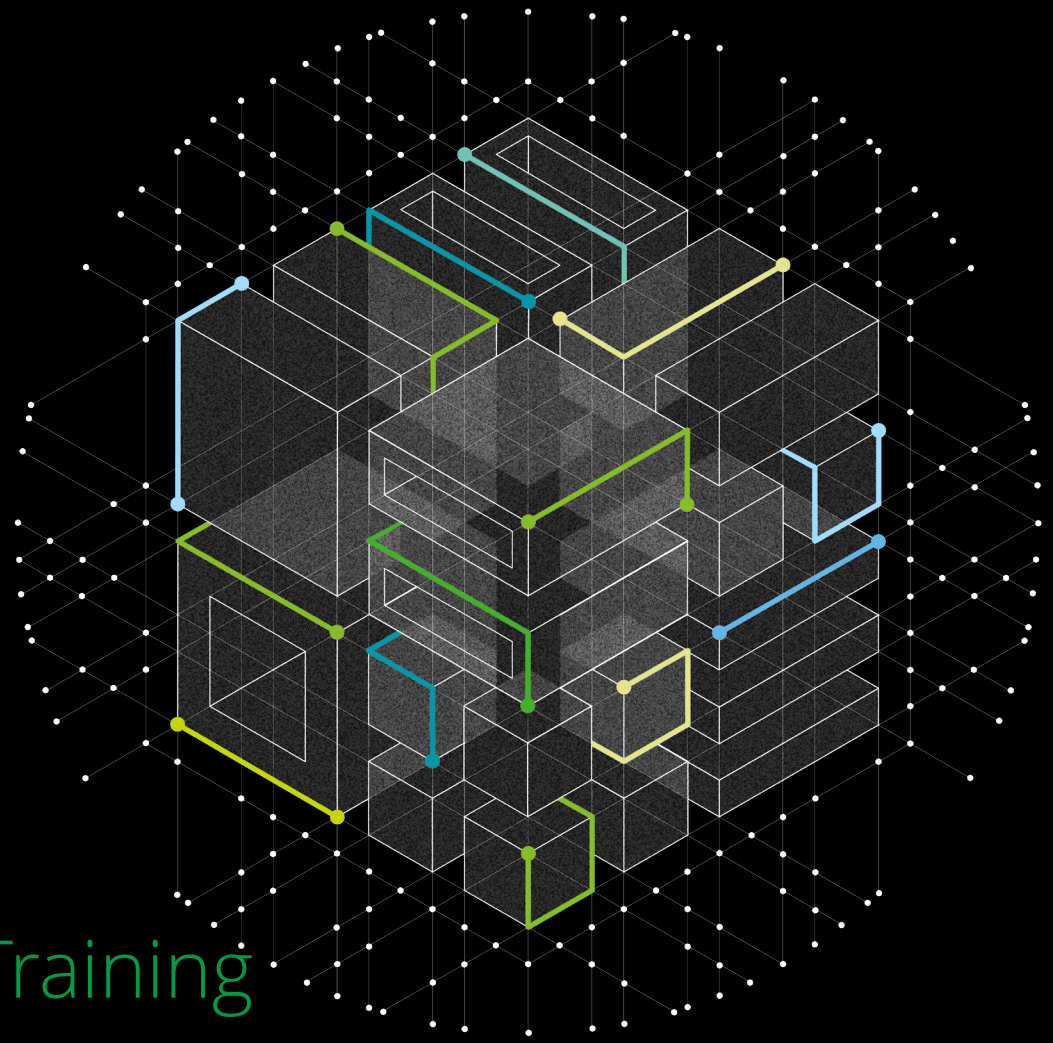
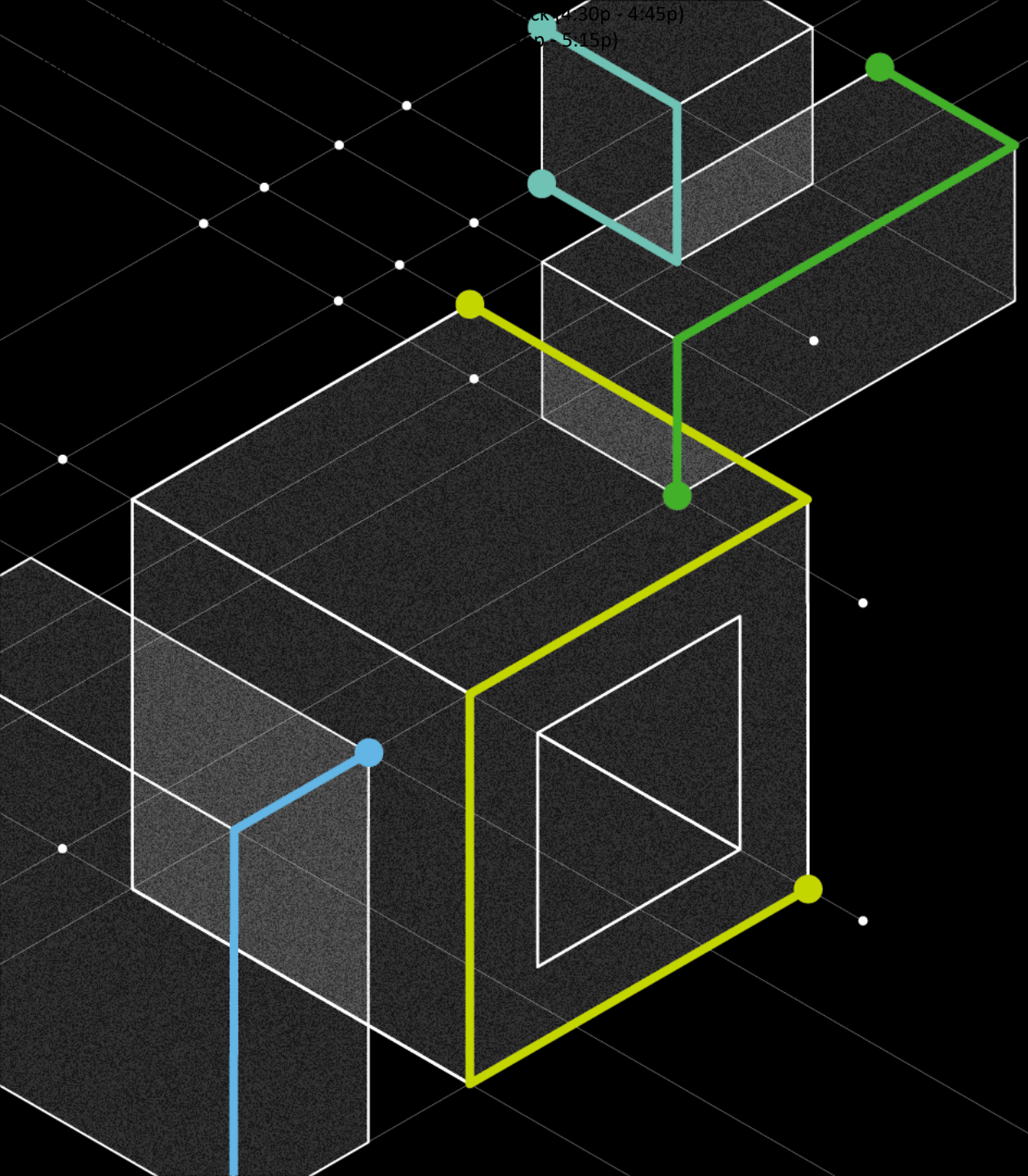
Deloitte.

databricks



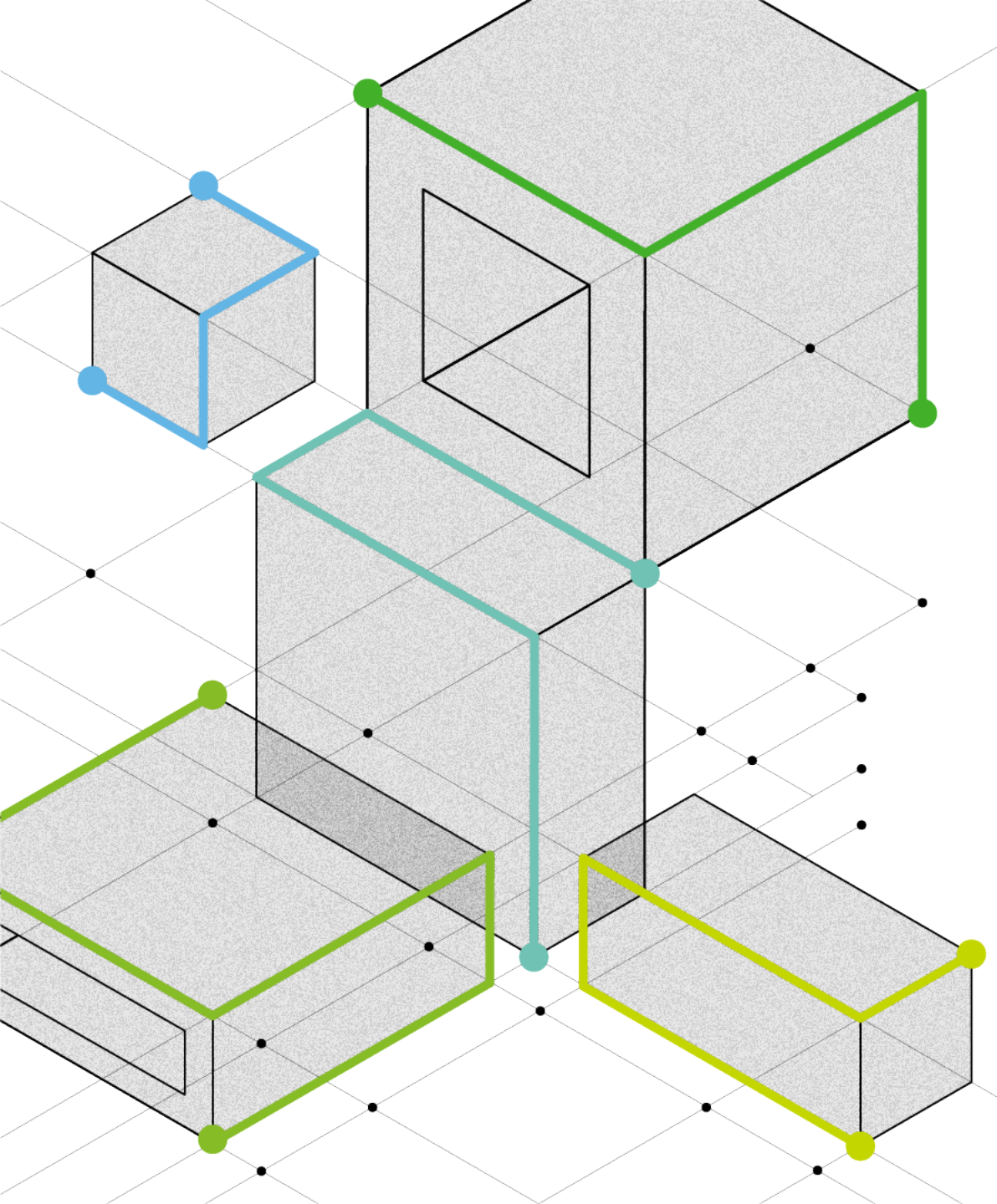
Databricks GenAI Training

Build LLM Apps in one shot with Databricks!
August 31, 2023



Agenda

- Introducing GenAI CoP & Databricks CoP (4:00p - 4:15p)
- Introducing Databricks LLM - Dolly, MPT-7B (4:15p - 4:30p)
- Introducing Databricks LLM Apps Architecture & Tech stack (4:30p - 4:45p)
- Hands on building your own LLM Apps in Databricks (4:45p - 5:15p)
- Wrap up and Q&A (5:15p - 5:30p)



Agenda

- Introducing GenAI CoP & Databricks CoP (4:00p - 4:15p)
- Introducing Databricks LLM - Dolly, MPT-7B (4:15p - 4:30p)
- Introducing Databricks LLM Apps Architecture & Tech stack (4:30p - 4:45p)
- Hands on building your own LLM Apps in Databricks (4:45p - 5:15p)
- Wrap up and Q&A (5:15p - 5:30p)

GenAI Community of Practice (COP): How to Get Involved?

Community of Practice (COP) Overview

What is the GenAI COP?

The GenAI COP fosters an active and engaged community through providing resources, informative communications, and exciting events to **promote GenAI awareness and adoption across the firm**. This consortium offers **GenAI training resources**, access to **virtual / in-person events**, as well as **GenAI resources and materials**

Get Involved

How do I sign up to the GenAI COP?

QR Code:*

1. Scan the QR Code to the right
2. Complete the GenAI COP Survey
3. The completed survey will add you to the GenAI Distro list

What to Expect

What will the GenAI COP offer to its members?

Newsletters:

- Highlights market trends and GenAI involvement opportunities

Training Opportunities:

- Participate in both virtual or in-person GenAI trainings

Proposal Toolkit:

- Collection of GenAI content and visuals intended for proposals

GenAI Project Opportunities:

- Connects COP members with GenAI project and initiative opportunities

***Survey Link & Shared Inbox:**

If the QR Code does not work, try the following survey link: https://deloittesurvey.deloitte.com/Community/se/3FC11B261535A929 or reach out directly to our shared inbox at USGenAIGPS@Deloitte.com

Databricks Alliance & COP

WHY DATABRICKS

LAKEHOUSE PLATFORM

- Cloud-based end-to-end data, analytics and insights platform providing flexibility, reliability, performance and security required for data lake/AI, data warehousing/SQL and data-mesh workloads and patterns
- Built for multi-cloud target state (PaaS on AWS, Azure, GCP)
- Comprehensive security and audit support, including native integration with cloud security and IAM providers. Data stays on customer's environment
- Native support for structured, semi-structured and unstructured datasets directly on top of cloud object stores to enable a single source of truth for all of today's enterprise data
- Native support for both batch and real-time analytics and insights

INDUSTRY SOLUTION ACCELERATORS

- Pre-built and fully-functional solutions that can be deployed with a click of a button to accelerate some of the most common and high-impact use cases

TRANSPARENT AND PAY-AS-YOU-GO PRICING

- Transparent (Databricks credits + hyper-scaler resources), elastic (auto-scale up/down) and pay-as-you-go pricing at per-second granularity

OUR EXPERIENCE

155 GLOBAL ENGAGEMENTS

to-date leading data modernization work leveraging Databricks

Global Elite Partner

Only 10 partners have this designation amongst a network of 8,000+ partners

OUR CAPABILITIES

- Lakehouse Implementation
- Legacy Data Warehouse Migration
- Hadoop Migration
- AI/ML
- Solution Accelerators

RESOURCES

- **Robust Community of Practice**
- **Assets & Accelerators**
- **Training & Certification**
- **Pursuit Support**

SECTORS

- FEDERAL HEALTH
- CIVILIAN
- DEFENSE, SECURITY & JUSTICE
- STATE & LOCAL

FY24 Initiatives

- Migration Factory
- Data Mesh
- Cyber Security
- Trustworthy AI

OUR TALENT

120 Databricks-trained/experienced GPS practitioners

14 certified GPS practitioners *(on track to badge 150+ within next 12 mo.)*

2 GPS Champions

5K+ Spark trained/certified practitioners

ALLIANCE TEAM

Dave Thomas
Lead Alliance Partner

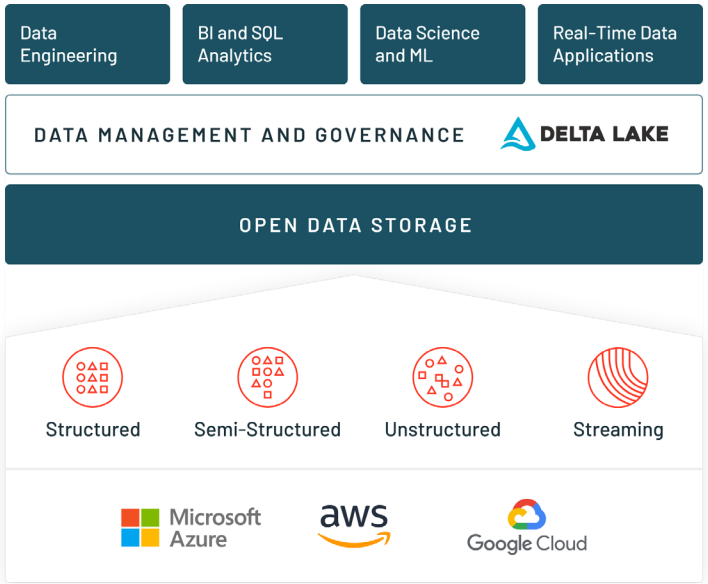
Yogesh More
GPS Pursuits Leader

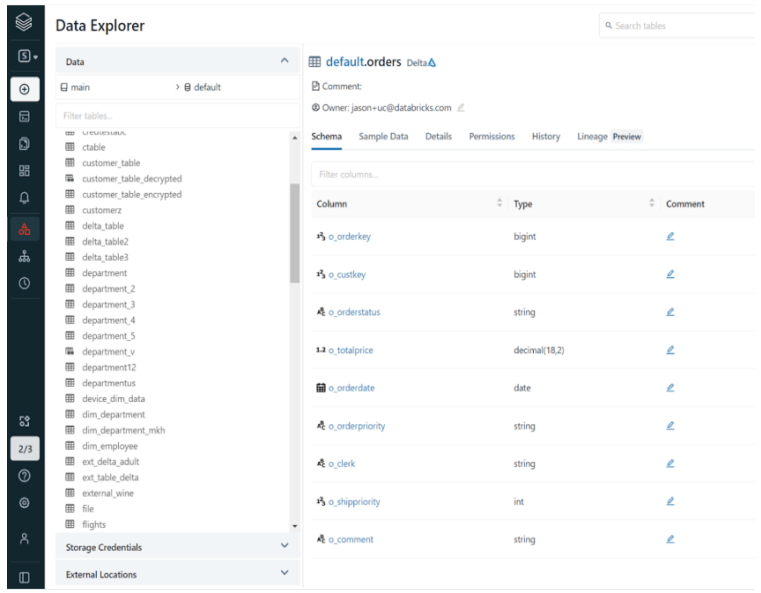
Emily Cole
Alliance Manager

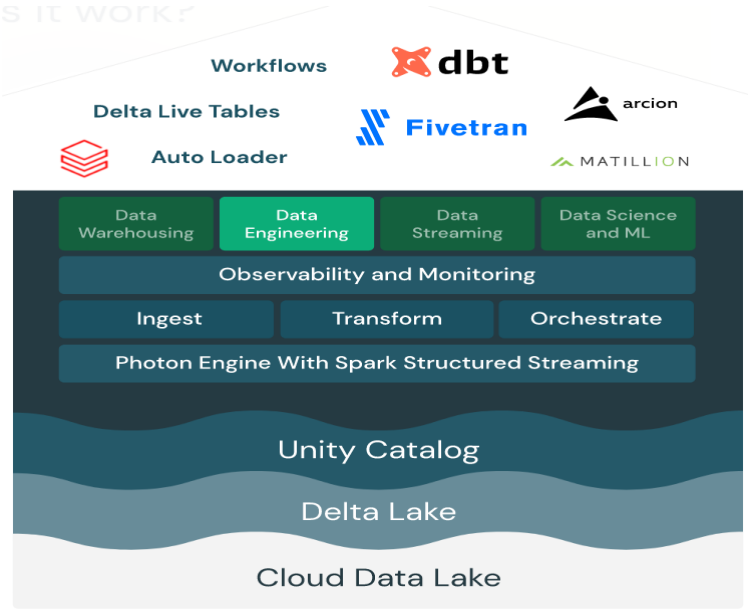
Ashvic Godinho
GPS COP Leader

**Databricks promotes badges as the key metric of enablement capabilities, as badges require hands-on capstone exams that demonstrate ability to apply knowledge to solve code-based challenges in product*

Databricks – Platform Details







Delta Lake

- Delta Lake is an open format storage layer that delivers reliability, security and performance on your data lake — for both streaming and batch operations
- By replacing data silos with a single home for structured, semi-structured and unstructured data, Delta Lake is the foundation of a cost-effective, highly scalable lakehouse

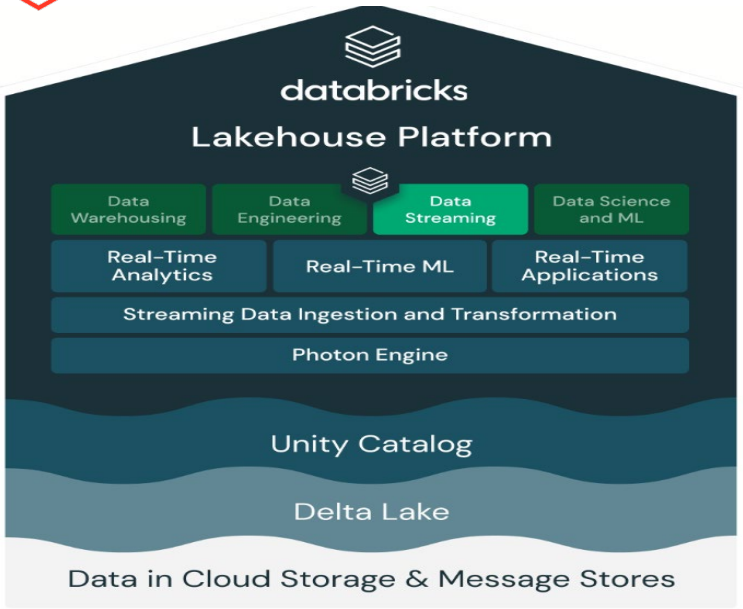
Data Governance

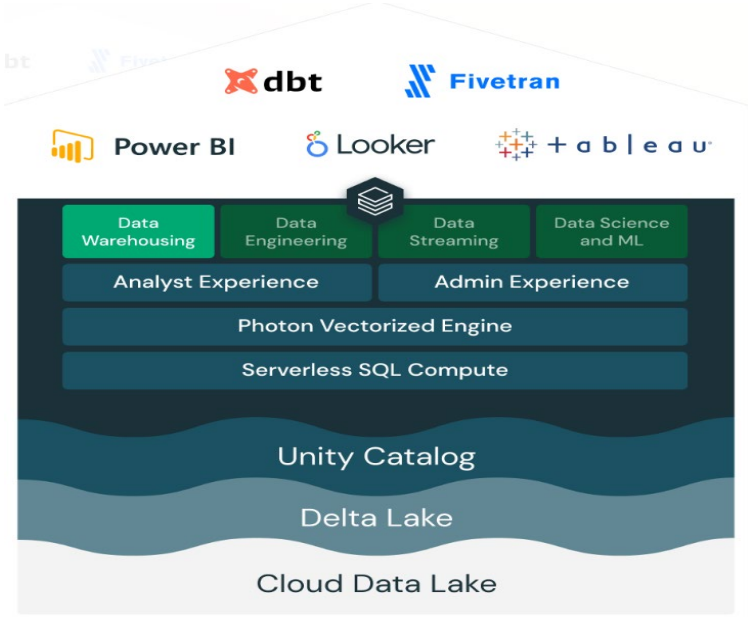
- With a common governance model based on open standard ANSI SQL, simplify governance for files, tables, dashboards and ML models on any cloud
- Unity Catalog also provides centralized fine-grained auditing by capturing an audit log of actions performed against the data and helps you meet your compliance and audit requirements

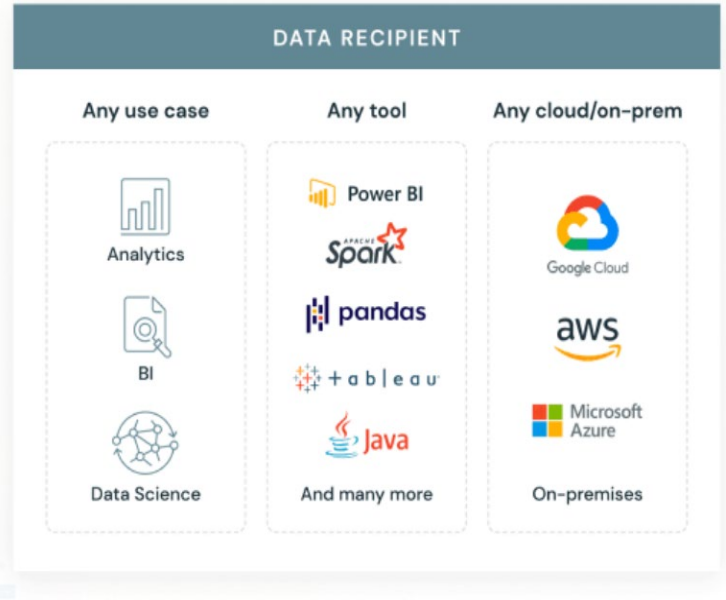
Data Engineering

- Simplified data ingestion
- Automated ETL processing
- Reliable workflow orchestration
- End-to-end observability and monitoring
- Next-generation data processing engine
- Foundation of governance, reliability and performance

Databricks – Platform Details







Data Streaming

- Streaming data ingestion and transformation
- Real-time analytics, ML and applications
- Automated operational tooling
- Next-generation stream processing engine
- Unified governance and storage

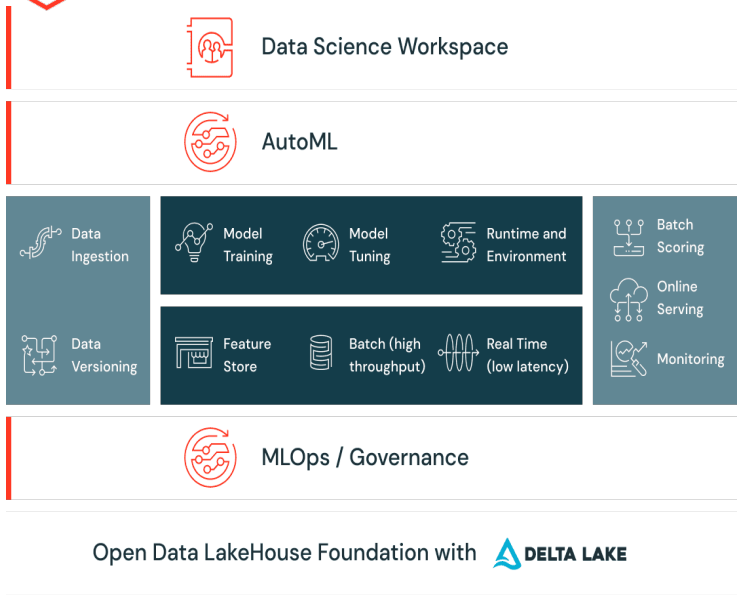
Data Warehousing

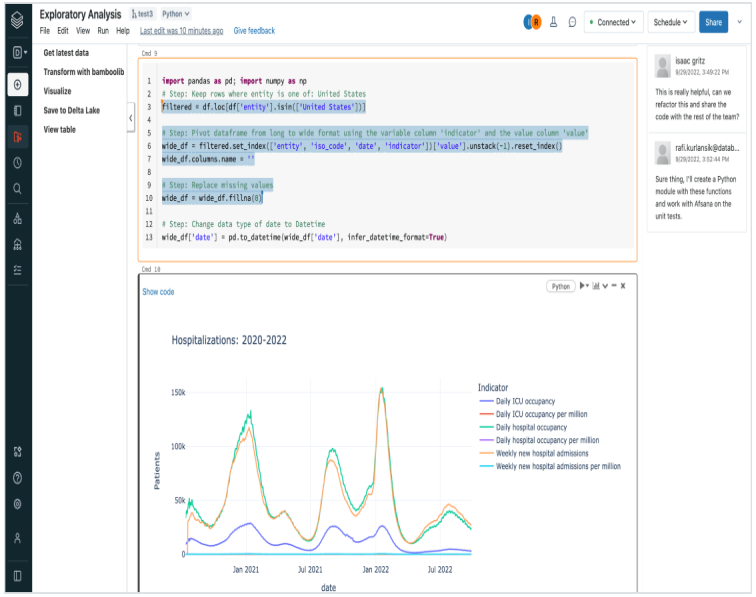
- Seamless integrations with the ecosystem
- Ease of use
- Real-world performance
- Centralized governance
- Open and reliable data lake as the foundation

Data Sharing

- Open cross-platform sharing
- Share live data with no replication
- Centralized governance
- Marketplace for data products
- Privacy-safe data cleanrooms

Databricks – Platform Details





Machine Learning

- Built on an open lakehouse architecture, Databricks Machine Learning empowers ML teams to prepare and process data, streamlines cross-team collaboration and standardizes the full ML lifecycle from experimentation to production

Data Science

- Streamline the end-to-end data science workflow — from data prep to modeling to sharing insights — with a collaborative and unified data science environment built on an open lakehouse foundation
- Get quick access to clean and reliable data, preconfigured compute resources, multi-language support, and built-in advanced visualization tools for maximum flexibility for data analytics teams.

Databricks COP Contact

COP Contact:

gpsdatabrickscop@deloiite.com

SharePoint:

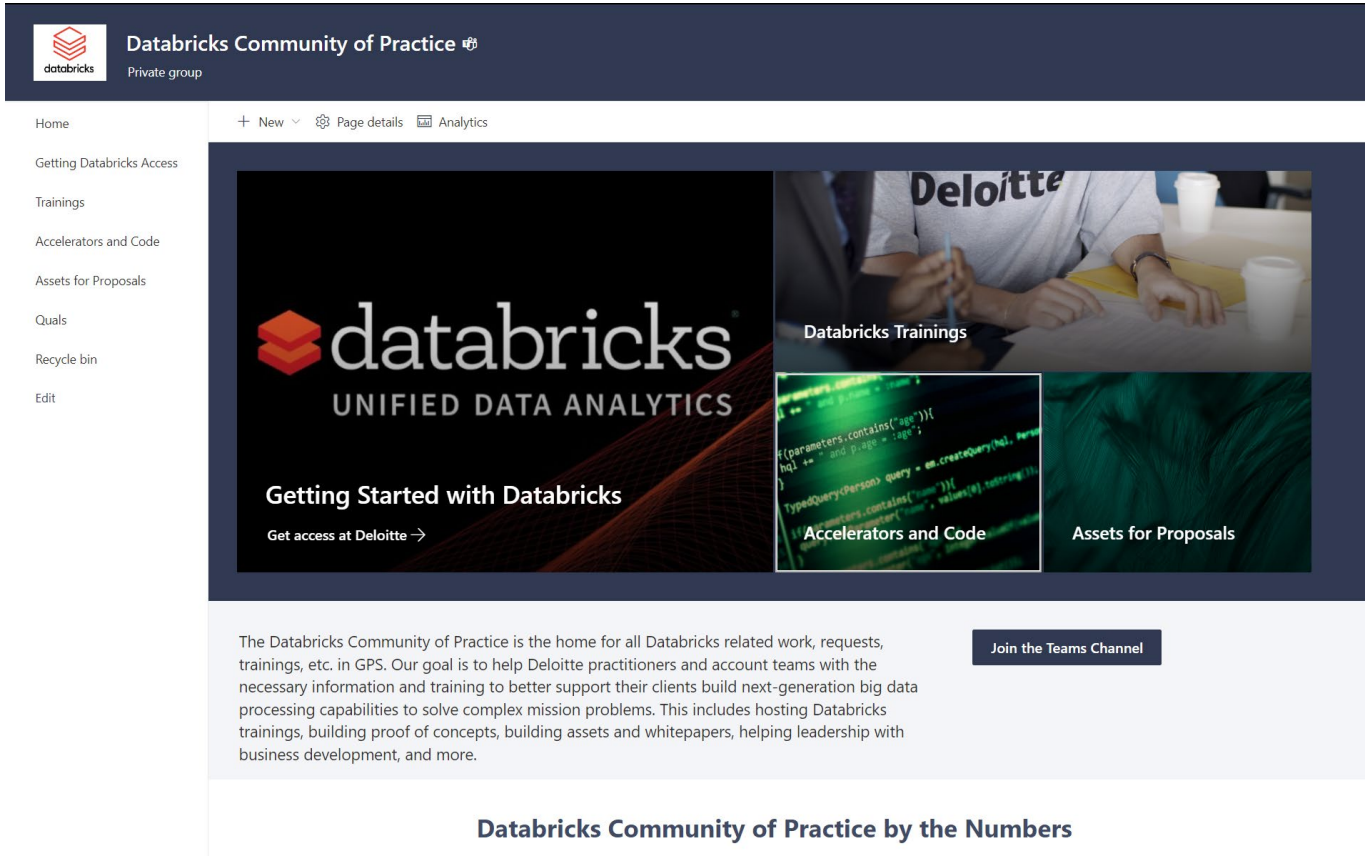
https://amedeloitte.sharepoint.com/sites/DatabricksCoP/

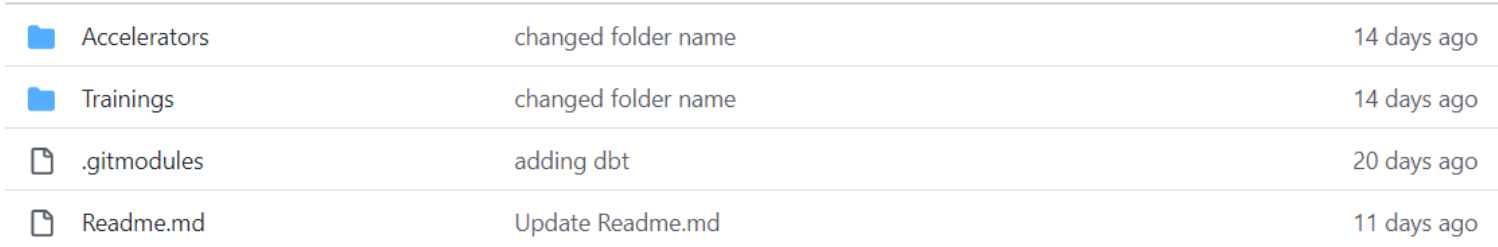
Teams Channel:

https://teams.microsoft.com/l/team/19:95b5f9a43e3a4760baa83d53b4ce5ec1%40thread.tacv2/conversations?groupId=6d208c08-8320-4e83-8cd9-ccc826553395&tenantId=36da45f1-dd2c-4d1f-af13-5abe46b99921

GitHub:

https://github.com/Deloitte/Databricks-Community-Of-Practice





“The hottest new programming language is English” (A. Karpathy)

- 2023 Databricks Summit topic was focusing on GenAI.
- Many new features were introduced in the area of GenAI app. Databricks CEO Ali announced that there will be a significant investment in LakehouseIQ.
- CEO Ali also emphasized the importance of using Unity Catalog with Delta Lake and build unified end to end AI solutions.

Democratize data analytics across the enterprise

LakehouseIQ | Databricks Marketplace | Lakehouse Apps

Develop Generative AI applications

MosaicML | Vector Search | Model Serving | Unity Catalog for AI

Implement unified governance for all data

Unity Catalog | Lakehouse Federation | Lakehouse Monitoring | Delta Lake 3.0

Databricks GenAI Offerings

Databricks' architecture is built to create, manage, and deploy custom GenAI models **quickly and easily**. These models are hosted **securely and internally in your client's environment**.

Deploy GenAI Models with a few lines of code

- Databricks uses Mlflow to manage and deploy large language models (LLMs) and integrate LLMs into the rest of your ML operations (LLMOps).
- Using automatic integration with the defacto GenAI repository, HuggingFace, Databricks can pull an open-source model and integrate with your pipelines in a minutes.
- Mlflow allows models to be tracked in your custom pipeline automatically for easy tracking, testing, and deployment.

```python
import transformers
import mlflow

chat_pipeline = transformers.pipeline(model="microsoft/DialoGPT-medium")

with mlflow.start_run():
  model_info = mlflow.transformers.log_model(
    transformers_model=chat_pipeline,
    artifact_path="chatbot",
    input_example="Hi there!"
  )

# Load as interactive pyfunc
chatbot = mlflow.pyfunc.load_model(model_info.model_uri)
```

Deploying an LLM model directing in your Databricks environment with 10 lines of code

Easy Model Integration with your pipeline

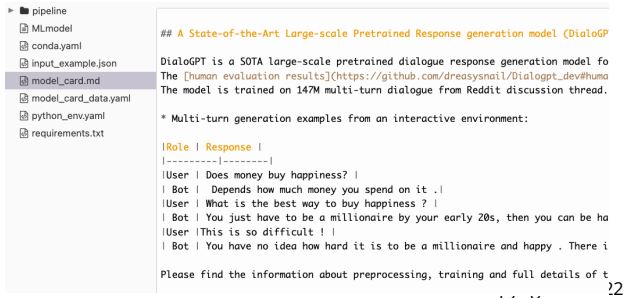
- Using LangChain, custom models can be integrated to run predictions and give responses on your data
- Provided templates for building prompts, running predictions on your Databricks data frames and integration within SQL allows non-Data Scientists to building LLM pipelines
- Databricks provides the necessary compute/GPUs, open-source models, and libraries to use without needed to set up any infrastructure

```sql
CREATE
OR REPLACE FUNCTION summarize(text STRING) RETURNS STRING RETURN llm_generate(
  concat('Summarize this to 1 sentence: ', text),
  'openai/gpt-3.5-turbo',
  'apiToken',
  secret('username', 'openai_api_token'),
  'temperature',
  0.0
);
SELECT
  product_name,
  summarize(long_product_description) AS product_summary
FROM
  products;
```

Automatic integration with SQL to run model predictions on your data

Model Cards for explanation in GenAI model types

- Databrick's provides easy and formatted descriptions of the models being used so the end user knows how to run a model
- Using Model Cards which are downloading in your Databricks environment when a model is pulled provides example prompts to use and information about the model



Open-source models used in Databricks come up a Model Card in Markdown that describes the model

Databricks Model Management

Using Mlflow, GenAI models, LLMs, and traditional data science models can be managed in Databricks using the same framework to easily manage models from development to production

Model Development

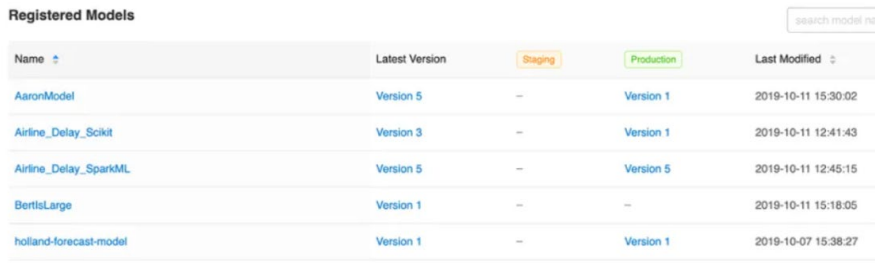
Bootstrap ML projects, perform rapid iteration with ease and ship high-quality models to production at scale.

```python
import transformers
import mlflow

summary_pipeline = transformers.pipeline(model="databricks/dolly-v2-12b")

with mlflow.start_run() as run:
    mlflow.transformers.log_model(
        transformers_model=summary_pipeline,
        artifact_path="your_path",
```

Taking any model and integrating it into your development pipeline

Model Management

Central place to share models, move from experimentation, testing, and production. Integrated with workflows and CI/CD pipelines, and monitor ML deployments and their performance.



Easily saving the model to Databricks to track history of models, and each revision

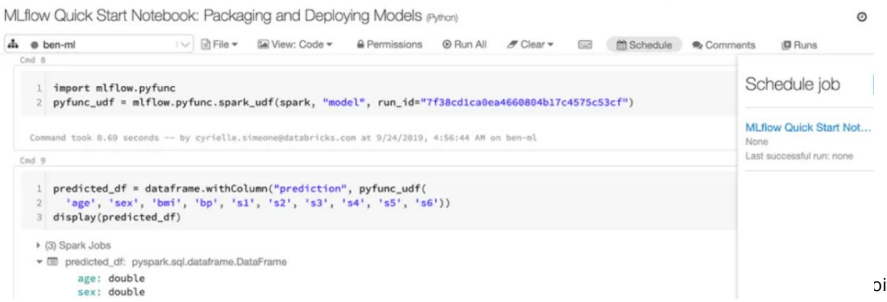
Experiment Tracking

Run experiments with any ML library, framework or language, and keep track of parameters, metrics, code and models from each experiment. Share, manage and compare results along with code versions integrated with the Databricks notebooks.

Example of Databricks tracking each model, parameter, and evaluation metric in the user interface

Model Deployment

Deploy production models for inference on as APIs using built-in integration with Docker containers, Azure ML or Amazon SageMaker. With Databricks, you can operationalize and monitor production models using Databricks Jobs Scheduler and auto-managed Clusters to scale based on the business needs.



Deploying models to productions with an API endpoint

Read about the acquisition here!

Future GenAI features with MosaicML

MosaicML's platform allows Databricks to easily train and deploy generative AI models on your data, **in your environment.**

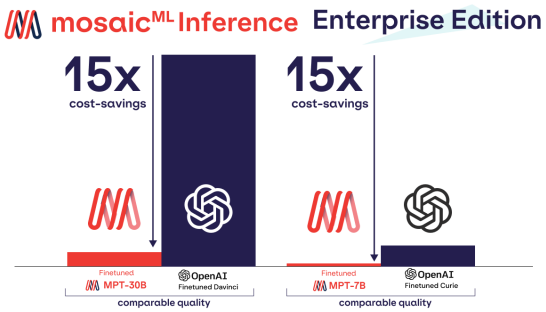
MosaicML Training

- Train your own LLMs and other generative AI models.
- Maintain full control of your data in your secure environment.
- Orchestrate across multiple clouds.

```
> mcli run -f gpt3-70b.yaml --gpus 512

------------------------------------------------------------
Let's run this run
------------------------------------------------------------

i  Run gpt3-70b-leaping-octopus submitted. Waiting for it to start...
i  You can press Ctrl+C to quit and follow your run manually.
✓  Run gpt3-70b-leaping-octopus started
i  Following run logs. Press Ctrl+C to quit.

Cloning into 'composer'...
remote: Enumerating objects: 17992, done.
remote: Counting objects: 100% (335/335), done.
remote: Compressing objects: 100% (260/260), done.
Receiving objects:   8% (1440/17992), 836.01 KiB | 793.00 KiB/s
```

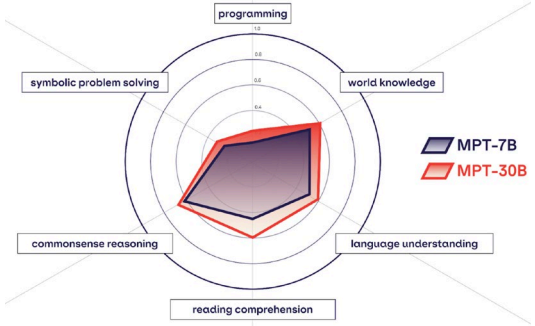
MosaicML Inference

- Add AI to your apps up to 15x cheaper. Choose any model, any size, in your secure environment.



MosaicML MPT Models

- Open-source **MPT-7B model**
  - Extension models: -Instruct, -Chat, and -StoryWriter have collectively been downloaded over 3M times
- **MPT-30B**, a new, more powerful member of, trained with an 8k context length on NVIDIA H100 Tensor Core GPUs.



Our view of MosaicML :

1. From their website About Us: ***MosaicML is the generative AI platform that empowers enterprises to build their own AI. Our research and engineering teams use cutting-edge scientific research to develop products that make it fast, cost-effective, and easy to train today's most popular machine learning models. MosaicML enables developers to maintain full control over the AI models they build, with model ownership and data privacy built into the platform's design.***
2. Databricks and Mosaic believe in a world where **everyone is powered to train their own LLM models, imbued with their own data**, wisdom and creativity, rather than have this capability centralized in a few generic models.
3. The Mosaic Training product is used to **train your own LLMs and** other GenAI models **at scale while maintaining full control** in your secure environment with orchestration across multiple clouds.
4. The Mosaic Inference product **makes large models accessible to all organizations** where you can **turn any saved model into a secure, inexpensive API** within a Mosaic managed cluster or within your own VPC in under a minute.
5. Mosaic also **provides open source models with commercial license terms that are hosted by Mosaic and available through an API**: text embedding models and text completion models

Deloitte. Databricks GenAI Models

There are several use cases being built out in GPS using Databricks's GenAI capabilities

1

Pretrained (w/o context)

Foundational Dolly for General Knowledge Q&A

- Open source – can download directly in your own environment
- Pretrained on 15k instruction/response

"Did Nero burn down Rome?"

"Yes"

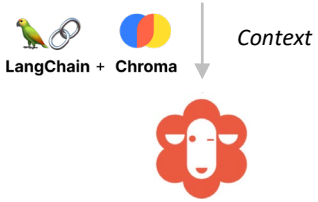
2

Prompt Engineering w/ context

Instruction-Fed Dolly for Context Bound Q&A

- Dolly model that is given *context* to a question
- Uses other AI toolsets such as VectorDB and LangChain to create better responses

"Did Nero burn down Rome?"



"The answer is more complicated"

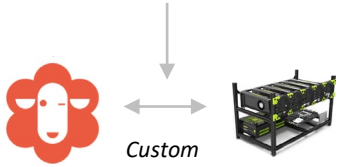
3

Full Scale Fine Tuning

Domain Adapted & Fine-Tuned Dolly for Specific Use Cases

- Retraining layers of NN (for Dolly) using full volume of context/training data
- Requires lots of GPU power

"Did Nero burn down Rome?"



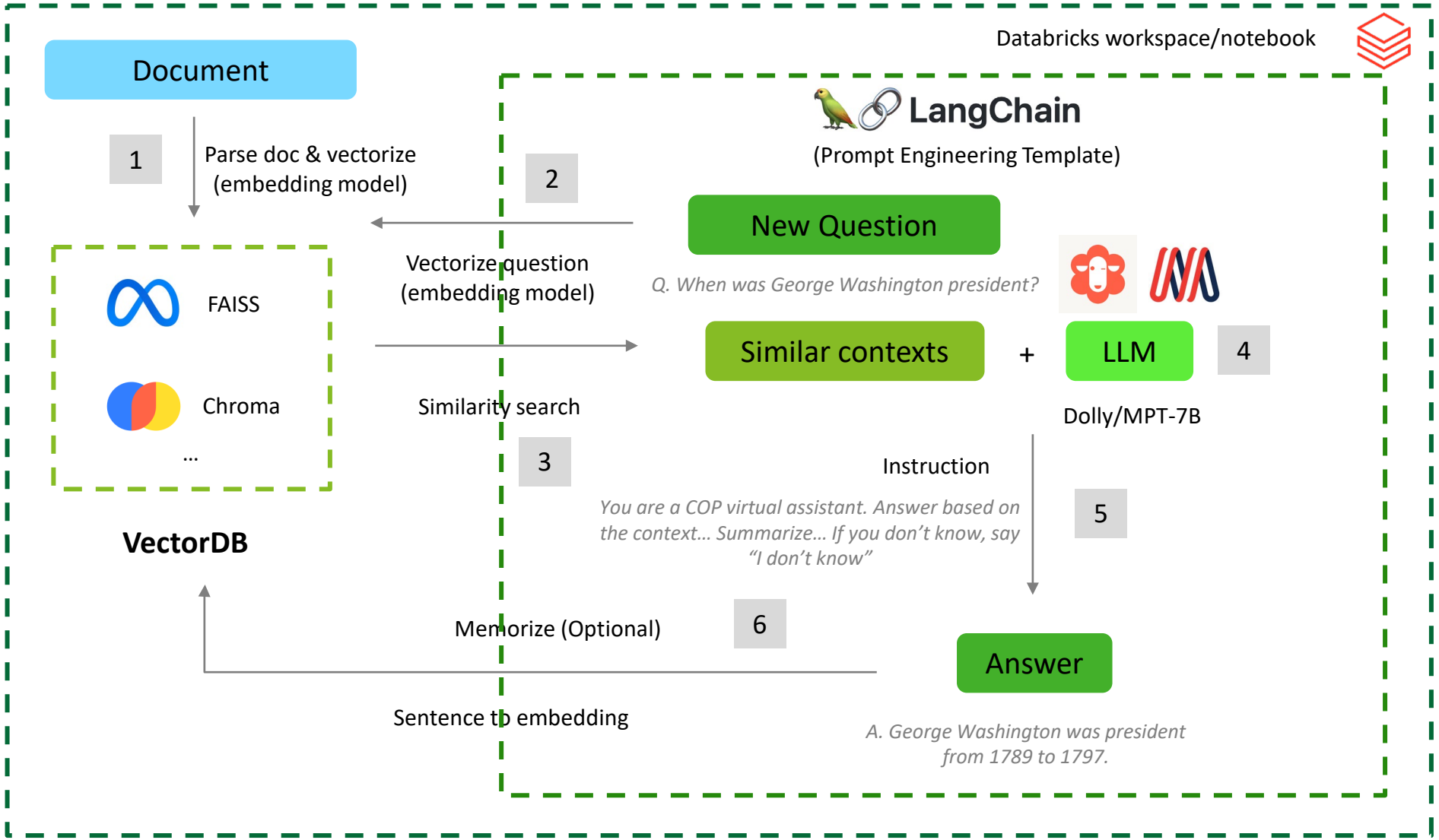
"This is what historians say exactly"

GPS Use Cases

- **CDER/FDA has deployed Dolly and other LLMs pipeline** integrated with VectorDB, LangChain within Databricks environment
- Other GPS accounts are in experimentation mode
- **Internal GPS efforts**:
  - Databricks COP built Dolly POC
  - preparing Proposal Response Initiative;
  - providing advisory for AI Center of Excellence, etc
  - GenAI meeting minutes summarization

Deloitte. Databricks Architecture

Example architecture of Databricks Dolly **hosted internally** at Deloitte



Step by Step

1. Parse the (PDF) document and store it into **VectorDB** using embedding model.
2. Get the question and transform it as **vectors** using the same sentence embedding model.
3. Do a **similarity search** within Vector DB to find most relevant contexts.
4. Engineer **a prompt** containing the question & the similar Q&A as context.
5. Send the "engineered prompt" to LLM.
6. Return contextual answers!

Deloitte.

Thank you