



# Intelligent Data Engineering

Slide Deck



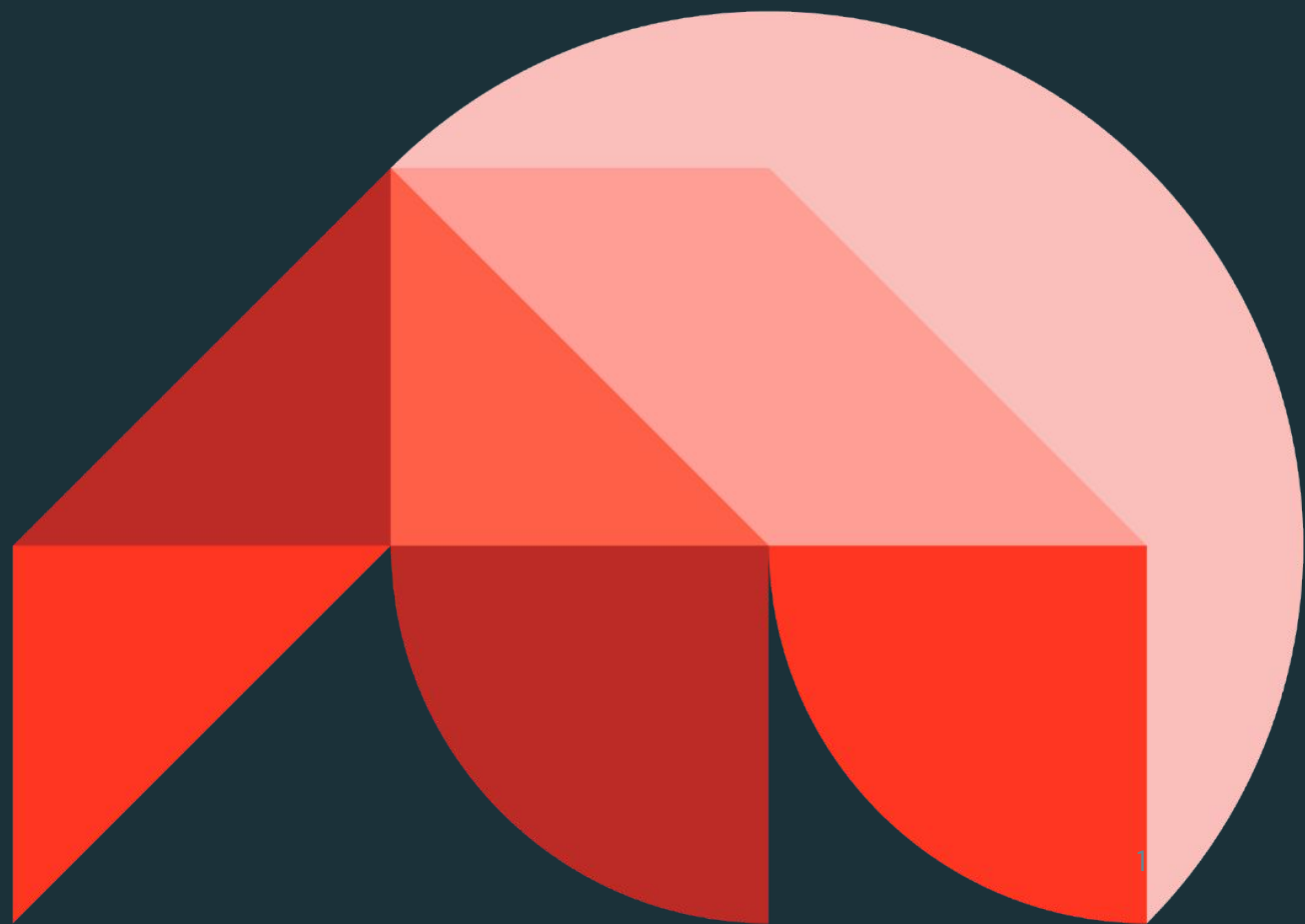
carahsoft.

For more information, contact Carahsoft or our reseller partners:

[Databricks@carahsoft.com](mailto:Databricks@carahsoft.com) | 703-581-6693



# Intelligent Data Engineering



A hand is shown in the foreground, interacting with a futuristic digital interface. The interface consists of several floating data cards, each representing a different work item. The cards are arranged in a grid-like pattern and feature various data visualizations such as progress bars, circular gauges, and status indicators. The background is a dark, blue-toned environment with glowing particles, suggesting a high-tech or data-driven setting. The overall aesthetic is clean and modern, with a focus on data visualization and user interaction.

Every AI innovation hinges  
on reliable data



## AI initiatives are top of mind...

By 2026, **over 80%** of enterprises will be using GenAI in production environments, up from **less than 5%** in 2023

—2023 Gartner Hype Cycle  
for Generative AI

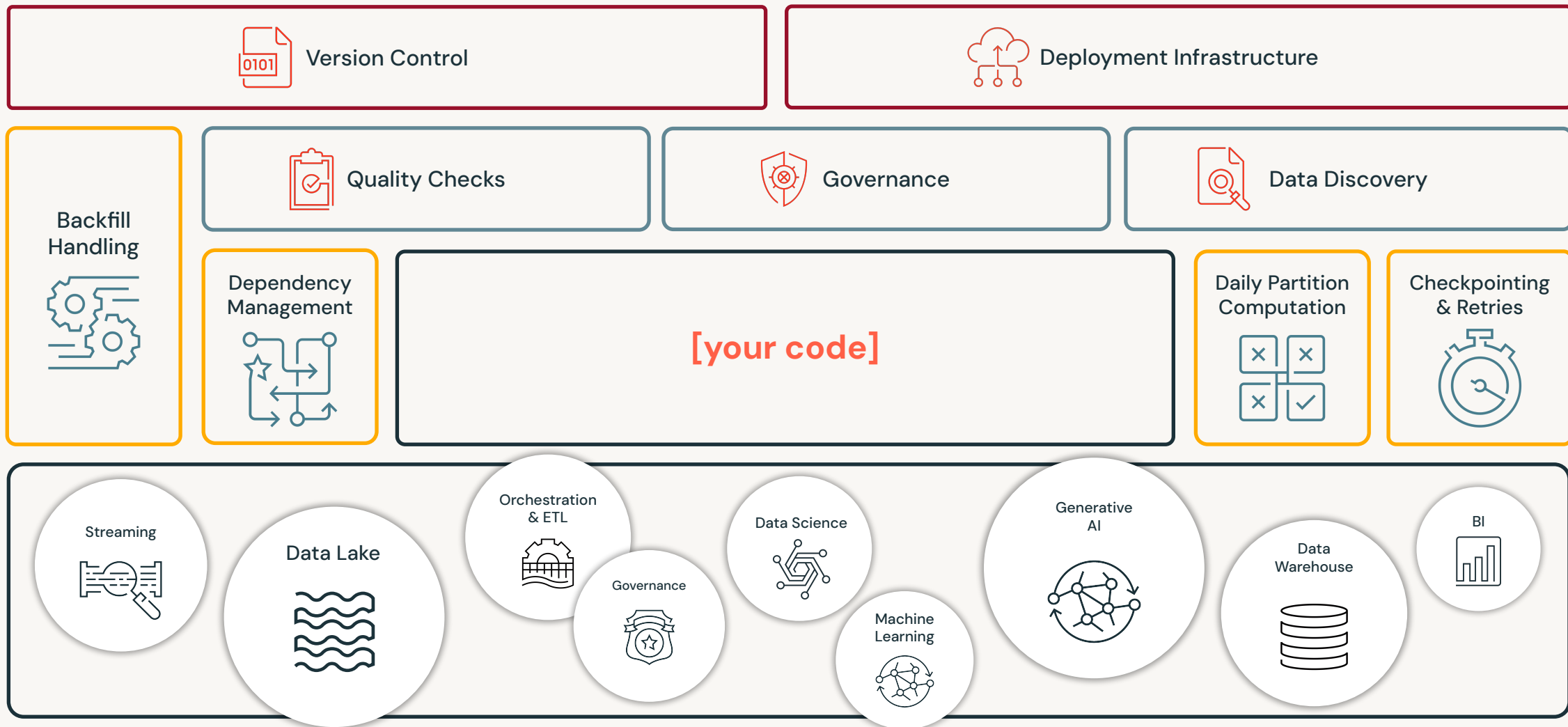


## ...but good models can't overcome bad data

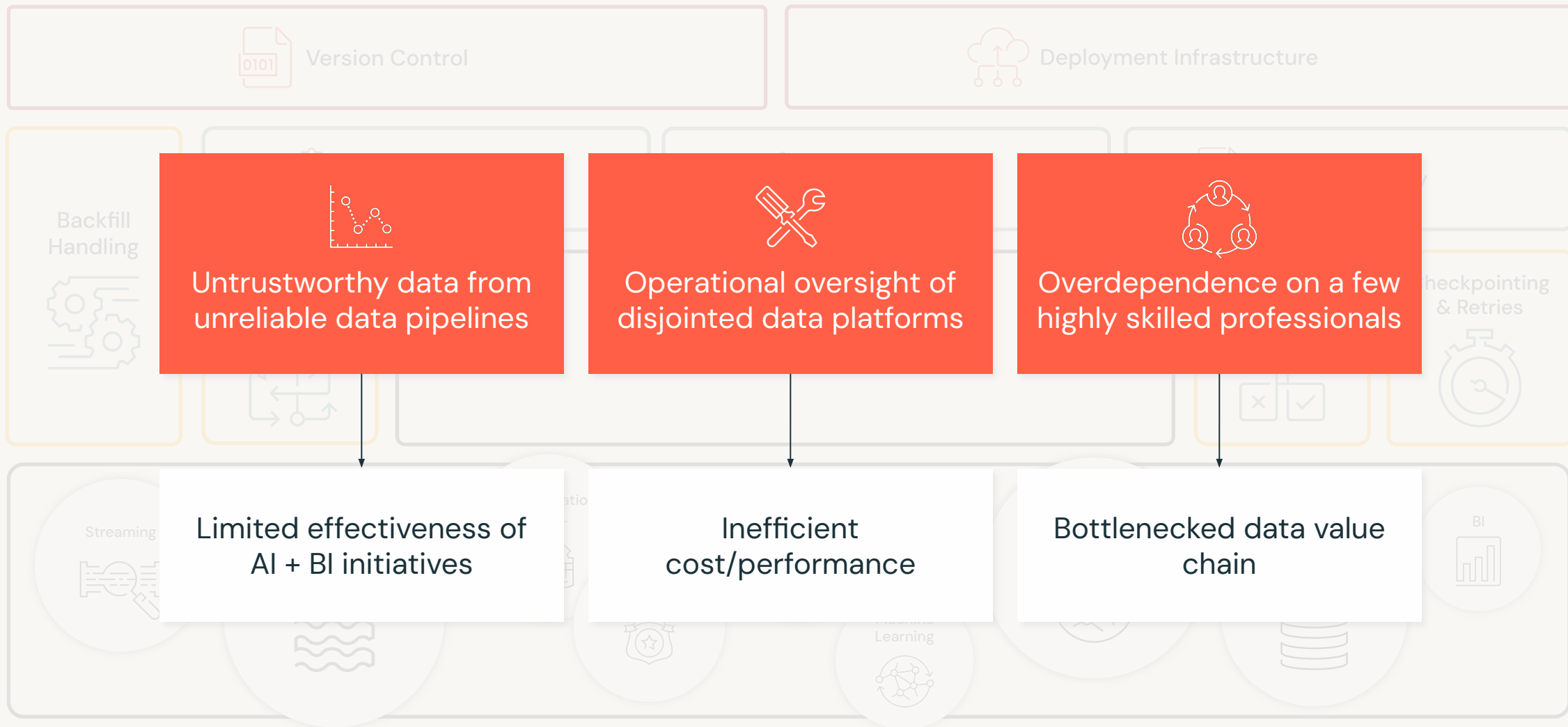
“Data problems are  
**the most likely factor**  
to jeopardize our  
AI/ML goals”

—MIT Technology Review  
Insights survey, 2022

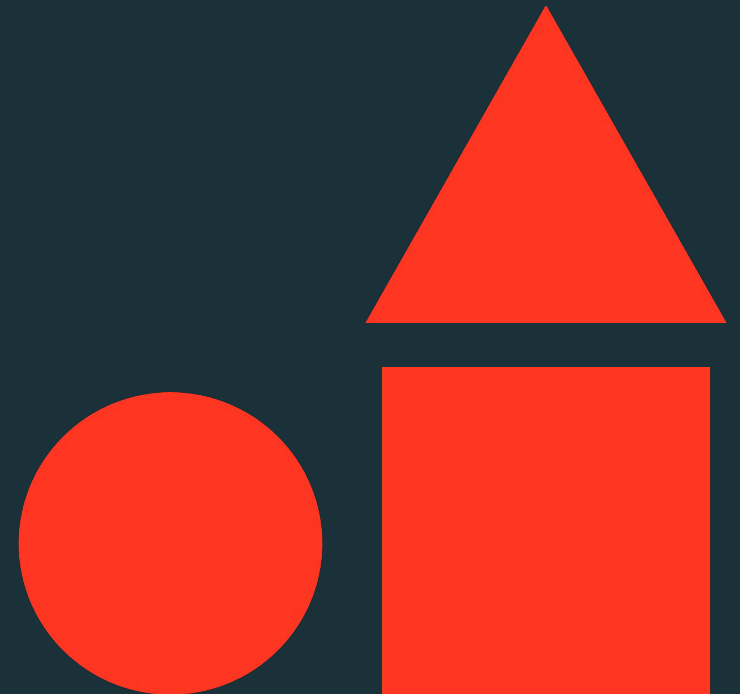
# It's difficult to build and operate reliable data pipelines



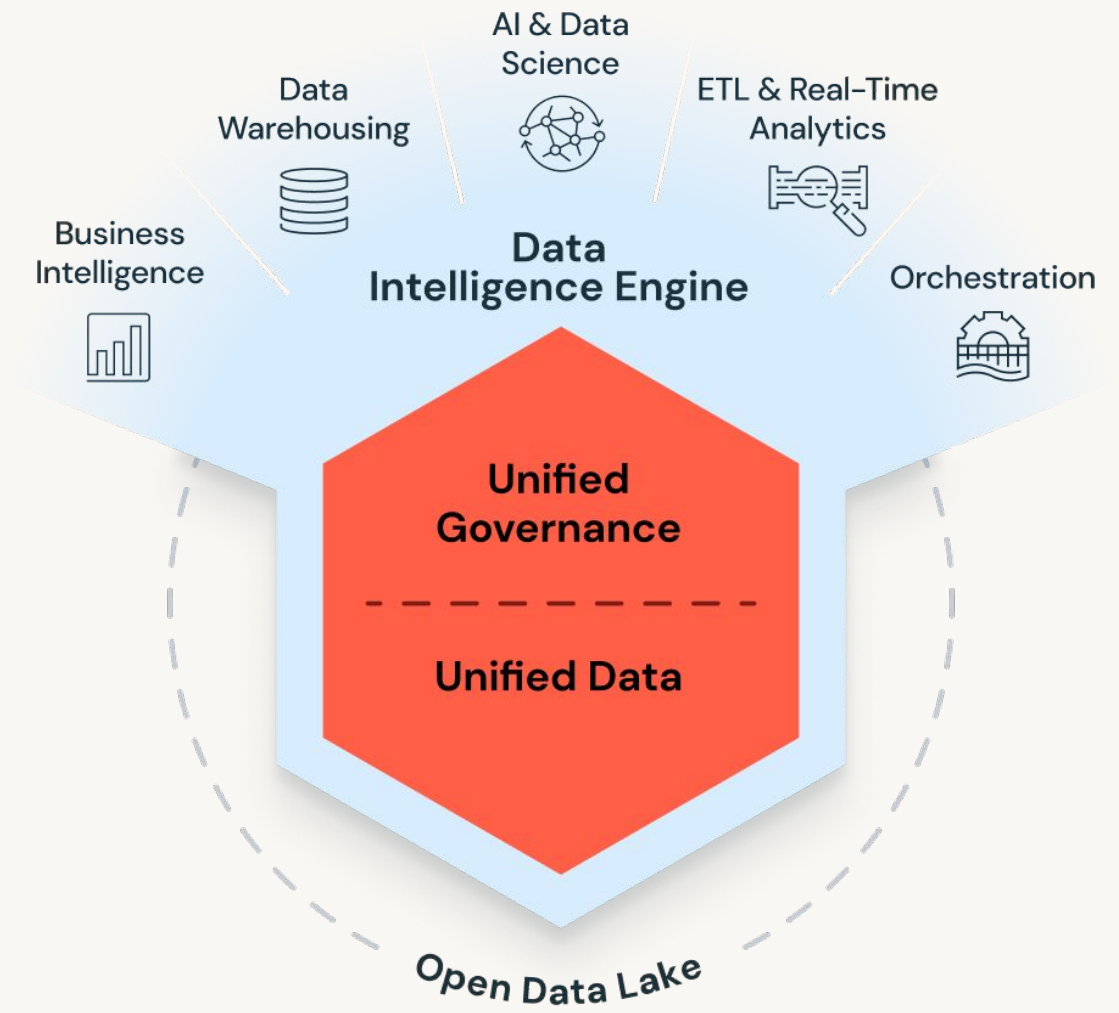
# It's difficult to build and operate reliable data pipelines



Reliable data  
pipelines require a  
unified platform with  
data intelligence

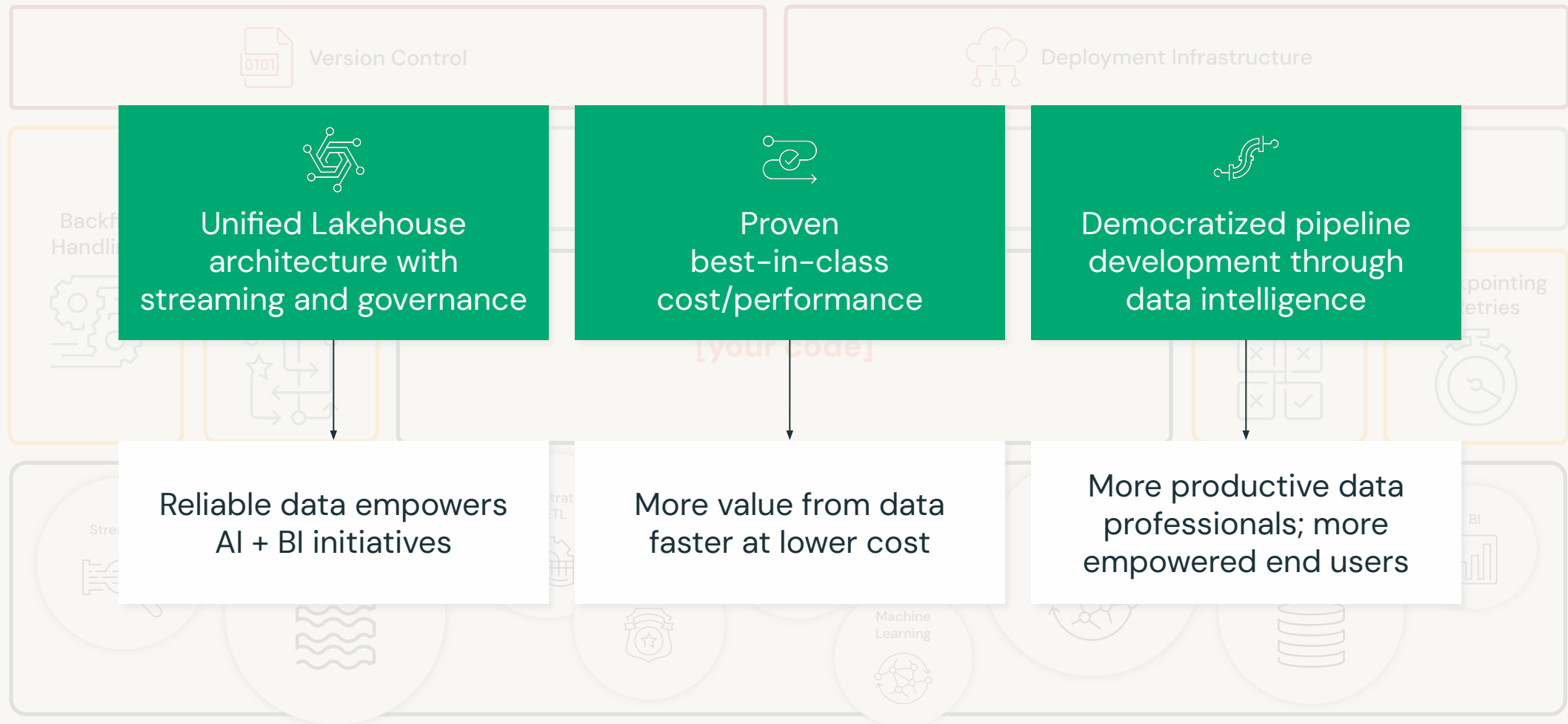


# The Databricks Data Intelligence Platform provides the foundation for Data Engineering in the age of AI



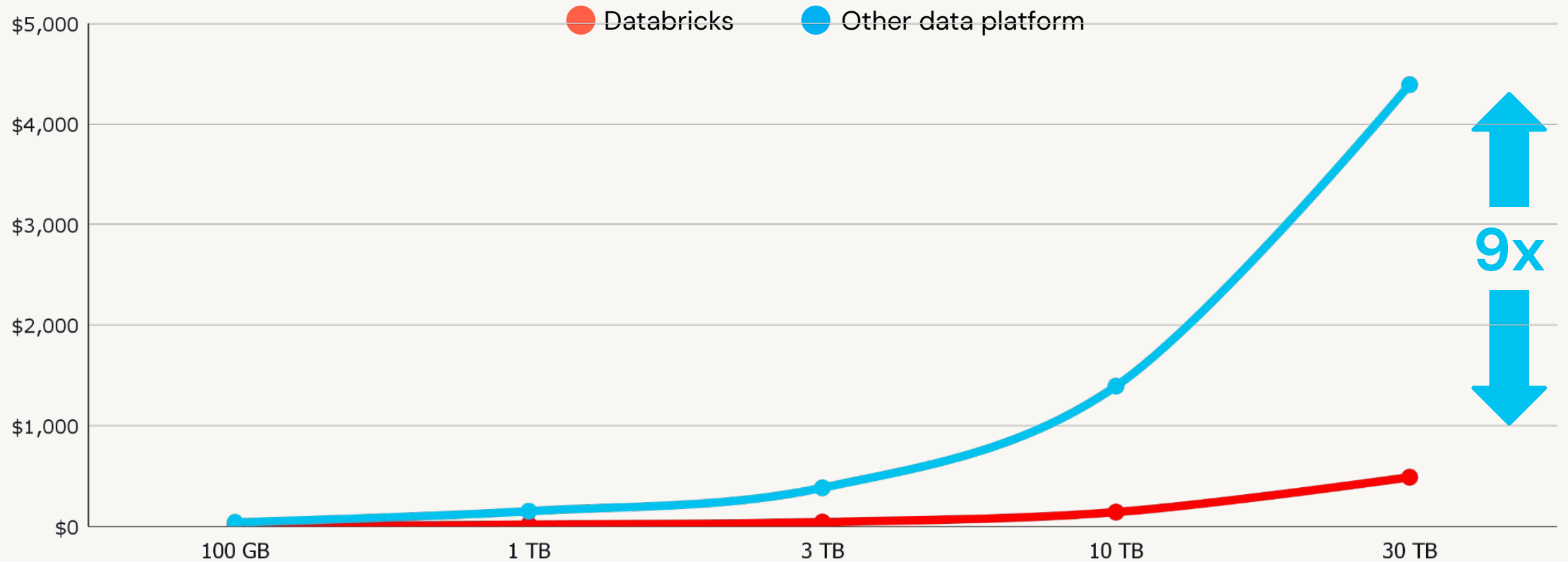


# The Databricks Data Intelligence Platform is the best place to do Data Engineering



# Leading price/performance for ETL

ETL Benchmark shows that Databricks is much more efficient when compared to other data platforms especially as you scale



# Trusted by organizations of all types

6,500+ Data Engineering customers across industries

Walgreens

HSBC

Adobe

Akamai

AKTIFY

AMGEN

Intuit

AstraZeneca

AT&T

ATLASSIAN

jetBlue

Barilla

ESTÉE LAUDER

yipitDATA

BUTCHER  
B—O—X

COMPASS

COMCAST

CRED

DEVSISTERS

CareSource

grammarly

CBC Radio-Canada

ExxonMobil

Johnson & Johnson

CONDÉ NAST

Husqvarna

LALIGA

VIZIO

GSK

SEGA

T Mobile

Columbia

Ahold  
Delhaize

edmunds

RIVIAN

TEXAS  
RANGERS

SNCF

hp

DELL

WB

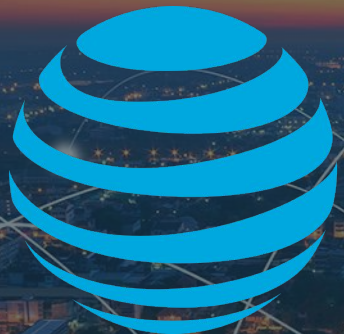
Shell

BAYER

bp

wejo





# AT&T

**Kate Hopkins**

Vice President  
AT&T

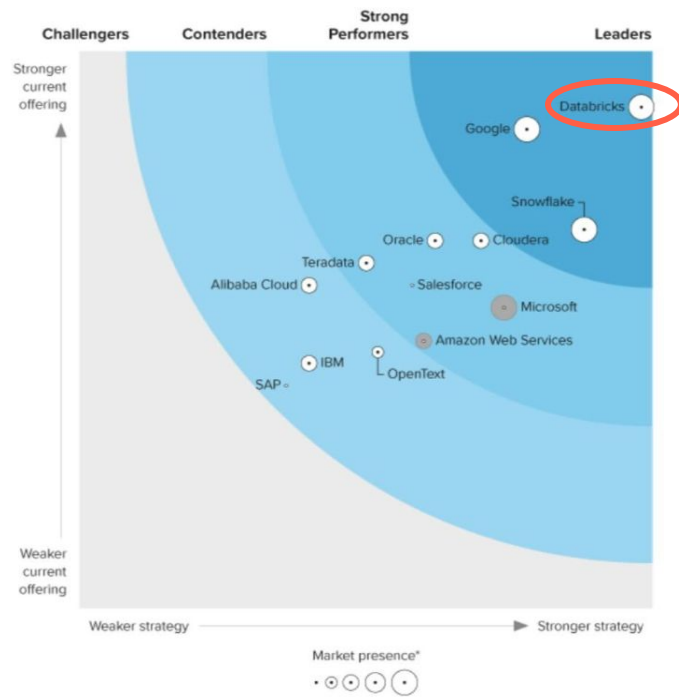
“We’re able to ingest **huge amounts of structured and unstructured data** coming from different systems, standardize it, and then build ML models that deliver alerts and recommendations that empower employees in our call centers, stores, and online.”



# Recognized as a leader in the industry

## Forrester Wave: Data Lakehouses

Q2 2024



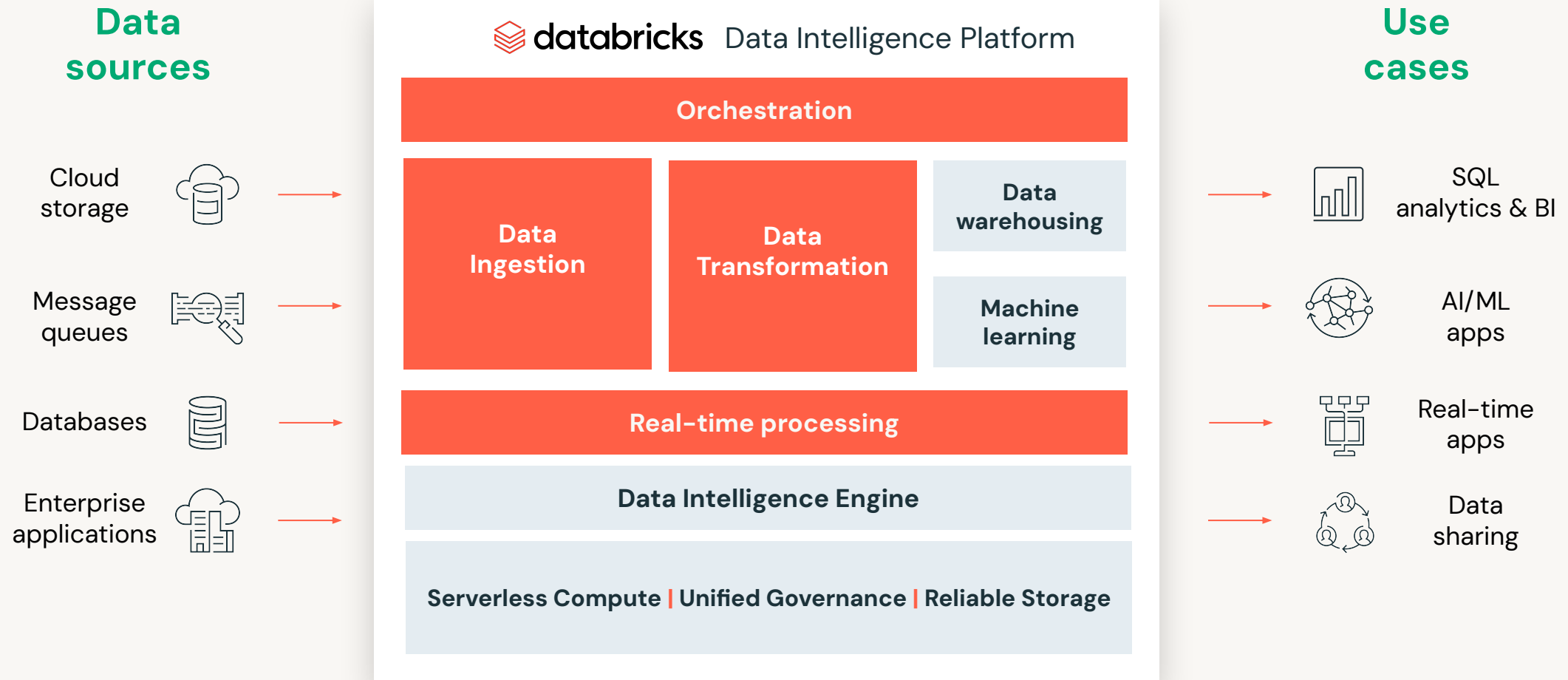
## Forrester Wave: Cloud Data Pipelines



## IDC MarketScape: Analytic Stream Processing



# Data Engineering on Databricks



# Ingestion

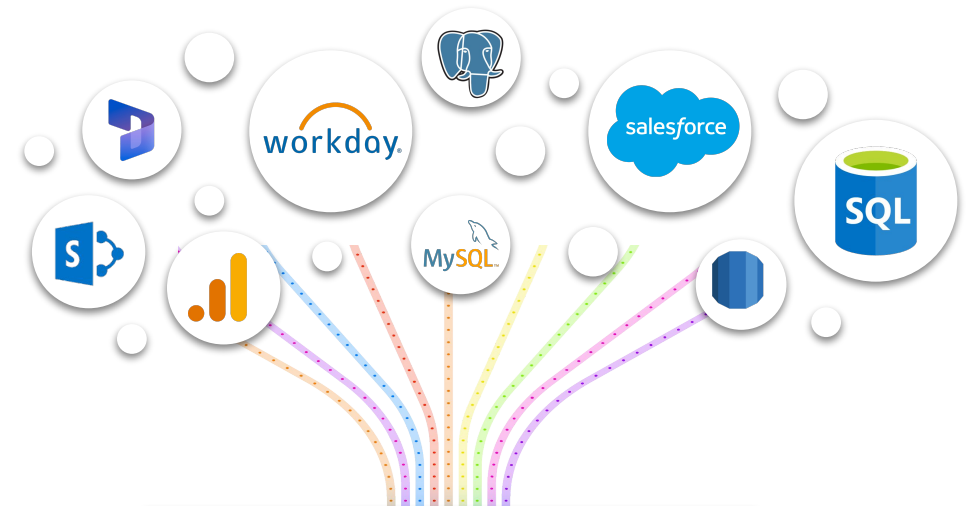
## Efficient ingestion for all

Cost efficient, highly performant data ingestion with incremental reads

Native ingestion support for cloud storage, databases, message buses and business applications.

Large ecosystem of partner solutions

## LakeFlow Connect



# Transformation

## Declarative data pipelines with Delta Live Tables

Faster and cheaper pipelines with automatic incremental processing

Reliable data through automated quality checks

Rapid pipeline development through automated infrastructure and operations

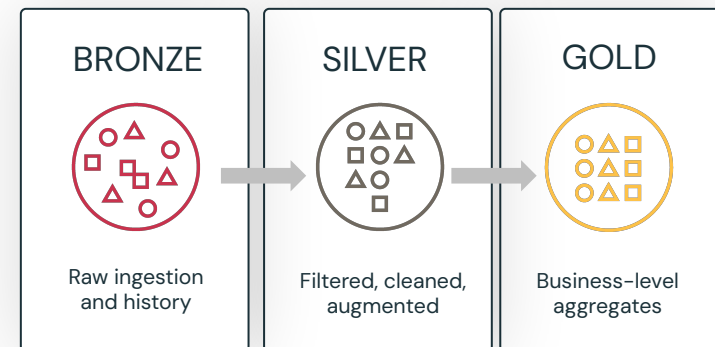
Streamlined data engineering with unified batch and streaming

## Delta Live Tables

```
CREATE STREAMING TABLE raw_data
AS SELECT *
FROM cloud_files ("/raw_data", "json")
```

```
CREATE MATERIALIZED VIEW clean_data
AS SELECT ...
FROM raw_data
```

SQL or  
Python





# Orchestration

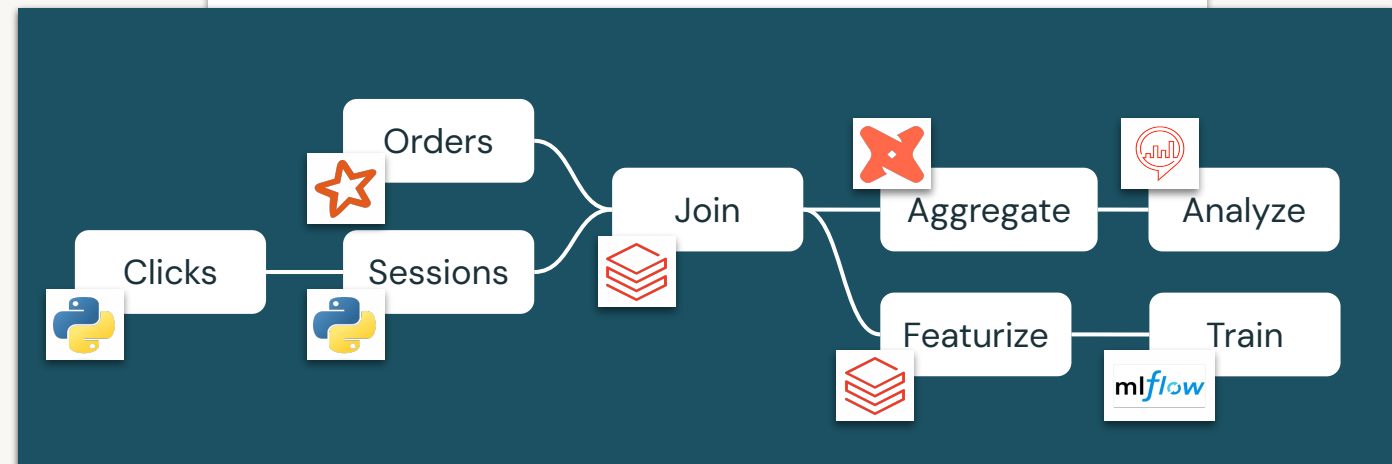
Unified orchestration for Data and AI  
with Databricks Workflows

Simple workflow authoring  
for any practitioner

Actionable insights with  
end-to-end observability

Reliability at scale in production

## Databricks Workflows



# Near Real-time processing

Tunable streaming architecture built on the world's most popular streaming engine

Simplified development with unified batch and streaming APIs

Reliable operations through automatic checkpointing and failure recovery

Easily adjustable throughput and latency for enterprise workload requirements

## Databricks Streaming

Up to **sub-second latency**



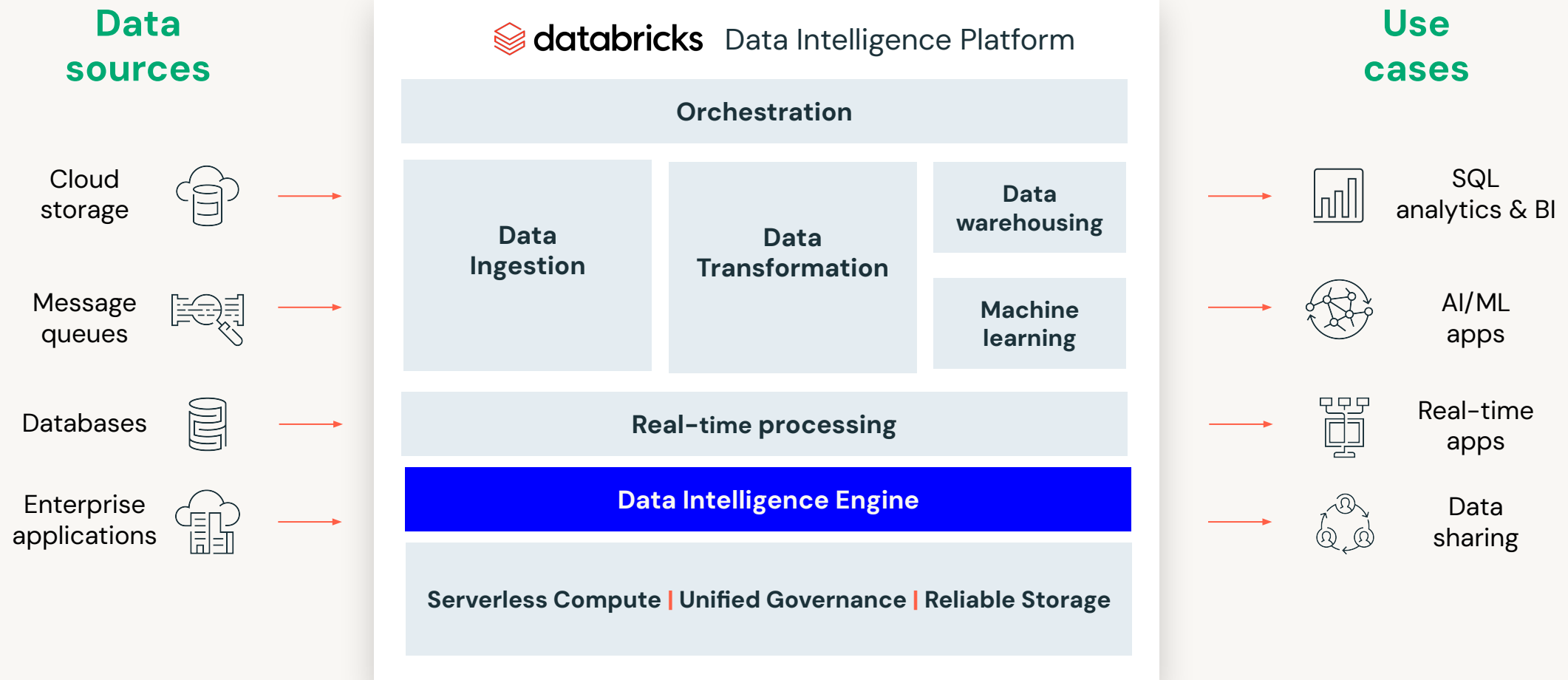
Optimize for Latency



Optimize for Cost



# Data Engineering on Databricks



# Empower every data engineer with AI

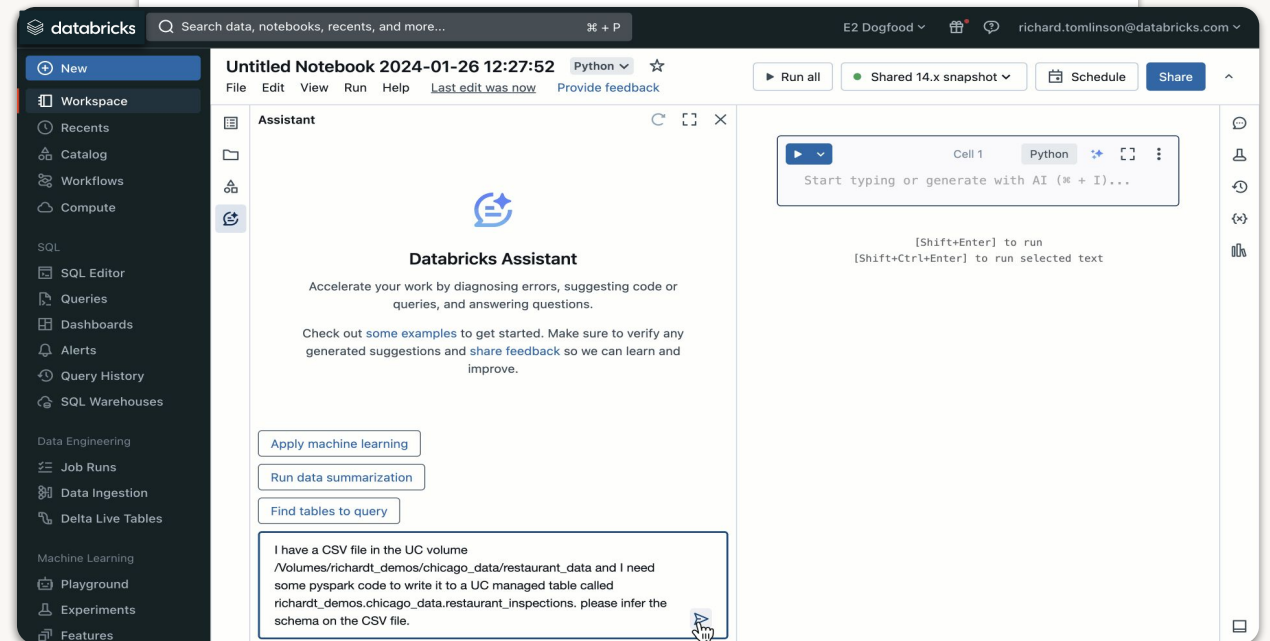
Boost productivity with your context-aware AI assistant

Generate, explain, and fix code with natural language

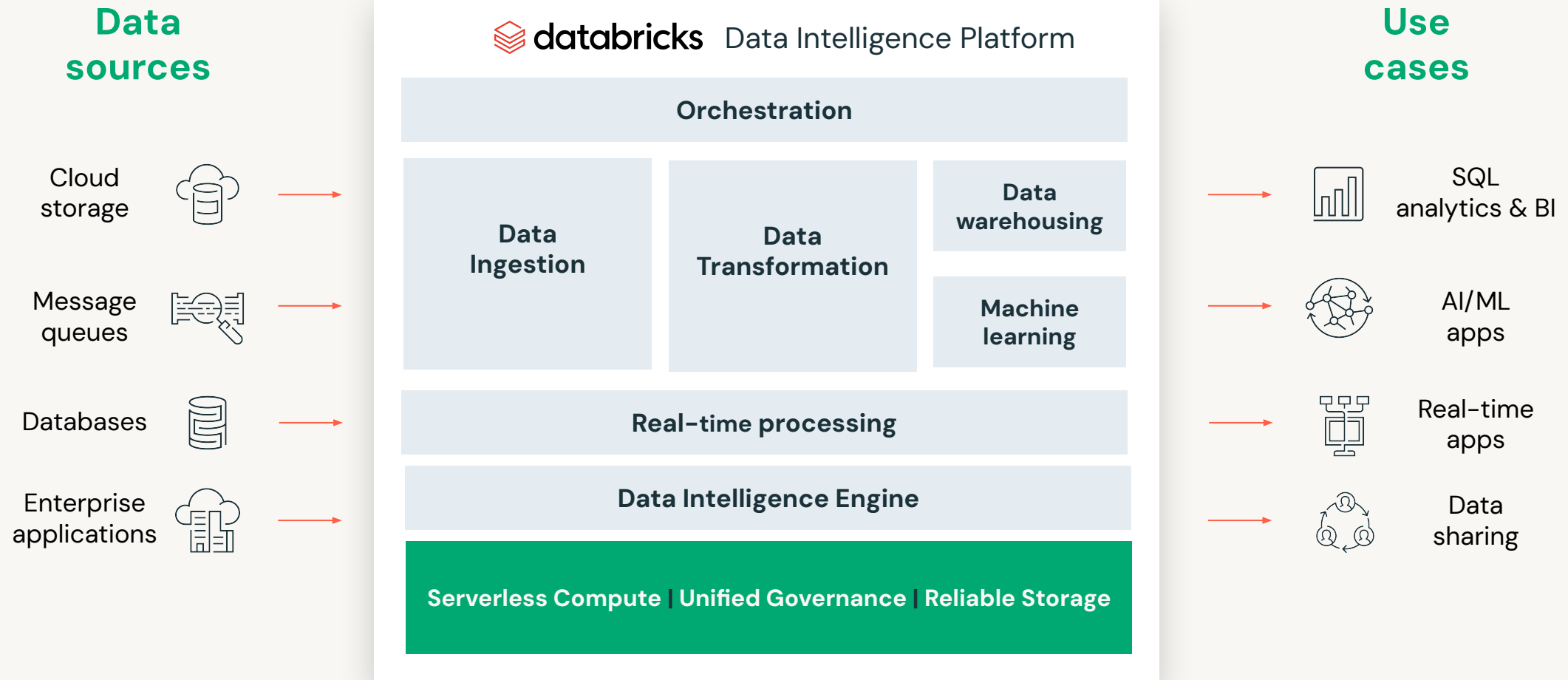
AI assistance in every user experience

Powered by the Data Intelligence engine for highly relevant answers

## Databricks Assistant



# Data Engineering on Databricks



# Built on a solid, open foundation

Fully managed serverless compute, unified governance and reliable storage



## Serverless Compute

**Hands off, fully managed compute**

Fast startups

Cost efficient with smart autoscaling

Stable and secured by default



## Unity Catalog

**Unified governance across all workloads**

Single permission model for data and AI

AI powered monitoring, observability and lineage

Secure open data sharing



## Delta Lake UniForm

**The open format storage layer**

High reliability and performance

Automatic instant translation across open formats with UniForm

Open source

# Highly productive data engineering teams

Empower data engineers with developer experience that meets them where they are

Build data pipelines in **Python, SQL, Scala or Java**

**Automate deployments** with CI/CD and git integrations

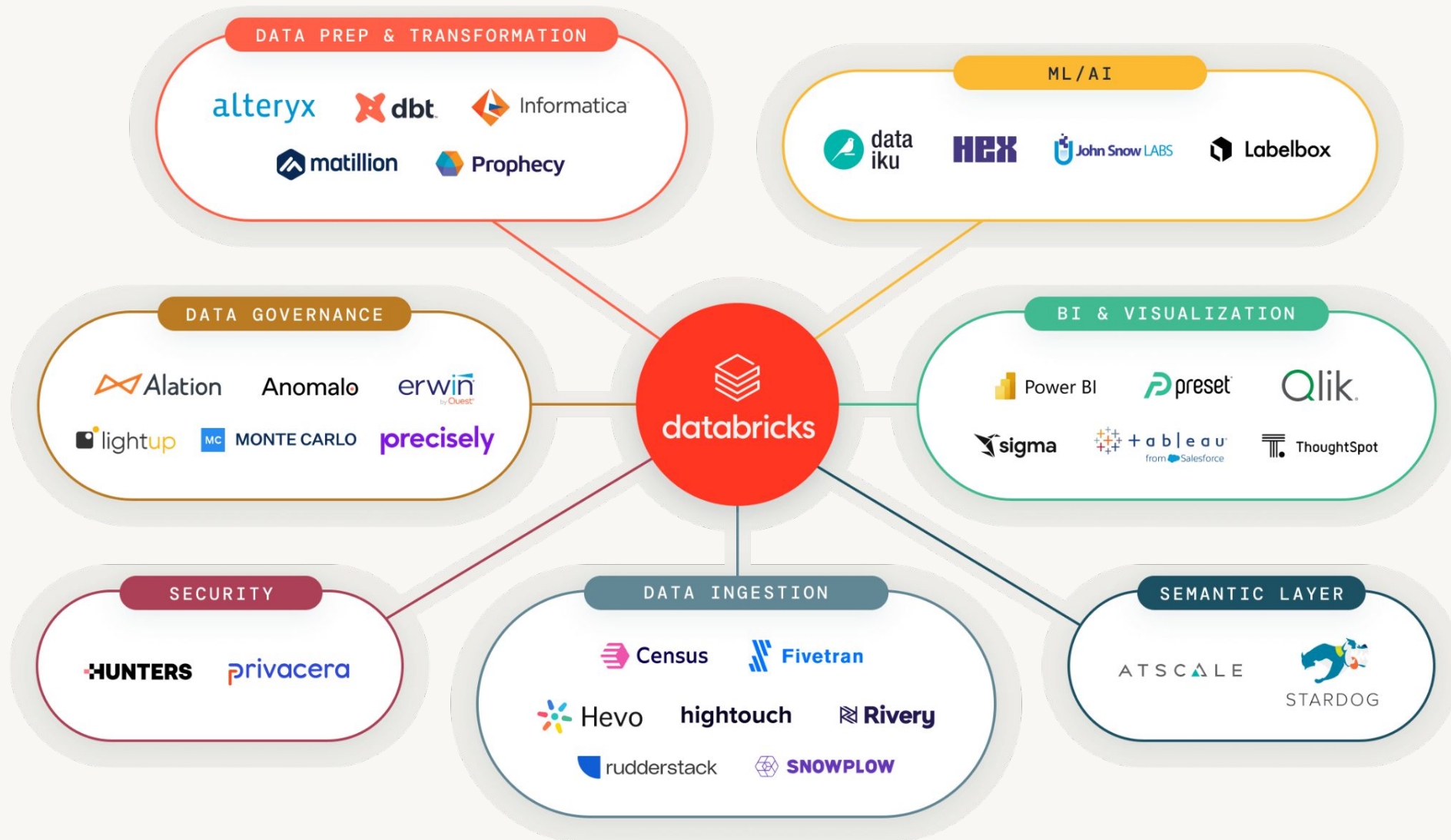
Full integration with **your existing tooling**

A screenshot of a Databricks notebook interface. The window title is "[Extension Development Host] hello.py — notebook-best-practices". The notebook has two tabs: "hello.py U" and "covid\_trends\_job.py M". The code in the notebook is:

```
jobs > hello.py > ...
1 from pyspark.sql import SparkSession
2
3 spark: SparkSession = spark
4
5 print("Hello from Databricks")
6 spark.sql("select * from covid_stats").show(5)
7
```

The interface includes a left sidebar with icons for file operations, search, and a "4k+" badge. The bottom status bar shows the user "fabian\*", a refresh icon, "0 ↓ 2 ↑", "0 ▲ 0", "1", and buttons for "Run on Databricks (notebook-best-practices)", "Connect", and "Git Graph".

# Integrated with the tools you know and love





# Data engineering: Side by side comparison



Price/performance at scale

❌ Up to 9x higher TCO at high scale

✅ Highly efficient at any scale

Declarative pipelines

✅ Dynamic tables

✅ Delta Live Tables

Serverless with autoscaling

✅ Define warehouse size  
No autoscaling at job level

✅ No need to select VM size  
Autoscale at job level

Developer experience

❌ Build pipelines in SQL, Python support limited or in private preview

✅ Build pipelines in SQL, Python, Scala, or Java

DevOps

✅ Git Integration for notebooks only  
Can execute files saved in git

✅ Git integration, CI/CD, Databricks Asset Bundles

Data Streaming

❌ Latency of 1 min

✅ Sub-second streaming latency

Orchestration

✅ Basic native orchestration  
complimented with third-party tools

✅ Native orchestrator with advanced features





**Akamai**

**Tomer Patel**

Engineering Manager  
Akamai

“Because of our **scale and the demands** of our SLA, Databricks was the right solution for us. If we went with another solution, we couldn't achieve the same level of performance.”

Akamai's web security analytics tool handles approximately **10GB of data related to security events per second. That's almost 1PB a day!**



Let's see the  
product!







---

Thank you for downloading this Databricks presentation! Carahsoft serves as the Master Government Aggregator® and Distributor for Databricks, offering expertise in government procurement processes and practices with purchasing available via GSA, SEWP V, ITES-SW and other contract vehicles.

To learn how to take the next step toward acquiring Databricks' solutions, please check out the following resources and information:



For additional resources, please visit [carah.io/DatabricksResources](https://carah.io/DatabricksResources)



For additional solutions, visit [carah.io/DatabricksSolutions](https://carah.io/DatabricksSolutions)



To speak with our team directly, email [Databricks@carahsoft.com](mailto:Databricks@carahsoft.com) or reach out at 703-581-6693.



To view our upcoming Databricks events, visit [carah.io/DatabricksEvents](https://carah.io/DatabricksEvents)



For additional Open Source solutions, visit [carah.io/OpenSourceSolutions](https://carah.io/OpenSourceSolutions)



To purchase, check out the contract vehicles available for procurement at [carah.io/DatabricksContracts](https://carah.io/DatabricksContracts)

**carahsoft**

For more information, contact Carahsoft or our reseller partners:  
[Databricks@carahsoft.com](mailto:Databricks@carahsoft.com) | 703-581-6693