acceldata

How to Improve Data Performance with Multidimensional Data Observability

The Complicated Mess That Is Most Modern Data Operations

Your data operations used to be simple.

There was your on-premises database, which was the single source of truth for all your company's traditional quantitative data, such as sales figures, inventory levels, and financial results.

From that, data would be ingested — think overnight batch jobs — into data warehouses using extract, transform and load (ETL) and other basic data pipeline tools.

Once complete, small teams of Business Intelligence (BI) analysts generated weekly, monthly or quarterly reports for their bosses. More rarely, they would tweak and update 'live' executive dashboards, albeit with data that was already out of date.

The Data Stack Evolution

The situation started to change fifteen years ago. Slowly at first, and then, over the past few years, extremely rapidly.

Data supply and demand exploded.

All of a sudden, there were a massive number of new data sources: social media, IT logs, IoT sensors, real-time streams, and many more.

Enterprise applications including new, interactive data visualization tools made data more accessible.

Business stakeholders began to clamor for real-time, embedded analytics to inform increasingly sophisticated business decisions.

Because of their mission-critical, real-time nature, lines of business began asking for Service Level Agreements (SLAs) with vendors to ensure data flowed consistently and reliably at a fast rate.

To enable this major digital transformation, new employees were brought on board specifically to caretake, analyze, and ensure smooth data operations.

Many became part of a new data operations team. "Data" started showing up in job titles: Data architects. Data engineers. Data stewards. And more.

Data operations teams by and large embraced new cloud-native storage and compute engines and repositories, such as data lakes.For many enterprises, the cloud was more scalable, efficient, and cost effective than on-premise servers and data centers.

However, to keep existing operations humming, new cloud implementations were integrated with legacy on-premise relational databases and warehouses. This hybrid architecture created greater complexity and continues to do so even now.

- **ELT (Extract, Load, Transfer)**: ELT uses a data warehouse to perform basic transformations. Data does not need to be staged. It should be noted that this is different from ETL (Extract, Transform, Load), which moves data from staging into a data warehouse.
- **CDC (Change Data Capture)**: This is a process that tracks data changes that occur in a database or data warehouse.
- **API (Application Programming Interface)**: An API is a set of functions that enables one application to access features, data, or functionality from another application, operating system, or service.
- **Event Streaming:** This is a process where an application or other source can ingest events at high volume from various data sources.

The Modern Data-driven Organization

Figure 1: *The Modern Data-driven Organization*

Data operations teams now oversee the building and management of a complicated, fragile set of data pipelines employing a wider variety of technologies than ever – ELT, CDC, APIs, event streaming, and more. These all interface with an increasingly demanding set of analytics applications, including predictive analytics and AI.

Modern Problems

This complexity and scale create a whole host of challenges for your data operations team.

Data quality can deteriorate due to structural changes, such as schema drifts.

The cost of storing and processing data can escalate beyond control or oscillate unpredictably.

To guarantee high availability and performance, you need visibility and control over your complex infrastructure of on-premise and cloud data storage connected to spider-webs of data pipelines.

There is no shortage of data monitoring tools on the market. Some are bundled with or work only for a particular data platform, such as Hadoop, and only offer partial visibility into your data infrastructure.

Without a single-pane-of-glass over all data activity, however, data ops teams are forced to adopt multiple monitoring solutions. They use these to analyze and reconcile large volumes of patchy, conflicting data. It typically requires significant manual work and time.

The bigger issue is that data monitoring solutions lack the strong management, predictive analysis, and automation capabilities needed for efficient performance management.

Visibility, however, is only one small piece of the data observability puzzle. Without multidimensional data observability, you may identify performance problems, but not until it's too late. Fixing issues after they've reared their ugly head requires extensive manual work by your data operations and IT teams, and makes guaranteeing high availability and performance for your now-mission-critical analytics applications nearly impossible.

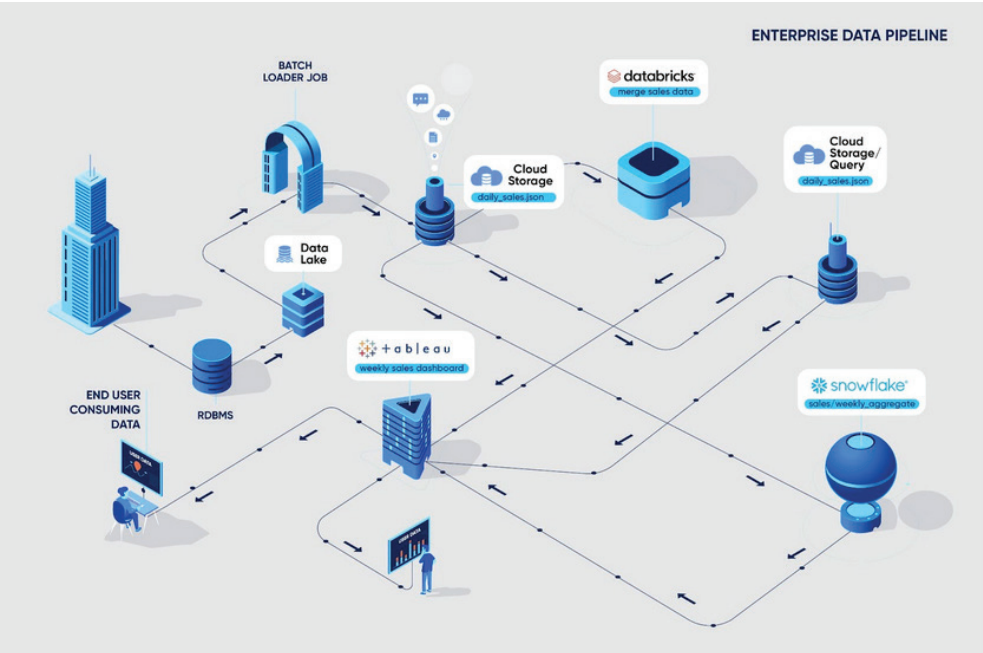

Figure 2. *Enterprise Data Pipeline*

...With Inadequate Solutions

Don't Application Performance Management (APM) tools provide the power to manage data performance?

APM tools offer a form of observability; they can monitor applications, diagnose performance issues within those applications, and, when aided by AI and machine learning, even predict problems before they occur.

Some APM solutions can even remediate problems automatically, or provide simple solutions for IT teams to implement.

However, because of APM's broader nature and ultimate focus on applications rather than data and data pipelines, none of them provides the in-depth visibility, control, predictive analysis, and time-saving automation that data operations teams need to efficiently maintain high data performance.

What's the result? Data operations teams find themselves plagued with bottlenecks, slowdowns, and failures. Responding to performance issues becomes a daily firefighting ritual.

Data ops team members then find themselves in conflict with their business-side peers, for whom these analytics applications have become mission-critical. And that leads to blame and finger pointing all around.

Besides failing to meet performance SLAs, the data ops teams have no time for longer-range projects that can add scale and new capabilities that are critical to meet business requirements and strategic objectives.

The Solution: Multidimensional Data Observability

Data observability is an all-encompassing approach to monitoring and managing today's complex data storage and pipelines. It helps predict, prevent and automatically resolve data performance problems and their associated AI and analytics workloads. It uses automation and machine learning to correlate thousands of events across multiple servers, nodes, clusters, containers and applications.

Data observability provides a 360-degree view of data at rest and data in motion, data processing, and the pipelines through which data travels. It creates solutions and provides optimizations to keep your data performance accurate, consistent, and available, so data operations and engineering teams can easily meet all of their SLAs.

In other words, data observability enables data ops teams and IT engineers to:

- **Observe:** Detect patterns of potential problems (including unknown unknowns) across complex data environments by analyzing the external outputs of data operations, including performance metrics, metadata, utilization, and more.
- **Analyze:** Infer from observations how to make data performance more reliable, scalable and cost effective.
- **Act:** Once you know what the issues are, you can manually take action or employ automated actions to fix problems before they impact performance, costs, and business results.

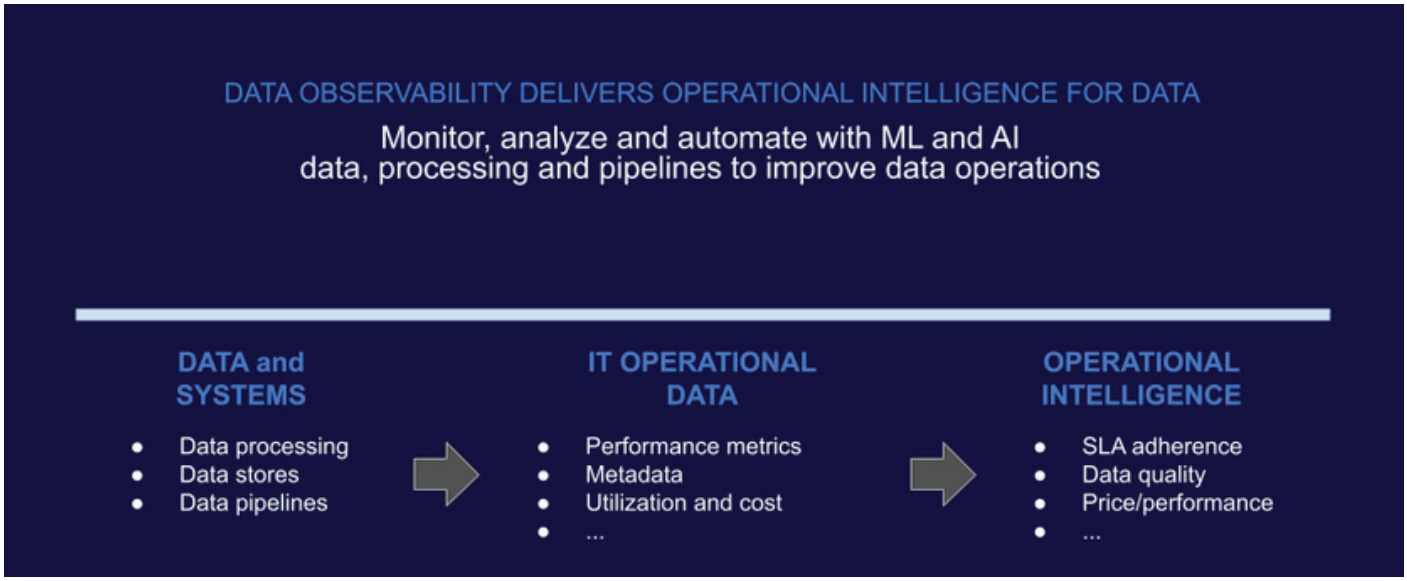

Figure 2. *Enterprise Data Pipeline*

A Layered Approach

Data observability addresses specific performance needs at each layer of the stack:

1. **Infrastructure layer —** platform engineers, devops engineers, and site reliability engineers (SREs) can monitor storage and compute availability, utilization, performance, and their impact on data flows. Effective monitoring enables data teams to correlate activity with any data or analytics pipeline issues, and prevent performance hiccups or system outages.

2. **Data layer —** data architects and engineers can monitor database and network applications such as Apache Spark and Apache Kafka that underlie data and analytic pipelines. This enables them to remediate any processing or network delays and prevent SLA failures.

3. **Application layer —** BI/data analysts, data scientists, and business managers use data observability to understand root causes of data-related performance issues where APM tools fall short.

Eight Key Use Cases

Let's look at eight key scenarios where multidimensional data observability can help performance.

1. **General performance management —** monitor systems and components to measure memory, CPU/storage consumption, and cluster/node status. Define alert thresholds and notifications. This level of granularity is key to help your engineers identify data congestion, outages, and runaway users or applications. It's also key for troubleshooting and remediation performance issues.

2. **Infrastructure streamlining —** it's common for the same small set of data to be used over and over. Data observability can illuminate when such "skew" occurs. Files that are not accessed often can be archived in less expensive, slower storage, while "warm" data can be stored in faster, premium tiers, to optimize performance.

3. **Capacity planning —** meeting SLAs may not be difficult if you have an unlimited budget. Data observability gives infrastructure engineers the monitoring and management tools to maintain strong performance and

availability at the lowest cost. Machine learning analysis of historical performance and capacity data helps calculate future capacity requirements, while eliminating the need to overbuy.

4. **Data pipeline performance —** collect thousands of pipeline events and analyze them for spikes and other anomalies to help data architects and engineers track and predict the performance of data in motion. Such tools also recommend ways to tune and optimize performance to improve reliability and reduce costs.

5. **Data quality —** sending poor-quality data is as bad as a data bottleneck. However, over time, data can "drift" and lose accuracy, completeness, and consistency. Data observability tools can inspect data transfers for these factors, creating rules to compare source and target data, and flagging mismatches for further review.

6. **Architectural design —** to get out of the trap of daily firefighting, data architects and engineers must design better architectures. Data observability tools can provide recommendations into past and predicted pipeline performance and utilization in order to enable successful redesigns.

8. **Service Level Agreements —** Data observability arms both users and data engineers with the information to agree on the most accurate capacity estimates, the fastest, most reliable data pipelines, and most realistic SLAs.

9. **Cost modeling and chargeback —** enable line of business and IT leaders to come up with precise cost estimates for their cloud and on-premises infrastructure that meet their SLAs.

Accelerate Your Data Performance

Acceldata is the only true multidimensional data observability cloud on the market. It gives enterprises comprehensive, cross-sectional visibility into the complex modern data stack across hybrid data-lakes and warehouses.

We deliver a single pane of glass that correlates events across components and across the entire data system to predict, identify, and fix critical performance issues across every stage of the enterprise data journey. Acceldata observes data pipelines, validates data quality, and provides deep computational insights to ensure reliable, high-performing data operations.

All of this enables Acceldata customers to measure what matters most and provide visibility to re-design, re-implement, and tune data pipelines and applications to deliver on your SLAs.

"Acceldata supports our hyper-growth and helps us manage one of the world's largest instant payment systems. PhonePe's biggest-ever data infrastructure initiative would never have been possible without Acceldata."

Burzin Engineer
Founder & Chief Reliability Engineer
PhonePe (Walmart)

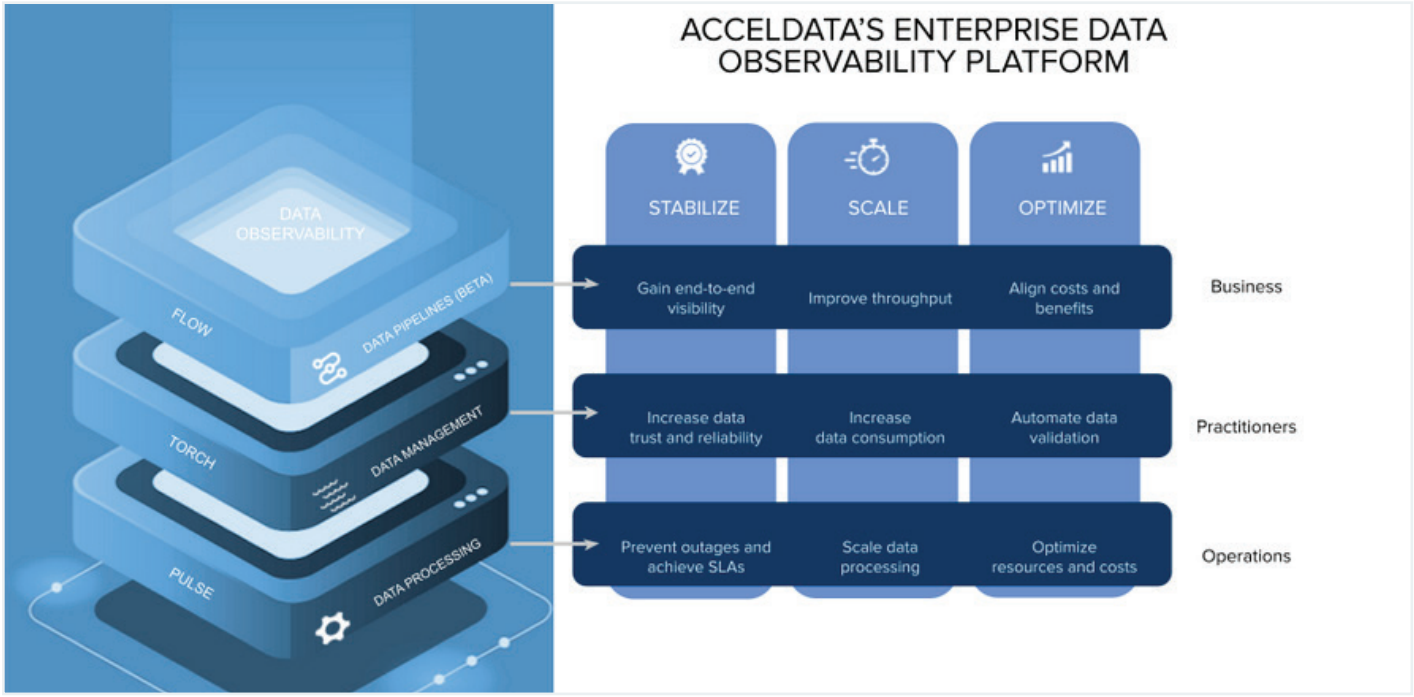

Figure 4. *Enterprise Data Observability Platform*

Acceldata comes in the form of three products, all of which have a key role in boosting and maintaining data performance.

Acceldata Pulse allows you to monitor all of your connected data sources against user-defined metrics as data flows throughout your infrastructure, data, and application layers. A rich, customizable visual dashboard makes monitoring efficient and easy. Pulse helps enterprises transition from resolving and troubleshooting incidents reactively to predicting and preventing incidents before they occur. Eliminate unplanned outages and scale your workloads while meeting SLAs and saving money with Pulse.

With Acceldata Pulse, you can run simulations to model the least-costly configuration that handily meets your SLAs.

"We partnered with Acceldata while we were on Hadoop. As we re-platformed, Acceldata enabled us with deep technical insights and allowed us to modernize seamlessly. We areexpanding the use of Acceldata across our data ecosystem for observability at GE's scale and complexity."

Diwakar Goel
Chief Data Officer and VP,
GE Digital

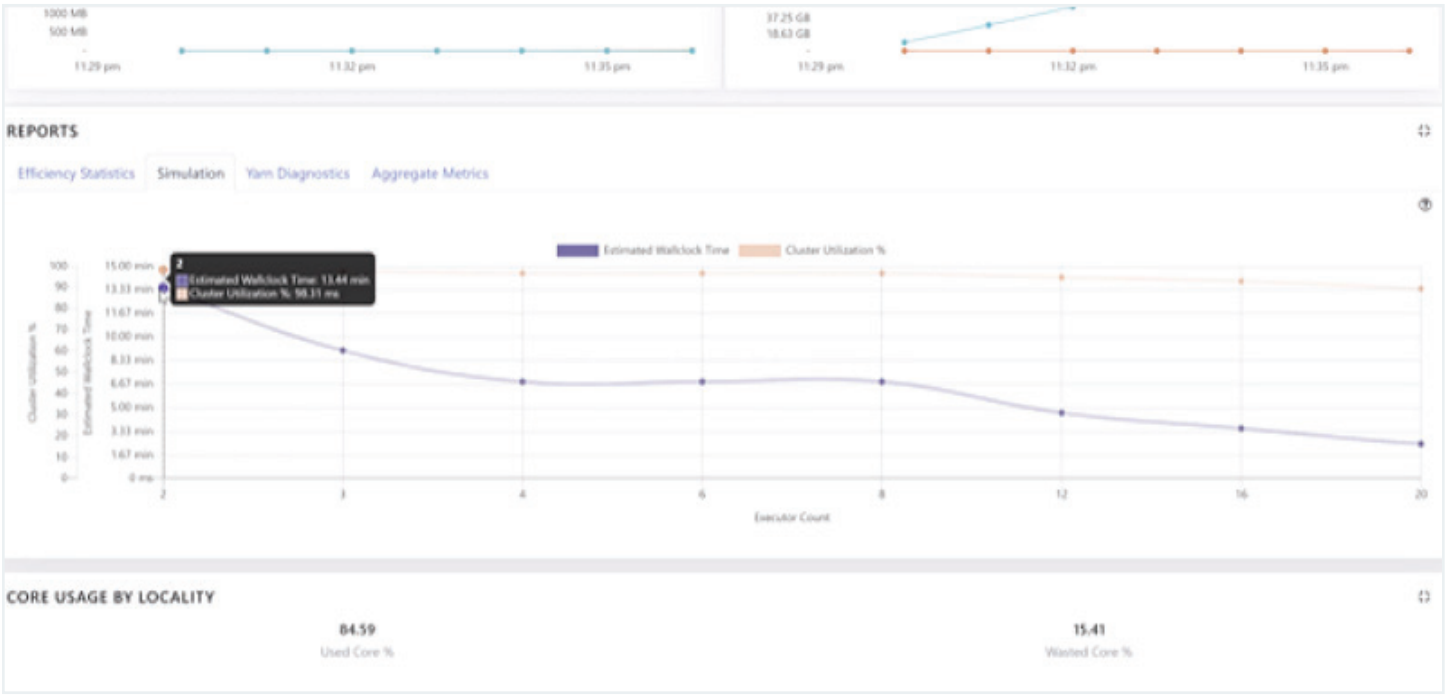

Figure 5. *With Acceldata Pulse, you can run simulations to model the least-costly configuration that handily meets your SLAs.*

Acceldata Torch eliminates data downtime by ensuring data reliability and quality. It continuously and intelligently validates quality, and lineage of structured, semi-structured, and unstructured data. Torch uses machine learning algorithms to predict data outages, report data issues in real-time, and prevent data errors across on-premises and cloud data warehouses, data lakes, and other data technologies.

Acceldata Flow builds upon Pulse to provide the most in-depth data pipeline monitoring. Flow integrates with leading data systems including Apache HBase, Spark, Hive and Kafka, Snowflake, Databricks, MySQL, and many others. This allows Flow to track every data transaction, handshake, and transformation from origin to consumption to ensure reliability and performance in your hybrid data lakes and data warehouses. When bottlenecks occur, you can easily hunt them down, and perform trend analyses to prevent future occurrences. Flow auto-aggregates information from data stores, pipelines, processing engines, and infrastructure into a single view to make monitoring of SLAs and anomaly identification easy.

Get a demo

Get a personalized demo ***here.***