acceldata

Choosing the Right Data Observability Solution:

A Feature Checklist for Complex Data Environments

Exploding data volumes and increased operational demand for data has led to significant complexity in modern data environments. There are no excuses, however, for data teams who do not successfully manage their data and put it to use towards achieving business outcomes. In today's data environments, that means continuous awareness and understanding of data and data systems within an environment, all with the intent of optimizing data supply chains so they deliver reliable, high-quality data when it is needed.

How do modern data teams achieve this? They use data observability.

Data observability can be defined as the ability of an organization to completely understand the health of its data. Put another way, it is a systematic solution to the problem of data complexity. It monitors and correlates data workload events across application, data, and infrastructure layers to resolve issues in production analytics and AI workloads.

Implementing a multidimensional data observability platform, such as Acceldata, is the most effective and efficient approach for data engineers to ensure the health of their data and the optimal operations of their data infrastructure investments. Unlike traditional APM tools that only monitor the organization's application layer, multidimensional data observability platforms provide visibility across your data, infrastructure, and pipelines.

What features should you look for in a data observability platform?

Our checklist (on the last page of this ebook) provides an at-a-glance introduction to nineteen important data observability features—all of which are core to the Acceldata platform. (Request a **demo** to learn more about these capabilities.)

We've categorized these features into three groups:

1 **Data**

2 **Compute performance**

3 **Data pipelines**

Let’s take a quick look at each category.

Data

Bad data leads to bad decision-making. Fixing the bad data problem can be extremely challenging in a business environment that’s plagued by dark, redundant, and cold data. What started out as a pristine data lake can quickly turn into a data swamp—a swamp that few are brave enough to wade through and clean up.

Implementing a multidimensional data observability platform can make it easier to predict, prevent, and resolve data quality issues, especially when it offers:

- Built-in data discovery
- Data quality & reconciliation rules
- Data drift & schema drift detection

Compute performance

Continuously adding more environments, more technology, and more data-driven use cases can lead to an unscalable situation that is both costly and prone to outages. Rising costs paired with frequent downtime creates additional complexity, causes friction with users, and erodes return on data.

A multidimensional data observability platform can help data teams overcome the traditional roadblocks to performance monitoring by providing:

- Trend analysis
- Auto-remediation and alerting
- Root cause analysis
- Configuration recommendations
- Performance simulation
- Workload analysis
- Capacity optimization
- Data temperature & utilization monitoring

Data pipelines

One problematic data pipeline can wreak substantial havoc on an organization’s data quality and efficiency. Unfortunately, identifying the culprit is no small task without the proper transparency into your multi-cloud and hybrid cloud environments.

A reliable data observability solution provides end-to-end visibility, making it possible to trace the flow of data (and the cost of data) across your interconnected systems.
The Acceldata platform makes this possible by delivering:

- Performance analytics
- Pipeline monitoring
- Cost-benefit analysis
- ETL integration
- Flexible API

Take the Next Step Toward Multidimensional Data Observability

Multidimensional data observability offers a scalable approach to monitor, detect, predict, prevent, and resolve issues across your data, processing, and pipelines.

Use the handy checklist on the following page to ensure you’re making the right decisions about data observability.

acceldata

Data Observability Checklist

Multidimensional data observability enables organizations to monitor, detect, predict, prevent, and resolve issues across their data, processing, and pipelines. Acceldata makes this possible by providing:

Data

❑ **Data Discovery**

Create a centralized inventory of data assets across environments and technologies. Enable self-service data discovery with simple faceted search, navigation across similar/elated data assets, and RBAC.

❑ **Data Quality Rules**

Improve data quality and reliability with ML-driven recommendations. Automated data quality rules and alerts make it easier to identify missing data, data type violations, incorrect values and formats, sensitive data, etc.

❑ **Data Reconciliation Rules**

Ensure data arrives as expected and in accordance with your data reconciliation policies. Email notifications keep you informed of rule failures in near real-time.

❑ **Data Drift Detection**

Increase the accuracy of AI/ML workloads by monitoring for unexpected content changes. Data drift rules automatically validate changes against tolerance thresholds for key metrics. Get notified of excessive data drift.

❑ **Schema Drift Detection**

Detect structural changes to schemas and tables that can break pipelines or impact downstream applications. Know when columns are added, modified, and deleted.

Compute Performance

❑ **Trend Analysis**

Predict anomalies and potential problems before they impact your operations. See which jobs are taking longer to run compared to historical norms.

❑ **Auto-Remediation & Alerting**

Prevent slowdowns and outages with prebuilt and customizable runbooks. Automate tuning, cleanup, configuration, and provisioning.

❑ **Root Cause Analysis**

Correlate events across complex environments to quickly resolve incidents. Get to the root cause by considering resource contention, environmental health, and historical comparisons.

❑ **Configuration Recommendations**

Convert historical usage data into actionable insights for optimizing resource configuration, data distribution, and query performance.

❑ **Performance Simulation**

Simulate expected performance and automatically tune systems to minimize resource utilization, increase performance, eliminate waste, and hit SLAs.

❑ **Workload Analysis**

Analyze workload efficiency to identify bottlenecks, redundancies, and performance improvement opportunities.

❑ **Capacity Optimization**

Leverage resource contention analytics to optimize scheduling, simplify planning and chargebacks, and improve your hybrid cloud environment.

❑ **Data Temperature & Utilization Monitoring**

Automatically monitor data temperature and detect “hot spots” that may indicate future issues.

Data Pipelines

❑ **End-to-end Visibility**

Trace the flow of data and cost of data a cross interconnected systems.

❑ **Performance Analytics**

Identify bottlenecks, analyze historical comparisons, and optimize data pipeline performance. Drill down for an in-depth view of data and processing issues.

❑ **Pipeline Monitoring**

Monitor SLAs/SLOs, data schemas, distributions, and business events. Track data transactions, handshakes, and transformations.

❑ **Cost-Benefit Analysis**

Evaluate price and performance trade-offs to ensure scalability and ROI as you make technology decisions.

❑ **ETL Integration**

Out-of-the-box ETL integrations reduce complexity and save your data engineering team time.

❑ **API For Integration**

Integrate existing infrastructure and connect to existing data, processes, pipelines, and applications with a flexible API.