LTIMindtree

POV

The Fast-Track is Getting Faster:

Why the Data Intelligence Platform is the Best Place for Your Gen AI Projects

“AI is the new electricity” – Andrew Ng

Just as the advent of harnessing electric power ushered in a new age, AI is driving incredible changes across industries like never before. AI is an emerging and rapidly evolving technology that will significantly impact all industry sectors from healthcare to finance.

Generative AI is a branch of AI that creates new data instances from data that it is trained on, in order to solve real-life problems. Generative AI models use neural networks to identify the patterns and structures within existing data to generate new and original content. This raises exciting possibilities for building solutions across industries, but it requires a data platform that can harness large volumes of data to train Gen AI solutions.

The objective of this Point of View (POV) is to provide insights into how the Databricks Data Intelligence Platform can be leveraged to build better Gen AI applications for enterprise value, the features and capabilities within the platform that can be utilized to this end, and how the entire Large Language Models (LLM) lifecycle process can be governed.

Key challenges in developing Gen AI solutions:

First, let us explore some of the key challenges that businesses encounter while developing and integrating Gen AI models into their production environment:

1. Optimizing model quality

Poor data can lead to biases, hallucinations, and toxic output. It is difficult to effectively evaluate Large Language Models (LLMs) as these models rarely have an objective ground truth label. Due to this, organizations often struggle to understand when the model can be trusted in critical use cases without supervision.

2. Cost and complexity of training with enterprise data

Organizations are looking to train their models using their own data and control them. However, they are unable to decide on how many data examples are adequate, which base model they should start with, how to manage the complexities of the infrastructure required to train and fine-tune models, and how to think about costs.

3. Trusting models in production

With a rapidly evolving technology landscape rapidly and new capabilities being introduced, it is a challenge to get these models into production. Sometimes these capabilities require new services such as a vector database, while at other times, they may require new interfaces such as deep prompt engineering support and tracking. Trusting models in production is difficult without robust and scalable infrastructure and a stack fully instrumented for monitoring.

4. Data security and governance

Organizations want to control what data is sent to and stored by third parties to prevent data leakage as well as to ensure that responses conform to regulations. We've seen cases where teams have unrestricted practices that compromise security and privacy or have cumbersome processes for data usage that impede the speed of innovation.

Focus areas for building a Gen AI solution:

As clients are looking for the right platform to build and productionize their Gen AI solutions, we will focus on three areas for successful implementation and how Databricks features and offerings can support this journey.

1. Robust platform for LLM Lifecycle

Databricks Mosaic AI is a Databricks offering which helps customers develop generative AI solutions rapidly by using foundational Software-as-a-Service (SaaS) models to securely train their own custom models with enterprise data.

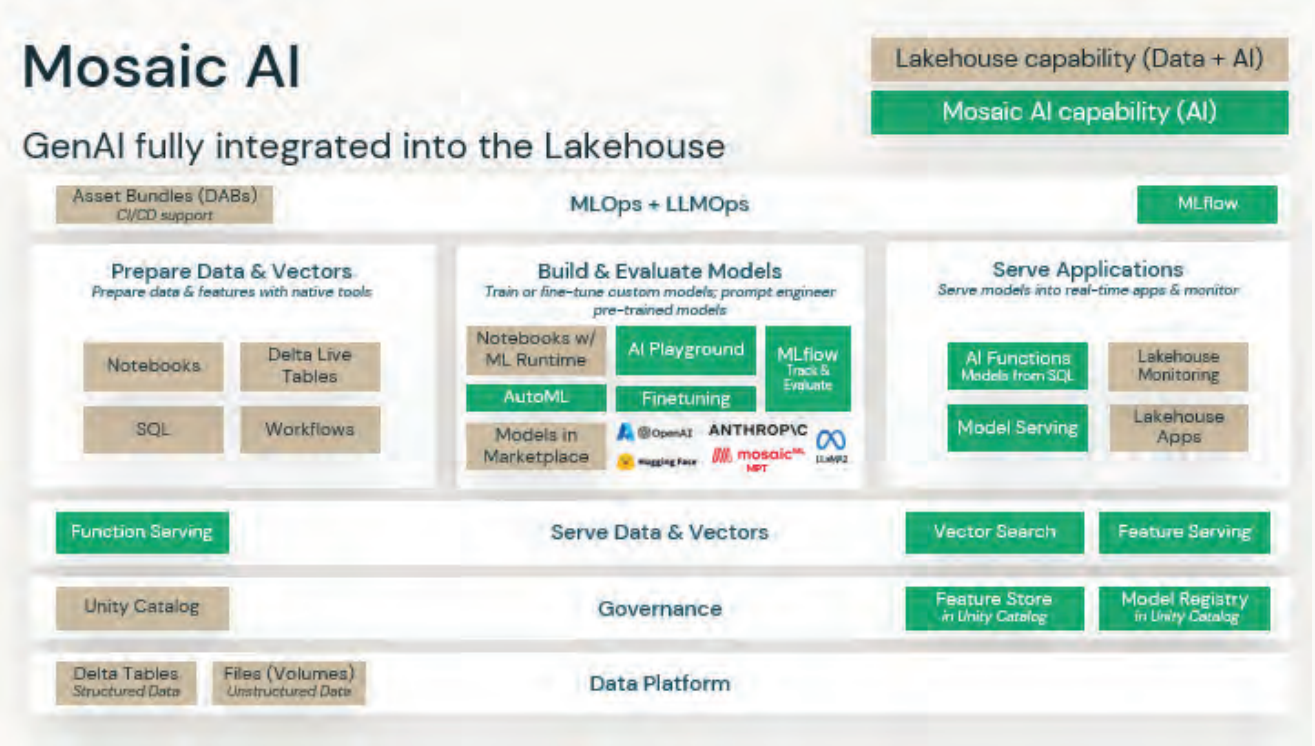


Figure 1: Mosaic AI – GenAI fully integrated into the Lakehouse

Mosaic AI

Databricks Mosaic AI offers the following capabilities for customers to accelerate their journey from POC to production and maintain data governance and security.

(Please note that * denotes a feature available in public preview and ** denotes a feature available in private preview)

a. **Vector search***

With vector embeddings, organizations can leverage the power of Gen AI and LLMs across many use cases. A vector database helps teams quickly index their organizations' data by embedding vectors and performing low-latency vector similarity searches in real-time deployments.

b. **Curated models**

Databricks Marketplace provides curated models for some of the common use cases. These task specific LLMs can be directly used or fine-tuned for your own data. Optimized model serving will provide highly available and low-latency service for deploying models.

c. **Fine-tuning***

AutoML offers support for fine tuning generative AI models for text classification as well fine-tune base embedding models with your data. AutoML enables non-technical users to fine-tune models with point-and-click ease using your organization's data. It also increases the efficiency of technical users doing the same.

d. **AI functions***

AI Functions is a built-in DB SQL function, allowing you to access Large Language Models (LLMs) directly from SQL. Common use cases include summarization, topic identification, entity extraction, and new content creation.

e. **Inference tables***

The incoming requests and outgoing responses to serving endpoints are logged to Delta tables in your Unity Catalog. This automatic payload logging enables teams to monitor the quality of their models in near real-time, and the table can be used to easily source data points that need to be relabeled as the next dataset to fine-tune your embeddings or other LLMs.

f. **Managed MLflow***

Managed MLflow provides innovative features that enhance its capability to manage and deploy large language models. MLflow's integration with LangChain and Prompt Engineering UI enables simplified model development for creating generative AI applications for a variety of use cases, including chatbots, document summarization, text classification and sentiment analysis.

Recently, it was also enhanced with LLM Evaluation and UI for training models, you can compare multiple models and prompts visually, iteratively test new queries during development and logging and visualization improvements in model training.

g. **LLMOps***

LLMOps platform provides data scientists and software engineers with a collaborative environment that facilitates iterative data exploration, real-time coworking capabilities for experiment tracking, prompt engineering, and model and pipeline management, as well as controlled model transitioning, deployment, and monitoring for LLMs. LLMOps automates the operational, synchronization and monitoring aspects of the machine learning lifecycle.

h. **RAG Studio:** RAG Studio is a managed RAG service, provides tools and an opinionated workflow for developing, evaluating, and iterating on Retrieval-Augmented Generation (RAG) applications in order to build apps that deliver consistent, accurate answers. RAG Studio is built on top of MLflow and is tightly integrated with Databricks tools and infrastructure.

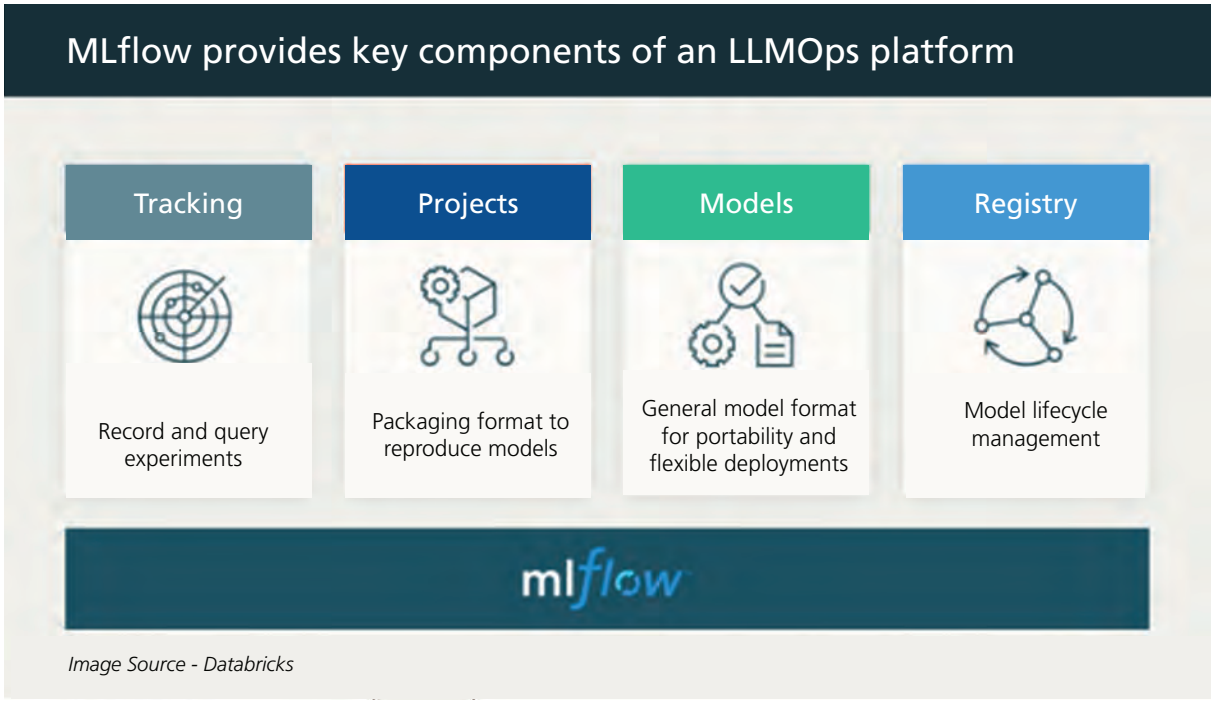


Figure 2: mlflow – Components of an LLMOps platform

Databricks glossary – What is LLMOps

i. **Model serving**

It is a fully managed service which provides a single solution to deploy any AI model without needing to manage complex infrastructure. This means you can deploy any natural language, vision, audio, tabular, or custom model, regardless of how it was trained – whether built from scratch, sourced from open-source, or fine-tuned with proprietary data. Databricks Model Serving support serving custom models (include scikit-learn, XGBoost, PyTorch, and Hugging Face transformer models), open sources models (made available by Foundation Model APIs), as well as External models (like OpenAI's GPT-4 and more).

The service will automatically scale instances to meet traffic patterns, saving infrastructure costs while optimizing latency performance.

2. Governance

a. **Lakehouse monitoring***

This data and AI monitoring service allows users to simultaneously track the quality of their data and AI assets. This fully managed service maintains profile and drift metrics on your assets, let's you configure proactive alerts, auto-generates quality dashboards to visualize and share across your organization, and facilitates root-cause analysis by correlating data-quality alerts across the lineage graph.

b. **External models ****

As organizations are empowering their employees to leverage OpenAI and other LLM providers, they are running into issues managing rate limits and credentials, burgeoning costs, and tracking what data is sent externally. The MLflow AI Gateway, part of MLflow 2.5, is a workspace-level API gateway that allows organizations to create and share routes, which then can be configured with various rate limits, caching, cost attribution, etc. to manage costs and usage.

c. **Unity Catalog**

The Unity Catalog provides comprehensive governance and lineage tracking of both data and AI assets in a single unified experience. The Model Registry is now provided via Models in Unity Catalog. The benefits of Unity Catalog are applied to ML models, which includes Centralized access control, Auditing, Lineage, Model sharing and discovery across workspaces.

d. **Databricks CLI for MLOps**

Databricks CLI allows data teams to set up projects with infra-as-code and get to production faster with integrated CI/CD tooling. Organizations can create "bundles" to automate AI lifecycle components with Databricks workflows.

e. **Delta sharing**

Clients can share live access to data, AI models and notebooks directly with consumers without costly or complicated replication. Delta sharing gives providers an easy way to manage access permissions to any consumer regardless of their cloud, region, or platform.

3. MosaicML for optimizing costs

MosaicML (acquired by Databricks) is a leading platform for creating and customizing generative AI models for your enterprise. MosaicML has an optimized stack to build your own LLMs. It allows you to train multi-billion parameter models in days, not weeks, reduce training costs by 10x and provide optimized LLM serving for reduced deployment cost.

MosaicML architecture:

Figure 3: MosaicML Architecture

MosaicML Platform: The software infrastructure for generative AI, Hagay Lupesko and Ajay Saini, February 28, 2023

Conclusion:

In summary, the demand for generative AI is driving disruption across industries. Thus, it is imperative for technical teams to build generative AI models and LLMs on top of their own data to differentiate their offerings.

The Databricks Data Intelligence platform along with Unity Catalog and Lakehouse AI provide a unified platform for data and AI so that customers can develop their Gen AI solutions faster, deploy them with minimum effort and reduce costs. By bringing together data, AI models, LLM operations (LLMOps), monitoring and governance on the Databricks Data Intelligence Platform, organizations can accelerate deriving business value from their generative AI journey.

References

- Databricks Vector Search, February 15, 2024
  https://docs.databricks.com/en/generative-ai/vector-search.html
- Lakehouse AI: A Data-Centric Approach to Building Generative AI Ap-plications, Patrick Wendell, Matei Zaharia, Xiangrui Meng, Craig Wiley, Kasey Uhlenhuth, Eric Peter and Prem Prakash, June 28, 2023
  https://www.databricks.com/blog/lakehouse-ai
- Build GenAI Apps Faster with New Foundation Model Capabilities, Ahmed Bilal, Asfandyar Qureshi, Margaret Qian, Jianwei Xie, Sue Ann Hong, Vladimir Kolovski, Mingyu Li and Ankit Mathur, December 11, 2023
  https://www.databricks.com/blog/build-genai-apps-faster-new-foundation-model-capabilities
- Introduction to Databricks Lakehouse Monitoring, February 15, 2024
  https://docs.databricks.com/en/lakehouse-monitoring/index.html
- Model serving with Databricks, February 07, 2024
  https://docs.databricks.com/en/machine-learning/model-serving/index.html
- Evaluate large language models with MLflow, January 12, 2024
  https://docs.databricks.com/en/mlflow/llm-evaluate.html#what-is-mlflow-llm-evaluate

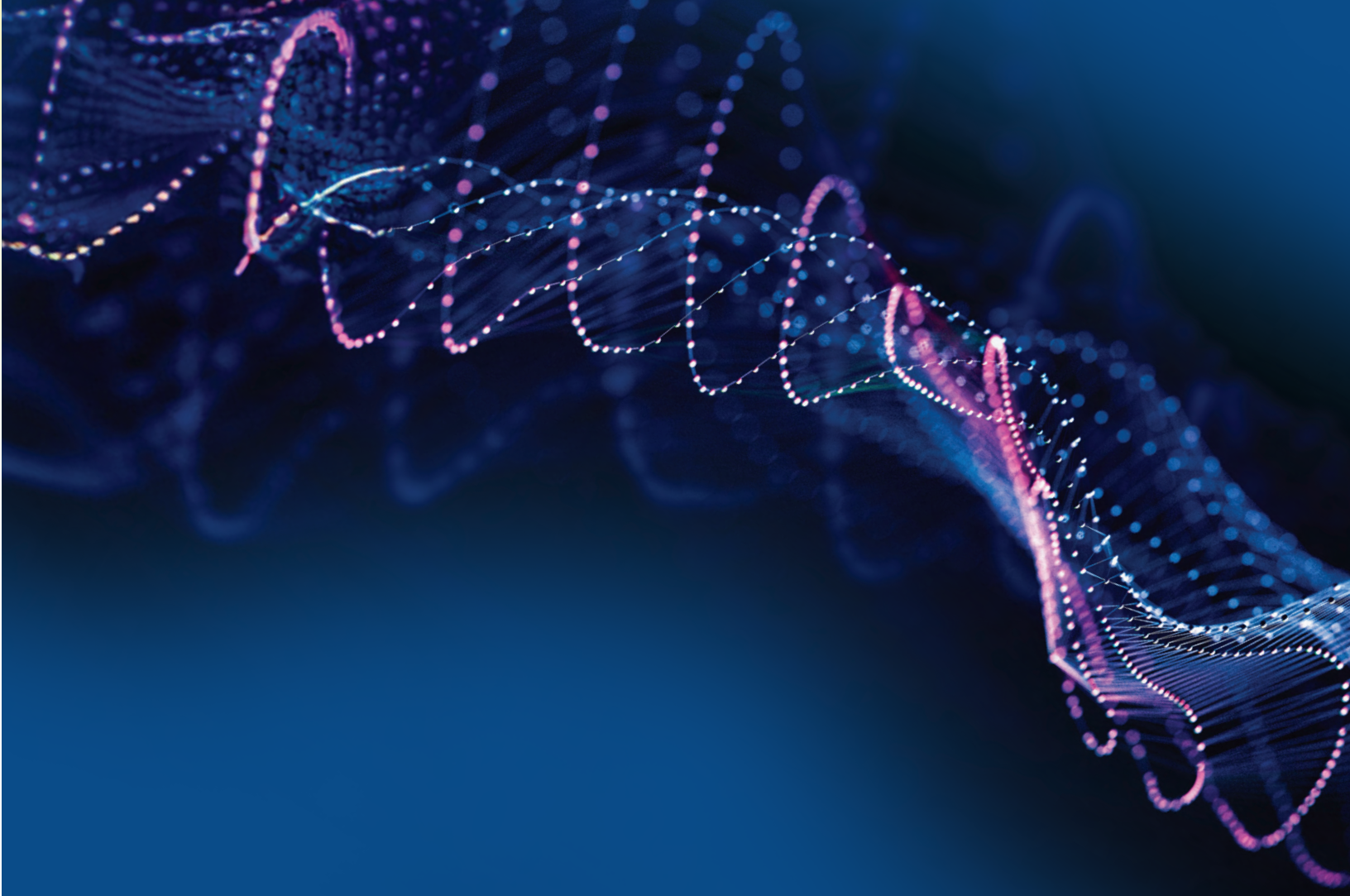
- Databricks glossary – What is LLMOps
  https://www.databricks.com/glossary/llmops
- MosaicML Platform: The software infrastructure for generative AI, Hagay Lupesko and Ajay Saini, February 28, 2023
  https://www.mosaicml.com/blog/train-custom-gpt-diffusion-models
- Databricks Unity Catalog — Unified governance for data, analytics and AI
  https://www.databricks.com/product/unity-catalog

About Author

Shankar Mahadevan

Designation: Senior Principal - Architecture

Shankar has 23+ years of experience helping customers build data platforms and analytical systems. In recent years, he has been helping organizations on their data modernization and migration journey to cloud data platforms like Azure and Databricks. He has also managed large scale multi-year programs across geographies and verticals like manufacturing, BFSI and healthcare. His core competencies include Data Strategy consulting, architecture design, data integration, data warehousing and reporting. He is certified as an Azure Solution Architect Expert and Databricks Champion. He holds a bachelor's degree in Electronics and Post Graduate Diploma in Finance



LTIMindtree

LTIMindtree is a global technology consulting and digital solutions company that enables enterprises across industries to reimagine business models, accelerate innovation, and maximize growth by harnessing digital technologies. As a digital transformation partner to more than 700 clients, LTIMindtree brings extensive domain and technology expertise to help drive superior competitive differentiation, customer experiences, and business outcomes in a converging world. Powered by 82,000+ talented and entrepreneurial professionals across more than 30 countries, LTIMindtree — a Larsen & Toubro Group company — combines the industry-acclaimed strengths of erstwhile Larsen and Toubro Infotech and Mindtree in solving the most complex business challenges and delivering transformation at scale. For more information, please visit **https://www.ltimindtree.com/.**