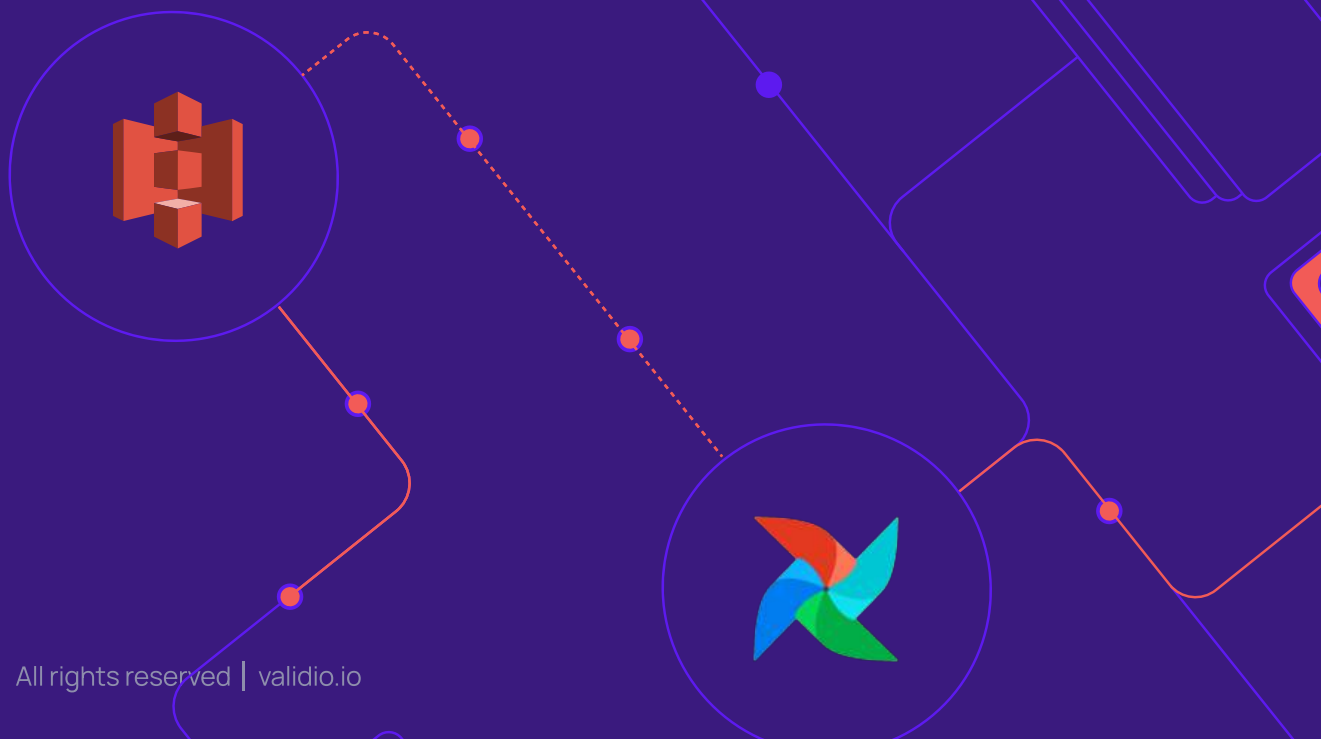
validio

The data leader's guide to Data Observability

- [x] How to measure and improve data quality
- [x] When to go broad and when to go deep
- [x] How to assess the need for data observability



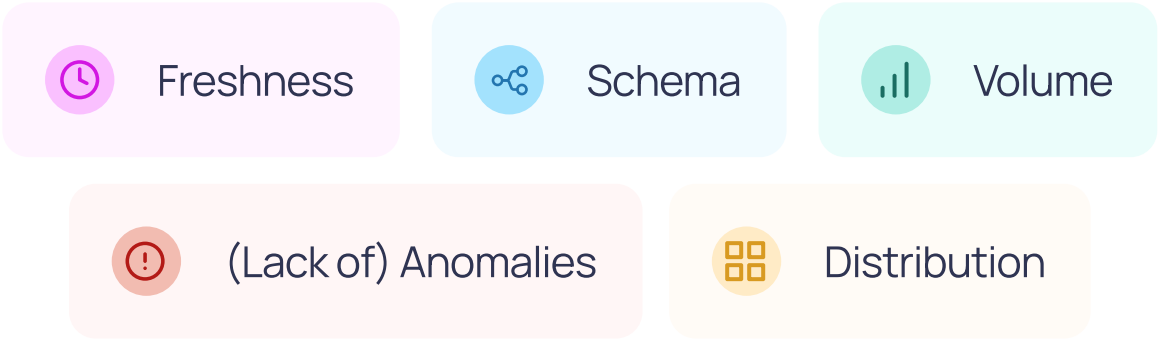
01 Introduction

With poor data quality costing organizations an average of $12.9 million annually, it's clear that data quality is an issue – not only for the data team but also for the wider business.

"Poor data quality costs organizations an average of $12.9 million annually."

Gartner report

And, as for most issues, a solution starts with awareness. When it comes to data quality, Data leaders should can become aware of the level of data quality by measuring it across five pillars: freshness, volume, schema, (lack of) anomalies, and distribution.



Data observability is one way to improve data quality. Data observability has several dimensions, including data sources, data formats, validator cadence, and user focus. The more of these dimensions you have observability into, the more efficient it will be to improve your data quality.
In this guide, we'll cover:

- ☑ How to measure data quality across the five observable pillars
- ☑ How to improve data quality through data observability
- ☑ When to go broad and when to go deep in data observability
- ☑ Assessing the business needs to data observability

By the end, you'll have a clear roadmap for evaluating data observability - and how it can be used to unlock the full potential of your organization's most valuable asset.

Let's get started.

Contents

01 Introduction to "Deep" Data Observability

Data quality and data observability are two related but distinct concepts.

Whereas data quality is a measurement of the extent to which data is fit for the intended purpose, data observability is a solution to allow data practitioners to improve the quality of the data.

2022 was the year when data observability really took off as a category, with the official Gartner terminology for the space. Nevertheless, the industry is nowhere near fully formed. In his 2023 report titled "Data Observability—the rise of the data guardians", Oyvind Bjerke at MMC Ventures discusses the space as having massive amounts of untapped potential for further innovation. One example of such innovation is the discussion on data contracts driven by thought leaders like Chad Sanderson and Andrew Jones. All in all, data observability still holds a lot of best practices yet to be uncovered.

In recent years, we've seen definitions start to converge. we define data observability as:

> → The degree to which an organization has visibility into its data pipelines.
> A high degree of data observability enables data teams to improve data quality.
>
> Definition of data observability

However, not all data observability tools, tools specifically designed to help organizations reach data observability, are equal. The tools differ in terms of the degree of data observability they can help data-driven teams achieve. While there's cases with a need to go broad with surface-level checks, the real value emerges when being able to go deep and uncover issues otherwise hidden. When determining the depth of data observability, we look at the following dimensions: Data sources, data formats, data granularity, validator configuration, validator cadence, and user focus.

In the rest of this article, we introduce two frameworks that serve as a guide for data leaders on how they should navigate this new space. First, we lay the foundation of data quality by introducing the five pillars data leaders should aim to observe. Later, we dive deep on data observability and explain the six dimensions relevant to data leaders when choosing tools.

02 The 5 observable pillars of data quality

Let's start with the foundation. Good data quality is the end goal when it comes to any data observability effort. But what does data quality really mean? We define data quality as:

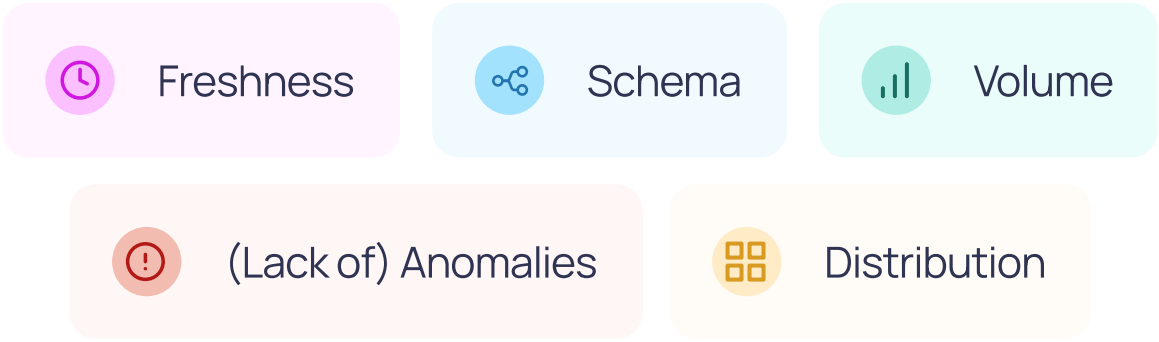
> → The extent to which an organization's data can be considered fit for its intended purpose. Data quality should be observed along five dimensions: freshness, volume, schema, (lack of) anomalies, and distribution. It is always relative to the data's specific business context.
>
> Definition of data quality

In other words, one single dataset can be considered high quality in one context and low quality in another. It's worth noting that business context can also depend on what stakeholder is using the data and for what purpose.

The five pillars of data quality are derived based on findings from conversations with 300+ data professionals and are defined to be mutually exclusive and collectively exhaustive.

Unlike some other pillars, they can be meaningfully measured and observed in the context of data observability. As a result, they are the pillars of data quality that matter the most for companies that want to become data-driven and able to fully trust their data. We will take a look at each one in more detail.

The 5 observable pillars of data quality
are the goal of data observability



Freshness

Data is updated and available at the expected time, with expected cadence.

That is to say data is available when you need it. Freshness is an important aspect of data quality because it answers two questions: when is this data from (data freshness)? and when was this data loaded into my system (source freshness)? If any of the two is off, it means something is probably broken upstreams and needs to be looked into.

Volume

The expected amount of data is available. Bad data manifests itself in the form of too much data (e.g. duplicates) or too little data (e.g. missing records).

Volume is a hugely important pillar of data quality since it can quickly indicate if something has gone wrong in transformations or ingestions of data. For example, low volume can be an indication that a pipeline job was not executed, and too high volumes can indicate duplicates were introduced with a new transformation.

Schema

A dataset holds the expected fields in the expected formats.

Schema adherence is another backbone of data quality, because it defines what can be done with data later in the pipelines. For example, transformation scripts and/or machine learning pipelines might break if a field is missing or of a value that's expected to be an integer type is in fact a string type. Note that we define schema in the broadest sense to also include schema for nested data types and not just for structured data.

(Lack of) Anomalies

The values of individual datapoints are expected.

Anomalies are defined in a broad sense and in many different ways; ranging from whether a datapoint, e.g. price = $39.99, surpasses a manually set business threshold, whether the datapoint is statistically different from what has been observed in the data so far (including multivariate analysis), or whether the cardinality of a field (i.e. column in a table) unexpectedly goes from three ({UK, SE, DK}) to four ({UK, SE, DK, NO}, in this case, 'NO' would be an anomalous datapoint).

An ML-based approach is important for obtaining high data quality at scale since it is often impossible to manually define and maintain thresholds and rules for hundreds of features across thousands of datasets.

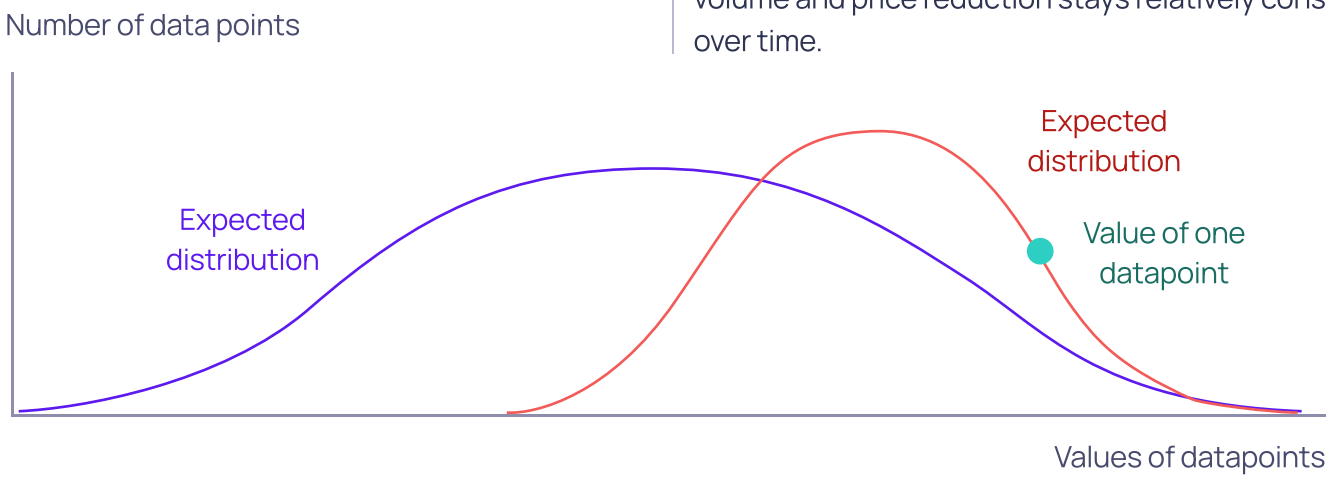
Validity

Datasets have expected distributions,

Each datapoint individually might look fine, but when grouped together they can show unexpected distributions that can have large implications for the business.

For example, let's say the "price" field in a dataset usually holds datapoints between $9.99 and $99.99. If all of a sudden you get a value of $1,000, this would be an anomaly (previous section) that doesn't necessarily shift the distribution (e.g. mean) of the data in a significant way. However, if all of a sudden 70% of the datapoints in the "price" column take on the value of $89.99, you will most certainly have a distribution shift because e.g. the mean and the shape of the distribution will change.

Worth noting here is that distribution must be considered in a univariate as well as a multivariate way. Real data has dependencies between fields, and ensuring these dependencies behave as expected is an important part of data quality. For example, you might want to ensure covariance between sales volume and price reduction stays relatively constant over time.



The value of an individual datapoint (green circle) might not be anomalous, but as a whole the distribution of the dataset (red line) might shift versus the expected distribution (purple line).

03 The 6 dimensions of "Deep" Data Observability

We've defined the five pillars of data quality. Next, we dive deeper into the six dimensions of data observability needed to improve data quality in a proactive and future-proof way.

These dimensions are data sources, data formats, data granularity, validator configuration, validator cadence, and user focus. The more of these dimensions you have observability into, the deeper your data observability.

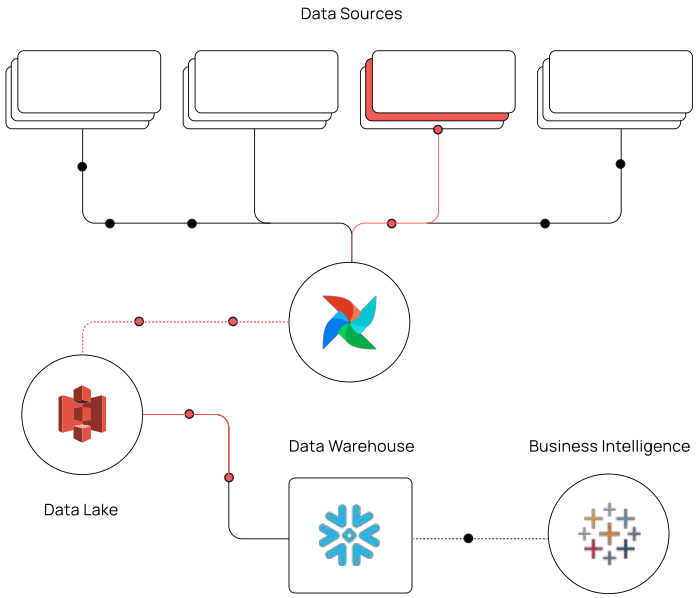
	Surface-level Data Observability	Deep Data Observability
01 Data Sources	Data warehouse only	Data streams, -lakes and -warehouses
02 Data formats	Structured data only	Semi-structured (nested) & structured data
03 Data granularity	Focused on aggregate data	Univariate & multivariate validation of individual datapoints and aggregate data
04 Validator configuration	Automatic, one-size-fits-all	Default is automation, full customization easy
05 Validator cadence	Daily	As frequently as needed, including real-time
06 User focus	Business users	Both business users and technical users

01 Data sources
Truly end-to-end

Surface-level data observability solutions tend to focus only on the data warehouse through SQL queries. Deep Data Observability solutions on the other hand, provide data teams with equal degrees of Observability across data streams, data lakes and data warehouses.

This is important in order to be truly proactive in ensuring good data quality for two reasons. Firstly, data does not just magically appear in the data warehouse. It often comes through some streaming source and lands in a data lake before it gets pushed to the data warehouse. Bad data can appear anywhere along the way, and you want to be able to identify the issue as soon as possible and be able to accurately pinpoint its origin.

Secondly, in an increasing amount of data use cases such as for machine learning and automated decision making, data never even touches the data warehouse. For a Data Observability tool to be proactive and future-proof, it needs to be truly end-to-end.



02 Data formats
Structured & semi-structured

Data streams and lakes are nice segways into the next section: data formats. Since Shallow Data Observability is focused on the data warehouse, it is thus adept at obtaining observability for structured data. However, in order to reach a high degree of Data Observability end-to-end in your data stack, the Data Observability solution needs to support data formats that are common in data streams and lakes (and increasingly warehouses). With Deep Data Observability, data teams can obtain high-quality data by monitoring the five pillars of data quality not only in structured datasets, but also for semi-structured data in nested formats.

03 Data granularity

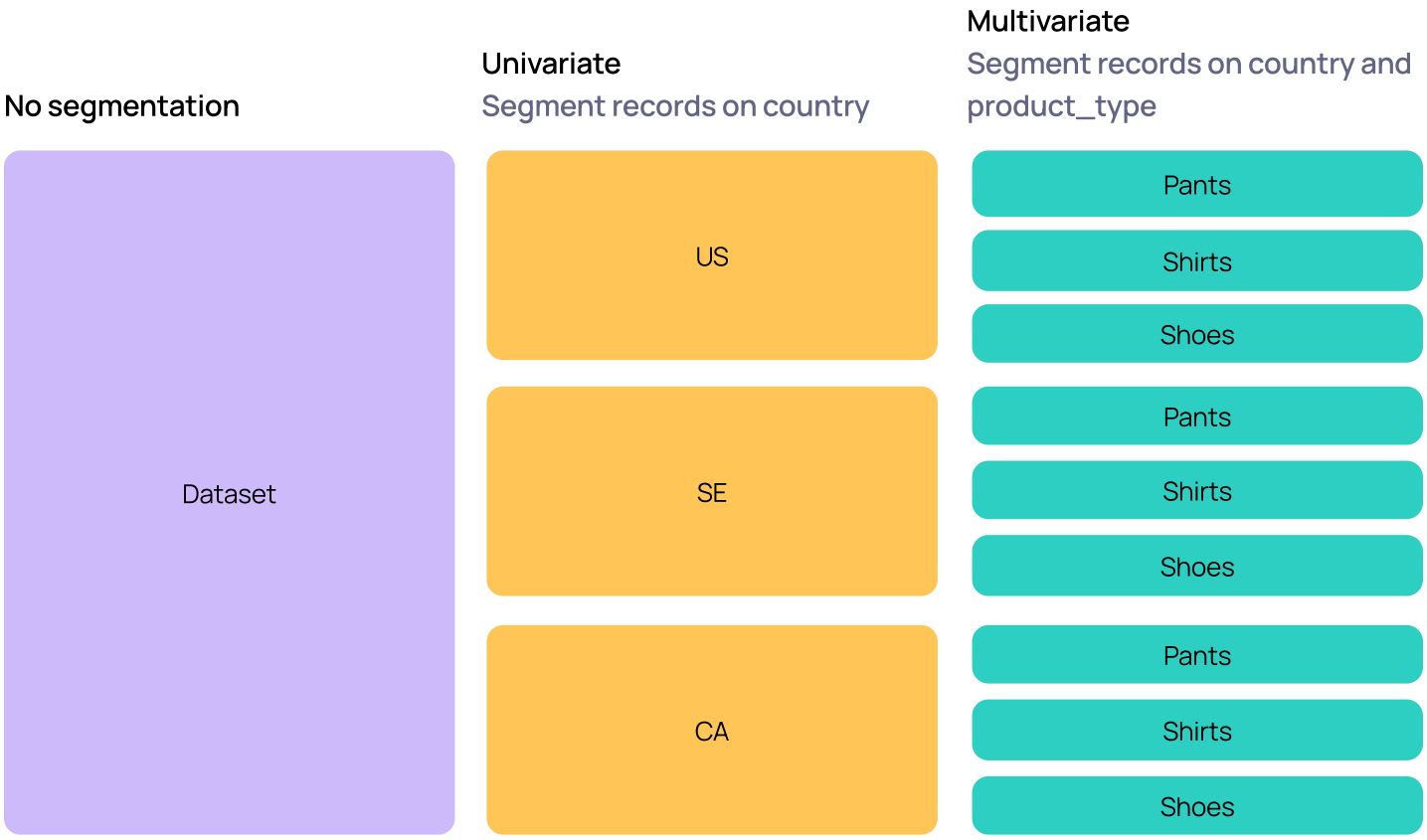
Univariate & multivariate validation of individual datapoints and aggregate data

Shallow Data Observability originally rose to fame based on analyzing aggregate data (e.g. metadata) and calculating summary statistics about datasets in a scalable way.

As has become abundantly clear over time, there is a need to not only look at summary statistics and distributions, but also each individual datapoint. Shallow Data Observability is sometimes complemented with custom SQL statements that try to go slightly deeper than metadata and summary statistics. However, this approach doesn't scale for most data teams. In order to obtain Deep Data Observability, data teams must take a data-centric approach where each datapoint can be validated in its own right, when needed in a simple way (more on "simple way" in the next section).

In addition, the ability to validate each of the five pillars of data quality from a multivariate perspective is important, since most real-world data quality issues are multivariate in nature. A Data Observability solution is not "Deep" if it does not support multivariate validation. For example, Deep Data Observability means a data team can observe the distribution of the individual segments where segments are determined by multiple variables.

In the following example, the dataset is segmented on **country** and on **product_type** (multiple variables, not just one) which is necessary in order to validate each individual subsegment (set of records). Each subsegment is likely to have unique volume, freshness, anomalies and distribution, which means it must be validated individually.



Each sub-segment has its own unique characteristics that must be validated in order to reach Deep Data Observability

04 Validator configuration: Automatically suggested as well as manually configured

Depending on what your organization looks like, you might be looking for various degrees of scalability in your overall data system. If what you're looking for is a "set it and forget it" type of solution that will alert you whenever something out of the ordinary happens, then Shallow Data Observability is what you're after. You will get a bird's eye view of e.g. all tables in your data warehouse and whether they behave as expected.

Conversely, it might be the case that your organization has some business logic, or custom validation rules you'll want to set up. The degree to which you're able to do this custom setup in a scalable way determines the degree to which you have Deep Data Observability. At the end of the day, it needs to be easy enough to set things up in a totally automated approach. Similarly, it needs to be easy to tailor validations to your business logic, so that it does not become a hurdle to set up fully comprehensive Data Observability.

For example, if each custom rule requires a data engineer to write SQL, you're looking at a not-so-scalable setup, and it will be very challenging to reach the state of Deep Data Observability. On the other hand, if you have a quick-to-implement menu of validators that can be combined in a tailored way to suit your business, then Deep Data Observability is within reach. Setting up customized validators should not be reserved for code-savvy data team members only.

05 Multi-cadence validation: As frequently as needed, including real-time

Again, depending on your business needs, you might have different requirements for Data Observability on various time horizons. If you use a standard type of setup where data is loaded into your warehouse every day, then Shallow Data Observability which only supports a standard daily cadence will be all you need.

Instead, if your data infrastructure is more complex with some sources being updated in real-time, some sources being updated daily, and others being updated less frequently, e.g. weekly or monthly, you will need support to validate data in all of these cadences. A Deep Data Observability tool has full support for this. This ensures that you get proper insights into your data at the right point in time that makes sense for your business context. It also means that you will be able to react on bad data before it hits your downstream use cases.

04 Validator configuration: Automatically suggested as well as manually configured

Data quality is an inherently cross-functional problem, which is part of the reason why it can be so challenging to solve. The person who knows what “good” data looks like in a CRM dataset might be a sales person with their boots on the ground in sales calls. Thus, the person that moves (or ingests) data from the CRM system into the data warehouse might have no insight into this at all, and might naturally be more concerned with whether the data pipelines ran as scheduled.

Shallow Data Observability solutions primarily cater to one single user group. They either focus on the data engineer who cares mostly about the nuts and bolts of the pipelines and whether the system scales. Alternatively, they focus on the business users who might care mostly about dashboards and summary statistics.

Deep Data Observability is obtained when both types of users are kept in mind. In practice, this means providing multiple modes of controlling a Data Observability platform: through a command line interface as well as through a graphical user interface. It might also entail multiple access levels and privileges. In this way, all users can obtain a high degree of visibility into data pipelines. It also means configuring Data Observability for high data quality can become a truly cross-functional effort. This in turn effectively democratizes data quality within the whole business.

04 Concluding thoughts

We've now covered the five observable pillars of data quality as well as the six dimensions of data observability. Our hope is that this report gives you two frameworks to rely on when evaluating your business needs for data quality and data observability tooling.

At the heart of your business success lies a not-so-secret ingredient: pristine data quality. Imagine a world where every piece of data at your fingertips is accurate, comprehensive, and as consistent as your favorite morning routine. This is the world where informed decision-making and customer satisfaction thrive. The key? A steadfast commitment to honing your data quality, armed with robust data quality processes, tools, and monitoring key metrics.

Are you ready to embark on your data quality journey? Take the first step today and ensure your data remains your biggest asset.

Start your free trial →

About Validio

Validio provides an automated data observability and quality platform to help you stay on top of your data and metrics - boosting your data team productivity and making your data ready for AI. Data-led companies like Truecaller, OfferFit, AB CarVal, and Volt rely on Validio to find and fix data issues before they become business issues.

Contact us
Validio.io

HQ

Linnégatan 78, 115 23
Stockholm