The top data quality metrics you need to know

Contents

01

Data quality overview

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 10 other categories all relate to data quality.

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

We 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 provided screenshots that highlight each data quality metric you can view in the IBM® Databand platform.

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

Metric 1: Null counts

Whom it's for
– Data engineers
– Data analysts

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

Why it's important
Because 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 because the change is more gradual. Monitoring anomalies in the percentage of nulls across different columns should make it easier to see.

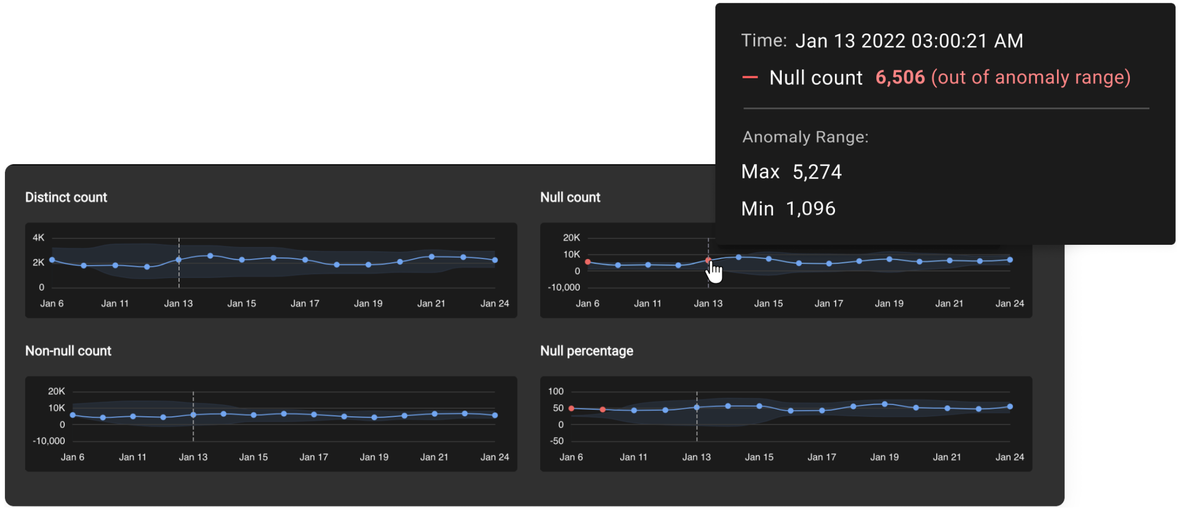

Figure 1. Null counts

Metric 2: Schema changes

Whom it's for
– Data engineers
– Data scientists
– Data analysts

How to track it
Track all changes in the schema for all the data sets related to a certain job.

Why it's important
Schema changes are key signals of poor-quality data. In a healthy situation, schema changes are communicated in advance and are infrequent because 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 include:
– 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 data sets.

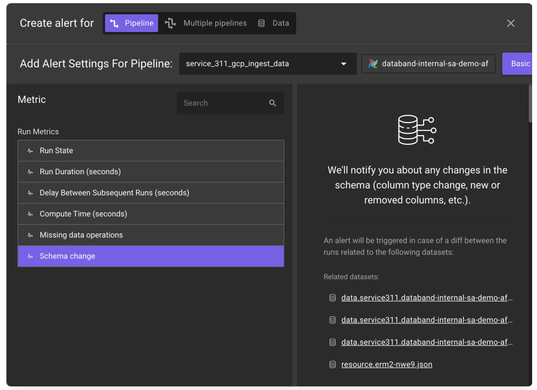

Figure 2. Schema change alert

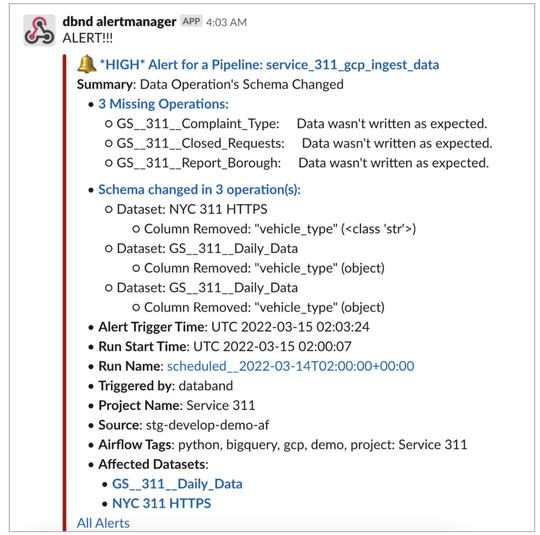

Figure 3. Schema change alert on Slack

Metric 3: Data lineage

Whom it's for
– Data engineers
– Data analysts

How to track it
Track the data linage with assets that appear downstream from a data set with an issue. This includes data sets and pipelines that consume the upstream data set's data.

Why it's important
The more damaged data assets are downstream in terms of data sets and pipelines, the bigger the impact of the data issue. This metric helps data engineers understand the severity of their issue and how fast they should fix it.

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

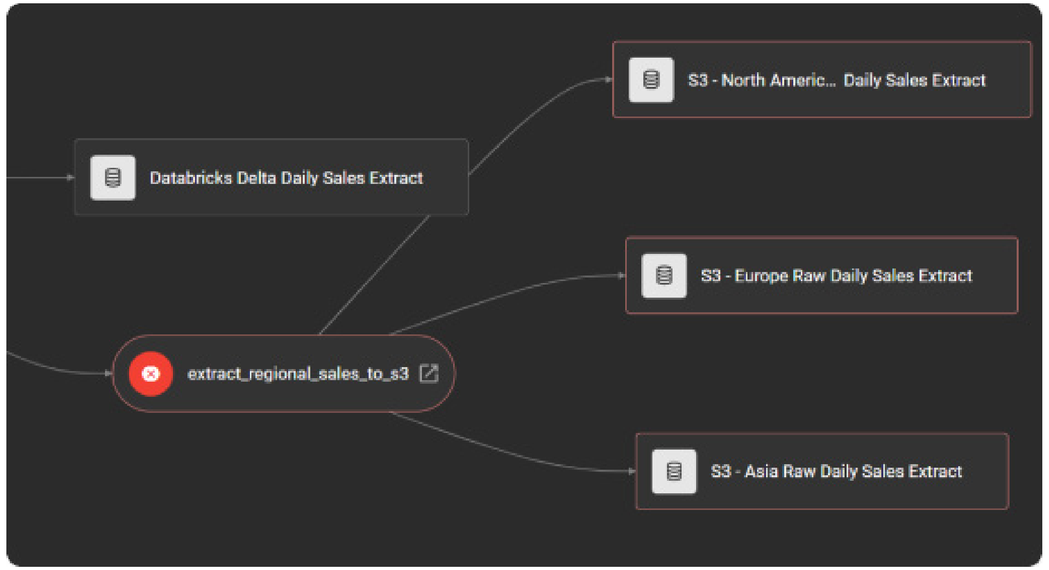

Figure 4. Data lineage and impact analysis

Metric 4: Pipeline failures

Whom it's for
– Data engineers
– Data executives

How to track it
Track the number of failed pipelines over time. To understand why pipelines fail, use tools that highlight root cause analysis and show deep dives inside all the tasks that the directed acyclic graph (DAG) contains.

Why it's important
The more pipelines fail, the more data health issues you'll have. Each pipeline failure causes issues such as missing data operations, schema changes and stale data. If you're experiencing many failures, this indicates severe problems at the root that must be addressed.

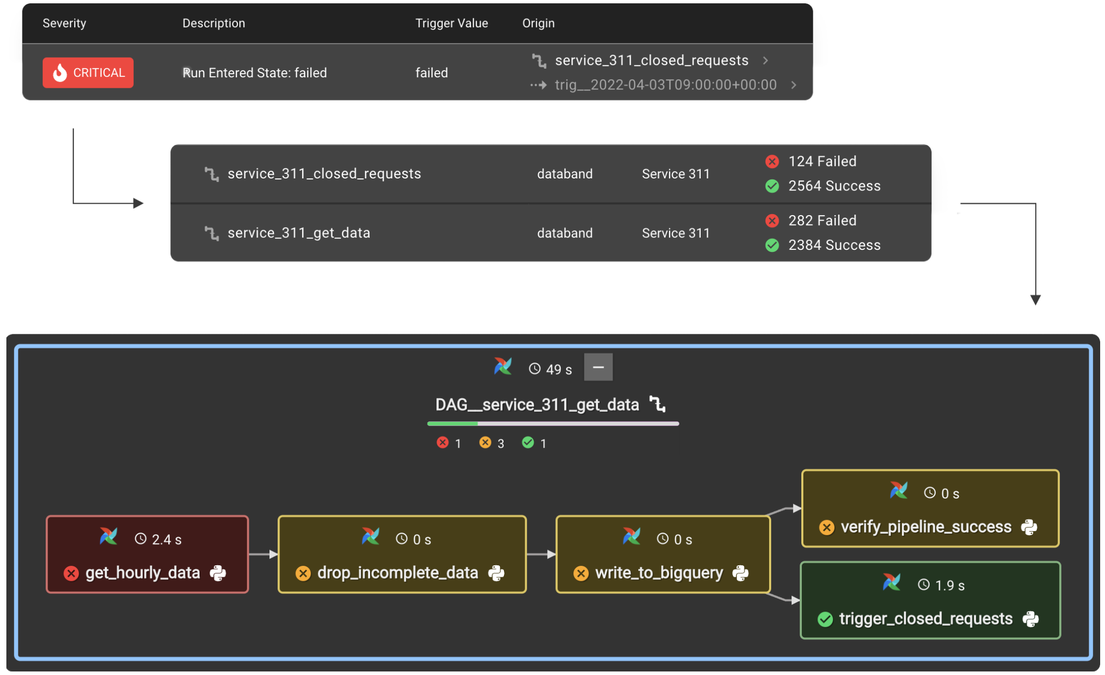

Figure 5. Error widget, pipeline, tasks

06

Metric 5: Pipeline duration

Whom it’s 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 IBM Databand software development kit (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 service level agreements (SLAs). The effect of extreme changes in the pipeline’s duration can be anything from the processing of stale data to the failure of downstream processes.

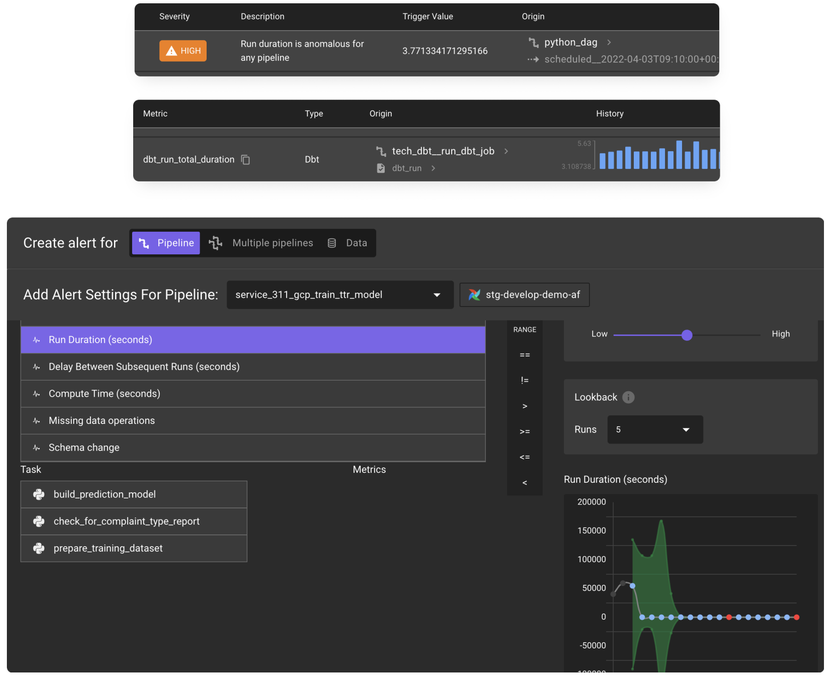

Figure 6. Pipeline duration

Metric 6: Missing data operations

Whom it's for
– Data engineers
– Data scientists
– Data analysts
– Data executives

How to track it
Track all the operations related to a particular data set. 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 failed pipelines, changes in the schema and delays.

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 machine learning models used by the data scientist
– The data engineers in charge of the data

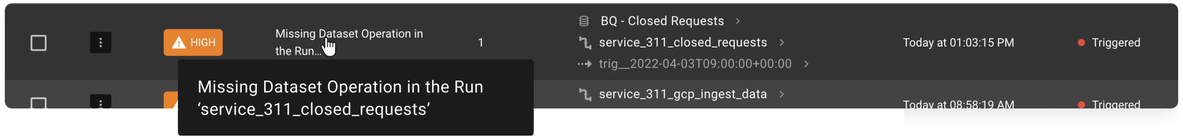

Figure 7. Pipeline operations alert

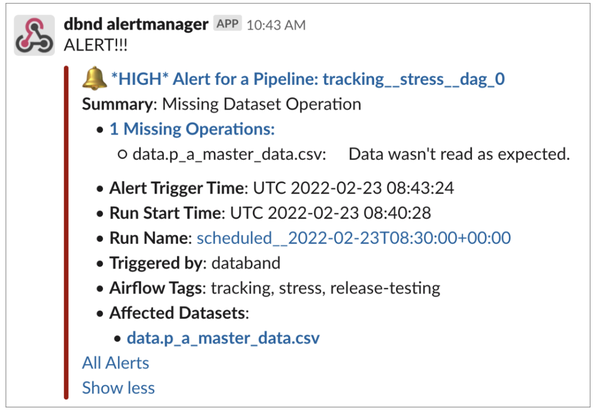

Figure 8. Pipeline operations alert on Slack

Metric 7: Record count in a run

Whom it's for
– Data engineers
– Data analysts

How to track it
Track the number of rows written to a data set.

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 data set provides a good way of checking that nothing suspicious has happened.

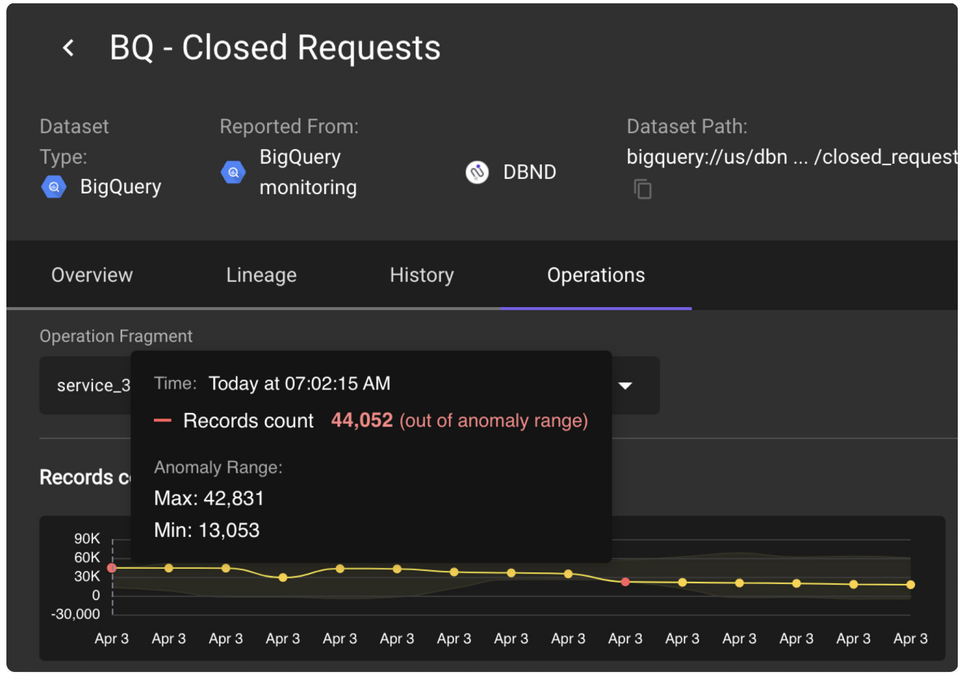

Figure 9. Record count in a run

Metric 8: Tasks read from data set

Whom it's for
– Data engineers

How to track it
The more tasks that are read from a certain data set, the more central and important that data set is.

Why it's important
Understanding the importance of the data set is crucial for impact analysis and realizing how fast you should deal with the issue you have.

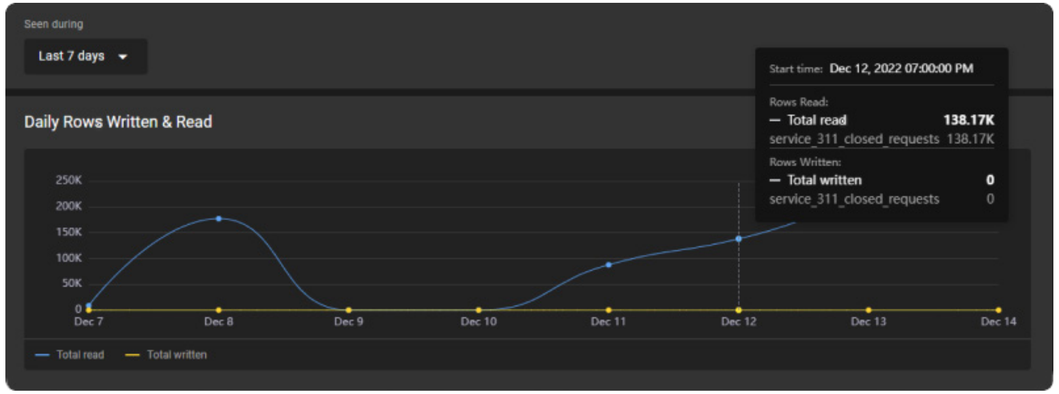

Figure 10. Tasks read from data set

Metric 9: Data freshness

Whom it’s for
– Data engineers
– Data scientists
– Data analysts

How to track it
Track the number of scheduled pipelines that write to a certain data set.

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 data set.

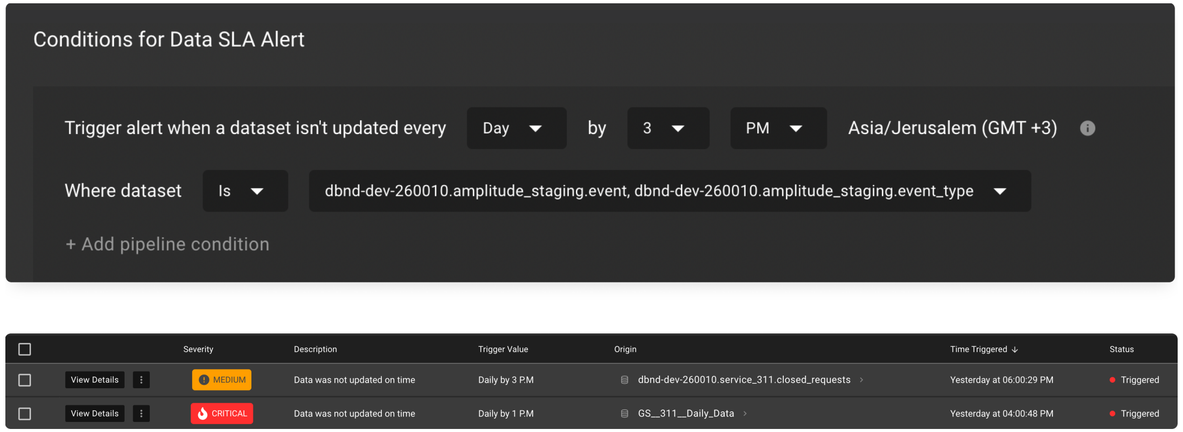

Figure 11. Data quality metrics SLA alert

11

Conclusion

It’s time for better data quality metrics.

Detect and resolve your data issues faster than ever with IBM Databand, the proactive data observability platform that resolves bad data issues before they turn into costly surprises for your business.

Why IBM?
IBM Databand delivers trusted data to your business. Learn more about how IBM Databand can help your organization automatically observe dynamic data pipelines, promote data quality and reliability, and continuously monitor AI and machine learning reliability.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
January 2023