databricks

Generative AI / LLMs using Databricks Data Intelligence Platform

10/01/2024

Dr. Caio Moreno | Solution Architect @ Databricks

About me

Dr. Caio Moreno

Solutions Architect @ Databricks

Please connect with me on LinkedIn!

Agenda

- Databricks vision
- Databricks Data Intelligence Platform
- Generative AI
- MIT Report
- Next steps

Databricks vision

6000+
global employees

$1.5B+
in revenue

$4B
in investment

Inventor of the **lakehouse**
&
Pioneer of **generative AI**

databricks

The data and AI company

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

Creator of

DELTA LAKE

mlflow™

APACHE Spark™

The winners in every industry will be data + AI companies

jetBlue®

Uses AI to model real-time digital twins of every flight

Delivers unified decision making across operations, maintenance, and customer service

intuit®

Simplifies the loan application experience with AI

$1.4B in loans facilitated to 40,000 small businesses through personalized applications

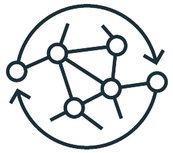
AT&T

Protects 182M wireless subscribers and 15M broadband households with AI

Predictive applications stop 80% of fraud in real-time before it can happen

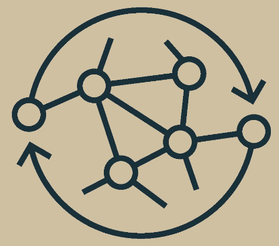
Data Lake

Machine Learning



Streaming

Generative AI

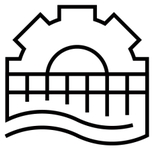


Data Science

Most organizations struggle to realize this vision

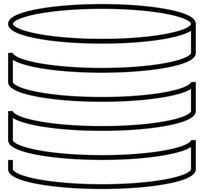
Governance

Orchestration & ETL



BI

Data Warehouse



Data Lake

Machine Learning

Streaming

Generative AI

Data S

Data, AI, and governance are siloed

Data privacy & control are challenged by AI

Dependent on highly technical staff

Governance

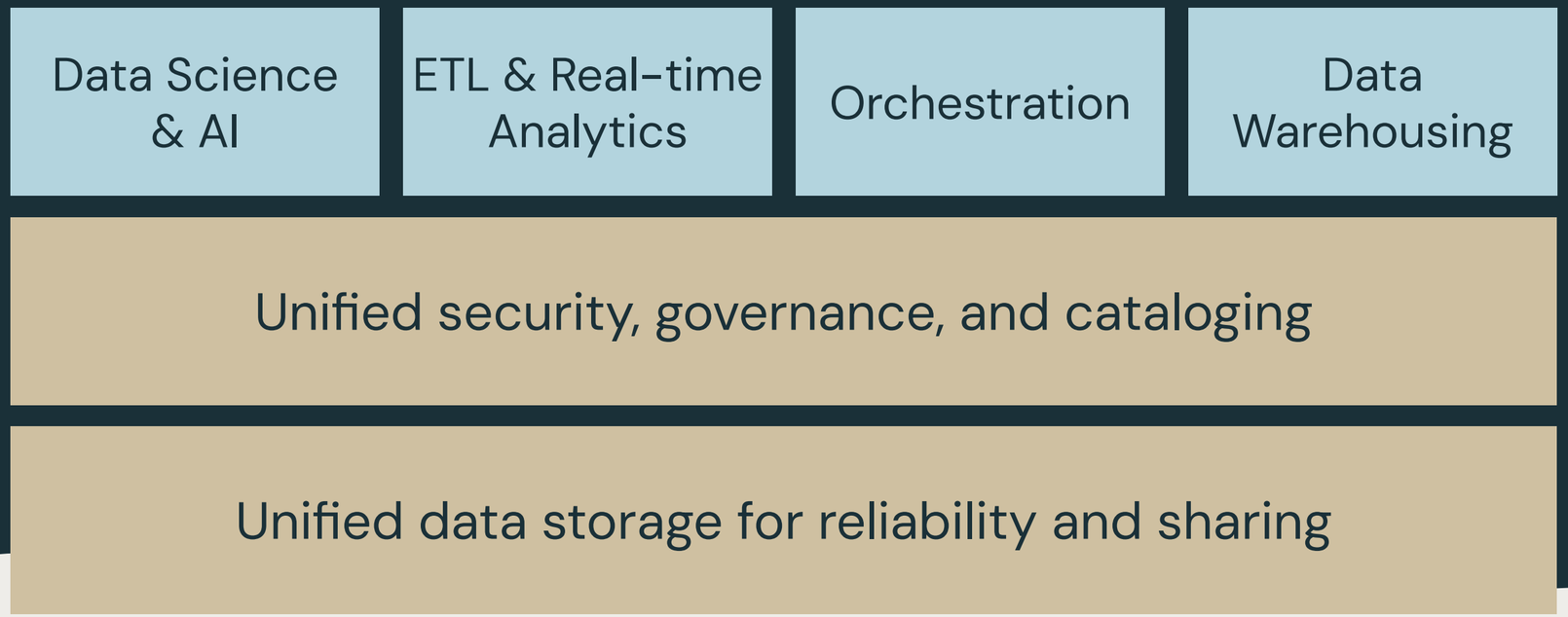
Orchestration & ETL

BI

Warehouse

The Data Lakehouse

An open, unified foundation for all your data



Open Data Lake

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

The Data Lakehouse

An open, unified foundation for all your data

Data Science & AI	ETL & Real-time Analytics	Orchestration	Data Warehousing
Databricks AI	Delta Live Tables	Workflows	Databricks SQL

Unified security, governance, and cataloging

Unity Catalog

Unified data storage for reliability and sharing

Delta Lake

Open Data Lake

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

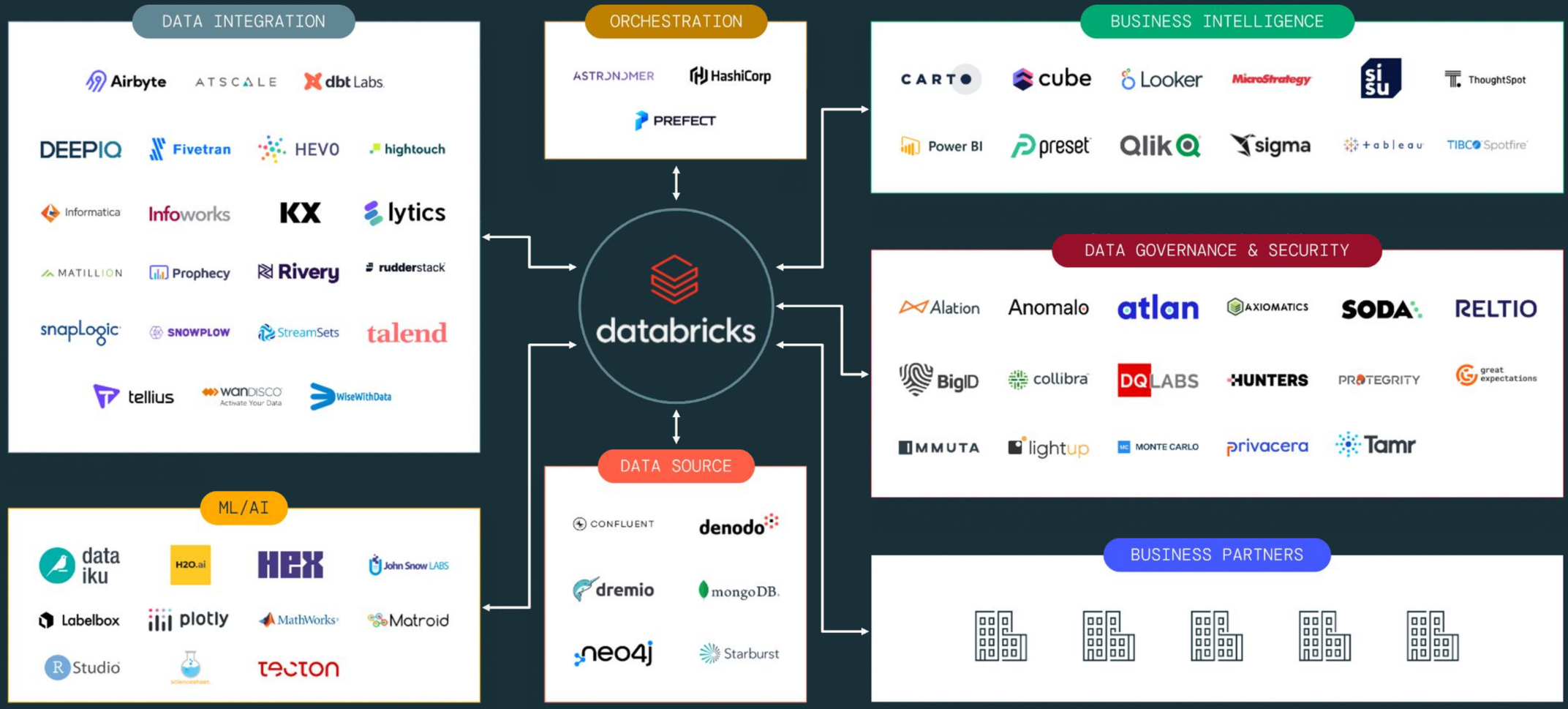
2020
Databricks pioneered the lakehouse architecture

Today
74% of global enterprises have adopted lakehouse

MIT Technology Review Insights, 2023

Built on an open foundation

Easily integrate with the entire data and AI ecosystem



Data Lakehouse

An open, unified foundation
for all your data

+

Generative AI

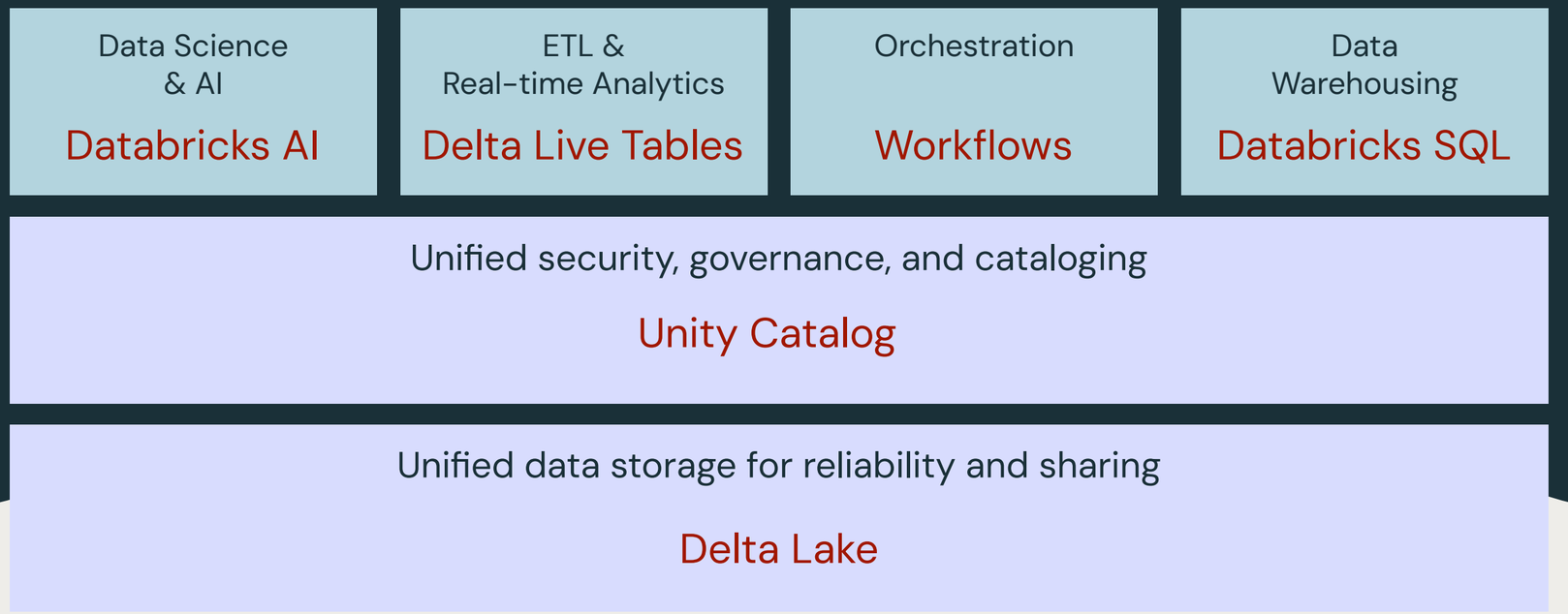
Easily scale and use data and AI

=

Data Intelligence Platform

Democratize data + AI across
your entire organization

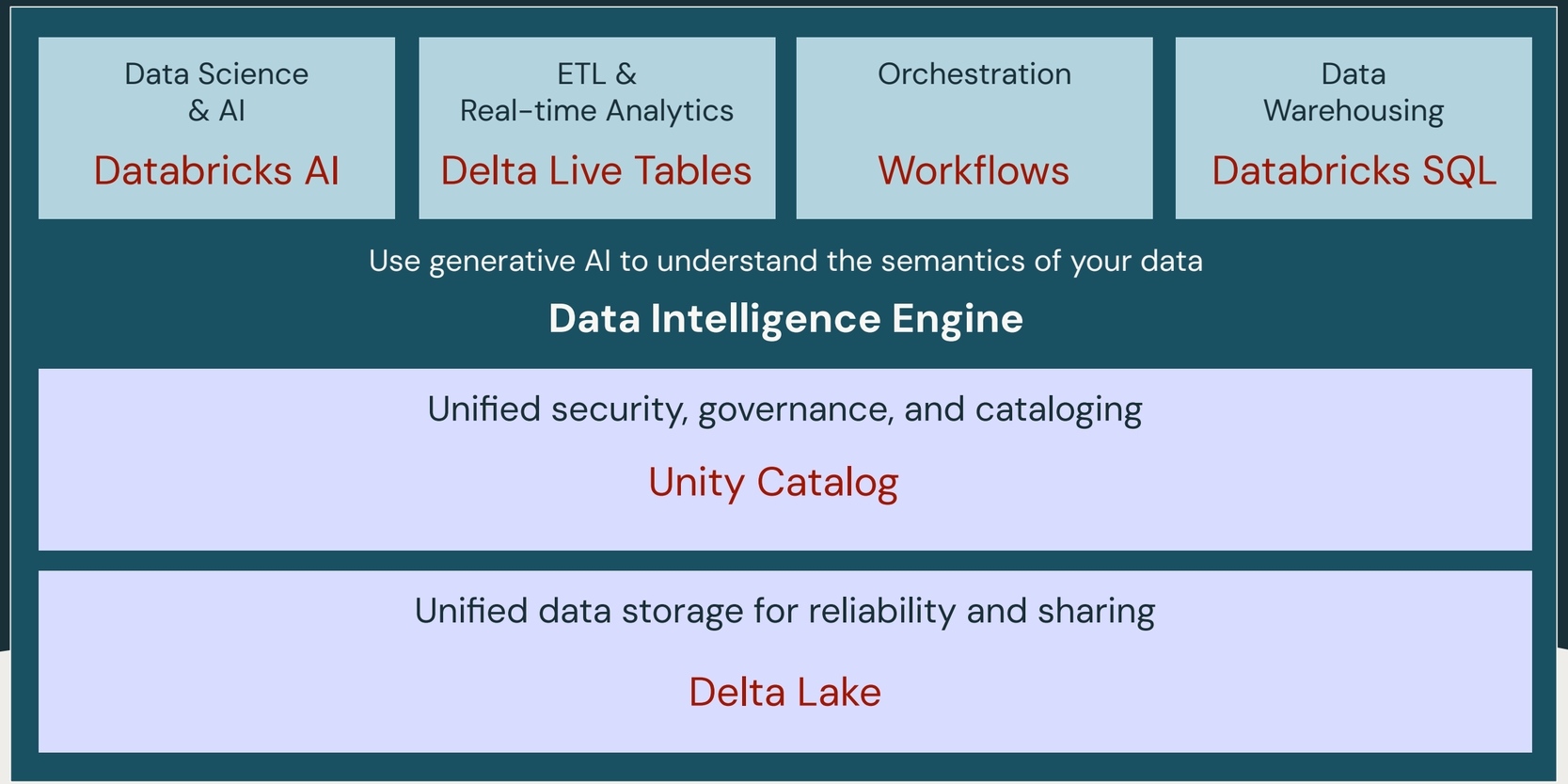
Databricks Data Intelligence Platform



Open Data Lake

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

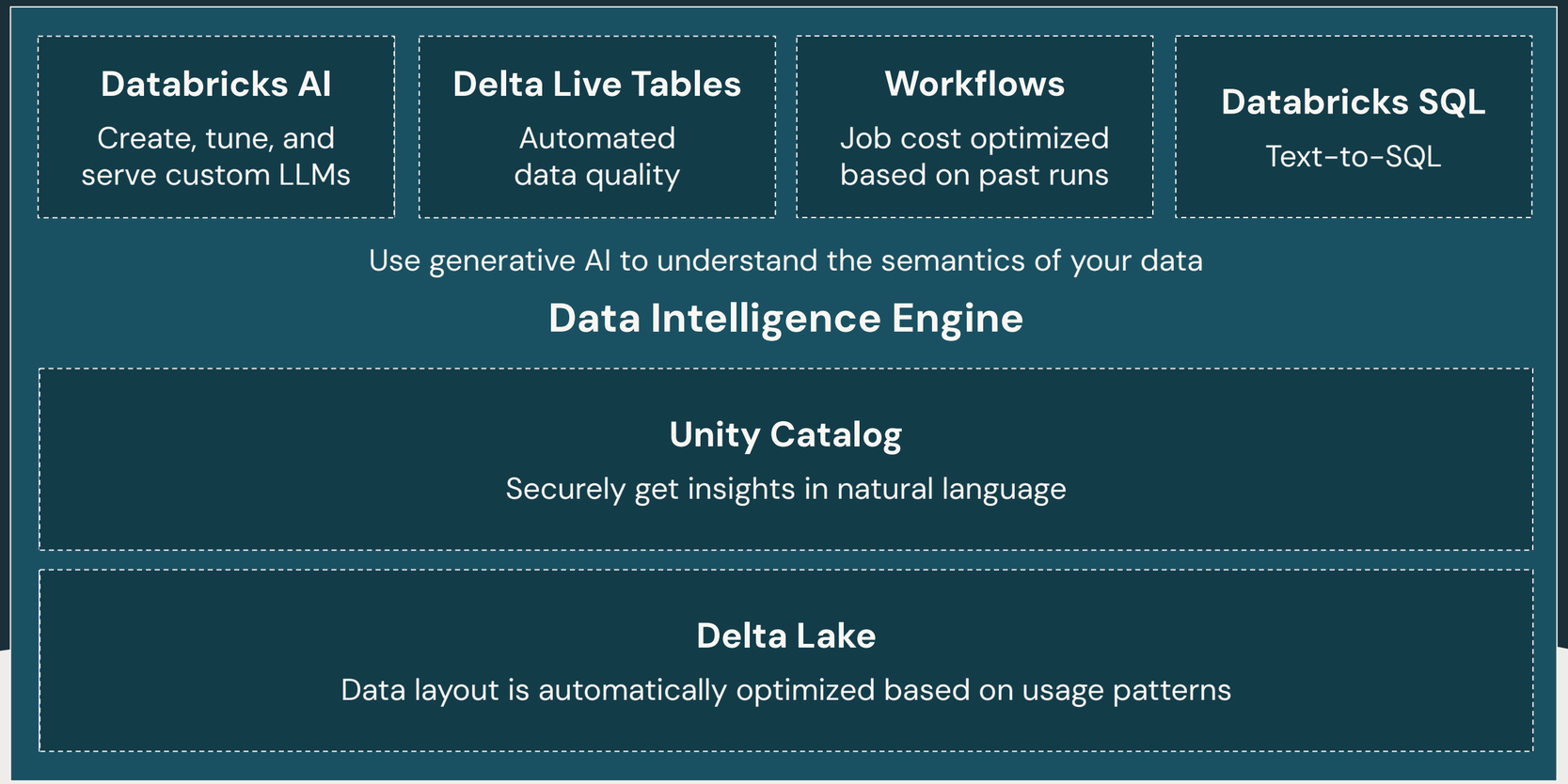
Databricks Data Intelligence Platform



Open Data Lake

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

Databricks Data Intelligence Platform



Open Data Lake

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

Databricks Data Intelligence Platform

Databricks AI

Gen AI

- Custom models
- Model serving
- RAG

End-to-end AI

- MLOps (MLflow)
- AutoML
- Monitoring
- Governance

Databricks AI
Create, tune, and serve custom LLMs

Delta Live Tables
Automated data quality

Workflows
Job cost optimized based on past runs

Databricks SQL
Text-to-SQL

Use generative AI to understand the semantics of your data

Data Intelligence Engine

Unity Catalog
Securely get insights in natural language

Delta Lake
Data layout is automatically optimized based on usage patterns

Open Data Lake

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

Databricks Data Intelligence Platform

Finance

Operations

Marketing

Customer Service

"Project Genie"

How [illegible] extend[illegible] AI to eve[illegible]he organi[illegible]n?

Data and AI for all with natural language

Databricks AI

Create, tune, and serve custom LLMs

Delta Live Tables

Automated data quality

Workflo

Job cost op based on pa

Use generative AI to understand the semantic

Data Intelligence Engin

Unity Catalog

Securely get insights in natural langu

Delta Lake

Data layout is automatically optimized based on

Open Data Lake

All Raw Data
(Logs, Texts, Audio, Video, Ima

A Data Intelligence Platform enables you to truly democratize data and AI

SIMPLE
Natural language provides ease of use and efficiency for all

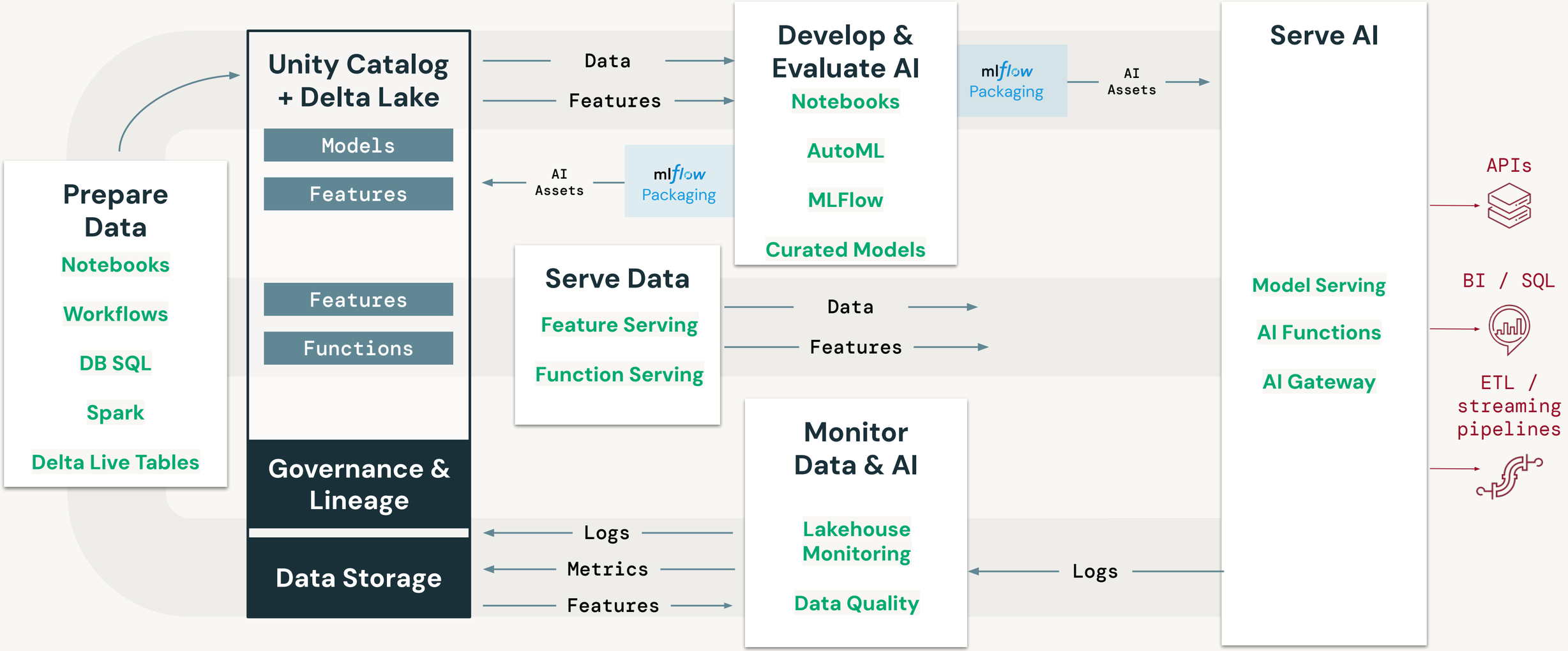
INTELLIGENT
AI is integrated end-to-end to uniquely understand your data

PRIVATE
Custom models are easily built on your private data

Databricks CEO present the Data Intelligence Platform at Microsoft Ignite 2023

https://www.youtube.com/watch?v=i6qgXunxuas

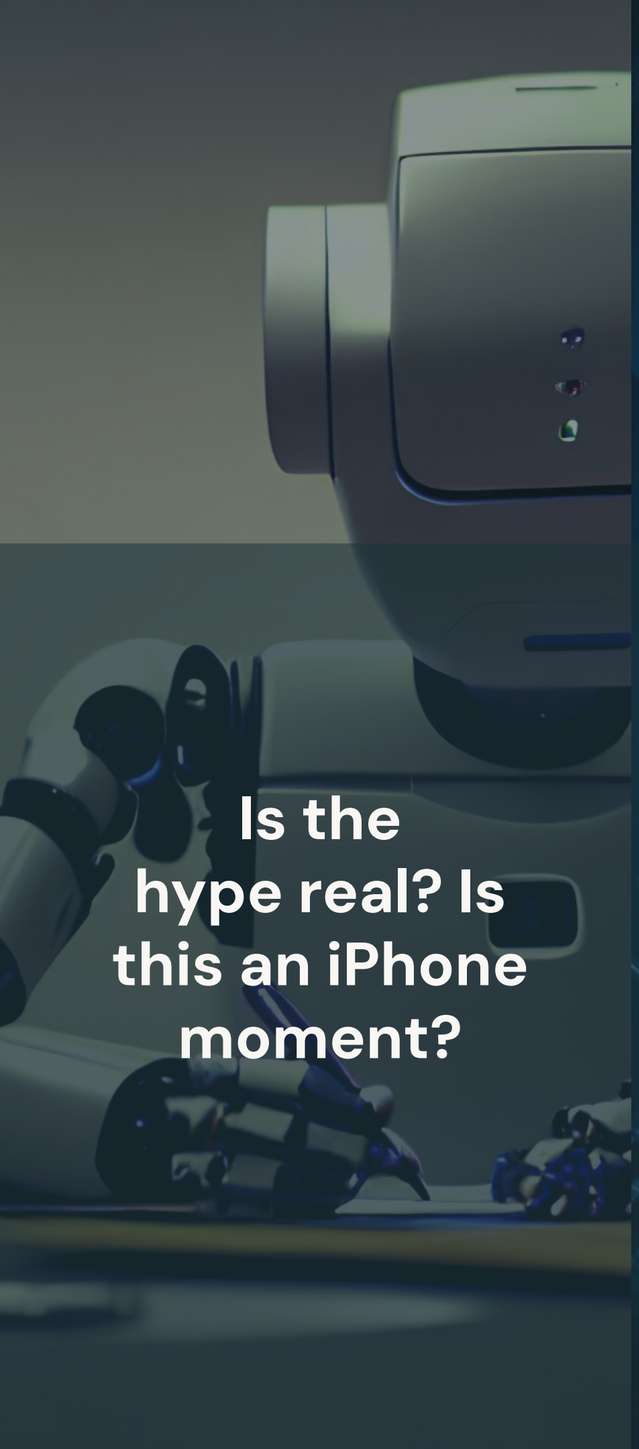
Lakehouse AI capabilities



Unity Catalog ❤️ AI

Generative AI

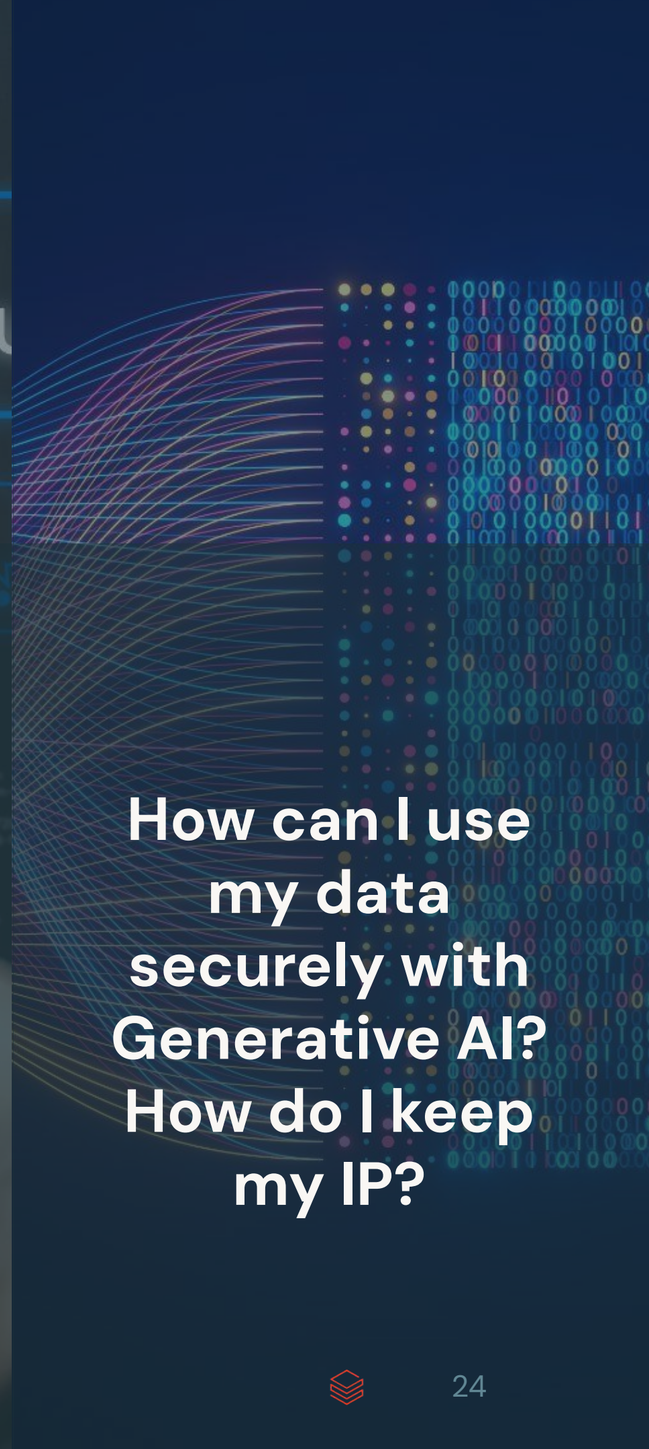
The AI Revolution



Is the hype real? Is this an iPhone moment?

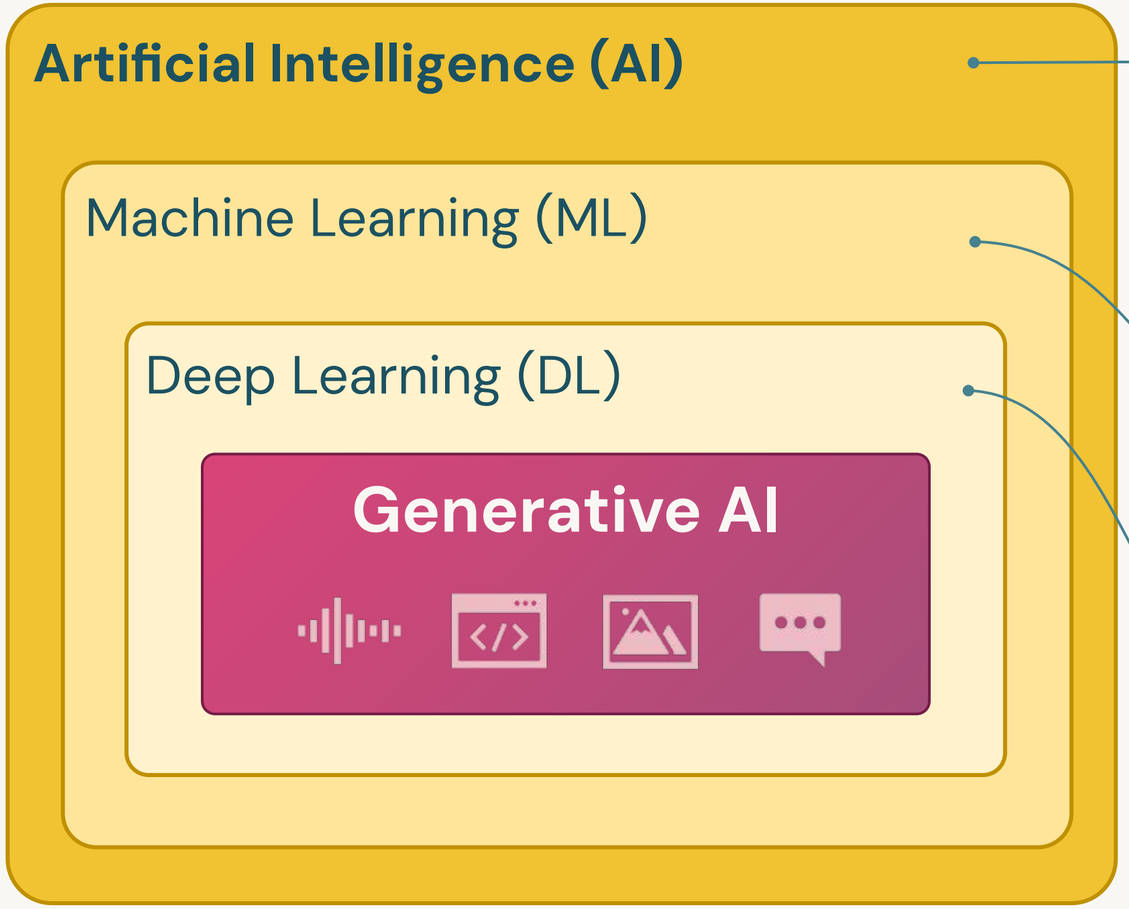
Is Generative AI a threat to or an opportunity for my business?

How exactly can I use Generative AI to gain a competitive advantage?



How can I use my data securely with Generative AI? How do I keep my IP?

What is Generative AI?



Artificial Intelligence:

A multidisciplinary field of computer science that aims to create systems capable of emulating and surpassing human-level intelligence.

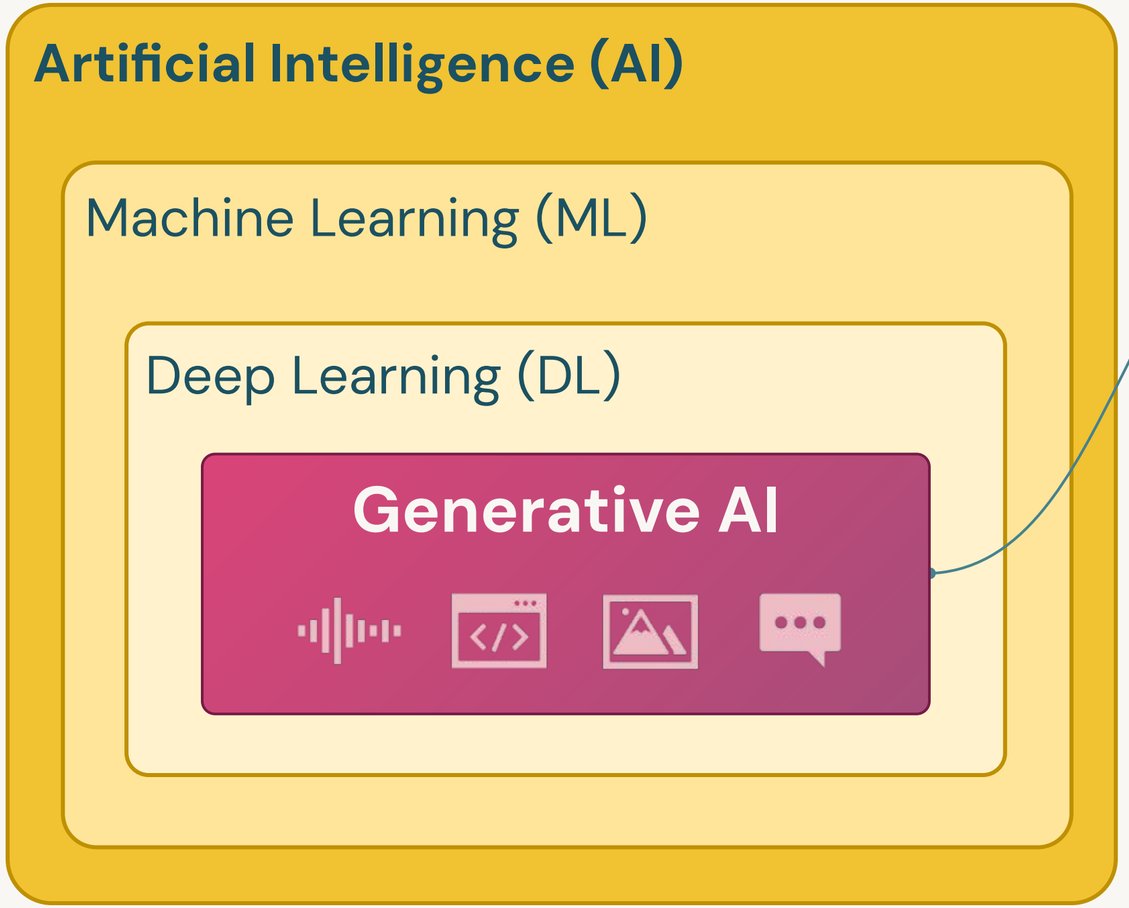
Machine Learning:

Learn from existing data and make predictions/prediction without being explicitly programmed.

Deep Learning:

Uses “artificial neural networks” to learn from data.

What is Generative AI?



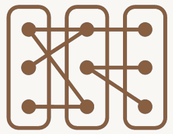
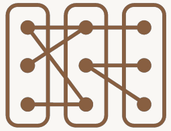
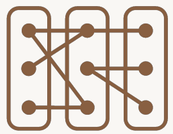
Generative Artificial Intelligence:

Sub-field of AI that focuses on **generating** new content such as:

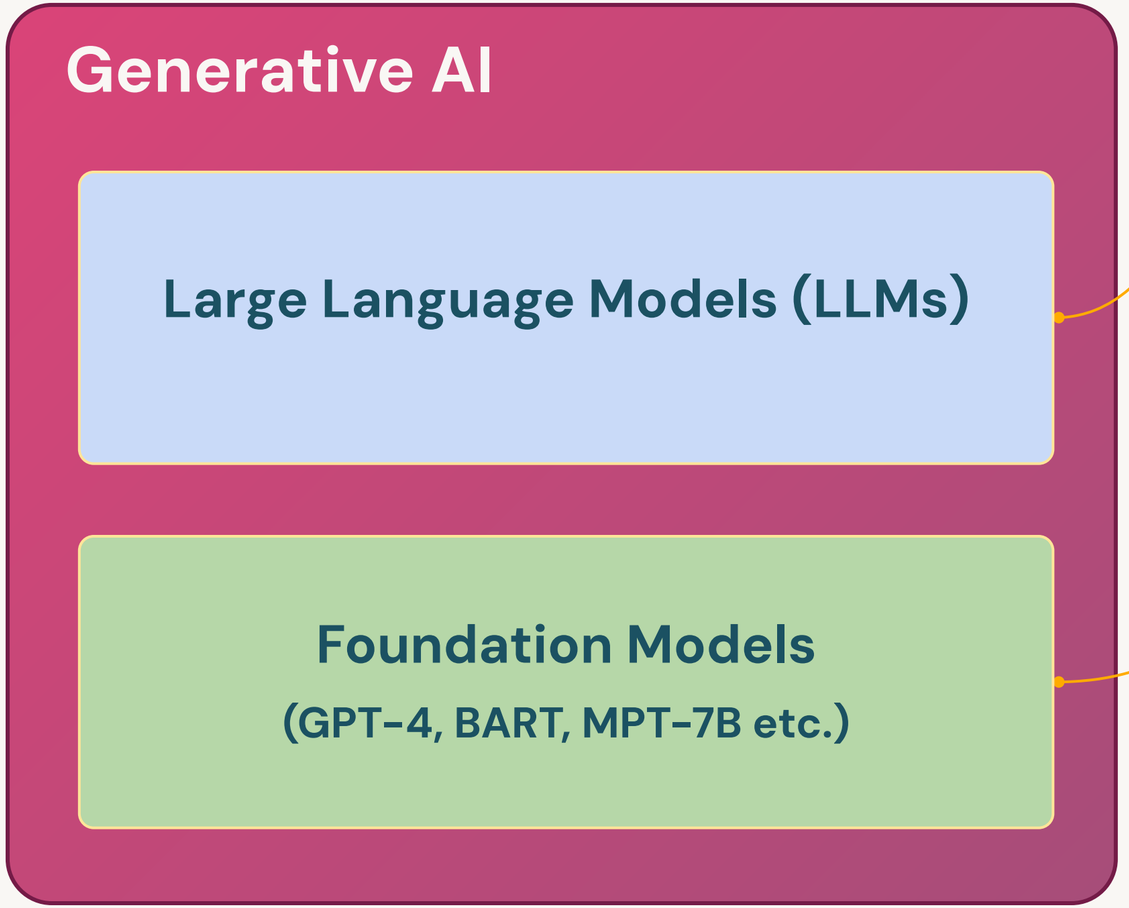
- Images
- Text
- Audio/music
- Video
- Code
- 3D objects
- Synthetic data

Generative Models: what are they?

A branch of ML modeling (LLM) which mathematically approximates the world

Data objects	Deep Neural Network	Tasks
	 [0.5, 1.4, -1.3, ...]	• Synthetic image generation • Style transfer / edit
	 [0.8, 1.4, -2.3, ...]	• Translation • Question Answering • Semantic search
	 [1.8, 0.4, -1.5, ...]	• Speech-to-text • Music transcription

What is a LLM?



Large Language Model (LLM):

Model trained on massive datasets to achieve advanced language processing capabilities

Based on deep learning neural networks

Foundation Model:

Large ML model trained on vast amount of data & fine-tuned for more specific language understanding and generation tasks

Why Now?

Factors making Generative AI possible now

Large Datasets

- Availability of large and diverse datasets
- AI models learn patterns, correlations, and characteristics of large datasets
- Pre-trained state-of-the-art models

Why Now?

Factors making Generative AI possible now

Large Datasets

- Availability of large and diverse datasets
- AI models learn patterns, correlations, and characteristics of large datasets
- Pre-trained state-of-the-art models

Computational Power

- Advancements in hardware; GPUs
- Access to cloud computing
- Open-source software, Hugging Face

Why Now?

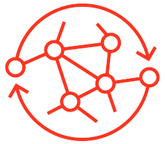
Factors making Generative AI possible now

Large Datasets

- Availability of large and diverse datasets
- AI models learn patterns, correlations, and characteristics of large datasets
- Pre-trained state-of-the-art models

Computational Power

- Advancements in hardware; GPUs
- Access to cloud computing
- Open-source software, Hugging Face



Innovative DL Models

- Generative Adversarial Networks (GANs)
- Transformers Architecture
- Reinforcement learning from human feedback (RLHF)

Why should I care now?

ML/AI has been around for a while, why it matters now

Generative AI models' accuracy and effectiveness have hit a tipping point

- Powerful enough to enable use cases not feasible even a year ago
- Economical enough for use even by non-technical business users

Generative AI models and tooling are readily available

- Many models are open source and customizable
- Requires powerful GPUs, but are available in the cloud

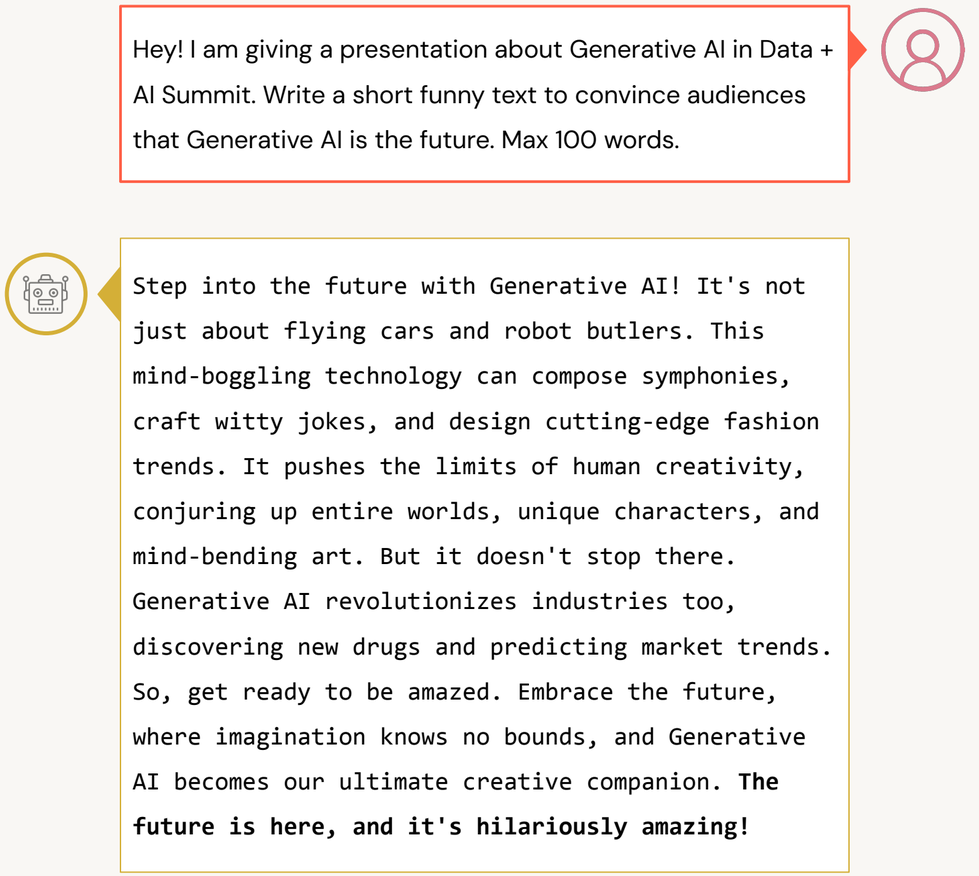
"Chegg shares drop more than 40% after company says ChatGPT is killing its business"

05/02/2023
Source: CNBC

Generative AI Use Cases

Intelligent conversations, creative text creation, code generation

- Content generation
- Question/answers
- Virtual assistants
- Content personalization
- Language style transfer
- Story telling, poetry, creative writing
- Translation
- Code generation / auto-completion



LLMs Generate Outputs for NLP Tasks

Common LLM tasks

	**Content Creation and Augmentation**	Generating coherent and contextually relevant text. LLMs excel at tasks like text completion, creative writing, story generation, and dialogue generation.
	**Summarization**	Summarizing long documents or articles into concise summaries. LLMs provide an efficient way to extract key information from large volumes of text.
	**Question Answering**	Comprehend questions and provide relevant answers by extracting information from their pre-trained knowledge.
	**Machine Translation**	Automatically converting a text from one language to another. LLMs are also capable to explain language structure such as grammatical rules.
	**Classification**	Categorizing text into predefined classes or topics. LLMs are useful for tasks like topic classification, spam detection, or sentiment analysis.
	**Named Entity Recognition (NER)**	Identifying and extracting named entities like names of persons, organizations, locations, dates, and more from text.
	**Tone / Level of content**	Adjusting the text's tone (professional, humorous, etc.) or complexity level (e.g., fourth-grade level).
	**Code generation**	Generating code in a specified programming language or converting code from one language to another.

Winners will use Gen AI to succeed

New use cases will affect platform decisions

Financial Services	Healthcare & Life Sciences	Comm, Media & Entertainment	Retail & Consumer Goods	Manufacturing	Public Sector
Fraud monitoring and predictions	Biomedical literature summarization & discovery	Hyper-personalization for customer experience (CX)	Try before you buy with virtual fitting rooms	Delightful, personalized customer experiences	Analysis of open-source Intelligence
Automating compliance data gathering	Clinical trial optimization	Enhancing customer support and self-service	Optimizing demand prediction and inventory	Increasing productivity and efficiency in operations	Modernizing legacy code bases
Accelerate underwriting and claims processing in insurance	Health insurance claim processing	Intelligent content creation and curation	Generate innovative product designs	Prescriptive and proactive field service	Regulatory compliance assistance

Customers are accelerating adoption of GenAI with Databricks

replit

replit is an online integrated development environment

Challenge

They want to use LLMs in their product to assist developers

Training LLMs are prohibitively expensive and error prone

Solution

replit built Ghostwriter, code-generation models, from scratch by training a 2.7B parameter LLM from MosaicML

Impact

3 days

to train LLM (versus weeks/months)

1 day

Raw data to model deployed in production

Lower costs

jetBlue®

Challenge

Lack of online feature store hydration, inability to scale quickly, and high latency in its cloud data warehouse, JetBlue's data scientists were prevented from constructing scalable ML training and inference pipelines, hindering its ability to provide seamless customer experiences

Solution

With Databricks Lakehouse, JetBlue is using LLMs built upon their own data to deliver better passenger experiences. Leveraging real-time streams of weather, IoT, and FAA data, JetBlue now operates the world's first digital-twin for efficient and safe operations, significantly minimizing delays

Impact

Increased

Innovation in LLMs and generative AI, powering safer operations

databricks

How Do LLMs Work?

A simplified version of LLM training process

Input

Books

Wikipedia

Scientific Research

Crawled data from the Internet

Custom Curated Datasets ...

Tokenize
(Encode text into numeric rep.)

```
Dolly is a language model that can help
you generate text for a variety of tasks.
```

```
Dolly is a language model that can help you
 generate text for a variety of tasks.
```

```
[35, 5098, 318, 257, 3303, 2746, 326, 460,
1037, 345, 7716, 2420, 329, 257, 4996, 286,
8861, 13]
```

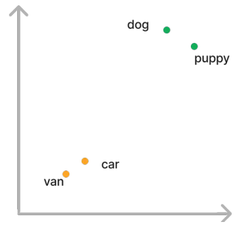
Tokens: 18, Characters: 81
(100 tokens ~= 75 words)

Token Embeddings
(Put words with similar meaning close in vector space)

Embedding Functions
(Pre-trained model)

[0.2, 1.5, 0.6 .... 0.6]

When done well, similar words will be closer in these embedding/vector spaces. Example 2D representation;



Encoding

Human Feedback

Pre-Trained Transformer Model

Billions of parameters

Output Text

[4.2, 1.2, -1.9, …]

This | is | ... | ...

Predicted next word

Decoding

LLM Flavors

Thinking of building your own modern LLM application?

Open-Source Models

- Use as **off-the-shelf** or **fine-tune**
- Provides flexibility for customizations
- Can be smaller in size to save cost
- **Commercial / Non-commercial use**

Open-source LLMs:

Commercial Use

Meta AI LLaMA 2 | databricks Dolly | mosaicML MPT

Proprietary Models

- Usually offered as **LLMs-as-a-service**
- Some can be **fine-tuned**
- Restrictive licenses for usage and modification

Proprietary LLMs:

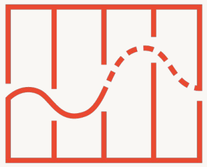
ANTHROP\C | OpenAI | ChatGPT | PaLM 2

Choose the right LLM model flavor

There is no “perfect” model, trade-offs are required.

LLM model decision criteria

Privacy



Quality

Cost

Latency

An Overview of Common LLMs

Open-source and Closed LLMs

Model or model family	Model size (# params)	License	Created by	Released	Notes
Falcon	7 B - 40 B	Apache 2.0	Technology Innovation Institute	2023	A newer potentially state-of-the-art model
MPT	**7 B**	**Apache 2.0**	**MosaicML**	**2023**	**Comes with various models for chat, writing etc.**
Dolly	**12 B**	**MIT**	**Databricks**	**2023**	**Instruction-tuned Pythia model**
Pythia	19 M - 12 B	Apache 2.0	EleutherAI	2023	Series of 8 models for comparisons across sizes
GPT-3.5	175 B	proprietary	OpenAI	2022	ChatGPT model option; related models GPT-1/2/3/4
BLOOM	560 M - 176 B	RAIL v1.0	BigScience	2022	46 languages
FLAN-T5	80 M - 540 B	Apache 2.0	Google	2021	methods to improve training for existing architectures
BART	139 M - 406 M	Apache 2.0	Meta	2019	derived from BERT, GPT, others
BERT	109 M - 335 M	Apache 2.0	Google	2018	early breakthrough

For up-to-date list of recommended LLMs : https://www.databricks.com/product/machine-learning/large-language-models-oss-guidance

Please note: Databricks does not endorse any of these models - you should evaluate these if they meet your needs.

MIT Report

MIT Report

In this report, you will learn:

- How to establish a data, analytics and AI infrastructure that is efficient, scalable, well-governed and future-proof
- Strategies for striking a balance between leveraging third-party capabilities and developing in-house models
- How to choose between open source and proprietary technologies
- Techniques for identifying suitable use cases, delivering tangible business value, and fostering trust in AI-driven solutions
- How to proactively prepare your organization for an exciting yet uncertain future

https://www.databricks.com/resources/ebook/mit-cio-generative-ai-report

MIT Report

LLMs and the opportunity

"LLMs now have the capability to achieve the necessary accuracy, and at a faster pace."

Andrew Blyton, Vice President and Chief Information Officer, DuPont Water & Protection

"Generative AI evolves the possibility and promise of AI exponentially. You can transform the conversation between the creator and the computer."

Cynthia Stoddard, Senior Vice President and Chief Information Officer, Adobe

"In the next five to ten years we will see how quickly we can adapt, and companies that fail to adapt, no matter how big, are going to disappear."

Noriko Rzonca, Chief Digital Officer, Cosmo Energy Holdings

MIT Report

Data & AI Unified Platform

"We have aggregated data across a lot of different technologies over time, and I think what we're finding now is that the lakehouse has the best cost performance straight off."

Andrew Blyton, Vice President and Chief Information Officer, DuPont Water & Protection

"Typical databases are designed for only one type of data. Lakehouse allows us to move much quicker."

Owen O'Connell, Senior Vice President and Chief Information Officer (Information Digital Services and Operations), Shell

"We are seeing the need to have very integrated governance models, integrated governance structures for all data and all models."

Richard Spencer Schaefer, Chief Health Informatics Officer, Kansas City VA Medical Center

MIT Report

LLMs: Buy, build? Open, closed?

"If your entire business model is based on the IP you own, protection is everything."

Andrew Blyton, Vice President and Chief Information Officer, DuPont Water & Protection

"All the large models that you can get from third-party providers are trained on data from the web. But within your organization, you have a lot of internal concepts and data that these models won't know about."

Matei Zaharia, Cofounder and Chief Technology Officer, Databricks, and Associate Professor of Computer Science, University of California, Berkeley

"If you care deeply about a particular problem or you're going to build a system that is very core for your business, it's a question of who owns your IP."

Michael Carbin, Associate Professor, MIT, and Founding Advisor, MosaicML

Next steps

Build better GenAI applications on Databricks

You need more than just good models

Complete control

Complete ownership over models and data

Production quality

Faster, more reliable deployment across multiple use cases

Lower cost

Cost-effective to build LLMs at scale

Start Building Your Generative AI Solution

Databricks is the only provider that enables every architectural pattern

Prompt engineering

Crafting specialized prompts to guide LLM behavior

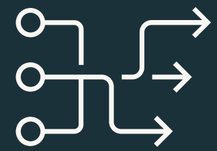
Retrieval Augmented Generation (RAG)

Combining an LLM with enterprise data

Fine-tuning

Adapting a pre-trained LLM to specific datasets or domains

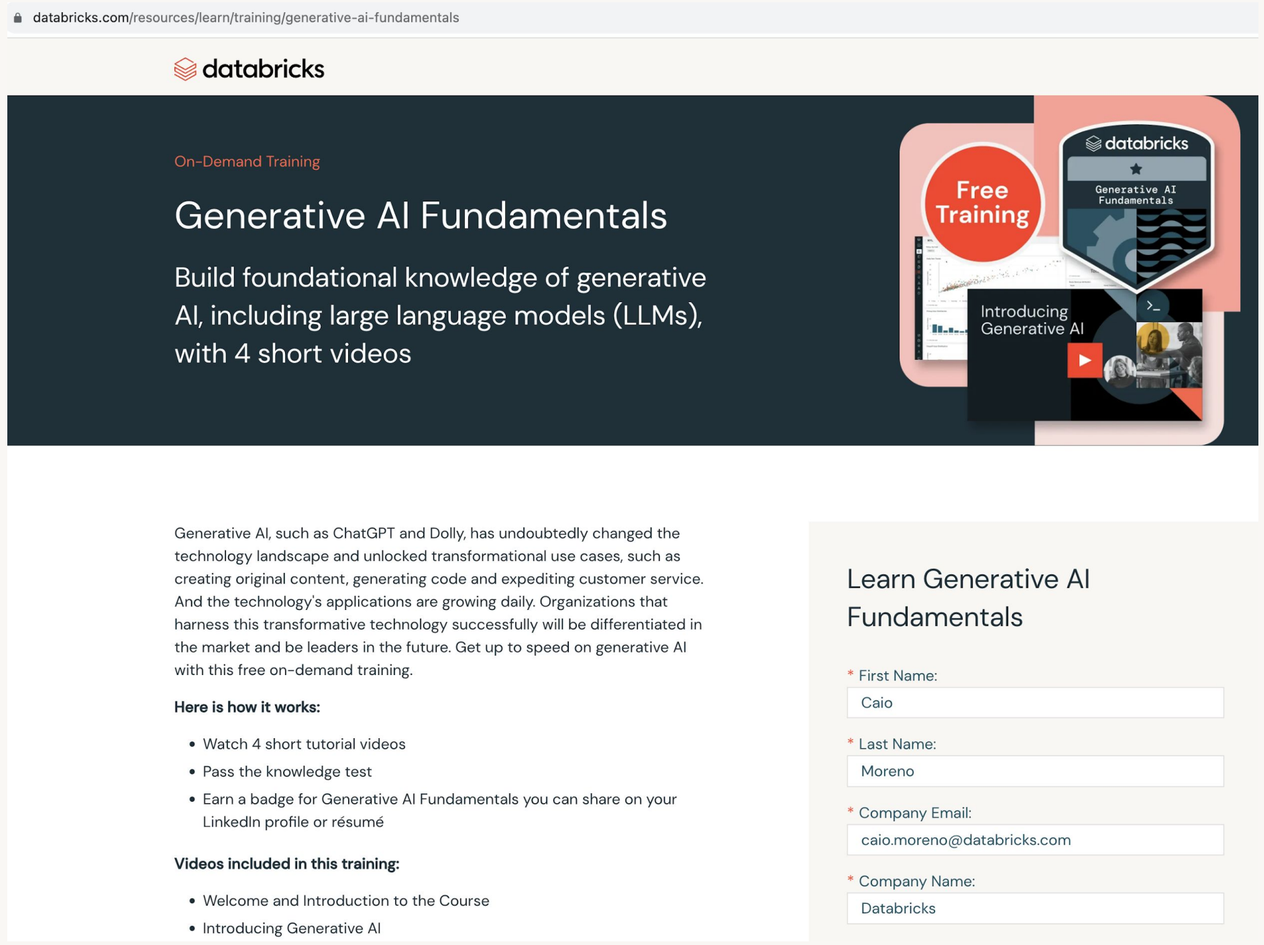
Pre-training



Training an LLM from scratch

Complexity/compute-intensiveness

Get Certified today!



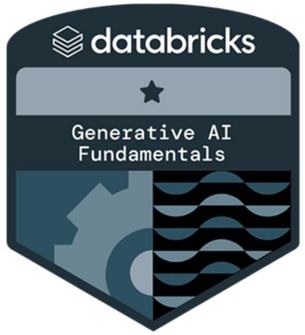
Here is how it works:

- Watch 4 short tutorial videos
- Pass the knowledge test
- Earn a badge for Generative AI Fundamentals you can share on your LinkedIn profile or résumé

Videos included in this training:

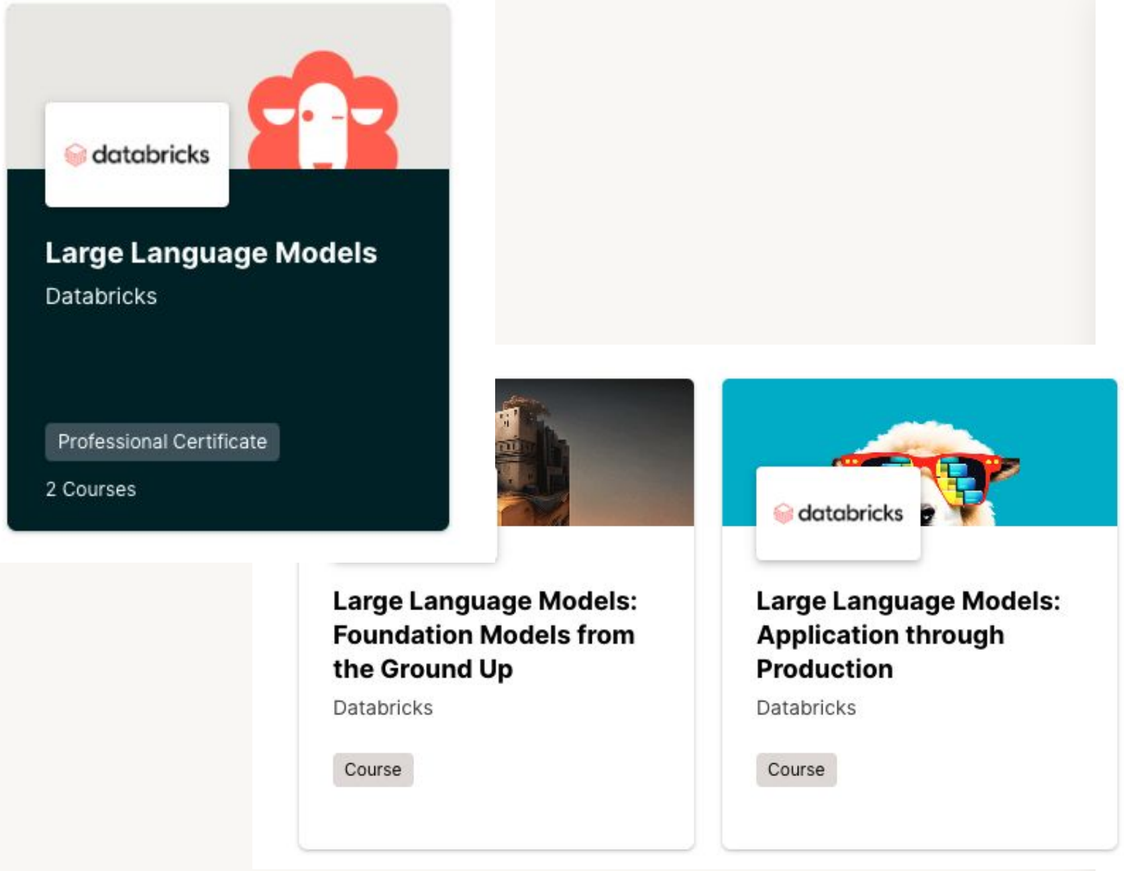
- Welcome and Introduction to the Course
- Introducing Generative AI
- Finding Success With Generative AI
- Assessing Potential Risks and Challenges

Earn your badge today and share your accomplishment on LinkedIn or résumé.



https://www.databricks.com/resources/learn/training/generative-ai-fundamentals

Expert-Led Large Language Models Courses on edX



- The first course, *LLMs: Application through Production* is aimed at developers, data scientists, and engineers looking to build LLM-centric applications with the latest and most popular frameworks.
- The second course, *LLMs: Foundation Models from the Ground Up* is aimed at data scientists interested in diving into the details of foundation models and the key innovations that led to the proliferation of transformer-based models.

Learn more about Gen AI on Databricks

- Explore our HowTo blog posts
- Train a Hugging Face model on DB
- Tune your own copy of Dolly on DB
- Hear and provide feedback on our LLM roadmap during DAIS
- Get access to our previews

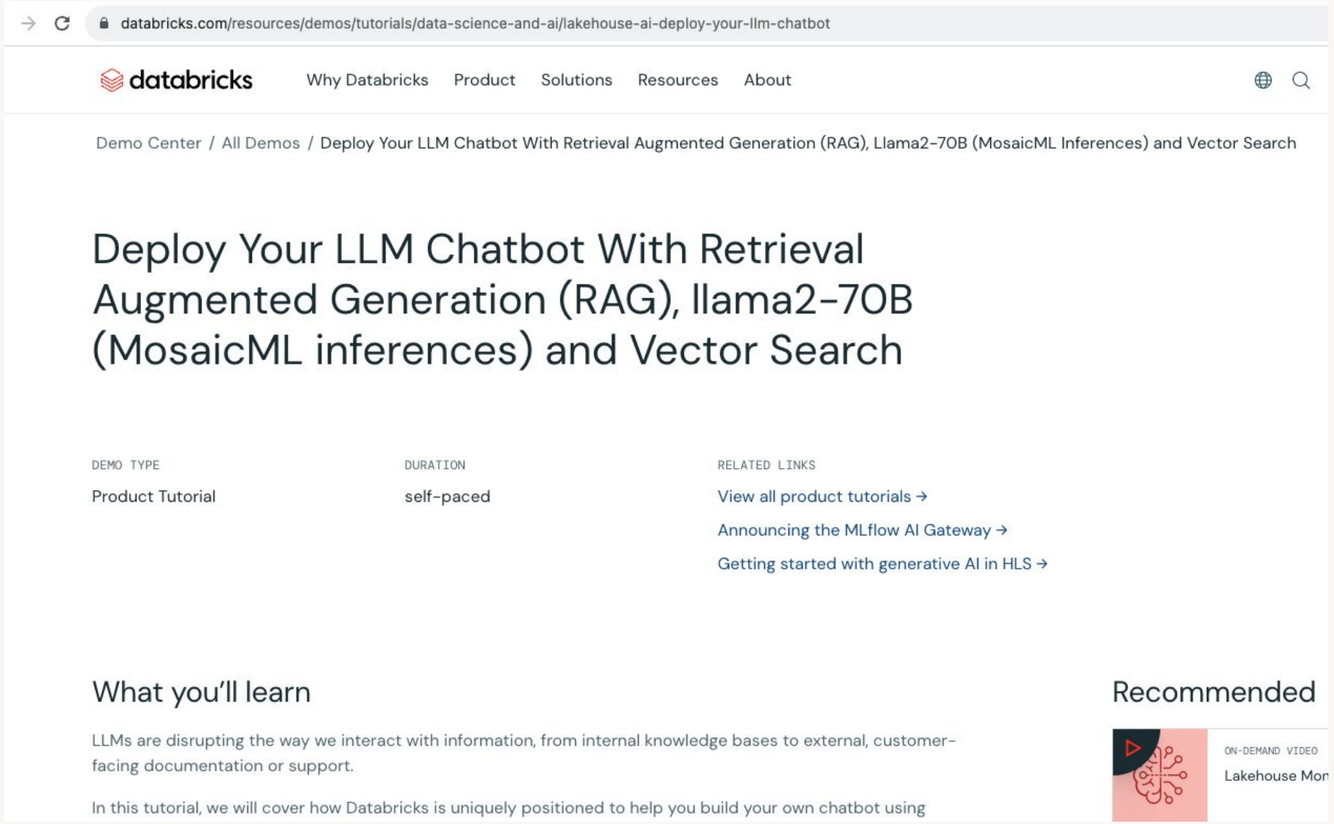
Explore our blogs

- Hello Dolly: Democratizing the magic of ChatGPT with open models
- Getting started with NLP using Hugging Face transformers pipelines
- How Outreach Productionizes PyTorch-based Hugging Face Transformers for NLP
- Fine-Tuning Large Language Models with Hugging Face and DeepSpeed

Databricks Webinars

- Build Your Own Large Language Model Like Dolly
  - How to use off-the-shelf pretrained models with tools like Hugging Face and GPU resources
  - How to fine-tune a model on your data
  - How Dolly was built
- Life After ChatGPT: Large Language Models (LLMs) at Scale on Databricks
  - Transformers & Attention: Why are they all you need
  - GPT Models, BERT Models, ChatGPT
  - Open source alternatives: Dolly, T5, FLAN-T5, Blenderbot, LLaMA
  - Fine tuning and deploying LLM applications on Databricks
- Scaling Large Language Models and Deep Learning on Databricks
  - Typical Deep Learning and LLM project lifecycle and Operations
  - Scaling simple Deep Learning training pipelines using SparkTorchDistributor
  - Fine-tuning LLMs: Using DeepSpeed for fine-tuning HuggingFace models
  - Using PyTorch FSDP for scaling LLM fine-tuning

Retrieval Augmented Generation (RAG) Demo



https://www.databricks.com/resources/demos/tutorials/data-science-and-ai/lakehouse-ai-deploy-your-llm-chatbot


databricks