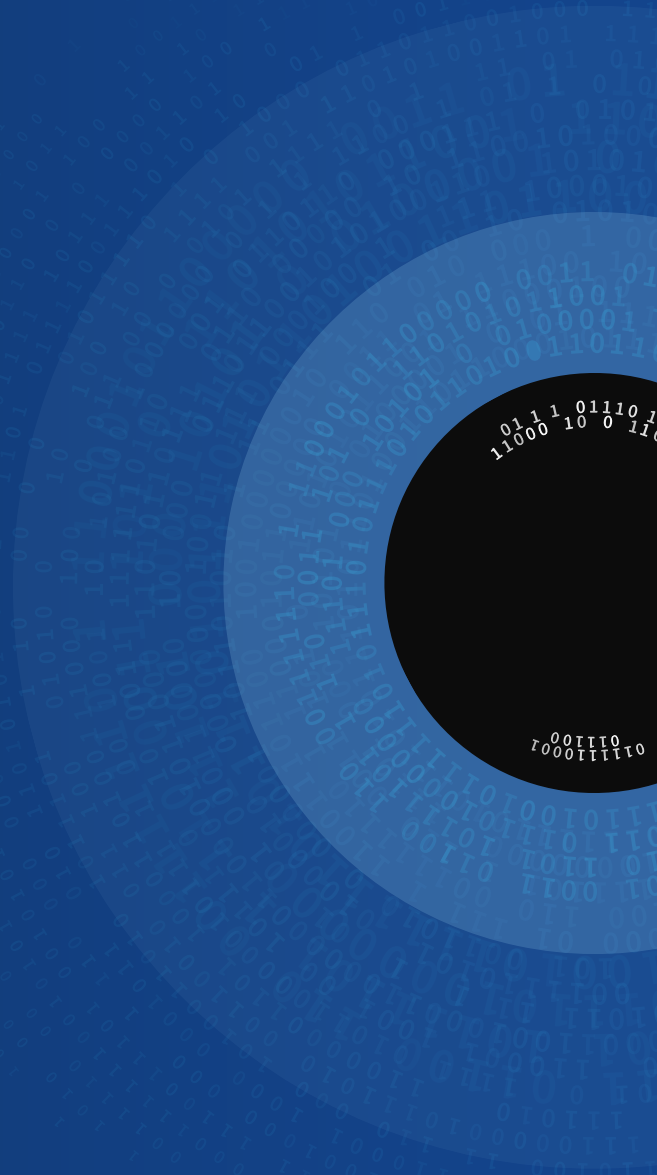
acceldata

The Definitive Guide to Data Reliability for Enterprise Data Teams



February 2023

Introduction

As analytics has evolved from a supporting role to a business-critical one for all organizations, the processes for delivering the necessary data for analytics have become top priorities. Data must now be treated as mission-critical, and for it to have optimal impact, data must have the highest degree of data reliability.

Analytics have evolved from traditional data warehouse approaches to modern, cloud-based analytics. As a result, the types of data captured and used, and the data stack that delivers the data have all had to evolve as well. Modern analytics deal with different forms of data: data-at-rest, data-in-motion, and data-for-consumption. As the data stack moves and transforms data in near real-time, it requires data reliability that keeps pace with the continuously changing
range and scope of data.

In this paper, we'll explore what data reliability is and what it means for modern analytics. We'll look at why a fresh new approach to data reliability is required to keep enterprise data and analytics processes agile and operational.

Limitations of Legacy Data Quality Systems

Historically, data processes that were built to deliver data for analytics were batch-oriented and focused on highly structured data. Data teams had very limited visibility into the data processes and processing and focused their data quality efforts on the data output from the processes: data-for-consumption.

Legacy data quality processes were limited in scope and functionality; they were equipped to only do a few things:

- Performed batch runs, which were done as semi-regular “data checks” (usually weekly or monthly)
- Only performed your basic quality checks
- Was only run on the structured data in the data warehouse
- Were manual queries or performed by “eyeballing” the data

Legacy data quality tools and processes had limitations of data processing and warehousing platforms of the time. Performance limitations constrained how often data quality checks could be performed and limited the number of checks that could run on each dataset.

Why Data Processes Require a New Approach

Modern analytics and modern data stacks are more complex, and as a result, the potential issues with data and data processes have grown:

The volume and variety of data make datasets much more complex and increase the potential for problems within the data

The near real-time data flow could introduce incidents at any time that could go undetected

Complex data pipelines have many steps, each of which could break and disrupt the flow of data

Data stack tools can tell you what happened within their processing but have no data on the surrounding tools or infrastructure

> To support modern analytics, data processes require a new approach that goes far beyond data quality: data reliability.

What is Modern Data Reliability

Data reliability is a major step forward from traditional data quality. Data reliability includes data quality but covers much more functionality that data teams need to support for modern, near-real-time data processes.

Data reliability takes into account the new characteristics of modern analytics. It provides:

- More substantial data monitoring checks on datasets such as data cadence, data drift, schema drift, and data reconciliation to support the greater volume and variety of data
- Continuous data asset and data pipeline monitoring and real-time alerts to support the near real-time data flow
- End-to-end monitoring of data pipeline execution and the state of data assets across the entire data pipeline to detect issues earlier
- 360-degree insights about what is happening with data processes from information that is captured up and down the data stack to drill down into problems and identify the root cause

Key Characteristics of Data Reliability

Many data observability platforms with data reliability capabilities claim to offer much of the functionality of modern data reliability mentioned above. When looking for the best possible data reliability platform, there are some critical elements that need to be considered, as we outline below.

Traditional data quality processes were applied at the end of data pipelines on the data-for-consumption. One key aspect of data reliability is that it performs data checks at all stages of a data pipeline across any form of data: data-at-rest, data-in-motion, and data-for-consumption.

End-to-end monitoring of data through your pipelines allows you to adopt a “shift-left” approach to data reliability. Shift-left monitoring lets you detect and isolate issues early in the data pipeline before it hits the data warehouse or lakehouse. This prevents bad data from hitting the downstream data-for-consumption zone and does not corrupt the analytics results. Early detection also allows teams to be alerted to data incidents and remediate problems quickly and efficiently.

Here are five additional key characteristics that a data reliability platform should support to help your team deliver the highest degrees of data reliability:

Automation

Data reliability platforms should automate much of the process of setting up data reliability checks. This is typically done via machine learning-guided assistance to automate many of the data reliability policies.

Data team efficiency

The platform needs to supply data policy recommendations and easy-to-use no- and low-code tools to improve the productivity of data teams and help them scale out their data reliability efforts.

Scale

Capabilities such as bulk policy management, user-defined functions, and a highly scalable processing engine allow teams to run deep and diverse policies across large volumes of data.

Operational Control

Data reliability platforms need to provide alerts, composable dashboards, recommended actions, and support multi-layer data to identify incidents and drill down to find the root cause.

Advanced data policies

The platform must offer advanced data policies that go far beyond basic quality checks such as data cadence, data drift, schema drift, and data reconciliation to support the greater variety and complexity of data.

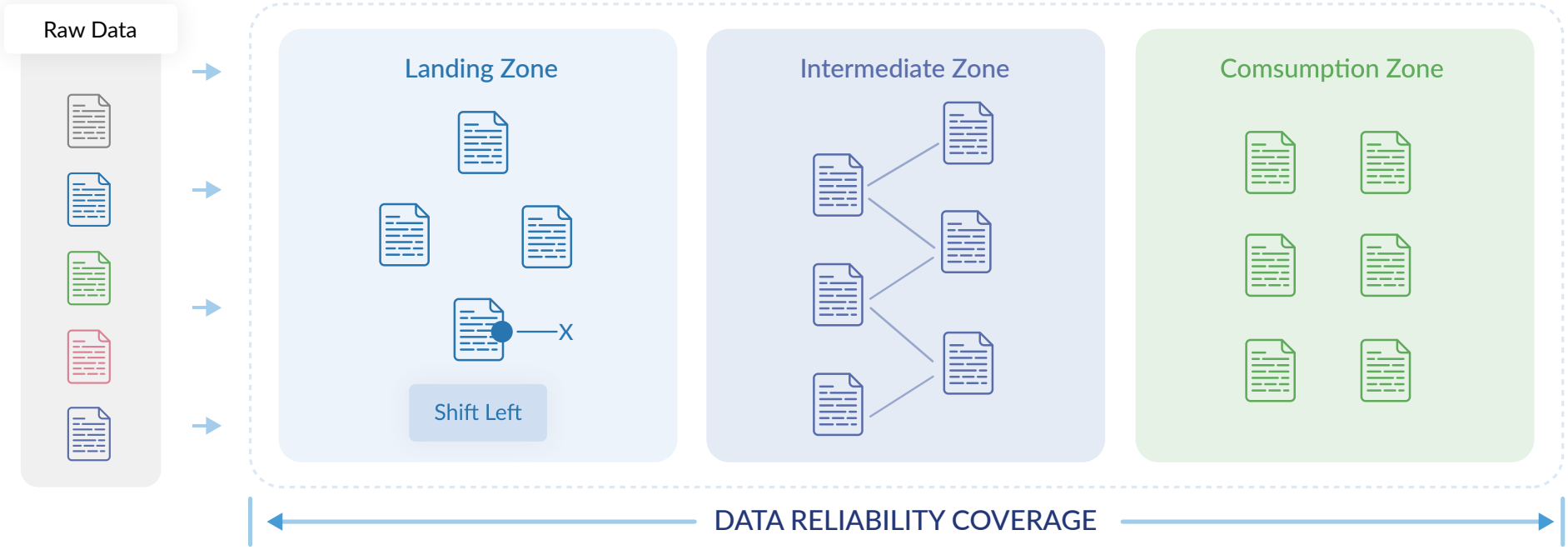
How to Operationalize Data Reliability

Data reliability is a process by which data and data pipelines are monitored, problems are troubleshot, and incidents are resolved. A high degree of data reliability is the desired outcome of this process.

Data reliability is a data operations (DataOps) process for maintaining the reliability of your data. Just like network operations teams would use a Network Operations Center to gain visibility up and down their network, data teams can use a data reliability operations center in a data observability platform to get visibility up and down their data stack.

With data reliability, organizations derive important advantages, including benign able to:

- Set up data quality and monitoring checks on all your critical data assets and pipelines using built-in automation to do this efficiently and increase the coverage of data policies.
- Monitor your data assets and pipelines continuously, getting alerts when data incidents occur.
- Identify data incidents, review and drill into data related to these incidents to identify the root cause and determine a resolution to the problem.
- Track the overall reliability of your data and data processes and determine if the data teams are meeting their service level agreements (SLAs) to the business and analytics teams who consume the data



Data Reliability in Acceldata Data Observability Cloud

The Acceldata Data Observability Cloud provides data teams with end-to-end visibility into your business-critical data assets and pipelines to help you obtain the highest degrees of data reliability. All your data assets and pipelines are continuously monitored as the data flows from source to final destination and checks are performed at every intermediate stop along the way for quality and reliability.

Acceldata helps data teams better align their data strategy and data pipelines to business needs. Data teams can investigate how a data issue impacts business objectives, isolate errors impacting business functions, prioritize work, and resolve inefficiencies based on business urgency and impact.

The Data Observability Cloud supports the end-to-end, shift-left approach to data reliability by monitoring data assets across the entire pipeline and isolating problems early in the pipeline before poor-quality data hits the consumption zone. It works with data-at-rest, data-in-motion, and data-for-consumption to work across your entire pipeline.

Data teams can dramatically increase their efficiency and productivity with the Data Observability Cloud. It does this via a deep set of ML- and AI-guided automation and recommendations, easy-to-use no- and low-code tools, templatized policies and bulk policy management, and advanced data policies such as data cadence, data-drift, schema-drift, and data reconciliation

With the Data Observability Cloud, you can create a complete data operational control center that treats your data like the mission-critical asset that it is and helps your team deliver data to the business with the highest level of data reliability.

About Acceldata

Acceldata is the market leader in enterprise data observability for the modern data stack. Founded in 2018, Campbell, CA-based Acceldata, enables data teams to build and operate great data products, eliminate complexity, and deliver reliable data efficiently.

Our Customers

PhonePe ORACLE Verisk true dun & bradstreet & many more

acceldata

www.acceldata.io

What's Next

Request a demo | Tour the product | Contact us