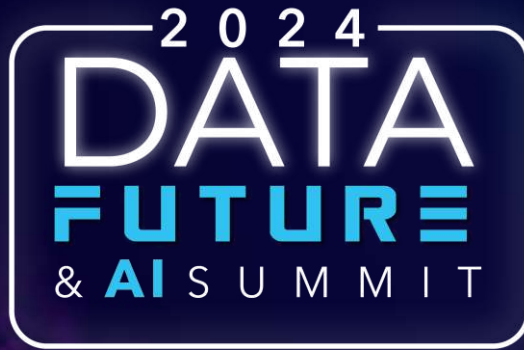




Data observability beyond the hype – for the reliability of your enterprise data.

Takeshi Kamimura

IBM Global Brand Executive, Data Observability



Why

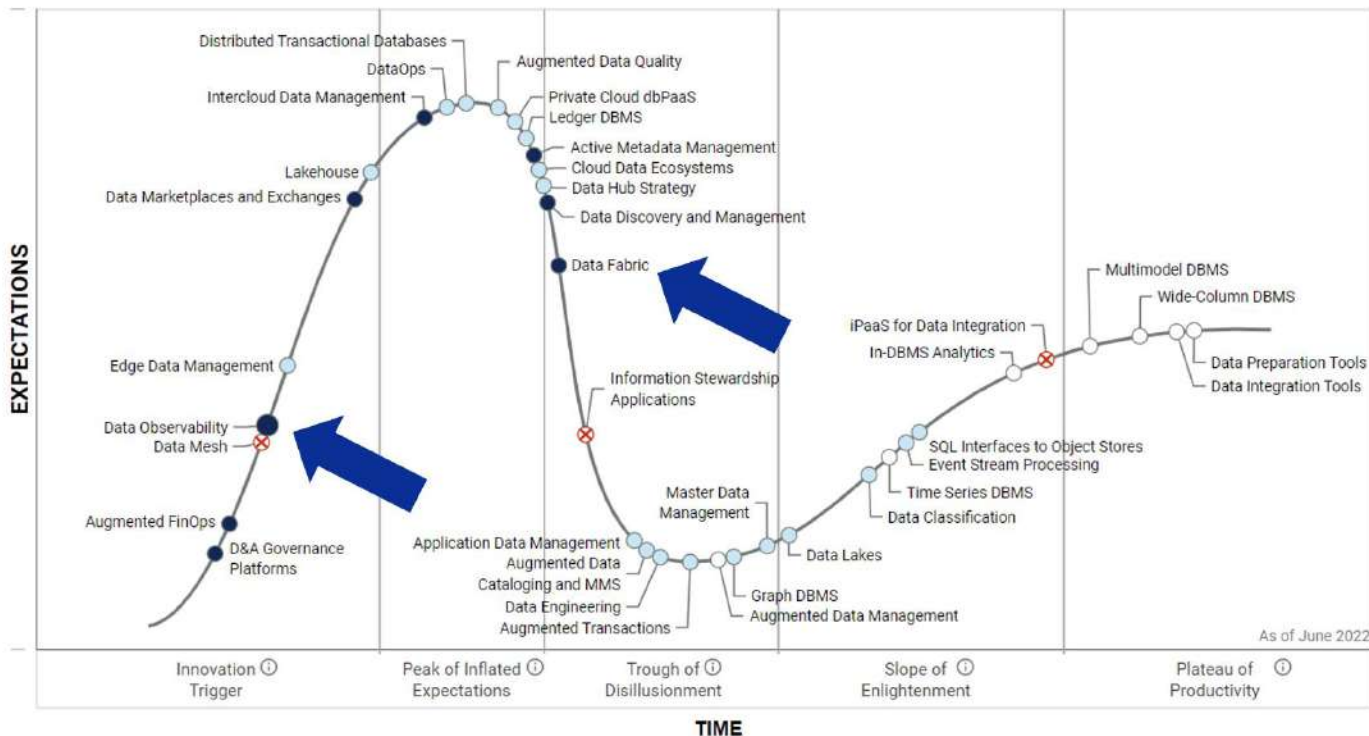
A little story



A full-page background image of an underwater scene. On the left, a large, dense school of silver fish, possibly jackfish, swims towards the right. On the right, a scuba diver in a blue wetsuit and mask swims towards the left, with a trail of bubbles rising from their regulator. The bottom of the frame shows a rocky seabed covered in green algae or coral. The water is a deep, clear blue.

Data observability
is the foundation
for reliable data

Why Now? Why IBM



Source: Gartner, Hype Cycle for Data Management, 2022, Donald Feinberg, Philip Russom, Nina Showell, 30 June 2022

Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.



However, many organizations consider their **data unreliable**

82%

say data quality concerns represent a barrier to their data integration projects

80%

of business executives surveyed do not trust their data

73%

of business executives surveyed are unhappy with data quality

And, this **unreliable data** is costly from a time and financial perspective

6%

impact on annual revenues from inaccurate customer data

40%

average proportion of time data teams spend on data cleaning, integration and preparation

\$15M

average cost of bad data

Thus, many organizations are unable to leverage **their data for innovation**

61%

of organizations unable to harness data to create a sustained competitive advantage

63%

of organizations unable to harness data to introduce a new business model

63%

of organizations unable to introduce a new product/service

The Importance of Data Observability

MarketWatch

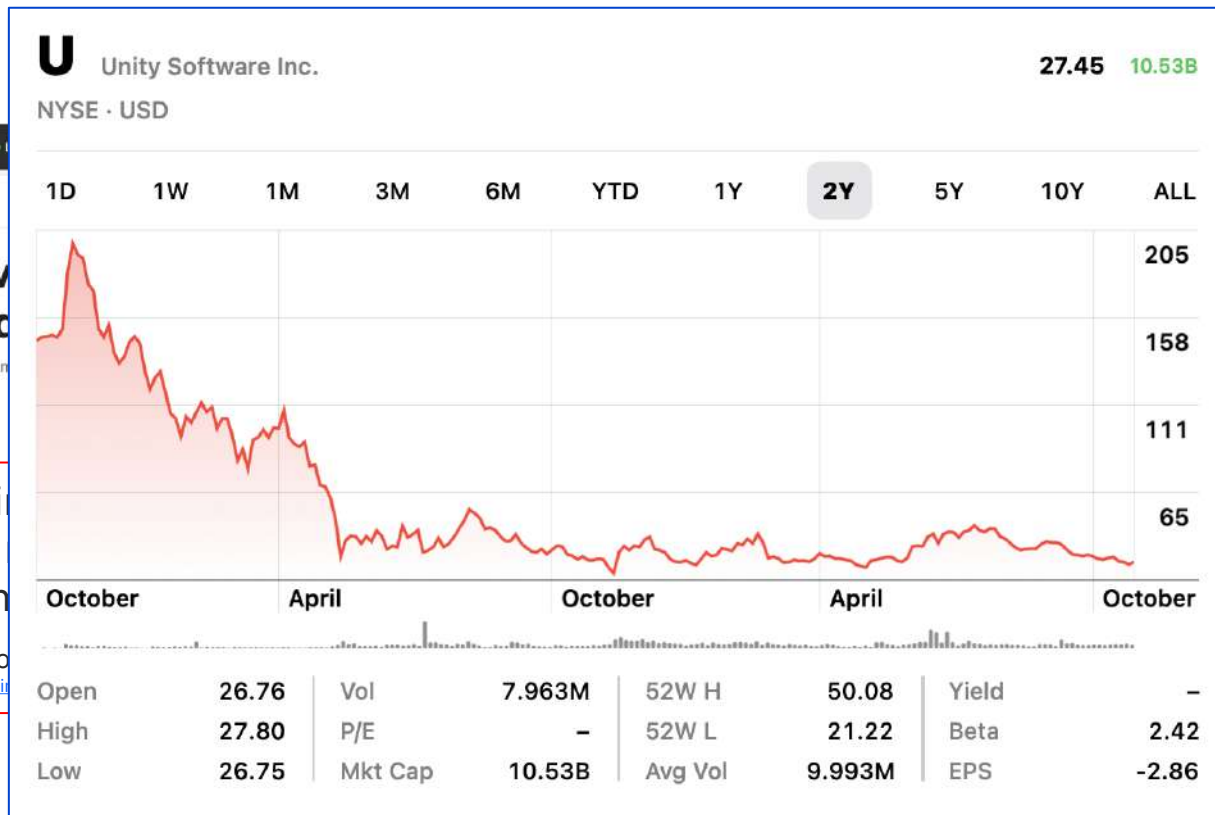
Home > Industries > Software

Unity Software 30% on wed

Published: May 10, 2022 at 8:49 p.m.

“In the meantime,
expects \$110
from this one n

– John Quast, The Motley Fool
[its-recovery-with-this-1-surprise](#)



tion of our
to us
large
very to go
data
nd
enue

0, Q2 2022

What

Introducing [Databand.ai](#)



Leverage data observability
for the reliability
of your enterprise data



On June 27, 2022, IBM acquired [Databand.ai](#) to enable clients dynamically and automatically manage the reliability of all data spanning the enterprise data lifecycle to provide quality data products for trusted business outcomes.

Data monitoring vs data observability

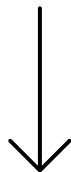


Logging, Return Codes,
Dashboards, home made
data quality

Abnomalies, Profiling,
Trending, Statistical
deviation

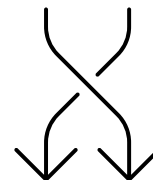
70%

“The pipeline **failed**...”



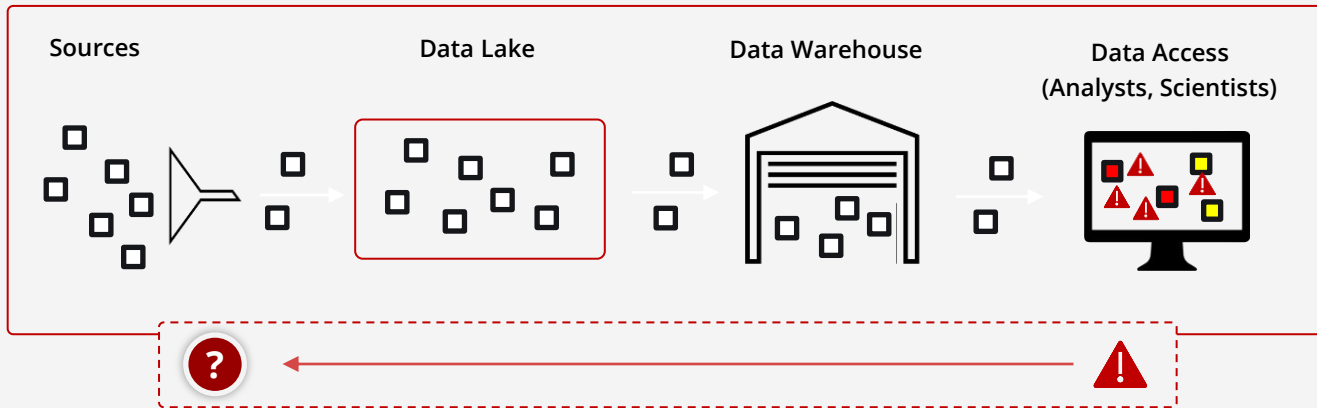
Data Monitoring	Data Observability
Limited View	Holistic View
What Happened	What, Where, and Why it Happened
Isolated System	Dependent Systems
Static Detection	Dynamic Detection
Reactive	Proactive

“The pipeline **failed**,
here’s the **root
cause** and how this
will impact **XYZ
dependencies**...”



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**

The root causes



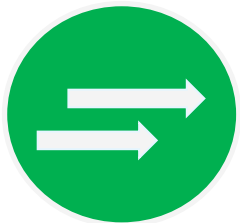
Fragmented toolchain



Volume of data



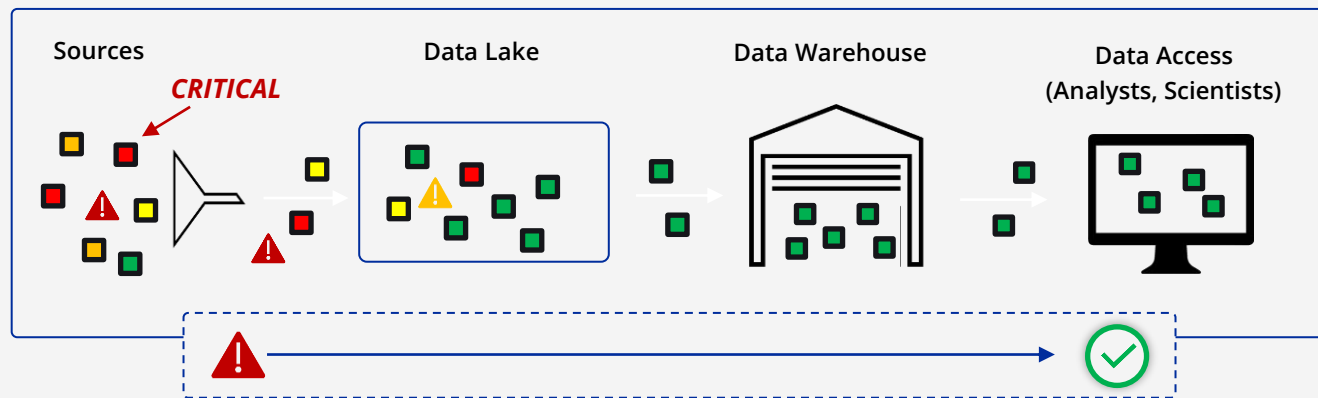
Flooding of noise



Our solution
focuses on
observing **data**
in motion

Solution

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



Improve **MTTD**

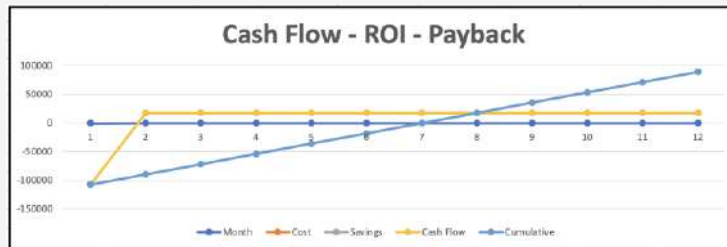
*Discover issues in real
time, early as ingest*

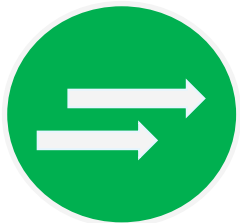
Improve **MTTR**

*Identify the cause of issues
instantly*

Improve data product
quality

*Enhance trust and
consumer satisfaction*

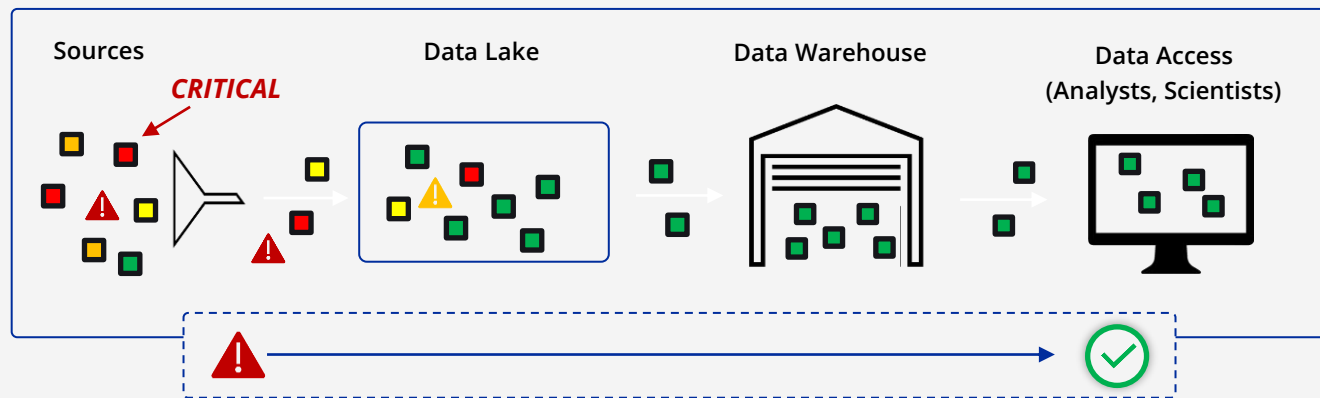




Databand
focuses on
observing data
in motion

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*

Observe **Lineage**

- *Impact Analysis*
- *Pipeline Discovery*
- *Root Cause Identification*

Client Reference Databand + ML Engineering



Qaish Kanchwala

ML Engineer Manager and ML Architect
The Weather Company

“Databand provides an operational view of how well we’re running our Airflow and Sagemaker pipelines. Without this operational view, it would be extremely hard to understand the overall health of our ML pipelines.”

Roles and Responsibilities

- Qaish manages a team of eight engineers, including DevOps, ML, and Data engineers.
- Responsible for building segments and training the ML models used in production for marketing ad campaigns.
- Revenue generation and business decisions rely on the accuracy of data used with these segments.

ML/AI Tech Stack

- Databand (SaaS) for Data Observability
- Airflow (EC2) for Data Pipelines on EC2
- Sagemaker (AWS) for ML
- Jenkins for CI/CD

Challenges and Business Impacts

Poor decisions and declining revenue based on inaccurate model predictions resulting from:

- Manual pipeline monitoring
- Limited alerting and lineage views
- Wrong data values from data drift or broken pipelines

How Databand’s observability helps

- Continuous visibility and transparency for Airflow & Sagemaker pipelines
- SLA alerting and automated lineage tracking
- Data quality monitoring for better prediction models

How

The Databand.ai solution

1

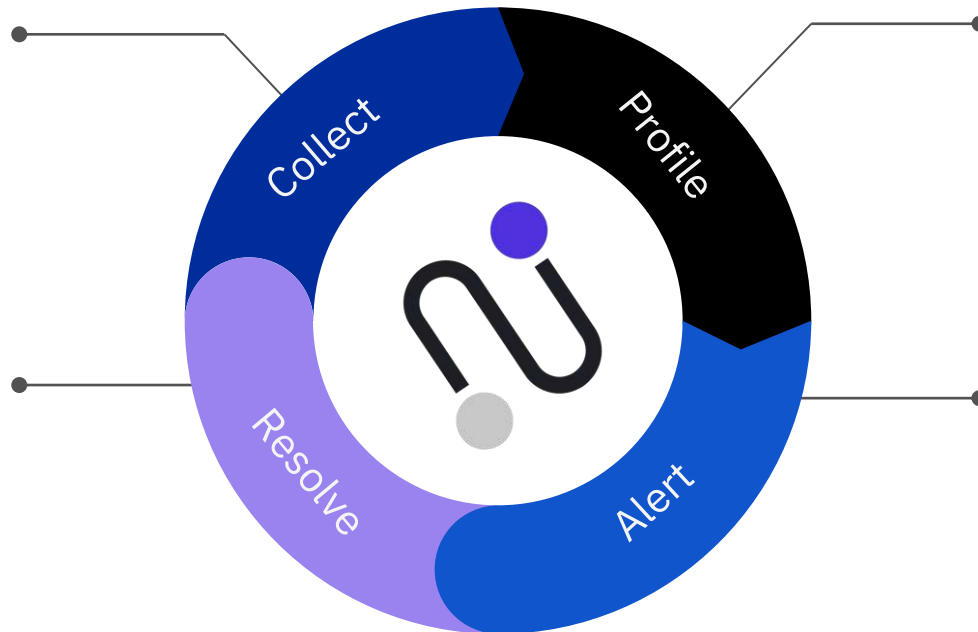
Automatically collect metadata

From all key solutions in the modern data stack.

4

Resolve through triage

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



2

Build historical baseline

Based on common data pipeline behavior.

3

Alert on anomalies and rules

Based on deviations or breaches.

The Databand.ai solution

Automatically collect metadata

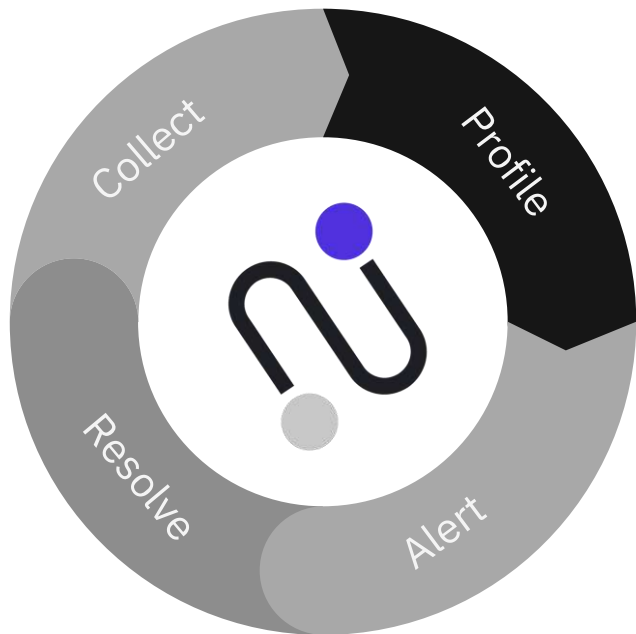
From all key solutions in the modern data stack.



Databand.ai automatically collects standard data quality and pipeline performance metrics instantly after integration, empowering data engineers to gain immediate visibility into mission critical metadata.

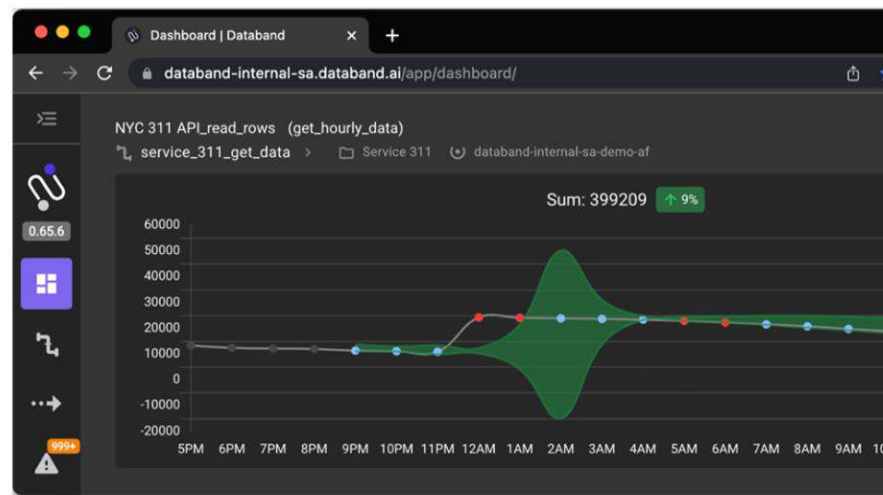
For additional metrics and data validations, analysts and scientists can define custom monitors specific to their key data sets.

The Databand.ai solution



Builds historical baseline

Based on common run and data behaviors.



The Databand.ai solution



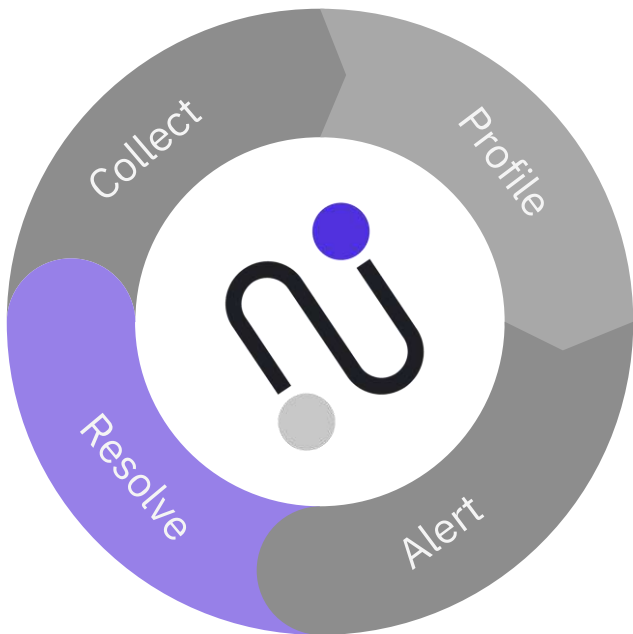
Alerts on anomalies and rules

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
- **Triggered by:** Databand
- **Affected Datasets:**
 - Writes into dataset [BQ - Hourly Data](#)
- **Affected Pipelines:**
 - [service_311_closed_requests](#) (reads from [BQ - Hourly Data](#))

The Databand.ai solution



Resolve through automation

Create smart workflows that automatically remediate data quality issues and keep your data deliveries on track.

Automated scale-up relating to SLAs

- Monitor real time data volumes
- Automatically provision higher resources if data increases

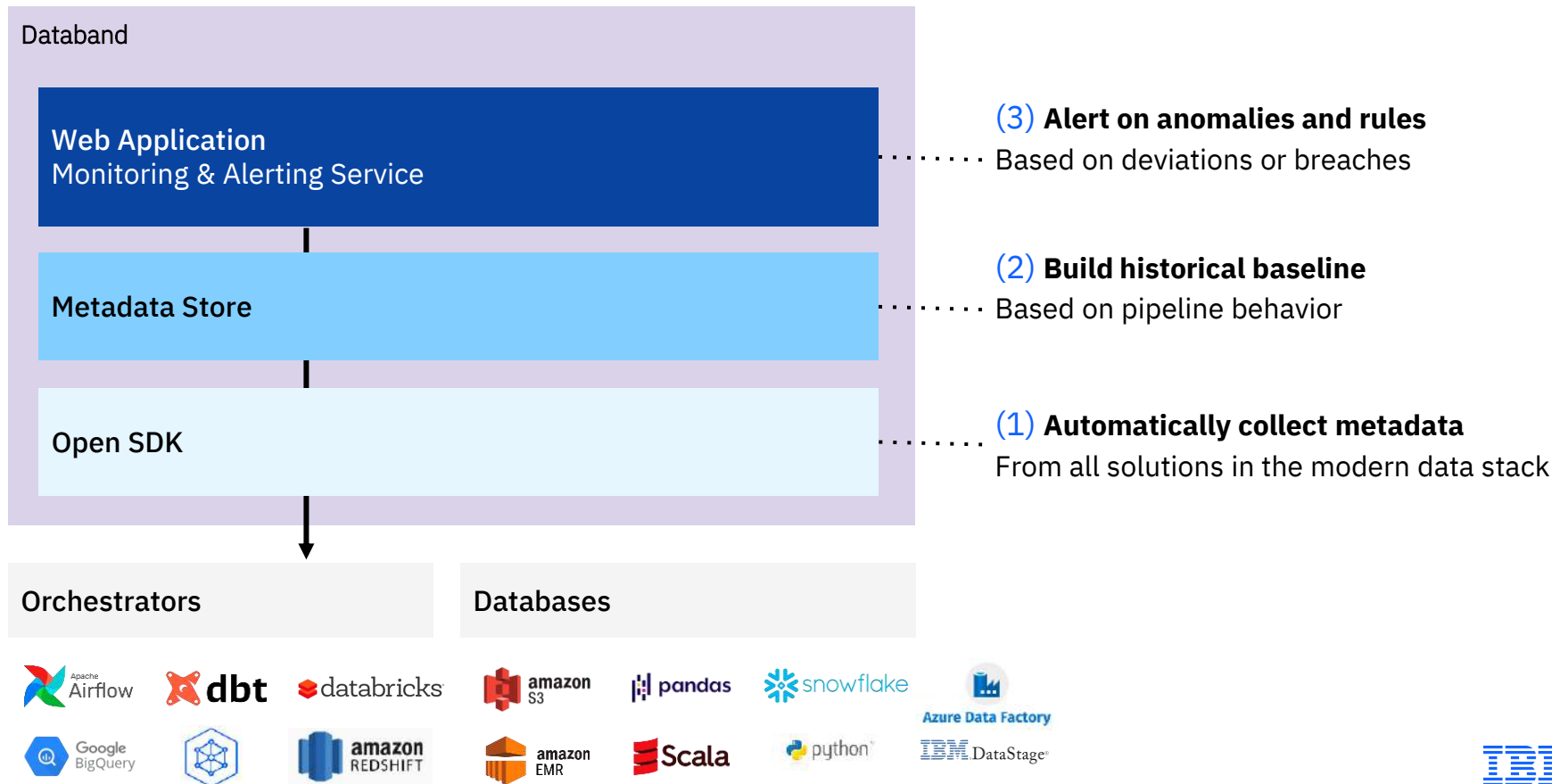
Automatically route bad data

- Monitor data profiles on ingestion
- Halt downstream processing, point low quality data to staging

Automated retraining to improve model accuracy

- Measure models for performance
- Retrain and push to prod if under threshold

How it works



Key Differentiators

High variety of critical metadata captured in motion

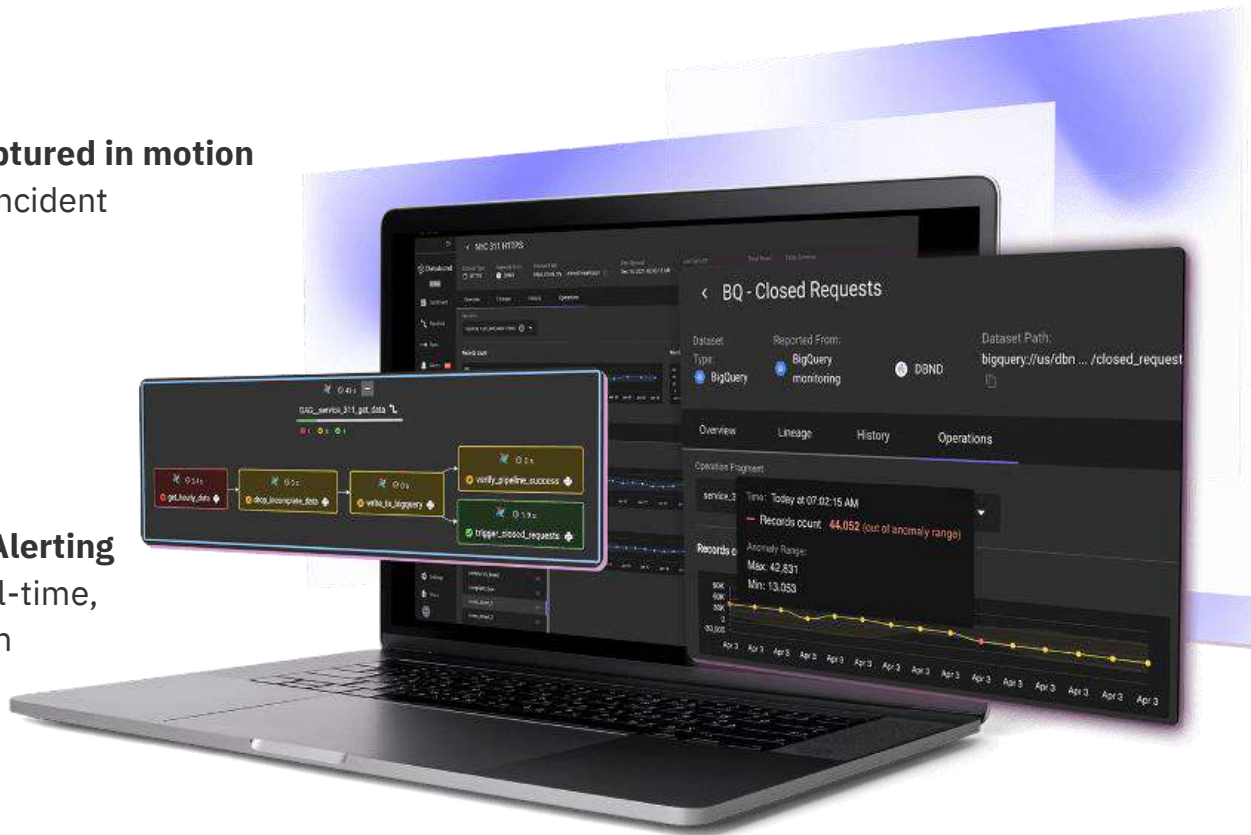
Databand collects all key sources of incident information

Coverage across technologies

Databand traces issues from the lake through the warehouse

Anomaly detection with Real Time Alerting

Databand raises incident alerts in real-time, providing the fastest time to detection



Next Steps

Data Observability Assessment

- Share the data observability value proposition with the rest of the organization
- Review the drivers and benefits of data observability
- Show (demo) not tell
- Assess the pre-reqs for implementation
- **Prioritize use case** to implement and scope of activities for the one-week PoC from the following topics:
 - Identify PoC Sponsor
 - Determine deployment option
 - Determine customer admin for the PoC
 - List Technology stack (Airflow, ADF, Spark, Python, etc)
 - Define Volumes of Data Pipeline definitions

Prerequisites:

- Technology Stack
- Above 200 Data Pipeline definitions in Prod

Duration:

- 4-8 hours





5 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 analytics and science



www.gbm.net    GBM Corp

Desde cualquier país marque la extensión 3840 (Contact Center)
GT (502)2424-2222 | ES (503)2505-9600 | HN (504)2232-3219/09
NI (505)2255-6630 | CR (506)2284-3999 | PA (507)300-4800
RD (809)566-5161 | MI +1(305)597-3998