (.csv compressed by gzip): 0.07 Gb