acceldata

The Three Critical Pillars of Data Reliability

Endless streams of data are collected from various sources and then deployed into complex pipelines to move that data to an ever-expanding number of targets. Each layer of technology adds more complexity and with it more potential points of failure. When data doesn't get to its destination on time, it's more than an inconvenience – it can be detrimental to the business. Employees, customers, suppliers, and partners need reliable data to make correct decisions, and without it, they may take actions that result in undesirable outcomes.

The concept of data reliability addresses the challenges of delivering data in increasingly complex systems. It builds on the paradigm of multidimensional data observability, which enables the monitoring and analysis of data operations across complex environments so that data engineers can identify, predict, prevent, and resolve problems.

> **"When data is unreliable, business executives puzzle over inaccurate results, data scientists need to identify alternate data sets, and data engineers must figure out the sources of bottlenecks or corrupted data."**
>
> Bill Schmarzo
> Author, Big Data MBA, Prof Univ. of San Francisco

Data reliability goes beyond performance and classic data quality concerns. Data must be reconciled as it moves across the pipeline. Structural changes ("schema drift") must be monitored which could break pipelines. Data anomalies and trends ("data drift") must also be tracked, either to inform the business or to ensure the accuracy of analytic models.

To ensure data reliability, organizations must be able to monitor pipelines end-to-end, identify early warning signs of data issues, quickly pinpoint underlying causes, and automate preventative maintenance to avoid disruption to the business. To operate effectively in modern environments, organizations must be able to balance coverage of data quality, performance, and cost. Insight into data operations can be used to drive efficiency as well.

The Shift to Real-Time and Continuous Data Analytics

For decades, the handling of business data was a linear process driven by IT. Skilled data engineers constructed fairly simple data pipelines in which extract, transform, and load (ETL) and change data capture (CDC) tools ingested raw data from databases and applications. The data was loaded into data warehouses and processed in pre-scheduled batches, typically hourly or nightly. When an individual business user or department requested a report, data analysts would use business intelligence (BI) software to perform database queries and create a static report— which typically took several days to build— based on historical data and past results. If the requesting individual or department had a follow-up question about the report, it would be placed at the bottom of the request queue and the whole process would start over again.

While this process was straightforward and easy to manage, it also created its share of inefficiencies. Data was typically housed in information silos and relied on skilled data experts to extract and analyze it so it could make sense for business users. These restrictions and the time required to respond to new reporting requests or modifications to a data mart, (often measured in weeks and months) resulted in frustratingly slow reporting cycles that made it infeasible to take advantage of data in real-time.

The explosion of available data over the last two decades has spurred a dramatic shift in how data is used in the enterprise. Thanks to a confluence of technological evolutions, organizations in every industry collect innumerable, and continuously-growing data points from a range of business activities. When data pipelines can scale to meet the volume and complexity of data and its many sources, an organization gets an accurate picture of their business operations. With that insight, data engineers can quickly adapt to capitalize on what they learn. These are usually the organizations that show outpaced returns, based on growth and market execution.

In this environment, businesses have turned to real-time and continuous data analytics to efficiently deliver information and improve business decision-making ability. Real-time data analytics eliminates the time delay between data collection and processing endemic to traditional business intelligence processes. Organizations now have access to data as soon as it's acquired.

As organizations strive to operationalize their data, they are extending access to BI and analytics to every corner of the enterprise. No longer solely the domain of analysts and executives, data is now shared freely with frontline employees, business partners, and customers. Business data analytics are integrated within the full spectrum of business operations to optimize processes, decision-making, customer relationships, and products.

The advantages of real-time and continuous data analytics are clear. Organizations that can mine and analyze data quickly are capable of making faster and better decisions, improving their operational processes, reducing security risks, and ultimately, increasing revenue and digital differentiation in their respective marketplaces.

If data pipelines can scale to serve the volume and complexity of data and it's many sources, an organization can operate from an accurate picture of their business operations and can quickly adapt to capitalize on what they learn.

But as data volume and analytic demands have grown, so too has the complexity of data pipelines and processing engines. Modern pipelines are constructed with a variety of components, including extract/load/transform (ELT) and change data capture (CDC) tools, online transaction processing (OLTP) systems, application programming interfaces (APIs), and event streaming platforms such as Apache Kafka. They ingest structured, semi-structured, and unstructured data from applications, websites, third-party data feeds, social media, databases IT logs, and Internet of Things (IoT) devices. The data is transformed and stored in data warehouses, data lakes, NoSQL databases, and event streaming platforms, where it is analyzed by machine learning algorithms. Insights, in the form of dashboards and reports, are now delivered via self-service BI platforms such as Snowflake.

Instead of being run in on-premises environments using commodity hardware, modern data pipelines leverage the scalability and cost efficiencies of the cloud. This requires that cloud object stores and compute engines integrate with legacy on-premise systems, further increasing complexity. The result is typically a brittle data pipeline that must feed data to an ever-expanding set of targets, including BI tools, artificial intelligence (AI), and embedded analytics workflows.

> **Thousands of traders manage over $35T of assets on our platform. We can't afford a second of downtime for insights from our data.**
>
> CDO, BLACKROCK

Data Reliability Defined

Data reliability seeks to ensure the dependable delivery of quality data, on-time processing, and end-to-end pipelines at scale.

It's important to differentiate data reliability from data quality. Classic data quality is a measurement of how fit a particular data set is at meeting the needs of its users. Data is considered of high quality when it satisfies a range of requirements, some of which include:

- **Accuracy**—The data contains no errors and conveys true information
- **Completeness**—The data set includes all the information needed to serve its purpose
- **Consistency**—Data values from different sources are the same
- **Uniformity**—All measurements in the data are uniform, i.e. all in kilograms or all in pounds
- **Relevance**—The data is compatible with its intended use or purpose

High-quality data is essential for making good business decisions. If data quality is low or suspect, organizations don't have a complete and accurate picture of their organization, and they risk making poor investments, missing revenue opportunities, or impairing their operations.

However, in modern data pipelines, data is in constant motion. As data flows through a pipeline from its source to its destination, it goes through several different stages. Integration stages see multiple data sources combined. Transformation stages are where data is cleansed and validated. There are simple processing stages where data is summarized, aggregated, and filtered. And finally there are more sophisticated types of processing stages using machine learning such as predictive modeling. At any one of these stages, a process can fail or slow down, preventing data from getting to its intended destination and creating a potential risk to the business. Because of this, high-quality data does not necessarily guarantee data reliability.

The Three Pillars of Data Reliability

Reliability requires end-to-end observability of the data pipeline and the ability to predict, prevent and optimize the flow of data while taking into account bottleneck detection, resource efficiency and finally, cost.

Data reliability revolves around three pillars:

1. Pipeline performance management

When the flow of data through the pipeline is compromised, it can prevent users from getting the information they need when they need it, resulting in decisions being made based on incomplete, or incorrect, information. To identify and resolve performance issues before they negatively impact the business, organizations need data reliability tools that can provide a macro view of the pipeline. Monitoring the flow of data as it moves among a diversity of clouds, technologies, and apps is a significant challenge for organizations. The ability to see the pipeline end-to-end through a single pane of glass enables them to see where an issue is occurring, what it's impacting, and from where it is originating.

Compute performance monitoring is critical for managing and optimizing pipeline performance. To ensure data reliability, data architects and data engineers must automatically collect and correlate thousands of pipeline events, identify and investigate anomalies, and use their learnings to predict, prevent, troubleshoot, and fix a host of issues.

Compute performance monitoring enables organizations to:

- **Predict and prevent incidents—**Compute performance monitoring provides analytics around pipeline performance trends and other activities that are early warning signs of operational incidents. This allows organizations to detect and predict anomalies, automate preventative maintenance, and correlate contributing events to accelerate root cause analysis.

- **Accelerate data consumption—**Monitoring the throughput of streaming data is important for reducing the delivery time of data to end users. Compute performance monitoring allows organizations to optimize query and algorithm performance, identify bottlenecks and excess overhead, and take advantage of customized guidance to improve deployment configurations, data distribution, and code and query execution.

- **Optimize data operations, capacity, and data engineering—**Compute performance monitoring helps optimize capacity planning by enabling DevOps, platform, and site reliability engineers to predict the resources required to meet SLAs. They can align deployment configurations and resources with business requirements, monitor and predict the costs of shared resources, and manage pipeline data flow with deep visibility into data usage and hotspots.

- **Integrate with critical data systems—**With the right observability tools, compute performance monitoring can provide comprehensive visibility over Databricks, Spark, Kafka, Hadoop, and other popular open-source distributions, data warehouses, query engines, and cloud platforms.

2. Data reconciliation

As data moves from one point to another through the pipeline, there's a risk it can arrive incomplete or corrupted. Consider an example scenario where 100 records may have left Point A but only 75 arrived at Point B. Or perhaps all 100 records made it to their destination but some of them were corrupted as they moved from one platform to another. To ensure data reliability, organizations must be able to quickly compare and reconcile the actual values of all these records as they move from the source to the target destination.

Data reconciliation relies on the ability to automatically evaluate data transfers for accuracy, completeness, and consistency. Data reliability tools enable data reconciliation through rules that compare sources to target tables and identify mismatches—such as duplicate records, null values, or altered schemas—for alerting, review, and reconciliation. These tools also integrate with both data and target BI tools to track data lineage end to end and when data is in motion to simplify error resolution.

3. Drift monitoring

Changes in data can skew outcomes, so it's essential to monitor for changes in data that can impact data quality and, ultimately, business decisions. Data is vulnerable to two primary types of changes, or drift: schema drift and data drift.

Schema drift refers to structural changes introduced by different sources. As data usage spreads across an organization, different users will often add, remove, or change structural elements (fields, columns, etc.) to better suit their particular use case. Without monitoring for schema drift, these changes can compromise downstream systems and “break” the pipeline.

Data drift describes any change in a machine learning model with input data that degrades that model’s performance. The change could be caused by data quality issues, an upstream process change such as replacing a sensor with a new one that uses a different unit of measurement, or natural drift such as when temperatures change with the seasons. Regardless of what causes the change, data drift reduces the accuracy of predictive models. These models are trained using historical data; as long as the production data has similar characteristics to the training data, it should perform well. But the further the production data deviates from the training data, the more predictive power the model loses.

For data to be reliable, the organization must establish a discipline that monitors for schema and data drift and alerts users before they impact the pipeline.

Comprehensive data reliability must be able to support enterprise data teams by:

Eliminating downtime

Monitor enterprise data across data lakes, warehouses, and other repositories to eliminate issues that impact reliability.

Scaling Workloads

Ensure availability for mission-critical data and workloads.

Automating Validation

Classify, catalog, and manage business rules for data at rest and data in motion.

Use Cases

PhonePe

PhonePe is one of the world's largest digital payments services, processing 400 million cash transactions per month and peaking at 1,100 cash transactions per second. To support its explosive business growth, PhonePe needed to rapidly scale its Hadoop-based clusters while adding other open source data technologies, such as Apache Hbase, HDFS, Kafka, Spark, and Spark Streaming, to run their high-volume, real-time payments and cash transfer platform. This infrastructure expansion put tremendous pressure on system performance and reliability.

Acceldata brought real-time visibility to PhonePe's Hbase, Hive, and Spark data pipelines, enabling PhonePe's engineers to monitor its entire data infrastructure from a single application. PhonePe's data reliability team was able to use Acceldata Pulse to monitor its modern data infrastructure performance so that it could easily distinguish between, for example, changes created by infrastructure issues and those driven by seasonal or campaign-related surges. Having increased visibility into its performance-intensive data workload compute performance helped PhonePe improve reliability significantly. The company has been able to smoothly grow its data infrastructure by 13x, from 70 nodes to 1500+ nodes, while maintaining 99.98% availability across its Hadoop data lakes. PhonePe also eliminated unplanned outages and Severity Level 1 issues.

TrueDigital

TrueDigital was struggling with poor performance in its proprietary data software and storage solutions. More than 50% of the company's data went unprocessed on a daily basis because its data infrastructure couldn't handle the data volume fast enough. Not to mention the system suffered from nagging reliability issues that impacted processing ability even further.

Acceldata brought real-time visibility to PhonePe's Hbase, Hive, and Spark data pipelines.

With Acceldata in place, TrueDigital's 8-petabyte data lake now runs smoothly over a 100+ node cluster using Hortonworks Data Platform (Apache Hadoop, Hive, and Spark), as well as Apache Ranger, Kafka, and open-source HDP. TrueDigital now processes 500 million user impressions/month, and streams almost 70,000 messages/second, all without suffering a single unplanned outage or Severity Level 1 issue.

PubMatic

PubMatic's Internet advertising platform helps publishers and app developers around the world reach their target audiences. Its Hortonworks Data Platform-based cluster was massive — thousands of nodes handling hundreds of petabytes of data. But that scale was causing frequent performance issues that were resulting in a critically high Mean Time to Resolve (MTTR). And despite relying on HDP and other Apache open-source software, PubMatic's infrastructure and support costs were also through the roof.

Acceldata Pulse helped PubMatic's engineers isolate data bottlenecks, automate performance improvements, and distinguish between mandatory and unnecessary data. This enabled Pubmatic to reduce its HDFS block footprint by 30% and consolidate its Kafka clusters, saving costs. Overall cost savings from reduced software licenses alone were in the millions of dollars a year. Even while trimming its infrastructure, PubMatic was able to boost the reliability and scale of its data pipelines, enabling its engineers to focus on supporting the growth of its mission-critical analytics business. Thanks in large part to Acceldata, PubMatic every day now handles more than 200 billion ad impressions and one trillion advertiser bids, while processing two petabytes of new data.

Market Landscape

Data reliability expands the capabilities of the application performance monitoring (APM) tools enterprises use to ensure that applications and their related infrastructure operate at maximum efficiency. APM tools continuously monitor application environments and leverage machine data to detect anomalies, identify trends, optimize resource usage, and troubleshoot performance issues before they impact end-users. Using machine learning, APM tools correlate infrastructure events to pinpoint root causes of slowdowns, blockages, and failures, then help administrators remediate them to improve and maintain operational workloads.

While these traditional APM tools are sufficient for providing visibility into the performance and health of microservices and web applications, they weren't designed to support complex data systems and their multitude of use cases. Without the ability to correlate and contextualize events across these complex ecosystems, data architects and engineers struggle to identify and solve problems quickly.

Data reliability also intersects with two other market segments. It shares capabilities with DataOps, which applies DevOps and agile development principles to the development and optimization of data applications and data pipelines. DataOps orchestrates people, processes, and technology to deliver data to users quickly, while data reliability brings the monitoring, data validation, and lineage pieces to help ensure data quality and data delivery performance. Data reliability also addresses the monitoring, diagnosis, and remediation aspects of ITOps and AIOps, which provision, manage, monitor, and tune IT infrastructure resources.

How to Improve Data Reliability

- **Start observing**—The first step to increasing data reliability is to reduce the complexity of your data systems. A data observability platform will provide comprehensive visibility of your pipelines, regardless of architecture, and improve your control of all the elements that handle AI and analytics workloads. This high-level view of end-to-end processes enables you to identify and drill down into data and processing issues that cause latency, failures, and other impediments to data reliability.

- **Eliminate data downtime**—It's important to monitor data across hybrid data lakes and warehouses to ensure high data quality and reliability. Compute performance monitoring improves the reliability of your data processing by correlating events across the environment for rapid root cause analysis, analyzing performance trends to predict potential failures, and automating fixes to prevent incidents before they impact operations.

- **Automate data validation**—Data drift can affect AI and machine learning accuracy. It's important to detect drift before impacting operations through continuous automated validation that addresses data quality, schema drift, and data drift to eliminate disruption and improve the accuracy of analytics and AI.

- **Monitor Data in Motion**—Data is never static. Organizations need to classify, catalog, and manage business rules for data in motion through the entire data pipeline with an enterprise architecture that is data source, infrastructure, and cloud provider-agnostic.

Improving Data Reliability with Acceldata

Acceldata launched the market's first multidimensional Data Observability Cloud as an easy-to-access, instantly available service for enterprises adopting hybrid data lakes and cloud data warehouses. The Acceldata multidimensional data observability platform enables enterprises to take advantage of real-time observability to:

- Build, operate, and optimize complex data systems across environments—e.g., hybrid data lakes and warehouses—and cloud providers with the highest level of data team productivity and return on data investment.
- Scale rapidly without loss of data quality across technologies, workloads, and applications.
- Align data and business strategies to ensure that enterprises satisfy business objectives.
- Maximize internal data team expertise to do more with fewer resources.

Get a demo

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