Audience Builder - Google BigQuery

Client

Our client is an advertising technology company focused on providing solutions to marketers to enable them to plan, execute and measure their digital media campaigns. The client has a huge data bank of 1.2 billion registered user profiles which serves as the primary foundation for their people based advertisements.

Objective

The client's aim was to develop an Audience builder tool, the crown jewel of a deterministic identity management platform. The tool is intended to build an actionable customer list based on factors such as age, gender, ethnicity, purchase history, online interests, TV viewing behaviour, etc.

The crux of the solution is to build a propensity score model for the 1.2 billion people data set by leveraging the huge offline customer purchase data from Direct match data partners such as Nielsen Catalina Solutions, Neustar.

Challenges

The problem was one of handling "BIG" data volumes. For one run, the application should be capable of handling 134 raw segment files; each file having data of 1M customers, with attributes running to 900 columns – totally amounting to 120 Billion data points.

Solution

Congruent designed a solution using Google BigQuery and Python as the primary technology stack.

The client's people data set from Google BigQuery was downloaded into hash tables. The segment files from partners were downloaded from the FTP folder, unzipped. The data from the two sources was used to arrive at propensity scores for each household member. Members with propensity scores above a set threshold were identified and pushed to the audience builder file in csv format. The csv files were finally uploaded to GBQ using the chunking approach.

Some highlights of the solution design are:

- Distributed and asynchronous processing in order to improve throughput
- Complies with no single point of failure principal
- Fault tolerant, capable of recovering and restarting from any unforeseen environmental issues.

Benefits

The capability to handle big data volumes provided the customer the ability to identify target audience

- with a far greater degree of accuracy (100% better on-target average against an industry standard of 32%)
- at a significantly faster turn-around times (10 times as compared to previous processing times) and
- with almost no human oversight required to run and monitor the process

Technology Used

Google Big Query APIs, Python, Sqlite3