QBurst

GCP Powered Data Engineering Solution

Overview

A comprehensive data management solution that consolidates unstructured data from diverse sources into Google Cloud's Firestore. Our approach involved incorporating security measures such as face masking API, encryption, and ensuring compliance while safeguarding sensitive information. The centralized infrastructure enabled improved analytics, scalability, and controlled data lifecycle management, ensuring improved decision-making and collaboration, with controlled access for authorized users via Azure AD integration. The solution streamlined operations, enhanced security, and ensured data-driven growth.

Client Profile

A leading fashion retail company that owns several brands and operates across diverse markets.

Business Challenges

The client wanted to develop pipelines to manage unstructured data sourced from various channels such as Instagram, eCommerce sites, fashion magazines, store surveys, and third-party providers. The primary objective was to create a centralized storage system for this data to facilitate visualization and analytics.

- **Data integration challenges:** Dealing with diverse data formats and sources required robust extraction and transformation methods.
- **Quality assurance and real-time handling:** Ensuring data accuracy across sources while managing real-time data requires stringent validation and continuous monitoring mechanisms.
- **Scalability, security, and compliance:** Building scalable infrastructure, implementing strong security measures, and adhering to data privacy regulations are crucial for data management.

- **Analytics compatibility and maintenance:** Enabling seamless integration with analytics tools and establishing ongoing monitoring and maintenance protocols for reliable data pipelines.

QBurst Solution

We developed a data engineering solution to manage unstructured data from third-party sources. Leveraging Google Cloud Platform (GCP) services, we built pipelines for extracting, transforming, and loading data into Firestore, our centralized cloud storage database.

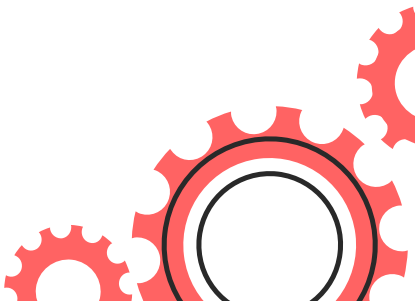
To protect Personally Identifiable Information (PII), we created a face masking API integrated into the data transformation process. This API ensured the security of PII data during processing. Additionally, robust data protection measures, including encryption and regular database backups, were implemented to safeguard sensitive information.

Continuous monitoring and feedback ensured the reliability, security, and compliance of the data pipelines and protection mechanisms.

Project Highlights

- **Analysis, Architectural Design, and Documentation:** Emphasis on comprehensive analysis, architectural design, and thorough documentation of various project components.
- **Utilization of GCP Services:** Implementation of services such as GCS buckets, Pub/Sub, and Cloud Run for data extraction and recovery from third-party APIs.
- **Airflow DAGs for Data Consumption:** Development of Airflow Directed Acyclic Graphs (DAGs) to consume unstructured data from third-party providers in various formats such as images, PDFs, text, and JSON.

- **Custom Pipelines and GKE Clusters:** Creation of custom pipelines leveraging GKE clusters for extracting unstructured data, particularly using Google Cloud Vision API for OCR on PDFs.
- **Firestore as Centralized Cloud Storage:** Loading extracted data into Firestore, a serverless document database on Google Cloud infrastructure, through component-specific Airflow DAGs.
- **Database Workload Protection:** Implementation of a full and incremental backup and restore mechanism for Firestore DB to prevent potential data loss or corruption.
- **Integration of Azure AD Permissions:** Implementation of Azure AD integration with specific permissions (view, edit, delete) for image APIs on Cloud Run services, enabling access for valid Azure AD account users from other applications.
- **PII Protection via Image Masking:** Development of face detection/masking APIs as a service to protect Personally Identifiable Information within images sourced from third parties, ensuring compliance with regulatory requirements.
- **Data Retraction and Deletion Components:** Airflow DAGs are designed for additional data protection by facilitating data retraction and deletion in compliance with regulatory guidelines.
- **Performance Enhancement Tasks:** ETL jobs focusing on performance improvements within GCP components through discussions and implementations (Cases) to optimize efficiency.
- **Export Jobs for Visualization and Analytics:** Development of export jobs enabling the use of centralized storage data (Firestore) for visualization and analytical purposes.



Technologies Used

- Google Cloud Platform
- GKE cluster
- Pub/Sub notification
- Secret Manager
- API Gateway
- VPC
- Alert Policy
- Google Drive API
- Terraform
- Azure AD
- GCS bucket
- Cloud Composer
- Dead Letter Queue
- Cloud Run
- Cloud Firestore
- Cloud Logging
- Monitoring Dashboard
- Google Cloud Vision API
- Python 3.11

Benefits

- **Centralized Data Management:** Consolidation of unstructured data from diverse sources into Firestore enables easier access and management.
- **Enhanced Security and Compliance:** Robust measures such as face masking API, encryption, and access controls ensure compliance with data regulations and bolsters data security.
- **Improved Analytics and Decision-making:** Organized data infrastructure enables better visualization and analysis, fostering informed strategic decisions.
- **Efficient Data Handling:** Streamlined extraction, transformation, and loading processes through GCP services and custom pipelines enhance operational speed and agility.

- **Scalability and Performance:** GKE clusters and performance enhancements ensure scalability and optimized efficiency even with increased data volumes.
- **Controlled Data Lifecycle Management:** DAGs for data retraction and deletion align with regulatory guidelines, facilitating effective data lifecycle management.
- **Collaboration and Access Control:** Integration with Azure AD enables controlled access for authorized users, promoting collaboration while maintaining security.
- **Actionable Insights:** Export jobs allow easy retrieval of data for visualization and analytics, ensuring data-driven decision-making and business growth.