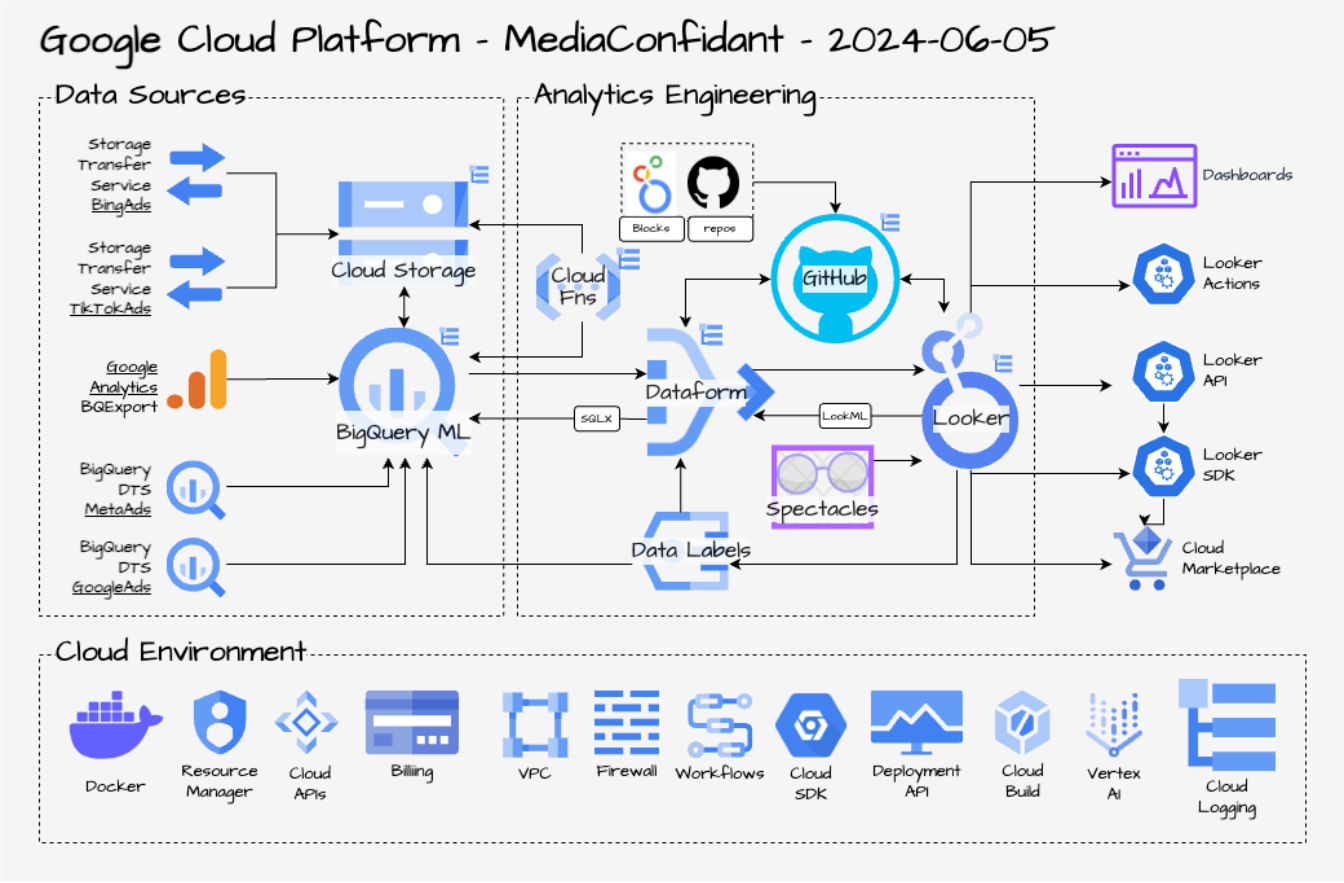
Google Cloud Platform - MediaConfidant

Analytics Pipeline as Code for Today's Data Product

Overview

This repository contains the code and documentation for an analytics pipeline built for MediaConfidant using Google Cloud Workflows and other Google Cloud services. This pipeline automates data ingestion, transformation, analysis, and reporting to provide actionable insights for MediaConfidant's business. Below is the updated architecture diagram as of 2024-06-05:



Architecture

This project utilizes a modular workflow approach to manage the entire data pipeline. Each workflow represents a distinct stage of the process, ensuring a clear and organized structure.

Data Sources

- BingAds: Storage Transfer Service
- TikTokAds: Storage Transfer Service
- Google Analytics: BQExport
- MetaAds: BigQuery DTS
- GoogleAds: BigQuery DTS

Analytics Engineering

1. Data Ingestion:
    - Data from various sources like BingAds and TikTokAds is ingested into Cloud Storage.
    - Google Analytics data is exported to BigQuery ML.
2. Data Processing:
    - BigQuery ML processes data from Cloud Storage and other sources.
    - DataForm integrates with GitHub for version control and uses Cloud Functions for automation. SQLX is used within DataForm for data transformations.
3. Data Management:
    - Looker is used for data visualization and reporting.
    - LookML integrates with Looker for data modeling.
    - Spectacles is used for LookML validation.

Outputs

- Dashboards: Generated by Looker for data visualization.
- Looker Actions: Custom actions integrated into Looker.
- Looker API: Used for programmatically accessing Looker data.
- Looker SDK: For integrating Looker with other applications.
- Cloud Marketplace: For publishing and accessing Looker applications.

Cloud Environment

- Docker: Containerization for applications.
- Resource Manager: For managing cloud resources.
- Cloud APIs: Various APIs provided by Google Cloud.
- Billing: Managing billing and costs.
- VPC: Virtual Private Cloud for networking.
- Firewall: Security and access control.
- Workflows: Orchestrating complex workflows.
- Cloud SDK: Tools and libraries for cloud management.
- Deployment API: Automated deployment of applications.
- Cloud Build: CI/CD for building and deploying applications.
- Vertex AI: For machine learning and AI applications.
- Cloud Logging: Centralized logging for monitoring and troubleshooting.

Getting Started

To get started with the project, follow these steps:

1. Set up your environment:
    - Ensure you have the required tools installed, such as Docker, Cloud SDK, and other dependencies.
      Follow the setup instructions in the relevant directories.
2. Deploy the infrastructure:
    - Use Deployment Manager API scripts to set up the infrastructure on GCP.
    - Ensure all services are correctly configured and accessible.
3. Run the data pipeline:
    - Use DataForm to execute the data ingestion and transformation processes.
    - Monitor the pipeline using Cloud Logging and Cloud Monitoring.

Supporting Documentation

Article 1: “Modern Data Pipeline Building with BigQuery Dataform — Part 1” by Peter Bavinck

In this article, Peter Bavinck introduces the foundational aspects of using BigQuery Dataform for building modern data pipelines. Dataform, now integrated within the BigQuery environment, facilitates SQL-based data transformations, making it accessible for data analysts and engineers. Bavinck outlines the benefits of the ELT (Extract, Load, Transform) approach, emphasizing the simplicity and efficiency of writing data pipelines in SQL.

The article walks through a basic transformation example, where a customer table is extended with a new segment column based on predefined value thresholds. Bavinck explains the structure of Dataform’s SQLX files, which include both SQL statements and metadata. He demonstrates how to make Dataform aware of existing data sources, enabling it to build and visualize execution graphs that define transformation dependencies.

Additionally, Bavinck highlights the use of JavaScript within SQLX files to create reusable code snippets and constants, enhancing the maintainability and flexibility of data transformations. By centralizing values and logic in JavaScript modules, changes can be applied across multiple transformations with minimal effort.

Overall, the article provides a clear and practical introduction to leveraging BigQuery Dataform for efficient data pipeline development, emphasizing the ease of use and integration capabilities within the Google Cloud ecosystem.

Article 2: “Modern Data Pipeline Building with BigQuery Dataform — Part 2: Incremental Tables” by Peter Bavinck

In the second part of his series, Peter Bavinck delves into the concept of incremental tables in BigQuery Dataform, a feature crucial for handling large, growing datasets efficiently. Incremental tables allow for only new records to be processed after the initial build, significantly optimizing performance.

Bavinck begins by setting up example tables to illustrate how incremental processing works. He demonstrates the SQL statements required to create orders and products tables, which are then joined to form a transformed table. The key to incremental processing is the use of a WHERE clause to filter out old records, based on the maximum ingestion time of previously processed data.
The article introduces Dataform’s self() function, which references the result of its own action, and the when() function, which conditionally applies the WHERE clause only during incremental

executions. Bavinck also covers the use of partitioned tables, explaining how partitioning can further optimize incremental processing by reducing the need for full table scans.

The article concludes with a detailed explanation of configuring incremental tables, including the use of BigQuery scripting for declaring variables and defining partitioned incremental tables. Bavinck emphasizes the flexibility and efficiency gains achieved through incremental processing in Dataform, providing a comprehensive guide for implementing this advanced feature.

Article 3: “Introduction to Data Product Blueprint Model” by Jarkko Moilanen (PhD)

Jarkko Moilanen explores the evolution of data management into data products, emphasizing a customer-centric approach and iterative development. Traditional data management, focused on internal operational needs, has transformed into interactive data catalogs that facilitate data discovery and consumption.

Moilanen discusses the concept of Moore’s Chasm, highlighting the challenge of transitioning data products from early adopters to the mainstream market. Successful data product management requires crossing this chasm by addressing the concerns of risk-averse mainstream customers.
The article outlines the importance of data reuse and monetization, distinguishing between internal and external data products. Internal products enhance data reuse and decision-making within organizations, while external products involve commercializing data assets for external customers. Moilanen introduces the concept of Data Agreements, which extend technical Data Contracts with legal elements to manage risks when data products are exposed to external customers.

Feedback loops are emphasized as crucial for continuous improvement in data product development. Moilanen advocates for a unified process model that allows internal data products to be easily commercialized when needed. The article also calls for greater cooperation between emerging data economy standards to streamline metadata management and ensure interoperability.

Article 4: “The Modern Data Products are Programmable” by Jarkko Moilanen (PhD)

Jarkko Moilanen explores the evolving landscape of data products, asserting that the “everything as code” model, a staple in software development, is becoming integral to the data economy. He advocates for the programmability of data products, which means they should be able to accept instructions to perform a range of tasks, similar to programmable computers.

Moilanen highlights the importance of open dialogue and interoperability between different data standards, such as the Open Data Product Specification and the Data Product Descriptor

Standard (DPDS). He emphasizes that despite their differences, these standards aim to enhance the functionality and interoperability of data products.

A key focus is on automation and programmability through APIs, which allow for flexible and efficient management of data products. Moilanen discusses how modern enterprise networks have evolved from manual CLI efforts to automated, programmable systems, enhancing efficiency and adaptability. This shift towards “infrastructure as code” is mirrored in data product specifications, where APIs play a crucial role.

The article also covers the integration of Service Level Agreements (SLAs) and data quality monitoring as code within data product specifications. This allows for the automatic execution of data quality assessments and SLA monitoring, improving the reliability and transparency of data products.

Moilanen concludes by calling for greater cooperation between emerging data economy standards to avoid redundancy and enhance interoperability, advocating for a unified approach to defining metadata and functional components.

Article 5: “Introduction to Data Product Blueprint Model” by Jarkko Moilanen (PhD)

Jarkko Moilanen delves into the transformation of traditional data management into modern data product management. He explains how traditional data management focused on internal operational needs, often resulting in siloed systems. In contrast, data product management emphasizes delivering value to external customers through iterative development and a customer-centric approach.

Moilanen introduces Moore’s Chasm, a concept that describes the gap between early adopters and the mainstream market. He explains how data products must cross this chasm to achieve mass-market success, requiring different marketing and sales strategies.

The article outlines the importance of data reuse and monetization, distinguishing between internal and external data products. Internal data products enhance decision-making within organizations, while external products involve commercializing data assets. Moilanen also introduces Data Agreements, which extend Data Contracts with legal elements to manage risks when data products are exposed to external customers.

Feedback loops are highlighted as crucial for continuous improvement in data product development. Moilanen advocates for a unified process model that allows internal data products to be easily commercialized when needed. He emphasizes the need for cooperation between emerging data economy standards to streamline metadata management and ensure interoperability.

Overall, Moilanen provides a comprehensive overview of the data product blueprint model, emphasizing the need for a customer-centric approach, iterative development, and effective data monetization strategies.

Examplecc: “Building Your Own GA4 Rules-Based Marketing Attribution Models using Google BigQuery and Looker” by Mark Rittman

In this article, Mark Rittman guides readers through creating custom marketing attribution models using Google Analytics 4 (GA4), BigQuery, and Looker. He begins by explaining the limitations of default attribution models in GA4 and the need for customized solutions to better reflect marketing strategies and customer journeys.

Rittman details the process of exporting GA4 event data to BigQuery, where users can leverage SQL to define custom attribution logic. He provides a step-by-step guide on writing SQL queries to calculate attribution based on user-defined rules, such as multi-touch or time-decay models. Rittman emphasizes the flexibility of SQL in tailoring attribution models to specific business needs.

The final step involves visualizing the custom attribution data in Looker. Rittman explains how to create LookML models to define the data structure and build intuitive dashboards that provide actionable insights. He highlights the importance of validating the custom models to ensure accuracy and reliability.

Overall, the article demonstrates the power of integrating GA4, BigQuery, and Looker to build sophisticated marketing attribution models, offering a comprehensive guide for marketers seeking to enhance their analytics capabilities.