Data Architecture

Brief

DatastreamX runs a set top box service which display ads and gives free view channels, they are currently collecting data from 30,000 set top boxes into a system running on premise. The data collected is analyzed and appropriate ads are pushed back to the set top box. The ad campaign is managed by a third party and they give the ad content based on parameters like channel being viewed, last channel viewed and the last ad displayed. Currently the system is collecting data and able to send it near real time to the ad aggregator.

The business has indicated that they will grow 10x in the next year and also want to expand the service support multiple ad aggregators to improve their revenue. The business is also in the process of hiring a few data scientists and analyst to build models that will help the select the right ad aggregator.

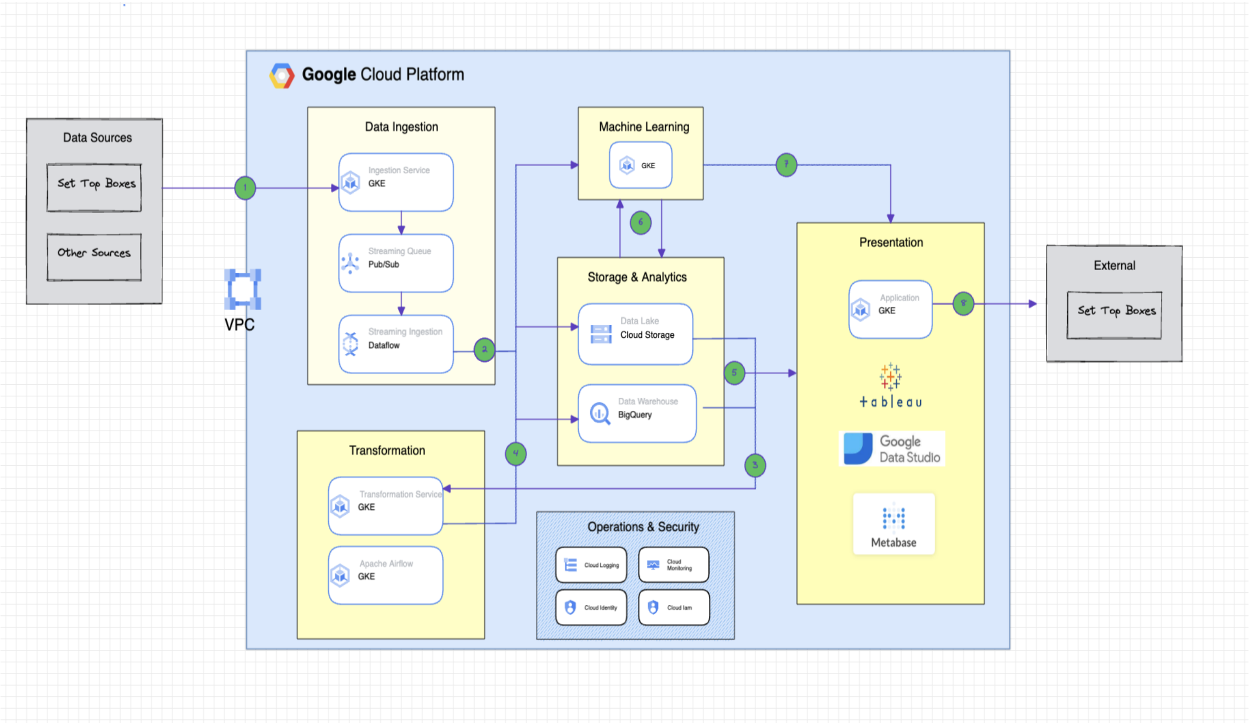
As part of the proposed solution the business analysts needs more detailed reporting in as close to real time as possible. The IT on the other hand are very concerned about cost, and would prefer to work on product improvement and automation than manage servers.

Objective

Recommend a manageable, secure, scalable, high performance, efficient, elastic, highly available, fault tolerant and recoverable architecture that allows DatastreamX to organically grow and harness data to derive insights. The architecture should specifically address the requirements/concerns as described above.

Deliverables

A PDF document no greater than three or four pages in length that clearly and succinctly present an analysis of the organizations requirements and the proposed architecture diagram. Clearly state all assumptions made during the design and explicitly state the referenced websites.



Use Case

1. Near Real Time Detail Reporting

   It achieve with diagram number 1->2->5

2. Data scientist build model

   It achieve with diagram number 1->2->6 / 1->2->7

Constraint

1. Cost and Focus on Product Improvement

   All the service using Google Cloud Platform (GCP), it will shifting the focus from maintaining server operational to deliver feature and insight. For cost, BigQuery will generate the biggest cost, and we will use some best practice to lowering the cost of BigQuery:

   1. Using Partition and Cluster on every table
   2. Using slot rather than pay as you go if it more affordable

2. Manageable

   All the service using Google Cloud Platform (GCP) and GCP has the capability to manage the service through dashboard, API, or command line. Hence, this make this architecture is manageable

3. Secure
   - The architecture build on top of VPN, so it provide security for accessing all the service.
   - It use cloud identity for all the services that need authentication to access, for instance Metabase, and Airflow
   - It use cloud IAM to govern BigQuery access, so the access to data will be govern respective to team and security clearance.

3. Scalable,Highly Available, Fault Tolerant
   - Using Cloud Managed service, automate scaling feature on GKE, and HA feature on every services