**Data Engineer Project Requirements**

ACME Corporation sells a number of products. As a part of their product expansion, the company added toothbrushes and toys to their product mix. We’ve been asked by the stake holders to product some insights based on the sample data that represents the entire set of data. In particular, the stake holders want to understand:

1. Which product offering is performing better
2. Provide information on user demographics (age range, region) would help them hone in on marketing spend
3. Are the sales seasonal or does it remain consistent throughout the year
4. In the year 2021, our nationwide marketing spend was £200,000,000. Marketing activities covered the following regions - UK South East, UK North East, UK North, UK South. Can we justify the spend or was it a wasteful expenditure?
5. We need to find out by region if our CPA – cost per acquisition is higher or lower than the average as per the research found on <https://mystaticwebsite-3.s3.amazonaws.com/index.html>

**Technical aspects**

The operations team has put together a sales report in CSV format and uploaded the report to Amazon S3 – the file is publicly accessible and is at this location: <https://mystaticwebsite-3.s3.amazonaws.com/test-data.csv> .

The order\_id is a unique identifier in the report. Since the report is produced manually, some of the values might be incorrect. We should sanitize the data before any import.

**Devops**

We can run this task locally but we’ve been provided with a MySQL database to push our data into tables. The stake holders would like us to publish these insights as either graphs or tables. For the reports produced as tables, the stake holders would like to see the underlying SQL queries.

**Recommendations**

This database has potential to be expanded in the future so the data model needs to be scalable, robust and fault tolerant. Design your database in such a manner that we ensure we don’t truncate or delete data. The CSV file might change in the future and we have been given assurances that the order\_id field is unique. To insert data into MySQL to avoid duplication, use INSERT IGNORE. This will work as long as the order\_id field is set as primary.