Full Stack Data Engineer Assessment

RkxBR3s3Mjc0ZTVkYy1hYjYwLTRmYzItOGY1MS1lZTZlY2ViNmJlZjd9

Date: 2025-07-28

The following exercises are designed to be a quick assessment of your affinity with handling data and general-purpose software/scripting tooling. This is an engineering assessment, and we are interested in your code (a zipped git repository) which should include a summary write-up (`readme.md` file). Please do not send us presentations or Word documents.

We recommend sending your submission within one week of receiving this, and spending no more than one afternoon. It is okay if you do not complete all exercises in this time. It is up to you on which exercises you spend your time. The exercises can be completed in any order. The choice of tooling to conclude any of the exercises is up to you. Some pointers that might influence your choices:

- Favor fast/pragmatic results over rigid engineering
- How is your solution impacted by any form of automation?
- What is the "resistance to change" of your solution? (e.g. new requirements, other insights, etc.)

Note that some assignments are intentionally vague. You're free to make any assumptions you think are necessary, but be prepared to defend the choices you made. Also note that there are no requirements regards tools or techniques, use whatever you think will be a good fit within our team. We would like to be able to run your code from our end, so please prepare your submission with that in mind.

During the assignment evaluation, you might be asked to make small adjustments or make additional data queries; it would help if you have some environment wherin those changes can be made and evaluated quickly.

Consider the following 3 datasets. These are all used in some way in subsequent exercises:

- airports.csv
- countries.csv
- runways.csv

To make the assignment *even more enjoyable*, 7 flags have been hidden throughout this assessment. Can you find all of them? A flag looks like this: `FLAG{ uuid }`.

1. Acquire this data and make it suitable for insights

One of your colleagues from the business development department is analyzing a possible market fit for an innovative new product idea. He'd like to derive some insights from the supplied datasets.

Answer the following questions, providing the code that you used to do so:

- Which are the top 3, and bottom 10, countries by number of airports (with counts)?
- For every country that has an airpoirt, list the width and length of the longest runway (and name of the airpoirt with that runway).

2. Upload airports dataset to cloud storage

For loading the data into our state-of-the-art Data Warehouse, the data should be supplied, well-formatted, in a block-storage service.

We request a dataset containing at least the columns one of the resultsets from (1) to be supplied to an Azure Blob Storage service (details below, format of choice). If you have no result sets, use some attributes of choice from the airports.csv set.

Details:

```
Azure Storage Account Name           : sacodeassessment
Azure Storage Account Shared Signature: sv=2022-11-02&ss=b&srt=o&sp=wc&se=2034-11-
11T11:00:00Z&st=2024-11-
10T23:00:00Z&spr=https&sig=D%2BgRbWPJDTmsbPtyfTEiTnb7gg594uNsvm62oQK49Yg%3D
```

```
Blob Storage Container                : results
Target blob path                      : /ingest-assessment-{date:yyyyMMdd}-{your
initials}
```

3. Acquire data from vendor API

Our supplier of high quality, curated country data no longer provides access to a file-based dataset (csv). Instead, they request you to gather this data through their next-generation data broker platform, in exchange for a detailed overview of revenues generated from this data.

Gather country data for every country listed in the airports dataset and store this in a single file. Next, upload an empty file named 'revenues.txt' to the upload endpoint (HTTP POST).

Which countries information is missing from the broker platform?

Details:

```
client id                   : abc123
host address                : code001.ecsbdp.com
HTTP path (get country info): /countries/{iso_code}
HTTP path (upload data)     : /revenues?client={client_id}
```

4. Navigate eCommerce RDBMS

As we're diversifying into digital media, an eCommerce ERP system (Chinook) has been build to handle all inventory tracking and sales for this new retail market.

Larry, a colleague marketeer, pitches an idea involving cloud hyper-scale machine learning based customer segmentation for targeted advertising. He assumes more revenue can be generated from 'more sophisticated' customer - those who enjoy jazz music.

Prove, or disprove, Larry's hypothesis by showing the average of the total value of all sales for a customer who has purchased any jazz music track versus the average value of all sales for customers who have never bought any jazz. Use Larry's credentials to access the chinook system.

Details:

```
PostgreSQL Server v15
hostname: code001.ecsbdp.com
port    : 5432
database: chinook
username: larry
password: iddqd
note    : when too many failed connection attemps are made, your client IP address
will be blocked for 1 hour!
```

Also, Larry confides in you and reveals his next killer-app prototype. It's an app where users can enter any track name and all albums that include that track are listed in the app. Unfortunately Larry is struggling with the performance of a query he needs to run. Can you give Larry or the database administrators a suggestion beneficial to this use case?

```sql
SELECT album.title
FROM   track
JOIN   album ON (track.album_id = album.album_id)
WHERE  LOWER(track.name) = LOWER('Enter Sandman')
```

5. HTTP service integration

You receive an email from a colleague, who is developing a software product that uses the HTTP api you helped building and which was released last week.

```
From: bob@eneco.com
Date: Tue Mar 23 10:21:58 AM CET 2021
Title: Your API is broken!

Hi!

I just tried to invoke your API, but I'm only receiving 'Unauthorized' responses. I
executed the call exectly as you showed me when you demonstrated the API's usage
yesterday.

What's going wrong? Did you run out of server disk space because you completely
filled it with funny cat pictures _again_?
```

```
Gr, Bob
```

You take a look at the access log of your server and notice a lot of 401 responses where the following text is used as part of the request:
`eyJ0eXAiOiJKV1QiLCJhbGciOiJSUzI1NiIsIng1dCI6Im5PbzNaRHJPRFhFSzFqS1doWHNsSFJfS1hFZyIsImtpZCI6Im5PbzNaRHJPRFhFSzFqS1doWHNsSFJfS1hFZyJ9.eyJhdWQiOiJodHRwczovL0VuZWNvLm9ubWljcm9zb2Z0LmNvbS9lY3Nhei1hcGlnZWUtb2RwLXQiLCJpc3MiOiJodHRwczovL3N0cy53aW5kb3dzLm5ldC9lY2EzNjA1NC00OWE5LTQ3MzEtYTQyZi04NDAwNjcwZmMwMjIvIiwiaWF0IjoxNjE2NDE2NzA1LCJuYmYiOjE2MTY0MTY3MDUsImV4cCI6MTYxNjQyMDYwNSwiYWlvIjoiRTJaZ1lFZzNWZG1mNmhzcHYvdVBNYmZmOC9NQ0FBPT0iLCJhcHBpZCI6IjFlMGZiMzU0LWE3OGQtNGY1Yi05OTY2LWVkNjIzYTYyNDU5OSIsImFwcGlkYWNyIjoiMSIsImlkcCI6Imh0dHBzOi8vc3RzLndpbmRvd3MubmV0L2VjYTM2MDU0LTQ5YTktNDczMS1hNDJmLTg0MDA2NzBmYzAyMi8iLCJvaWQiOiJkYjZlODlmOS1kYjE2LTQyNTItOTQyOS1jNGQ1ZTQ4YWRhYmQiLCJyaCI6IjAuQVFVQVZHQ2o3S2xKTVVla0w0UUFad19BSWxTekR4Nk5wMXRQbVdidFlqcGlSWmtGQUFBLiIsInJvbGVzIjpbIlJlYWRXb29uRW5lcmdpZSIsIlJlYWRFbmVjbyIsIlJlYWRPeHhpbyIsIlJlYWRFbmVjb0J1c2luZXNzIiwiUmVhZEFsbCJdLCJzdWIiOiJkYjZlODlmOS1kYjE2LTQyNTItOTQyOS1jNGQ1ZTQ4YWRhYmQiLCJ0aWQiOiJlY2EzNjA1NC00OWE5LTQ3MzEtYTQyZi04NDAwNjcwZmMwMjIiLCJ1dGkiOiJTZ0JhN1EzS1RVeTNRZjhoek9ZM0FBIiwidmVyIjoiMS4wIn0.AZRHBIxtI9uOT99Q0WRII1wPLKbrcBU-BHmQfUvaCCAnKApOZyrH3kxxtYjejgDTnoPIS0I1NnH0pxNd2ATN_50fcHjsXkCr3DaspJggLS_p2rT2nBMkTDyPBKIzw6rZ9tRrwsuFVXh2cYP1kIgoX-_PxpWfzIsyXctcSzbgYnLOJpiVDiZuz_hi4YWbmvc_lO5iezLLpXvQ0ER034tco9An6LtAGH0mK0-4FHW4McQGLpWPXhwAVLRzNdCXx4TNnOK7eDAVyvxf_fD_lbG9OSmd-PnLbFJKHwWNan06D7hTC8yQa4k-k0gcnHjDYKirG7CfnR5b6dKSst-BCa3X9w`. There are no observed layer 4 anomalies.

What could this snippet represent and how is it typically used? Can you derive some details from this snippet? Describe the steps you would take to further analyze Bob's issue before you will respond to his email.

6. Capture the flag

Consider this exercise optional. Did you find any flags already?

Here's some hints:

- Flags can be hidden in plain sight, you might be looking at one right now
- Do you understand what consitutes a HTTP request?
- Check the DNS records associated with `code001.ecsbdp.com`. What mechanism can be used to store arbitrary additional data?
- Some binary formats are just containers for data
- There are many objects living within an RDBMS. Tables, sequences, views, etc.