AlpacaTech Take-home Exercise

This assignment contains 2 parts `Data Provider Selection` and `Data ETL`. There is no time limit for this assignment, but for reference, we found out that it usually takes people between 1 hours 30 minutes and 2 hours 30 minutes to complete both tasks.

Note: If you use AI to complete the assignment, please include all AI prompts and the corresponding responses as part of your submission.

Your submission should be separated and stored into:

- `${WORKDIR}/AT_take_home_exercise/data_provider_selection` and
- `${WORKDIR}/AT_take_home_exercise/data_etl`

Any other files and folders initially provided such as `AT_take_home_exercise.pdf` and `tick_data/**` can be removed in your submission.

Data Provider Selection

You are provided with three FOREX tick datasets each from a different data provider: A, B and C, located at `AT_take_home_exercise/tick_data/date=2024-03-01`. Each dataset is provided in a CSV format with the following columns:

- datetime: Date and time of the data point
- currency_pair: Identifier for the currency pair in the format of XXXYYY (e.g., "USDJPY")
- bid: Bid price of the currency pair at the given timestamp
- ask: Ask price of the currency pair at the given timestamp
- volume: The trading volume during the time interval

Please perform data analysis on the sample datasets, and select the most appropriate data provider.

Deliverables expected:

a. **Data investigation report :** A detailed data analysis report in the form of Jupyter notebook (**`analysis.ipynb`**) and HTML output (**`analysis.html`**), with a conclusion indicating and justifying the chosen data provider with reason.

b. **[Optional] AI Usage Disclosure**: AI prompts and corresponding responses, if AI tools were used in completing this task.

Please leave your deliverables in the following folder: `${WORKING_DIR}/AT_take_home_exercise/data_provider_selection`.

Data ETL

Input

By contract, on every weekday, the data provider should deliver a CSV file containing full historical FOREX tick data up to 9am JST of the same day via a private network connection. For the purpose of this assignment, you may assume that the file is made available on your local machine at the following path:
`${WORKDIR}/AT_take_home_exercise/tick_data/date=YYYY-MM-DD/sample_fx_data_{A|B|C}.csv.gz`

Output

Using sample data from the **selected** provider, build a production-quality process that converts secondly tick data into **production-ready 1-minute OHLC data for dates starting from 2024**. The output should be stored in your local machine at the following path:

`${WORKDIR}/AT_take_home_exercise/data_etl/ohlc_data/date=YYYY-MM-DD/data.csv.gz`

The resulting dataset will be consumed by a live model in a production environment, which generates predictions once every weekday, one hour after the OHLC data becomes available. If the latest data for today is missing, the model will fail noisily. Not making any predictions at all is preferable to making wrong predictions with stale/incorrect data.

The model expects the following attributes at minimum:

- timestamp: Date and time of the data point
- currency_pair: Identifier for the currency pair in the format of XXX/YYY (e.g., "USD/JPY")
- open: Opening mid price of the currency pair at the given timestamp
- high: Highest mid price of the currency pair during the time interval
- low: Lowest mid price of the currency pair during the time interval
- close: Closing mid price of the currency pair at the given timestamp

* Note: mid price = (bid + ask) / 2

Deliverables expected:

a. **Python Program / Scripts**: A python program that contains methods that perform the required ETL process and writes the results to an output file.

b. **Output File**: A readable output file in CSV format or compressed GZIP format containing the desired output.

c. **Readme**: Documentation with minimum instructions for running the python program. State any assumptions and limitations if necessary.

d. **[Optional] AI Usage Disclosure**: AI prompts and corresponding responses, if AI tools were used in completing this task.

Please leave your deliverables in the following folder:
`${WORKDIR}/AT_take_home_exercise/data_etl`.