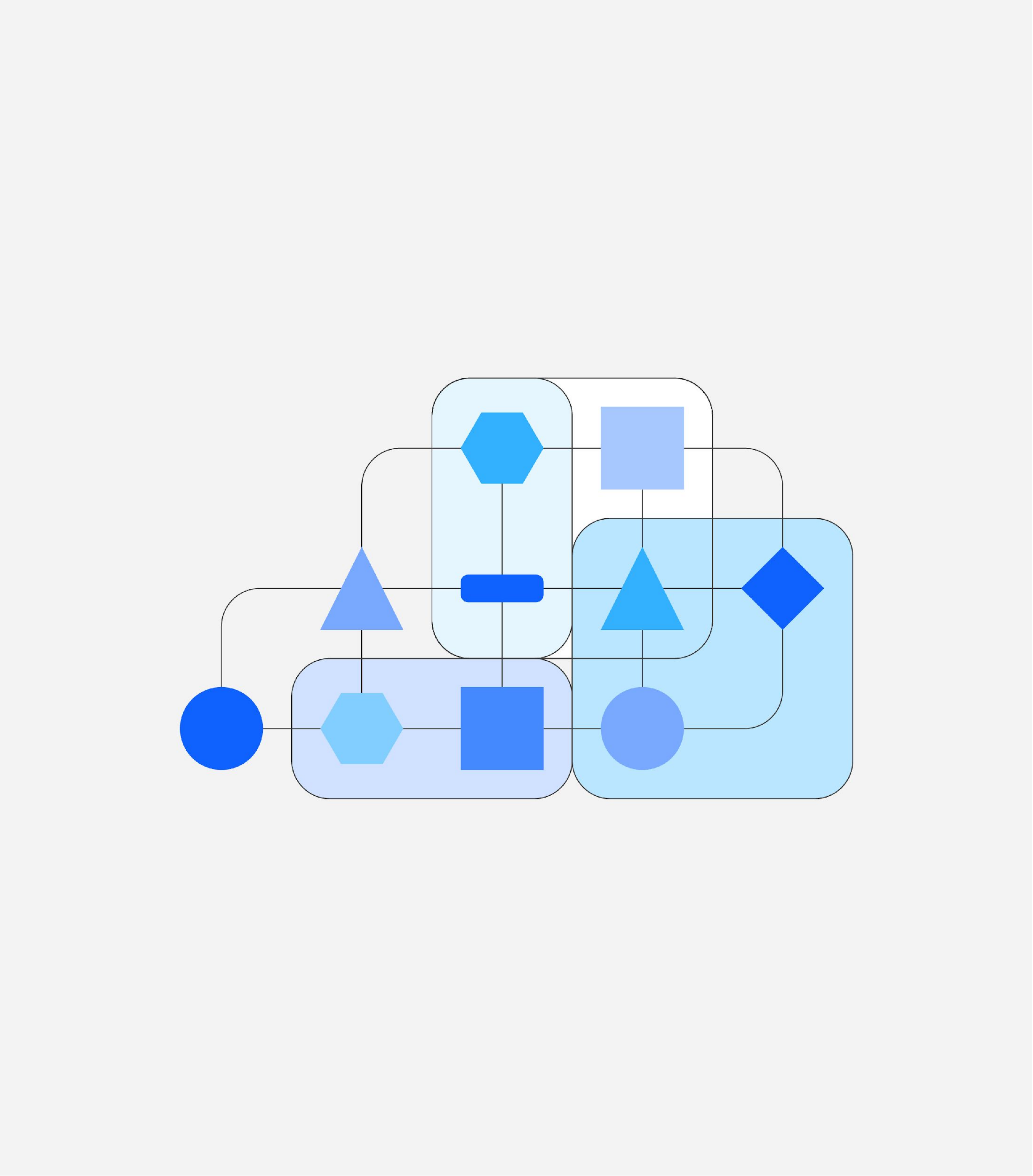
Ensuring Data Quality in Streaming Data Pipelines at Scale



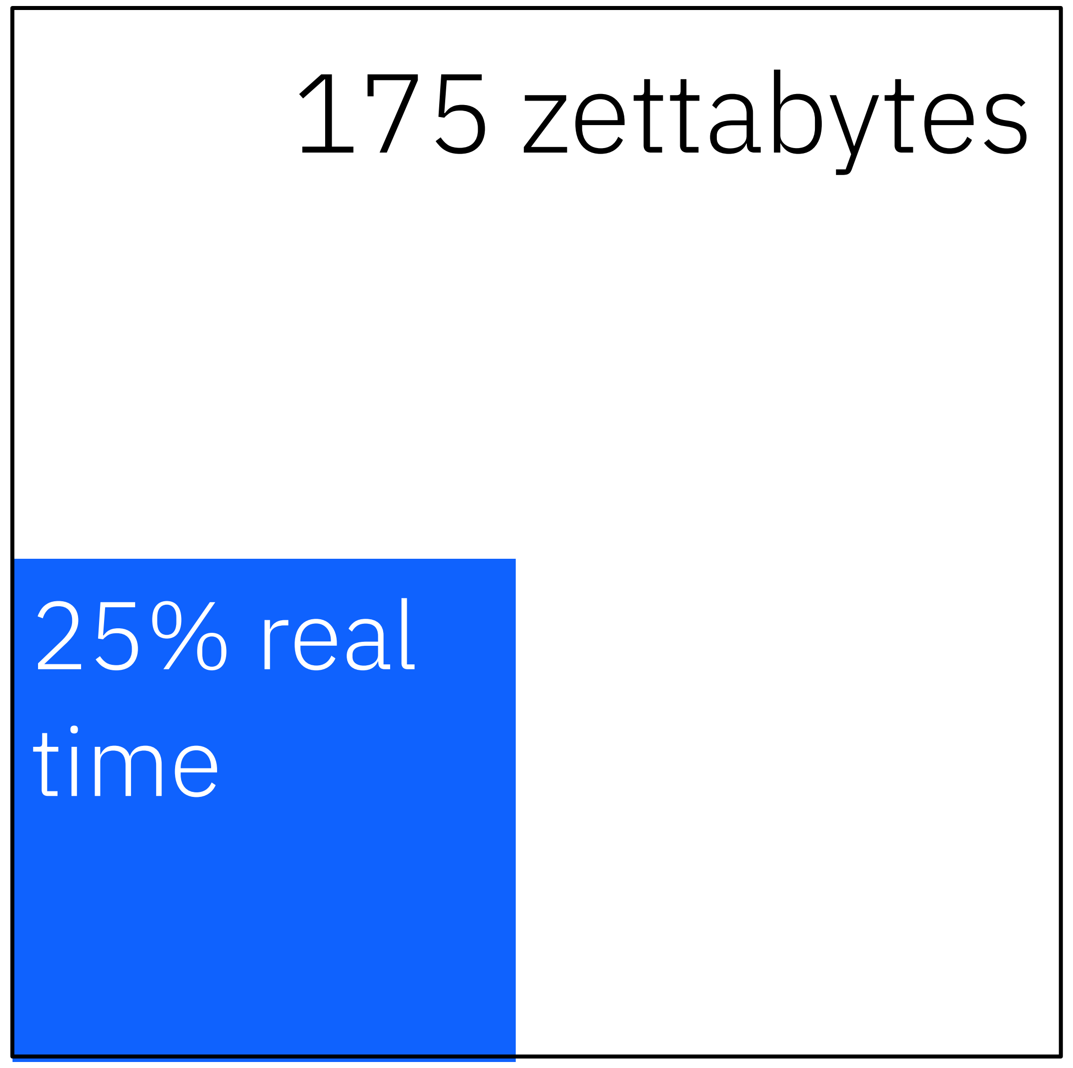
Ryan Yackel
Ryan.Yackel@ibm.com
GTM Product Manager
IBM Databand

Marc Sabate
Marc.Sabate@ibm.com
Technical Product Manager
IBM StreamSets

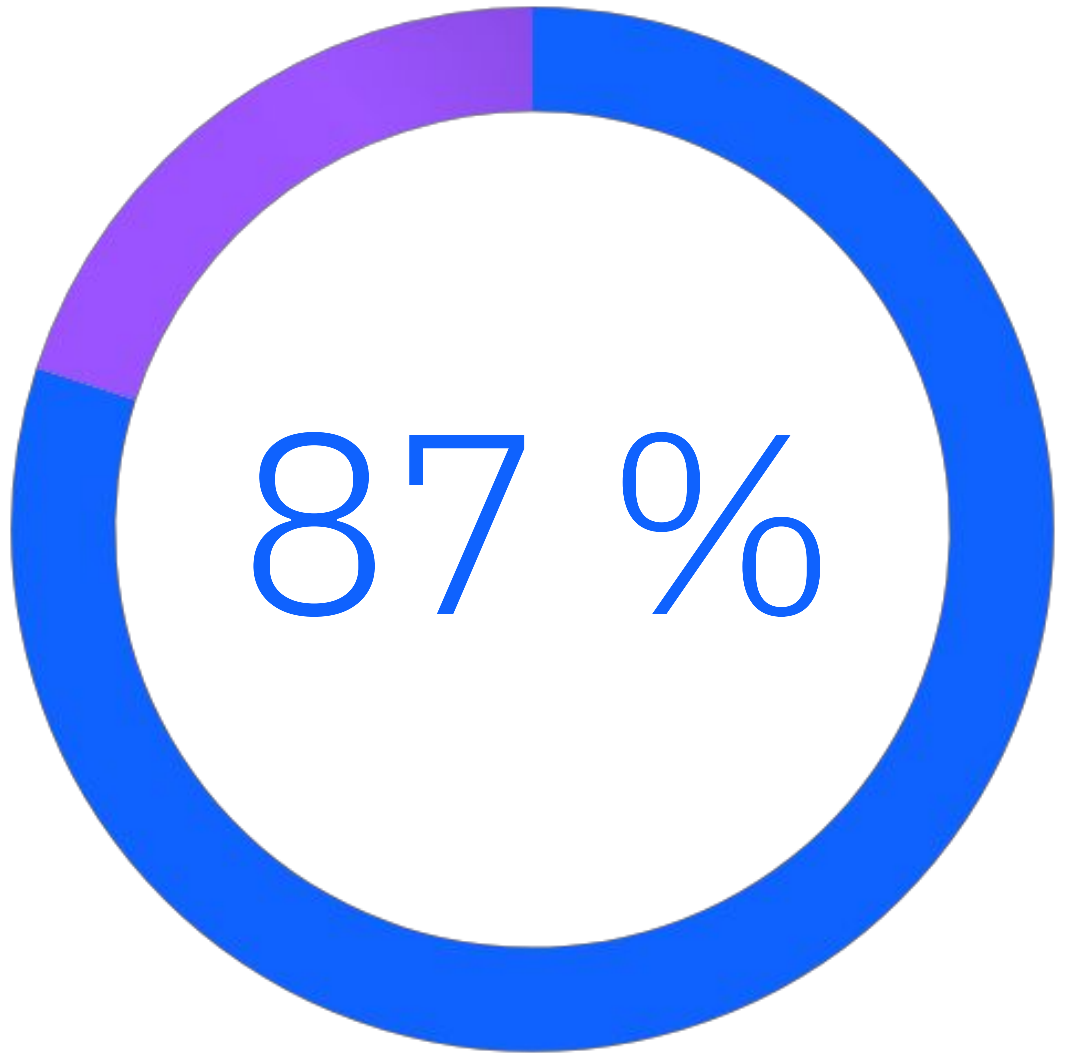
According to the latest estimates, 402.74 million terabytes of data are created each day.

Source: Exploding Topics https://explodingtopics.com/blog/data-generated-per-day#how-much

Global data volume is projected to grow to 175 zettabytes this year, with nearly 25% of this data being real-time



87% of organizations require data to be ingested and analyzed within one day or faster



Source: 1. https://www.seagate.com/files/www-content/our-story/trends/files/dataage-idc-report-final.pdf
Source: 2. Worldwide Enterprise Global DataSphere by Vertical Industry Forecast, 2023–2027

While data integration remains a priority for organizations, challenges remain

67% of data leaders believe prioritizing integrating data from various sources is a key priority[1]

Data access

50% of data analyst time is spent working reactively or trying to find or get access to data[2]

Data drift

80% of data engineering time is spent on maintaining & fixing existing pipelines due to data drift, creating a costly maintenance burden[3]

Data proliferation

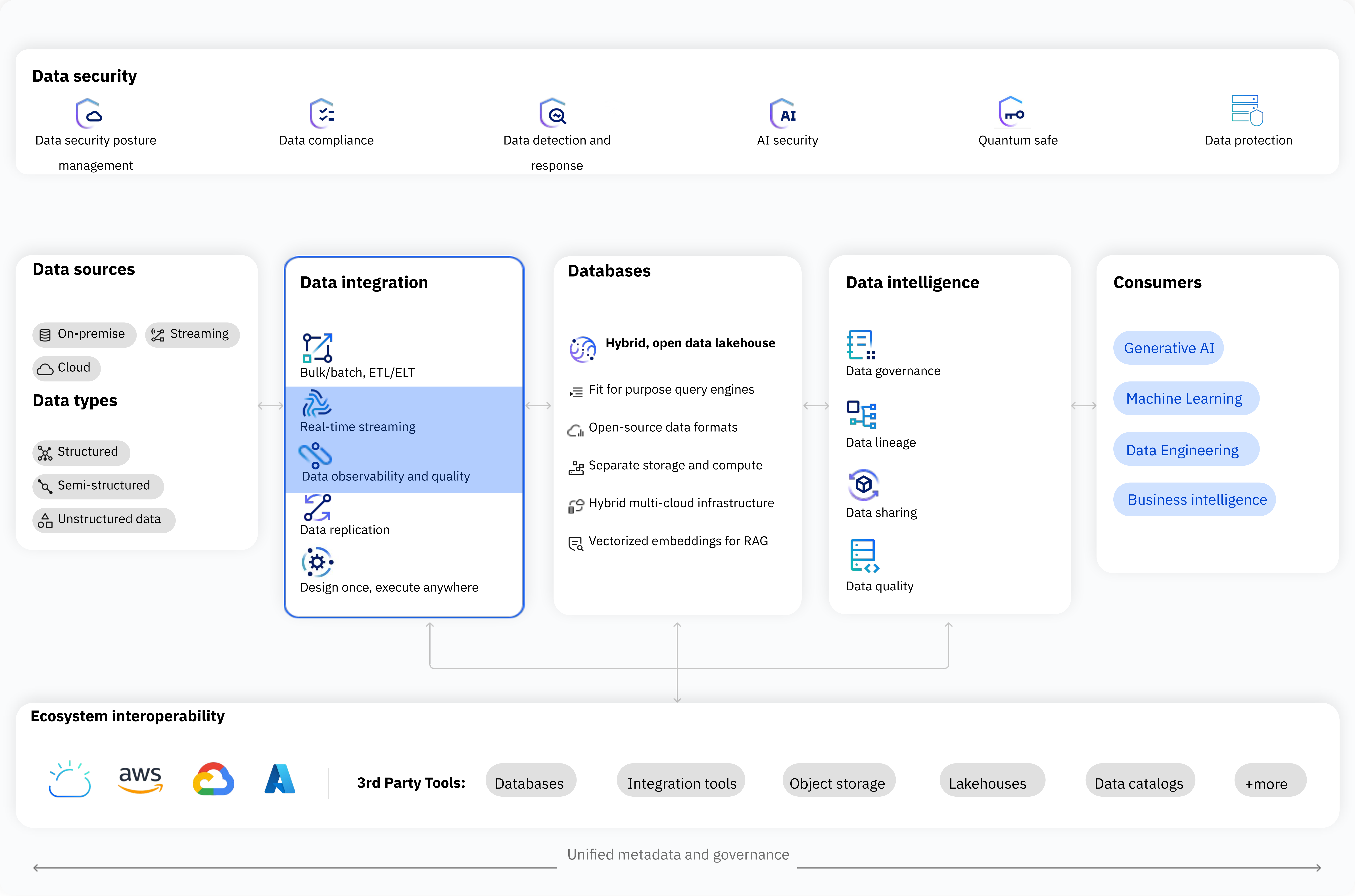
With growing data volume and datasets organizations require scalability and faster response times

Data latency

Need for processing data as it arrives to reduce delays in business decision and operations

A reimagined data strategy is imperative to meet the demands of contemporary use cases and democratize data access across the enterprise at scale

Integrate, access, govern, and secure all data types with an open and hybrid data architecture



Data security

Data security posture management

Data compliance

Data detection and response

AI security

Quantum safe

Data protection

Data sources

On-premise

Streaming

Cloud

Data types

Structured

Semi-structured

Unstructured data

Data integration

Bulk/batch, ETL/ELT

Real-time streaming

Data observability and quality

Data replication

Design once, execute anywhere

Databases

Hybrid, open data lakehouse

Fit for purpose query engines

Open-source data formats

Separate storage and compute

Hybrid multi-cloud infrastructure

Vectorized embeddings for RAG

Data intelligence

Data governance

Data lineage

Data sharing

Data quality

Consumers

Generative AI

Machine Learning

Data Engineering

Business intelligence

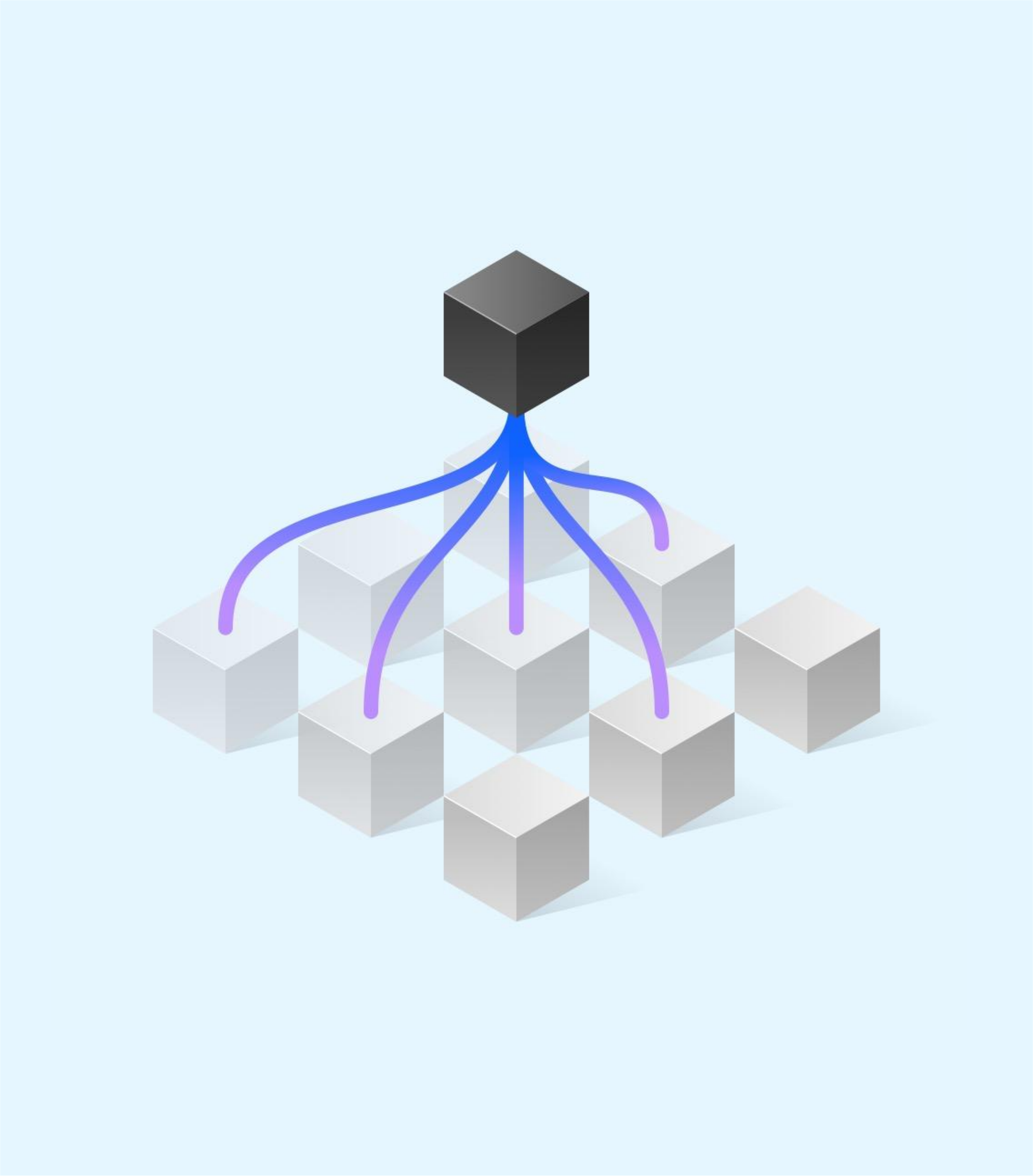
Ecosystem interoperability

3rd Party Tools: Databases | Integration tools | Object storage | Lakehouses | Data catalogs | +more

Unified metadata and governance

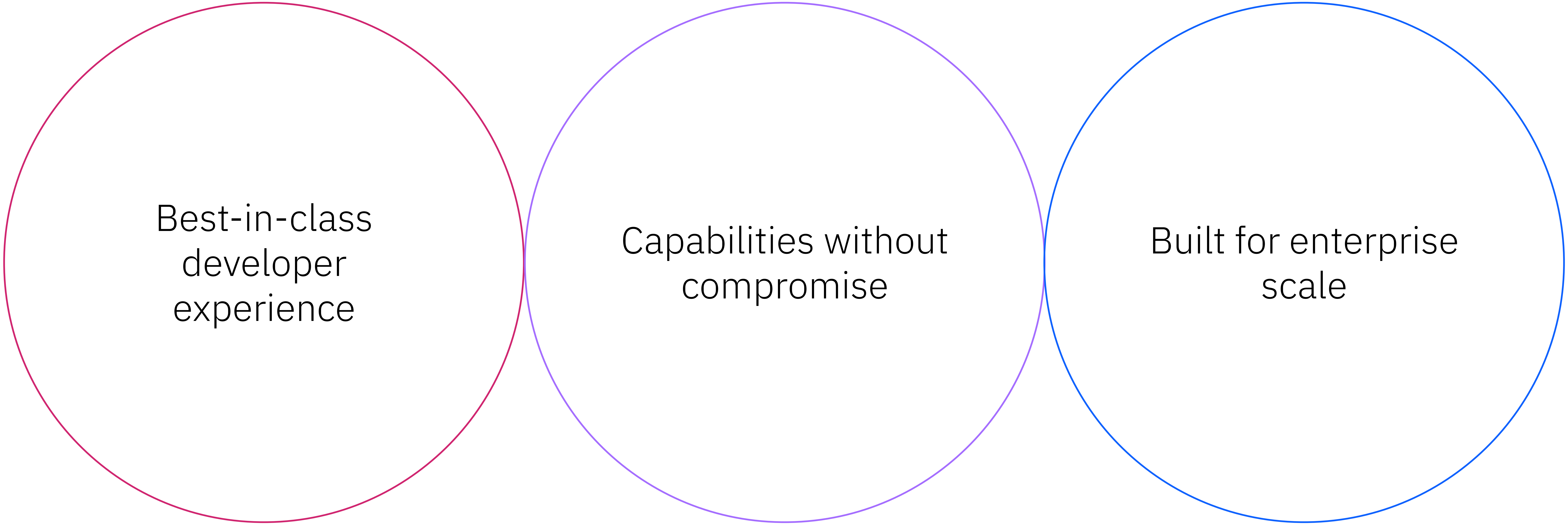
What is Real-Time Data Integration?

IBM defines real-time data integration as the ability to ingest, process, and write data as soon as it's available instead of on an intermittent or scheduled basis.



IBM StreamSets

A no-code/low-code streaming data integration tool for engineers who *hate* no-code/low code tools



Best-in-class developer experience

Capabilities without compromise

Built for enterprise scale

IBM StreamSets

Benefits

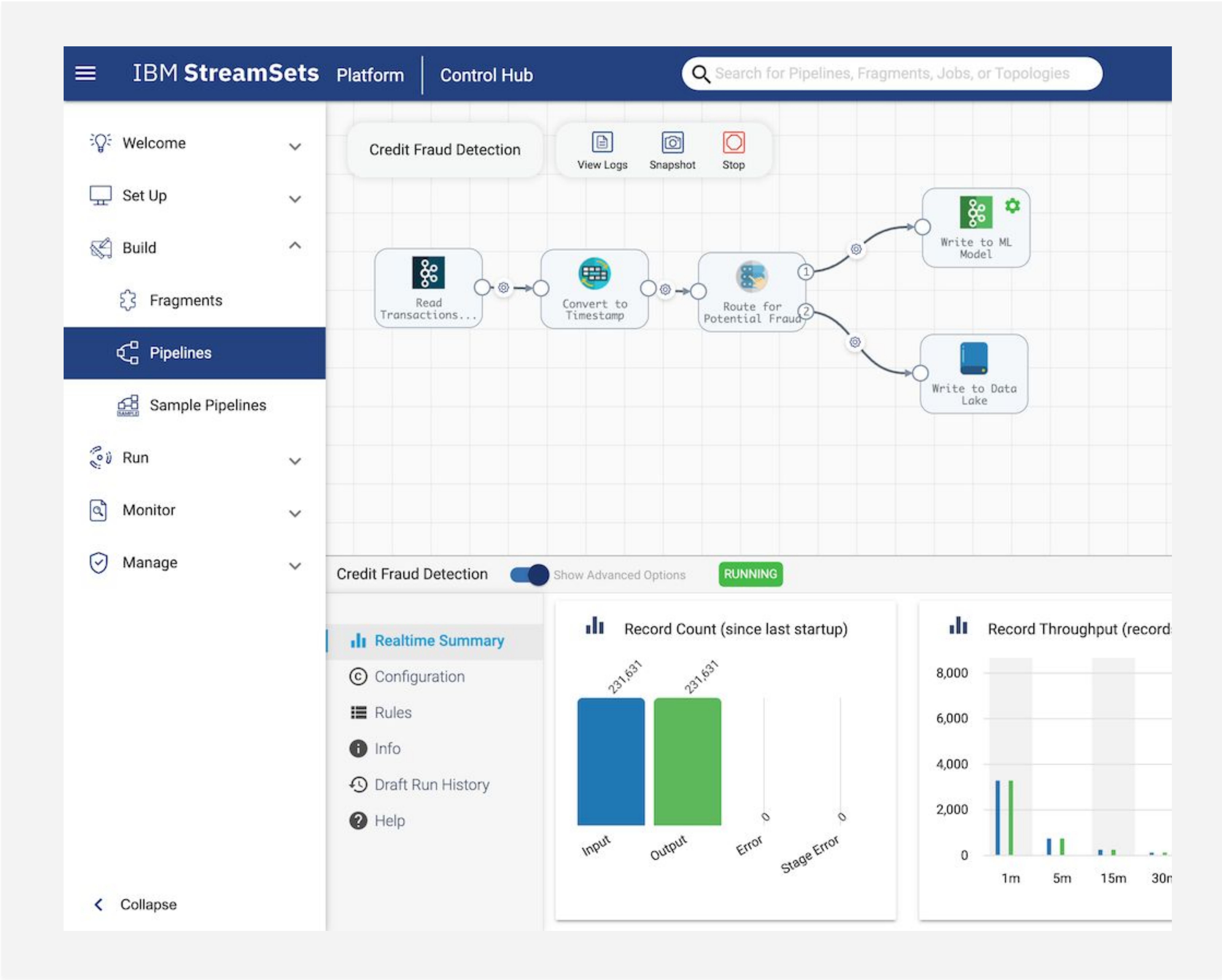
↓

Real-time data integration solution for building streaming data pipelines to enhance real-time decision-making and mitigate risks

Enable real-time data ingestion at scale

Reduce data drift with intelligent streaming data pipelines

Stream any data from any source



What about data quality?

IBM Databand

Benefits

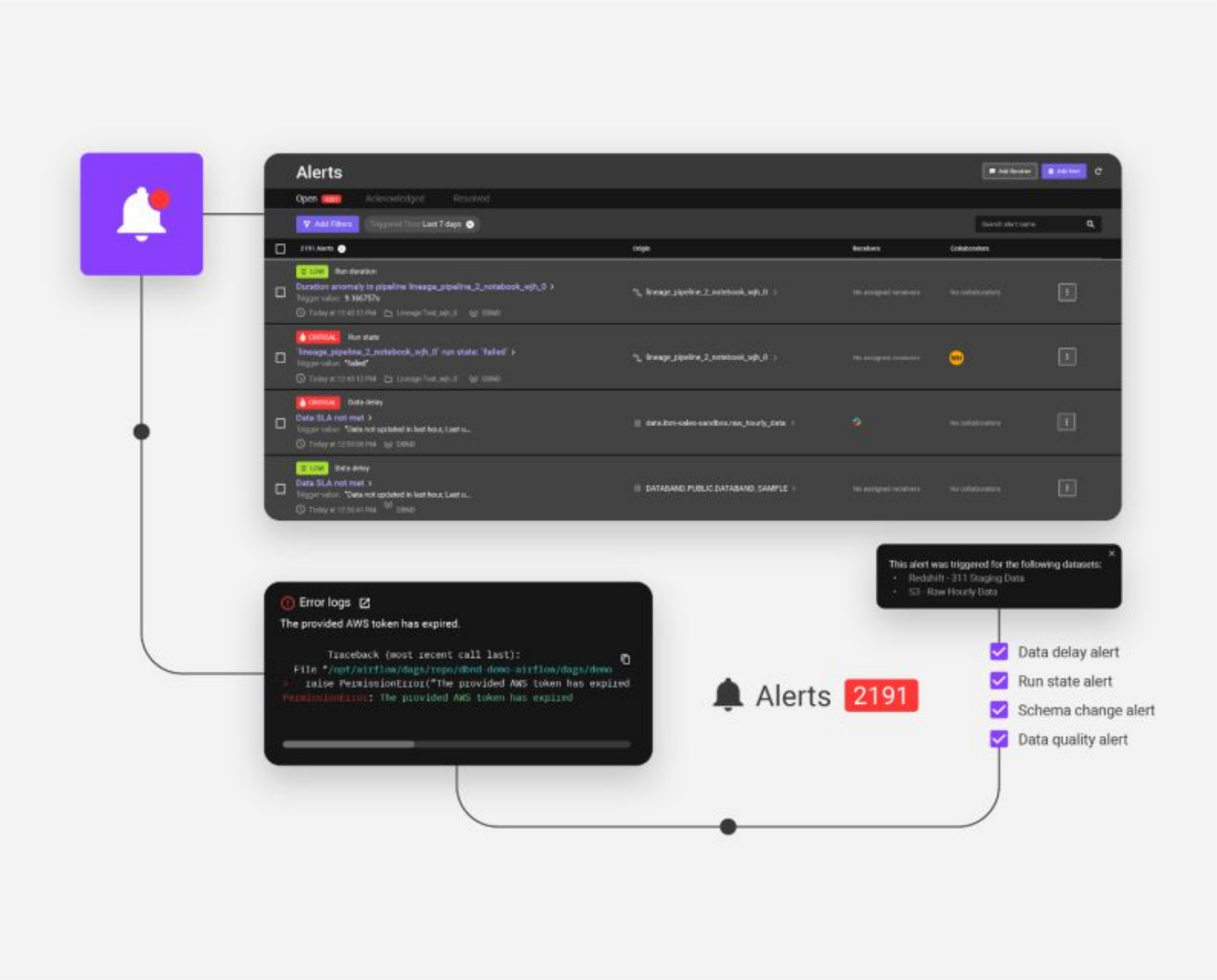
↓

Databand empowers data platform teams to deliver reliable and trustworthy data with continuous data observability.

Detect issues earlier

Resolve issues faster

Improve data reliability



Alerts

Open | Acknowledged | Resolved

Add Filters | Triggered Time: Last 7 days | Search alert name

2191 Alerts	Origin	Receivers	Collaborators
LOW Run duration Duration anomaly in pipeline lineage_pipeline_2_notebook_wjh_0 › Trigger value: 9.166757s Today at 12:40:13 PM · Lineage Test_wjh_0 · DBND	lineage_pipeline_2_notebook_wjh_0 ›	No assigned receivers	No collaborators
CRITICAL Run state 'lineage_pipeline_2_notebook_wjh_0' run state: 'failed' › Trigger value: "failed" Today at 12:40:12 PM · Lineage Test_wjh_0 · DBND	lineage_pipeline_2_notebook_wjh_0 ›	No assigned receivers	WH
CRITICAL Data delay Data SLA not met › Trigger value: "Data not updated in last hour, Last u... Today at 12:39:08 PM · DBND	data.ibm-sales-sandbox.raw_hourly_data ›		No collaborators
LOW Data delay Data SLA not met › Trigger value: "Data not updated in last hour, Last u... Today at 12:38:41 PM · DBND	DATABAND.PUBLIC.DATABAND_SAMPLE ›	No assigned receivers	No collaborators

Error logs

The provided AWS token has expired.

```
        Traceback (most recent call last):
  File "/opt/airflow/dags/repo/dbnd-demo-airflow/dags/demo
>   raise PermissionError("The provided AWS token has expired
PermissionError: The provided AWS token has expired
```

This alert was triggered for the following datasets:
- Redshift - 311 Staging Data
- S3 - Raw Hourly Data

Alerts 2191

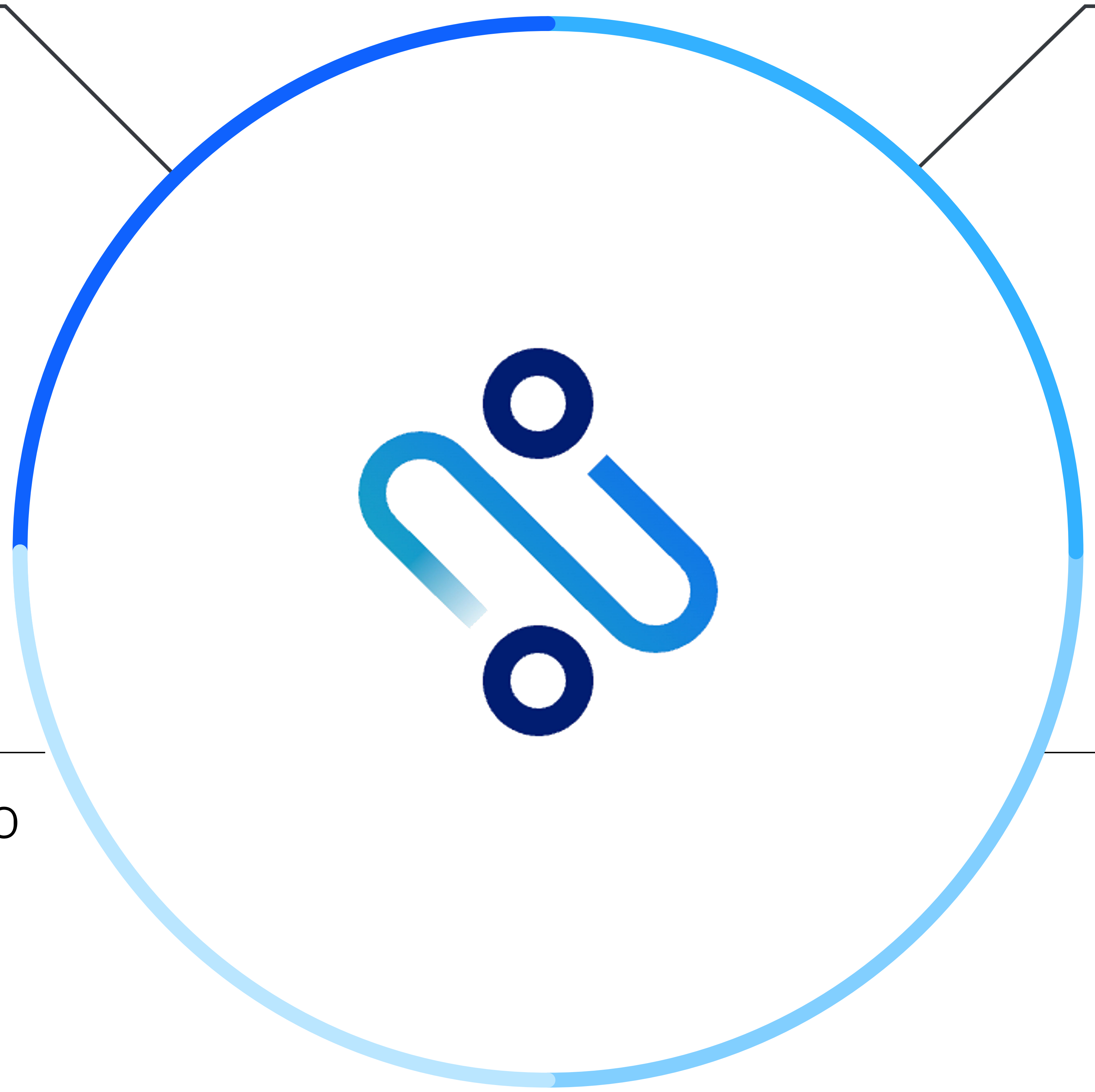
☑ Data delay alert
☑ Run state alert
☑ Schema change alert
☑ Data quality alert

How Databand works

1. Collect

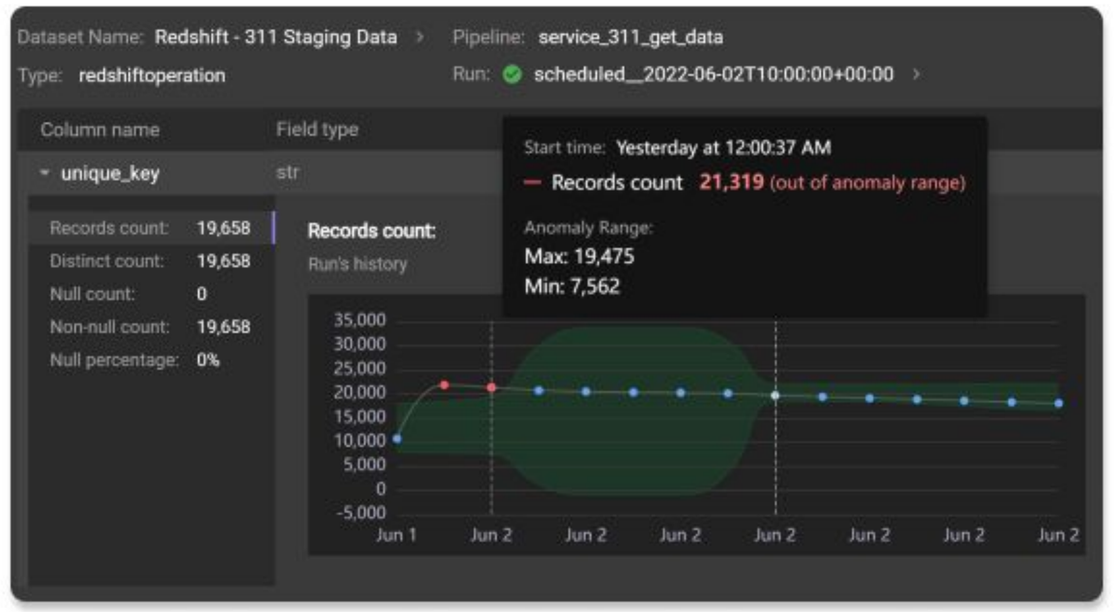
Automatically collect metadata.
From all key solutions in the modern data stack.

Apache Airflow · databricks · dbt · Azure Data Factory · bmc Control-M · Apache Spark



2. Profile

Build historical baseline.
Based on common data pipeline behavior.



Dataset Name: Redshift - 311 Staging Data › Pipeline: service_311_get_data
Type: redshiftoperation Run: scheduled__2022-06-02T10:00:00+00:00 ›

Column name	Field type
unique_key	str

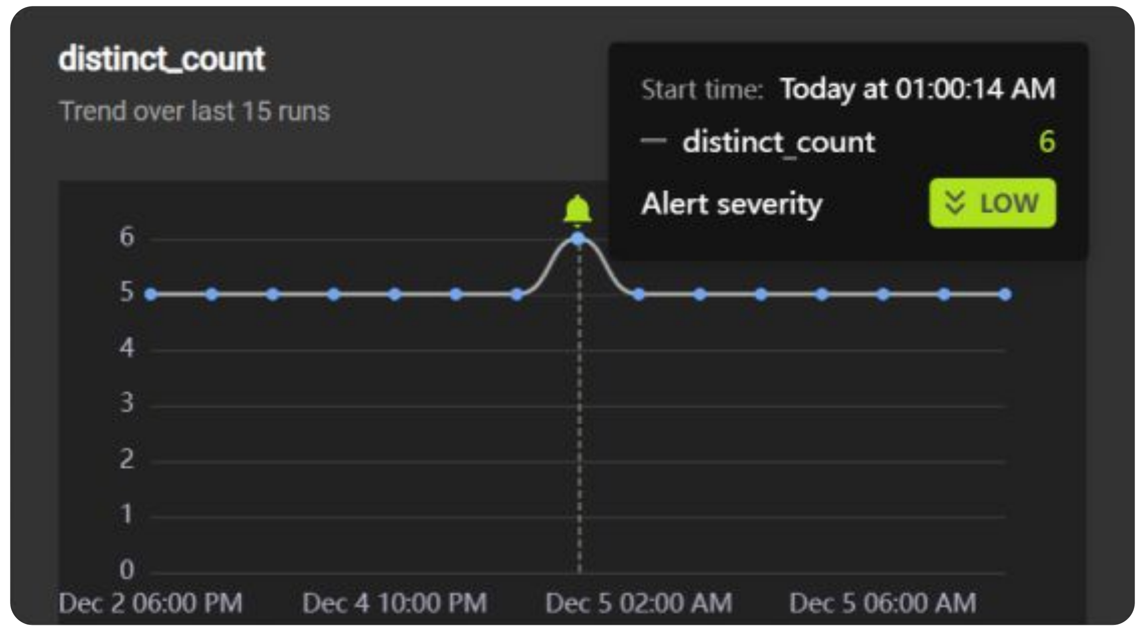
Records count: 19,658
Distinct count: 19,658
Null count: 0
Non-null count: 19,658
Null percentage: 0%

Records count:
Run's history

Start time: Yesterday at 12:00:37 AM
Records count 21,319 (out of anomaly range)
Anomaly Range:
Max: 19,475
Min: 7,562

3. Alert

Alert on anomalies and rules.
Based on deviations or breaches.

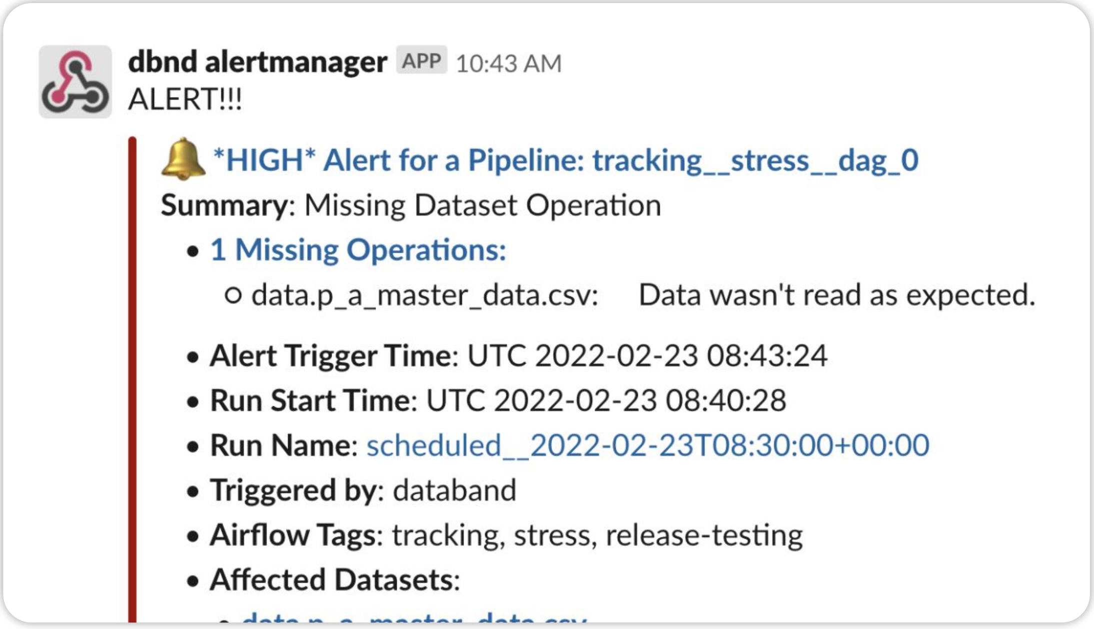


distinct_count
Trend over last 15 runs

Start time: Today at 01:00:14 AM
distinct_count 6
Alert severity LOW

4. Resolve

Resolve the root cause.
Create smart communication workflows to resolve data quality issues & meet SLAs.



dbnd alertmanager APP 10:43 AM
ALERT!!!

🔔 *HIGH* Alert for a Pipeline: tracking__stress__dag_0
Summary: Missing Dataset Operation

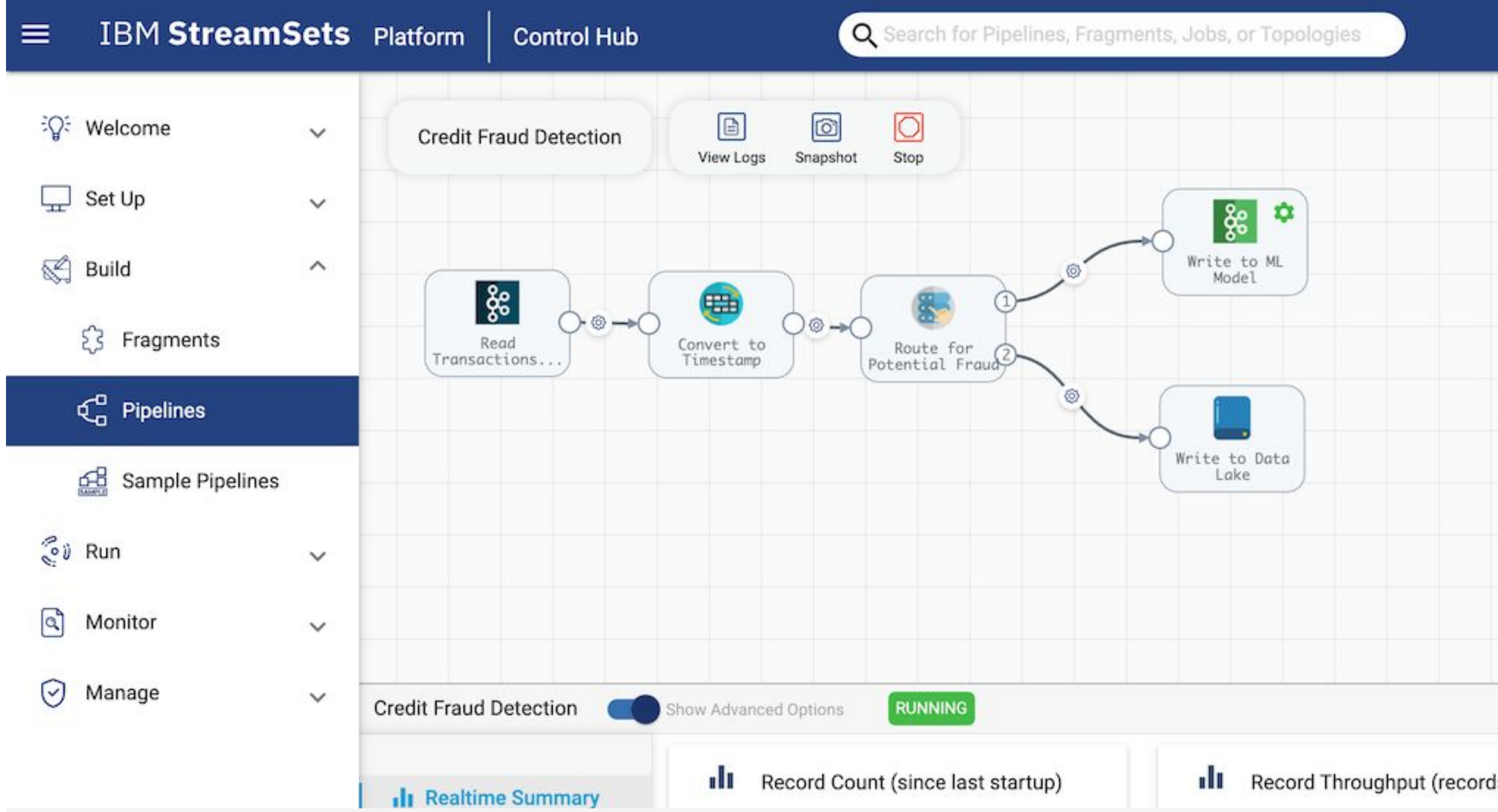
- 1 Missing Operations:
  - data.p_a_master_data.csv: Data wasn't read as expected.
- **Alert Trigger Time**: UTC 2022-02-23 08:43:24
- **Run Start Time**: UTC 2022-02-23 08:40:28
- **Run Name**: scheduled__2022-02-23T08:30:00+00:00
- **Triggered by**: databand
- **Airflow Tags**: tracking, stress, release-testing
- **Affected Datasets**:
  - data.p_a_master_data.csv

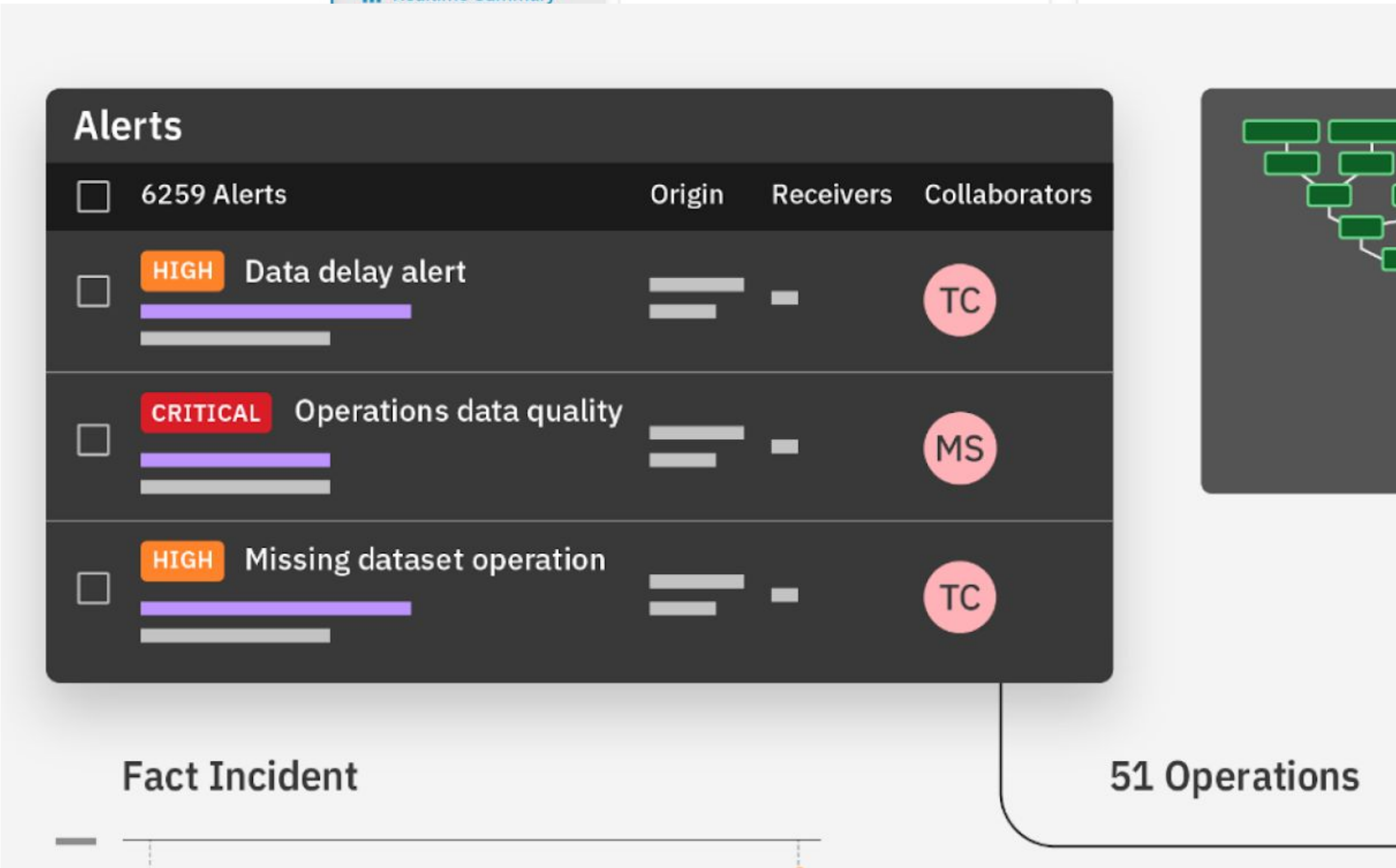
DEMO

Overview

1. Streaming pipeline ingestion with masked data with AWS S3 destination
2. Streaming pipeline ingestion with product data with Snowflake destination
3. Data quality alerts based on Json format errors
4. Data quality alerts based on infrastructure errors

Marc Sabate
Marc.Sabate@ibm.com
Technical Product Manager
IBM StreamSets





Alerts

□	6259 Alerts	Origin	Receivers	Collaborators
□	HIGH Data delay alert			TC
□	CRITICAL Operations data quality			MS
□	HIGH Missing dataset operation			TC

Q&A Time

Ryan Yackel
Ryan.Yackel@ibm.com
GTM Product Manager
IBM Databand

Marc Sabate
Marc.Sabate@ibm.com
Technical Product Manager
IBM StreamSets

Three ways to get started with IBM data integration today

Want to read more about IBM data integration?

Visit the IBM data integration website →

See how it works for you.

Start a data integration trial →

Kickstart your project with the IBM technical experts

Book here →

IBM

Autodesk and IBM

Moving from reactive to proactive data quality

> “We got tired of being caught off guard by repeated types of data incidents with no owner to tackle these incidents. With Databand, we’ve been able to reduce our mean time to detection down to almost zero.
>
> At Autodesk, we encourage innovation, so we saw this as an internal opportunity to bring Databand’s data observability to the business.”

Senior Manager for Data Engineering
and Visualization
Autodesk

Technologies

– IBM Databand
– Apache Airflow
– Apache Spark
– dbt Core
– Snowflake

Use cases

– Batch processing monitoring
Databand extensively monitors production batch processing > 1,000 DAGs for ML and AI teams.

– Inline testing
The team uses Databand’s inline testing capabilities to detect data quality issues in real time.

– Data products support
Databand supports pipelines that deliver insights and in-product messaging for Autodesk’s customers.

AUTODESK

Results

– Reduction in MTTD
Databand reduced the time to detect data quality issues from days to minutes. This immediate detection allowed the team to address problems before they could cause major disruptions.

– Reduction in MTTR
The mean time to resolving data issues dropped from weeks to days. Detecting incidents like late-arriving data, schema changes and pipeline failures helps maintain trust and efficiency.

– Cost savings
Autodesk saw a decrease in cloud consumption costs by detecting issues early and avoiding reruns.

Alerts

□ 6259 Alerts	Origin	Receivers	Collaborators
□ HIGH Data delay alert			TC
□ CRITICAL Operations data quality			MS
□ HIGH Missing dataset operation			TC

Run Status: ✓ Success
Start time: Yesterday at 09:2
– Total Read 12.1
– Total Write 14.

NatWest Group and IBM

Modernize and compete with challenger banks using innovation and smarter data

NatWest Group PLC (formerly The Royal Bank of Scotland Group PLC), is a majority state-owned British banking and insurance holding company based in Edinburgh, Scotland. The group operates a wide variety of banking brands offering personal and business banking, private banking, insurance and corporate finance, and offers its services to over 19 million retail customers across the UK and Ireland. It also provides business banking services for around 1 in 4 businesses across the UK and Ireland, from startups to multinationals.

Technologies

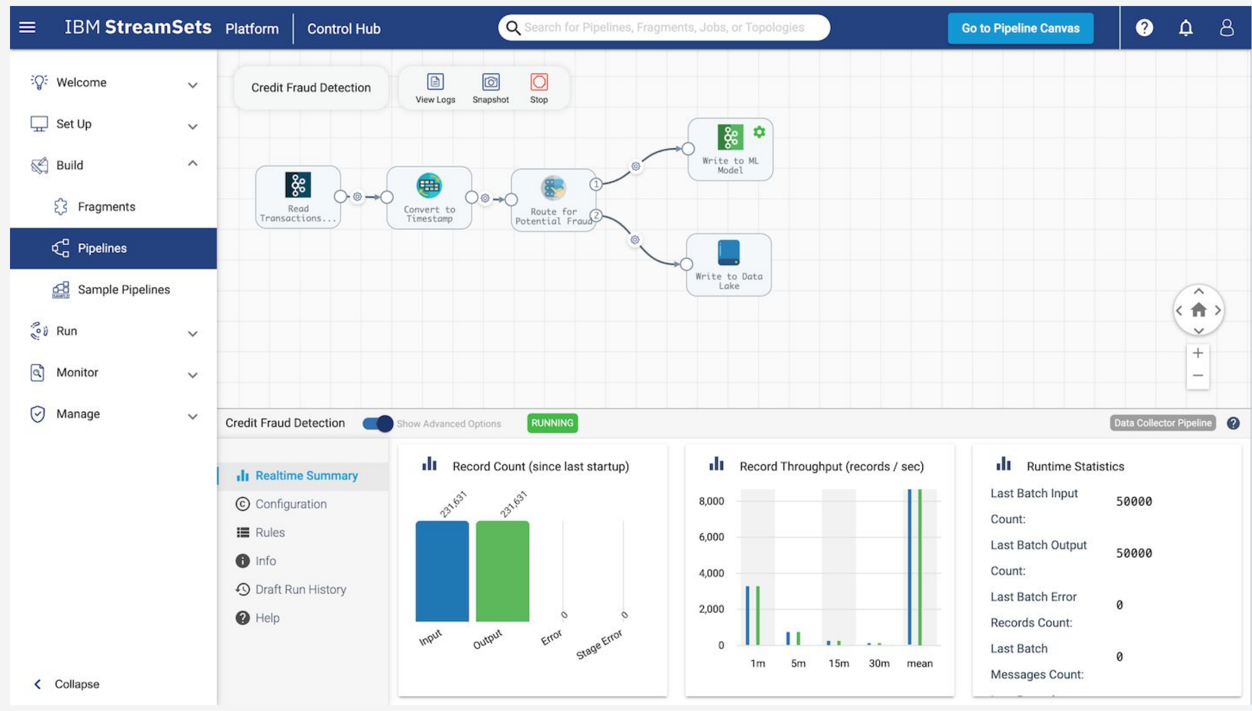
- IBM StreamSets
- Kafka
- Hadoop, AWS (S3, EMR)
- Snowflake, MongoDB

Use cases

- **Risk management and compliance:** Comply with regulations, reduce operational and credit risk exposure, reduce fines
- **Improve customer interactions:** Provide class-leading notification services
- **Optimize customer service:** Understand customer interactions and lifecycles, and offer an industry-leading experience

Results

- **Cost savings:** Reduce customer messaging application cost by £400k
- **Regulatory compliance:** PS2 regulation compliance achieved, delivering 18 million alerts daily. Financial crime compliance achieved
- **Single view** of customer achieved



Last Batch Input Count: 50000
Last Batch Output Count: 50000
Last Batch Error Records Count: 0
Last Batch Messages Count: 0