

acceldata

Customer Case Study

**Global Information Provider Replaces
Legacy Vendor with Acceldata to Scale and
Automate Data Quality and Reliability**

Customer Case Study

The Customer

A multinational provider of data and AI-driven insights to more than 90 percent of Fortune 500 companies. The 6,000-employee company is a long-time leader in this space, and reaps \$2 billion in annual revenue. It boasts a massive repository of comprehensive information on 460 million organizations around the world which it provides to its clients in almost all industries.

Its Legacy Data Infrastructure

The company's multi-petabyte database stores hundreds of millions of business records in AWS EMR's cloud big data platform. That data is gathered from directly-sourced public data, proprietary sources, and external business activity indicators. Its team of data scientists combine this data with AI models to create ratings and scores that can intelligently guide strategies and decisions. For its business customers, these cloud-based insights can help accelerate sales, anticipate and alleviate procurement risk, prevent compliance penalties and brand damage, and more.

Data is the foundation of this company's business. To protect and manage it, the company had long used software from a legacy data integration vendor. However, the company was growing and modernizing, and their legacy solution could not keep up. For instance, it had no global view into or control over data quality, as it had to perform its DQ checks for each individual application. The solution's data cataloging features were also not up to snuff. It was also having trouble scaling and providing automation features that the company needed.

Why it Chose Acceldata

The company chose Acceldata to provide multi-dimensional data observability from the cloud. In particular, it valued Acceldata's ability to:

1. Catalog and create metadata for all data sources and data files
2. Write all data quality and data validation rules using SQL, Python or Spark
3. Run data quality and data validation rules on both static and streaming data
4. Detect schema, and attribute-based changes
5. Track all rejected or bad records and send alerts/reports based on data
6. Handle reference data management requirements
7. Handle large data volumes, such as GSRL7 data files with 100+ million records

The company is deploying Acceldata's data performance monitoring module on 200 data nodes. This will help eliminate bottlenecks and outages, enable single-click scalability, and optimize the price-performance of data workloads. Acceldata will ingest operational metrics from the company's data ops platform, Streamsets, and link it to the data set or data source.

The company is also deploying Acceldata's data reliability solution to automate monitoring, validation, and remediation of more than 4 PB of data as it flows through the company's data pipeline. This will push metadata into the company's cloud data access control solution, Immuta, so that it can apply data access policies.



Building a Brand New Data Architecture

Acceldata will play a key role in the company's transformed internal data supply chain. All data will be validated and cleansed by Acceldata before it enters the data supply chain. As data is extracted from GSRL7 files and other source buckets, it is checked to see if it already exists on AWS. If data is rejected for being duplicative or for other quality reasons, an alert will be sent to the company's data engineering team.

If the data passes this second validation/quality check, then it will be uploaded to a central data lake connected to the company's big data and analytics platforms, including AWS EMR, Databricks, and Spark. Here, the data is refined before being made available to internal users and customer-facing data applications. Data and job latency, cluster performance, and price performance are also all tracked and optimized by Acceldata.

“ Acceldata will play a key role in the company's transformed internal data supply chain. All data will be validated and cleansed by Acceldata before it enters the data supply chain.

Acceldata integrates strongly with modern data platforms such as Databricks, Spark on Kubernetes, Snowflake, and more. We have helped many other petabyte-scale customers improve their data operations, including Walmart fintech subsidiary, PhonePe, advertising tech firm Pubmatic, and Asian communications provider True Digital. The Acceldata Data Observability Platform helps companies improve data quality at scale.

Get a demo

Get a personalized demo [here](#).