

How to Process and Analyze Audit Logs with Delta Lake and Structured Streaming

Craig Ng | Miklos Christine

Agenda

- Why are audit logs important?
- Why Delta Lake?
- Why Structured Streaming?
- Demo

Why are audit logs important?

- Databricks audit logs are the main way for admins to track any behavior on the platform
 - Efficient use of resources
 - Malicious activity
 - User engagement
- Combined with cloud provider logs, they provide a full picture of usage patterns in the platform and underlying infrastructure

Why Delta Lake?

- Allows us to gracefully handle schema evolution, specifically with the **requestParams** field (an evolving struct)
- Provides Schema Validation on Write instead of on Read
 - Removes the risk of schema inconsistency
- Take advantage of specific performance optimizations like **OPTIMIZE** to maximize read performance

Why Structured Streaming?

- State management, state management, state management
- We can utilize Structured Streaming's checkpoints and write-ahead log to ensure that we're only processing the newly added audit log files
- We designed our streaming queries as trigger once daily jobs which are like pseudo-batch jobs
 - Blog explaining the benefits of Trigger Once to reduce costs

DEMO

