**Team Four**

*ACME company is a large financial analytics company with more than 4000 employees. ACME company’s business works with hedge funds, investment banks and holding companies and provides them with information about potential acquisitions, trades and mergers. For example, if ACME company’s customer wants to know if* ***Pomegranate Inc. (Ticker: PMGN, + 0.45 (2.3%))***  *would be a good investment, ACME company would provide a financial report using public and private data sources. The financial report would then help ACME company’s customer decide if it’s a good prospect or not. ACME company approach your team to create a document summarization system using Generative AI.*

**Key Stakeholders**

**VP of Engineering**

The end user of this system would be financial analysts who would like to speed up the time it takes to submit their deliverables to a central report ready for a customer. Using an input such as “**describe** **Pomegranate Inc’s performance in the global printer market and provide information such as market share and potential competitors”**, analysts would expect a summary to be about 500 words long. The output should provide data sources that have been used for the response (to make the response understandable and somewhat explainable) and to some degree the reasoning logic that has been used to come to a certain conclusion.

The model used would be training across a huge selection of structured and unstructured text data totalling about 2.3TB, growing at a rate of about 30GB per day. We have about 1,000 analysts and would like each response from the model to take less than 1 minute at a time, we expect analysts will be using this systems dozens of times per day. Since this application is business critical, it must have high availability and no downtime.

**Head of Data Science**

We have a large data science and engineering team here at ACME with 300-350 members. They are very experienced at developing ML models and have a background in the financial services industry. They are comfortable with Python and Java programming languages and we would like to understand how this model would be deployed using AWS.

**Chief Financial Officer**

We realise the importance of a good document summarization system for our company and appreciate that it is expensive to operate, from a financial perspective, we are willing to allow inference and training production systems to have very powerful infrastructure however, we are very keen that strict budgets are enforced and monitored in non-production environments.