Anomalo

The definitive guide to data quality monitoring vendor selection

Are you searching for a data quality monitoring solution that fits the specific needs of your team and data? You're not alone – many businesses today face this same challenge. However, without a clear picture of what you need from your monitoring, it can be difficult to decide which solution fits your unique needs.

Here, we've compiled a comprehensive guide to help you navigate the data quality monitoring landscape. We'll explain what data quality platforms can offer you and help you understand what you should look for. We'll walk you through the key factors you need to consider when selecting a data quality monitoring vendor, so you can protect your data and make informed decisions.

Contents

What is data quality monitoring?

Data quality monitoring is the process of detecting and resolving complicated data issues that may impact a business due to the inaccuracy of BI dashboards, reports, or predictive models.

Companies now use data throughout all business aspects to unlock next-generation capabilities. However, just as it's easier to take advantage of more data sources and volume than ever before, it's also easier for issues to creep into your data.

Whether internally or externally obtained, source data is never perfect – it can contain errors that can skew your insights or, worse, cause customer issues. Unexpected changes in data values can be tough to detect and understand since there are often many possible explanations to evaluate. Was a spike in purchases caused by an error in your data pipeline or by a seasonal trend? Is your ML model performing poorly because its input data has suddenly changed or because of a flaw in the model design?

The most common problems enterprises experience are:

- **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 data inputs because one of your data sources changed its schema without letting you know beforehand.
- **Unexpected changes in the values** of data because somewhere along the data pipeline of sourcing, ingesting, cleaning, extracting, joining, and transforming the data, an error was introduced.

Data quality issues can quickly spiral out of control if not adequately managed, and worse, they can be hard to spot. Data quality monitoring is crucial in ensuring your data is accurate, complete, and consistent. By implementing a data quality monitoring solution, businesses can detect data quality issues early on and take corrective measures before they impact their operations.

How to evaluate a data quality monitoring platform

When evaluating a data quality monitoring platform, it's essential to consider how it fits the needs of your business. Here's an overview of the factors you should keep in mind for evaluating a solution. We'll then expand on these in the next section.

Three key pillars of data quality solutions

Fundamentally, data quality monitoring solutions provide three key capabilities:

1. **Detect:** A quality data monitoring platform should be able to detect data quality issues like inconsistencies, errors, or drift.
2. **Alert:** The platform should be able to alert the right person or team at the right time of data quality issues.
3. **Resolve:** The platform should help investigate, triage, and ultimately remediate data issues.

In order to protect your data quality and empower your team, make sure that your platform of choice can help you do all three of these.

Additional considerations

In addition to those three key pillars of functionality, other details to consider when selecting a data quality monitoring partner will depend on your needs. These include:

1. **Security:** The technical and security measures executed by the data quality partner should meet the needs of your data. For instance, if you have compliance requirements for your data.
2. **Support:** The level of enterprise support provided by the data quality partner should be sufficient to meet the needs of your team.

Accessibility

An overarching point to keep in mind is to evaluate the accessibility of the platform for your team. You should ensure the data quality monitoring platform you choose is user-friendly for your whole team.

As your data engineers consider all of the above functionality, they should consult analysts or other non-technical users to ensure the solution will also work for them. Is engineering effort needed to set up new monitoring or analyze an issue?

What to look for

Now that we have a high-level picture, let's describe the features you can look for to evaluate data quality monitoring solutions: detect, alert, resolve, security, and support.

Detect: The ability to easily detect data quality issues

Foundational data quality checks	Checks that confirm data in a table is complete and has arrived on time.
Validation rules	Allows experts to specify hard and fast rules about their data.
Metric checks	Monitors changes in key business or data quality metrics.
Unsupervised data quality monitoring	Leverages an algorithm to learn the structure of your data and monitor for significant unexpected changes; tests can learn a representation of the typical data in a table. As new data arrives, the platform will use machine learning to detect if that data is meaningfully different from what appeared in the table before.
Detect missing data	Detects increases in NULL, Zero or Empty string values, drops in segments.
Detect unusual anomalies	Detects duplicate records, dropped columns, or other unusual anomalies.
Automation	Automatically executes data quality checks without requiring user input.
Low-code / No-code UI	Offers self-service creation and execution of data quality rules by non-technical stakeholders. Could include features like a predefined rule library.
Coverage of data	Offers a breadth of integrations into different cloud data warehouses.
Cross database coverage	Monitors and compares data within a single enterprise across multiple cloud environments ("hybrid-cloud data quality").
Prioritization of data quality rules	Allows you to prioritize quality issues by setting the priority level for each individual check, as well as aggregate checks.
In-pipeline data quality checks	Uses an APIs to run data quality checks in data pipelines. For example, in Apache Airflow, you could use an API to execute data quality checks on transformed data, poll the check results, and publish the data if there are no failures. If a check does fail, you could run automated tasks to fix the bad data, then abort the remainder of the DAG or quarantine bad records.
Automating/ scheduling of data quality checks	Offers flexible configuration to monitor a table based on a set schedule, data updates, or when initiated by a user.
Scale	Supports an unlimited number of data profiles, table configurations, and data quality rules.
API	Provides an API to modify all UI changes.

Alert: The ability to alert the right user at the right time of data quality issues

Visualizations	Provides notifications that include relevant information and samples of raw data that highlight good and bad values.
Supported platforms	Allows you to receive alerts to the platform your team uses to run daily business (Slack, Microsoft teams, Gmail, PagerDuty, etc).
False positive suppression	Has mechanisms to reduce false positive and false negative alerts and to mitigate their impact when they do occur.
Team and role-based management	Allows targeting of alerts to select teams and people within your organization.
No-code configuration changes	Allows users to adjust their checks easily, with an audit trail of changes to allow for easy reversion.
Feedback loop	Offers machine learning tools for tuning the platform and suppressing alerts that are not useful.
Data catalog and BI integrations	Integrates data quality with data governance/catalog providers and BI tools to show data quality checks within important dashboards.

Resolve: The ability to help investigate, triage, and ultimately remediate data issues

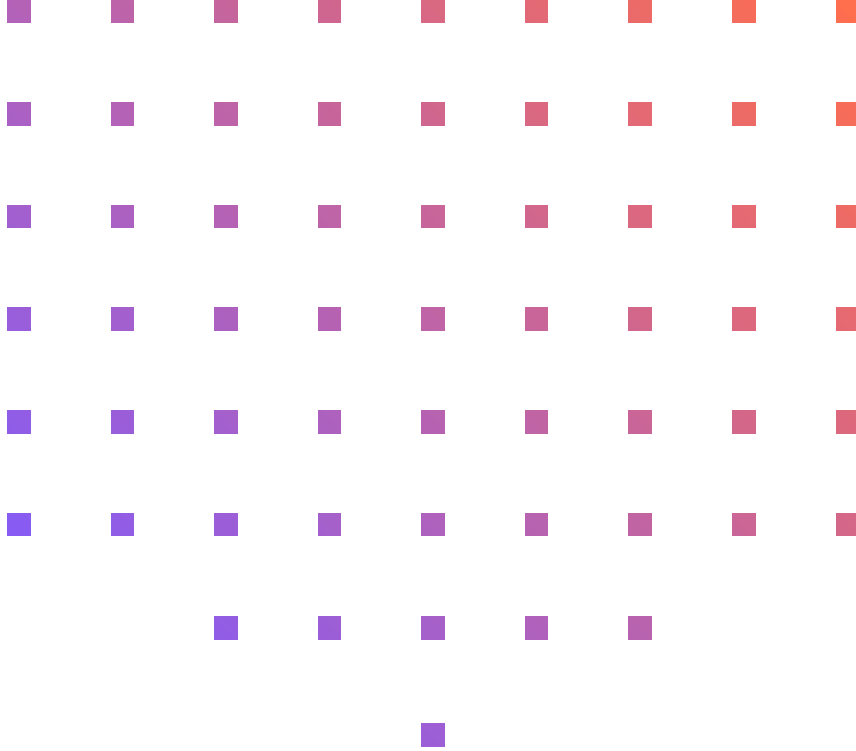
Root cause analysis	Offers tools to diagnose the root cause of data problems. For example, the platform could analyze all the possible data segments independently for the presence of an issue or anomaly and cluster the results, then surface the segments with the greatest explanatory power.
Visuals	Provides rich visuals with context, so you can quickly understand anomalies within your data and drill down as needed.
Triage	Provides a triage flow and integrates with your organization's remediation steps (e.g. ticketing systems).
Interoperability	Integrates your metadata repository, business glossary, and existing data to identify data quality issues or extract critical data elements.
Metadata checks / data lineage	Allows tracking of upstream tables that are causing data quality issues as well as their downstream tables, in order to alert the user to the potential effects of the failed data quality checks.
Dashboarding / custom report	Provides macro-statistics and trends related to data quality coverage.

Security: The technical and security measures executed by the data quality monitoring vendor

In-VPC	Keeps data inside your environment. Can be deployed in your cloud environment (i.e. on a VM instance or Kubernetes) with read-only access to data, limited access to other cloud resources, and no data leaving the cloud environment.
SOC 2	Offers SOC 2 compliance.
SSO	Offers Single Sign On as a default for new users.
GDPR	Offers GDPR compliance.
HIPAA	Offers HIPAA compliance.
ISO 27001	Is ISO 27001 certified.
Role-based access	Supports multiple levels of role-based access control such as admin, editor, restricted editor, viewer and restricted viewer that can be assigned and controlled via the UI.
Data processing and retention	Ensures data that is accessed and processed by the data quality vendor will be processed in memory only and will never be stored locally.

Support: The level of enterprise support provided by the data quality partner

Dedicated support	Provides your team a dedicated business partner to answer questions and give product guidance.
Data solutions architect support	Provides a technical resource who can advise you on how your data is structured, and what configuration and test options are most appropriate for advanced use cases.
Real-time support	Offers a dedicated Slack channel for asynchronous support.
Query optimization	Optimizes queries to alleviate unnecessary consumption costs for your cloud data warehouse.



Choose a solution with confidence

With this knowledge in hand, you should be able to look at your options and speak fluently about what best fits your team's needs. As you evaluate, consult with both technical and non-technical users to make sure everyone will find the solution helpful. When you make an informed decision about which data quality monitoring solution to use, you can rest assured that your team is making a wise investment in the protection of your data.

To learn more about Anomalo and how we make detection, alerting, and resolution easy for everyone on your team, visit anomalo.com.