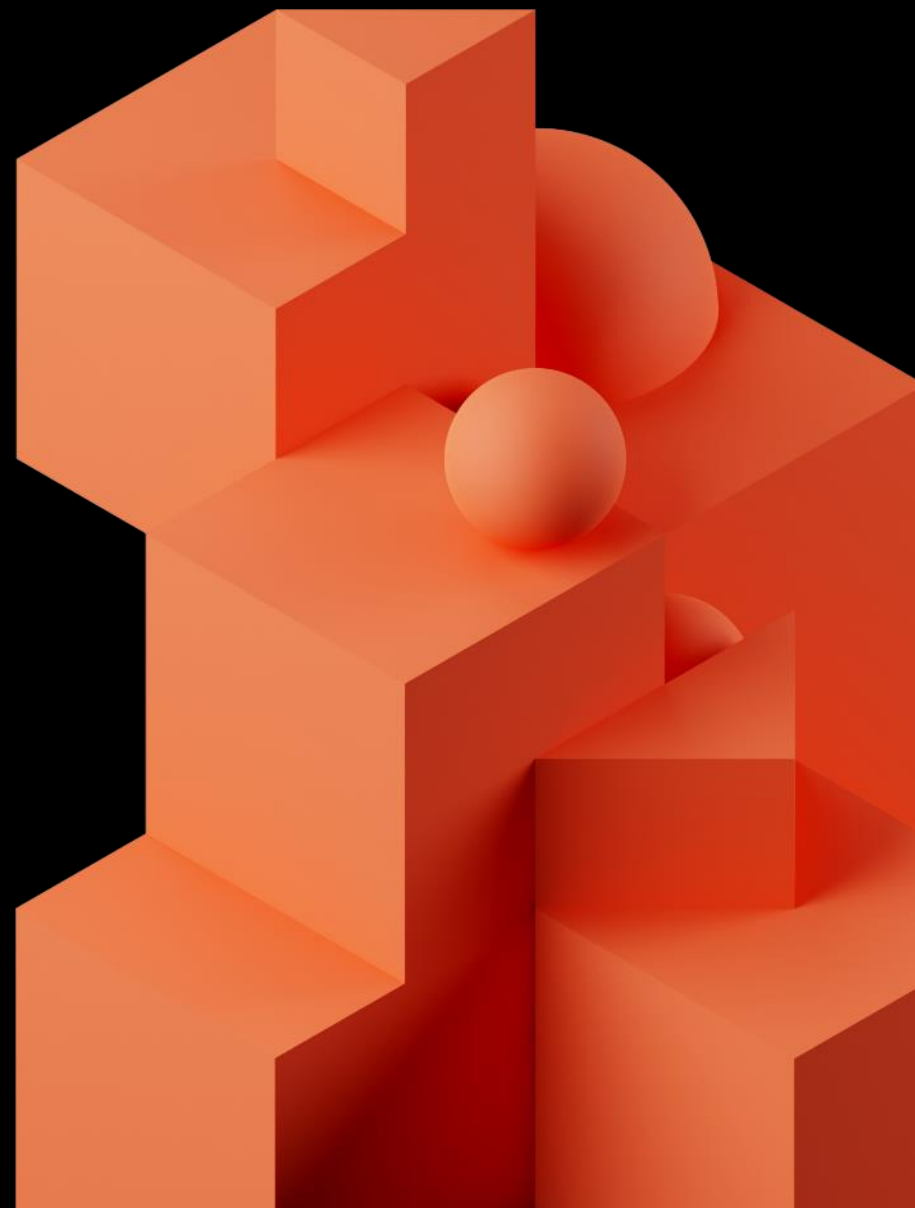


How Comcast's Effectv Drives Data Observability with Databricks and Monte Carlo

Databricks
2023



Introductions



SCOTT LERNER

Customer Success Manager, Monte Carlo

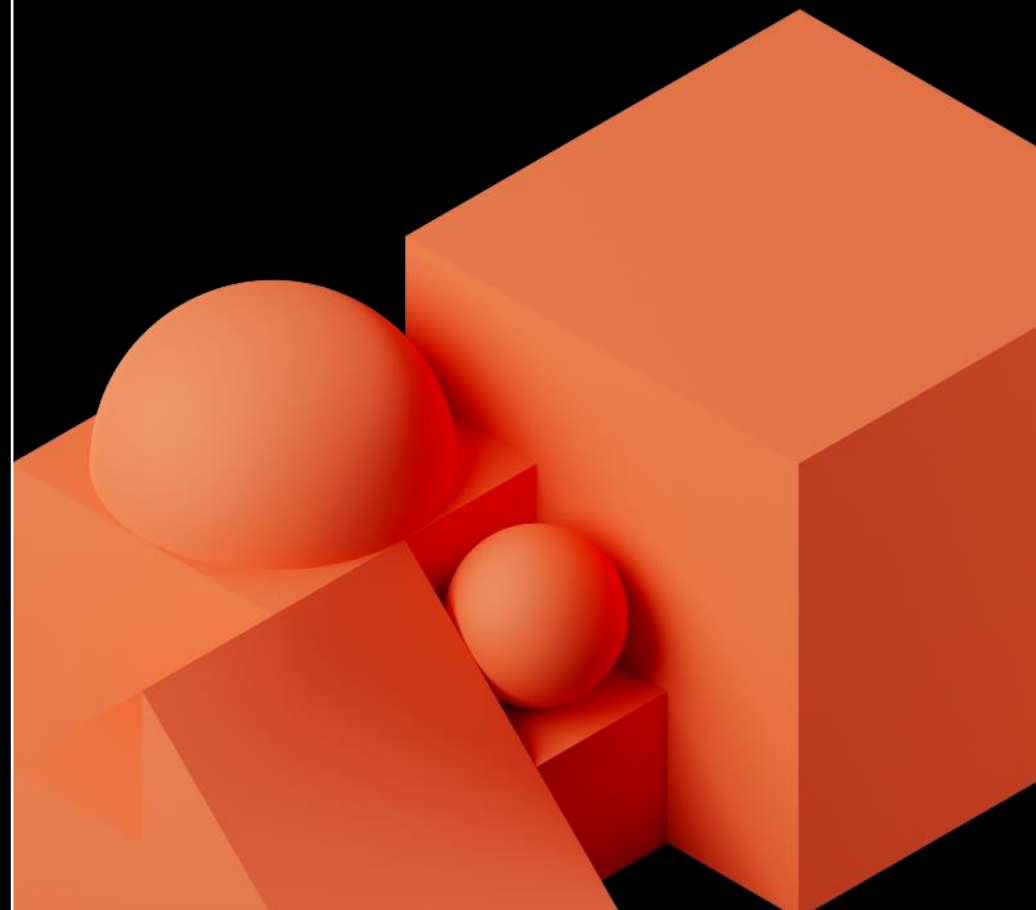


ROBINSON CREIGHTON

Sr. Manager DataOps, Effectv



What is Monte Carlo?



What is Data Observability?

Good pipelines, bad data

- Is the data up-to-date?
- Is the data complete?
- Are fields within expected ranges?
- Is the null rate higher or lower than it should be?
- Has the schema changed?
- ... and many more



DATA OBSERVABILITY PILLARS



Freshness

Volume

Quality

Schema

Lineage

What is Data Observability?

Good pipelines, bad data

Detect

- ML-powered anomaly detection
- Rule-based detection
- Targeted alerts to impacted owners & downstream users

Resolve

- Automated field-level lineage
- Impact radius assessment
- Code, data, and operational diagnostics

Prevent

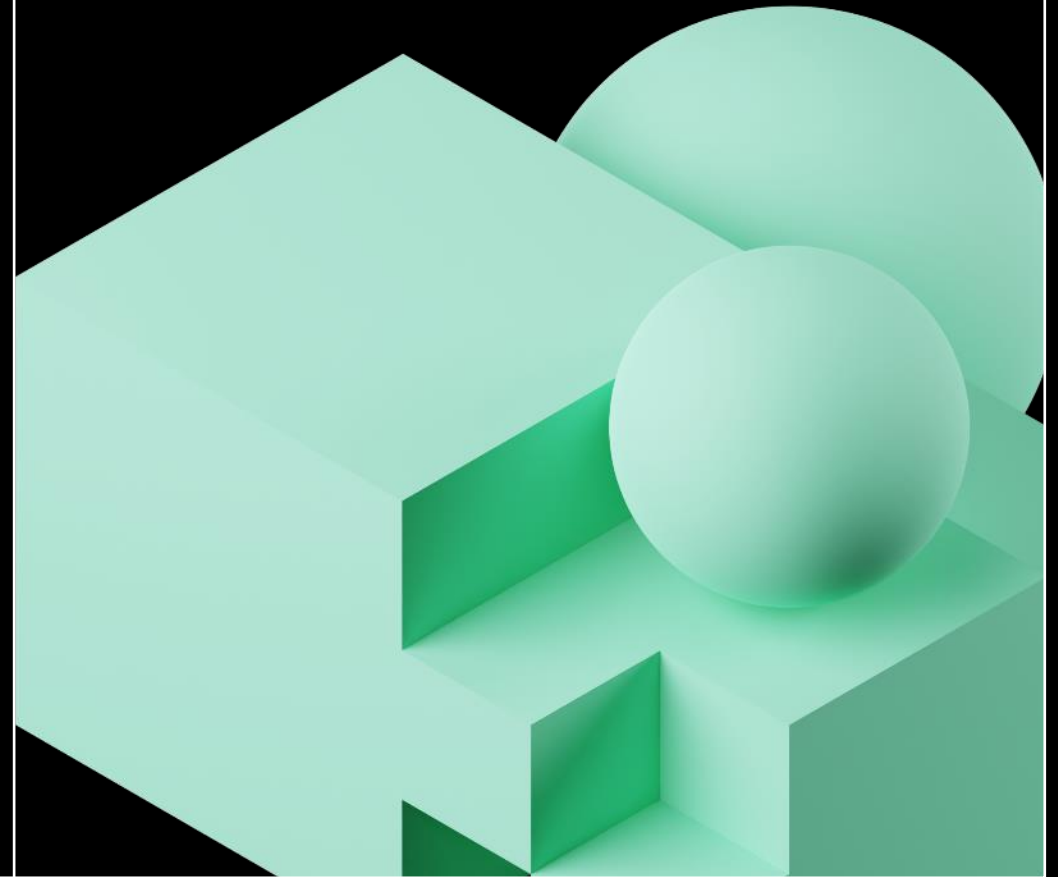
- Auto-generated and on-demand insights
- Schema change notifications
- Automated circuit breakers



DATA OBSERVABILITY PILLARS

Freshness | Volume | Quality | Schema | Lineage

Effectv: Who We Are & What We Do



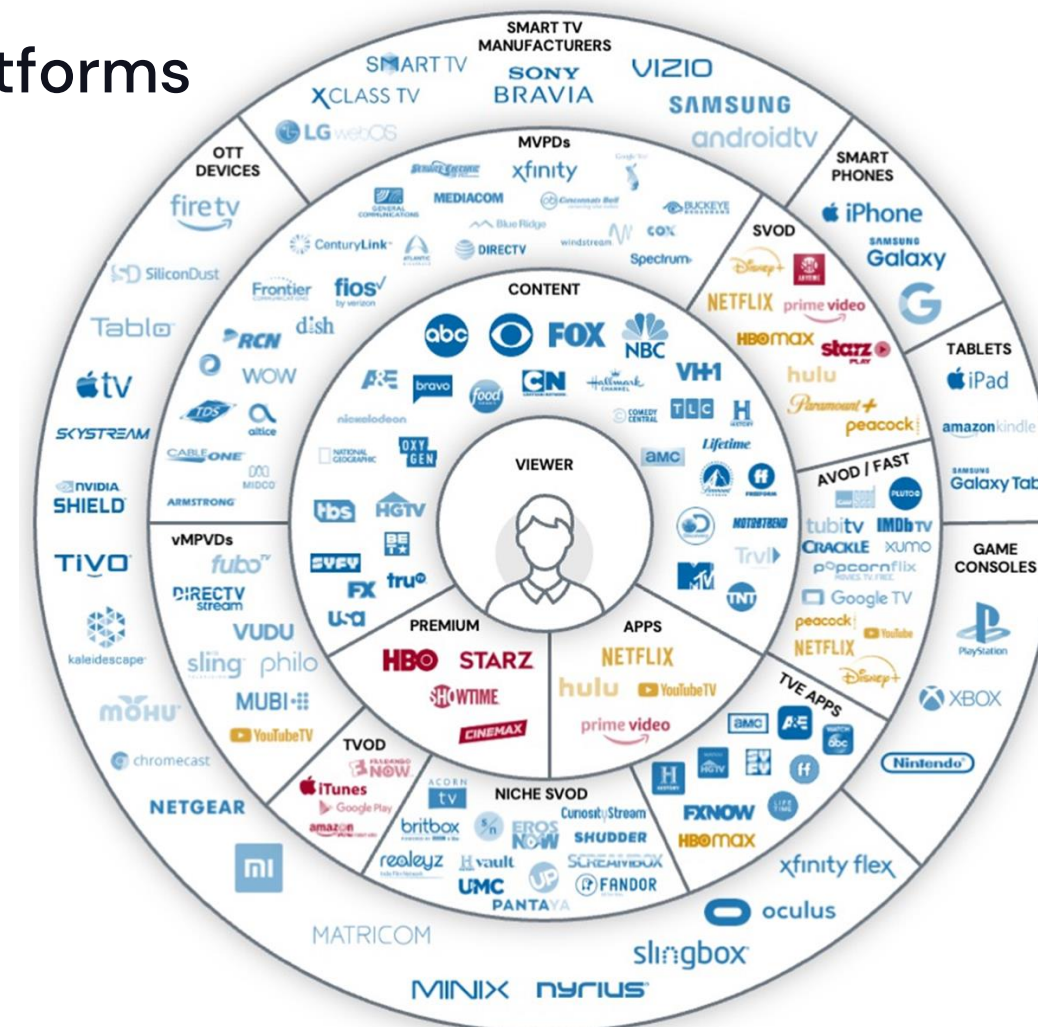
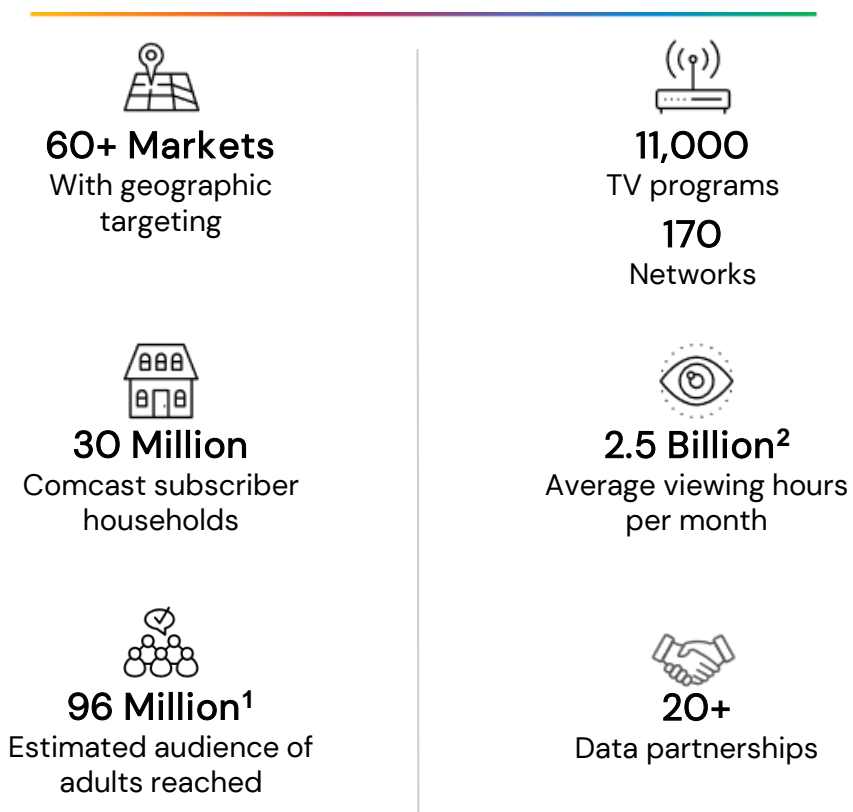
Effectv is an
audience delivery
company that combines
the best of digital with
the power of TV



Data at Effectv



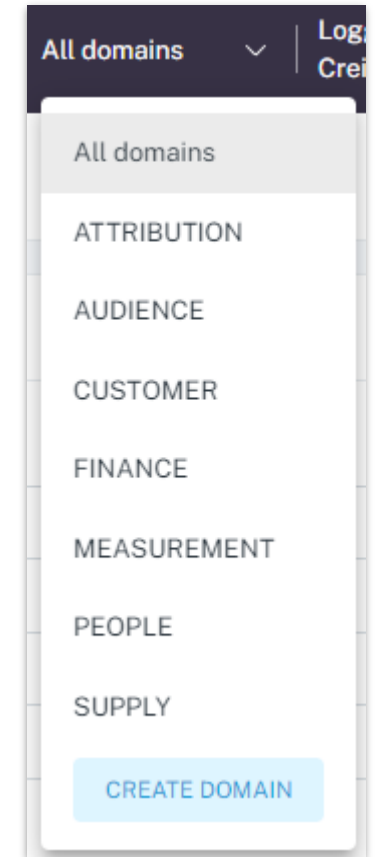
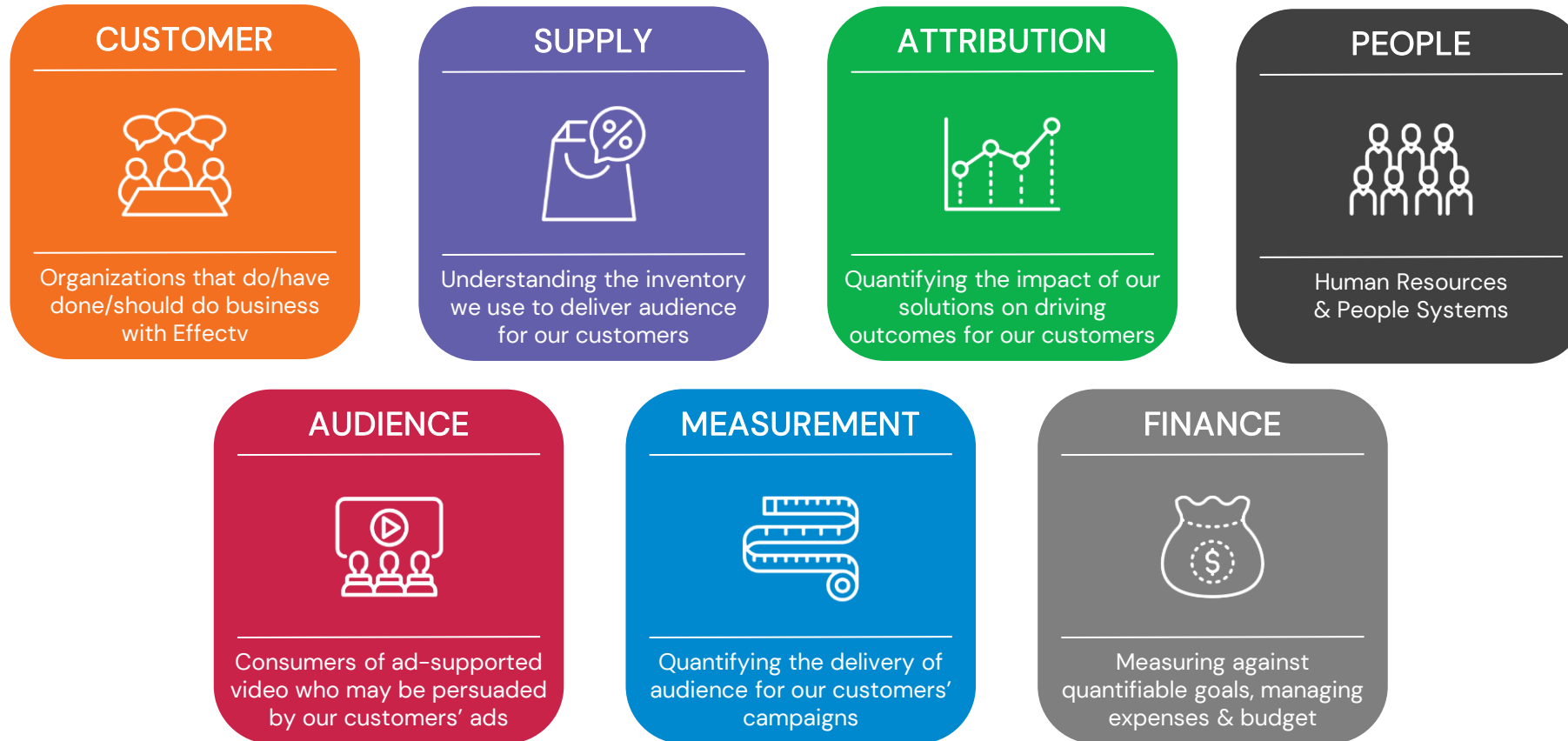
5 billion streams monthly across our platforms



Ad supported by Effectv | Not ad supported by Effectv | Not ad supported at all

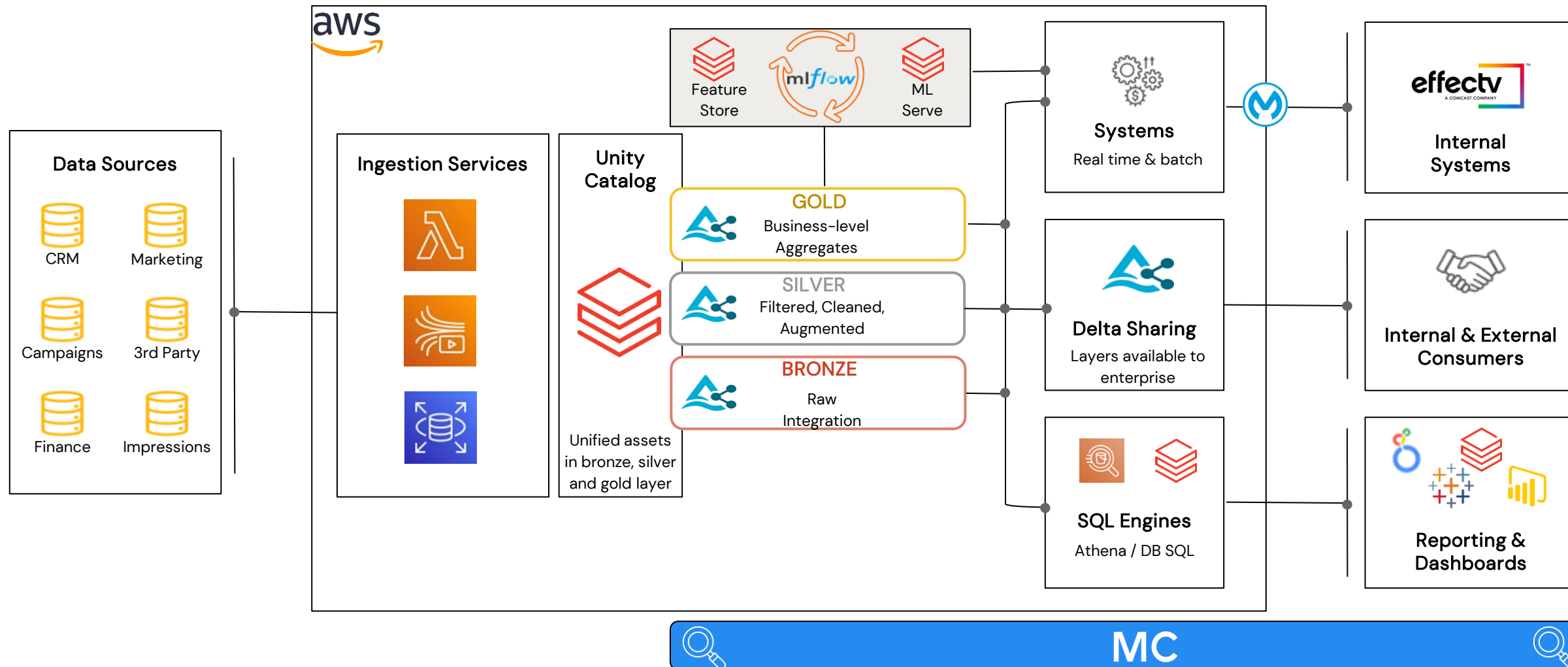
Data Domains to Business Value

Segment data using Monte Carlo **domains**

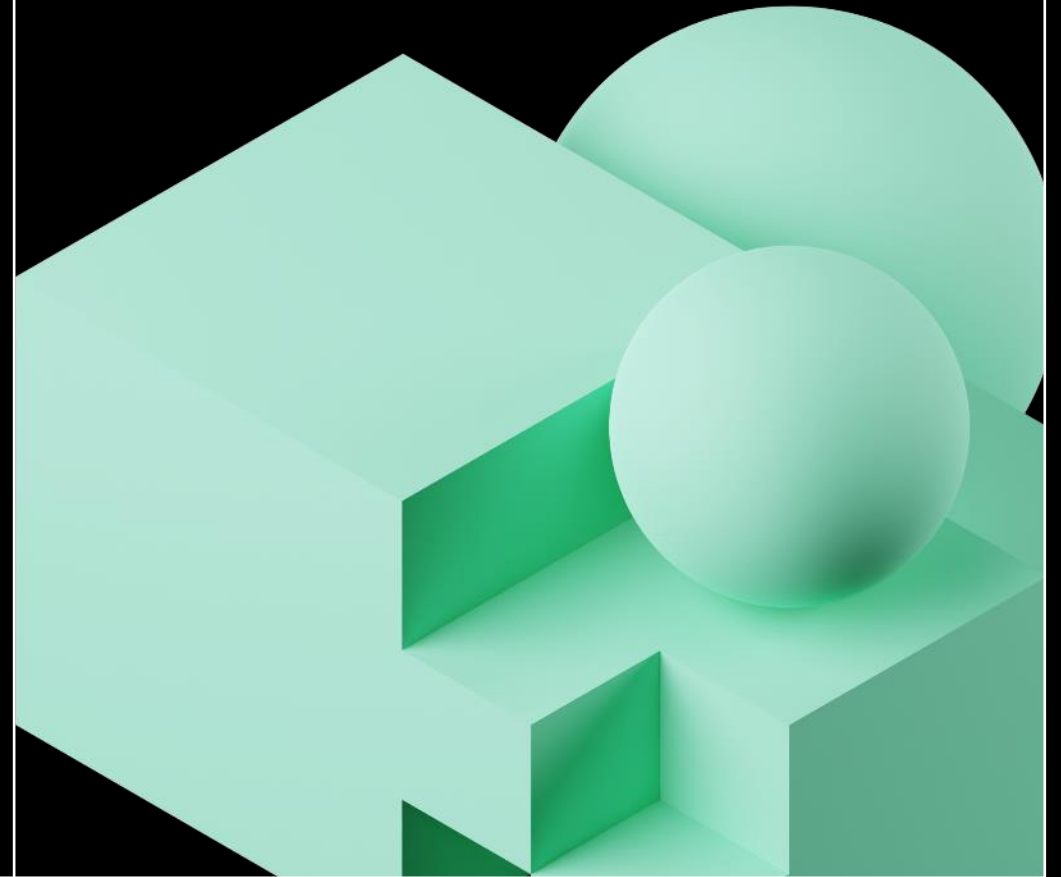


Effectv Data Stack

Enabling an analytics platform with **Databricks**



Integrating Monte Carlo Into Your Data Stack



Setting up Monte Carlo

MC

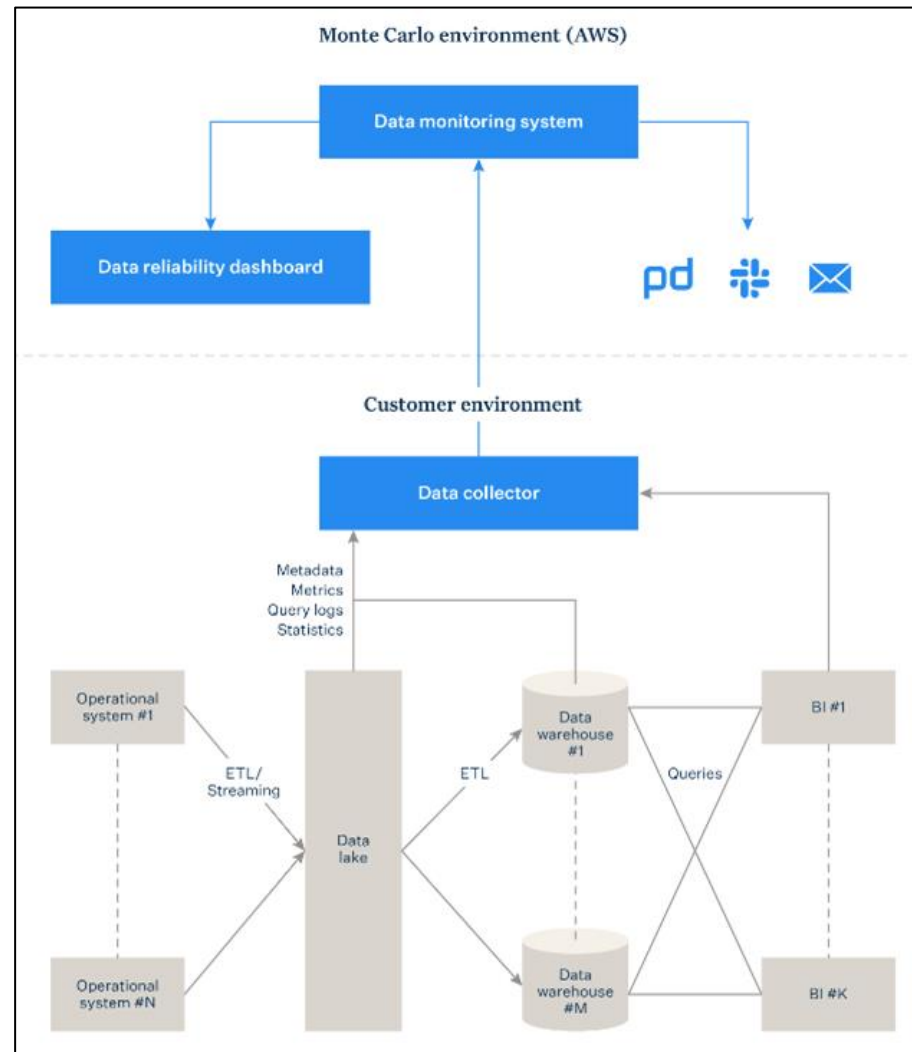
Designed with **security** and **data privacy** top-of-mind

Ease of Implementation

Deploy Data Collector
Connect Sources

Speed to Value

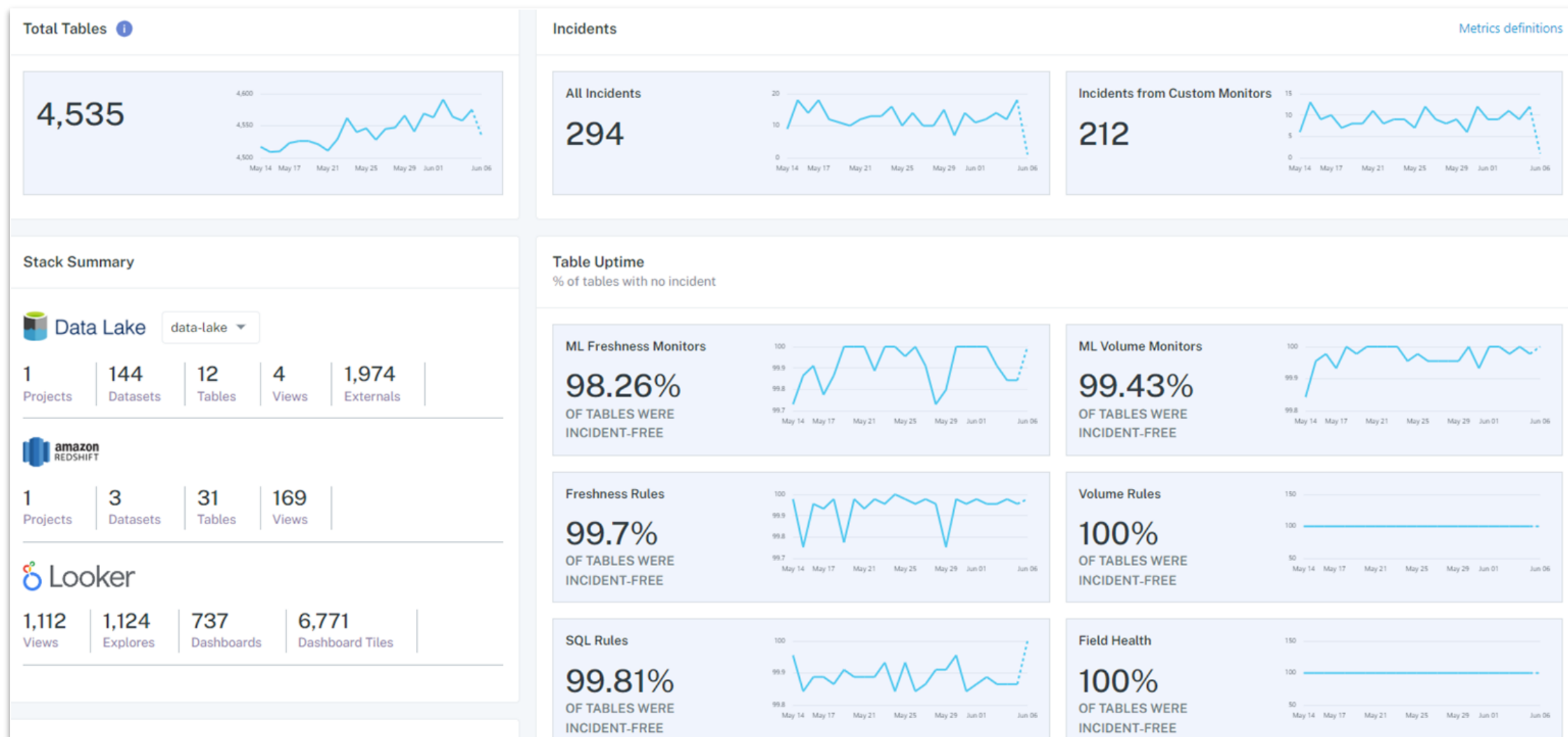
Start Training Models
Detect Incidents



Data Reliability Dashboard

MC

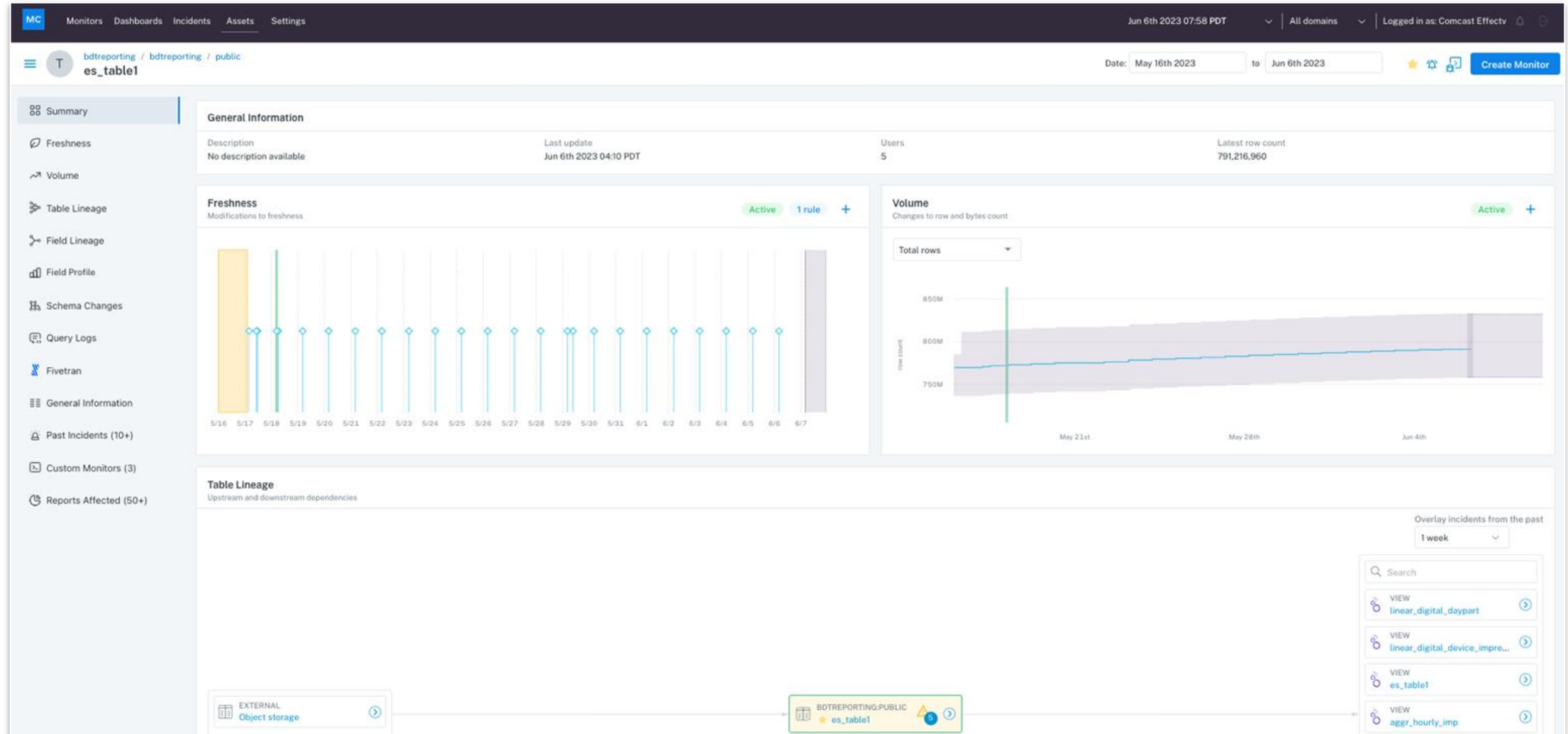
A **single pane of glass** across multiple teams and stacks



Asset Manager

MC

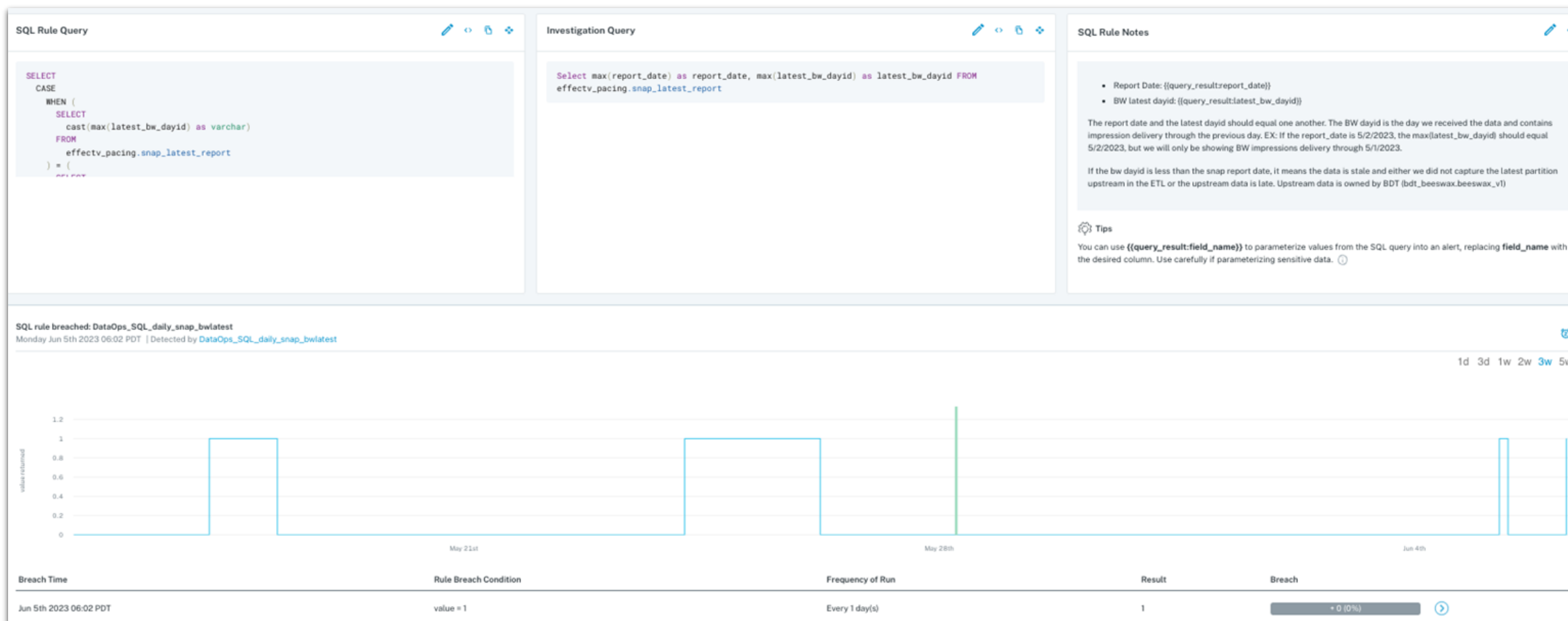
Search and summarize any dataset



Custom Alerting

MC

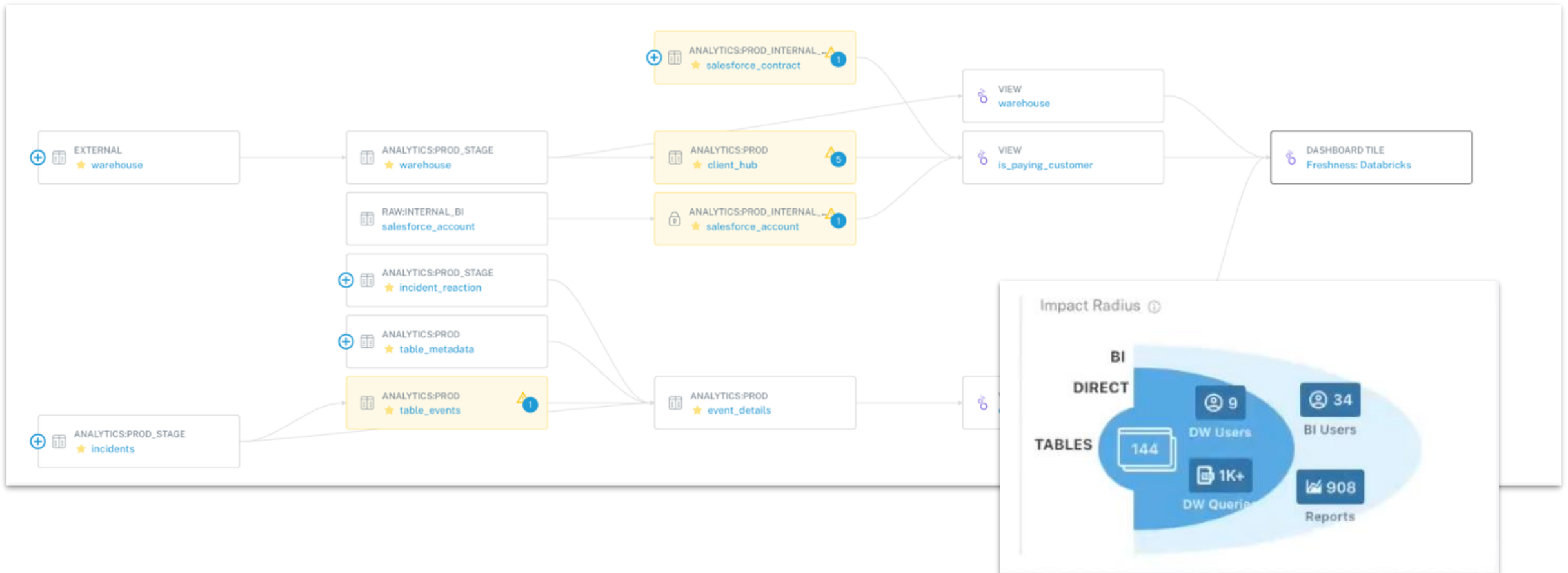
Define specific rules with **Custom SQL Monitors**



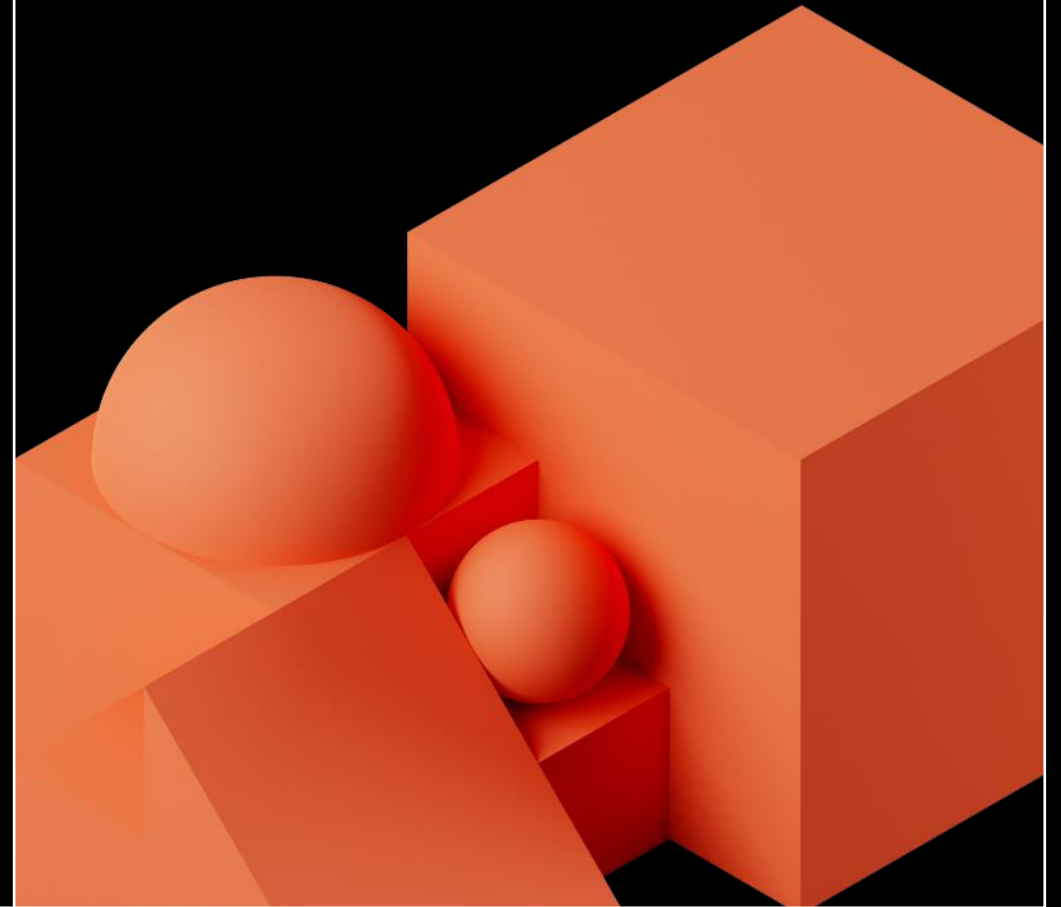
Lineage Impact

MC

See **upstream** and **downstream** for any dataset

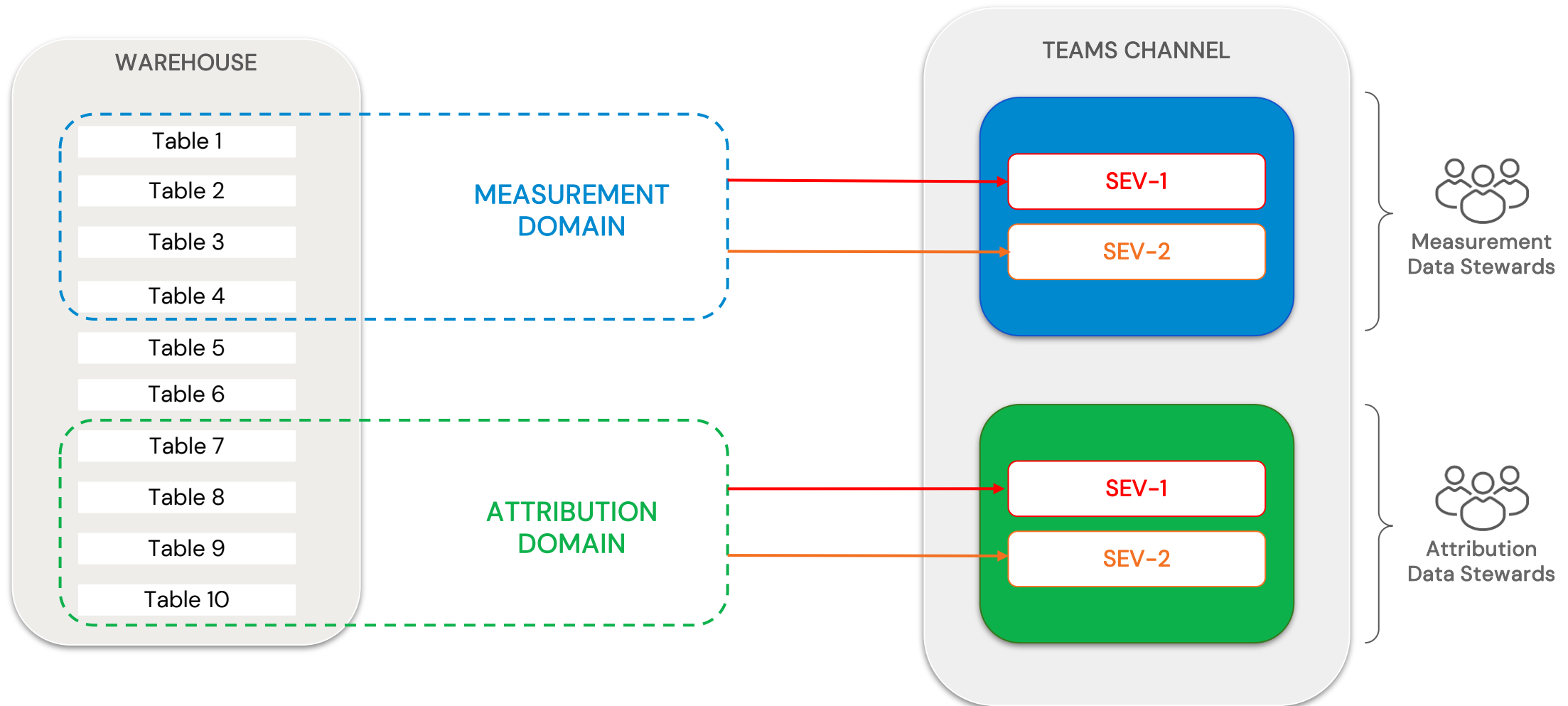


Operationalizing Monte Carlo With Your Business



Domain-oriented Notification Strategy

Stakeholders subscribe to **Microsoft Teams**





Notification Levers

OOTB features to avoid **alert fatigue**



Importance

★ Key Asset	Yes 
Importance score 	0.915
Authors	No recent changes.
Recent schema changes	1

Defining the most important data assets in the business



Muting

[bdt-datalake] awsdatacatalog:adexposure	<input checked="" type="checkbox"/>	^
Schema Change	<input checked="" type="checkbox"/>	
Unchanged Size	<input checked="" type="checkbox"/>	
Freshness	<input checked="" type="checkbox"/>	
Volume	<input checked="" type="checkbox"/>	

Muting data assets that are less relevant or don't require monitoring



Tagging

Search for tables, reports or fields

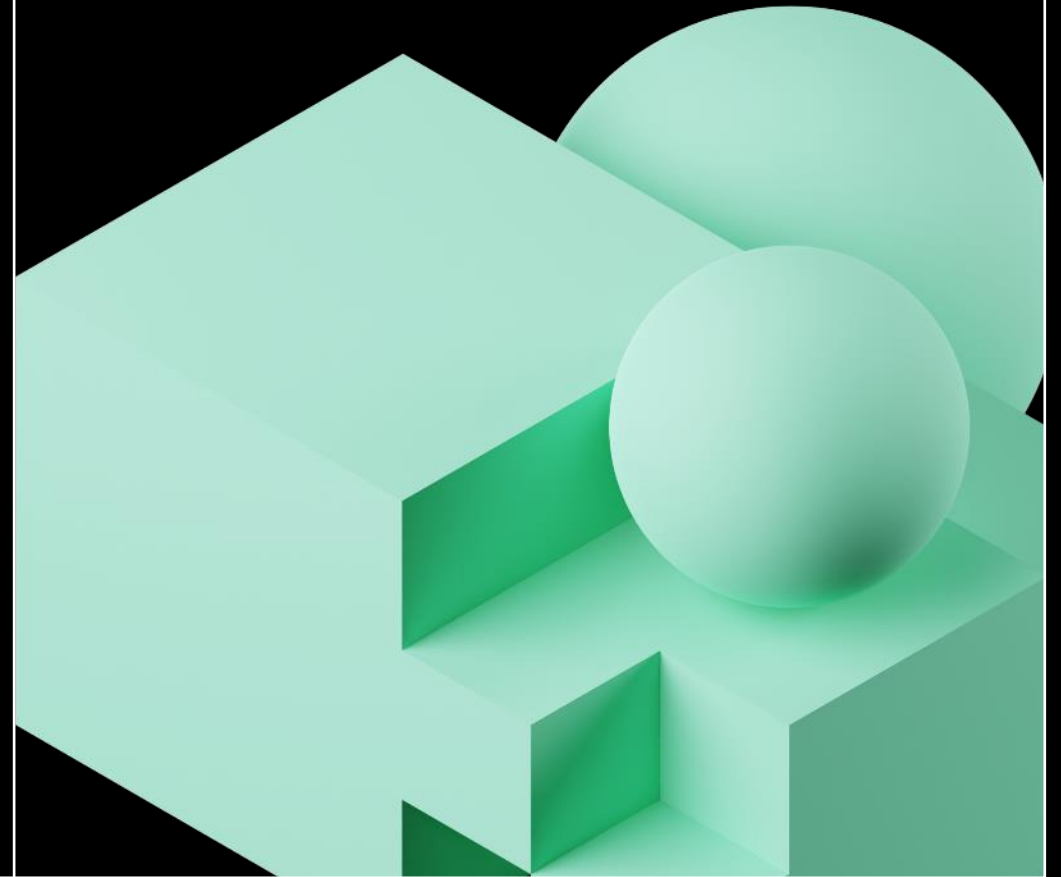
[Search operator help](#)

Integration Domain

Owner Project

Tagging data assets to manage, identify, and organize our instance

Monte Carlo in Action at Effectv



Campaign Pacing Performance

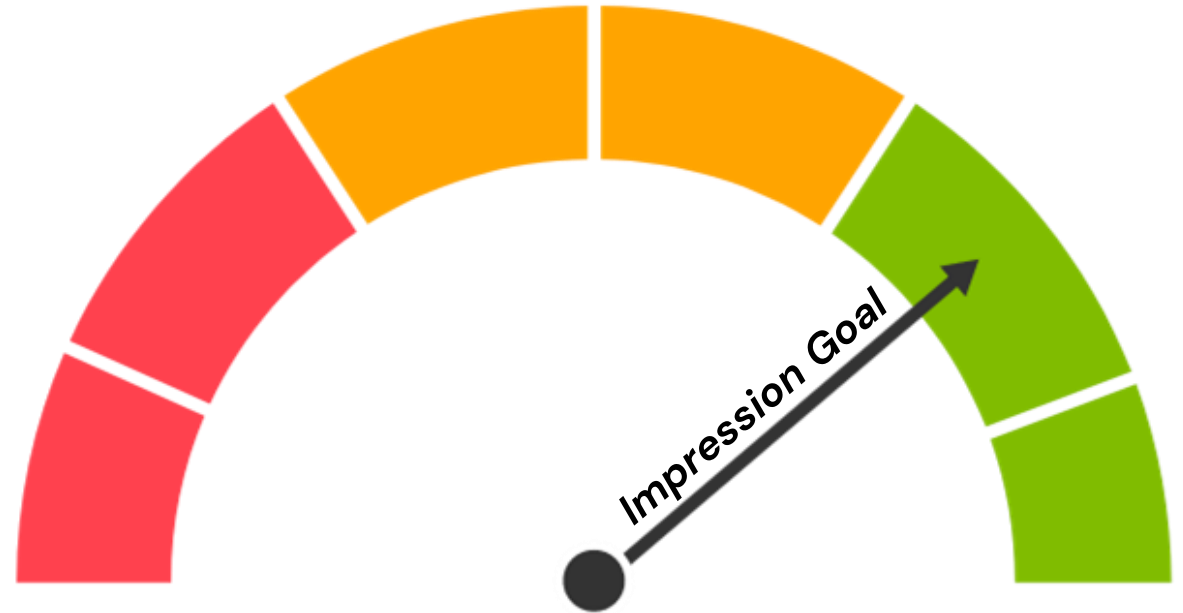
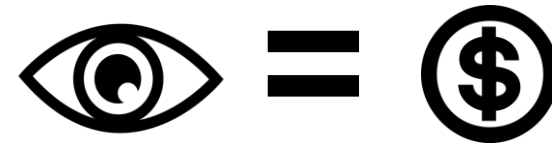
Effectv Streaming makes good on **impression** delivery



AD IMPRESSIONS

Metric used to quantify the number of times the digital video ad was viewed.

The impression count is measured based on the response from a video ad delivery system to an ad request from the digital video content host.



Campaign Pacing Performance

Daily impression aggregation for all active campaigns with **Databricks**



Campaigns

K

Thousands of active
campaigns monitored daily



Impressions

B

Billions of impression
events targeted



Revenue

M

Millions in gross
contracted revenue



Consumers

K

Thousands of advertisers,
account executives, and
campaign managers

Campaign Pacing Performance

Daily impression aggregation for all active campaigns with **Databricks**



Campaigns

K

Thousands of active campaigns monitored daily



Impressions

B

Billions of impression events targeted



Revenue

M

Millions in gross contracted revenue



Consumers

K

Thousands of advertisers, account executives, and campaign managers

Campaign Pacing Performance

Daily impression aggregation for all active campaigns with **Databricks**



Campaigns

K

Thousands of active campaigns monitored daily



Impressions

B

Billions of impression events targeted



Revenue

M

Millions in gross contracted revenue



Consumers

K

Thousands of advertisers, account executives, and campaign managers

Campaign Pacing Performance

Daily impression aggregation for all active campaigns with **Databricks**



Campaigns

K

Thousands of active
campaigns monitored daily



Impressions

B

Billions of impression
events targeted



Revenue

M

Millions in gross
contracted revenue



Consumers

K

Thousands of advertisers,
account executives, and
campaign managers

Campaign Pacing Performance

Daily impression aggregation for all active campaigns with **Databricks**



Campaigns

K

Thousands of active campaigns monitored daily



Impressions

B

Billions of impression events targeted



Revenue

M

Millions in gross contracted revenue



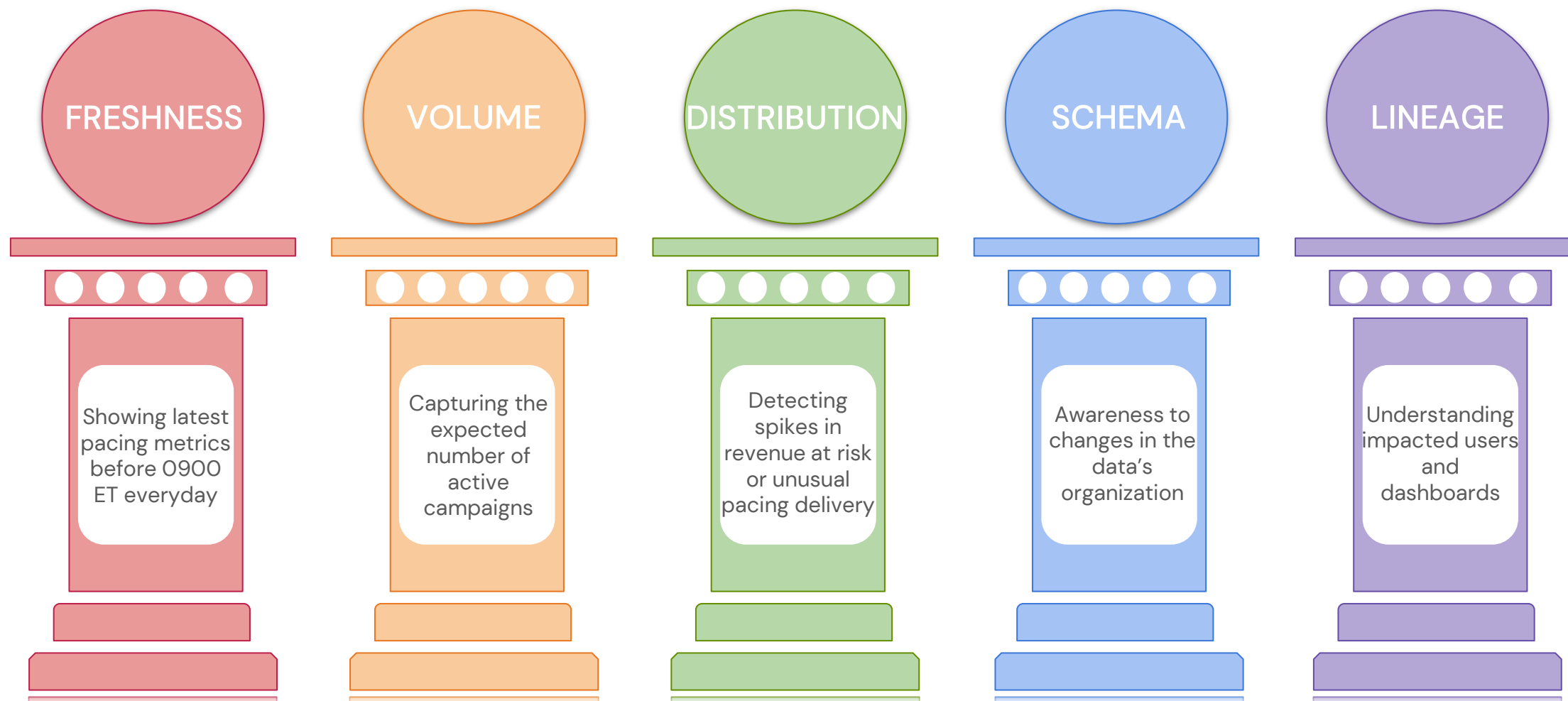
Consumers

K

Thousands of advertisers, account executives, and campaign managers

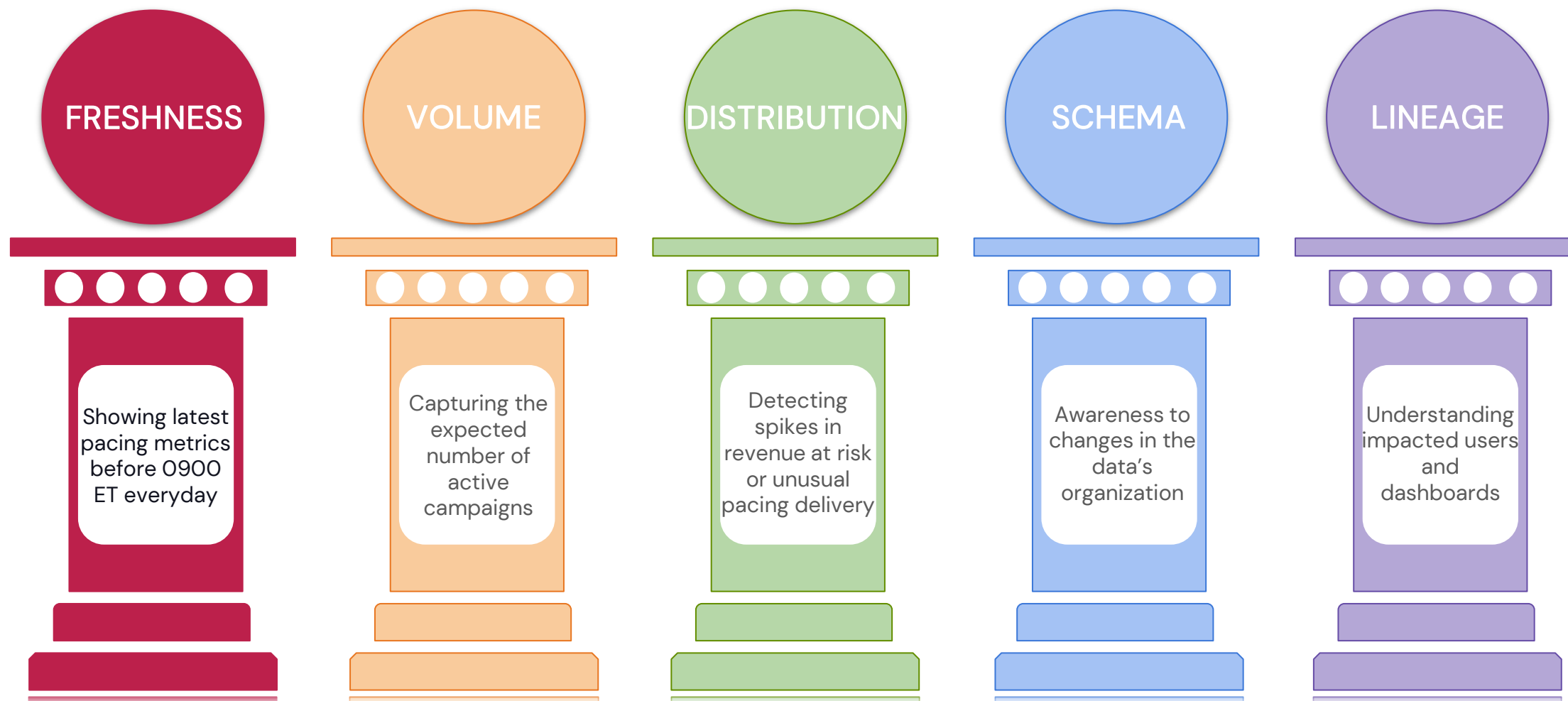
The 5 Pillars with Monte Carlo

No longer manually collecting the **answers** behind data downtime



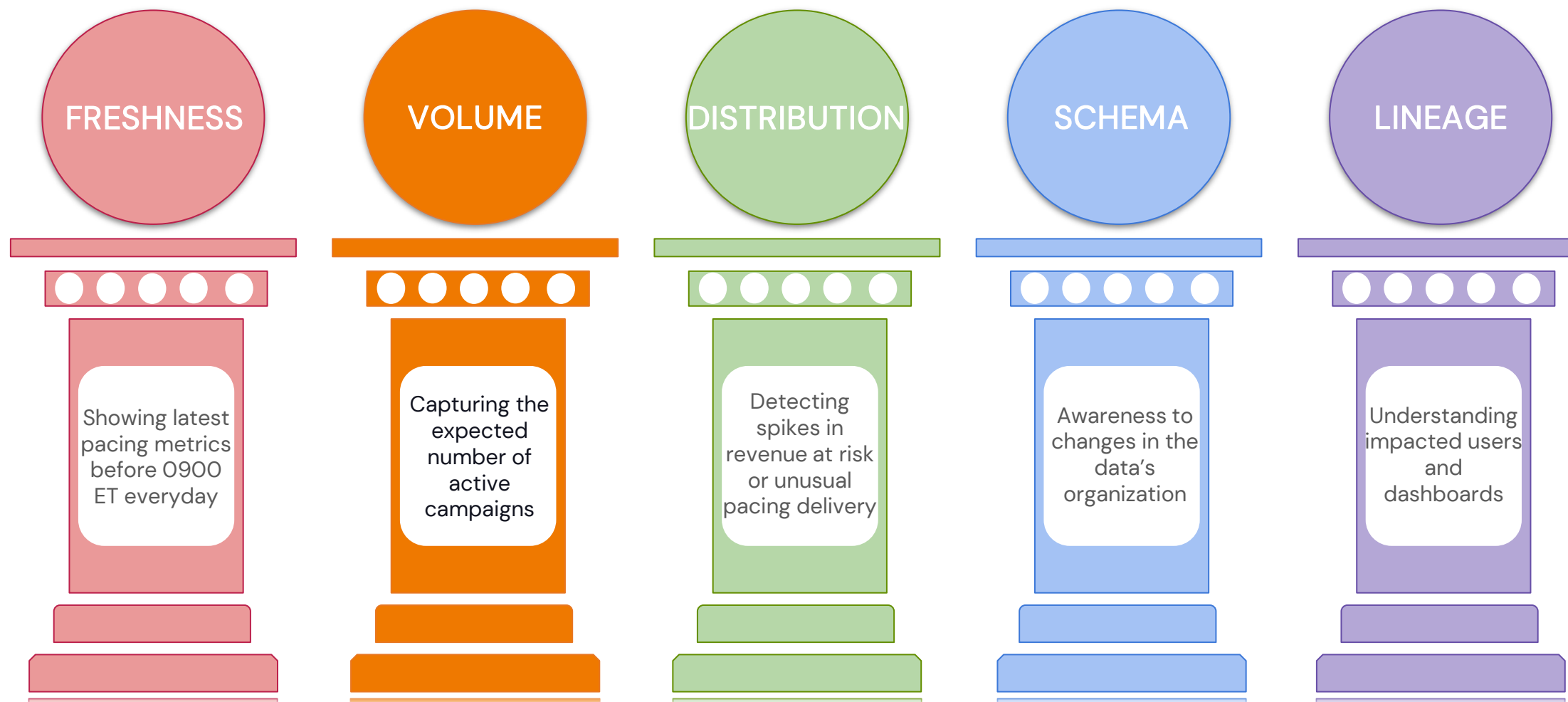
The 5 Pillars with Monte Carlo

No longer manually collecting the **answers** behind data downtime



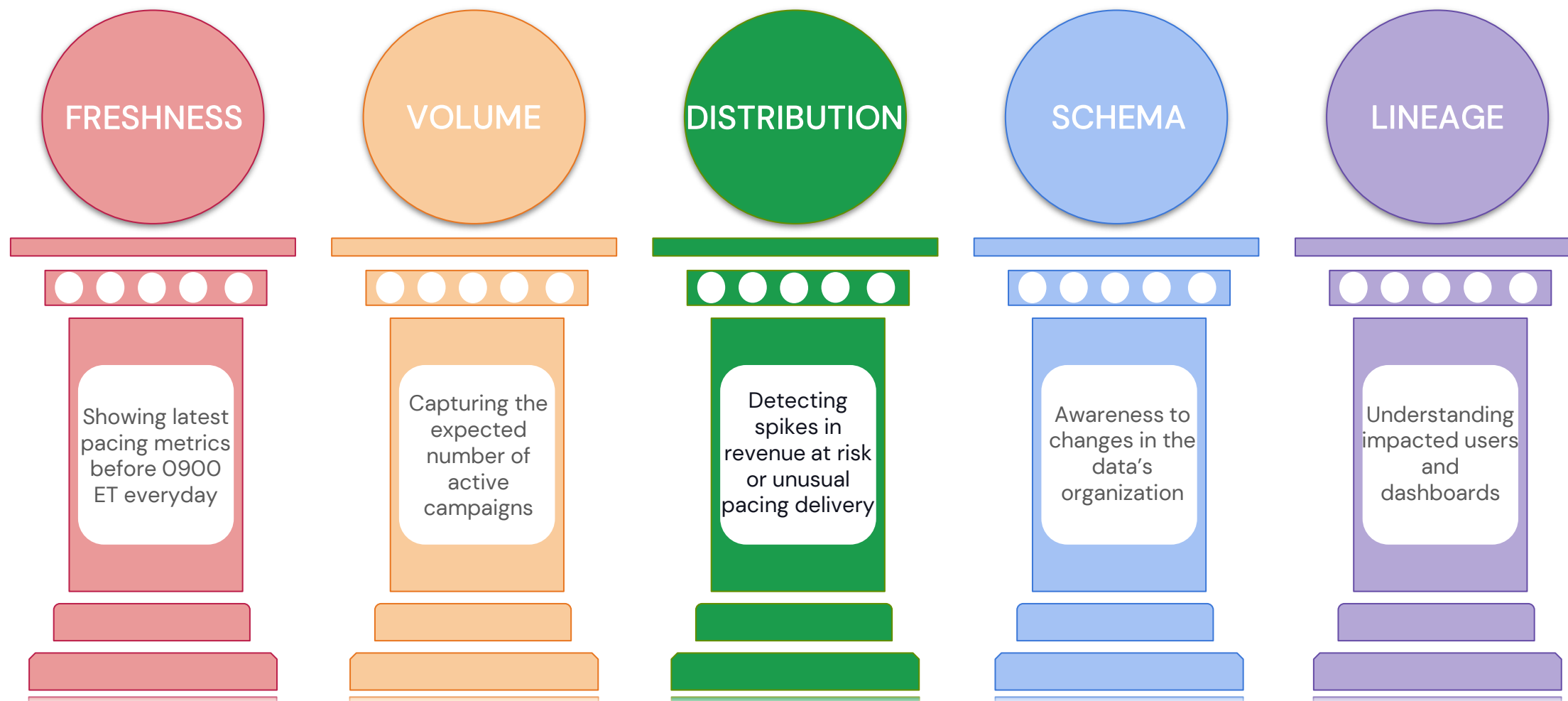
The 5 Pillars with Monte Carlo

No longer manually collecting the **answers** behind data downtime



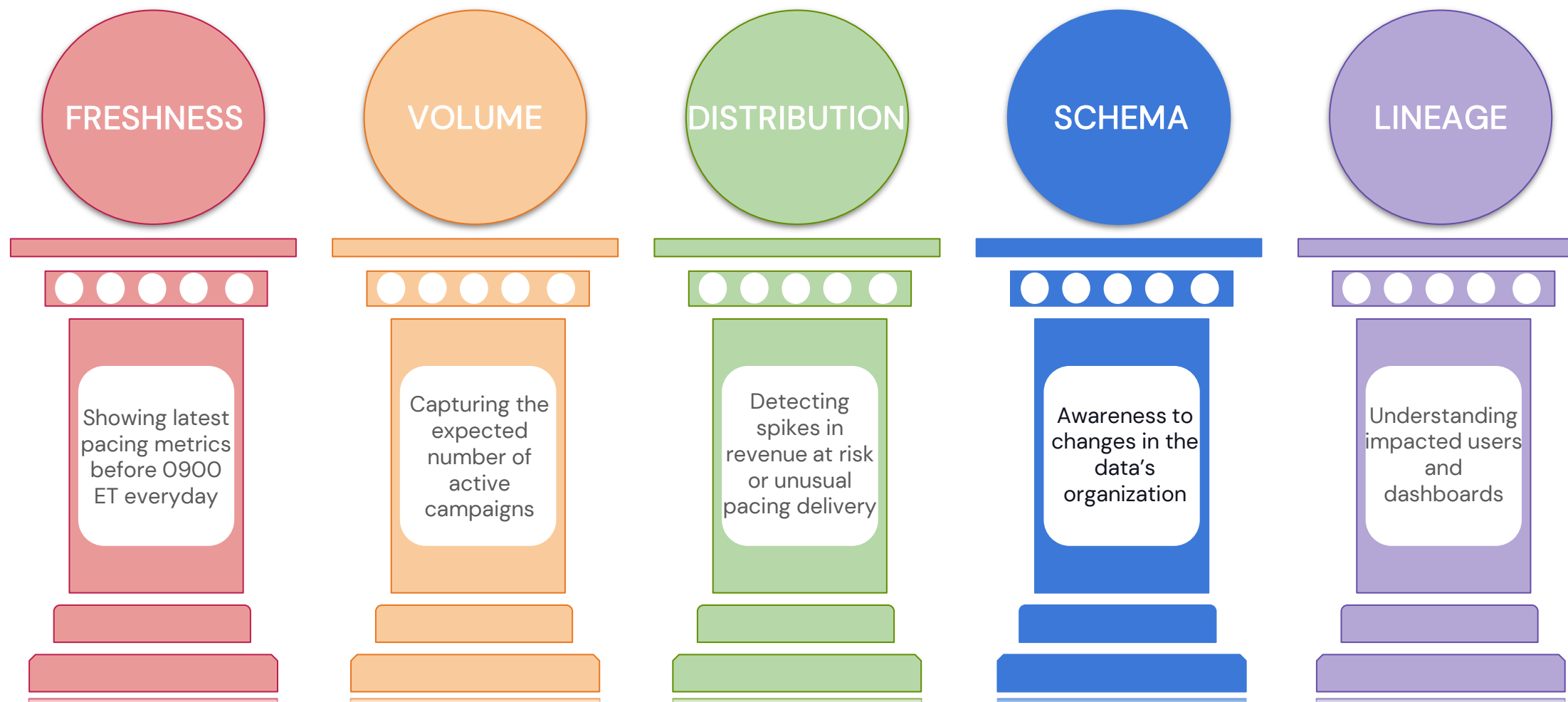
The 5 Pillars with Monte Carlo

No longer manually collecting the **answers** behind data downtime



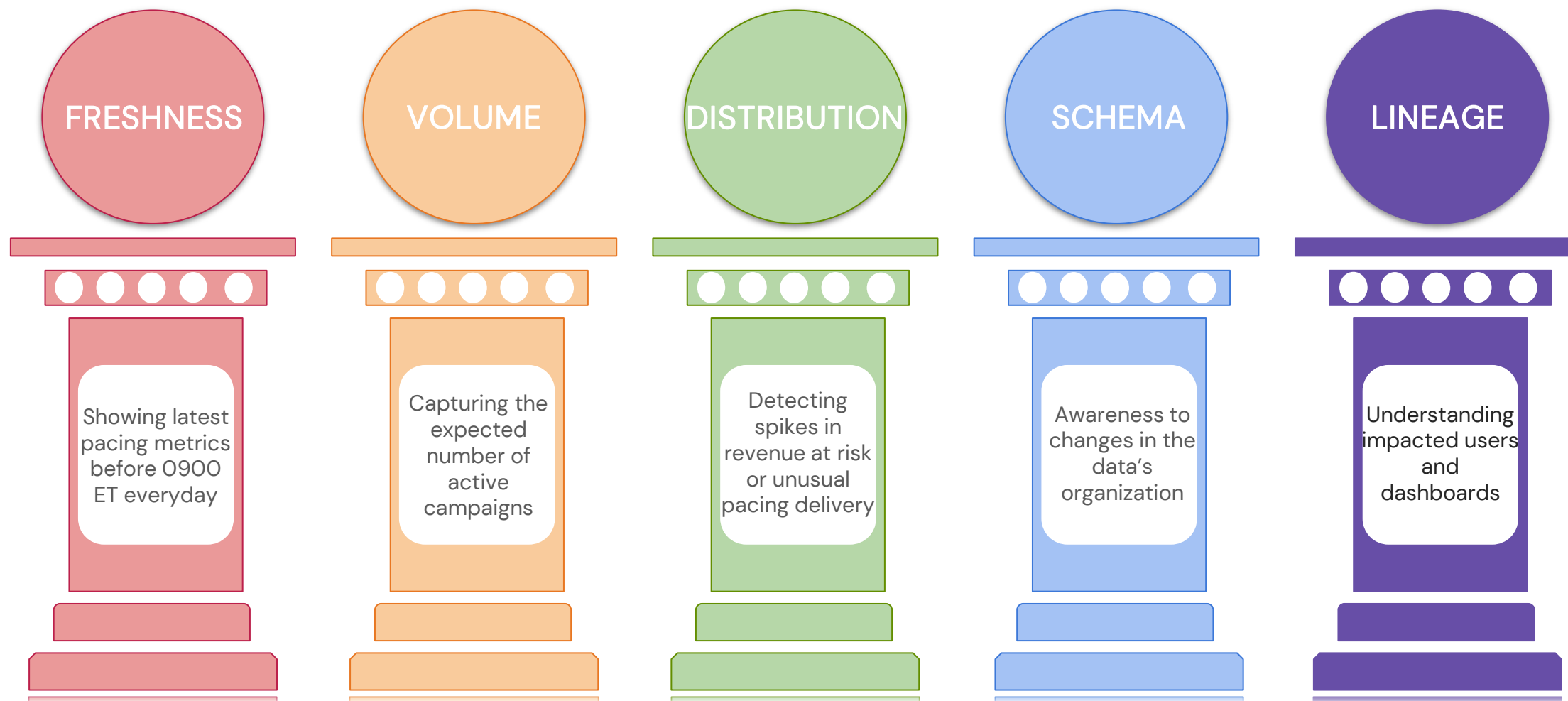
The 5 Pillars with Monte Carlo

No longer manually collecting the **answers** behind data downtime



The 5 Pillars with Monte Carlo

No longer manually collecting the **answers** behind data downtime



Scaling Data Trust with Monte Carlo

Overwhelmingly **positive** response

“

Having up-to-date data is critical to ensuring we can quickly respond to changes in campaign pacing and can adjust our delivery strategy to recapture revenue at risk.

Juliette Jones

Sr. Director, Media Operations

”

Building trust in timely data and decisions being made of it!

Neeraj Grover

Sr. Director, DataOps & MLOps

“

A task that in the past could have taken teams a week to do, Monte Carlo gives us the answers in minutes!

Heather Blythe

Director, Enterprise Data Analytics

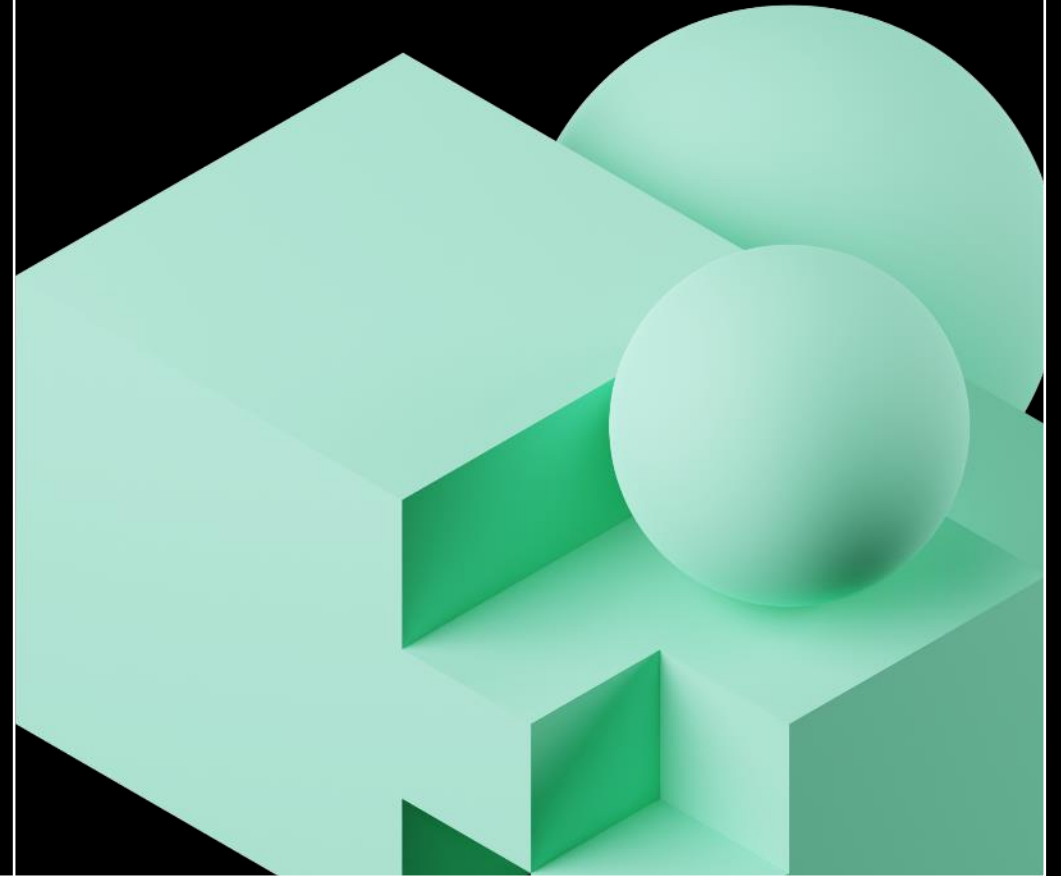
”

Empowering ops teams with the latest insights to make data-driven optimizations for high-risk campaigns.

Erica Fellin

Sr. Manager, Product Management

What's Next



ServiceNow integration, Circuit Breakers

Example

```
from datetime import datetime, timedelta

from airflow import DAG

try:
    from airflow.operators.bash import BashOperator
except ImportError:
    # For airflow versions <= 2.0.0. This module was deprecated in 2.0.0.
    from airflow.operators.bash_operator import BashOperator

from airflow_mcd.operators import SimpleCircuitBreakerOperator

mcd_connection_id = 'mcd_default_session'

with DAG('sample-dag', start_date=datetime(2022, 2, 8), catchup=False, schedule_interval=timedelta(1))
    task1 = BashOperator(
        task_id='example_elt_job_1',
        bash_command='echo I am transforming a very important table!',
    )
    breaker = SimpleCircuitBreakerOperator(
        task_id='example_circuit_breaker',
        mcd_session_conn_id=mcd_connection_id,
        rule_uuid='<RULE_UUID>'
    )
    task2 = BashOperator(
        task_id='example_elt_job_2',
        bash_command='echo I am building a very important dashboard from the table created in task1!',
        trigger_rule='none_failed'
    )

    task1 >> breaker >> task2
```

Monitors as Code, Leverage API

Monitor YAML

For use with [Monitors as code](#)

Access monitor YAML [via API](#)

```
custom_sql:
- description: Test DS SQL
  schedule:
    type: fixed
    interval_minutes: 1440
    start_time: '2023-03-03T13:00:00.656000+00:00'
    timezone: Africa/Bangui
  sql: asdasd
  comparisons:
  - type: threshold
    operator: GT
    threshold_value: 0.0
```

API Explorer

Explore the Monte Carlo GraphQL APIs using the client below. [API Documentation](#)

```
1 # Explore the Monte Carlo API!
2 #
3 # API Documentation can be found here: https://apidocs.getmontecarlo.com/
4 #
5 # If you want to run these queries outside of the dashboard (e.g. via curl),
6 # you can follow these docs to generate and use an API token:
7 # https://docs.getmontecarlo.com/docs/creating-an-api-token
8 #
9 # These queries can also be run using our Python SDK (with first class
10 # objects, pythonic snake_case, automatic retries, and more):
11 # https://pypi.org/project/pycarlo/
12 #
13 # For reference here is an example query to get user information.
14 # Press the play button to try it out:
15
16 query getUser {
17   getUser {
18     email
19     firstName
20     lastName
21     createdOn
22     role
23     account {
24       uuid
25     }
26   }
27 }
28
```

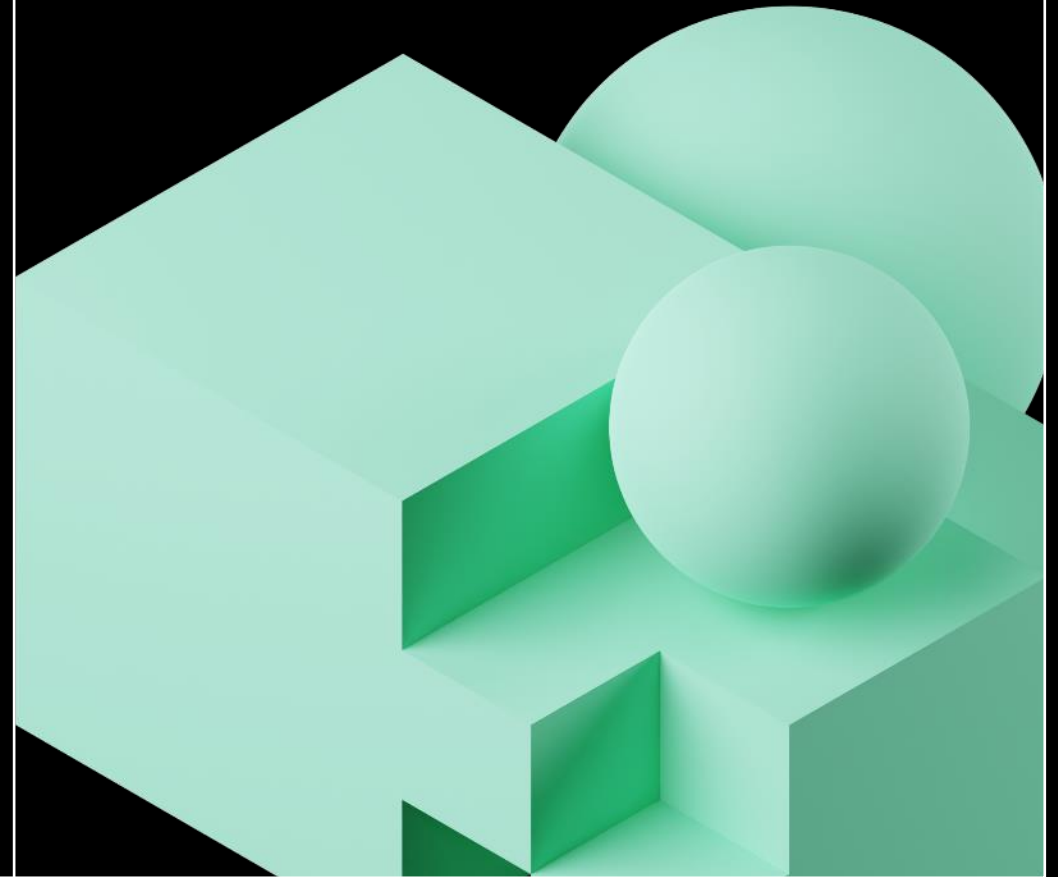

Want to know more?

**Enabling a Data Governance
Program on Databricks Using Unity
Catalog and ML-Driven MDM**

June 28th | 12:30 PDT | Hall D Room 10

**Ad Measurement: From
Impressions to Attribution**

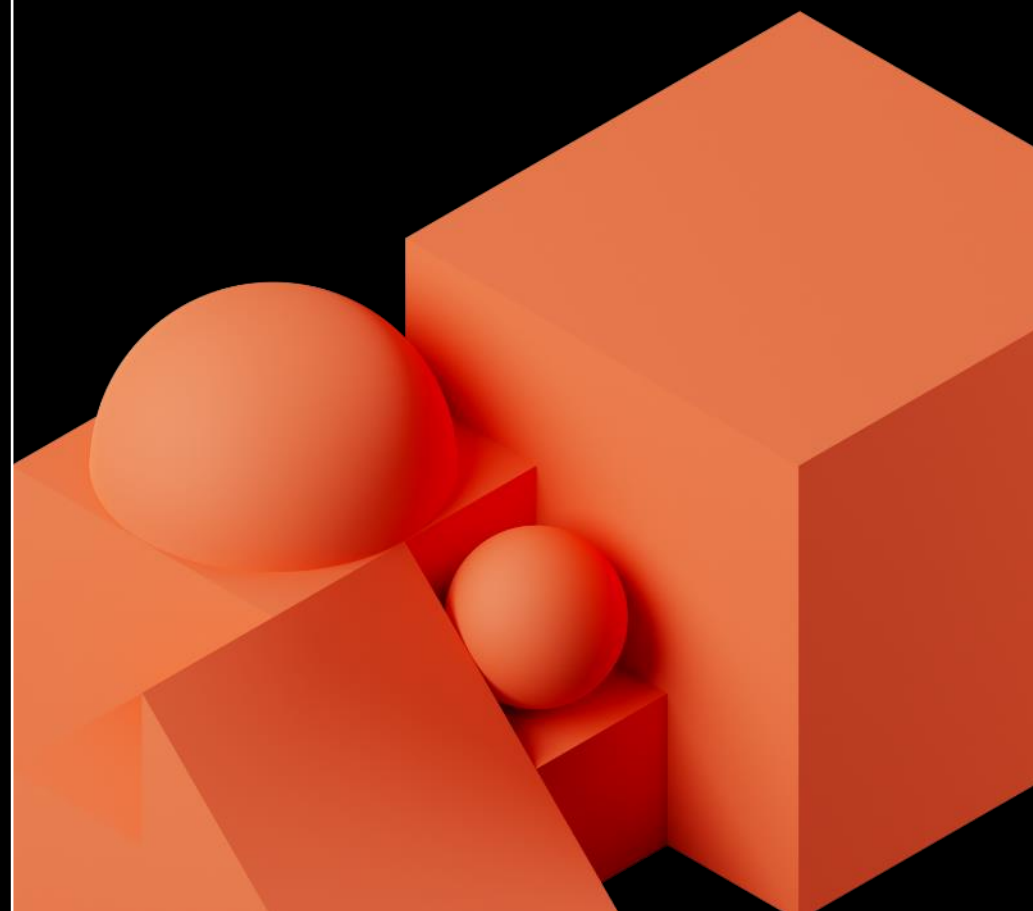
June 29th | 12:30 PDT | Hall D Room 4



Thanks

Learn more about data observability for the lakehouse:
<https://www.montecarlodata.com/meet-monte-carlo-at-databricks-data-ai-summit/>

Visit us at Monte Carlo's booth:
#407



Any Questions?



