# ML Ops Engineer - Problem Set

Axpo Iberia
Advanced Analytics

March 8, 2021

This problem set consists of 3 differentiated parts that assess your coding skills as well as your experience / background in the area of ML Ops Engineering.

## Part One - ETL Pipeline

We want to extract, transform and load some data of the Spanish power market. This data can be accessed via the public API of ESIOS (more information and documentation can be found here: `https://api.esios.ree.es/`). Please create a python script (`your_name_etl.py`) that contains your code as well as all imports needed to run your file.

In order to send data requests to the ESIOS API, use the following access Token:
`Token = "0a3ec68a89579911bd7610e20df0f79c83f79b10275b19025aaf414136c91f9a"`

### Extract

We would like to load the hourly power demand forecast (Previsión diaria de la demanda eléctrica peninsular) from 2017-01-01 to 2018-12-31. Please refer to the Spanish term in parentheses to look up the id needed to request that particular data from the ESIOS API.

### Transform

Transform the loaded data into a pandas DataFrame and adjust it accordingly in order to match the following column structure:

- ”name” of the time series
- ”id” of the time series
- ”datetime”, format: ”yyyy-mm-dd HH:MM:SS”
- ”demand_forecast” values of the time series
- ”update_timestamp” values updated at, format: ”yyyy-mm-dd HH:MM:SS”

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

### Load

Lastly, we would like to load the data into a SQL Server database. Hence, please build a generalized python connector that could connect to a particular SQL DB and load the previously transformed DataFrame into an already existing table called `demand_forecast_esios`. In order to mock the connection to the SQL DB you can refer to those details:

- server = ”server.database.com”
- database = ”mock_db”
- user = ”user1”
- password = ”123abc”

## Part Two - Code Optimization

In this section we want to evaluate your ability to adjust and optimize some python code. Please create a python script (`your_name_optimization.py`) that contains your code as well as all imports needed to run your file.

### Running Time Optimization

In the following code snippet a function named `function_to_apply` is applied to all the rows in a DataFrame and creates a new column called `distance`. Please try to optimize the code to run as fast as possible. The dataset to be used is called `dataset_speed_optimization.csv`.

```
import pandas as pd
import numpy as np

# Function to apply to all rows in the DataFrame
def function_to_apply(lat, lon):
  a = np.sin(lat/2)**2 + np.cos(lat) * np.cos(lon) * np.sin(lon/2)**2
  return a

# Add new column to the DataFrame
list_results = []
for i in range(0, len(df)):
    r = function_to_apply(df.iloc[i]['latitude'], df.iloc[i]['longitude'])
    list_results.append(r)
df['distance'] = list_results
```

### Memory Optimization

Load the dataset called `dataset_memory_optimization.csv` in memory and try to reduce the memory taken up as much as possible.

## Part Three - Open Questions

Finally, we would like to assess your experience in ML Ops specific problems / situations. Please write distinctive answers to the following questions and clearly state your assumptions. You can write your answers into a plain pdf file called (`your_name_questions.pdf`) that should not extend more than 2 pages.

### Question 1

Let's assume you are responsible for monitoring a simple machine learning model, which is running on a productive cloud cluster. The model is scheduled to run every hour and predicts the average power price for the following day.

- Which would be the most important KPI's to monitor that model?
- What part of the monitoring could be automated in order to have the model running in operation 24/7, and how would you do that?

### Question 2

Imagine that you are responsible to monitor a set of ETL pipelines, which load structured time series data from external sources into the cloud. All of this data is needed as input for ML models, that are running in production.

- What would you do if one of those ETL pipelines starts to fail?
- What measures could be implemented to make those ETL pipelines as robust as possible?
- Which characteristics of the different data sets could be used to evaluate the particular data quality?

## Submitting

Please hand in all your solution files by email. Many thanks for taking the time to go through this assessment, and please get back to us if you have any question regarding the problem set.