Data Intake Report

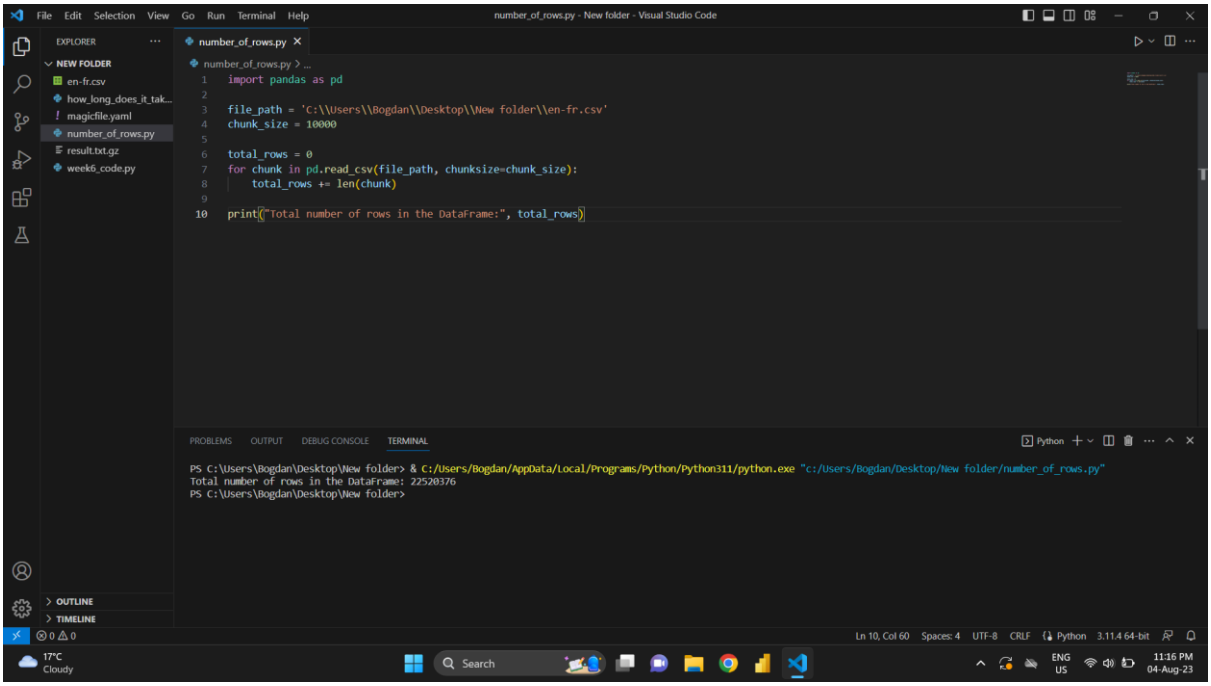
Name: en-fr-translation-dataset
Report date: 12/06/2023
Internship Batch: LISUM21
Version: 1.0
Data intake by: Bogdan-Remus Pintilie Data
intake reviewer:
Data storage location: https://www.kaggle.com/datasets/dhruvildave/en-fr-translation-dataset

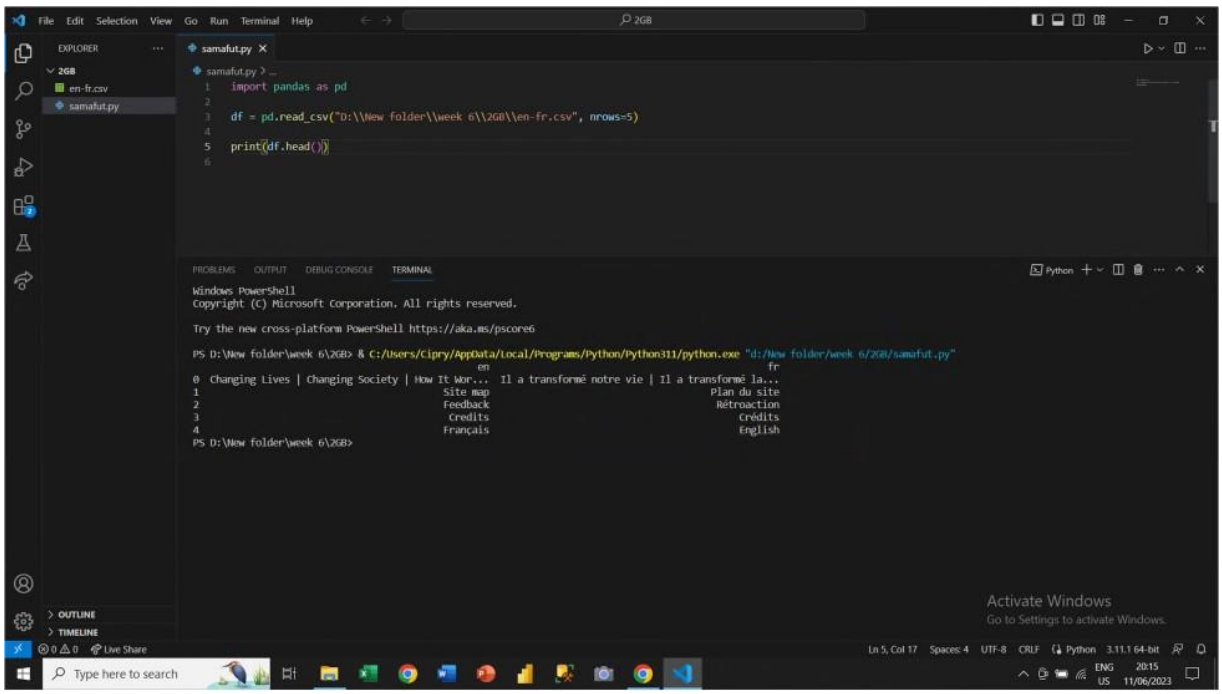
Tabular data details:

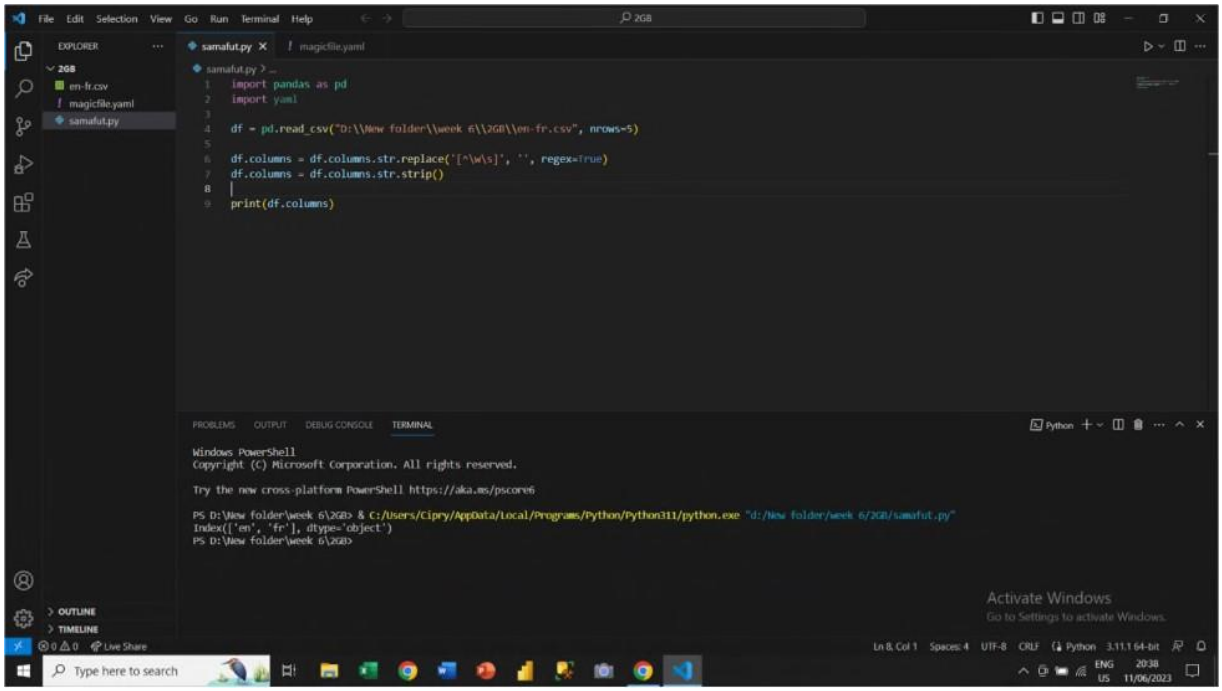
Total number of observations	22520376
Total number of files	1
Total number of features	2
Base format of the file	.csv
Size of the data	8.2 GB

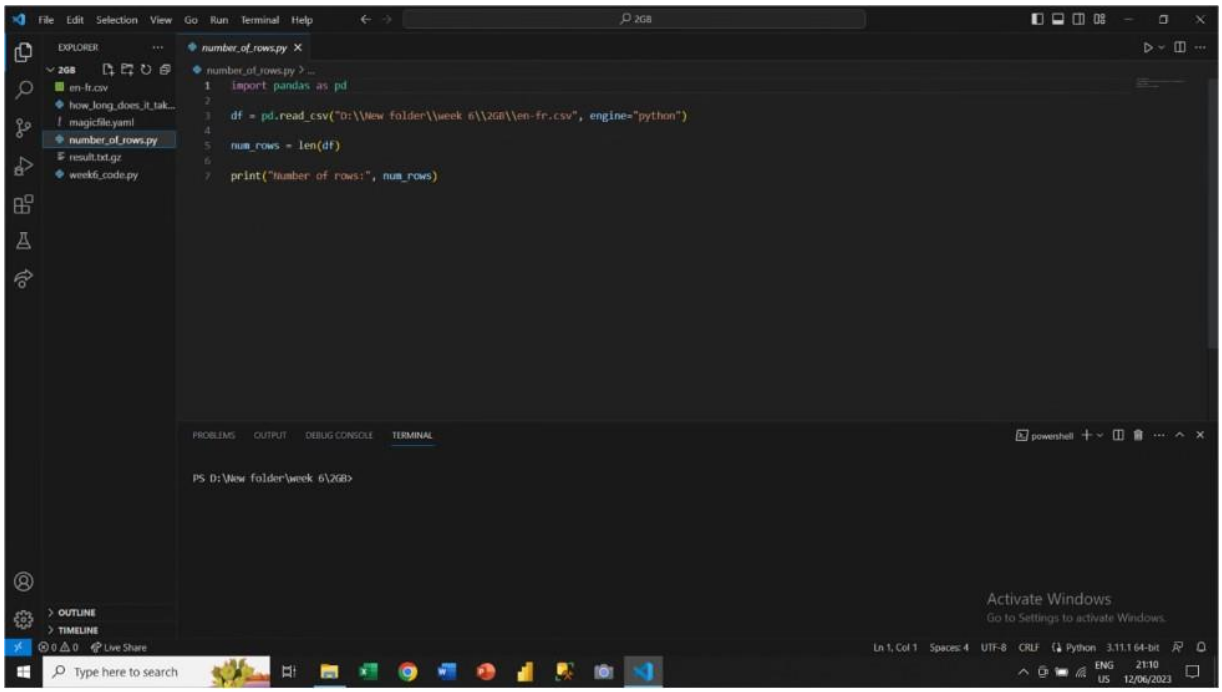
Proposed Approach:

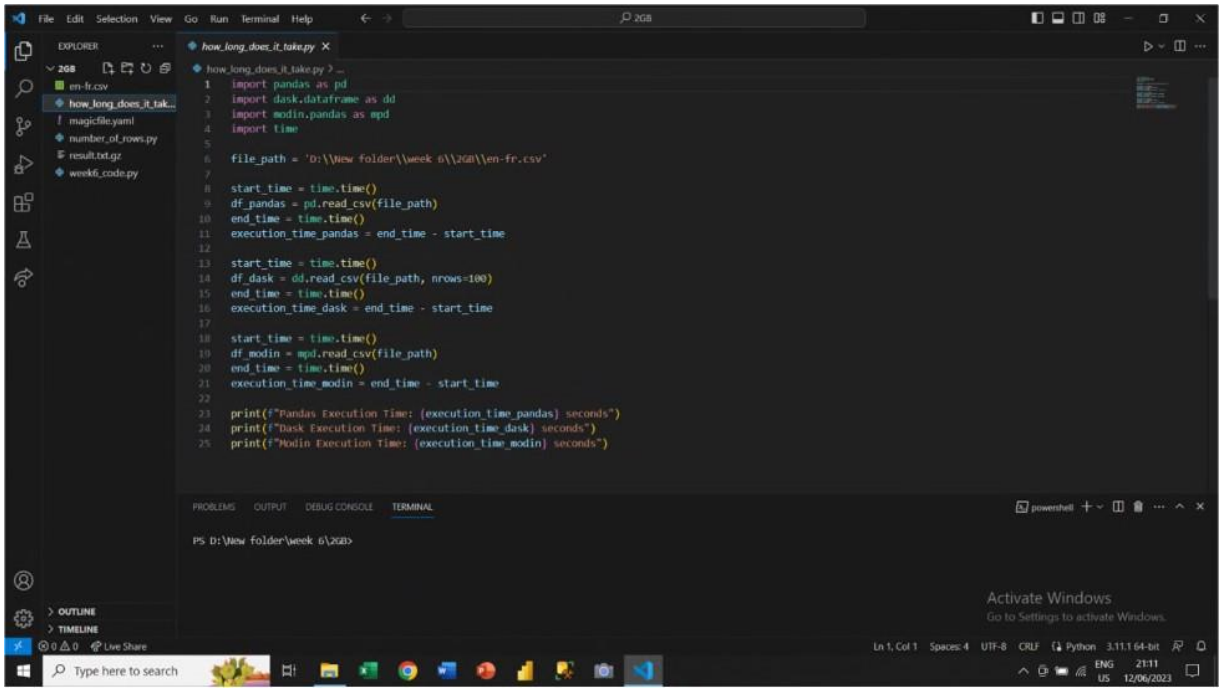
- Set up an environment
- Upload csv file
- Read the file using different methods (pandas, dask, modin)
- Use a smaller subset of the data, reading the entire file at once can be memory-intensive
- Measure the computational efficiency of each method and document findings
- Perform basic validation on data columns (remove special characters and white spaces)
- Create a YAML file defining the separator and column names
- Read the YAML file and validate the number of columns and column names
- Write the file in pipe-separated (“,”) text format in gzipped format

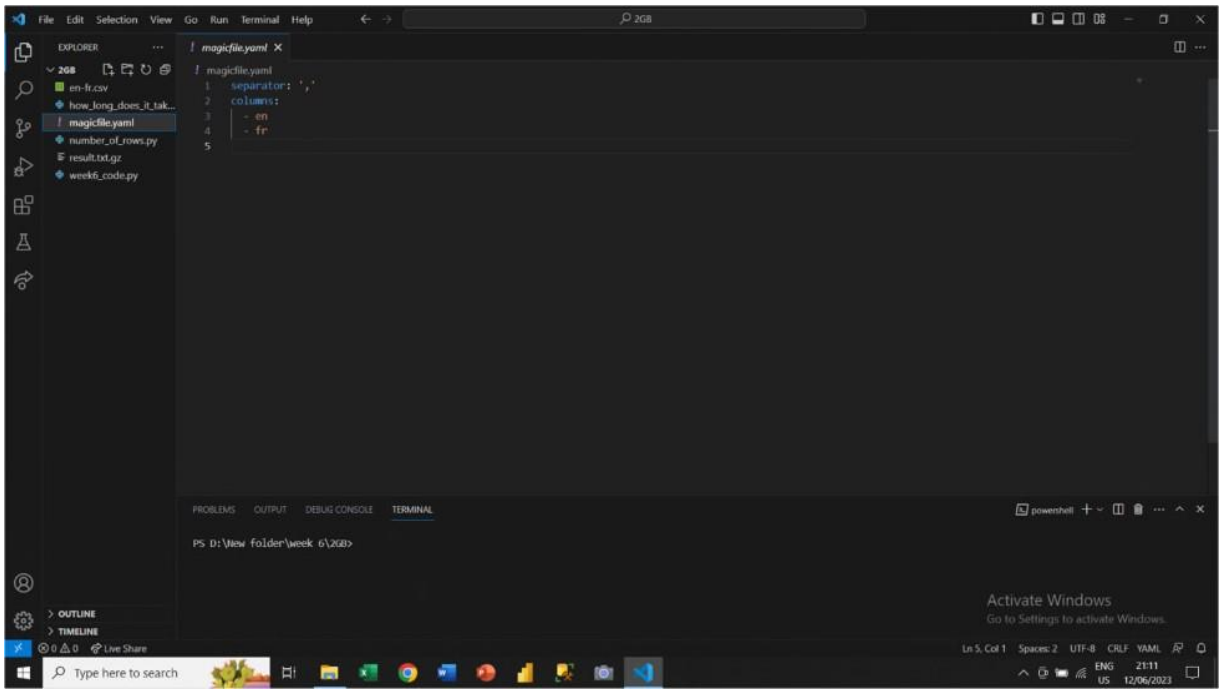

File Edit Selection View Go Run Terminal Help
number_of_rows.py - New folder - Visual Studio Code
EXPLORER
NEW FOLDER
en-fr.csv
how_long_does_it_tak...
magicfile.yaml
number_of_rows.py
result.txt.gz
week6_code.py
number_of_rows.py
import pandas as pd
file_path = 'C:\\Users\\Bogdan\\Desktop\\New folder\\en-fr.csv'
chunk_size = 10000
total_rows = 0
for chunk in pd.read_csv(file_path, chunksize=chunk_size):
total_rows += len(chunk)
print("Total number of rows in the DataFrame:", total_rows)
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
PS C:\Users\Bogdan\Desktop\New folder> & C:/Users/Bogdan/AppData/Local/Programs/Python/Python311/python.exe "c:/Users/Bogdan/Desktop/New folder/number_of_rows.py"
Total number of rows in the DataFrame: 22520376
PS C:\Users\Bogdan\Desktop\New folder>
OUTLINE
TIMELINE
Ln 10, Col 60 Spaces: 4 UTF-8 CRLF Python 3.11.4 64-bit
17°C Cloudy
Search
ENG US
11:16 PM 04-Aug-23

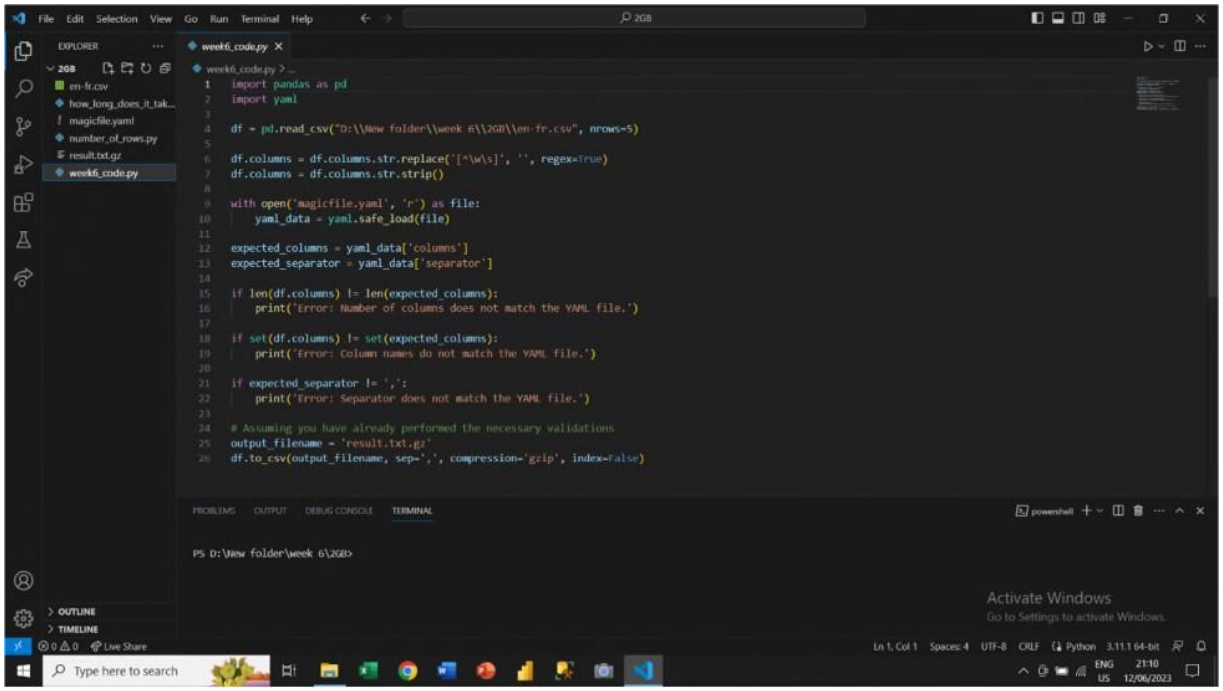

File Edit Selection View Go Run Terminal Help
2GB
EXPLORER
2GB
en-fr.csv
samafut.py
samafut.py
import pandas as pd
df = pd.read_csv("D:\\New folder\\week 6\\2GB\\en-fr.csv", nrows=5)
print(df.head())
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.
Try the new cross-platform PowerShell https://aka.ms/pscore6
PS D:\New folder\week 6\2GB> & C:/Users/Cipry/AppData/Local/Programs/Python/Python311/python.exe "d:/New folder/week 6/2GB/samafut.py"
en fr
0 Changing Lives | Changing Society | How It Wor... Il a transformé notre vie | Il a transformé la...
1 Site map Plan du site
2 Feedback Rétroaction
3 Credits Crédits
4 Français English
PS D:\New folder\week 6\2GB>
OUTLINE
TIMELINE
Activate Windows
Go to Settings to activate Windows.
Ln 5, Col 17 Spaces: 4 UTF-8 CRLF Python 3.11.1 64-bit
Live Share
Type here to search
ENG US
20:15 11/06/2023


File Edit Selection View Go Run Terminal Help
2GB
EXPLORER
2GB
en-fr.csv
magicfile.yaml
samafut.py
samafut.py
magicfile.yaml
samafut.py > ...
1 import pandas as pd
2 import yaml
3
4 df = pd.read_csv("D:\\New folder\\week 6\\2GB\\en-fr.csv", nrows=5)
5
6 df.columns = df.columns.str.replace('[^\w\s]', '', regex=True)
7 df.columns = df.columns.str.strip()
8
9 print(df.columns)
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
Python
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.
Try the new cross-platform PowerShell https://aka.ms/pscore6
PS D:\New folder\week 6\2GB> & C:/Users/Cipry/AppData/Local/Programs/Python/Python311/python.exe "d:/New folder/week 6/2GB/samafut.py"
Index(['en', 'fr'], dtype='object')
PS D:\New folder\week 6\2GB>
OUTLINE
TIMELINE
Live Share
Activate Windows
Go to Settings to activate Windows.
Ln 8, Col 1 Spaces: 4 UTF-8 CRLF Python 3.11.1 64-bit
Type here to search
ENG US 20:38 11/06/2023


File Edit Selection View Go Run Terminal Help
2GB
EXPLORER
2GB
en-fr.csv
how_long_does_it_tak...
magicfile.yaml
number_of_rows.py
result.txt.gz
week6_code.py
number_of_rows.py
number_of_rows.py > ...
1 import pandas as pd
2
3 df = pd.read_csv("D:\\New folder\\week 6\\2GB\\en-fr.csv", engine="python")
4
5 num_rows = len(df)
6
7 print("Number of rows:", num_rows)
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
powershell
PS D:\New folder\week 6\2GB>
OUTLINE
TIMELINE
Live Share
Activate Windows
Go to Settings to activate Windows.
Ln 1, Col 1 Spaces: 4 UTF-8 CRLF Python 3.11.1 64-bit
Type here to search
ENG US 21:10 12/06/2023


File Edit Selection View Go Run Terminal Help
2GB
EXPLORER
how_long_does_it_take.py
2GB
en-fr.csv
how_long_does_it_tak...
magicfile.yaml
number_of_rows.py
result.txt.gz
week6_code.py
how_long_does_it_take.py > ...
1 import pandas as pd
2 import dask.dataframe as dd
3 import modin.pandas as mpd
4 import time
5
6 file_path = 'D:\\New folder\\week 6\\2GB\\en-fr.csv'
7
8 start_time = time.time()
9 df_pandas = pd.read_csv(file_path)
10 end_time = time.time()
11 execution_time_pandas = end_time - start_time
12
13 start_time = time.time()
14 df_dask = dd.read_csv(file_path, nrows=100)
15 end_time = time.time()
16 execution_time_dask = end_time - start_time
17
18 start_time = time.time()
19 df_modin = mpd.read_csv(file_path)
20 end_time = time.time()
21 execution_time_modin = end_time - start_time
22
23 print(f"Pandas Execution Time: {execution_time_pandas} seconds")
24 print(f"Dask Execution Time: {execution_time_dask} seconds")
25 print(f"Modin Execution Time: {execution_time_modin} seconds")
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
powershell
PS D:\New folder\week 6\2GB>
OUTLINE
TIMELINE
Live Share
Activate Windows
Go to Settings to activate Windows.
Ln 1, Col 1 Spaces: 4 UTF-8 CRLF Python 3.11.1 64-bit
Type here to search
ENG US 21:11 12/06/2023


File Edit Selection View Go Run Terminal Help
2GB
EXPLORER
magicfile.yaml
2GB
en-fr.csv
how_long_does_it_tak...
magicfile.yaml
number_of_rows.py
result.txt.gz
week6_code.py
magicfile.yaml
1 separator: ','
2 columns:
3 - en
4 - fr
5
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
powershell
PS D:\New folder\week 6\2GB>
OUTLINE
TIMELINE
Live Share
Activate Windows
Go to Settings to activate Windows.
Ln 5, Col 1 Spaces: 2 UTF-8 CRLF YAML
Type here to search
ENG US 21:11 12/06/2023


File Edit Selection View Go Run Terminal Help
2GB
EXPLORER
2GB
en-fr.csv
how_long_does_it_tak...
magicfile.yaml
number_of_rows.py
result.txt.gz
week6_code.py
week6_code.py
week6_code.py > ...
import pandas as pd
import yaml
df = pd.read_csv("D:\\New folder\\week 6\\2GB\\en-fr.csv", nrows=5)
df.columns = df.columns.str.replace('[^\w\s]', '', regex=True)
df.columns = df.columns.str.strip()
with open('magicfile.yaml', 'r') as file:
yaml_data = yaml.safe_load(file)
expected_columns = yaml_data['columns']
expected_separator = yaml_data['separator']
if len(df.columns) != len(expected_columns):
print('Error: Number of columns does not match the YAML file.')
if set(df.columns) != set(expected_columns):
print("Error: Column names do not match the YAML file.')
if expected_separator != ',':
print('Error: Separator does not match the YAML file.')
# Assuming you have already performed the necessary validations
output_filename = 'result.txt.gz'
df.to_csv(output_filename, sep=',', compression='gzip', index=False)
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
powershell
PS D:\New folder\week 6\2GB>
OUTLINE
TIMELINE
Activate Windows
Go to Settings to activate Windows.
Ln 1, Col 1 Spaces: 4 UTF-8 CRLF Python 3.11.1 64-bit
Live Share
Type here to search
ENG
US
21:10
12/06/2023