**Process followed for the analysis of fetch AE assignment dataset**

At the beginning , I have looked at the data set and realized these are mongodb json files with three datasets , after close examination , I found that receipts has nested data inside , so I thought this could be divided into 2 making the over all count to 4 datasets, which are

* Users
* Brands
* Receipts
* Items

I decided to use python to convert the json files to csv files since GCP BIG QUERY can import these files with ease, additionally I have use python to do a basic data profiling task using the library ProfileReport from ydata-profiling library, This profilereport creates html files about the profile of the dataset , this report includes dataset statistics, Distribution of values , missing values etc., it is used to determines the structure of the date , detect patterns.

once I have the overall idea of how data looked like , I have exported them converted csv files from json to BIGQUERY

Here is the ERD that can be used to represent the data

A screenshot of a computer

Description automatically generated

Here , I thought of following medallion design in the Big query Datalake

So , I have created three datasets

* Fetch ( where data is brought in as is )
* Fetch\_prep ( here, the transformation takes place including de duplication , datatypes conversion and column names changes etc )
* Fetch\_serve ( once the cleaned data is available in prep , we do the aggregation and analysis in this layer )

The layers can be seen below

A screenshot of a computer

Description automatically generated

Following link should take you to the dataset in Big query

fetch :

<https://console.cloud.google.com/bigquery?ws=!1m4!1m3!3m2!1sae-devraw!2sfetch>

fetch\_prep:

<https://console.cloud.google.com/bigquery?ws=!1m4!1m3!3m2!1sae-devraw!2sfetch_prep>

fetch\_serve:

<https://console.cloud.google.com/bigquery?ws=!1m4!1m3!3m2!1sae-devraw!2sfetch_prep>

I have attached the data mapping excel file ( named Mapping\_Datatypes.xlsx ) to the repository wherever datatype and other changes were done.

I have used the queries in analysis to draw insights in Serve dataset.

User Analysis : SELECT

    u.userId,

    u.active,

    u.createdDateTime,

    u.lastLoginDateTime,

    u.role,

    u.signUpSource,

    u.state,

    COUNT(r.receiptId) AS receiptCount

FROM

    `fetch\_prep.users` AS u

LEFT JOIN

    `fetch\_prep.receipts` AS r ON u.userId = r.userId

GROUP BY

    u.userId, u.active, u.createdDateTime, u.lastLoginDateTime, u.role, u.signUpSource, u.state

ORDER BY

    receiptCount DESC;

A screenshot of a computer

Description automatically generated

2 )Brand Revenue

A screenshot of a computer

Description automatically generated

3) Category Analysis

A screenshot of a computer

Description automatically generated

4) Purchase Trends over time

A screenshot of a computer

Description automatically generated

5)

User bonus points

A screenshot of a computer

Description automatically generated

Further , these can be used in a Visualization platform like Power BI , Tableau and looker to show more insightful data , this attempt is just showing the overall process outline we usually follow.

Thank you very for your consideration !!!