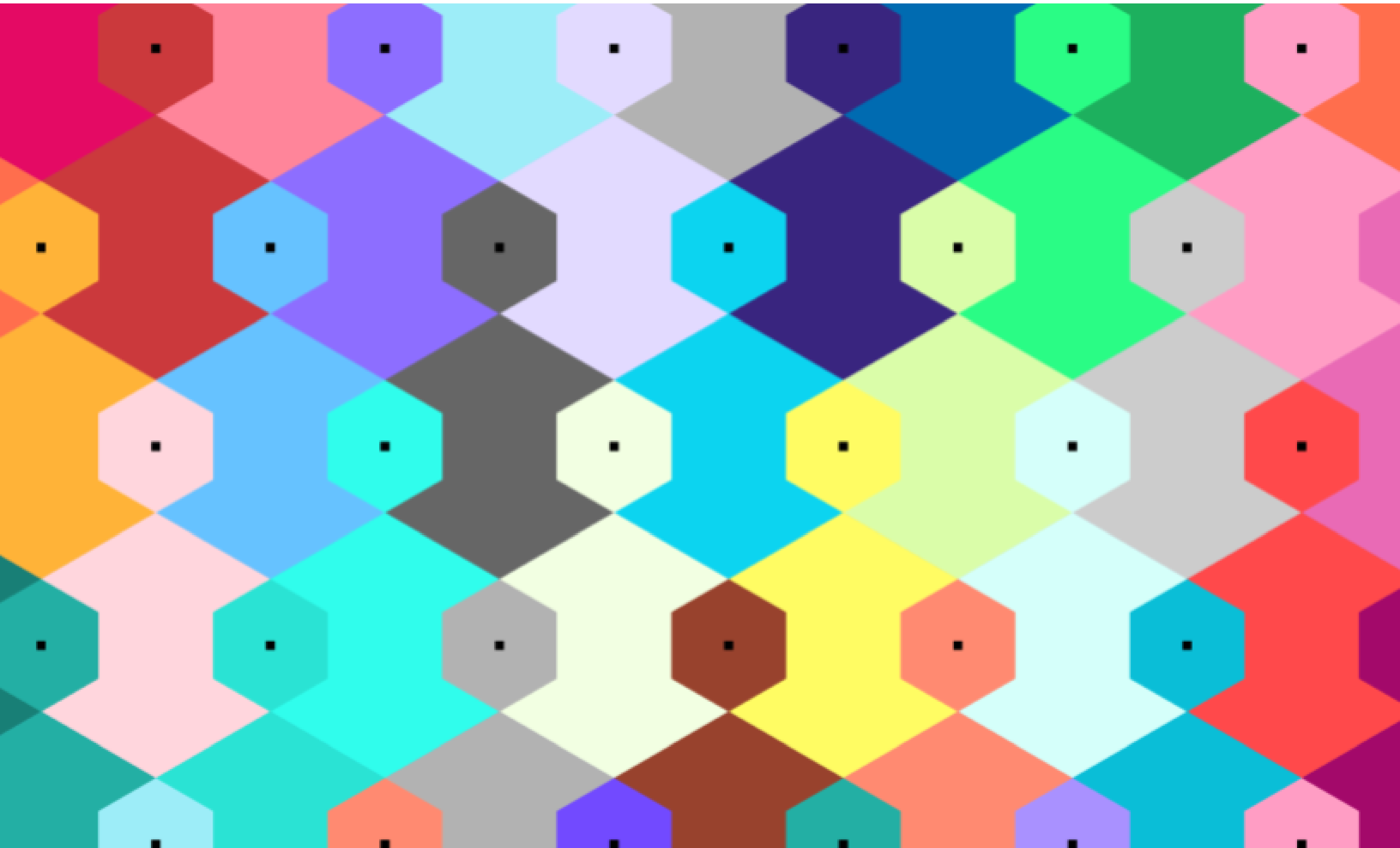
Anomalo

A Business Leader's Guide to Data Quality

White Paper

May 2022



What's lurking in your data lake?

With the rise of cloud data warehouses and lake houses, the data at your business's disposal has never been more diverse or more democratized. It's no longer surprising to be working with terabytes or petabytes of data as you connect to third-party data sources and ingest new information about your users and products. Furthermore, data is no longer the purview of a handful of engineers—now everyone from sales and marketing to operations has access and wants to use data to drive improvements in their respective domains.

But as anyone who's dipped beneath the shiny surface of their data lake can tell you, *here be monsters*. Corrupted data, dropped columns, stale tables, and a sudden proliferation of NULLs are all common data issues. And data issues are one of the top complaints for data-driven teams today. Data quality incidents can cause customer issues in your product, hamstring your analytics team, and feed your AI models with false information. Root-causing bugs can consume valuable analytics and engineering time, and even worse, it's easy for issues to silently wreak havoc for months before they're discovered.

In this whitepaper, we'll provide a framework for taking a proactive approach to data quality as your organization continues to scale and grow its stack. Read on to learn the following:

1. **Introduction to data quality**
    1. What data quality means
    2. Where data quality issues can appear in the modern data stack
    3. How and why data quality issues erode trust
2. **Improving your data quality**
    1. A three-pronged approach for monitoring data quality
    2. How to accelerate resolution with automated root cause analysis
3. **Considering a data quality platform**
    1. The build vs. buy decision
    2. Key features to look for in a data quality platform

Introduction to data quality

Defining data quality

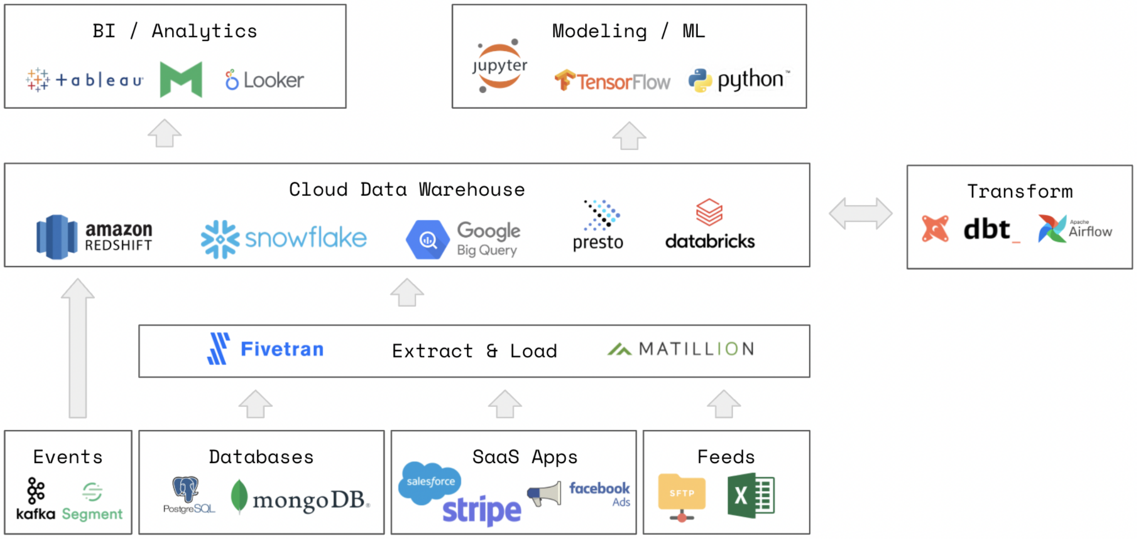
"Quality" can mean different things to different people. After decades of working on data quality, we believe trustworthy data meets the following four requirements:

1. **On Time:** Did we receive the data in the expected timeframe?
2. **Complete:** Do we have all the data we should have or is some of it missing?
3. **Consistent:** Is the data in the format we expect? Are the column names and types still the same? (Otherwise, our pipelines might break.)
4. **Error-free:** Did a human or machine introduce an error or corruption that changed the data from its true state into another false or misleading state?

©2022 Anomalo, Inc.

Points of failure in the modern data stack

The modern data stack—consisting of a variety of real-time and batch historical data sources feeding into a single cloud data warehouse or data lake—has been transformational from an analytics and machine learning perspective. From a data quality perspective, it's also introduced unprecedented ways for things to go wrong.



At every layer of the stack, data quality issues can be introduced

- **Late delivery** from third-party data sources or because internal batch jobs failed to complete on time.
- **Missing events and incomplete data** due to bugs or concurrency issues in your event streams and data feeds.
- **Unexpected changes in the format or definition** of your data inputs because one of your data sources changed its schema without letting you know ahead of time.
- **Corrupted data** because of a bug at the source or because somewhere along the data pipeline of cleaning, extracting, joining, and transforming the data, an error was introduced.

The impact of data quality issues

As your modern data stack grows, new data quality issues will inevitably arise. If you don't have robust systems and processes in place for addressing these issues, data quality will erode over time, slowly and mostly silently. Meanwhile, trust—in your product, your dashboards, and the decisions that power your business—will erode along with it. The types of impact we see can be characterized as follows:

User experience

One of the most devastating impacts a data quality issue can have is that something goes wrong in the product as a result. An integration with a SaaS app could fail because the format of the data is no longer as expected. Key data used in the product could be accidentally wiped out or dropped.

©2022 Anomalo, Inc.

AI model quality

Machine learning models are increasingly powering real-time business decisions, from in-product personalization to automated operations. They're trained on historical data to make predictions. Data quality issues often create sudden chunks of data that are *completely different* from what the model was trained on. When the model makes predictions using this data, it can behave wildly. For example, a fraud model could suddenly classify every affected transaction as fraudulent (ruining the customer experience) or every transaction as genuine (leading to fraud exposure).

Business intelligence and analytics

Your dashboards and reporting are only as good as your data quality. As the democratization of data leads to more internal stakeholders having access to BI and analytics, a common scenario is for an executive to check a metric and realize that it looks suspicious from a business perspective — resulting in a fire drill to fix the upstream data issue. This derails your team's productivity and worse, it makes it harder to feel confident about the data that's being used to make business decisions.

Improving your data quality

Data quality detection and monitoring

Data quality issues, like bugs in code, are best treated as inevitable. They can't be entirely prevented at the source. Instead, they need to be detected early so they can be quickly triaged and fixed. There are three main ways of monitoring for data issues:

1. Validation rules

Data engineering teams have long dedicated themselves to developing hard-coded rules that their data should conform to. For instance, an email might get sent to the engineering team if a certain column's values are outside of a given range.

Validation rules are easy to understand and to implement and will always be valuable as a way for experts to codify their expectations of the system according to the business's unique requirements. However, onboarding new data sources is time consuming, as existing rules need to be copied over and adapted and new rules have to be created. Furthermore, human-generated rules can't detect the "unknown unknowns": the issues in the data that no one anticipated.

2. Key metrics

A second way that teams detect data issues is to monitor key metrics and understand if they are deviating far outside of expectations.

Tracking key metrics is very valuable, providing a good high-level, high-visibility check on whether something is amiss in the data. Plus, dashboards and visualizations can be a great way to share data quality indicators with a wide variety of stakeholders. However, it's important to keep in mind that unless the data is already robust and well-known, it can be hard to distinguish a data quality issue from a true

©2022 Anomalo, Inc.

outlier or an unexpected—but real—trend in the data. And, it's easy to miss finer-grained issues that could nevertheless be impacting a segment of users or decisions.

3. Unsupervised learning models

A third, new way that teams are approaching data quality is to train unsupervised machine learning models to detect late, missing, and anomalous data. These models look at the trends in the data over a historical period, and then automatically learn what to expect and what to flag as suspicious. They can even detect unexpected changes in the relationships between multiple columns of data.

This approach requires little to no setup or maintenance to achieve a base layer of protection against obviously adverse changes in the data. Furthermore, unsupervised models can catch the "unknown unknown" issues that humans wouldn't have thought to look for. Similar to trends in metrics, it's possible—though less likely—for models to be fooled by sudden changes in the real world, but this can be counteracted by constraining the application to a well-defined data model.

How these approaches can work together

Ideally, the three approaches described above can be implemented in tandem, working together and complementing each other:

- Unsupervised learning can monitor **thousands of tables** for adverse changes, and drastically reduce the setup and maintenance costs of monitoring data at scale.
- Metrics monitoring can pay very close attention to **the most important data** that the business cares about — in contrast to unsupervised learning, which treats every row and column as equally important.
- Validation rules can find "**needles in the haystack**" when the data must be perfect in some way. Rules also allow users to express their expectations for the data from first principles, which can catch historical data quality issues that might already be present in the tables.

Root-cause analysis

Detecting that something is wrong with the data is only step one. From there, it's essential to understand *what* precisely is wrong, and *why*.

A data issue can often be traced to a single column inside a vast, multi-dimensional table. It's important that investigators can quickly unearth:

- Which table originated the error
- Which column had anomalous data
- How many rows were affected (the scope of the issue)
- A visualization of the timeline and impact of the issue
- An analysis of good and bad data samples, to determine if the issue is isolated to any specific data segments

Triage, alerting, and resolution

The next step is to triage data quality issues. A triage process uses the root cause analysis to determine two things: 1) what's the priority of the issue (Does it need to be fixed immediately? Does it need to be fixed at all?), and 2) who is the owner/channel that should be notified to resolve the issue.

Common mistakes to watch out for here include sending too many low-priority or false positive alerts, resulting in alert fatigue; not having clear owners, so everyone thinks it's someone else's responsibility to handle the fix; and not sending an alert to the right specific channel, whether that's MS Teams, Slack, email, PagerDuty, or elsewhere.

Finally, it's important to provide visibility into the resolution process, as many others might be depending on a data quality fix. There should be a link to the ticketing system (Jira, ServiceNow, etc.) where the work to resolve the root cause will be undertaken, so that interested parties, such as managers and teammates, can track progress.

When to consider a data quality platform

Build or buy?

When considering any third-party vendor, it's important to evaluate the long-term cost of the commitment. Here are some questions to help determine whether a third-party solution is right for your organization:

- Can non-technical stakeholders currently self-service by creating their own data quality checks if they're experts in their own data sources?
- Do you have machine learning expertise and are you willing to invest in building models that can detect anomalous data?
- How fast is your team at root-causing data issues? Can you quickly onboard new engineers and get them up to speed on fixing data quality failures?
- How much does your organization spend maintaining data quality rules? Do you have the resources to grow these rules as the data changes?
- Can you summarize data quality visually to your stakeholders (like managers and executives)?
- Have you had a data quality problem in the last 6 months that impacted the bottom line of the business or eroded trust in the data platform or data team?
- Do you have robust data quality coverage for your most critical enterprise-wide datasets, and do all of your data-driven teams have coverage for the datasets they rely upon?

Key features of a data quality platform

If you've decided to investigate a data quality platform rather than building a homegrown solution, the next step is to decide what features matter most to you. We've built a helpful table to navigate the different offerings—and at the risk of showing our bias here, we added a column explaining what Anomalo can deliver in each area.

FEATURE	DESCRIPTION	Anomalo
Warehouse Integrations	What data warehouses / data lakes the platform supports	Turn-key integrations with Databricks, Snowflake, BigQuery, Redshift, Presto, and Athena
Alerting	How easy it is to sign up for and customize alerts from the platform	Set up alerts for Slack, Microsoft Teams, Pagerduty, email. Webhooks + full-featured API available to consume alerts
Rules	How easy it is for any stakeholder, even if non-technical, to define data quality checks	Configure any number of no code rules from an user-friendly web interface. More technical users can define arbitrarily complex SQL-based rules
Key Metrics	The ability to define and track business KPIs alongside other quality metrics	Configure any number of no- code KPIs from the user interface. More technical users can define arbitrarily complex SQL-based KPIs. Metrics can be segmented to avoid creating alert fatigue from false-positive notifications.
Automation	What the platform does for you out of the box (e.g. via machine learning) without having to configure rules for your tables	Instantly start monitoring tables for late, missing, and anomalous data with one click. Unsupervised models detect "unknown unknowns" and automatically learn and improve over time
Resilience	Whether the automated checks configured by the platform can adapt as your data, business, and organization evolve	Unsupervised learning algorithms are re-trained every day and will automatically adjust as your data changes, without requiring human intervention
Root Cause Analysis	How capable the platform is at helping teams understand the source of the data quality issue	Detailed incident reports with information about the location of the issue in the table along with samples of good and bad data
Triaging	How the platform facilitates the triaging of issues, so that users can facilitate issue resolution and managers can track and report on resolution status	Supports issue acknowledgement and resolution via a triage flow, maintains a full history of these actions, and can trigger issue or ticket creation in other platforms via API
Visualizations	The richness of dashboards and visuals that the tool provides by default	Visualizations optimized for anomaly detection and root cause analysis. Quick to insight and quick to communicate
Executive Summary	How easy it is to use the tool to communicate data quality at a high-level for managers, executives, etc.	Anomalo Pulse is a dashboard that helps all users get on the same page about data quality (e.g. repeat offenders, how many tables are being monitored) and create strategies to improve quality across the organization
Customer Support / SLAs	What level of support and guarantees the company provides to customers	Support on dedicated Slack channels with max 24 hour response times, weekly office hours, full-service deployment aid, and a team that is focused on users getting as much value out of the platform as possible

©2022 Anomalo, Inc.

We built Anomalo to have all these features and more because we think quality data is essential for any team that wants to make better decisions with confidence. Anomalo is **intelligent**, automatic, and simple: intelligent because of features like unsupervised ML monitoring, root cause analysis, and time series modeling; **automatic** because we instantly scale and grow with your organization's data as the size and complexity increases; and **simple** because non-technical users can self-service as the role of data and data quality expands and impacts many different business functions.

Interested in learning what Anomalo can do for you? Sign up for a free demo.

©2022 Anomalo, Inc.