IBM Data Observability by Databand

Deliver reliable and trustworthy data with Databand

Highlights

- Understand pipeline execution health
- Alert on pipeline latency
- Check data sanity
- Analyze data trends
- Identify data access and business impacts

What is data observability?
Data runs the world. Seriously. It's estimated that we create 2.5 quintillion bytes of data each day. That's 18 zeros.

This influx of data has created challenges for data-driven organizations, largely because they are not fully prepared for today's volume of data, the variety of data sources, and the complex infrastructure.

Additionally, there's been too much of a focus on analytics engineering over data engineering—the latter of which is essential to making sure data quality is in a good place to power those advanced analytics.

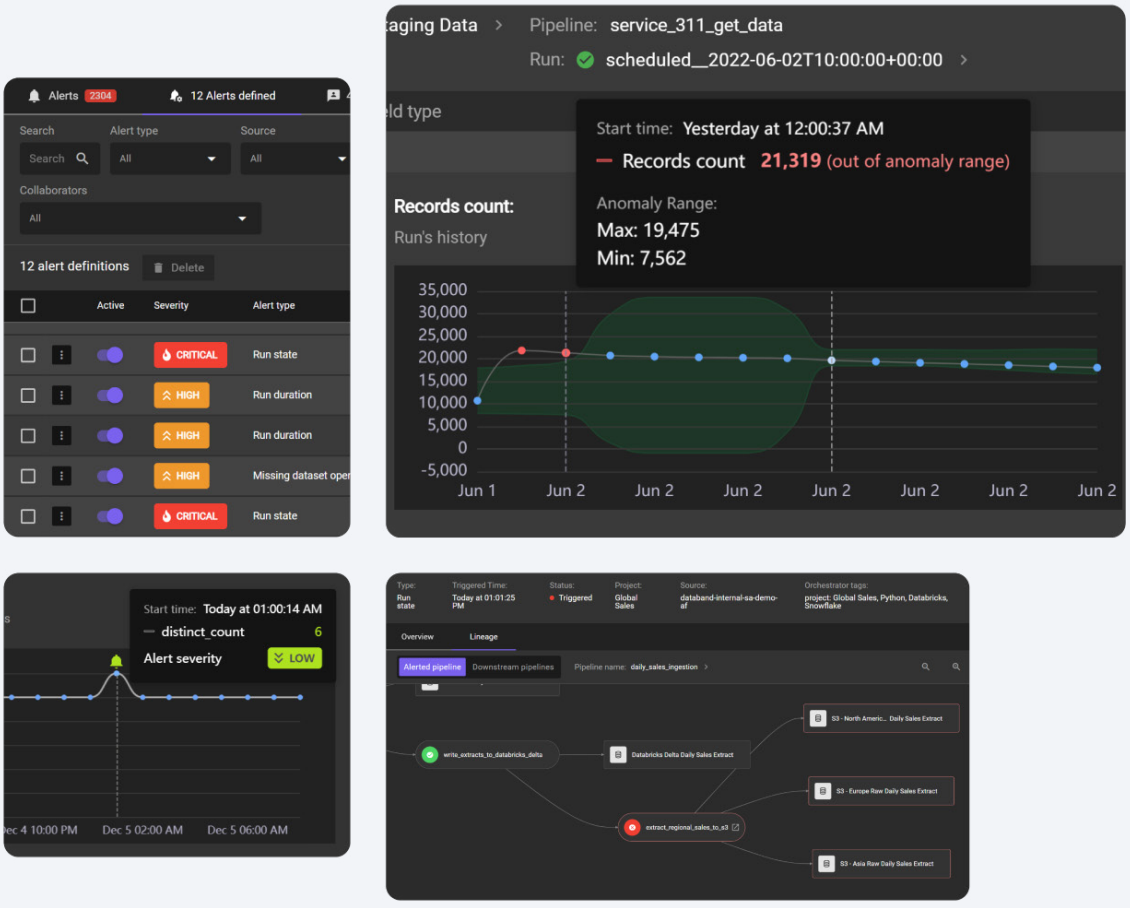
When so much hinges on having good data in place, a lot can go wrong in a highly visible way. Just ask any data or engineering team that's received a frantic call from the CEO. Enter data observability.

Data observability is about understanding your system's health and state of data. It relies on several activities and technologies to enable teams to collect, profile, alert, and resolve data issues in near real-time.

Data observability needs to be infused consistently throughout the end-to-end data lifecycle. That way, all activities involved are standardized and centralized across teams for a clear and uninterrupted view of issues and impacts across the organization.

Of course, achieving data observability is easier said than done (because isn't everything?). But it's not impossible by any stretch—it simply requires the right approach.

This solution brief outlines how Databand proactively resolves the common data observability challenges.



Solution description:

1. Automatically collect metadata to gain immediate visibility into critical metadata. This visibility gives analysts and scientists a standard method for customized data quality validations.
2. Build historical baseline based on common run and data behaviors, essentially profiling the data pipeline landscape.
3. Once the historical baseline is established, alert on anomalies and rules based on deviations relative to the profile and/or rule breaches.
4. Create smart workflows that automatically remediate data quality issues and keep data deliveries on track.

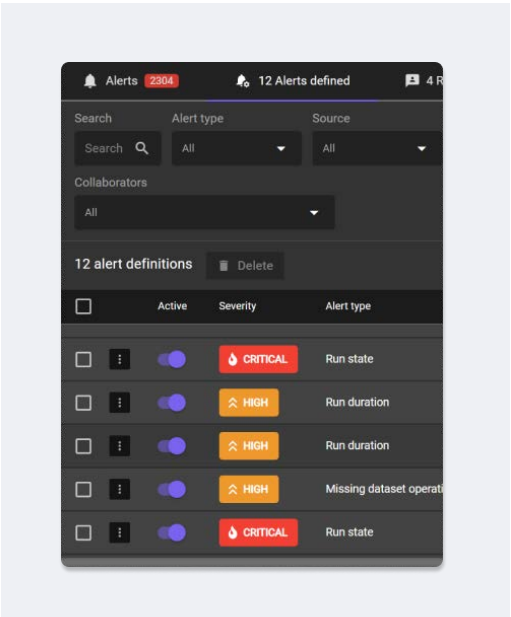
Pipeline monitoring

Understand pipeline execution health

The very first thing to understand is whether or not your data pipeline is executing correctly, as this impacts the flow of data. Specifically, a data orchestrator like Airflow or IBM DataStage® will help create pipelines to run data through, but how do you know whether the data is moving through them as planned?

If the pipeline fails, no data can come through, and everything else that follows won't execute correctly.

The Databand solution offers a clear way to visualize pipeline health and ensure your pipelines are executing as expected. Additionally, the solution makes it possible to quickly isolate where the issues are occurring so teams can jump into fix mode right away.



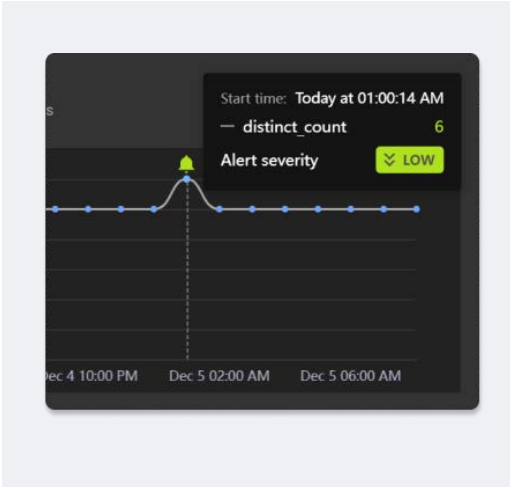
Alert on pipeline latency

Alert on pipeline latency

Next, it's important to determine if data arrives in a usable time window based on when it's expected to come through. For example, a company expects to have data in-hand for users at a certain time for them to make critical business decisions.

Any data that arrives late or not at all creates serious problems. As a result, pipeline latency—or the amount of time it should take data to arrive from point A to point B—is essential to nail down and maintain on an ongoing basis.

The Databand solution helps address pipeline latency issues by keeping tabs on how often a specific process runs and how long that run actually takes versus the average run. Through analysis of the expected and actual runtime, Databand provides a clear indication of an issue that requires further attention.



Check data sanity

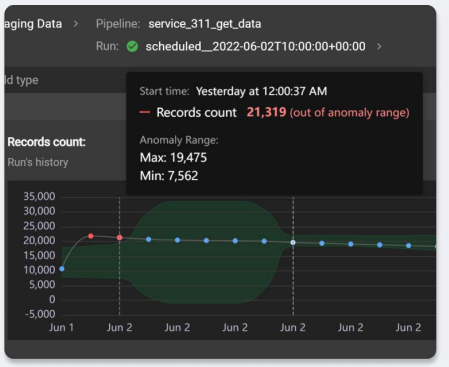
Check data sanity

Once the data arrives, does it come through valid and complete, or are there any errors? This data sanity check is critical because even if data arrives on time, it doesn't do any good if it's error-ridden.

Data sanity centers around any changes to the data schema, such as whether or not there are more or fewer columns than expected.

Checking data sanity at scale—think millions of data points moving through the pipelines—can get very difficult, requiring the right team and dedicated technology.

Databand provides an easy way to pull up each data schema and get a bird's eye view of the columns while allowing for a closer look at the data contained within to look for anomalies.



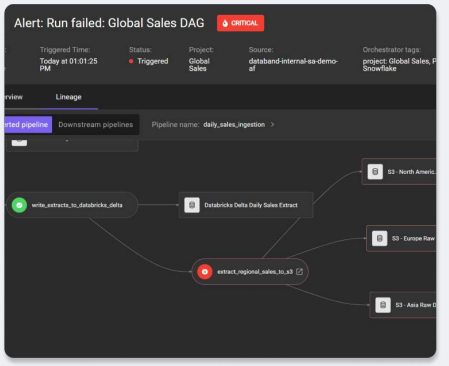
Analyze data trends

Analyze data trends

While data sanity looks at the data schema to confirm that the data coming in is structurally okay, there's still more to consider. From there, it's essential to dig into the dataset trends to determine what causes these changes to keep happening.

For instance, what if one of the columns is structurally right, but its data is wildly different from expected?

Databand provides the ability to look closely at not just how the data is changing but what within it is changing, including how any structural changes might impact the health of the data over time, where issues are occurring, and what needs to be corrected.



Identify business impact

Identify data access and business impact

The last step brings everything together to identify the business impact of issues at any point by understanding how the data maps to key decisions.

Gaining this understanding requires investigating how the business uses each set of data. That way, it's easy to identify the downstream impacts of any issues in terms of latency, sanity, trends, or anything else.

Databand provides a means of understanding the data flow. Looking at lineage, for example, can help identify the end destination and the business impacts.

It's important to look closely at this view and related alerts to understand the different data sets that will be affected by any issues associated with a particular pipeline or data run.

Get started with Databand

Implement proactive data observability today and know when there's a data health issue before your consumers do.

Discover how you can get alerts on leading indicators of data pipeline health issues and drill deep into the data to implement a fix before bad data gets through.

Why IBM?

Deployed through the IBM data fabric architecture, Databand helps you deliver reliable data to your business by ensuring your organization can

- Automatically observe dynamic data pipelines
- Proactively address data quality and reliability
- Continuously monitor AI/ML reliability across data and model

For more information:
To learn more about IBM Data Observability by Databand, please contact your IBM representative or IBM Business Partner, or visit databand.ai/request-demo to book a demo.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
December 2022