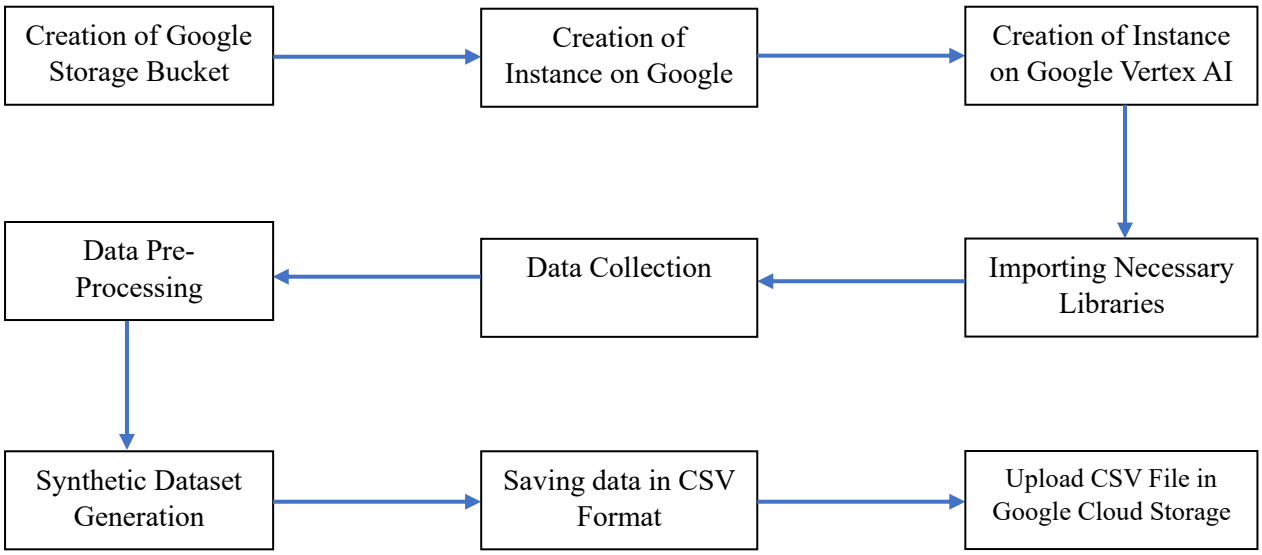
Synthetic Dataset Generation

Abstract **–** This project uses Google Vertex AI Test-Bison model to generate a synthetic dataset for product reviews using Product and Reviews dataset. The solution processes these datasets. The generated data is limited to 100 rows for efficiency with post-processing steps. The final synthetic dataset is stored in csv format and uploaded to Google Cloud Storage.

INTRODUCTION

Google Vertex AI is a machine learning platform that enables users to build, deploy and scale machine learning models efficiently. It integrates Google's AI tools providing pre-trained models and to train customized models on scalable infrastructure. One of the models is PaLM 2 model (Bison) which is the large language model designed for objectives like Text Generation [2]. Within its seamless integration into Vertex AI, developers can use this AI tool.

METHODOLOGY



(1) Creation of Google Storage Bucket and uploading the given datasets

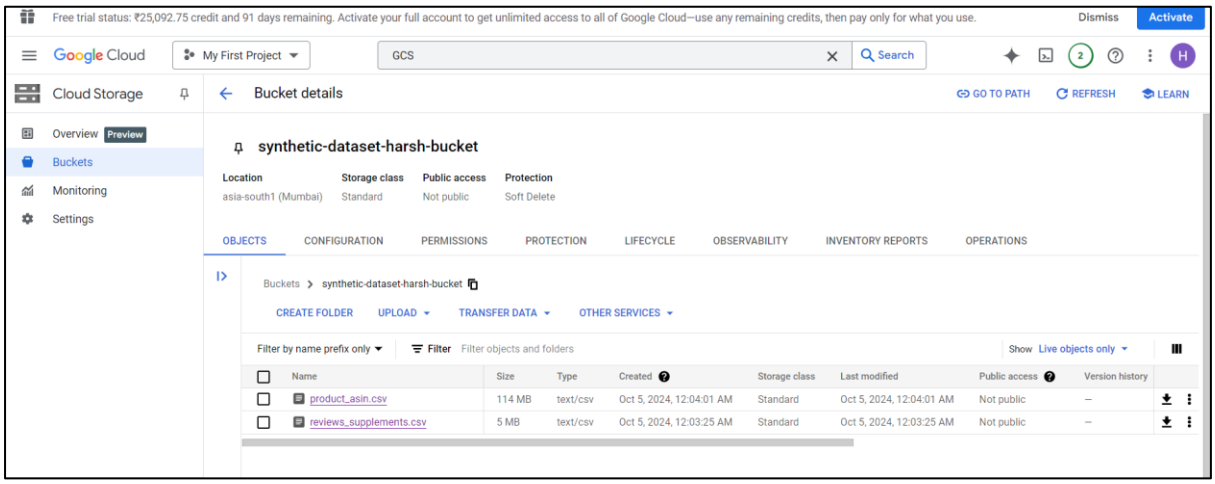


Fig 1: Uploading Product and Reviews Dataset

(2) Creating an Instance on Google Vertex AI and then launching the Jupyter Noteboook

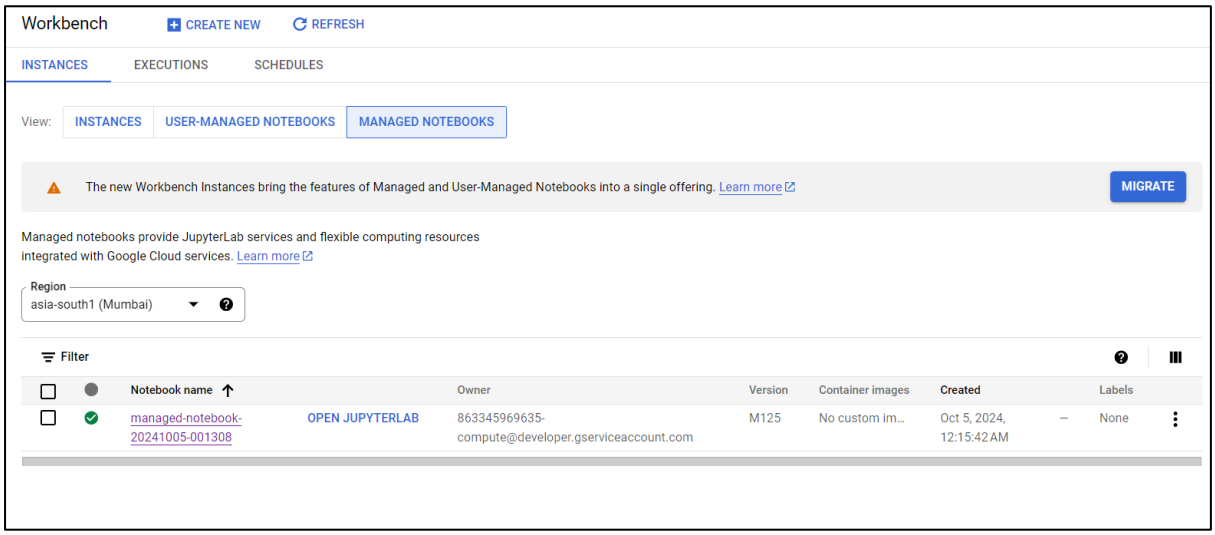


Fig 2: Instance on Google Vertex AI Workbench

(3) **Import necessary Libraries:** First step involves importing necessary libraries like pandas, sckit-learn etc.

(4) **Data Collection:** Upload dataset from the Google Cloud Storage bucket to Jupyter Notebook and then both the datasets were loaded into pandas data frames for easier manipulation and preprocessing.

(5) **Data Preprocessing:** The loaded data was cleaned by handling missing values.

(6) **Synthetic Data Generation using Google Vertex AI:** The solution uses the Vertex AI and its text-Bison Model to generate a synthetic dataset based on the attributes given in the dataset. The generated synthetic data includes fields like Product Title, Review Text, User ID. The model is used to generate 100 synthetic rows to reduce the time.

(7) **Post-Processing of generation of Synthetic Data:** After generating the synthetic data, the results were parsed and loaded into pandas' data frame.

RESULTS

The final synthetic dataset was made to saved in CSV format and uploaded to Google Cloud Storage.

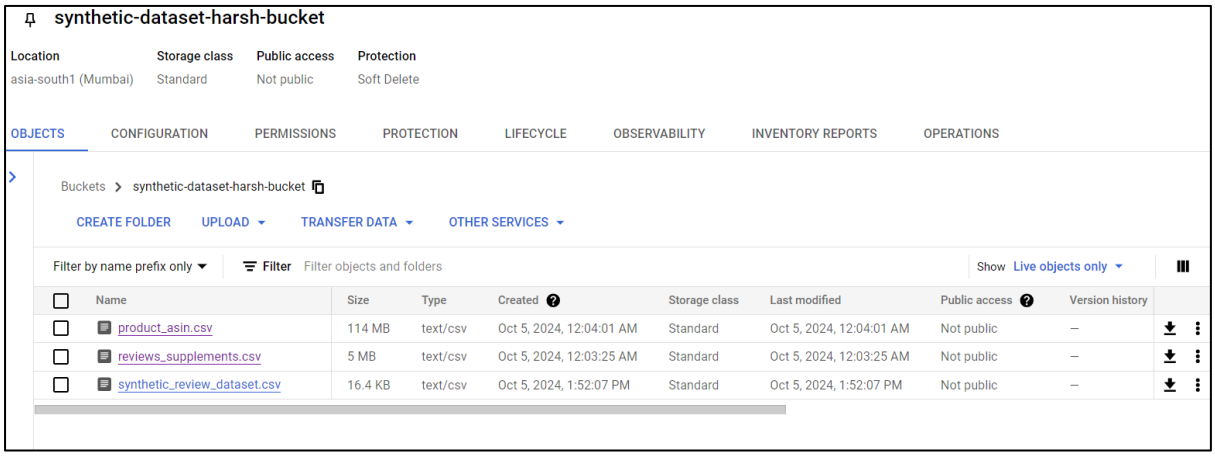


Fig 3: Synthetic Dataset in Google Vertex AI Workbench

CONCLUSION

Synthetic dataset gets generated using Product and Reviews Dataset using Vertex AI and Text-Bison Model.

Why was this model / architecture was used?

As we are dealing with text only so keeping this thing into consideration I've decided to use PaLM 2 model. Though other LLM models can also be used for text generation like Gemini 1.5 Flash, Gemini Nano, Gemini 1.0 Pro and so on but most of them are multimodal LLMs while PaLOM 2 Bison is specialized for text-based tasks. Since, as per the problem statement the objective is used to generate text so I believe Bison is more appropriate. [1]

What were the different factors considered for generating this dataset?

Generation of synthetic data for a limited number of rows. In this case, we limit it to 100 considering the time to generate the data.

Length of the product reviews will vary which will give a realistic impression to the dataset.

How do we measure the efficacy of a synthetic dataset?

Though in this solution I didn't measure the efficacy of the synthetic dataset but we can check the similarity rate using cosine similarity.

How do we ensure the synthetic dataset one generates is inspired from a source dataset but not an exact replica?

First by Prompt Designing

Second, it can be ensured by calculating the cosine similarity between original dataset and synthetic dataset.

What were the top challenges in solving for this problem statement?

Data Quality: Generating unique "user_id" which should be in alphanumeric format, "ratings" was challenging because it can not be repetitive but it is often repetitive in the generated synthetic dataset.

Performance: Generating synthetic data generation process took time due to the amount of data which slow down the iterations.

Fitting Generated Data in csv File: Fitting the generated data in csv was another challenge because some of the field's data were overlapping in generated csv file.

REFERENCES

[1] https://ai.google/discover/our-models/#:~:text=Discover%20the%20AI%20models%20behind%20our%20most%20impactful%20innovations,%20understand

[2] https://www.geeksforgeeks.org/palm-2-vs-gemini/