Sakina Banu

GCP Data Lake Project

**Project: Real-Time Data Lake Implementation for StyleNest on Google Cloud**

**1. Project Background**

As a Data Engineer for this project, I was responsible for designing and implementing a **real-time data lake architecture** for **StyleNest**, a growing sustainable fashion e-commerce brand.  
The client was dealing with scattered data sources (Shopify orders, Google Ads performance, inventory updates, customer chat logs) without a centralized system for storage, processing, and analysis.

**2. Business Requirements**

* Consolidate structured and unstructured data into a **single centralized cloud environment**.
* Enable **real-time and batch** data ingestion.
* Cleanse, standardize, and prepare data for **business intelligence** needs.
* Build **automated dashboards** for executive and marketing teams.
* Ensure **cost-efficiency**, **scalability**, and **security**.

**3. Solution Overview**

I proposed and implemented a **multi-layered data lake** on **Google Cloud Platform (GCP)** based on the Bronze–Silver–Gold architecture:

| **Layer** | **Purpose** | **Tools Used** |
| --- | --- | --- |
| Bronze Layer | Raw data ingestion | Cloud Storage, Pub/Sub |
| Silver Layer | Cleansed and standardized datasets | Dataflow |
| Gold Layer | Curated, business-ready datasets | BigQuery |
| Visualization | BI Dashboards | Looker Studio |

**4. Step-by-Step Implementation**

**4.1 Environment Setup**

* Created a dedicated GCP project.
* Enabled services: Cloud Storage, Pub/Sub, Dataflow, BigQuery, Looker Studio.
* Set up billing alerts and IAM roles (Principle of Least Privilege).

**4.2 Data Ingestion (Bronze Layer)**

**Batch Sources:**

* Shopify orders (CSV exports) → uploaded daily to Cloud Storage.
* Inventory reports (XLS/CSV) → scheduled ingestion into Cloud Storage.

**Streaming Sources:**

* Web clickstream logs → published in real-time to Cloud Pub/Sub.
* Google Ads data → streamed via Pub/Sub topics.

**Buckets and topics were structured by source and data type for organized ingestion.**

**4.3 Data Transformation (Silver Layer)**

* Created **Dataflow pipelines** to:
  + Parse raw data formats.
  + Standardize date formats, remove duplicates.
  + Enrich data (e.g., mapping user IDs, product categories).
* Batch and streaming pipelines were optimized for autoscaling based on load.

**Intermediate cleaned data was stored in:**

* Cleaned Cloud Storage folders (Silver Layer).
* Staging tables in BigQuery.

**4.4 Data Curation (Gold Layer)**

* Designed BigQuery datasets and tables for:
  + Sales metrics
  + Inventory tracking
  + Customer behavioral patterns
  + Marketing campaign effectiveness
* Applied **partitioning** (by date) and **clustering** (by region, product category) for optimized performance.

**4.5 Visualization and Reporting**

* Built real-time dashboards using Looker Studio connected to BigQuery datasets.
* Created dashboards for:
  + Daily sales and returns
  + Top-performing products
  + Customer lifetime value segmentation
  + Marketing spend vs revenue (ROAS)

**Dashboards refresh automatically every 15 minutes based on new data ingested.**

**5. Monitoring and Maintenance**

* Set up **Cloud Monitoring** to track:
  + Dataflow job health and latencies
  + Pub/Sub message backlog
  + BigQuery costs and query performance
* Alerts configured for:
  + Failed ingestion jobs
  + Budget threshold exceedance
  + Storage anomalies

**6. Security and Cost Management**

* Enforced **IAM roles** per principle of least privilege.
* Data encryption was enabled both **at rest** and **in transit**.
* Budget alerts ensured that monthly GCP costs stayed within forecasted limits.
* Audit logs enabled to maintain full traceability of actions.

A diagram of a cloud storage system

AI-generated content may be incorrect.

Note: This diagram represents a layered data lake architecture on Google Cloud. Raw data from sources like Shopify and CSV files is first ingested into **Cloud Storage (Bronze Layer)**. Using **Dataflow**, the data is cleaned and transformed, then stored in the **Silver Layer** (structured and semi-processed data). The refined data is then loaded into **BigQuery (Gold Layer)** for analytics and reporting. Finally, **Looker Studio** connects to BigQuery to create interactive dashboards and visualizations. This structure ensures data is securely stored, systematically processed, and easily accessible for decision-making, providing a scalable and efficient foundation for data-driven business insights.

**7. Project Outcomes**

| **Before Implementation** | **After Implementation** |
| --- | --- |
| Siloed raw data in local systems | Unified and structured cloud data lake |
| Manual Excel-based reporting | Real-time automated dashboards |
| No real-time insights | Streamed analytics updated every few minutes |
| Poor data quality and validation | Clean, standardized, queryable datasets |

**8. Challenges and Solutions**

| **Challenge** | **Solution** |
| --- | --- |
| Handling late-arriving streaming data | Implemented windowing and watermarking in Dataflow |
| Scaling ETL during flash sales | Enabled autoscaling in streaming pipelines |
| Organizing multiple source schemas | Created standardized naming conventions and metadata tracking |

**9. Technologies Used**

| **Component** | **Technology** |
| --- | --- |
| Cloud Storage | GCS Buckets (Bronze, Silver) |
| Real-Time Streaming | Pub/Sub Topics & Subscriptions |
| ETL Pipelines | Apache Beam (via Dataflow) |
| Data Warehouse | BigQuery |
| Dashboards | Looker Studio |
| Monitoring | Cloud Monitoring & Alerting |

**10. Future Enhancements**

* Automate all pipelines orchestration using **Cloud Composer** (Apache Airflow).
* Integrate **Vertex AI** for predictive analytics.
* Implement **Data Catalog** for data lineage and discoverability.

**Final Note**

This project not only provided StyleNest with a real-time data backbone but also prepared them for future growth by building a scalable, auditable, and analytics-ready environment using Google Cloud Platform best practices.