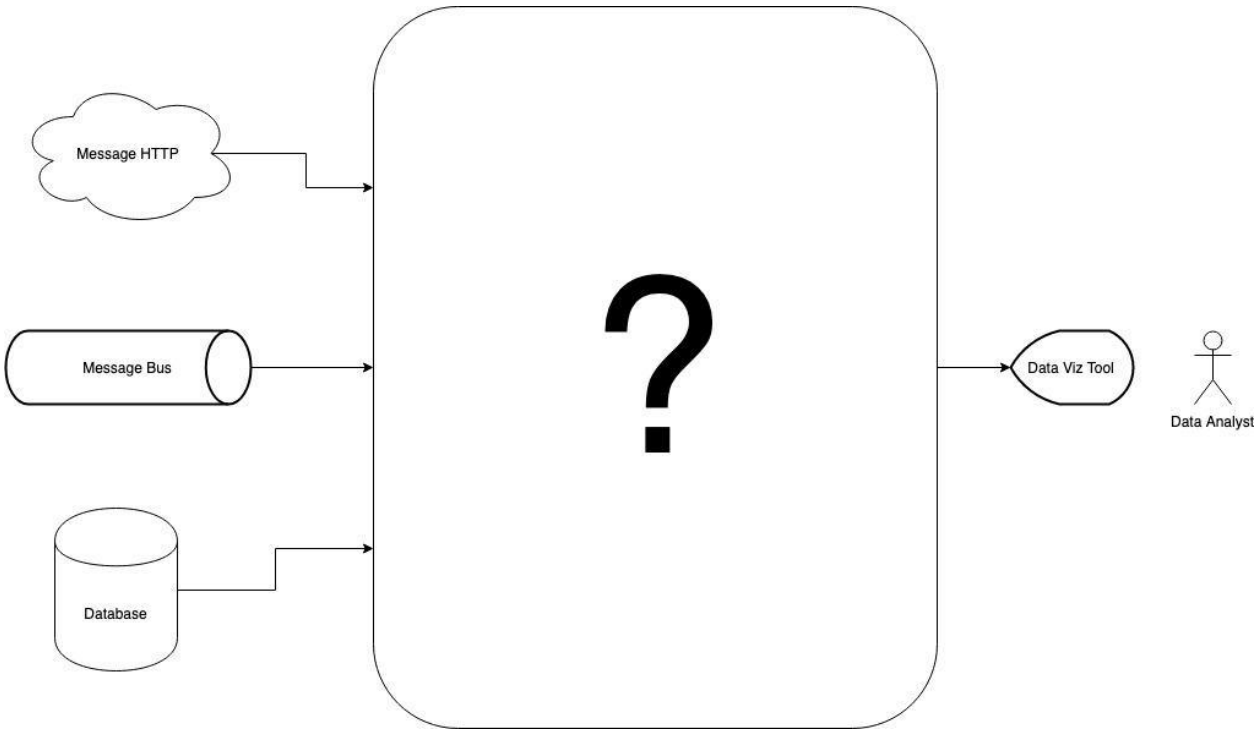
Data Engineer Challenge

The Goal of this challenge is to see your capabilities in capturing, mapping and aggregating data. We're also interested in seeing how you can solve problems by designing a solution based in the AWS Infrastructure.

The scenario below is one where you are encouraged to go as deep as you wish, ask any clarifying questions needed. Feel free to provide additional information, supporting documentation, and/or a demo.

Scenario:



We have multiple data sources in SafetyCulture, the main sources come from three places:

- **Message HTTP**: A 3rd party service that we use to enrich data which we extract via a HTTP request *(3rd_party_data.json)*.
- **Message BUS**: A messaging system where we receive real time data from multiple sources *(message_bus.json)*.
- **Databases**: Multiple types of Databases, SQL and NoSQL. These databases also provide data that we can read and maintain. (database.csv)

The Task:

You've been tasked to design a solution in a way that would aggregate the three data sources into one whilst providing a unique view of the consolidated data. Please include the technologies you would use to build it.

Expected Output:

1. The **solution design**, a draft of how you would build an end to end pipeline consuming from the three data sources, accompanied by an explanation for why you chose the specific technologies.
2. **An executable prototype** of the pipeline, using the example files managing the loading + aggregation of the data providing a unique view of the output. Not required to implement a database or API or Message Bus connection, processing the files direct from the disk is fine. Ensuring the **quality of the data and the code**.
3. Discuss the scalability and the deployment of the solution.