Databand

Data teams are growing rapidly but still immature in ops practices. Like software teams a decade ago.

DevOps teams emerged to enable software engineers to focus on building applications. Now, the data platform team has emerged to manage data infrastructure services and provides prepared data to analysts and data scientists, so that they can focus on building data products.

1990s
Legacy **software** teams

Computer scientists

1990s
Legacy **data** teams

Data scientists

2010s
Modern **software** teams

Developers
Software engineers

Infrastructure providers
DevOps and operations


splunk> ATLASSIAN

2010s
Modern **data** teams

Data consumers
Data analytics and science

Data providers
Data platform and engineering

Databand

When a data platform is inefficient, everyone is impacted

Databand

90%

of data flows through data platform and engineering.

Source: Gartner Data Engineering Essentials, Patterns and Best Practices, 2021

Data analytics
Inaccurate business analytics

Data product
Frustrated customers from bad client facing data

Data science
Underperforming machine learning

Unreliable data is the most costly problem enterprise data teams face today

Wasted time

44% Of day maintaining data pipelines

No confidence in results

71% said end users make business decisions with old or error-prone data

Lost revenues

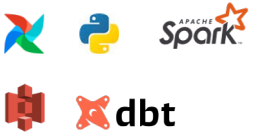
85% enterprises make bad decisions that cost them revenue

The modern data stack

Databand

Sources

Data lake



Data warehouse

Data access
(Analysts, scientists)

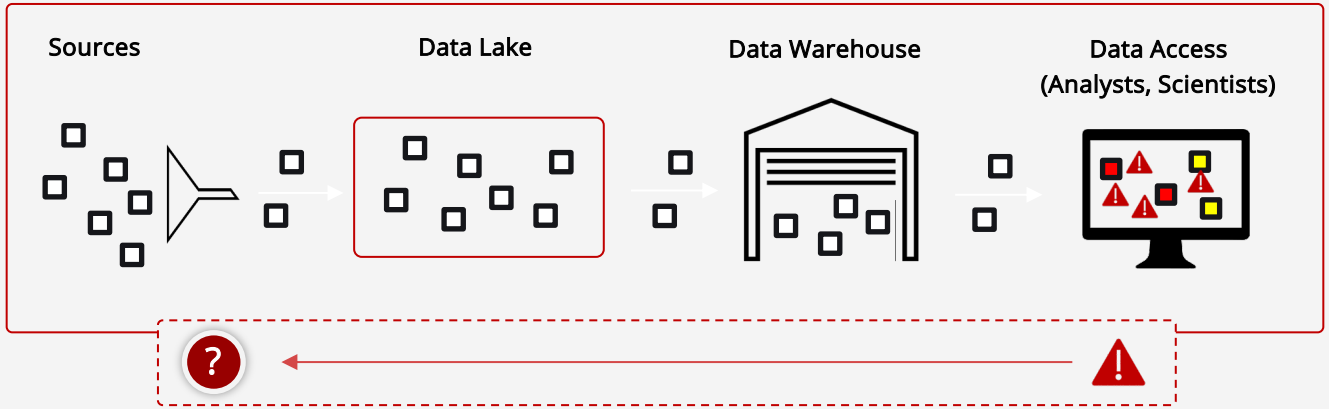
Data providers
Data platform and engineering

Data consumers
Data analytics and science

Data Observability

Problem Statement

Data engineers are reactive to data issues



Many data quality issues are overlooked

Platform only learns about issues when reported by data consumers

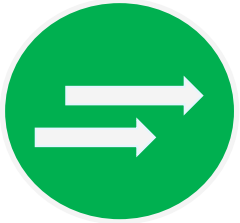
After issues are reported, they are not resolved quickly

Databand

Databand
an IBM Company

The root causes

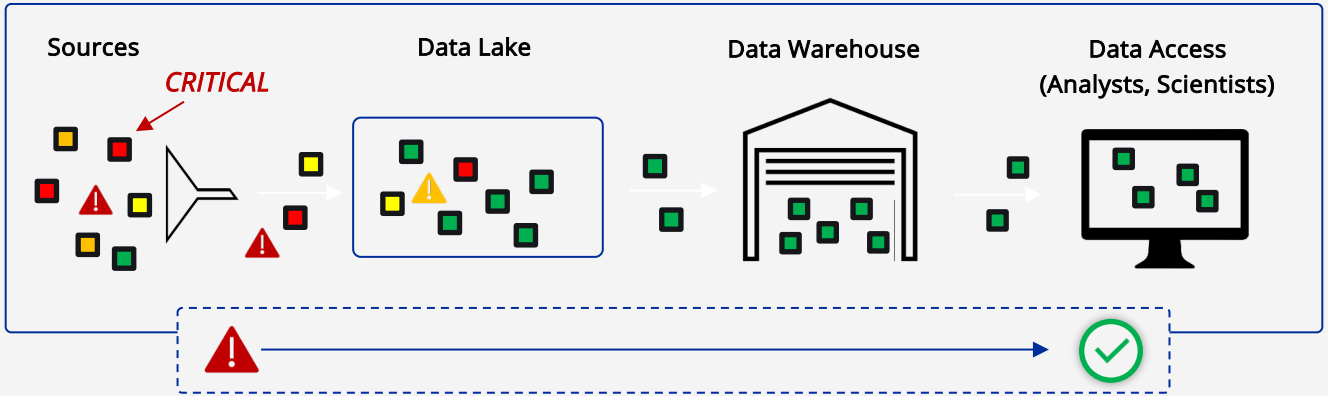
- Fragmented toolchain
- Volume of data
- Flooding of noise



Databand focuses on observing <u>data in motion</u>

Solution

Proactive Data Observability: *Shift left* and solve problems at the source

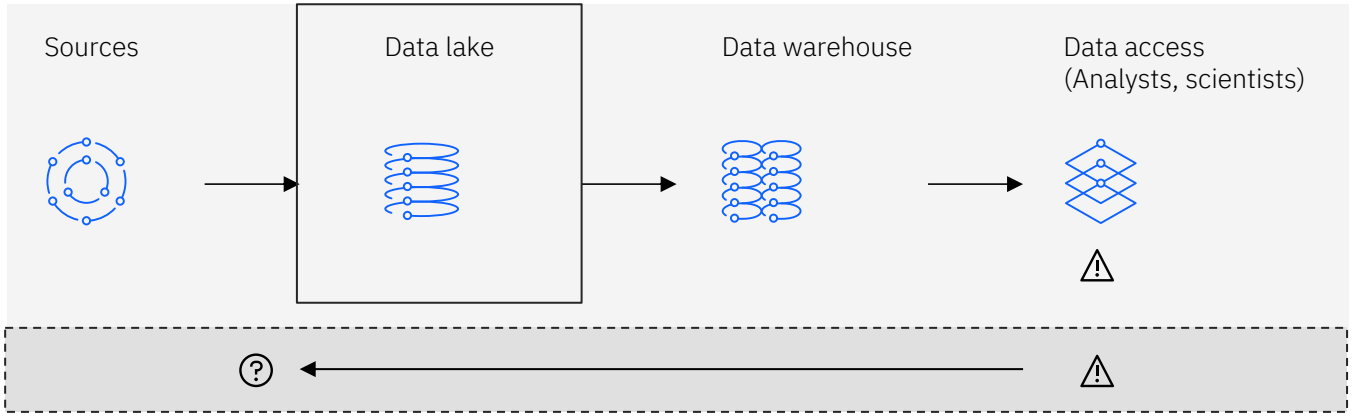


Observe data pipeline process quality
- Status
- Performance
- Latency

Observe data quality and reliability
- *Schema changes*
- *Data shape*
- *Data freshness*

Supported data pipelines and workflow managers
- *Airflow*: Python, Spark, dbt, SQL and other operators

Before databand. Data engineers are reactive to data issues



Most data quality issues are **never discovered**

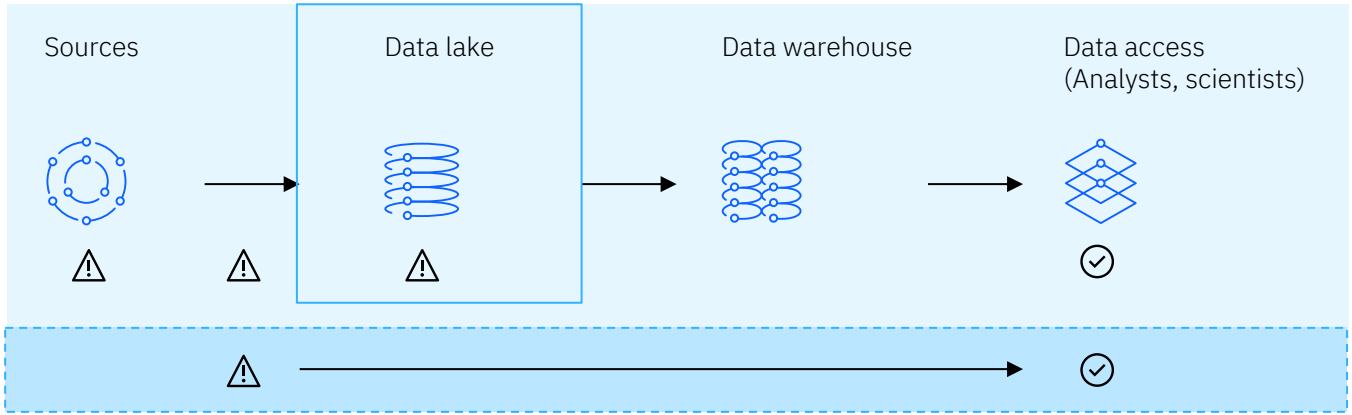
Platform only learns about issues when **reported by data consumers**

After issues are reported, they take weeks to fix

The root causes

- Fragmented toolchain
- Explosion of data
- Flooding of noise

After databand. Proactive data observability: Shift left and solve problems at the source



Improve **mean time to detect (MTTD)**

Discover issues in real time, early as ingest

Improve **mean time to resolve (MTTR)**

Identify the cause of issues instantly

Improve data product **quality**

Enhance trust and consumer satisfaction

The required solution

- Standard metadata collection
- Visibility into all data flows, from source to target
- Ability to separate signal from noise

Five steps to proactive data observability

1. Pipeline execution

Is data flowing?

2. Pipeline latency

Is data arriving on time?

3. Data structure

Is the data shape valid and complete?

4. Data content

Are there significant changes in the data profile?

5. Data validation

Does the data content conform to how it is being used?

Five steps to proactive data observability

Data SLAs—real-time alerting. **Managed by** data platform and engineering.

1. Pipeline execution

Is data flowing?

2. Pipeline latency

Is data arriving on time?

3. Data structure

Is the data shape valid and complete?

4. Data content

Are there significant changes in the data profile?

5. Data validation

Does the data content conform to how it is being used?

Custom metrics and assertions
—real-time or asynchronous alerting
Managed by data Data analytics and science.

Our solution

Databand

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

2. Profile
   Build historical baseline. Based on common data pipeline behavior.

3. Alert
   Alert on anomalies and rules. Based on deviations or braches.

4. Resolve
   Resolve through automation. Create smart workflows to remediate data quality issues and keep SLAs on track.

Our solution

Databand

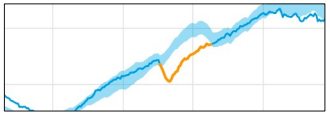
1. Collect

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

pandas databricks Apache Airflow dbt SQL

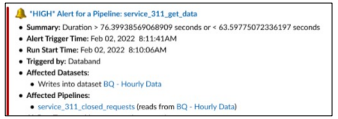
2. Profile

Build historical baseline. Based on common data pipeline behavior.



3. Alert

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



4. Resolve

Resolve through automation. Create smart workflows to remediate data quality issues and keep SLAs on track.

The Databand solution

1. Collect

Automatically collect metadata

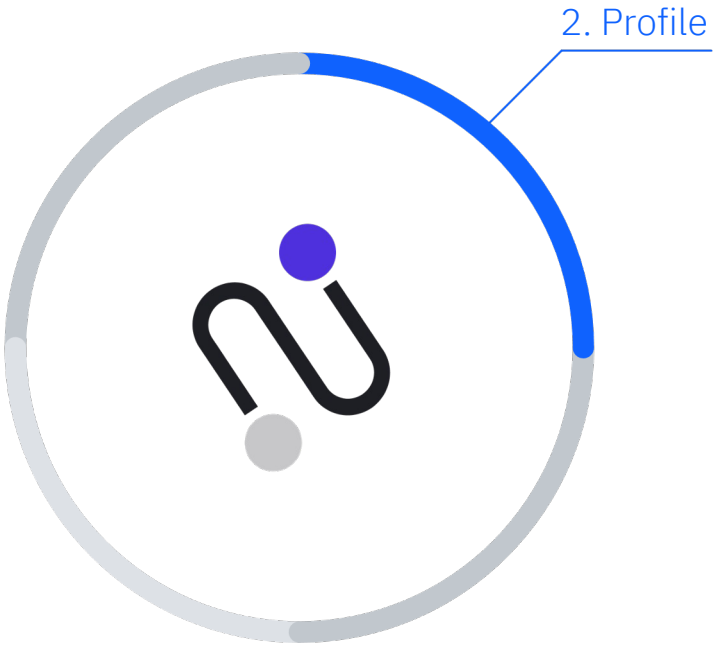
From all key solutions in the modern data stack.

pandas databricks Apache Airflow dbt SQL

Empower platform to gain immediate visibility into mission- critical metadata. Provide analysts and scientists with a standard method for customized data-quality validations.

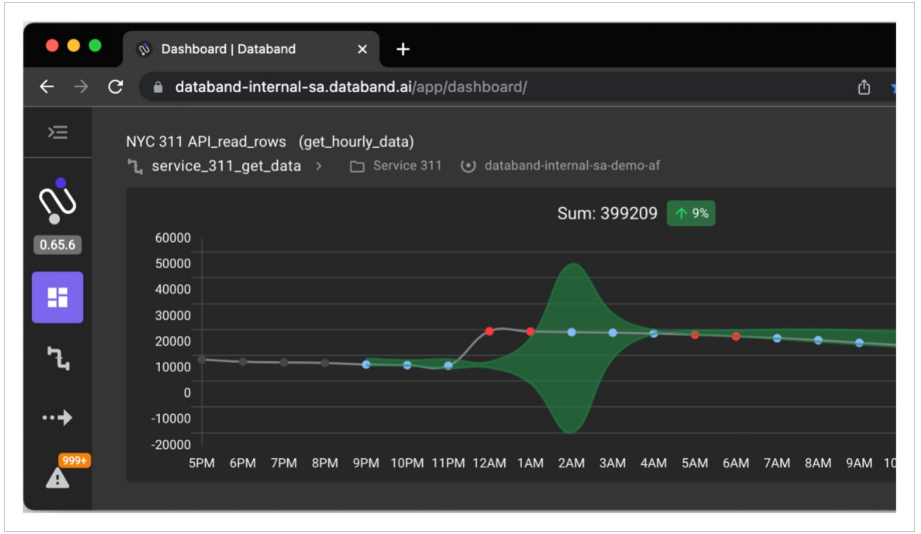
The Databand solution

Databand



Builds historical baseline

Based on common run and data behaviors.



The Databand solution

Databand

Alerts on anomalies and rules

Based on deviations relative to historical behavior or rule breaches.

HIGH Alert for a Pipeline: service_311_get_data

- **Summary:** Duration > 76.39938569068909 seconds or < 63.59775072336197 seconds
- **Alert Trigger Time:** Feb 02, 2022 8:11:41AM
- **Run Start Time:** Feb 02, 2022 8:10:06AM
- **Triggerd by:** Databand
- **Affected Datasets:**
  - Writes into dataset BQ - Hourly Data
- **Affected Pipelines:**
  - service_311_closed_requests (reads from BQ - Hourly Data)

The Databand solution

Resolve through automation

When you create smart workflows that automatically remediate data quality issues and keep your data deliveries on track.

Databand Customer Profile

Use Case Pattern: Operational Monitoring

Organization: Data engineers & Data leadership

Technologies:

- DataStage
- Apache Airflow
- DBT
- Apache Spark / Databricks

Challenges / Pains

- Poor Data Reliability
- Fragmented Data Quality Monitoring
- Unknown Impact Analysis

Questions to ask:

- Which data pipeline and orchestration technologies do you use? Examples: AirFlow, DataStage, Python data pipelines, Spark data pipelines, dbt
- How do you assess quality of your data pipelines?
- How do you verify that your ingestion processes are delivering the right data?
- Do you trust that your data ingestion framework is working?
- What do your data teams do when issues are found during ingestion?
- How long does it take to pinpoint and resolve them?

Deployment Options

In September, we GA'ed two deployment options of Databand for maximum flexibility

SaaS

New features, fast with SaaS running on Google Cloud Platform

Self Hosted

Enterprises can self host in environment of their choosing

What does this mean?
Customer example

“We are an AWS/Azure shop. Does Databand run on AWS/Azure?”

Self-hosted

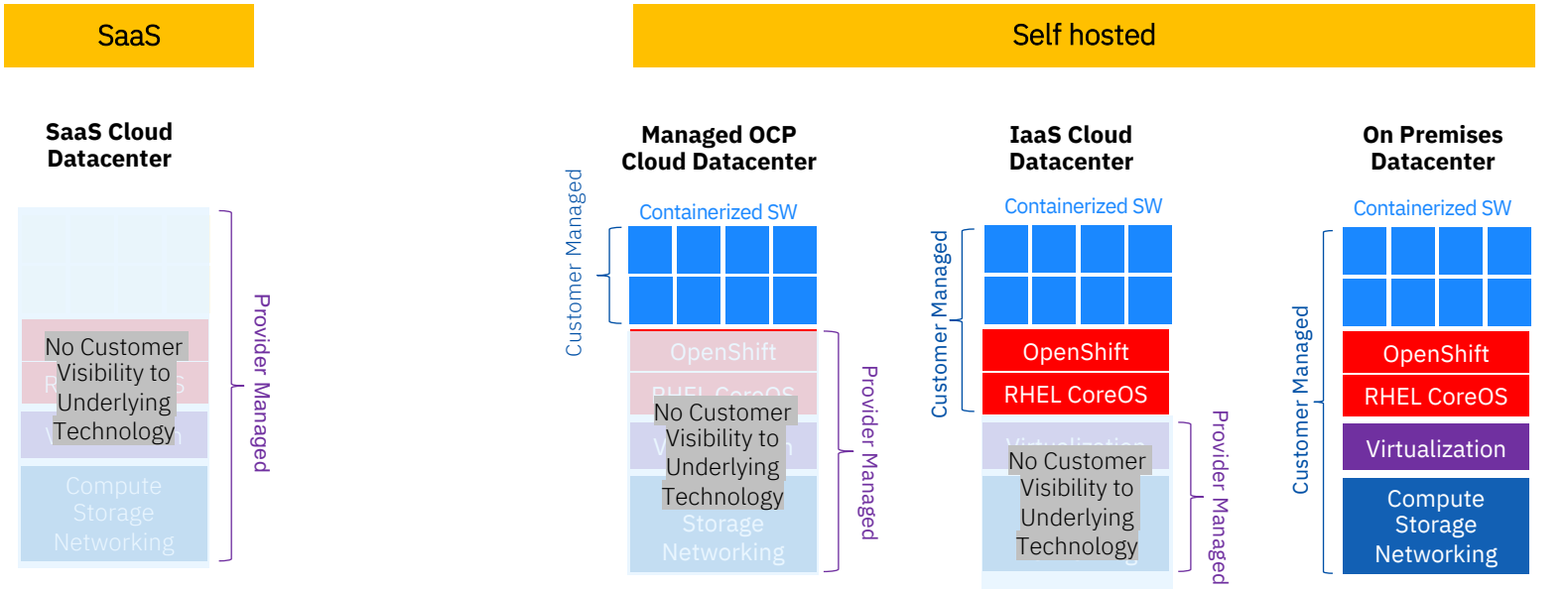
“Databand **Self-hosted** is a containerized application and can be deployed using Kubernetes in the cloud of your choice”

SaaS

“Alternatively we offer **Databand SaaS** which is available in three data centers globally – US East, Frankfurt, and Sydney. This runs on Google Cloud and meets IBM's security requirements"

Both options are cloud relevant

SaaS brings the greatest client value because it’s 100% managed



Databand and MLOps

MLOps - Observability