DATA ECONOMY

Cloud Infrastructure Observability (Data Lake)

Category: **Data Integration/Data Engineering**

Table of Contents

About the Customer

Customer is one of the USA's oldest and largest financial institutions, with $218.3 billion in assets as of June 30, 2025. Headquartered in Providence, Rhode Island, the bank offers a broad range of retail and commercial banking products and services to individuals, small businesses, middle-market companies, large corporations and institutions. In Commercial Banking, the group offers a broad complement of financial products and solutions, including lending and leasing, deposit and treasury management services, foreign exchange, interest rate and commodity risk management solutions, as well as loan syndication, corporate finance, merger and acquisition, and debt and equity capital markets capabilities.

The Challenge

Customer faced significant **challenges in maintaining real-time visibility across its AWS Data Lake environment**. Monitoring was fragmented across multiple cloud services, making it difficult for operations teams to **quickly identify and resolve performance bottlenecks**. The lack of an integrated observability framework led to delays in **detecting pipeline failures, resulting in missed SLAs for business reporting cycles and increased operational risk**.

Pain Points

01 Fragmented monitoring across EMR, Redshift, EC2, S3, and Talend workloads.

02 Limited reliability and traceability of data pipelines

03 High mean time to detect (MTTD) and mean time to recover (MTTR) incidents

04 Lack of unified metrics and transparency for leadership reporting

05 Rising costs associated with commercial monitoring tool

06 Absence of automated alerting and proactive health checks

Moderization Objective

To address these challenges, Customer embarked on an observability modernization initiative aimed at establishing end-to-end visibility and reliability across the AWS Data Lake stack. The goal was to build a cost-effective, metrics-driven, and automated monitoring framework that would:

01 Provide real-time insights into workload health and performance

02 Enable proactive detection and resolution of issues before they impacted reporting cycles

03 Reduce dependency on expensive third-party tools through native AWS observability services

04 Empower leadership with actionable dashboards for data-driven decision-making

Partner Solution

DataEconomy implemented a unified Data Lake Observability Framework combining AWS native services with open-source technologies to provide real-time visibility across metrics, logs, dashboards, and alerts. The solution enabled proactive monitoring, faster issue resolution, and cost optimization.

High Level Architecture Diagram of Data Lake

Key Components

Data Sources
EMR, EMR Serverless, Redshift, RDS, Talend, and S3 generate system and application telemetry.

Data Collection
Alloy agent and custom Python scripts collected metrics and logs from CloudWatch and S3 and written to Mimir and Loki

Processing & Insights
Correlation between metrics and logs is done through common labels and visualized through Grafana dashboards

Dashboards & Alerting
Aggregated and detailed dashboards are created in Grafana for each of the AWS services along with automated threshold and SLA alerts to send email notifications

Root Cause Analysis
Metrics and logs are correlated in Grafana to accelerate RCA and reduce recovery times

Design Principles and Implementation highlights

The DataEconomy Observability Framework was built on five key architectural principles to ensure long-term adaptability, operational efficiency, and resilience across the AWS Data Lake ecosystem.

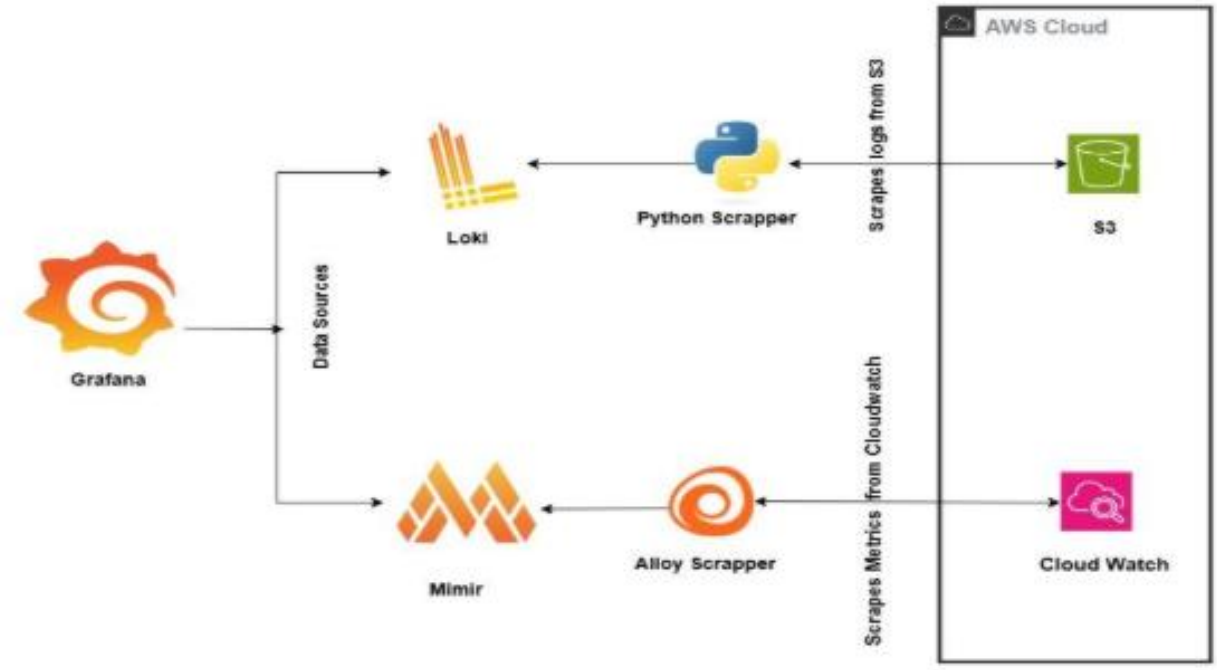
Principle	Implementation Example
Flexibility	• Integrates metrics/logs from CloudWatch, Mimir, Alloy & S3 • Supports hybrid/multi-cloud observability in future • Grafana templating for self-service dashboards • Decoupled architecture and new tools plug-in easily
Automation	• Alloy automates scraping - no manual exports • Continuous metrics collection with no downtime • CloudWatch auto-discovers new resources • Alerting automation triggers incident workflows
Scalability	• Mimir horizontally scales to unlimited ingestion • Loki index-less log model low-cost retention • S3 near-infinite scalable storage • Grafana scales users/dashboards without re-architecture

Principle	Implementation Example
Security	• IAM least-privilege controls • S3 encryption at rest + TLS encryption in transit • Grafana RBAC restricts sensitive data exposure • Audit-ready access governance
Cost Optimization	• Fully open-source stack - zero monitoring licensing • S3-tiered retention reduces storage expense • Metric retention policies optimize cost

As part of the DataEconomy Observability initiative, a lightweight, open-source monitoring pipeline was built to aggregate and visualize metrics and logs from AWS services in real time. The design eliminates fragmented visibility, reduces dependency on commercial tools, and provides a scalable foundation for analytics and alerting.

High Level Observability Platform Framework

The framework integrates AWS CloudWatch and Amazon S3 with Grafana, Mimir, and Loki through custom data scrapers to deliver unified observability across the Data Lake stack.



Principle	Implementation Example
Data Collection & Ingestion	• CloudWatch metrics ingested into Mimir via Alloy agents for long-term analytics • S3 and CloudWatch logs streamed to Loki using custom exporters • Continuous data ingestion from 15+ AWS services for complete platform coverage
Performance & Pipeline Monitoring	• Observability for EMR, EMR Serverless, Talend, and Redshift workloads • Tracking of latency, throughput, failures, and resource utilization to maintain reliability
Visualization & Insights	• Real-time Grafana dashboards offering a unified system and KPI view • Custom drill-downs for correlating workload performance with data pipelines
Monitoring, Alerting & RCA	• Automated alerts for SLA breaches with rapid root-cause correlation between metrics and logs • Improved detection and recovery times across production workloads
Automation & Security	• IaC-based deployment for faster rollout and consistency. • IAM + RBAC controls ensuring secure, least-privilege access to observability data
Technology Stack	• AWS - CloudWatch, S3, EC2, Redshift, EMR, EMR Serverless • Open Source - Grafana, Mimir, Loki, Alloy
Pipelines	• Metrics Pipeline -The Alloy Scraper collects service and infrastructure metrics from AWS CloudWatch, pushing them into Mimir for high-performance storage and time-series analytics • Logs Pipeline - Application and system logs stored in Amazon S3 are extracted using a custom Python Scraper and streamed into Loki for log indexing and correlation
Visualization Layer	• Grafana serves as the central observability console, connecting to both Mimir and Loki as data sources to display real-time dashboards, trends, and alert metrics
Dashboard Components	• Service Health overview (EMR, EMR Serverless, RedShift) • Resource Utilization Trends (CPU, memory, disk, network) • Anomaly Detection (ML-based alerts for deviations from baseline) • Failure rate and Error tracking (spark/hive job failures, Redshift query errors) • Cost optimization insights (idle clusters, underutilized resources) • Alerting Strategy – proactive alerts, severity-based notifications and auto-remediation triggers

Core Services Being Monitored

AWS EMR

- Cluster health (master/core/task nodes)
- YARN resource utilization (memory, vCPU)
- HDFS/EMRFS storage metrics
- Job failures/retries & step execution times
- Spot-instance interruptions (if applicable)

EMR Serverless

- Application startup times & failures
- Driver/executor resource usage
- Job duration trends & bottlenecks
- Throttling/API limit alerts

RedShift

- Query performance (long-running queries, queue time)
- WLM (Workload Management) slot utilization
- Storage growth & disk space alerts
- Connection limits & concurrency issues

Results and Benefits

The modernization of Customer's Data Lake observability framework delivered measurable improvements across reliability, productivity, and cost efficiency. By integrating AWS native services with open-source observability tools, the organization transitioned from reactive incident management to proactive operational governance.

Operational & Reliability

- Unified, real-time observability across all AWS Data Lake components
- Rapid detection and recovery from pipeline or infrastructure anomalies
- Improved SLA performance with reduced incident escalation cycles

Productivity Enhancement

- Troubleshooting accelerated by 5× through correlated metrics and logs
- Leadership gained transparent visibility into system health and KPIs
- Automated dashboards eliminated manual performance reporting

Cost Optimization

- 100% elimination of legacy APM licensing costs.
- Cost-efficient, high-volume log retention using S3 + lifecycle policies.
- Mimir/Loki scaling achieved with no hidden cost expansion

Customer successfully transitioned its cloud operations from reactive firefighting to proactive governance. The modernized observability framework improved reliability, accelerated recovery, and reduced total cost of ownership, enhancing transparency, resilience, and business confidence across the data-driven ecosystem.

About the Partner

Data Economy is a data-driven consulting and technology services organization helping enterprises accelerate their digital transformation through modern data, analytics, and AI solutions. As an AWS Advanced Consulting Partner, we specialize in designing and implementing scalable, secure, and cost-effective data platforms on AWS that unlock actionable insights and deliver measurable business value. Our expertise spans data modernization, data lakes, analytics engineering, machine learning, cloud migrations, SRE and generative AI, enabling organizations to transform raw data into strategic assets. We bring strong experience in AWS-native services such as Amazon Redshift, AWS Glue, Amazon Athena, Amazon EMR, Amazon SageMaker, and AWS Lake Formation ensuring performance, governance, and agility across the data lifecycle. At Data Economy we combine deep industry knowledge with cloud-native best practices to help clients build next-generation data ecosystems. Our solutions empower enterprises to make real-time, insight-led decisions, improve operational efficiency, and drive innovation at scale