Lakehouse AI

AI capabilities built directly into the data platform
Agenda

Databricks’ perspective on Generative AI & LLMs
Challenges faced building LLMs & AI
How Lakehouse solves these challenges
Creator of

Inventor and pioneer of the data lakehouse

5000+ global employees

$1B+ in revenue

Creator of

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

$3B in investment
Frankly, this is AI’s “iPhone” moment

Generative AI & LLMs are a once-in-a-generation shift in technology

“AI has become democratized”

“Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality”

“Smaller, more performant models such as LLaMA enable… further democratizing access in this important, fast-changing field…”

“GPT-4 beats 90% of lawyers trying to pass the bar”

“Falcon is now free of royalties for commercial and research use… Falcon 40B outperforms ... Meta’s LLaMA and Stability AI’s StableLM”
88% of enterprises are already investing in Generative AI

MIT Technology Review
### AI Application Themes

<table>
<thead>
<tr>
<th>AI model capabilities</th>
<th>Generative AI</th>
<th>Analytical AI</th>
</tr>
</thead>
</table>
| **AI facilitated knowledge and insight generation** | Knowledge Management  
Surfacing information and insights embedded in all manner of data (documents, e-mails, transcripts) | Summarisation  
generate concise and precise summary of collected text / other media |
| **AI augmented decision and action support** | Reasoning and planning  
Ability to plan and propose actions, explaining reasoning and decision logic | Generate new content  
Automatically create new, value-adding content (text, image) |
| **Fully autonomous AI agents** | Interaction and communication  
Adaptable, contextual communication between a user and a computer system (asynchronous or real-time) | Act and use tools  
AI understands desired actions and executes (e.g., email, Claims Centre etc.) |
| **Insights** | Prediction  
Predict future events based on historical data and diagnosis of potential actions | Prescription  
Automate decisioning based on a defined objective function / target |
SaaS models do not provide differentiation and limit

**Control**
SaaS vendors own the weights for your models

**Privacy**
Your data must be transferred to the SaaS vendor

**Efficiency**
Poor economics beyond generic use cases and proof-of-concepts
Risks and Challenges
Generative AI brings new risks and challenges for businesses and society

• Legal issues
  • Privacy
  • Security
  • Intellectual property protection
• Ethical issues
  • Bias
  • Misinformation
• Social/Environmental issues
  • Impact on workforce
  • Impact on the environment
Auditing Generative AI Models
Allocating responsibility and increasing model transparency

Governance Audit
- Model limitations
- Model characteristics
- Training datasets
- Model selection and testing procedures
- Impact reports
- Failure model analysis
- Model access
- Intended/prohibited use cases

Model Audit
- Model limitations
- Model characteristics
- Output logs
- Environmental data

Application Audit

Source: Mitsander et al 2023
Host your own model for a better solution

**Control**
Full ownership of your gen AI solution – from data to models

**Privacy**
Maintain compliance in your secure, private environment

**Efficiency**
Up to 7x less expensive to tune or train your own model
Host your own model for a better solution

Control
Full ownership of your gen AI solution—from data to models

71%

of technology executives plan to build some or all of their own models

MIT Technology Review

Cost
Up to 7x less expensive to tune or train your own model

Privacy
Maintain compliance in your secure, single environment
Lakehouse — a data-centric approach
Supports ALL data use cases from Engineering to BI to AI

Datasets
Prepare data & features with native tools

Models
Use Pre-trained Model or Build Custom Model

Applications
Serve Models into Real-Time Apps and Monitor

Governance — Unity Catalog

Data Platform — Delta Lake
Agenda

Databricks’ perspective on Generative AI & LLMs

Challenges faced building LLMs & AI

How Lakehouse solves these challenges
Delivering business value from LLMs requires tackling many challenges. How do we...?

- **Customize LLMs with our data**
  - ...rapidly prototype high-value LLM use cases?
  - ...pick the right LLM for each use case (proprietary, open source, ...)?
  - ...choose the right customization technique (fine-tuning, prompt engineering, ...)?

- **Securely connect our data to LLMs**
  - ...securely connect structured data sources to LLMs?
  - ...vectorize your unstructured data for LLMs?

- **Deploy LLMs without new infrastructure**
  - ...manage LLMOps?
  - ...deploy large models that require complex GPU configurations?
  - ...access large amounts of unstructured data in a vector databases?

- **Ensure LLMs deliver high quality answers**
  - ...prevent hallucinations and incorrect answers?
  - ...ensure compliance with ethics and business policies?

- **Integrate LLMs w/ data governance**
  - ...prevent a data privacy or security leak to third-party LLM vendors?
  - ...enforce existing data access controls & permissions?

- **Maintain flexibility to upgrade LLMs**
  - ...ensure we can upgrade our LLMs as the technology advances?
  - ...make sure we aren’t caught flat footed if a vendor changes the price?
Generative AI state of the art is rapidly advancing

No single model to rule them all—trade-offs are required to find the best model for each use case

The decision criteria are:

- Privacy
- Quality
- Cost
- Latency

Proprietary LLMs

ChatGPT
PaLM 2
ANTHROPV\C
OpenAI

Open Source LLMs

databricks
Dolly
mosaic
Hugging Face
stability.ai
Stable Diffusion

©2023 Databricks Inc. — All rights reserved | Confidential and proprietary
Using Proprietary Models (LLMs-as-a-Service)

**Pros**
- Speed of development
  - Quick to get started and working.
  - As this is another API call, it will fit very easily into existing pipelines.
- Quality
  - Can offer state-of-the-art results

**Cons**
- Cost
  - Pay for each token sent/received.
- Data Privacy/Security
  - You may not know how your data is being used.
- Vendor lock-in
  - Susceptible to vendor outages, deprecated features, etc.
Using Open Source Models

Pros

- Task-tailoring
  - Select and/or fine-tune a task-specific model for your use case.
- Inference Cost
  - More tailored models often smaller, making them faster at inference time.
- Control
  - All of the data and model information stays entirely within your locus of control.

Cons

- Upfront time investments
  - Needs time to select, evaluate, and possibly tune
- Data Requirements
  - Fine-tuning or larger models require larger datasets.
- Skill Sets
  - Require in-house expertise
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

**Proprietary Models**
- Usually offered as *LLMs-as-a-service*
- Some can be *fine-tuned*
- Restrictive licenses for usage and modification
Tackling these challenges requires integrating your data with full AI capabilities.
Agenda

Databricks’ perspective on Generative AI & LLMs
Challenges faced building LLMs & AI
How Lakehouse solves these challenges
What is a Lakehouse?

One platform that unifies all of data, analytics, and AI workloads

Link to Blog Post by Bill Inmon, Computer scientist, author, and technology pioneer. Best known as the Father of Data Warehousing

©2021 Databricks Inc. — All rights reserved
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)
Lakehouse: AI capabilities built directly into the data platform

Unified governance and unified tools that operate on single copy of your data

MLops + LLMops + DataOps
Shared interfaces and paradigms to deploy data and AI

One platform to support multiple personas

Prepare Data
One toolkit for data prep

Develop & Evaluate AI
Tools for every persona

Serve AI
Every model accessible via API, SQL, or Pipelines

Serve Data
One platform for real-time data serving

Monitoring
One toolkit for data and AI monitoring

Data Science & ML

Unity catalog
One security and governance model for all data and AI across the organization

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

Cloud Data Lake
All Raw Data
(Logs, Texts, Audio, Video, Images)
Lakehouse: AI capabilities built directly into the data platform
Unified governance and unified tools that operate on single copy of your data

MLops + LLMops + DataOps
Shared interfaces and paradigms to deploy data and AI

Prepare Data
One toolkit for data prep

Develop & Evaluate AI
Tools for every persona

Serve AI
Every model accessible via API, SQL, or Pipelines

Serve Data
One platform for real-time data serving

Monitoring
One toolkit for data and AI monitoring

Unity catalog
One security and governance model for all data and AI across the organization

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

The first AI platform built directly into the data layer
- Common tooling for all personas
- End-to-end governance, lineage, version control across data and AI
- Single copy of your data
- Models inherit the governance of the data they are trained on

Intelligently leverages semantic & lineage understanding for automation
- Automated error correction
- Smart feature suggestions
- ...

Cloud Data Lake
All Raw Data
(Logs, Texts, Audio, Video, Images)
Lakehouse AI works for all AI models

Classic, deep, proprietary or open source Generative AI + LLMs

<table>
<thead>
<tr>
<th>Deep learning models</th>
<th>Classical ML algorithms</th>
<th>Proprietary LLMs</th>
<th>Open source generative AI + LLMs</th>
<th>Chains &amp; agents</th>
</tr>
</thead>
<tbody>
<tr>
<td>PyTorch</td>
<td>learn</td>
<td>ChatGPT</td>
<td>databricks Dolly</td>
<td>Dolly</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>XGBoost</td>
<td>ANTHROPIC</td>
<td>mosaicML MPT</td>
<td>LangChain</td>
</tr>
<tr>
<td>AX</td>
<td></td>
<td>PaLM 2</td>
<td>Hugging Face</td>
<td>haystack</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>stability.ai</td>
<td>jina</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Stable Diffusion</td>
<td></td>
</tr>
</tbody>
</table>

Pick the best model(s) for your use case
LLMOps, unified with DataOps + MLOps

LLM Operations for end-to-end production
- Databricks unifies LLMOps with traditional MLOps & DevOps
- Teams need to learn mental model of how LLMs coexist with traditional ML in operations

Differences to MLOps
- Internal/External Model Hub
- Fine-Tuned LLM
- Vector Database
- Model Serving
- Human Feedback in Monitoring & Evaluation
# Lakehouse AI: the best platform for AI

AI = Generative AI, LLMs & Machine Learning

<table>
<thead>
<tr>
<th></th>
<th>Separate AI Platform + Data Platform</th>
<th>Many AI tools + Data Platform</th>
<th>Lakehouse AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unified data &amp; AI governance</td>
<td>✗ Separate governance</td>
<td>✗ Some tools don’t have governance</td>
<td>✓</td>
</tr>
<tr>
<td>Centralized search and discovery Data &amp; AI</td>
<td>~ Separate search interfaces</td>
<td>❌ Some tools don’t have search</td>
<td>✓</td>
</tr>
<tr>
<td>Unified toolkit across data &amp; AI</td>
<td>✗ Separate data / AI tools</td>
<td>✗ Separate data / AI tools</td>
<td>✓</td>
</tr>
<tr>
<td>Single copy of your data</td>
<td>✗ Copy of data in each platform</td>
<td>✗ Copy of data in each tool</td>
<td>✓</td>
</tr>
<tr>
<td>Unified, automated lineage tracking</td>
<td>~ Only within each platform</td>
<td>❌ Not provided</td>
<td>✓</td>
</tr>
<tr>
<td>Performance and scale</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Integration cost</td>
<td>~ Costly effort to integrate platform</td>
<td>❌ Stitch together 10s of tools</td>
<td>✓</td>
</tr>
</tbody>
</table>
We at Databricks are transforming our business

✓ Understand customer usage: evaluate existing usage and model future trends, spikes, and growth/decline

1 Talk to your data: natural language chats with structured + unstructured data

2 Make Databricks simpler: intelligent task assistant that understands your data and documents your code

3 Smarter developer support: contextually aware help and assistance in model accuracy & maintenance
Generative AI can transform your’s

Encoding your organization’s IP and data into generative AI models unlocks significant value and efficiencies

1. **Create conversational interfaces for everything**
   - Reduce employee time spent looking for information
   - Automate and improve business processes e.g., customer sales, support, etc

2. **Human-level comprehension – but at billions of words per second**
   - Drive revenue & proactively identify problems by extracting insights from every customer interaction
   - Reduce cost by automating language-heavy processes

3. **Generate human-quality text, images, and code**
   - Increase employee productivity by drafting marketing, support, sales content
   - Drive revenue by personalizing every customer interaction
Strategic Roadmap for AI Adoption

Formulate a strategy on how you will successfully integrate this technology into your business landscape

1. Define Gen AI Strategy
   - Identify AI strategy
   - Engage business units
   - Setup ethical and legal policies
   - Define success criteria

2. Business Use Cases
   - Identify business objectives
   - Research use-cases and prioritize high value use cases
   - Data availability and alignment with use cases

3. Design & Architecture
   - Choose the right AI model architecture
   - Integrate developed model into existing business systems

4. Operations & Monitoring
   - Align your operation model
   - Automation
   - Gather feedback, continues interactive improvements

5. People & Adoption
   - Refine roles and responsibilities
   - Training and support

Organization’s Strategy & Mission
How AI can be used for achieving or accelerating business objectives?
Recommendations

- Centralize your data estate to include AI
  - Allows for Collaboration and Productivity
- Use any and all models available
  - Combine some as needed
  - Provide your controls and weights
- Don’t lose control of your data
  - Develop in house if IP data sets
  - Provide secure sharing
- Experiment

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