The Top Data Quality Metrics You Need to Know (With Examples)

Contents

Data quality metrics can be a touchy subject, especially within the focus of data observability.

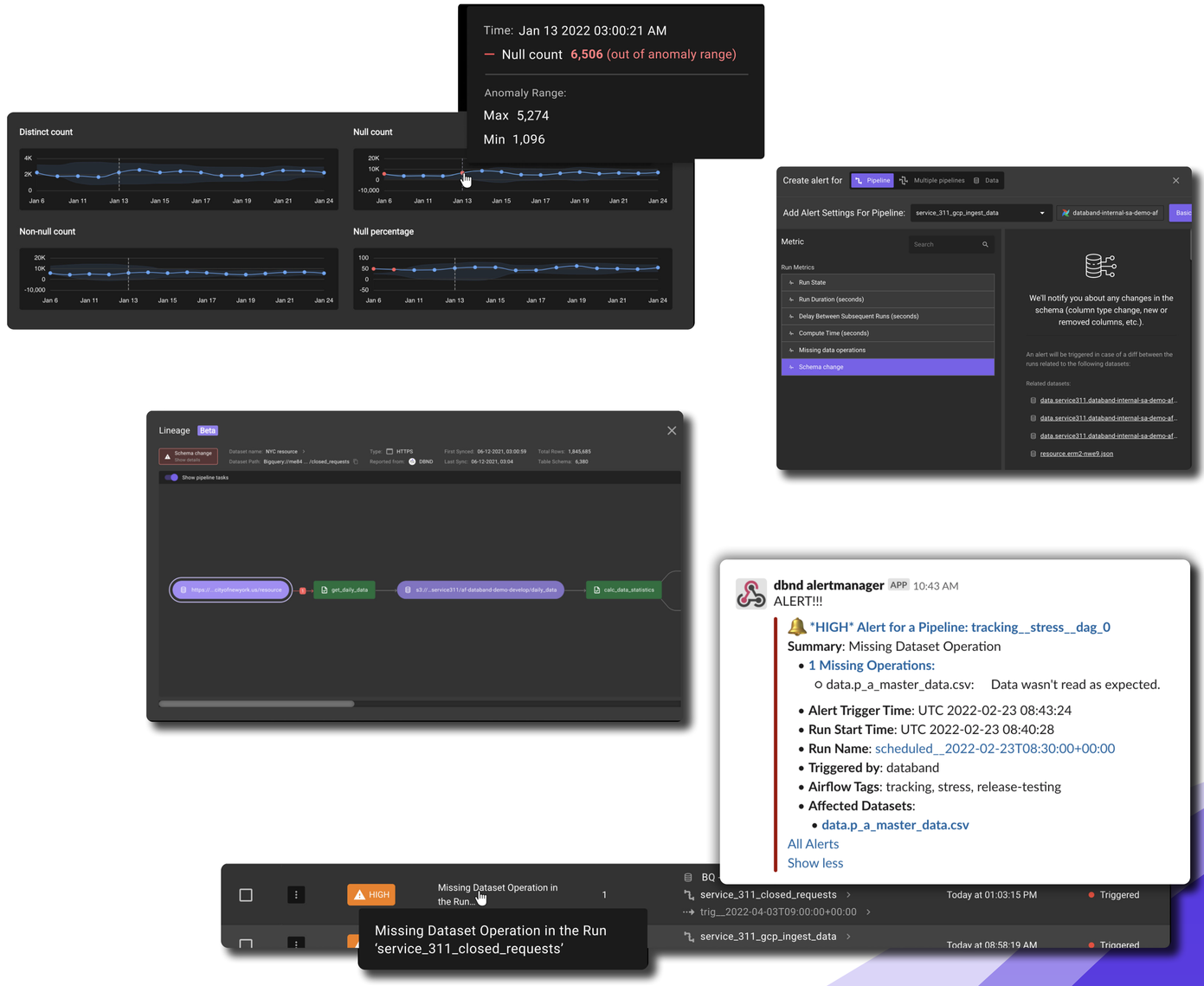
A quick google search will show that data quality metrics involve all sorts of categories.

For example, completeness, consistency, conformity, accuracy, integrity, timeliness, continuity, availability, reliability, reproducibility, searchability, comparability, and probably ten other categories I forgot to mention all relate to data quality.

So what are the right metrics to track? Well, we're glad you asked.

We've compiled a list of the top data quality metrics that you can use to measure the quality of the data in your environment. Plus, we've added a few screenshots that highlight each data quality metric you can view in Databand's observability platform.

Take a look and let us know what other metrics you think we need to add!



Metric 1 # of Nulls in Different Columns

Who's it for?

- Data engineers
- Data analysts

How to track it?

Calculate the number of nulls, non-null counts, and null percentages per column so users can set an alert on those metrics.

Why it's important?

Since a null is the absence of value, you want to be aware of any nulls that pass through your data workflows.
For example, downstream processes might be damaged if the data used is now "null" instead of actual data.

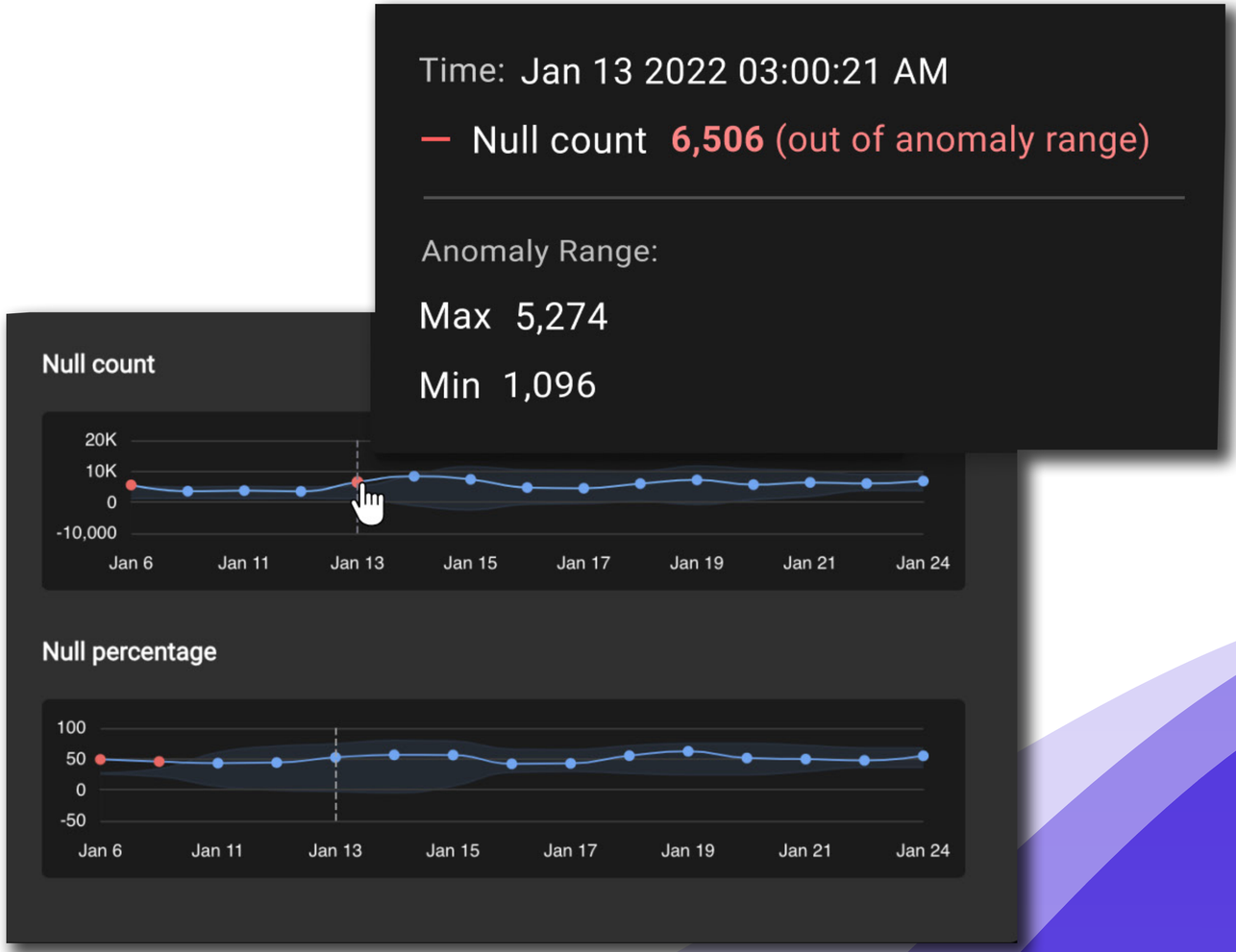
Dropped columns

The values of a column might be "dropped" by mistake when the data processes are not performing as expected.
This might cause the entire column to disappear, which would make the issue easier to see. But sometimes, all of its values will be null.

Data drift

The data of a column might slowly drift into "nullness."
This is more difficult to detect than the above since the change is more gradual. Monitoring anomalies in the percentage of nulls across different columns should make it easier to see.

What's it look like?



Metric 2 # Frequency of Schema Changes

Who's it for?

- Data engineers
- Data scientists
- Data analysts

How to track it?

Tracking all changes in the schema for all the datasets related to a certain job.

Why it's important?

Schema changes are key signals of bad quality data.
In a healthy situation, schema changes are communicated in advance and are not frequent since many processes rely on the number of columns and their type in each table to be stable.

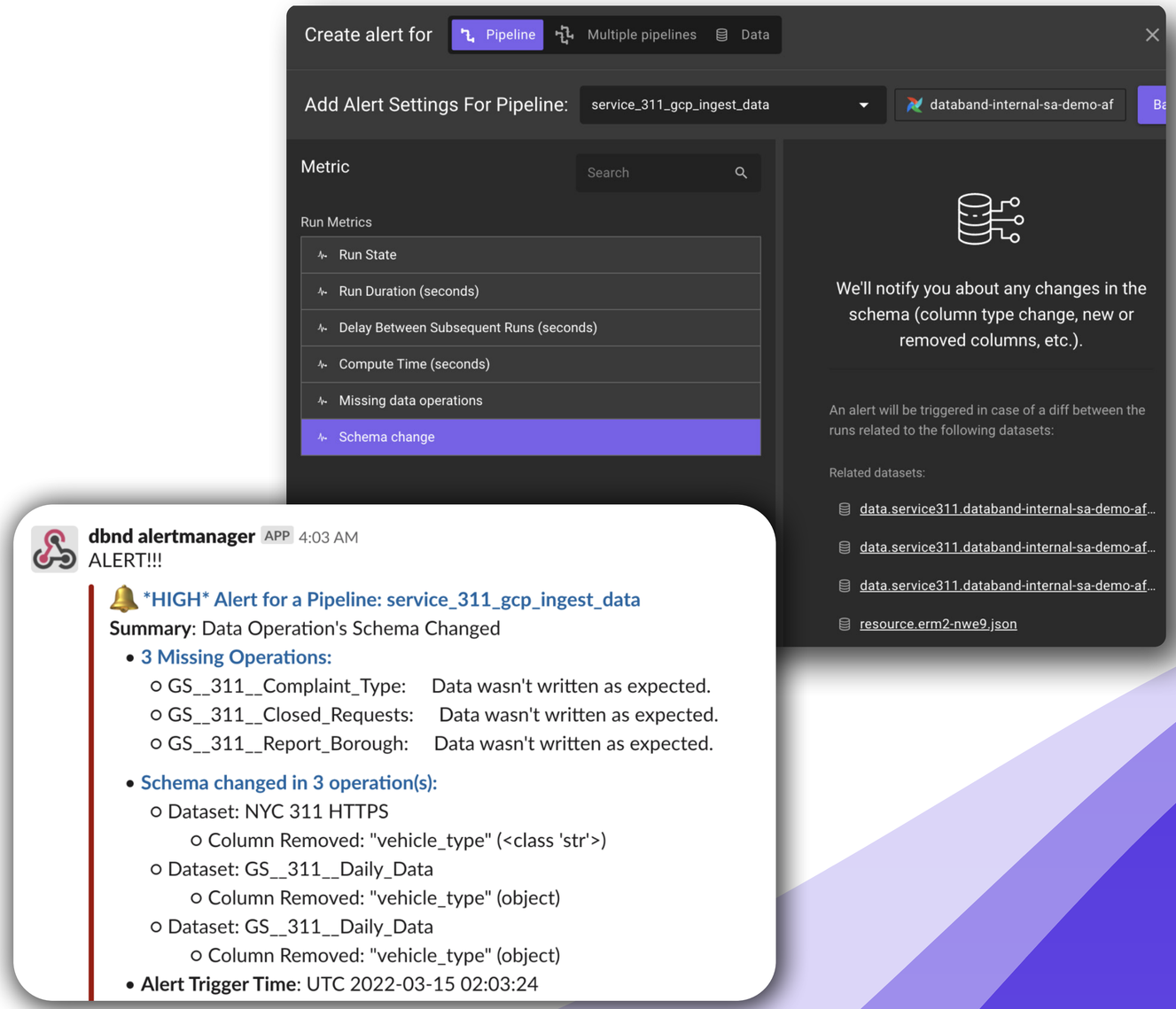
Frequent changes might indicate an unreliable data source and problematic DataOps practices, resulting in downstream data issues.

Examples of changes in the schema can be:

- Column type changes
- New columns
- Removed columns

Go beyond having a good understanding of what changed in the schema and evaluate the effect this change will have on downstream pipelines and datasets.

What's it look like?



Metric 3 # Data Lineage

Who's it for?

- Data engineers
- Data analysts

How to track it?

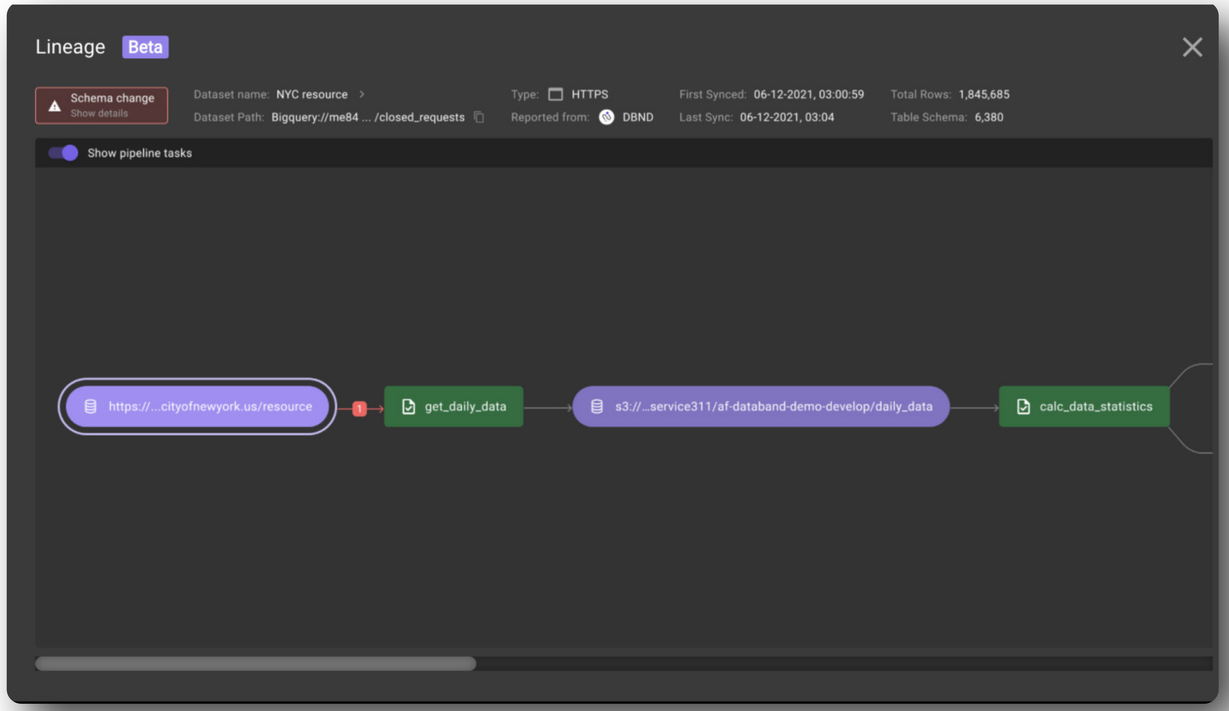
Track the data linage with assets that appear downstream from a dataset with an issue. This includes datasets and pipelines that consume the upstream dataset's data.

Why it's important?

The more damaged data assets are downstream (i.e., datasets, pipelines), the bigger the data issue's impact. This metric helps the data engineers understand the severity of their issues and how fast they should fix it.

It is also an important metric for data analysts because most downstream datasets make up their company's BI reports.

What's it look like?



Metric 4 # Of Pipeline Failures

Who's it for?

- Data engineers
- Data executives

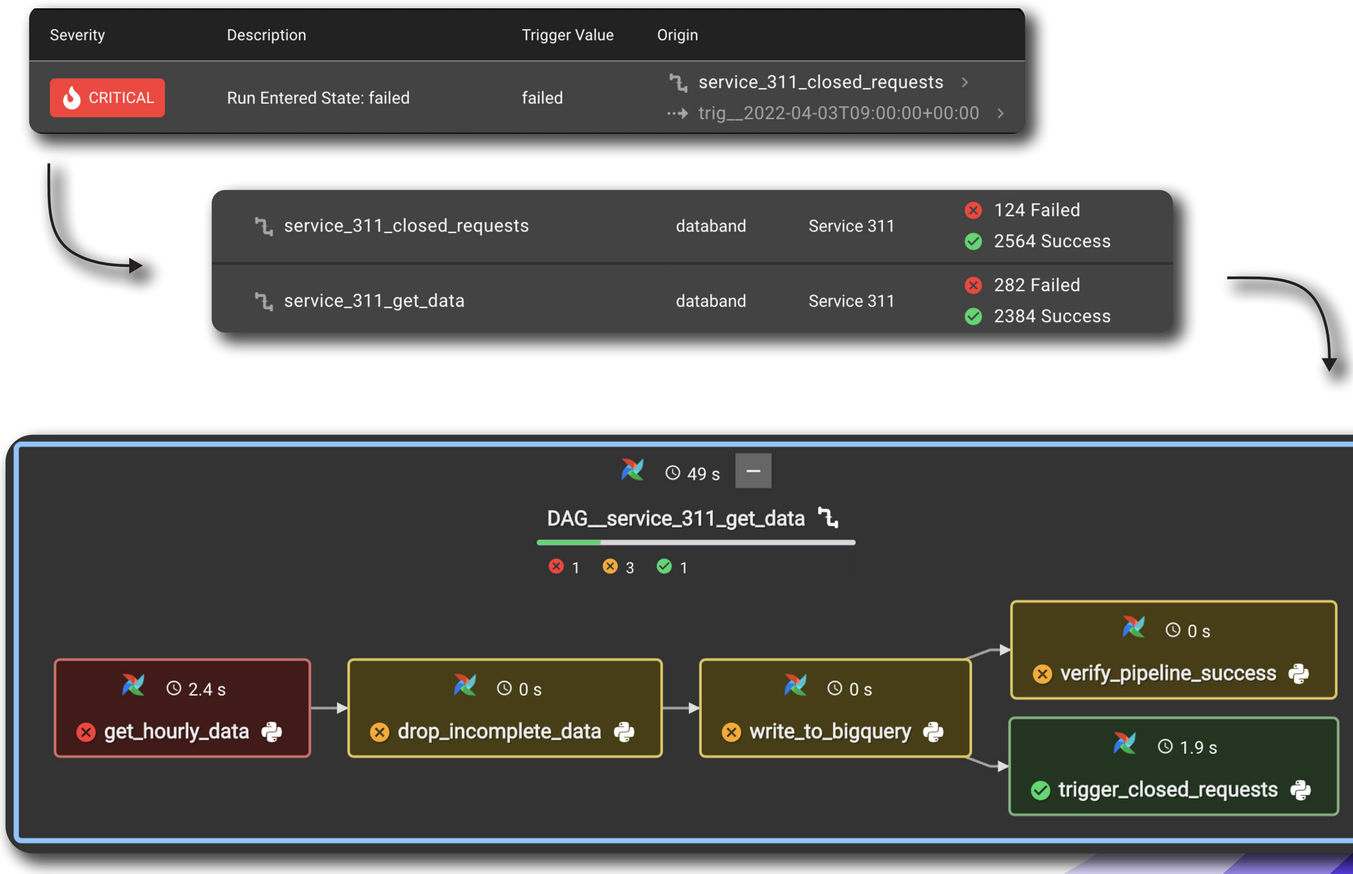
How to track it?

Track the number of failed pipelines over time. Use tools that highlight root cause analysis, and show deep dives inside all the tasks that the DAG contains to understand why pipelines fail.

Why it's important?

The more pipelines fail, the more data health issues you'll have. Each pipeline failure causes issues like missing data operations, schema changes, and data freshness. If you're experiencing many failures, this indicates severe problems at the root that needs to be addressed.

What's it look like?



Metric 5 # Pipeline Duration

Who's it for?

- Data engineers

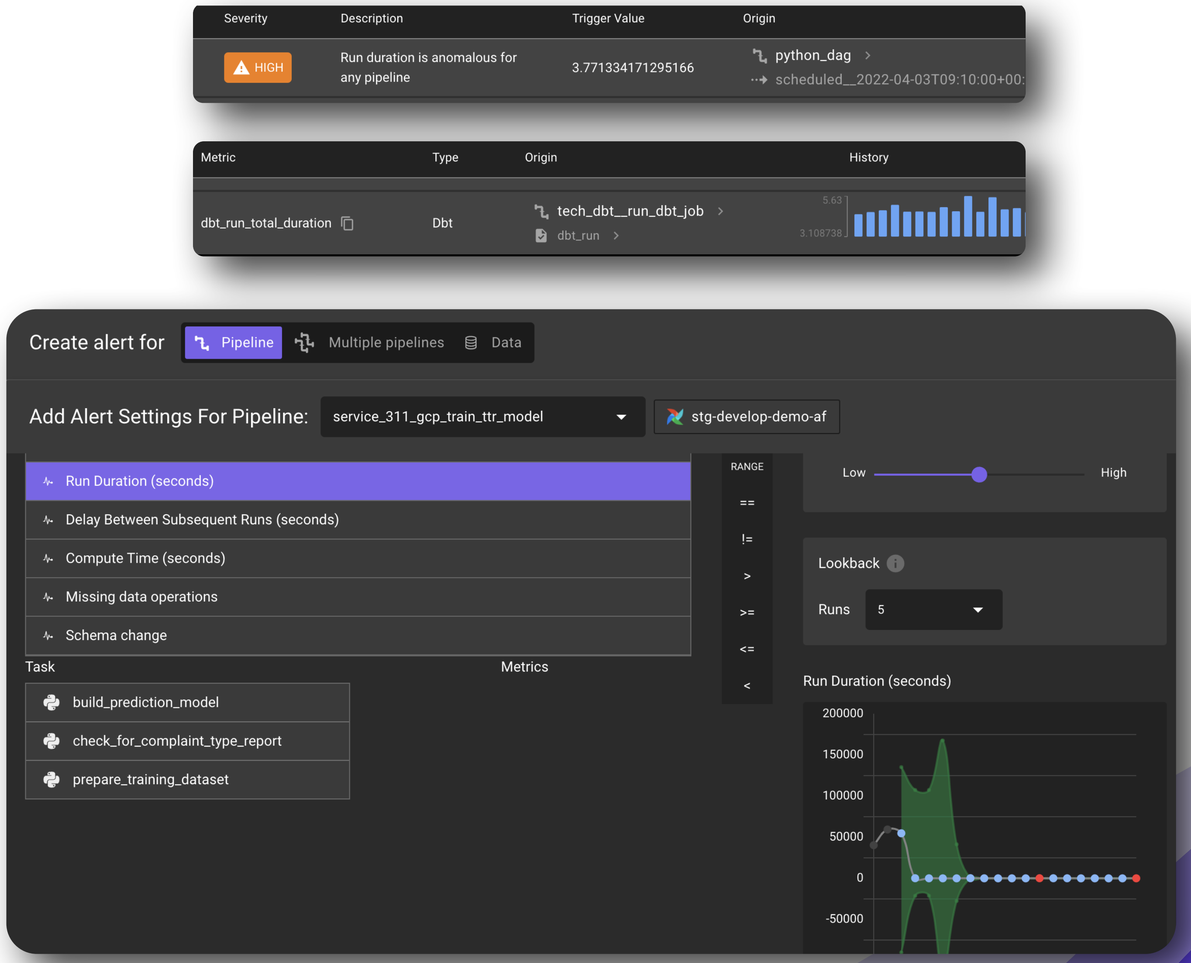
How to track it?

The engineering team can track this with the Airflow syncer, which reports on the total duration of a DAG run, or by using our tracking context as part of the Databand SDK.

Why it's important?

Pipelines that work in complex data processes are usually expected to have similar duration across different runs. In these complex environments, pipelines downstream depend on upstream pipelines processing the data in certain SLAs. The effect of extreme changes in the pipeline's duration can be anywhere between the processing of stale data to the failure of downstream processes.

What's it look like?

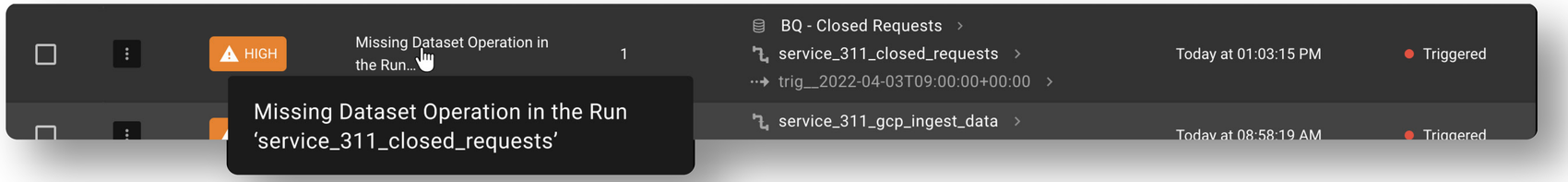


Metric 6 # Missing Data Operations

Who's it for?

- Data engineers
- Data scientists
- Data analysts
- Data executives

What's it look like?



How to track it?

Tracking all the operations related to a particular dataset.

A data operation is a combination of a task in a specific pipeline that reads or writes to a table.

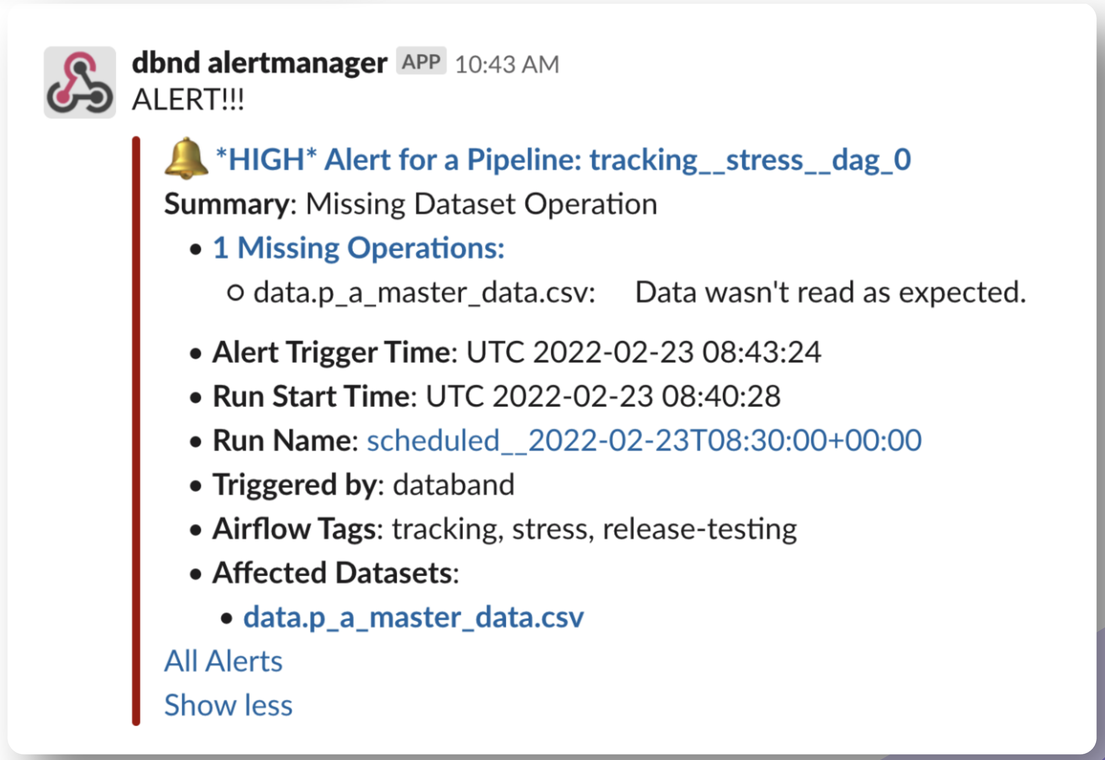
Why it's important?

When a certain data operation is missing, a chain of issues in your data stack will be triggered. It can cause pipelines to fail, changes in the schema, and delay problems.

Also, the downstream consumers of this data will be affected by the data that didn't arrive.

A few examples include:

- The data analyst who is using this data for analysis
- The ML models used by the data scientist
- The data engineers in charge of the data



Metric 7 # Record Count in a Run

Who's it for?

- Data engineers
- Data analysts

How to track it?

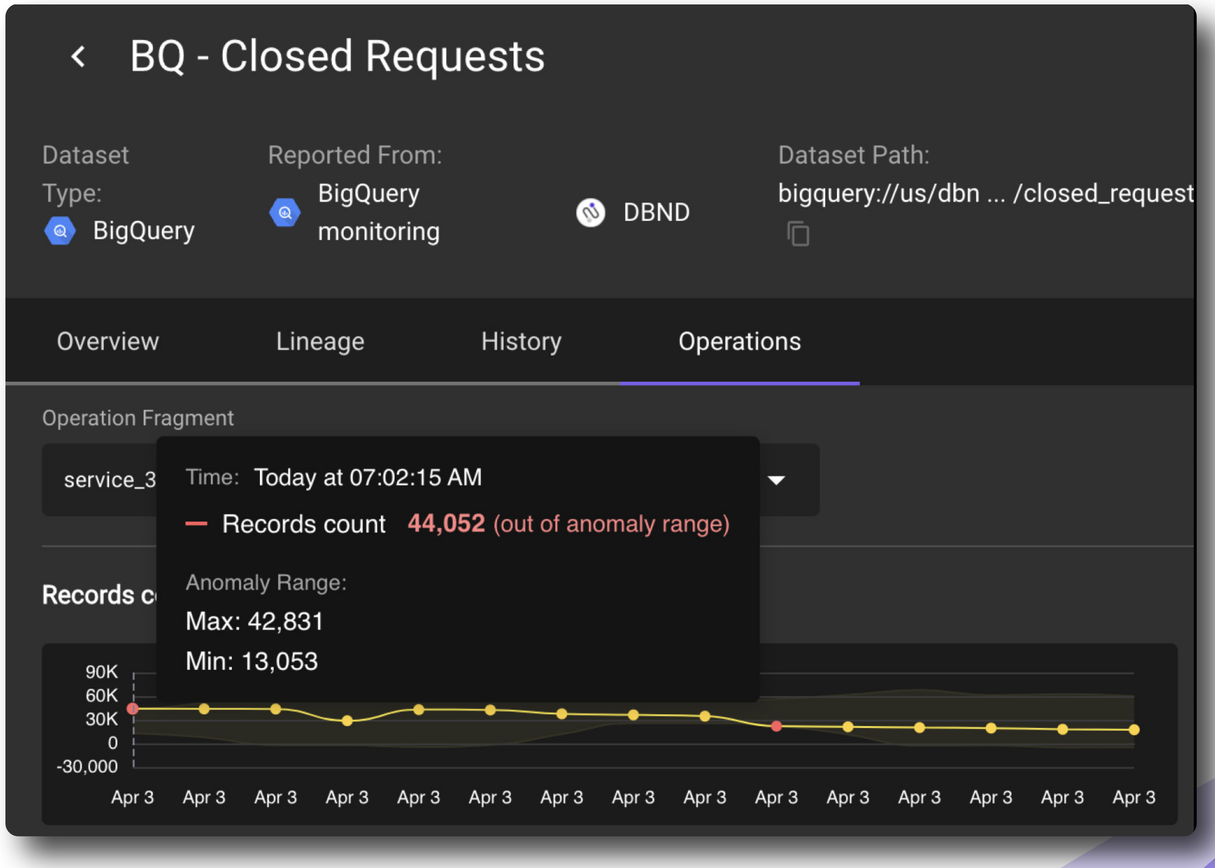
Track the number of raws written to a dataset.

Why it's important?

A sudden change in the expected number of table rows signals that too much data is being written.

Using anomaly detection in the number of rows in a dataset provides a good way of checking that nothing suspicious has happened.

What's it look like?



Metric 8 # Of Tasks Read From Dataset

Who's it for?

- Data engineer

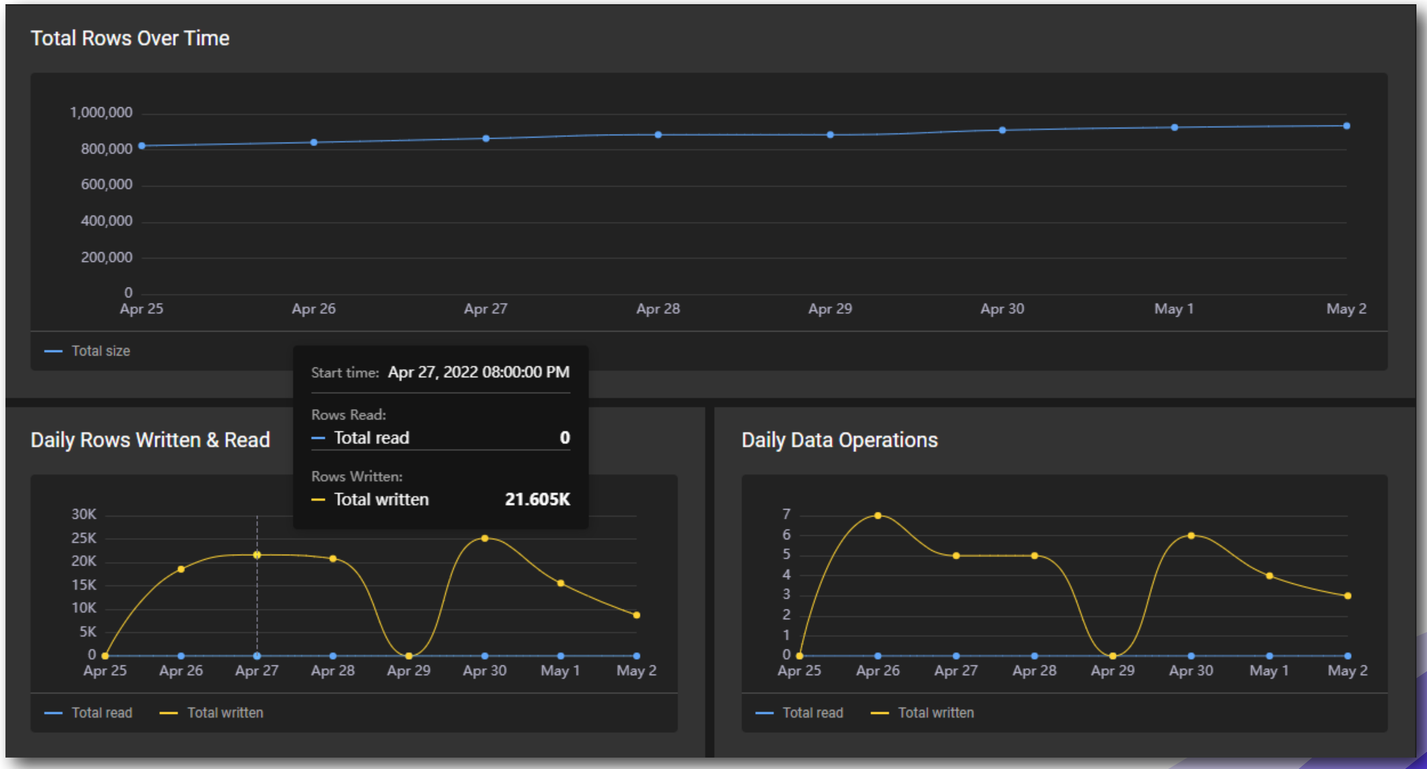
How to track it?

The more tasks read from a certain dataset, the more central it is and the more important this dataset.

Why it's important?

Understanding the importance of the dataset is crucial for impact analysis and realizing how fast you should deal with the issue you have.

What's it look like?



Metric 9 # Data Freshness (SLA alert)

Who's it for?

- Data Engineers
- Data Scientists
- Data Analysts

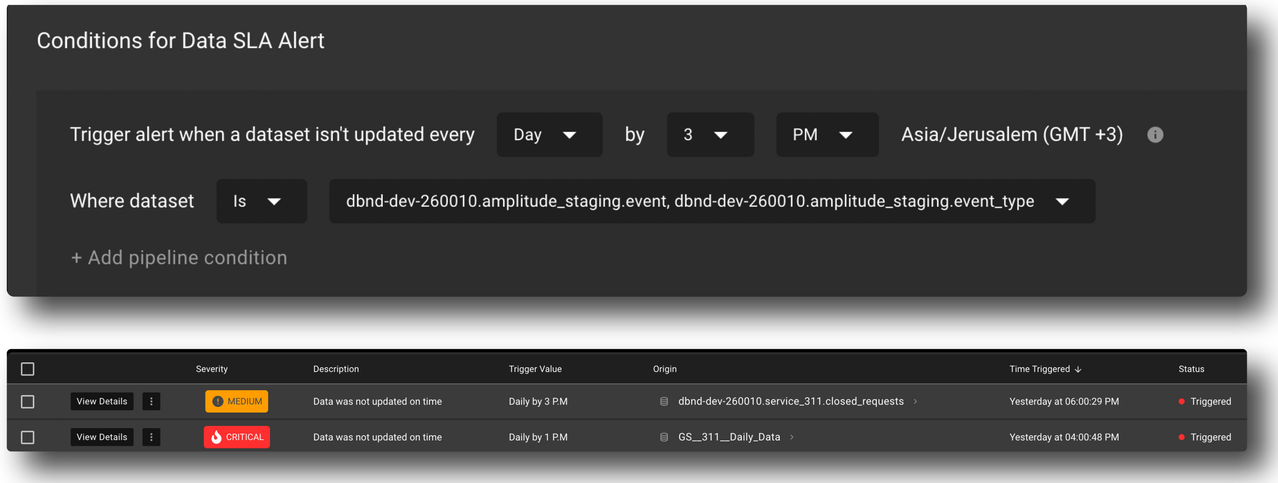
How to track it?

Track the number of scheduled pipelines that write to a certain dataset.

Why it's important?

When data isn't updated as expected, it can wrongly feed downstream reports. This results in consuming the wrong information. A good way of knowing data freshness is to monitor your SLA and get notified of delays in the pipeline that should be written to the dataset.

What's it look like?



Wrapping it up

It's time for better data quality metrics.

Detect and resolve your data issues faster than ever with Databand.

The only proactive data observability platform that resolves bad data issues before they turn into costly surprises for your business.

Request a demo today and see how you can eliminate data quality surprises and deliver more trustworthy data.

Request a Demo