Data Glacier

Your Deep Learning Partner

File ingestion and schema validation

<15/08/2022>

File

- Type: csv
- File type: csv
- Dataset name: test_file
- File name: animelist
- Table name: edsurv
- Inbound delimiter: ",“
- Outbound delimeter: "|“
- Number of columns: 11

Validation

WE CREATED A YAML FILE TO STORE OUR TARGET ATTRIBUTES TO MAKE SURE THAT NO DATA IS MISSED IN THE READING PROCESS.

WE HAVE ALSO CREATED A PIPELINE TO PREPROCESS THE DATA BY REMOVING CAPITAL LETTERS AND UNWANTED SYMBOLS

FINALLY IN OUR PIPELINE WE CHECK TO SEE IF OUR READ FILE MATCHES THE ATTRIBUTES IN THE YAML FILE.

Importing the file

- We have completed this project on a windows device therefore we were only able to use pandas and dask dataframe to read the file.
- Using pandas we can have a CPU runtime of 1 minute 39 seconds and using dask dataframe we have a CPU runtime of 50.2 seconds.
- After importing the required files we then exported the dataframe using pandas converting the file into a .gz file.

Thank You

Data Glacier

Your Deep Learning Partner