The Ultimate Data Observability Checklist

Getting started with data observability?
Here are the 5 things every data observability platform needs to help companies achieve data trust.

MONTE CARLO

Overview

What is Data Observability?

For the past decade or so, software engineers have leveraged targeted solutions like New Relic and DataDog to ensure high application uptime while keeping downtime to a minimum.

In data, we call this phenomena data downtime. Data downtime refers to periods of time when data is partial, erroneous, missing, or otherwise inaccurate, and it only multiplies as data systems become increasingly complex, supporting an endless ecosystem of sources and consumers.

The good news? By applying the same principles of software application observability and reliability to data, these issues can be identified, resolved and even prevented. Data Observability is an organization's ability to fully understand the health of the data in their system by eliminating data downtime via best practices of DevOps and software engineering.

DATA OBSERVABILITY PILLARS

Freshness | Distribution | Volume | Schema | Lineage

Building data observability platforms

Data observability platforms are the newest layer in the modern data stack, helping teams monitor the health of critical data assets and pipelines while building organizational trust in data at scale.

Read on for the definitive checklist to evaluate data observability platforms across five key areas: visibility, troubleshooting, self-serve tooling, data discovery and metadata management, and security.

End-to-End Visibility 

To ensure your data team is the first to know about data downtime through automated monitoring and alerting, your data observability platform should:

1. Infer information about table operations, such as load patterns and expected volume
2. Detect anomalies based on historical data and patterns
3. Track table updates and alert teams when updates don't occur as expected
4. Track changes in data volume in individual tables and alert teams to abnormal size changes
5. Track and alert on schema changes, distribution changes in low cardinality fields, and null rates, uniqueness, and other changes in values within select fields
6. Allow team members to create custom thresholds, including multiple/dual thresholds, for anomalies
7. Group related anomalies across tables based on inferred dependencies

"We have 10% of the incidents we had a year ago...I think every data engineer has to have this level of monitoring in order to do this work in an efficient away."

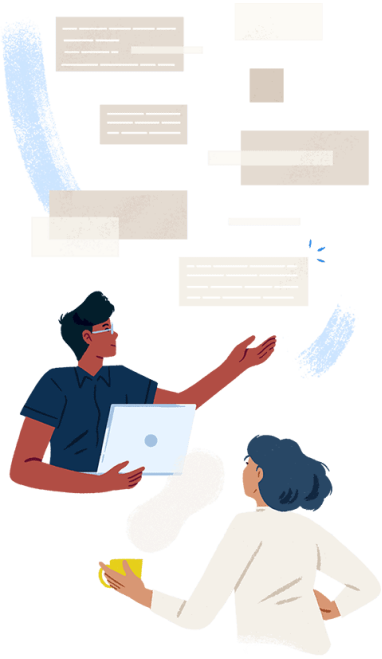
Daniel Rimon, Head of Data Engineering at Resident

MC

Rapid, ML-based detection and resolution of data downtime

To help your team resolve data quality issues swiftly and automatically, your data observability platform should:

1. Automatically create data lineages to display upstream and downstream data relations, including BI reports and dashboards
2. Filter and intelligently route alerts by dataset based on dataset owners
3. Automatically understand and prioritize issue resolution based on business impact
4. Enable incident management collaboration in a centralized interface with comprehensive activity logs to speed up root cause analysis across each stage of the pipeline
5. Offer API access to all information presented in the UI for customization and/or workflow integration



“Being able to quickly identify client-facing issues and be proactive is really the key to building trust in our data.”

Patrick Campbell, Lead Data Engineer at Optoro

MC

Unified, self-service platform

When it comes to data trust, you should be able to understand the health of your data from a central, all-in-one UI. Long gone are the days of data silos and playing the bad data name game between data engineer and analyst teams. With data observability, all stakeholders are able to collaborate in a single, self-service platform. This interface should:

1. Make it easy to search for and explore data assets with a simple UI
2. Collect and display information required for investigating and resolving issues
3. Deliver all the relevant information required to conduct root cause analysis, down to the field level
4. Maps out data incidents over time to that make it easy to view impacted tables, and every action that was taken to manage and resolve an incident
5. Share comprehensive query logs that reveal periodic ETL queries, ad hoc/backfill queries, changes in query patterns, and more hints that help teams identify the root cause of data incidents.
6. Seamlessly connect to Slack, Opsgenie, PagerDuty, webhooks, email, or your communication channel of choice to alert about downtime to the individuals who need to know
7. Display sample data, to help users immediately understand what data involved in the incidents looks like, and what typical data looks like

MC

Automated data discovery and metadata management

To support the growing demand for data democratization and decentralized data ownership, your data observability platform should:

1. Dynamically create a data catalog that enables data discoverability and searchability
2. Include self-service diagnostic tools that perform data profiling and understand data lineage
3. Provide standard reporting for data quality dimensions on data sets
4. Deliver value-add insights on table importance, monitor coverage, unused tables, and other information
5. Provide information on queries with deteriorating performance
6. Offer a centralized interface for self-service incident analysis, impact assessments, and cleansing requirements
7. Allow users to track and discover details on any dataset or environment
8. Automatically update schema metadata and information, without requiring any manual changes

"The self-service capabilities of data observability helped build back trust in data, as users were seeing us in action: going from a red alert to a blue "work-in-progress" to "resolved" in green. They knew who was accountable, they knew the teams were working on it, and everything became crystal clear."

-

Gopi Krishnamurthy, Director of Engineering at Blinkist

MC

Security-first architecture

To ensure your data's full protection and security, your data observability platform should:

1. Monitor data at rest by extracting query logs, metadata, and statistics about data usage—without exposing your data warehouse, lake, or other infrastructure to external environments
2. Offer SOC-2 Type II certification
3. Never extract or store individual records, PII, or other sensitive information outside of your environment
4. Allow you to comply with HIPAA, PCI, GDPR, CCPA, FINRA, and other compliance frameworks that you are subjected to
5. Allow easy and simple deployment with little to no ongoing operational overhead and frequent automatic upgrades

"Data Observability allows my team to understand what data is important for the business, as well as whether or not this data can be trusted. A unified interface helps draw these connections between critical data tables and the reports the company relies on to make decisions."

Satish Rane, Head of Engineering, ThredUp

MC

Interested in learning more about data observability?

- [Stay up-to-date will all things data on the Data Downtime Blog]

- [Register for IMPACT: The Data Observability Summit]

- [Request a demo of Monte Carlo]

MC MONTE CARLO