Data Ops Engineer Internship - Problem Set

Advanced Analytics (AA)
Axpo Solutions AG

May 27, 2024

Part One - ETL Pipeline

We aim to extract, transform, and load data from the Spanish power market. This data is accessible via the public API of ESIOS.

To send valid data requests to the ESIOS API, use the following token. If you encounter issues requesting data from the API, please let us know.

Token: `ca757527cb8381ad315cd72b02a0176f8842fa5b548d99e14f4de46f61bcb17a`

Extract

Load the following indicators from 2020-01-01 to 2021-12-31:

- Power demand forecast (indicator 460)
- Power demand real (indicator 1293)
- Nuclear power availability (indicator 474)

Refer to the IDs in parentheses to request the specific data from the ESIOS API.

Transform

Transform these three time series into a pandas DataFrame with hourly granularity and the following column structure:

- "**datetime**", format: "yyyy-mm-dd HH:MM:SS"
- "**demand forecast**": values of the time series for id 460
- "**demand real**": values of the time series for id 1293
- "**nuclear power availability**": values of the time series for id 474
- "**error demand forecast**": the error between the forecast of the demand and actual demand

- "**error demand forecast avg 24H**": average of the error between forecast demand and actual demand for the last 24 hours
- "**error demand forecast avg 12H**": average of the error between forecast demand and actual demand for the last 12 hours
- "**demand real lag 1D**": the value of the real demand for the same hour on the previous day
- "**demand forecast lag 1D**": the value of the forecast demand for the same hour in the previous day
- "**nuclear power availability avg 24H**": average of the nuclear power availability for the last 24 hours
- "**nuclear power availability avg 12H**": average of the nuclear power availability for the last 12 hours

Save the DataFrame as a csv file (`yourname_etl.csv`).

Part Two - Platform Design

In this section, we would like you to design a scalable data and analytics Platform. This Platform is meant to serve Data Scientists and Quantitative Analysts in different departments within the organization and enables them to develop and orchestrate simple models in the domain of energy trading.

However, in order to reduce the complexity of this exercise we want you to only design some parts of the Platform, **namely the data ingestion and storage components as well as the data access layer**, considering the following requirements:

- The platform should support batch data ingestion methods and handle various types of data formats from multiple internal and external data sources such as databases, APIs, or file systems.
- Define a storage solution capable of handling these different data formats and scaling efficiently with growing data volumes.
- The platform should enable user-based data access in a highly standardized and quality-assured way, allowing users to interact efficiently with the ingested data (assuming all users have a basic understanding of Python programming and SQL syntax).
- The designed platform should be highly available and maintainable, with continuous service monitoring and automated alerting capabilities.
- The platform should scale efficiently to accommodate a growing user base within the organization.

Describe the overall design, the selected technologies and how they satisfy the requirements stated above. You can include images to better communicate any design decisions made such as an architectural diagram of your solution showcasing the individual components and their interactions within this data and analytics platform.

There is no need to write any code for this exercise.

Submitting

Please hand in all your solution files in a zipped folder named as **firstname_lastname** by email. Many thanks for taking the time to go through this assessment and please get back to us if you have any questions regarding the problem set.