Data Observability Strategy for Real-time Monitoring and Anomaly Detection

Data observability monitors data as it flows through complex systems. Observability enables organizations to quickly detect and diagnose issues with their data pipelines, ensuring that data is accurate, consistent, and reliable.

Here are some of the challenges of data observability in real time:

Complexity

Real-time monitoring requires organizations to monitor vast amounts of data at blistering speeds.

Scaling

As data grows, monitoring and anomaly detection becomes harder, and organizations must invest in scalable infrastructure to support observability.

Data Quality Issues

Anomaly detection relies on accurate and consistent data, which is tough to accomplish in complex systems.

Cost

Implementing a holistic observability strategy isn't cheap, especially when real-time monitoring enters the picture.

The Backstory

We played a vital role in developing the data observability and error-logging strategy for a leading global life science company's messaging gateway system. With our top-notch implementation of data observability practices, we could closely monitor data pipelines in real time and quickly detect anomalies and potential issues before escalation. Our innovative approach allowed for proactive action and prevented any serious mishaps. It also made data management more efficient and effective at the global life sciences company.

The Challenge

Given the large and complicated data sets at the global life science company, we needed to ensure efficient error handling and data observability to track and address data pipeline issues of the time. By logging errors, we could identify the root cause of a problem, implement necessary changes to prevent future occurrences, and improve the system's overall performance.

The Solution

We leveraged several Google Cloud services for our project, including Google Kubernetes Engine for hosting internal services like translation and integration, BigQuery for storing patient and institution data, Dataproc for creating and resolving data pipeline issues, Logs Explorer for monitoring data pipelines and deploying hotfixes, GCS Buckets for storing data and scripts, and Pub/Sub for triggering pipeline re-runs via notifications.

The Results

Thanks to our data observability solution, we managed to slash the time it took to identify and fix errors. Previously, finding and resolving issues took five days, but now it only takes one. Furthermore, we reduced partner onboarding times from an average of three weeks to just one week. These improvements allowed the global life sciences company to onboard 19 new partners and resolve 1,350+ issues.

egen