Google Dataflow Use Case

Use Case: Real-Time and Batch Data Processing with Google Dataflow

GENERAL CHARACTERISTICS

Intent	To enable scalable, real-time, and batch data processing using Google Dataflow.
Scope	Implementation of a data processing pipeline for a client to handle large-scale data streams and batch processing.
Level	System-level.
Client	Confidential (Financial Services Provider).
Last Update	03/12/2024
Status	Finalized.
Stage	Implementation and Optimization.

ACTORS

Primary Actor	Data Engineer.
Secondary Actors	Data Analysts, Business Intelligence Teams, IT Operations Team.

PREREQUISITES

Static Preconditions	- Google Cloud Project set up with Dataflow API enabled. - Data sources identified, such as message queues, databases, and storage systems.
Dynamic Preconditions	- Data pipeline architecture designed and validated. - Dependencies packaged for deployment with Dataflow.
Assumptions	- Client requires near real-time insights and periodic batch processing. - Data volumes are high and require scalable processing capabilities.

TRIGGERS

Trigger Event	The client needed a solution to process large streams of transactional data in real-time and batch modes for fraud detection and compliance reporting.

EXPECTED OUTCOME

Success Postcondition	- Data pipelines handle real-time and batch processing seamlessly. - Insights are delivered with low latency, and compliance reports are generated on

	time.
Failed Postcondition	- Processing delays lead to missed opportunities for fraud detection and compliance violations.

OPERATIONS AND CONCEPTS

Operations	1. Designed a unified pipeline to process both real-time and batch data using Apache Beam. 2. Set up Pub/Sub as the message ingestion service for real-time data streams. 3. Integrated with BigQuery for storing processed data and enabling analytics. 4. Configured Cloud Storage for batch data ingestion and archival. 5. Deployed the pipeline on Google Dataflow for auto-scaling and managed execution. 6. Monitored and optimized pipeline performance using Dataflow Monitoring and Logs Explorer.
Concepts	- Dataflow: A fully managed service for stream and batch data processing. - Apache Beam: A unified programming model for building data processing pipelines. - Pub/Sub: A messaging service for real-time data ingestion.

MAIN SUCCESS SCENARIO

Step 1	Analyzed the client's requirements for real-time and batch data processing.
Step 2	Designed an Apache Beam pipeline to unify stream and batch processing workflows.
Step 3	Set up Pub/Sub for ingesting real-time transactional data.
Step 4	Configured Cloud Storage for ingesting and archiving batch data.
Step 5	Deployed the pipeline on Google Dataflow for managed execution and scalability.
Step 6	Integrated BigQuery for analytics and compliance reporting.
Step 7	Monitored and optimized the pipeline for low-latency processing and cost-efficiency.