Data Engineer Take Home Test

Objective

Develop a pipeline to process air quality data from OpenAQ and store the results in a database.

Description

To evaluate the quality of living, we have been approached by some cities to analyze the air quality. To do this, we can rely on a subset of the open dataset provided by OpenAQ. You are given a few datasets in *ndjson* format which we expect you to process. You can find them here https://drive.google.com/file/d/1bH6BM7hrVI9ufuJ5GVGE7QPEwIJAM1xX/view?usp=sharing

We would like to monitor the 24hr rolling average of the following measures for each city:

1. PM2.5
2. PM10
3. O3
4. NO2
5. CO

We are also interested in the Air Quality Index (AQI)values of the $PM_{2.5}$ AQI and PM_{10} AQI which are defined on the 24hr average of $PM_{2.5}$ and PM_{10} respectively (see formula in annex). We are only interested in the cities from the countries provided in **countries.csv**.

You are tasked with developing a data pipeline that takes in data from OpenAQ, processes it and eventually stores the resulting data in a database of your choice for various use cases. The database should contain one or more tables with the hourly value of the metrics mentioned above (24-hour rolling average of $PM_{2.5.}$, PM_{10}, O3 Ozone, NO2, and CO; And AQI $PM_{2.5}$, and AQI PM_{10}) for each city. You are free in how you approach this task regarding programming language, frameworks, platform, databases, etc. We expect you to be able to explain your choices either in a few lines or in a follow-up conversation.

Deliverables

We expect you to come back to us in 48hs with your solution providing:

1. Source code
2. Dump of resulting database or credentials to access it if it is online
3. Any documentation you believe is necessary to understand the solution

The challenge is designed to be completed in no more than 4 hours. Feel free to provide a list of hypotheses and assumptions, and future improvements that could be implemented if more time were available. If you have any blocking questions, please send an email to {CHALLENGE_EMAIL}@pythonpredictions.com

Annex

PM2.5 AQI

AQI Category	AQI Value AQI Low - AQI High	24hr Average PM2.5 Concentration (µg/m³) Breakpoint Low – Breakpoint High
Good	0 - 50	0 - 12.0
Moderate	51 - 100	12.1 - 35.4
Unhealthy for Sensitive Groups	101 - 150	35.5 - 55.4
Unhealthy	151 -200	55.5 - 150.4
Very Unhealthy	201 - 300	150.5 - 250.4
Hazardous	301 - 400	250.5 - 350.4
Hazardous	401 - 500	350.5 - 500.4
Hazardous	501 - 999	500.5 - 99999.9

PM10 AQI

AQI Category	AQI Value AQI Low – AQI High	24hr Average PM10 Concentration (µg/m³) Breakpoint Low - Breakpoint High
Good	0 - 50	0 - 54.0
Moderate	51 - 100	55.0 - 154.0
Unhealthy for Sensitive Groups	101 - 150	155.0 - 254.0
Unhealthy	151 -200	255.0 - 354.0
Very Unhealthy	201 - 300	355.0 - 424.0
Hazardous	301 - 400	425.0 - 504.0
Hazardous	401 - 500	505.0 - 604.0
Hazardous	501 - 999	604.0 - 99999.9

AQI formula[1]:

$$AQI = \frac{(AQIHigh - AQILow)}{(BreakpointHigh - BreakpointLow)} \times (PMobs - BreakpointLow) + AQILow$$

PM_{obs} = observed 24-hour average concentration in µg/m3
BreakPointHigh = maximum concentration of AQI category that contains PM_{obs}
BreakPointLow= minimum concentration of AQI category that contains PM_{obs}
AQIHigh = maximum AQI value for category that corresponds to PM_{obs}
AQILow = minimum AQI value for category that corresponds to PM_{obs}
For example, for a PM2.5 24hr average of 21.4 µg/m³:

$$AQI = \frac{(100 - 51)}{(35.4 - 12.1)} \times (21.4 - 12.1) + 51 \; = 70.558$$

[1] https://en.wikipedia.org/wiki/Air_quality_index#Computing_the_AQI