QBurst | A High AI-Q™ Company

GCP-Powered Data Engineering Solution for Data Management

A comprehensive data management solution on Google Cloud Platform (GCP) which consolidates unstructured data from diverse sources, ensuring robust security, scalability, and enhanced analytics capabilities.

Overview

QBurst delivered a data engineering solution on GCP to address a global fashion retailer's unstructured data challenges.

- **Centralized Data Hub:** Consolidated unstructured data from diverse sources into a secure, centralized Firestore database, enabling easier management and access.
- **Enhanced Security:** Implemented robust security measures, including a face masking API for PII protection and encryption, ensuring compliance with data privacy regulations.
- **Improved Analytics:** Created a scalable data pipeline that transformed raw data into a structured format, enabling actionable insights and data-driven decision-making.

Client Profile

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

Challenges

- **Data Integration:** The variety of data formats and sources (images, PDFs, text, JSON) required complex extraction and transformation pipelines.
- **Data Quality & Real-Time Handling:** Ensuring the accuracy of real-time data from various channels required stringent validation and continuous monitoring.
- **Security & Compliance:** The data contained Personally Identifiable Information (PII), requiring robust security measures like data masking and adherence to privacy regulations.

- **Scalability & Maintenance:** The solution needed to be highly scalable to handle ever-increasing data volumes while being easy to maintain and integrate with existing analytics tools.

QBurst Solution: GCP-powered Data Engineering

We developed a comprehensive data engineering solution leveraging the power of Google Cloud Platform (GCP) services. We built a series of automated pipelines for extracting, transforming, and loading unstructured data from third-party sources into Cloud Firestore, which serves as the centralized, secure data repository.

Our Solution Includes

- **Automated Data Pipelines:** We developed Airflow Directed Acyclic Graphs (DAGs) to consume unstructured data from third-party providers. Custom pipelines were created using GKE clusters to perform Optical Character Recognition (OCR) on PDFs using the Google Cloud Vision API.
- **PII Protection:** A custom face masking API was developed and integrated into the data transformation process to automatically detect and protect PII within images.
- **Centralized Data Storage:** Extracted and transformed data was loaded into Firestore, providing a single, organized source for analytics and visualization.
- **Robust Security and Compliance:** We implemented comprehensive data protection measures, including encryption, regular backups, and data retraction/deletion components to ensure regulatory compliance.

Technical Highlights

- **GCP Service Utilization:** The solution leverages a suite of GCP services, including GCS buckets, Pub/Sub, and Cloud Run, for efficient data extraction and processing.
- **Azure AD Integration:** We integrated with Azure AD to provide controlled, role-based access for authorized users, promoting secure collaboration.

- **Automated Data Lifecycle Management:** Airflow DAGs were designed to manage the entire data lifecycle, including data retraction and deletion, in compliance with regulatory guidelines.
- **Performance Optimization:** Implemented ETL jobs and a full and incremental backup/restore mechanism for the Firestore database to ensure high performance and data integrity.

Impact: Making Sense of Unstructured Data

- **Enhanced Security and Compliance:** The face masking API and robust encryption ensured the protection of sensitive PII, giving the client confidence in their data privacy measures.
- **Improved Analytics & Decision-Making:** By consolidating and structuring data, the solution enabled better visualization and analysis, fostering more informed strategic decisions.
- **Efficient Data Handling:** The automated data pipelines streamlined operations, enhancing operational speed and agility in managing vast volumes of unstructured data.
- **Scalability & Reliability:** The use of GKE clusters and scalable GCP services ensures that the platform can efficiently handle increasing data volumes up to 70%, while the backup and restore mechanisms guarantee data integrity.
- **Seamless Collaboration:** Seamless Collaboration: Controlled access via Azure AD integration promotes secure collaboration among teams.