

Week 6: File ingestion and schema validation

Name: Canmert Demir – File ingestion and schema validation

Submission date: 05.11.2023

Internship Batch: 30 Sept to 30 Dec 2023

Version: 3.11.6.final.0

Data intake by: <https://www.kaggle.com/datasets/devendra416/ddos-datasets>

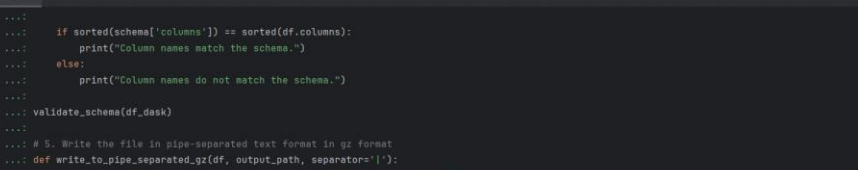
Data intake reviewer: Canmert Demir

Submitted location: <https://github.com/Canmertdemir/Canmertdemir-DataGlacierWeek6>

In this project, I worked on a Kaggle dataset called "DDoS Datasets," which provides data on internet interruptions caused by hacking. The dataset consists of two parts. Firstly, I read the data in three different ways and compared the efficiency of the libraries I used for data reading, namely pandas, Modin.pandas, Dask, and Ray. While using pandas, I encountered a memory error. Subsequently, I attempted to use Modin.pandas, but it failed to read the data. Finally, I turned to Dask, and successfully read the data in Python. I combined two different CSV documents, creating a 3 GB large dataset. Afterward, I cleaned spaces and character issues from the dataset. I addressed the challenge of a lack of recent and exclusively available DDoS datasets in the public domain. To overcome this limitation, I took the initiative to extract DDoS flows from various public Intrusion Detection System (IDS) datasets, specifically the CSE-CIC-IDS2018-AWS, CICIDS2017, and CIC DoS dataset(2016). These datasets were produced in different years and utilized different experimental DDoS traffic generation tools, introducing more variability into the collected DDoS data.

To create a comprehensive dataset for my analysis, I combined the extracted DDoS flows with "Benign" flows, which were also separately extracted from the same base dataset.

In the next step, I defined a code block that generates a YAML file in Python. Following that, I created a validation schema code and an output file in GZ format. Below are some screenshots from my work.



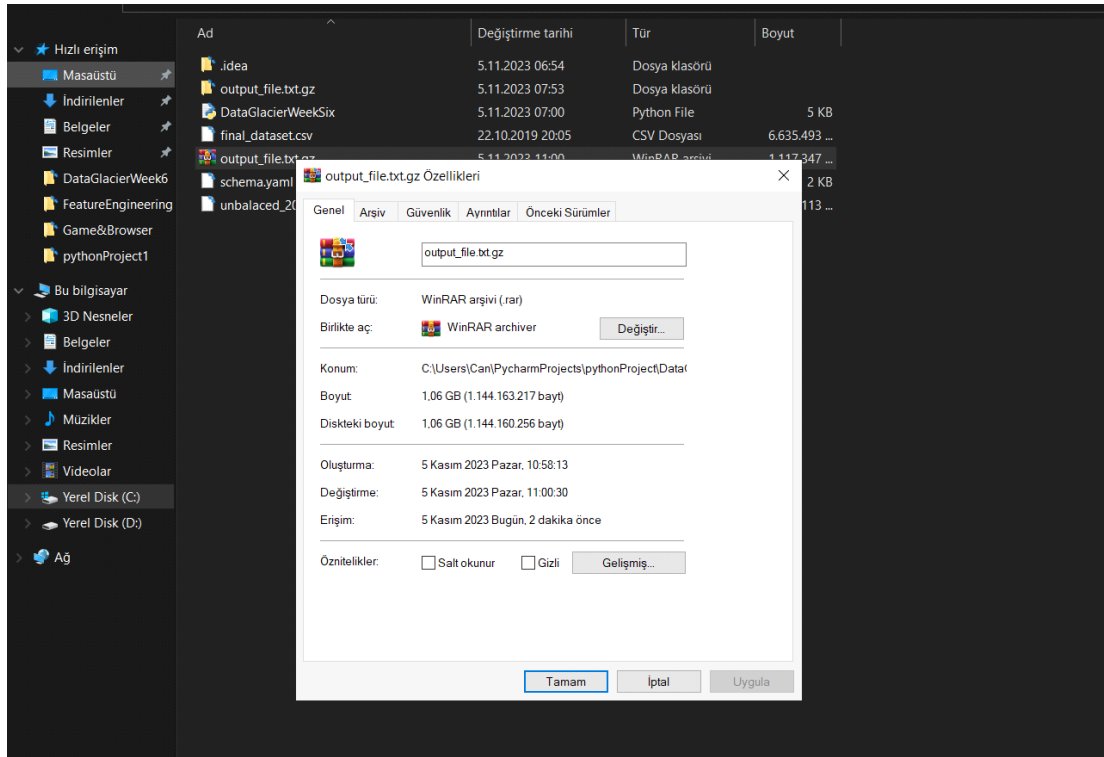
The screenshot displays a Jupyter Notebook environment with a dark theme. The 'Python Console' tab is active, showing a Python script that performs the following actions:

- Validates the schema of a DataFrame against a list of column names.
- Writes the DataFrame to a pipe-separated CSV file in gzipped format.
- Prints summary statistics: total number of rows, total number of columns, and file size in MB.

The output of the script is displayed below the code:

```
Column names match the schema.
Total number of rows: 7616509
Total number of columns: 85
File size: 0.6390625 MB
```

The bottom status bar indicates the current file is 'DataGlacierWeek6.py' and shows various editor settings like 'UTF-8' encoding and '4 spaces' for indentation.



Here is some information about the dataset required from Data Glacier:

-The DDoS dataset is divided into two parts, with a total size of 3 GB. Also data set is available at

CSE-CIC-IDS2018-AWS: <https://www.unb.ca/cic/datasets/ids-2017.html>

CICIDS2017: <https://www.unb.ca/cic/datasets/ids-2018.html>

CIC DoS dataset(2016) : <https://www.unb.ca/cic/datasets/dos-dataset.html>

- The column names in the dataset match the schema.

- The total number of rows in the dataset is 7,616,509.

- The dataset consists of a total of 85 columns.

- The file size of the dataset is 0.0390625 MB.(output_file.txt.gz)