Data Engineer - Code Challenge

Applicants for the Data Engineer role at Wave are asked to work through the following code challenge and submit a zip file (or a Github link) prior to the team interview.

This code challenge is a series of exercises that is common to the role of being a Data Engineer at Wave. The challenges are designed to tackle the entire 'data supply chain' from infrastructure, ingestion, data processing and output using different tools and libraries.

Answer as many exercises you can; spend no more than 3 hrs on the entire code challenge. This will be a springboard for conversation through the interview process to understand your proficiency in technical concepts and your approach to problem solving that is critical to the success in the role.

Code documentations including recommendations or optimizations that you wish to add and discuss in the interview can be included in the readme file.

Requirements per Exercise

- Python 3.8+
- Provide a configuration file per exercise (can be either JSON or YML). This will serve as the input for your program. A sample configuration file is included in this folder
- Any Python libraries that you wish to use
- Requirements.txt file

Data Source

1. Download Station Inventory CSV from here
2. Read guidelines for getting Weather data here

Exercise 1 - Extraction, Transformation, Load

A significant portion of the Data Engineering role is to develop end to end data pipelines. This exercise is designed to evaluate the proficiency in coding of a standard data process in retrieving, transforming and loading data.

Resources:

- Station Inventory File
- Guidelines for downloading the Weather data

Submit your code for the following tasks:

Data Retrieval
Read data from the Station Inventory file. Using this data as reference and the Weather data URL, download the Weather data for **Toronto, Ontario** for the **last 3 years** from input **year**.

Cleanup & Merge
Remove all unnecessary data such as future dates without temperature data.
Join Station Inventory and Weather data using the Climate ID

Merge
Generate an Excel file that contains all data retrieved. Data should be divided into years and stored in their respective tabs/sheets

Output: Excel file

Exercise 2 - Computation

The role of a Data Engineer requires analyzing the data by running calculations and computations that provide us insights into how to best design our systems.

Submit your code for the following tasks:

Data Retrieval
Input data: year

Read data from the Station Inventory file. Using this data as reference and the Weather data URL, download the Weather data for **Toronto, Ontario** for the input **year**.

Computation
Compute the following:

1. Max Temperature for **year**

2. Minimum Temperature for **year**
3. Average Temperature per month for **year**
4. Average Temperature overall for **year**
5. Number of days in **year** and **previous year** where temperature is
    a. Equal
    b. Within 1 degree

Output: Command line print

Commented [1]: equal to what?
or do you mean for each temp value, provide the number of days where temp equal to that number?

Commented [2]: provide the number of days where the temperature for year and previous year are equal

Commented [3]: comparing Feb 20th, current year to Feb 20th, previous year for example?

Exercise 3 - Packaging

The Data Operations functions have many similarities to the standard application engineering practices. As a Data Engineer, you are expected to understand components of the application development process and apply these concepts to building data systems.

Submit your code for the following tasks:

Create a Package
Create a Docker file with the following requirements:

- Python 3.8+
- OS: Linux
- Include requirements.txt file
- Should be able to run previous exercises

Output: **Docker file**

Exercise 4 - Infrastructure

The Data Operations team helps set up tools and technology to enable the organizational use of data. As a Data Engineer, you are expected to understand how to use and set up the infrastructure that stores and processes data.

Resources

- Getting started with AWS: https://aws.amazon.com/getting-started/
- DynamoDB Sample Code: https://github.com/aws-samples/aws-dynamodb-examples/tree/master/DynamoDB-SDK-Examples/python
- Lambda Sample Code: https://docs.aws.amazon.com/code-samples/latest/catalog/code-catalog-python-example_code-lambda-boto_client_examples.html

Submit your code for the following tasks:

Data Retrieval
Input:

1. **Year**
2. **S3 Bucket and Key**
3. **Data Range**

Task
Create a Lambda function for each task:

1. Ingest the Weather data for **year** to DynamoDB
    a. Creation of DynamoDB is codified
    b. Checking if DynamoDB exists before creation
2. Pull Weather data from DynamoDB with specified **date range** and write it to an Excel file. Upload this file to the specified **S3 Bucket and Key**

Output:

1. **DynamoDB Table**
2. **Excel File (.xlsx)**

Exercise 5 - Serverless

We find ways to leverage the latest in engineering technology and design principles to drive innovation and scale. The Serverless Framework is one of those principles we apply to our Data Operations practice and as a member of the team you will be applying this to Data engineering tasks.

Resources:

- Getting Started: https://www.serverless.com/framework/docs/getting-started/
- Examples: https://www.serverless.com/examples/ (filter to AWS and Python)

Submit your code for the following tasks:

Input:

1. **Year**
2. **S3 Bucket and Key**
3. **Data Range**

Task: Convert Exercise 4 to Serverless Framework (serverless.com)

Output:

1. **DynamoDB Table**
2. **Excel File (.xlsx)**

Commented [4]: how do they provide that? will they be given access to an AWS account or expected to give access to their AWS account?

Commented [5]: _Marked as resolved_

Commented [6]: _Re-opened_

Commented [7]: good point. since we asked to have the creation of dynamo db codified, we can just ask for a screenshot of the dynamo db table maybe? @twongkee@waveapps.com thoughts?

Commented [8]: Maybe have them provide a DynamoDB export: https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/DataExport.html ?