databricks

Simplifying Data, Analytics & AI

Elmer Cecilio - Sr Solutions Architect

Agenda

- Databricks Overview
- Databricks Core Capabilities
- Demo
  - Retail Review Response using LLM

Creator of

DELTA LAKE

APACHE Spark™

mlflow™

Inventor and pioneer
of the **data lakehouse**

databricks

The Lakehouse Company

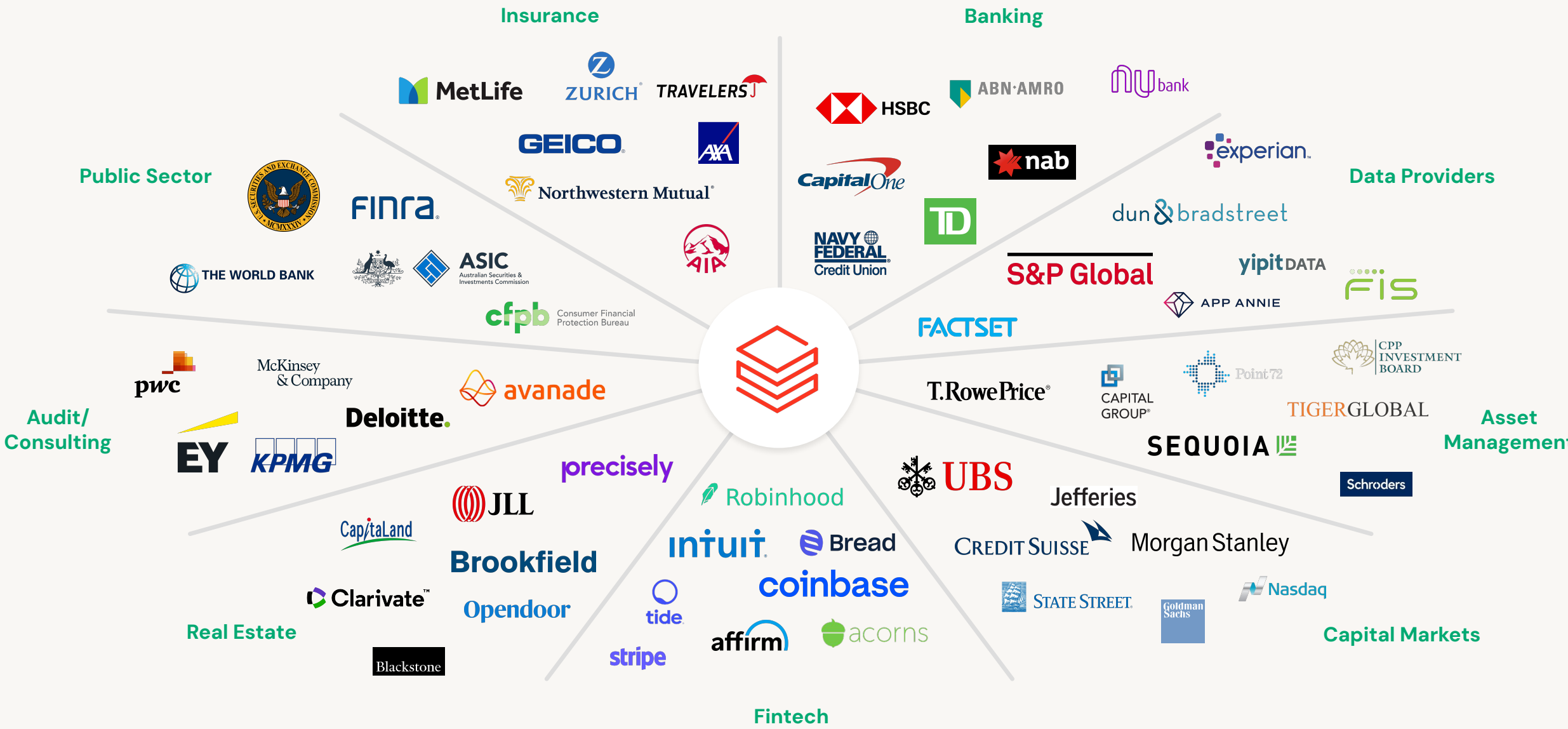
Gartner-recognized Leader
Database Management Systems
Data Science and Machine Learning Platforms

5000+
global employees

$1B+
in revenue

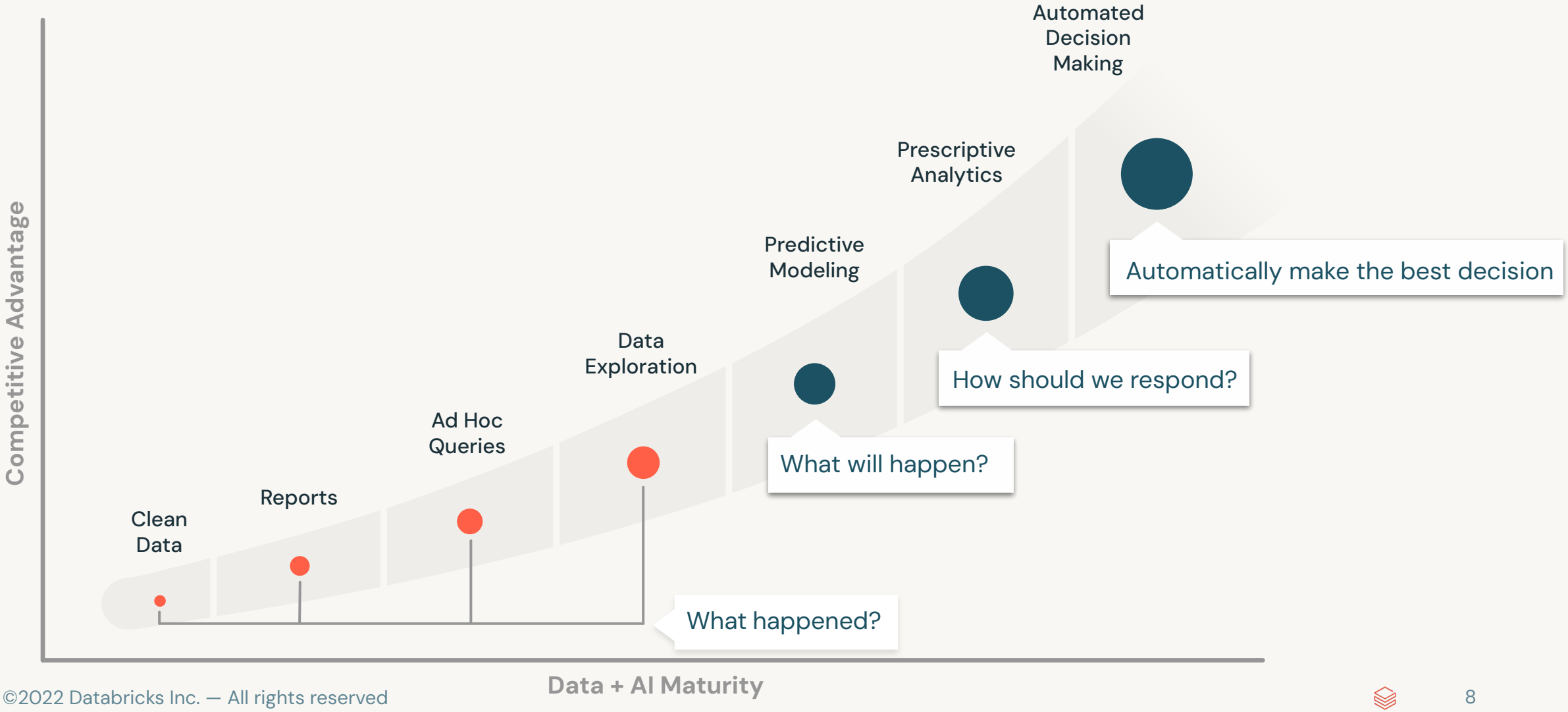
$3B
in investment

FS Customers: 600+ across the globe

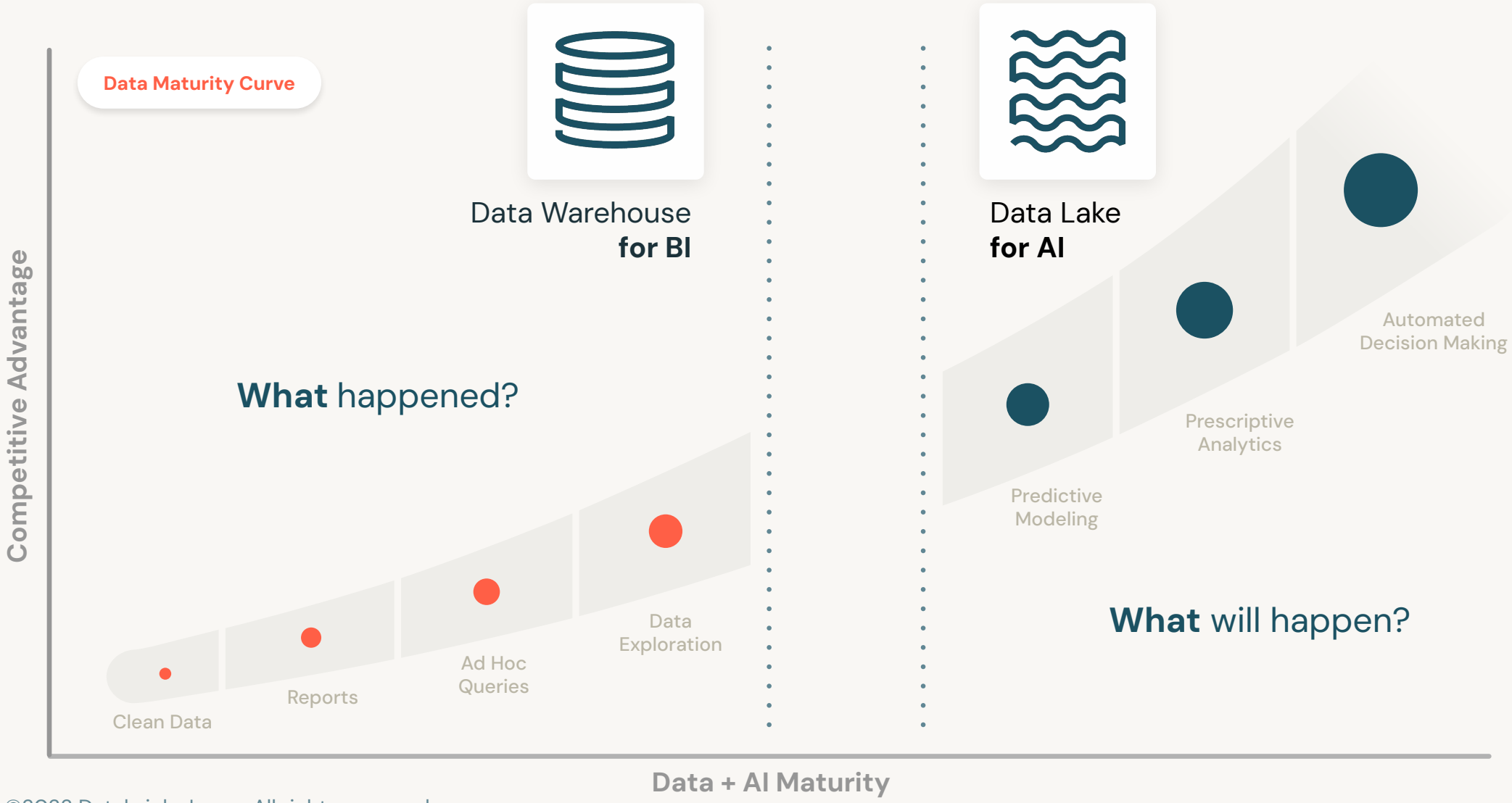


Tech leaders are to the right of the Data Maturity Curve

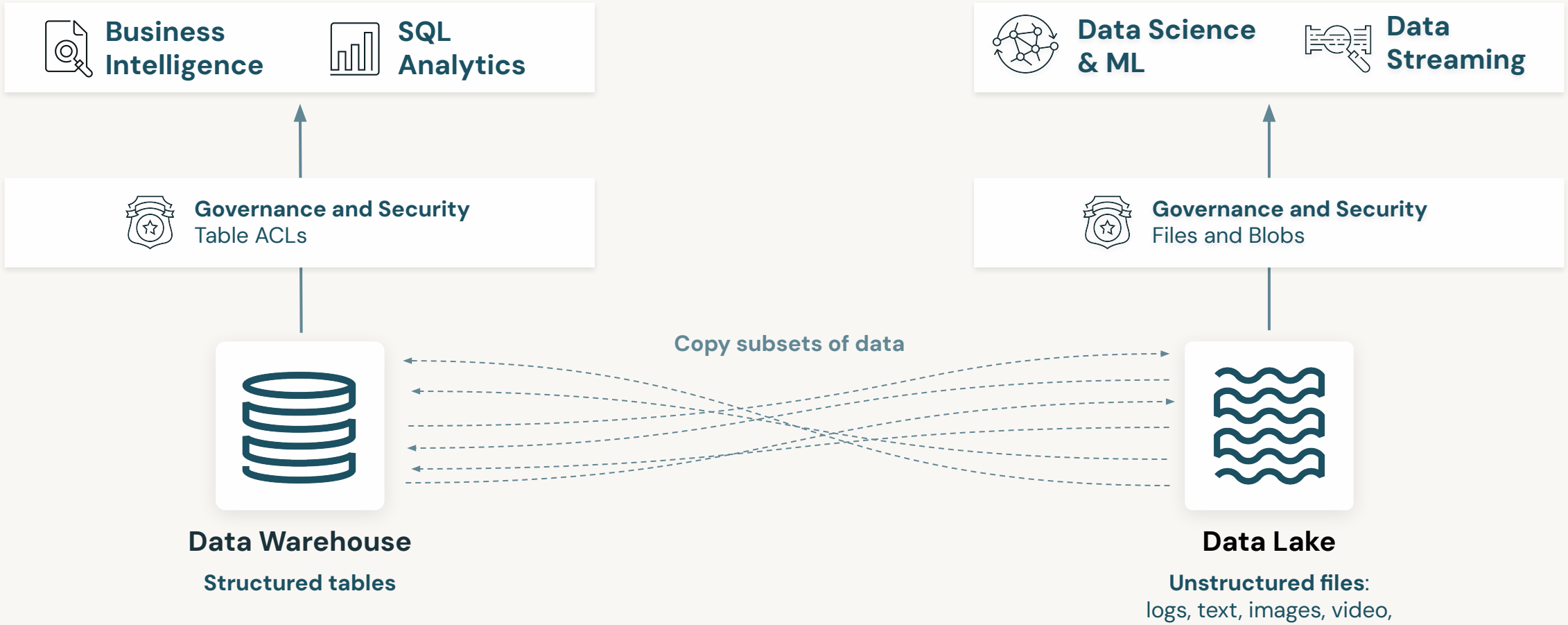
From hindsight to foresight



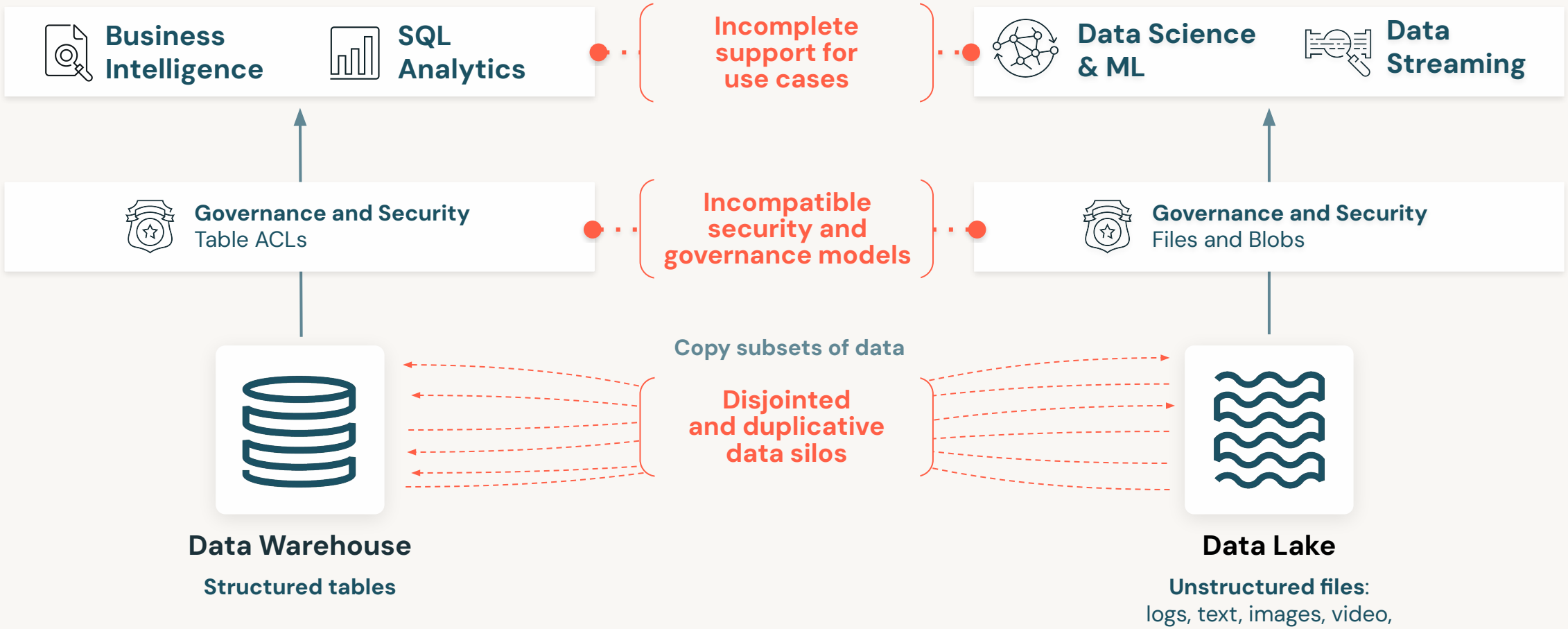
Realizing this requires two disparate, incompatible data platforms



Realizing this requires two disparate, incompatible data platforms



Realizing this requires two disparate, incompatible data platforms



A data lakehouse takes a different approach

One platform to support multiple personas

BI & Data Warehousing

Data Engineering

Data Streaming

Data Science & ML

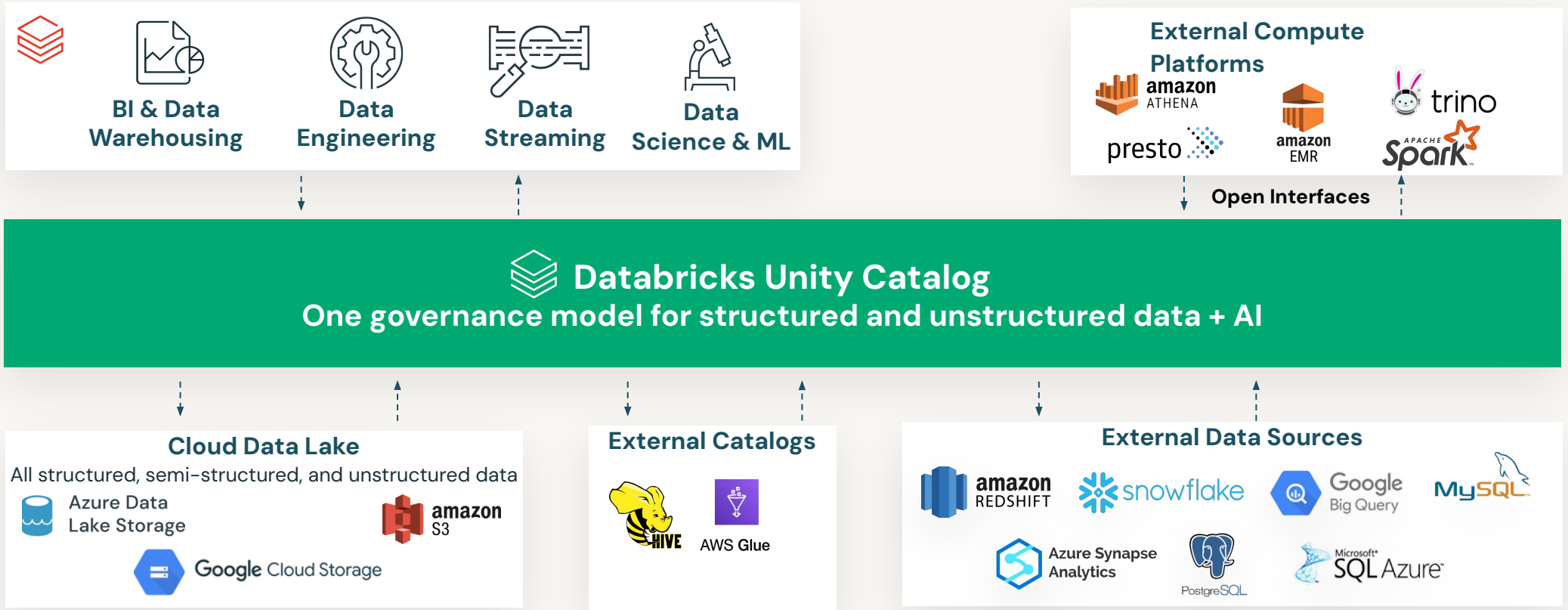
One security and governance model for all data access across the organization

One platform to store and manage all structured, semi-structured, and unstructured data

Cloud Data Lake

All Raw Data
(Logs, Texts, Audio, Video, Images)

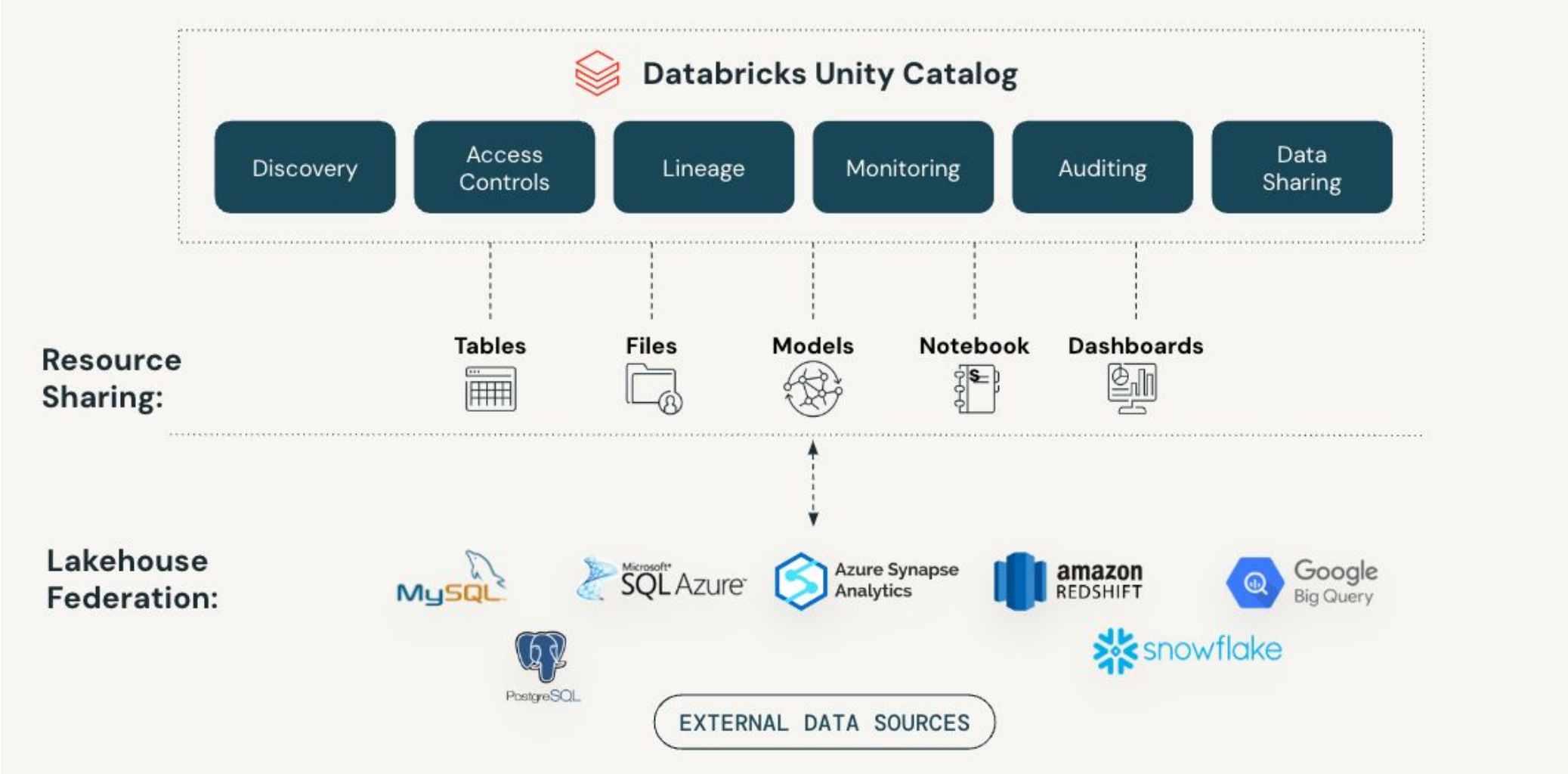
Databricks Lakehouse unifies data and AI governance



Databricks Core Capabilities

Innovations in Unity Catalog

Simplify, Secure, and Scale Your Data and AI Resources



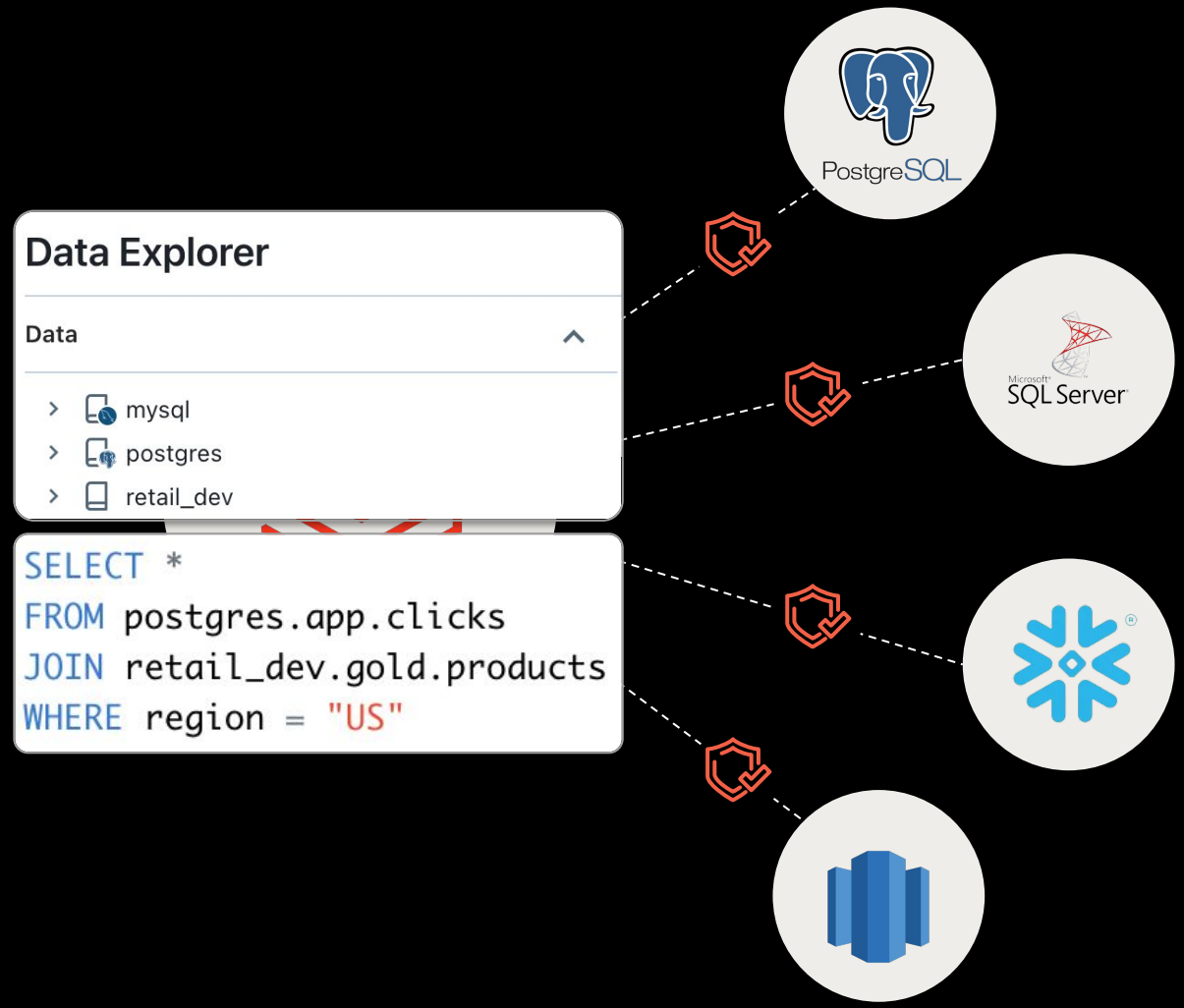
Lakehouse Federation

Discover, query, and govern your data wherever it lives

Connect data sources
to Unity Catalog and set policies across
them (access controls, tags, etc)

Efficient optimization & caching
across sources

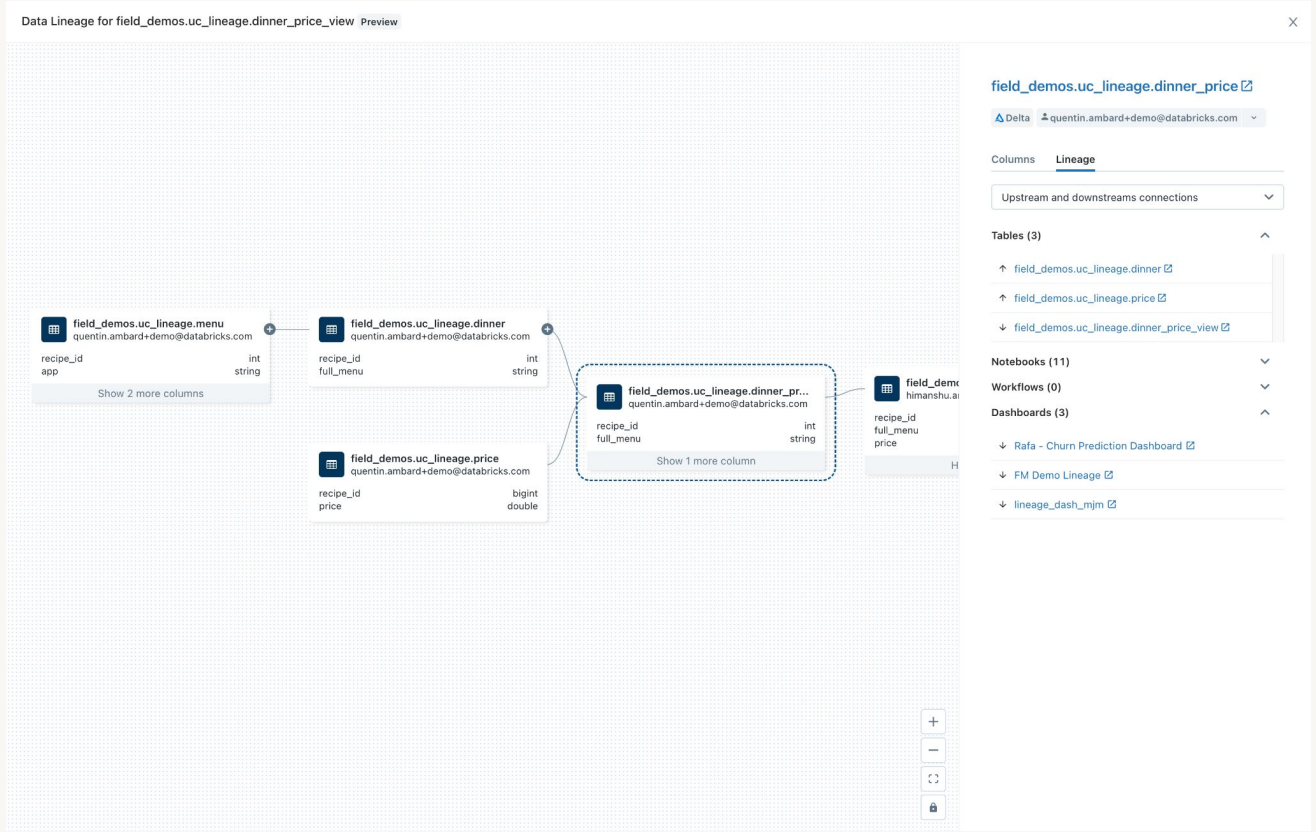
Coming later: policy push
into warehouses



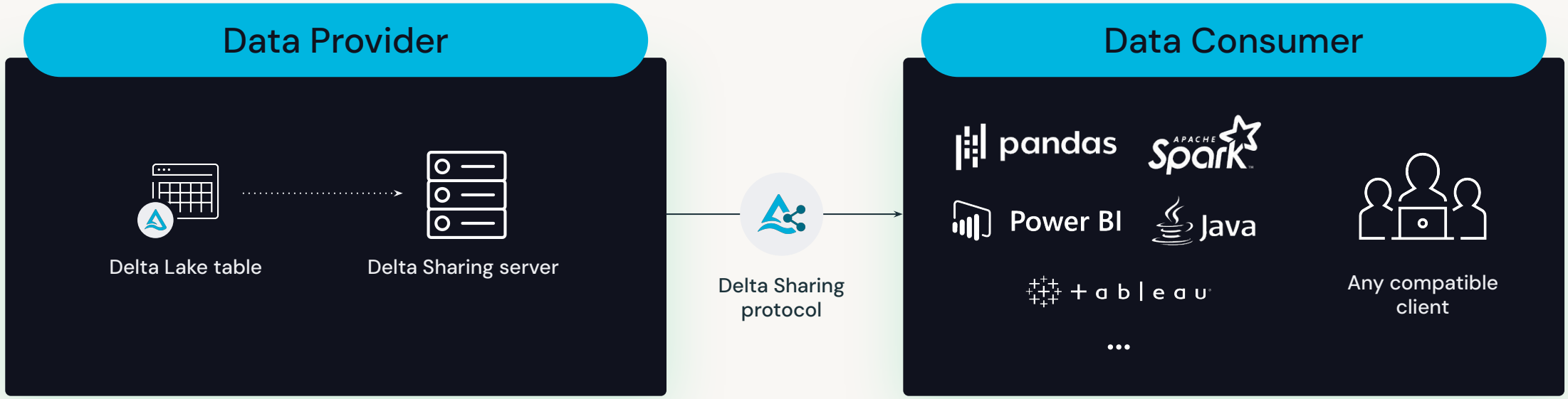
Automated lineage for all workloads

End-to-end visibility into how data flows and consumed in your organization

- Auto-capture runtime data lineage on a Databricks cluster or SQL warehouse
- Track lineage down to the table and column level
- Lineage across tables, dashboards, workflows, notebooks, feature tables, files, ML Models and DLT



Streamlined Sharing with Delta Sharing



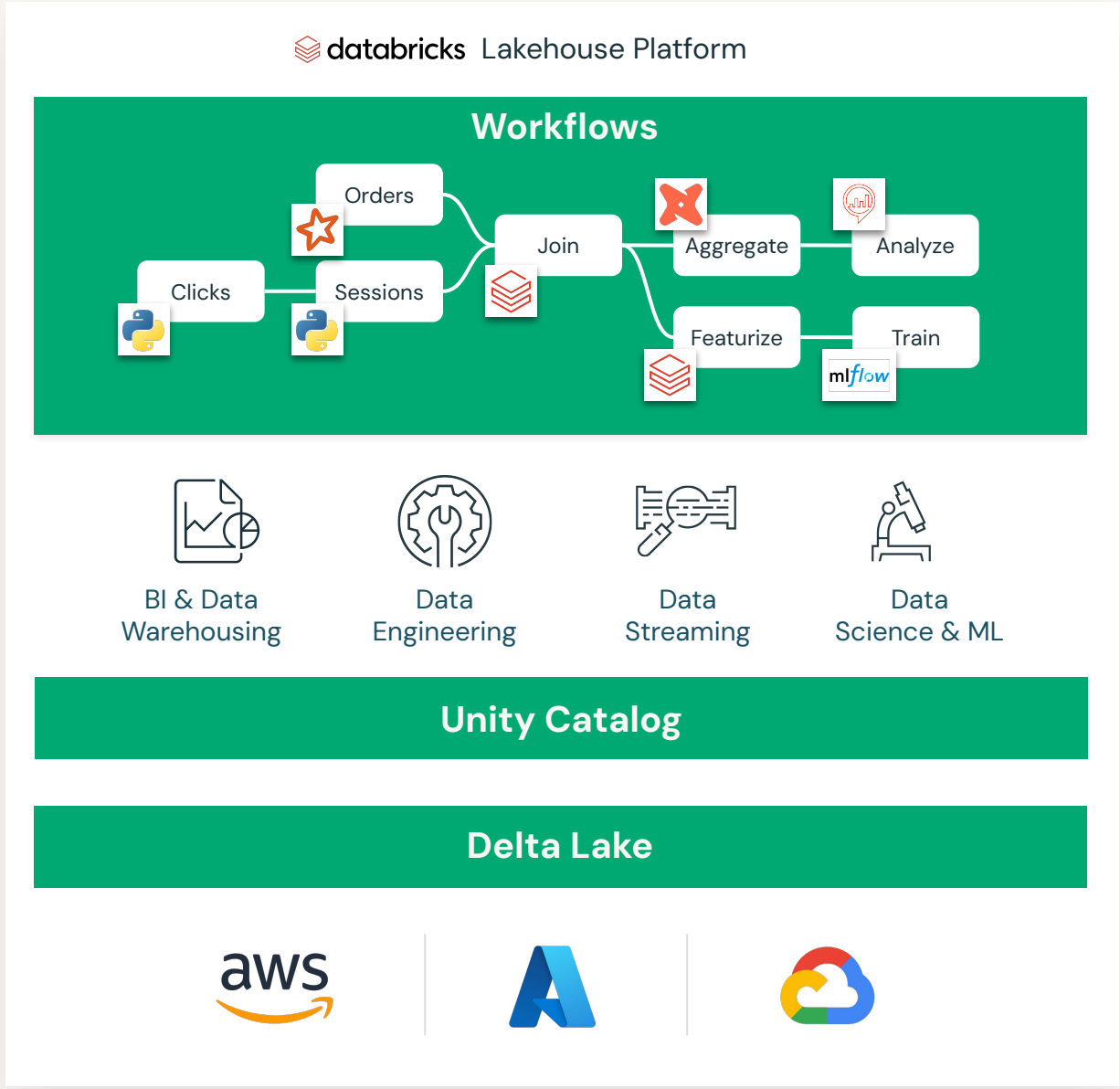
Open cross-platform sharing

Share live data with no replication

Centralized governance

Data engineering workloads on Databricks

- Operate on massive scale datasets with proven, reliable, distributed compute framework (Spark API)
- Delta Live Tables manage your full data pipelines
- Simplify data engineering with a curated data lake-centric approach with Delta Lake

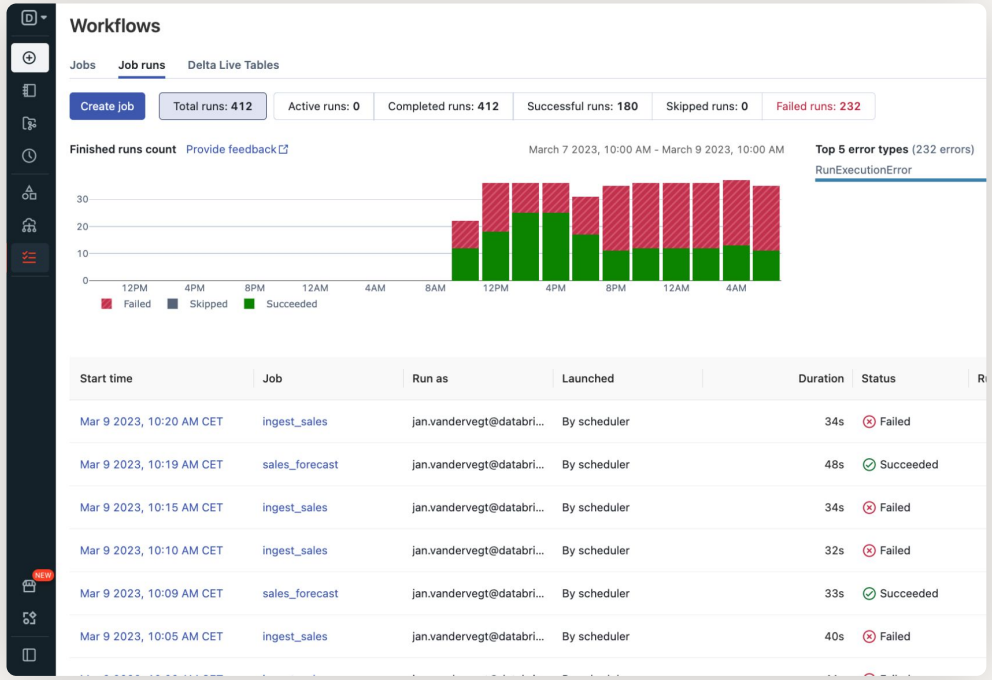


Databricks Workflows

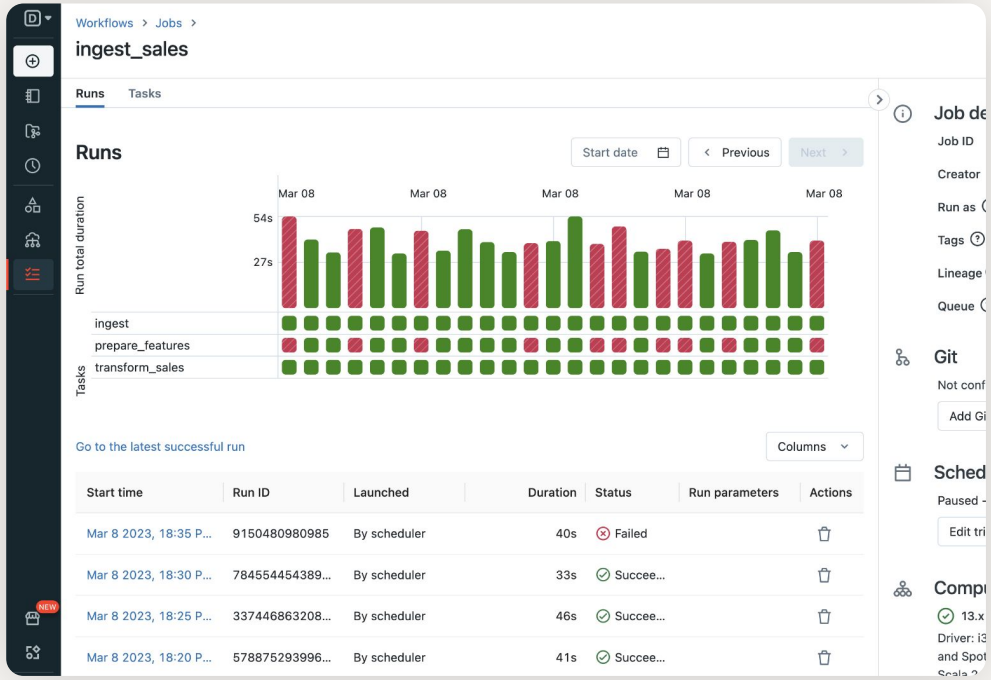
Unified orchestration for data, analytics, and AI on the Lakehouse Platform

- Simple authoring
- Actionable insights
- Proven reliability

Actionable insights from real-time monitoring



A simple and intuitive monitoring UI provides real-time metrics and detailed analytics for every workflow run



Drill down to understand which tasks are failing and why. Troubleshoot issues before your customers are impacted

Data Warehousing on the Lakehouse

Powered by Databricks SQL

Databricks SQL (DB SQL) is a serverless data warehouse on the Databricks Lakehouse Platform that lets you run all your SQL and BI applications at scale with up to 12x better price/performance, a unified governance model, open formats and APIs, and your tools of choice - no lock-in.

Best price/performance

Built-in governance

Rich ecosystem

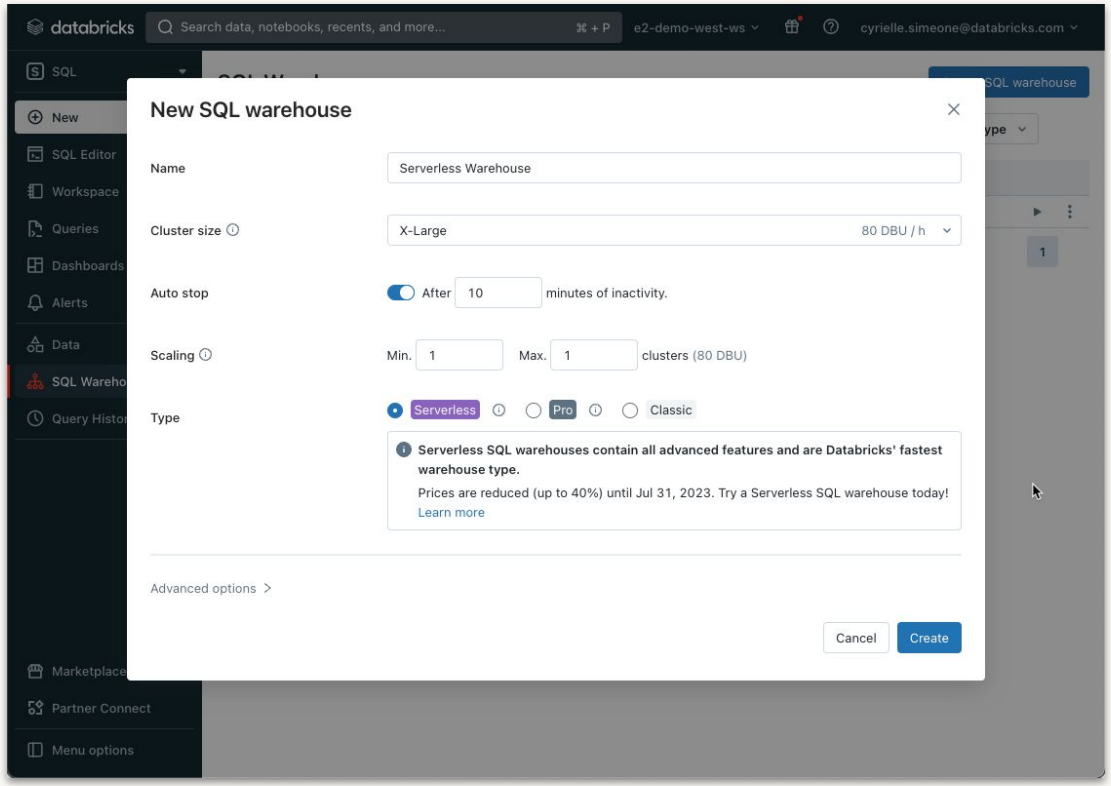
Break down silos

Serverless compute for Databricks SQL

Instant, elastic SQL compute decoupled from storage

Lower costs and increase productivity with **instant, elastic SQL serverless compute – decoupled from storage**.

Databricks automatically determines instance types and configuration for the **best price/performance** and scale for **high concurrency** BI needs.



GA! GA! Coming Soon

ML & data science workloads on Databricks

Machine Learning

- Model registry, reproducibility, productionization
- Leverages Delta Lake for reproducibility
- AutoML for citizen data scientists
- End-to-end integrated MLOps

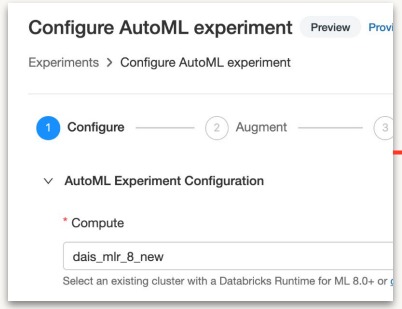
Data Science

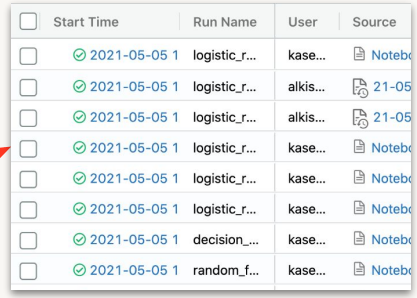
- Collaborative notebooks and dashboards for interactive analysis
- Native support for Python, Java, R, Scala
- Delta Lake data natively supported

AutoML on Databricks

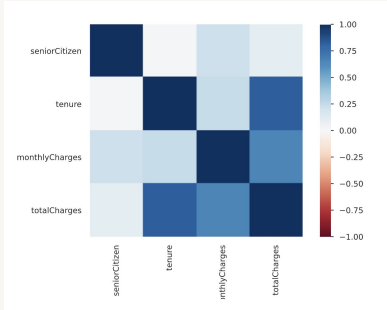
Rapidly generate baseline models for discovery and iteration

UI and API to start AutoML

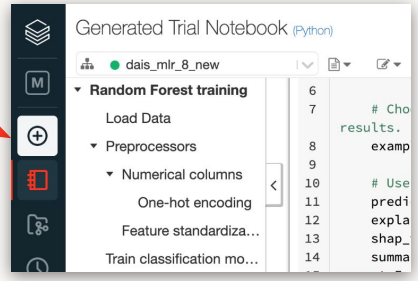




MLflow experiment
Auto-created MLflow Experiment to track models and metrics



Data exploration notebook
Generated notebook with feature summary statistics and distributions



Reproducible trial notebooks
Generated notebooks with source code for every model

Easily deploy to Model Registry

Understand and debug data quality and preprocessing

Iterate further on models from AutoML, adding your expertise

WSJ PRO
VENTURE CAPITAL

SUBSCRIBE SIGN OUT

INDUSTRY NEWS

Databricks Launches 'Dolly,' Another ChatGPT Rival

The data-management startup introduced an open-source language model for developers to build their own AI-powered chatbot apps

By Angus Loten
March 24, 2023 8:00 am ET | WSJ PRO

REUTERS
World Business Legal Markets More
My View Following Saved

Disrupted

2 minute read · March 24, 2023 8:02 AM EDT · Last Updated 5 days ago

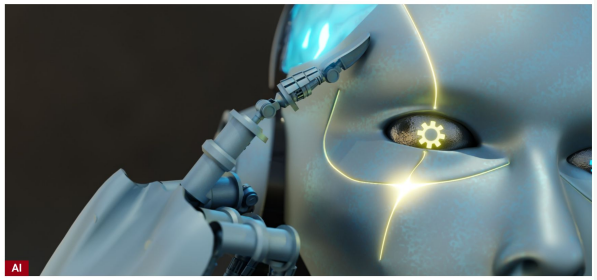
Databricks pushes open-source chatbot as cheaper ChatGPT alternative

By Krystal Hu and Stephen Nellis

siliconANGLE

AI/ML AWS STARTUP SHOWCASE.
Top Startups Building Generative AI on AWS

UPDATED 08:00 EDT / MARCH 24 2023



Databricks open-sources an AI it says is as good as ChatGPT, but much easier to train

Hello Dolly: Democratizing the magic of ChatGPT with open models

by Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Ali Ghodsi, Patrick Wendell and Matei Zaharia

March 24, 2023 in Company Blog

Share this post

Summary

We show that anyone can take a dated off-the-shelf open source large language model (LLM) and give it magical ChatGPT-like instruction following ability by training it in 30 minutes on one machine, using high-quality training data. Surprisingly, instruction-following does not seem to require the latest or largest models: our model is only 6

https://www.databricks.com/blog/2023/03/24/hello-dolly-democratizing-magic-chatgpt-open-models.html

Events Special Issues VentureBeat

Artificial Intelligence Security Data Infrastructure Automation Enterp

Databricks debuts ChatGPT-like Dolly, a clone any enterprise can own

The Information Sign In SUBSCRIBE

BRIEFING CLOUD ENTERPRISE AI

Taking Aim at OpenAI, Databricks Releases 'Dolly' Language Model

By Kevin McLaughlin · 4 days ago · Source: The Information

datanami
DATA SCIENCE • AI • ADVANCED ANALYTICS

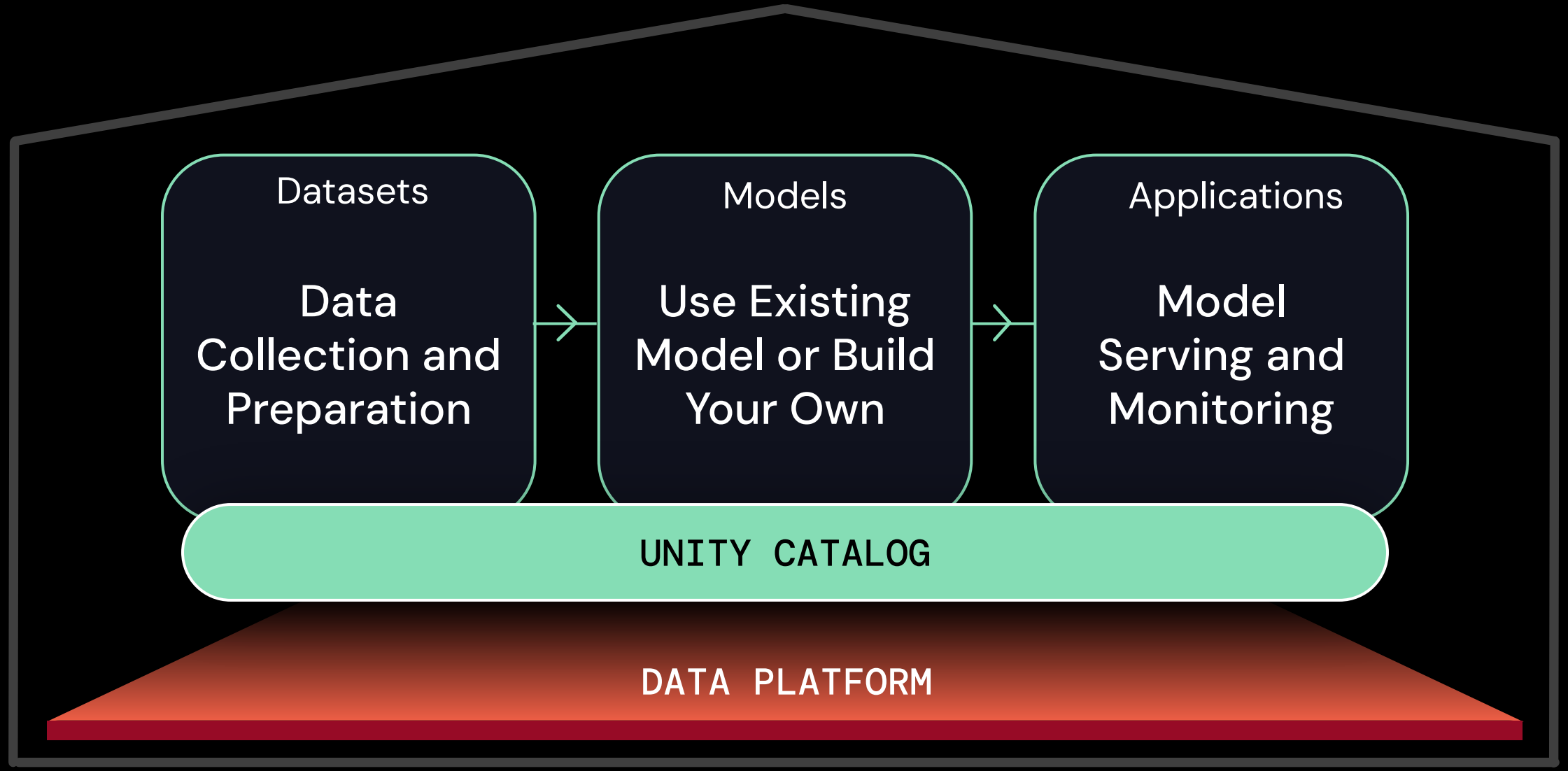
About Resources Subscribe

Menu

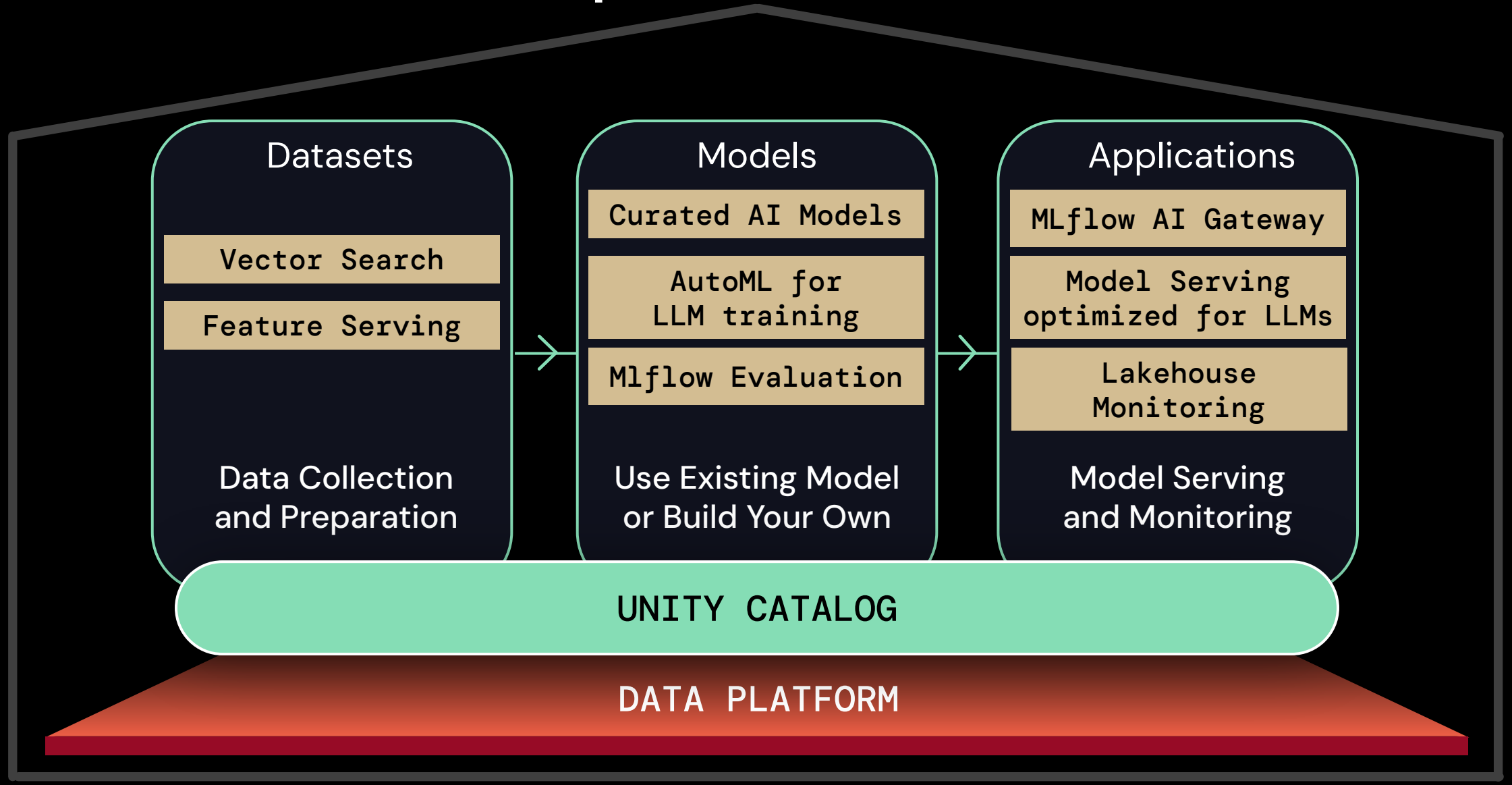
March 24, 2023
Databricks Bucks the Herd with Dolly, a Slim New LLM You Can Train Yourself
Alex Woodie


databricks
mosaicML

Lakehouse AI — a data-centric AI Platform



Lakehouse AI — optimized for Generative AI

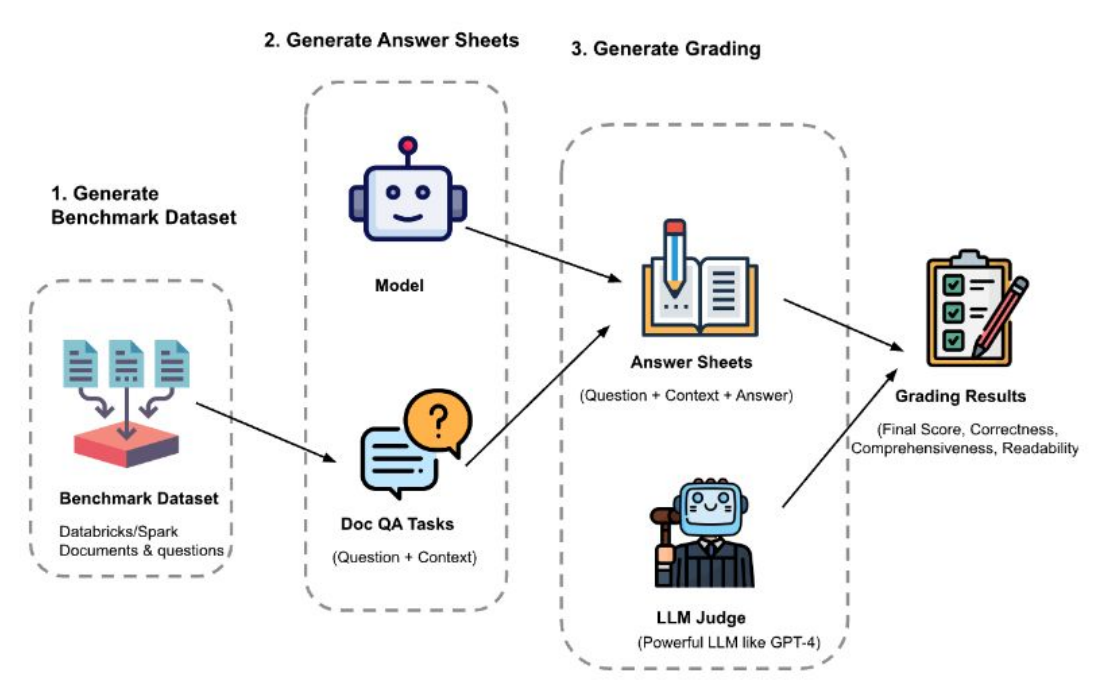


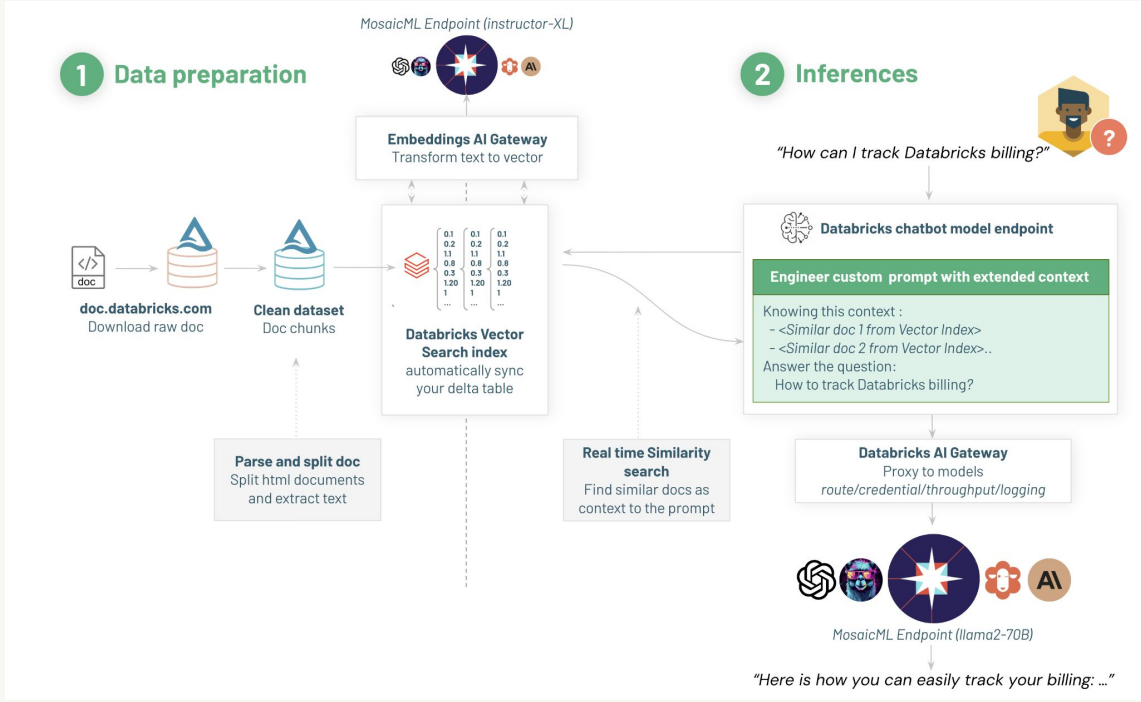
Best-in-class open source Gen AI models for free commercial use

Use case	Quality-optimized	Balanced	Speed-optimized	Notes
Text generation following instructions	MPT-30B-Instruct † Llama-2-70b-chat-hf	MPT-7B-Instruct † MPT-7b-8k-instruct Llama-2-7b-chat-hf Llama-2-13b-chat-hf		† Supervised fine-tuning using databricks-dolly-15k data set
Text embeddings (English only)		Bge-large-en-v1.5 (0.3B) e5-large-v2 (0.3B) instructor-xl (0.3B)	bge-base-en (0.4B) bge-small-en (0.1B) e5-base-v2 (0.1B)	
Transcription (speech to text)		whisper-large-v2 (1.6B) whisper-medium (0.8B, English only)		
Image generation		stable-diffusion-xl		
Code generation	CodeLlama-34b-hf CodeLlama-34b-Instruct-hf CodeLlama-34b-Python-hf (Python optimized) WizardCoder-Python-34B-V1.0	CodeLlama-13b-hf CodeLlama-13b-Instruct-hf CodeLlama-13b-Python-hf (Python optimized) CodeLlama-7b-hf CodeLlama-7b-Instruct-hf CodeLlama-7b-Python-hf (Python optimized) WizardCoder-Python-13B-V1.0 WizardCoder-15B-V1.0	WizardCoder-1B-V1.0	Code LLMs usually need fine-tuning to follow instructions and work on application-specific code

Link to page - https://www.databricks.com/product/machine-learning/large-language-models-oss-guidance

LLM RAG





[Best Practices for LLM Evaluation of RAG Applications]

[Deploy Your LLM Chatbot With Retrieval Augmented Generation (RAG), llama2-70B (MosaicML inferences) and Vector Search]

Databricks Marketplace

Open marketplace for all your data, analytics, and AI

Providers

Reach users on any platform

Monetize more than just data

Share data securely

Databricks Marketplace

Consumers

Discover more than just data

Evaluate data products faster

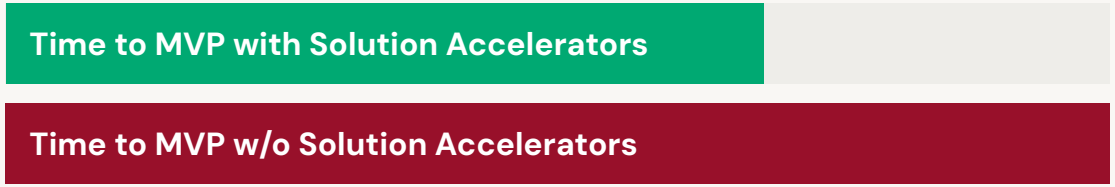
Avoid vendor lock-in

75+ providers

500+ listings

Get started with Solution Accelerators

Solution Accelerators quickly demonstrates the technical feasibility of using Databricks for an industry use-case

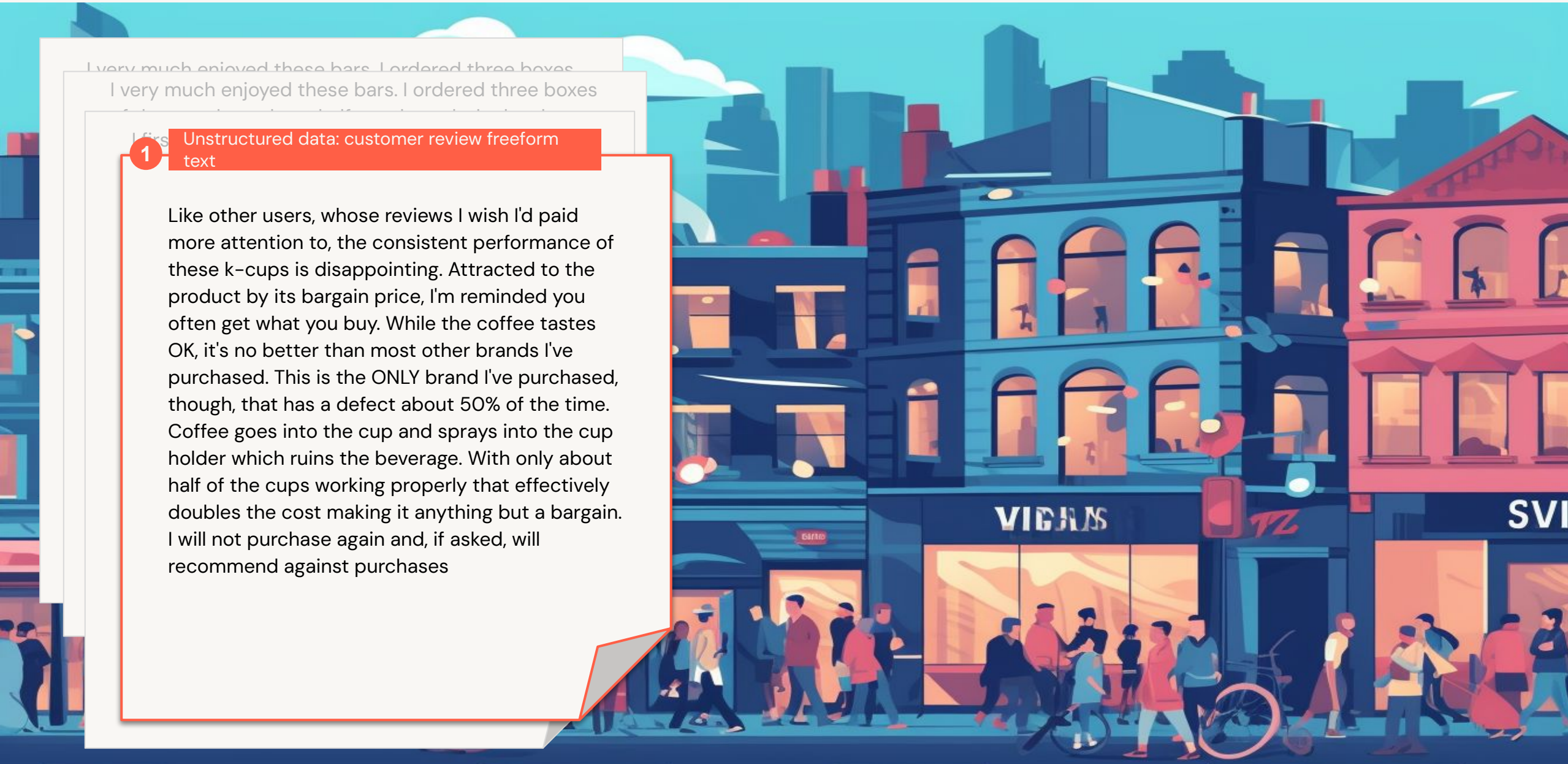


50+ Solution Accelerators

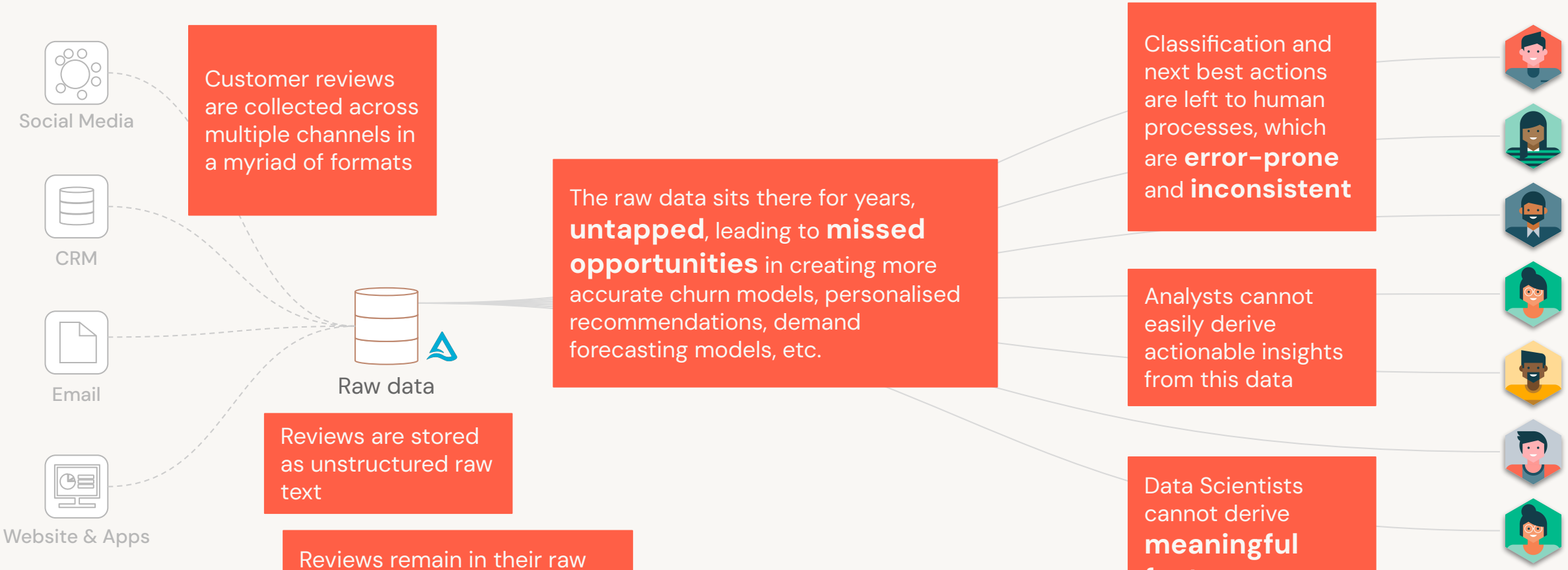
Market risk Simulation	**Alternative data**
Market risk Backtest	**ESG** Operationalizing
Investment data platforms	**Regulatory reporting**
Investment analytics	**C360** Forecasting

Demo

How to get started - Walk through a Retail use case



The Problem



The Fix

Social Media

CRM

Email

Website & Apps

Raw data

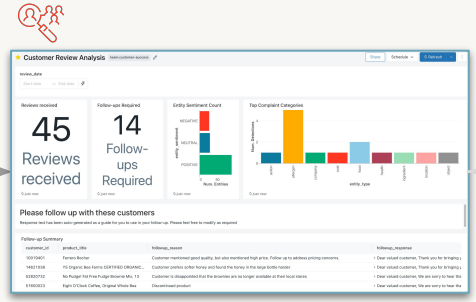
```
AI_GENERATE_TEXT()

"Tell me about the
sentiments expressed in each
review. Do I need to follow
up with my customer? Why?"

"Suggest a starting point
for a message I can send
back to my customer"
```

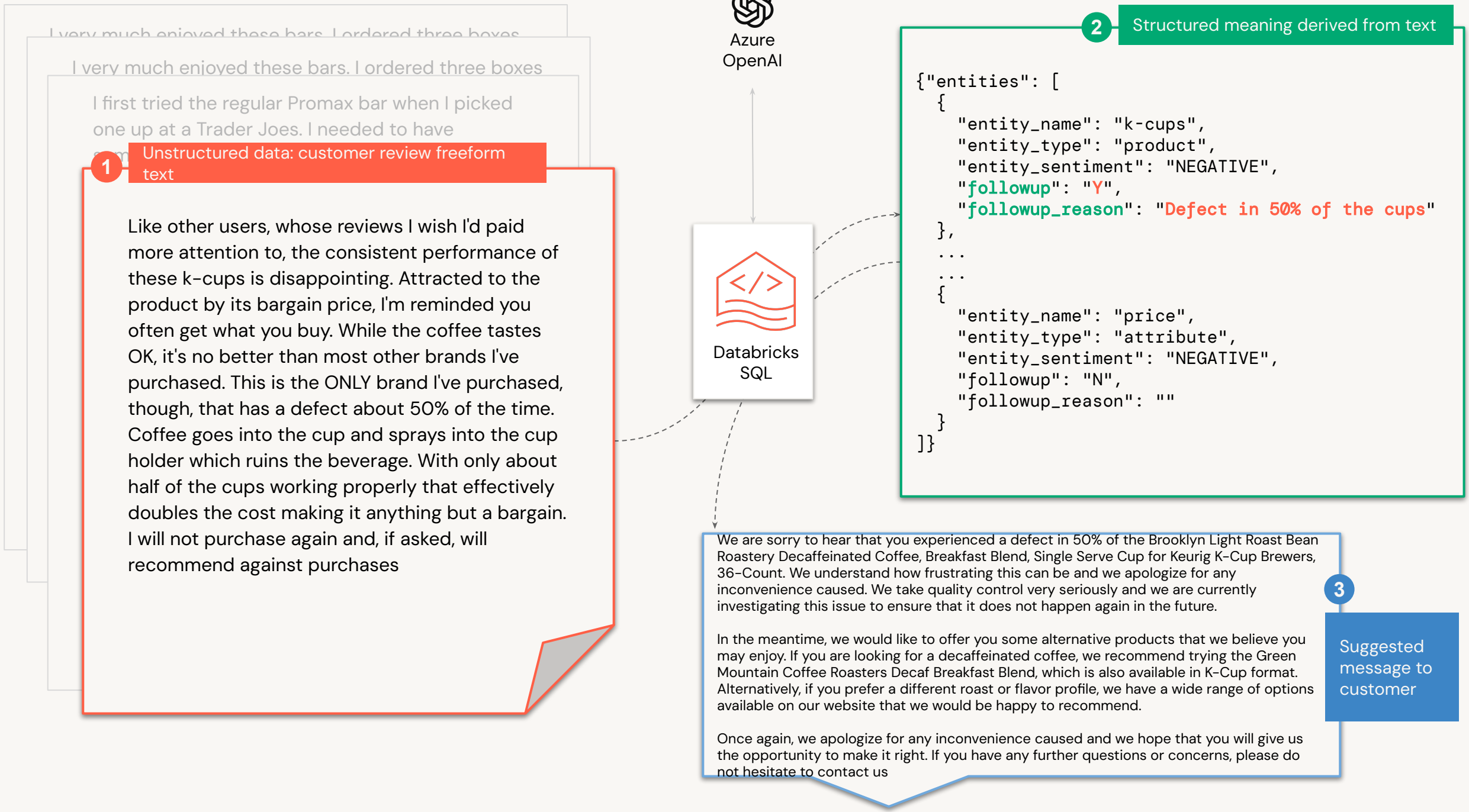
Enriched data

Entity sentiment, follow up required, messaging

Business-ready data

Next best actions




I very much enjoyed these bars. I ordered three boxes
I very much enjoyed these bars. I ordered three boxes
I first tried the regular Promax bar when I picked one up at a Trader Joes. I needed to have
1
Unstructured data: customer review freeform text
Like other users, whose reviews I wish I'd paid more attention to, the consistent performance of these k-cups is disappointing. Attracted to the product by its bargain price, I'm reminded you often get what you buy. While the coffee tastes OK, it's no better than most other brands I've purchased. This is the ONLY brand I've purchased, though, that has a defect about 50% of the time. Coffee goes into the cup and sprays into the cup holder which ruins the beverage. With only about half of the cups working properly that effectively doubles the cost making it anything but a bargain. I will not purchase again and, if asked, will recommend against purchases
Azure OpenAI
Databricks SQL
2
Structured meaning derived from text
{"entities": [
{
"entity_name": "k-cups",
"entity_type": "product",
"entity_sentiment": "NEGATIVE",
"followup": "Y",
"followup_reason": "Defect in 50% of the cups"
},
...
...
{
"entity_name": "price",
"entity_type": "attribute",
"entity_sentiment": "NEGATIVE",
"followup": "N",
"followup_reason": ""
}
]}
We are sorry to hear that you experienced a defect in 50% of the Brooklyn Light Roast Bean Roastery Decaffeinated Coffee, Breakfast Blend, Single Serve Cup for Keurig K-Cup Brewers, 36-Count. We understand how frustrating this can be and we apologize for any inconvenience caused. We take quality control very seriously and we are currently investigating this issue to ensure that it does not happen again in the future.
In the meantime, we would like to offer you some alternative products that we believe you may enjoy. If you are looking for a decaffeinated coffee, we recommend trying the Green Mountain Coffee Roasters Decaf Breakfast Blend, which is also available in K-Cup format. Alternatively, if you prefer a different roast or flavor profile, we have a wide range of options available on our website that we would be happy to recommend.
Once again, we apologize for any inconvenience caused and we hope that you will give us the opportunity to make it right. If you have any further questions or concerns, please do not hesitate to contact us
3
Suggested message to customer

Thank you