Data Engineer Technical Challenge

Questions

Part 1
You need to work as a data engineer as part of a team to design a Spotify-like system. Similarly to Spotify, the system needs to allow music creators to upload songs, and allow users to listen to these songs on various platforms (e.g. a mobile app, a desktop app, a web app etc.).

A key aspect of the Spotify-like system would be its recommendation engine. It needs to be state-of-the-art and recommend relevant songs to users so as to keep them engaged and increase the amount of time they spend using the system.

As part of your team, there will be data scientists who will be responsible for actually building the different machine learning models which will form the recommendation engine. From initial discussions with the data scientists, you've learned that they are thinking of exploring various features on which recommendations will be based. This includes, among others,

- Preferences explicitly provided by users (e.g. that they like rock music, they like 1990s music, their age etc.)
- Properties of songs which they actually listen to through the system (e.g. the singer/band, the genre, the language etc.)
- Properties which are to be extracted from the audio content of the song itself (e.g. tempo, danceability, speechiness etc.)

Devise a solution for getting such a system up and running. As part of the solution, you should think about and define:

- The different system blocks which would be required
- The database and schema you would use
- Caching and load balancing needs
- Data preprocessing pipelines required for feeding the data to the recommendation engines
- Potential scalability issues
- Potential security issues
- Potential trade-offs which you would have to make
- The main challenges you foresee in designing/building such a system

Note: Your final solution should include at least one clear diagram which depicts how the different parts of the system would work together. What we are looking for here is your ability to clearly scope down requirements, design a concept-level pipeline, and clearly communicate those.

Part 2

Assume that you have two tables defined by the following SQL queries.

```
CREATE TABLE products (
  sku INT NOT NULL,
  name VARCHAR(50) NOT NULL,
  price DECIMAL(10, 2) NOT NULL,
  PRIMARY KEY(sku)
);

CREATE TABLE orders (
  product_sku INT NOT NULL,
  price DECIMAL(10, 2) NOT NULL,
  quantity INT NOT NULL,
  tax_rate DECIMAL(3, 2) NOT NULL,
  shipping_rate DECIMAL(3, 2) NOT NULL,
  FOREIGN KEY(product_sku) REFERENCES products(sku)
);

INSERT INTO products VALUES (1, 'shirt', 25.99);
INSERT INTO products VALUES (2, 'sweater', 34.99);
INSERT INTO orders VALUES (1, 25.99, 1, 15.0, 3.0);
INSERT INTO orders VALUES (2, 34.99, 3, 13.0, 2.0);
```

1. How would you construct a query that shows the lowest priced product?
2. How would you construct a query that shows the average order cost?

Presentation

During your interview, you will be expected to give a short presentation (around 10 minutes) on the 2 parts of this challenge. This presentation will be targeted at an audience which includes both technical and non-technical people, so you should take this into account when creating your presentation.

You are welcome to prepare a short slide deck with relevant content from this challenge, if you think it would help you with your presentation. The slide deck **does not** have to be part of this submission.

Notes

- You can use any software to create the diagram for your solution or to write up your answers.
- Your submission should be **one** PDF file which contains all your explanations and diagrams for both part 1 and part 2 of the challenge.
- You shouldn't spend more than 6 hours on this challenge.
- Please email the PDF file to info@emoodie.com by 9pm on the 17th of September.