BeamData Ltd Project Summary

Extracting Data from Google Universal Analytics (ga3)

Ameenah Al-Haidari

alhaidari_am@outlook.com

Date: Nov 16, 2023

TABLE OF CONTENTS:

INTRODUCTION

This is the summary document of WeCloudData Academy's work on the Extracting Data from Google Universal Analytics project at Beam Data from Sep 20,2023 to Nov 30, 2023.

Our client is WeCloudData is a learning academy that providing the best quality data skills & AI training to the students with Career counselling services, and also for corporate clients.

The main service for this part from the project that BeamData team provided to them was on these parts:

- Retrieve the Historical user activity data from Google Universal Analytics starting from 2020-01-01 to 2023-07-24;
- Storage those Historical Data in PostgreSqL;
- Design the Universal Analytics Glossary;
- Break down the most differentiators between the two platforms " Universal Analytics and Google Analytics 4" with an aim to give a clear understanding of the new platform for better migration preparedness.

The project of Extracting Data from Google Universal Analytics will be discussed in this article.

PROBLEM STATEMENT

Google Analytics is a web analytics service offered by Google that tracks and reports website traffic and also the mobile app traffic & events, currently as a platform inside the Google Marketing Platform brand.

Starting on July 1, 2023, standard Google Universal Analytics properties stopped processing new data, and all customers will lose access to the Universal Analytics interface and API starting on July 1, 2024. To maintain your website measurement, you'll need a Google Analytics 4 property.

One of the primary questions on the minds of marketers, analysts, and developers is:

- How to Save your historical Universal Analytics data.
- What's the difference between Universal Analytics to Google Analytics 4?
- What is the best tool to storage your historical data?

The need to preserve this data is the driving force behind this project. Our responsibility will be

to help to ingest the historical [user activity data] from Google Universal Analytics. This data type focuses on user interactions with a website or app, capturing metrics such as page views, session duration, bounce rate, event tracking, and e-commerce transactions. User behavior data offers insights into visitor navigation patterns, the pages they visit most, the length of their stay, and actions they undertake, like filling out a form or making a purchase.

Tools used: Python(pandas, numpy, matplotlib, seaborn, sqlalchemy), BeautifulSoup (bs4), Selenium, google.oauth2, googleapiclient, oauth2client, apiclient, PostgreSqL, GitHub, Notion.

MAJOR WORKS

We have built different scripts and documents to Extracting Data from Google Universal Analytics, as following;

1. Google Analytics Reporting API from google.oauth2 and oauth2client;
2. Google's User Explorer in Universal Analytics (userActivity.search);
3. Google Analytics data with pagination and unsampled data;
4. Discover the Technical Challenges to get the Number of Indexed Pages on Google and Sitemap (SEO) using BeautifulSoup (bs4) and Selenium;
5. Script to create any config (ini) file for SQLAlchemy;
6. Script to importing Data from a Pandas to a PostgreSql using SQLAlchemy;
7. Wrote a document to collect the most important Universal Analytics Glossary;
8. Wrote a document to show the difference between the Universal Analytics and Google Analytics 4.
9. Wrote a document to Explain Key Errors and Duplicate Rows in fetched Google Universal Analytics reports API.

All scripts and documents of this project are posted to Beamdata repository on Github "beam-data/ga3_ameenah_a_202309"

https://github.com/beam-data/ga3_ameenah_a_202309

Google Analytics Reporting API

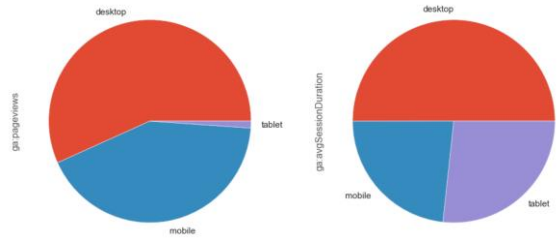
This script requests a report from Google Analytics Reporting API using google.oauth2 and oauth2client and returns the response as a DataFrame. It can handle pivot and dimensions reports, summary reports with no pivot and dimensions. With the Google Analytics Reporting API, we:

- Build custom dashboards to display Google Analytics data.
- Automate complex reporting tasks to save time.
- Integrate the Google Analytics data with other business applications.

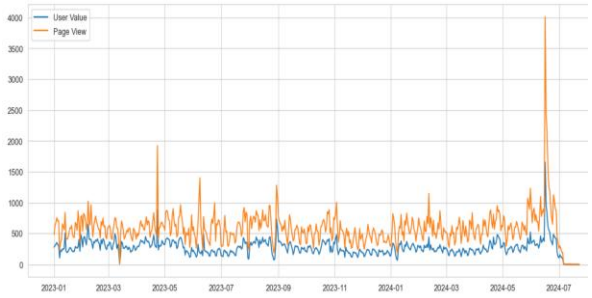
Samples from the Dashboard;

Device Categories

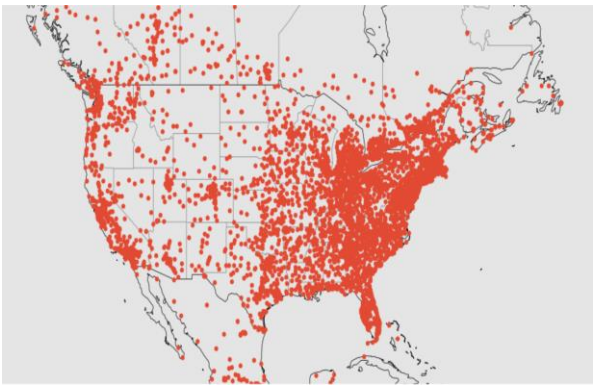
(pageviews vs avgSessionDuration)

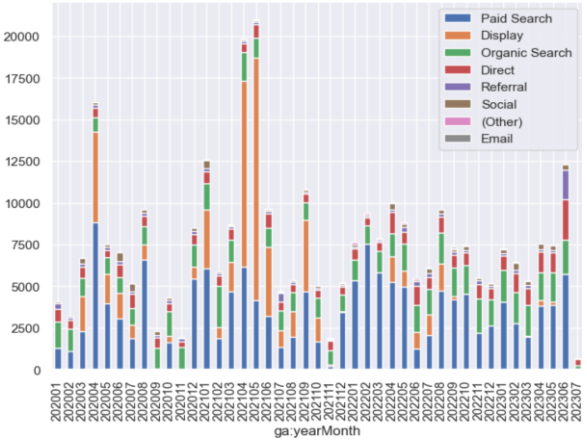


Get and visualize page views as well as the number of users.



Natural Earth Feature





Google's User Explorer in Universal Analytics (userActivity.search);

To pull out the User Explorer report with Python (userActivity.search), we did it in three stages;

First stage: retrieve and export Client Ids with number of sessions from a Universal Analytics using "analytics.reports().batchGet" and store that in csv file.

Second stage: read client IDs file and request Analytics Reporting API V4 using "analytics.userActivity().search"

```
def get_client_list_report(analytics,client_id):
    return analytics.userActivity().search(
        body = {
          'user': {
             'type': 'CLIENT_ID',
             'userId': client_id
          },
          'dateRange':
          {
             'startDate': '2021-11-01',
             'endDate': '2021-11-30'
          },
          'viewId': VIEW_ID,
          'pageSize': 100000,
          'pageToken': 'next_page_token',
        },
).execute()
```

The result is a table with nested dictionary under the "activities" column;

	client_id	sessionId	deviceCategory	platform	dataSource	activities	sessionDate
0	1.000033e+08	1634760666	desktop	Windows	web	[{'activityTime': '2021-10-20T20:11:06.703299Z', 'source': 'google', 'medium': 'organic', 'channelGrouping': 'Organic Search', 'campaign': '(not set)', 'keyword': '(not provided)', 'hostname': 'weclouddata.com', 'landingPagePath': '/', 'activityType': 'PAGEVIEW', 'customDimension': [{'index': 1}], 'pageview': {'pagePath': '/', 'pageTitle': 'Best Data Science and AI Courses in Canada | WeCloudData'}}]	2021-10-20
1	1.000819e+08	1634045131	desktop	Windows	web	[{'activityTime': '2021-10-12T13:25:31.196087Z', 'source': 'google', 'medium': 'organic', 'channelGrouping': 'Organic Search', 'campaign': '(not set)', 'keyword': '(not provided)', 'hostname': 'weclouddata.com', 'landingPagePath': '/building-digital-marketing-dashboard-using-python-docker-airflow-in-google-cloud-part-2/', 'activityType': 'PAGEVIEW', 'customDimension': [{'index': 1}], 'pageview': {'pagePath': '/building-digital-marketing-dashboard-using-python-docker-airflow-in-google-cloud-part-2/', 'pageTitle': 'Python and Docker used to Create a Digital Marketing Dashboard'}}]	2021-10-12

Third stage: flatten "activities" column, convert it to list and then to DataFrame. After that concat it to the original table. And the last step is export it to csv file.

The final table shape that exported to PostgreSql;

	client_id	sessionId	deviceCategory	platform	dataSource	sessionDate	activityTime	source	medium	channelGrouping	campaign
0	1.000033e+08	1634760666	desktop	Windows	web	2021-10-20	2021-10-20T20:11:06.703299Z	google	organic	Organic Search	(not set)
1	1.000819e+08	1634045131	desktop	Windows	web	2021-10-12	2021-10-12T13:25:31.196087Z	google	organic	Organic Search	(not set)

keyword	hostname	landingPagePath	activityType	customDimension	pageview.pagePath	pageview.pageTitle
(not provided)	weclouddata.com	/	PAGEVIEW	[{'index': 1}]	/	Best Data Science and AI Courses in Canada | WeCloudData
(not provided)	weclouddata.com	/building-digital-marketing-dashboard-using-python-docker-airflow-in-google-cloud-part-2/	PAGEVIEW	[{'index': 1}]	/building-digital-marketing-dashboard-using-python-docker-airflow-in-google-cloud-part-2/	Python and Docker used to Create a Digital Marketing Dashboard

The name of columns;

```
Index(['client_id', 'sessionId', 'deviceCategory', 'platform', 'dataSource',
       'sessionDate', 'activityTime', 'source', 'medium', 'channelGrouping',
       'campaign', 'keyword', 'hostname', 'landingPagePath', 'activityType',
       'customDimension', 'pageview.pagePath', 'pageview.pageTitle'],
      dtype='object')
```

Google Analytics data with pagination and unsampled data

Pulls Google Analytics data with pagination and unsampled data. Data was retrieved from 2020-01-01 to 2023-07-24. We counted and found the number of Sessions about 406864.0.

```
def get_report(analytics, pageToken='unknown'):

    return analytics.reports().batchGet(body={'reportRequests': [{
        'viewId': VIEW_ID,
        'pageSize': PAGESIZE,
        'samplingLevel': 'LARGE',
        'pageToken': pageToken,
        'dateRanges': [{'startDate': '2020-01-01',
                        'endDate': '2023-07-24'}],
        'metrics': [{'expression': 'ga:sessions'}],
        'dimensions': [
            {'name': 'ga:clientId'},
            ],
        }]}).execute()
```

Total number of Sessions

```
df.shape
```

```
(303121, 2)
```

```
pd.to_numeric(df['Sessions'], errors='coerce').sum()
```

```
406864.0
```

Discover the Technical Challenges to get the Number of Indexed Pages on Google and Sitemap (SEO) using BeautifulSoup (bs4) and Selenium

This Python script performs a search to check the number of indexed pages on Google for multiple sites using Selenium, bs4 and Python. And compare that with the real indexed pages from Sitemap-parser.

For the aim to know and to get an idea of how committed a site is in a market. Around how many pages does the business deal with 200 or 200M. And to have an idea of the competitor's index sizes when building an SEO strategy.

We found;

	indexed_pages
weclouddata.com	471
brainstation.io	13,700
lighthouselabs.ca	2,420
junocollege.com	502
metroc.ca	290

- the Number of Indexed Pages that scraped from weclouddata sitemap is 1008 whereas from google is about max 471, with poor performance (46.73%).
- the Number of Indexed Pages that scraped from brainstation.io sitemap is 11298 whereas from google is about max 13,700, with extra performance (121.26%) .
- the Number of Indexed Pages that scraped from lighthouselabs.ca sitemap is 1118 whereas from google is about max 2,420, with extra performance (216.46%).

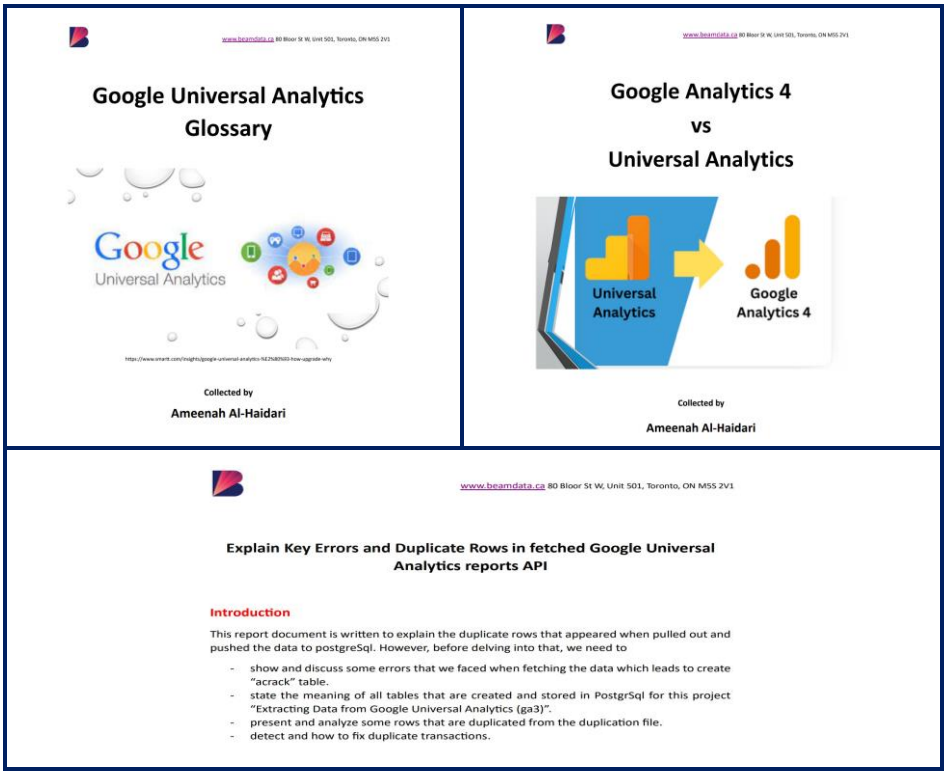
Wrote three Documentations;

- **to collect the most important Universal Analytics Glossary**
- **to show the difference between Universal Analytics and Google Analytics 4.**
- **to explain KeyErrors and Duplicate Rows in fetched Google Universal Analytics reports API.**

The Universal Analytics (UA) is the previous version of Google Analytics, and was used by many websites for tracking their traffic. After Google Analytics 4 (GA4) released, you might find it challenging to understand all the terminologies attached to the platforms related to the various Google Analytics versions.

There are so many reports and so much data inside the UA and GA4 (GA4). With huge information on the Internet sites of the different Google Analytics versions that makes kind of confusion especially for the beginners. Beside that, one of the primary questions on the minds of marketers, analysts, and developers is: What's the difference between Universal Analytics to Google Analytics 4? Thus, we;

- break down the most differentiators between the two platforms.
- Write Google Universal Analytics glossary.



CHALLENGES

Google Analytics put limits and quotas on API requests to protect the system from receiving more data than it can handle, and to ensure an equitable distribution of the system resources.

> The following quotas apply to all Reporting APIs, including the Core Reporting API v3, Analytics Reporting API v4, Real Time API v3, and Multi-channel Funnel API v3:
>
> - 10,000 requests per **view (profile)** per day (cannot be increased)
> - 10 concurrent requests per **view (profile)** (cannot be increased)

The project is done on Nov 16, 2023, but because of these limits and quotas, we only retrieve less than 10,000 per day. And with crowded traffic day, we can ingest less than 8,000 per day. Thus, the extracting data will continue for around 3 weeks after this date.

FUTURE WORKS

if any

ACHIEVEMENTS AND CONCLUSIONS

The key achievements that Beamdata made through the Extracting Data from Google Universal Analytics project were;

- help the client to ingest and export the historical [user activity data] before the aforementioned deadline;
- support the BI to build custom dashboards to display Google Analytics data;
- be able to compare between the two platforms the UA and GA4;
- build good knowledge to understand the most differentiators between the two platforms and cover the most important terminology with the aim to give a clear understanding of the new platform GA4 for better migration preparedness.
- recommend improving Keywords for SEO.