ETL with Python

Erickson Figueroa

Executive Summary: This project focuses on automating the ETL process for extracting, transforming, and loading car price data from diverse formats (JSON, XML, CSV) available at a specified URL. The process involves:

- Downloading a zip file.
- Extracting its contents into a staging area.
- Consolidating data into a unified Pandas Data Frame.

I've implemented robust logging mechanisms to track the ETL process, including start times, key events, and any encountered errors. I also performed data transformations, such as adding a new column ('modified_price') and saving the transformed data to a destination folder. The comprehensive logging ensures transparency, allowing easy identification of successes and potential issues in each step of the ETL process. This automated solution streamlines data management and enhances traceability for efficient decision-making.

Objectives:

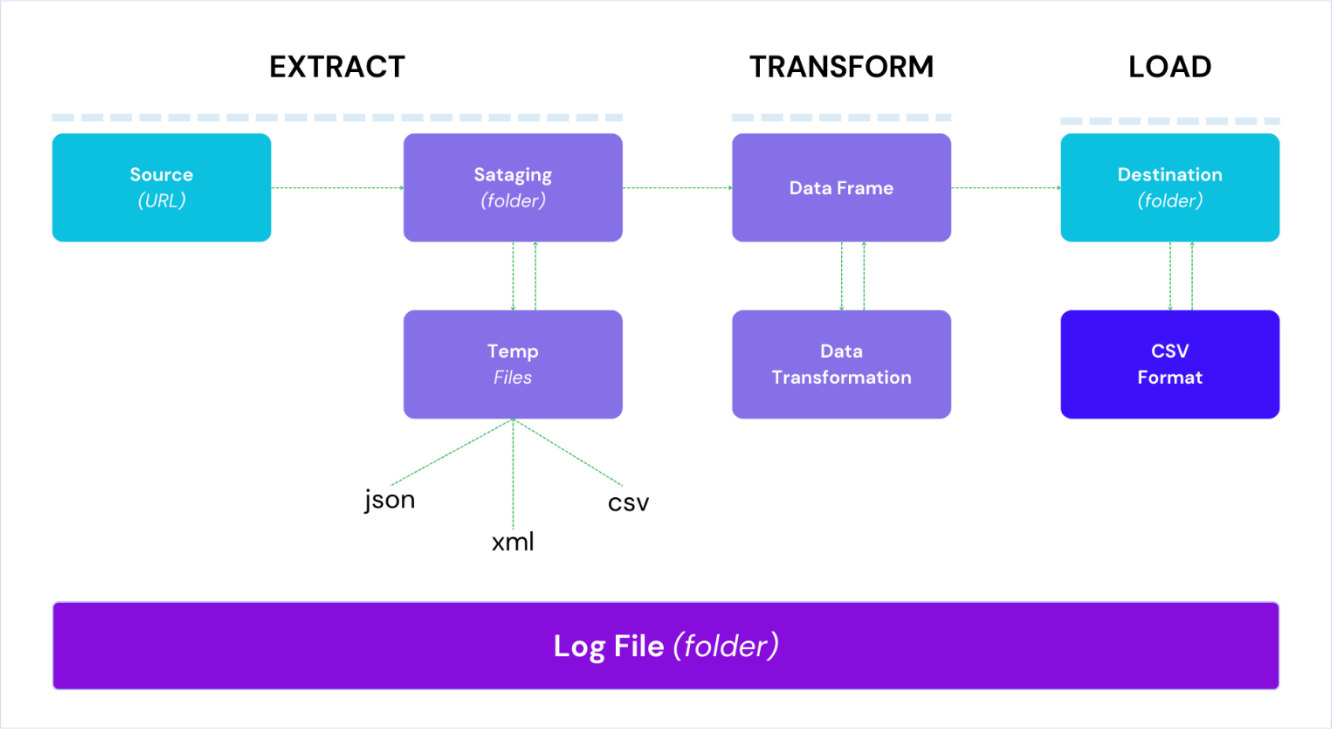
- **Download Data:** Download a zip file from a specified URL containing various file formats (JSON, XML, CSV) related to car prices.
- **Extract Data:** Extract the content of the downloaded zip file into a staging folder.
- **Logging:** Log critical events and errors during the ETL process, including the start and completion times.
- **Consolidate Data:** Read data from different file formats (JSON, XML, CSV) and consolidate it into a Pandas Data Frame.
- **Data Transformation:** Add a new column, modified_price, to the Data Frame by converting and manipulating the existing price column.

- **Save Transformed Data:** Save the transformed Data Frame to a CSV file in a destination folder.
- **Logging Results:** Log the success or failure of each step in the ETL process.

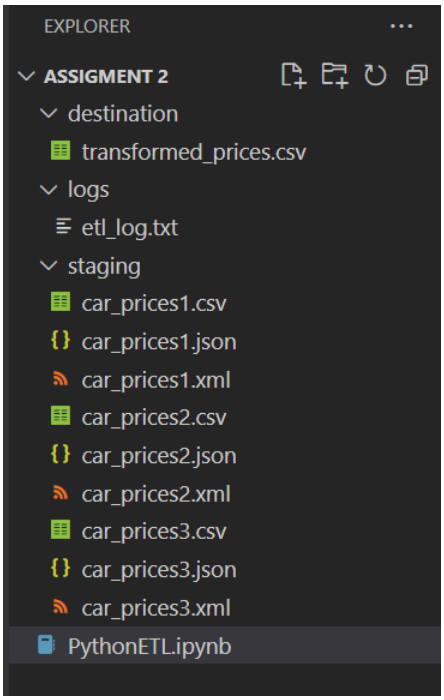
Brief code functionality:

- **Logging Setup:** Configures logging to record events and errors in an **etl_log.txt** file.
- **Datetime Function:** Defines a function **get_current_datetime()** to get the current date and time with milliseconds.
- **Download & Extraction:** Downloads a zip file from the provided URL, extracts its content to a staging folder, and logs success or failure.
- **Consolidation:** Reads data from JSON, XML, and CSV files, consolidates it into a Data Frame, and logs any unsupported file formats.
- **Data Transformation:** Adds a new column **modified_price** to the Data Frame based on the price column, doubling, and rounding the value.
- **Save Transformed Data:** Saves the transformed Data Frame to a **CSV** file in the destination folder and logs the path.
- **Final Logging:** Logs the completion of the ETL process, indicating rows loaded and success or failure.

Folder structure and ETL process



Folders:



Log file content:

```
[INFO]: ETL process started at: 2024-01-14 01:09:11.500
[INFO]: Zip file downloaded and extracted successfully to the path:'./staging/' 2024-01-14 01:09:12.209
[INFO]: Transformed data saved to: './destination/transformed_prices.csv' 2024-01-14 01:16:17.292
[INFO]: ETL process completed successfully. 10 rows were loaded. 2024-01-14 01:16:17.292
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:17:00.053
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:17:05.653
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:17:10.854
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:20:30.393
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:20:35.677
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:20:39.174
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:20:50.873
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:22:15.039
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:23:16.592
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:24:18.012
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:24:21.763
[INFO]: ETL process started at: 2024-01-14 01:24:24.928
[INFO]: Zip file downloaded and extracted successfully to the path:'./staging/' 2024-01-14 01:24:25.626
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:24:27.770
[INFO]: ETL process started at: 2024-01-14 01:25:13.290
[INFO]: Zip file downloaded and extracted successfully to the path:'./staging/' 2024-01-14 01:25:13.991
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:25:17.139
[INFO]: The consolidated dataframe was created successfully. 2024-01-14 01:25:29.423
[INFO]: Transformed data saved to: './destination/transformed_prices.csv' 2024-01-14 01:25:36.156
[INFO]: ETL process completed successfully. 90 rows were loaded. 2024-01-14 01:25:36.156
[ERROR]: An error occurred: [Errno 13] Permission denied: './destination/transformed_prices.csv' 2024-01-14 01:26:38.514
[ERROR]: An error occurred: [Errno 13] Permission denied: './destination/transformed_prices.csv' 2024-01-14 01:27:25.990
[INFO]: Transformed data saved to: './destination/transformed_prices.csv' 2024-01-14 01:27:51.009
[INFO]: ETL process completed successfully. 90 rows were loaded. 2024-01-14 01:27:51.009
```

Source Code