Practical exercise

Introduction

Please implement a solution to any number of the proposed questions below. Feel free to present additional insights on the dataset, if you find any. Your implementation, code, and results will be discussed as part of your interview, so make sure to send back all the necessary material or link to your repository to allow the interviewers to understand your approach.

 # 1st part -Ingest data into a relational database from JSON files

The goal of the exercise is to read a dataset in JSON format and insert the data into a RDBMS of your choice. The exercise can be solved using any kind of program or script of your choice that does the job.

The dataset you are going to use for the exercise is:
https://catalog.data.gov/dataset/air-quality-measures-on-the-national-environmental-health-tracking-network

On the website, you can find useful information about the dataset. The JSON file also contains metadata describing the dataset.

Some constraints:

1. The solution has to be as automated as possible in all the steps, including data gathering, transformation, ingestion, and result presentation.
2. The solution has to be reproducible
3. The solution has to be prepared to deal with incomplete/wrong/missing data (e.g. discard incomplete rows).
4. The database can be any open-source relational database of your choice.

 # 2nd part -Answer some questions using SQL

We already have the data into our RDBMS and now we want to generate some SQL queries to answer the following questions about the dataset:

1. Sum value of "Number of days with maximum 8-hour average ozone concentration over the National Ambient Air Quality Standard" per year
2. Year with max value of "Number of days with maximum 8-hour average ozone concentration over the National Ambient Air Quality Standard" from year 2008 and later (inclusive)
3. Max value of each measurement per state
4. Average value of "Number of person-days with PM2.5 over the National Ambient Air Quality Standard (monitor and modeled data)" per year and state in ascending order
5. State with the max accumulated value of "Number of days with maximum 8-hour average ozone concentration over the National Ambient Air Quality Standard" overall years
6. Average value of "Number of person-days with maximum 8-hour average ozone concentration over the National Ambient Air Quality Standard" in the state of Florida

7. County with min "Number of days with maximum 8-hour average ozone concentration over the National Ambient Air Quality Standard" per state per year

Infrastructure design

This is a design exercise.
As one of the first data architects at the Data:Lab you will have a large impact on the design and implementation of our data infrastructure. We plan to ingest data from numerous sources, and we want to make this data available to customers, to our in-house engineers, and to our data scientists.
We would like you to write a design proposal for how you might design and implement a data infrastructure for storing and representing the following data sources, given the desired use cases. You do not need to describe how the use cases are implemented - just describe how you might get the data to the consumer in a reasonable way.

Sources:
two data sources
data source A delivers data every day, with a 3-day delay
data source B delivers data every 14 days

- both sources deliver data via HTTP or FTP as a direct file download
- both sources use different formats, so data model from A cannot be used to denote data from source B
- one weather source
- weather source C delivers either the current weather or a 7-day forecast
- this source delivers data via a JSON REST API
- the data can be retrieved using two endpoints - one for the current weather and one for the forecast
- the forecast endpoint takes two parameters: latitude and longitude
- the current endpoint takes three parameters: latitude, longitude, and optionally a timestamp one market data source
- data source D delivers near real-time market data, available every 15 seconds
- this source delivers data via a JSON API for streaming (e.g. like the Twitter streaming API) each data set includes a coordinate with a latitude and longitude, where each coordinate represents a 50-mile market radius("region").

Use cases:
using a REST API, I want to get the average weather, average market interest, and accompanying data for a region

I want to get all daily weather data and daily market prices of the past 15 years
I want to get data for a single region of the past 15 years, run analyses on it, and store the results for consumption by a REST API

Bonus challenge:
as a customer, I want to upload proprietary data Z, and then run custom analyses on data Z and the data from data sources A and B,
and access it through the REST API

Submission

Please email us a document detailing your proposal as a README.md or a PDF. You might consider looking at some design proposal templates for ideas on how to structure your proposal.