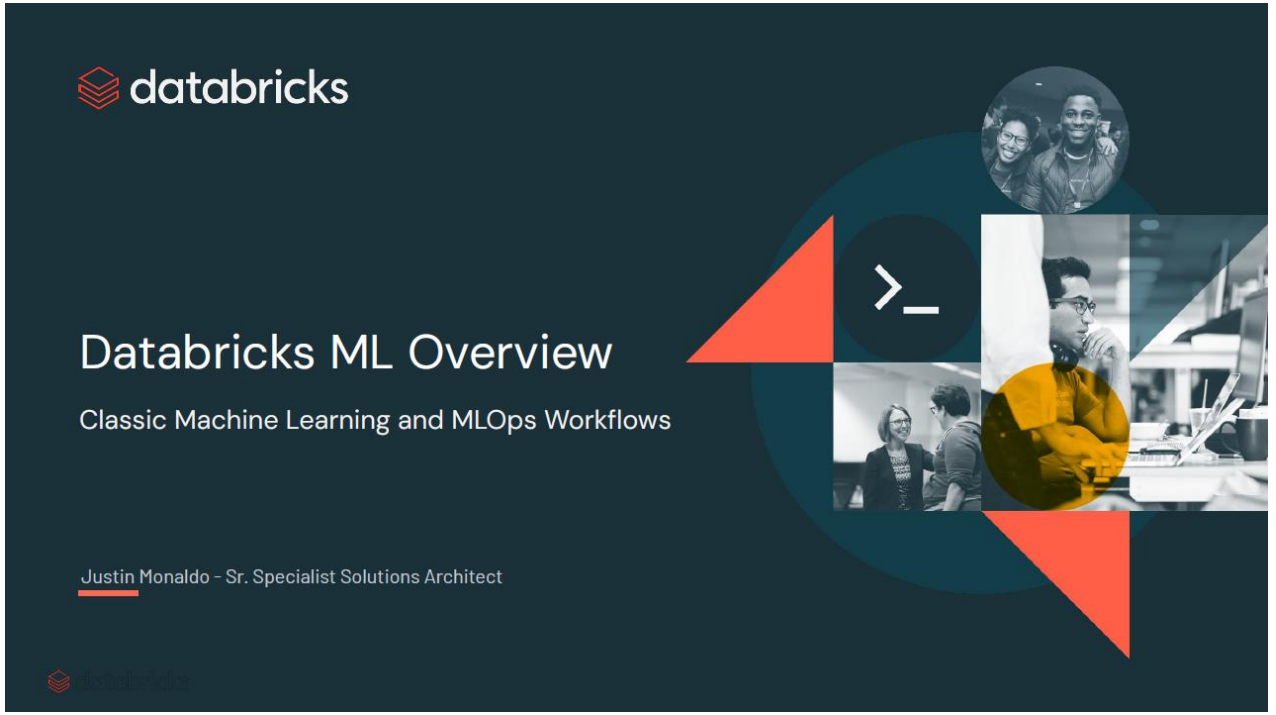
databricks

Databricks ML Overview

Slide Deck



databricks

Databricks ML Overview

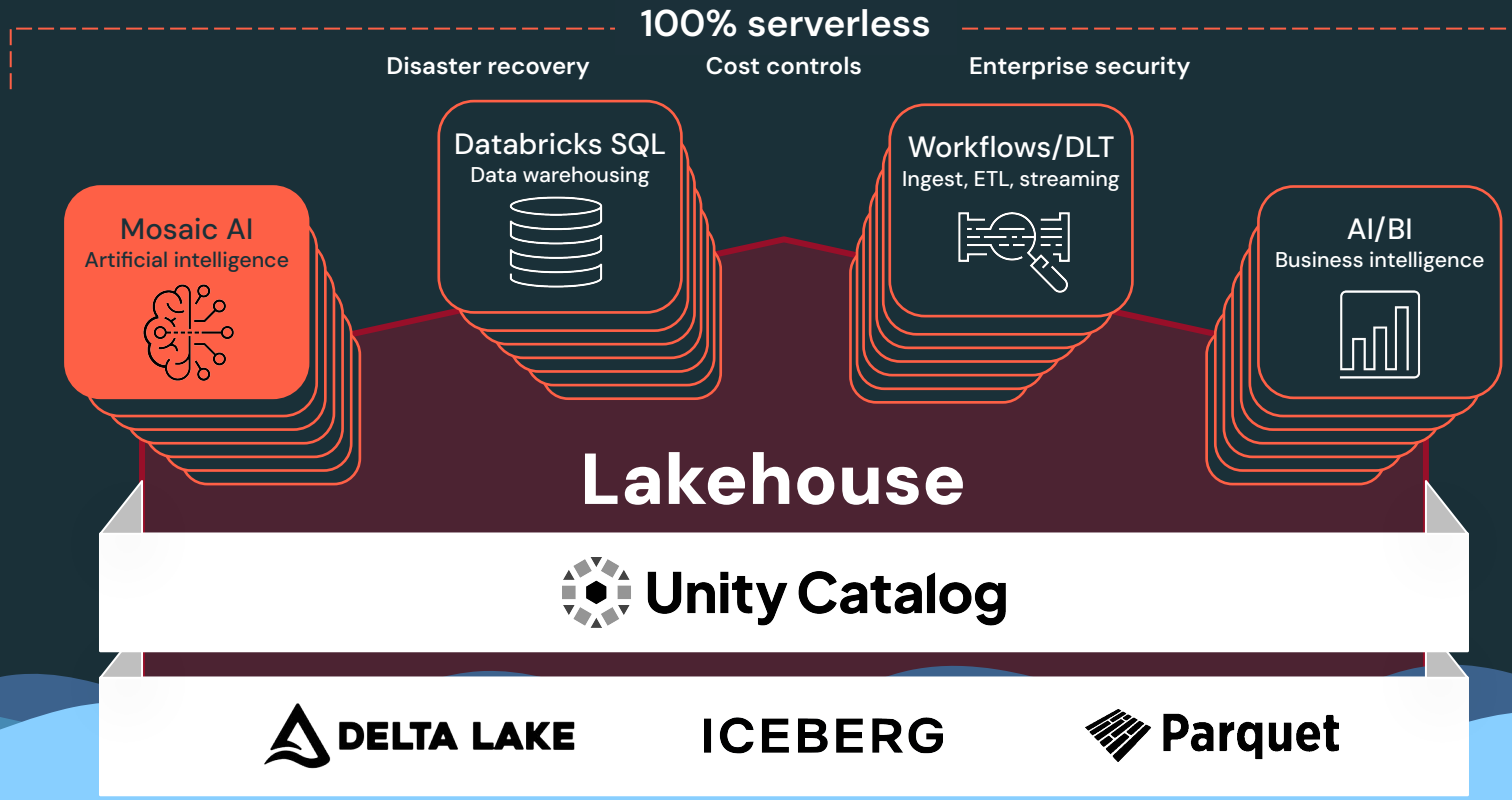
Classic Machine Learning and MLOps Workflows

Justin Monaldo - Sr. Specialist Solutions Architect

Agenda

- Overview
- MLflow
- Model Registry
- MLOps
- Feature Store
- AutoML
- Model Serving

Databricks Data Intelligence Platform



Databricks Machine Learning User Experience

Provide a collaborative environment for Unified Machine Learning & Data Analytics

Multi-Language

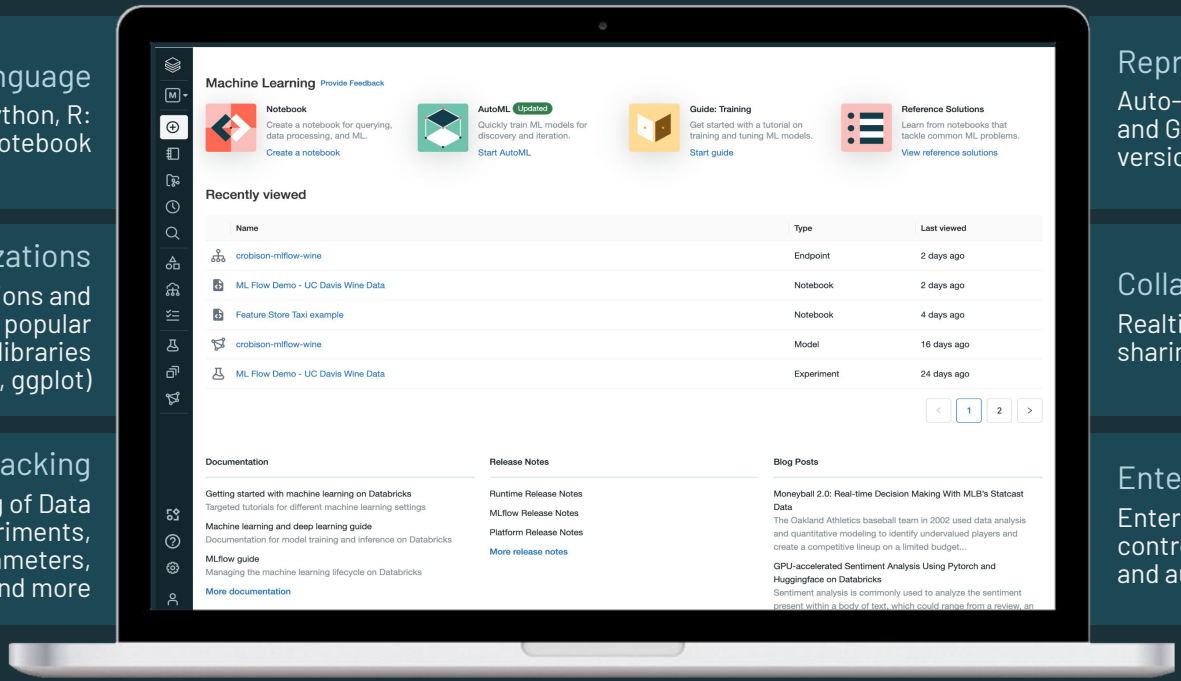
Scala, SQL, Python, R: All in one notebook

Visualizations

Built-in visualizations and support for the most popular visualization libraries (e.g. matplotlib, ggplot)

Experiment Tracking

Built-in tracking of Data Science and ML experiments, with metrics, parameters, artifacts, and more



Reproducible

Auto-logged revision history and Git integration for version control

Collaborative

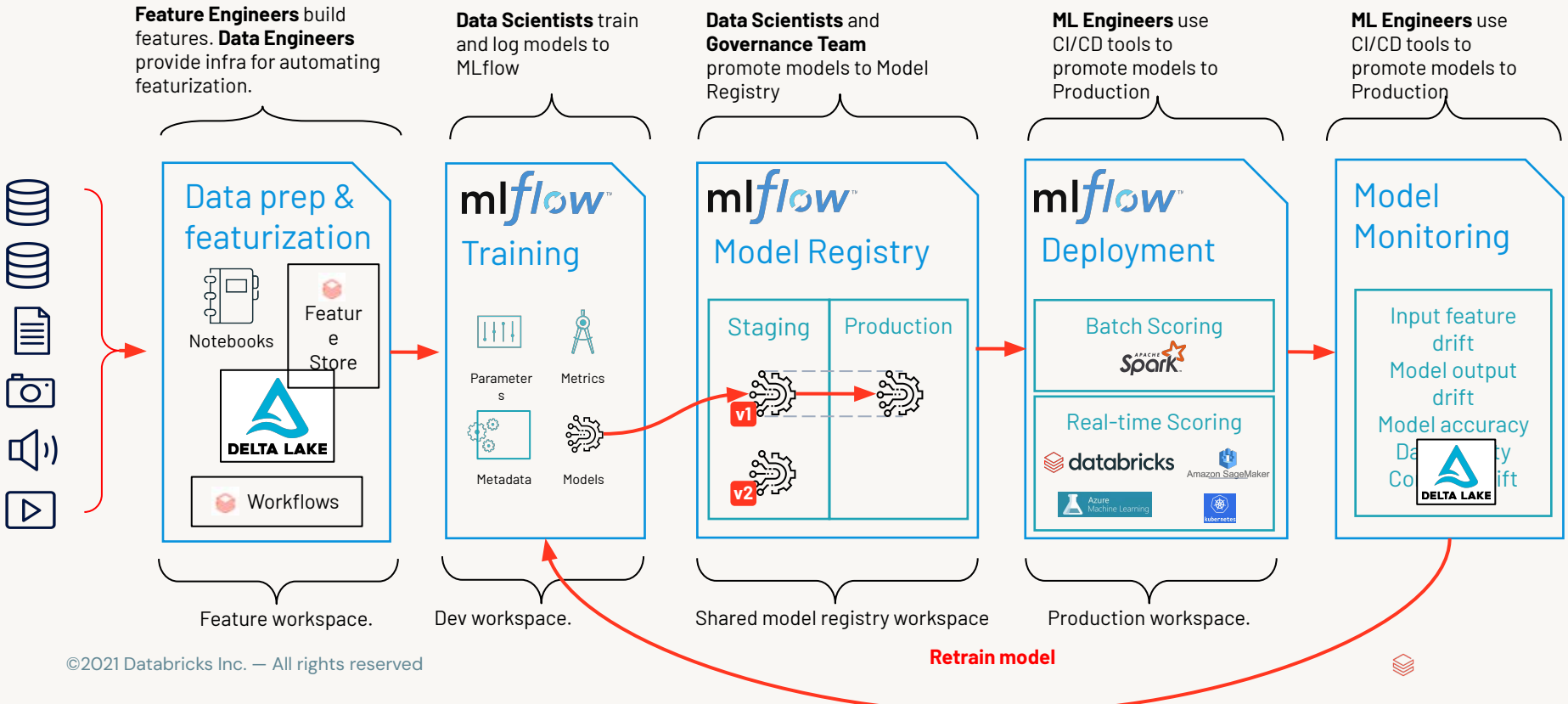
Realtime co-editing, with sharing and permissions

Enterprise Ready

Enterprise grade access controls, identity pass-through, and auditability

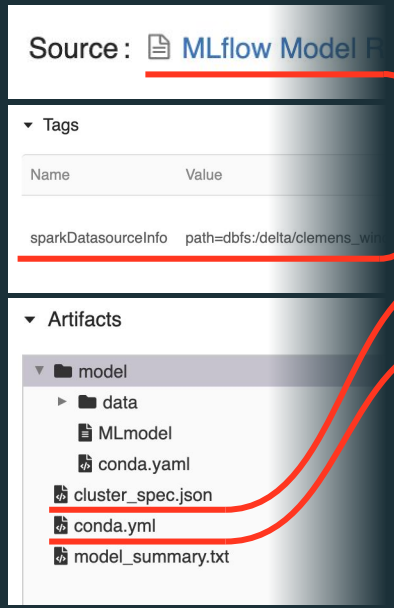
databricks

Databricks Machine Learning Pipeline



mlflow Auto-Logging for Reproducibility

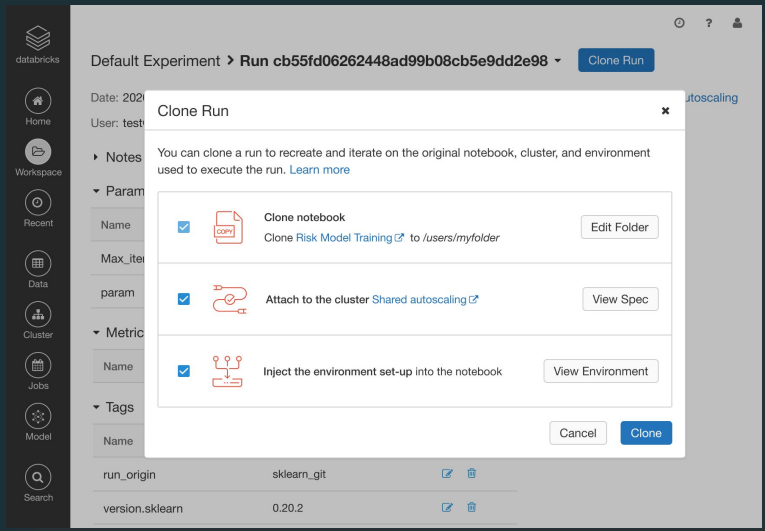
Auto-Logging of Cluster Specification and Environment Dependencies



Reproducibility Checklist:

- ✓ Code versioning
- ✓ Data versioning
- ✓ Cluster configuration
- ✓ Environment specification

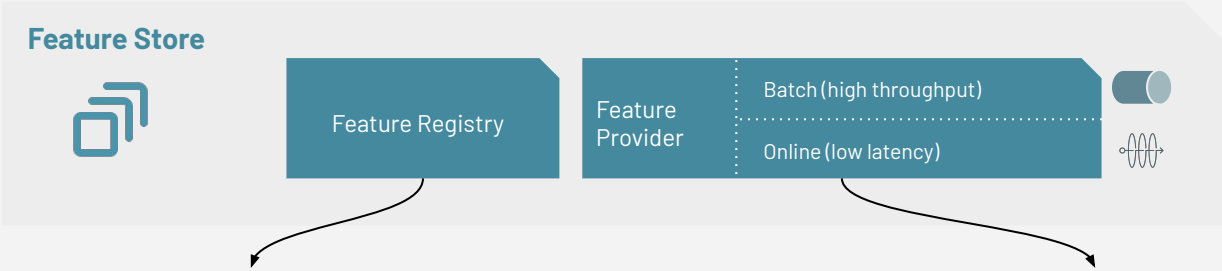
Reproduce Run Feature:



databricks

Databricks Feature Store

The first Feature Store codesigned with a data and MLOps platform



The Feature Registry allows reuse

- Allows reuse across projects
- Discoverability and Reusability
- Versioning
- Upstream and downstream Lineage

The Feature Provider

- Prevents offline/online skew
- Batch and online access to Features
- Feature lookup packaged with Models
- Simplified deployment process

Co-designed with DELTA LAKE

- Open format
- Built-in data versioning and governance
- Native access through PySpark, SQL, etc.

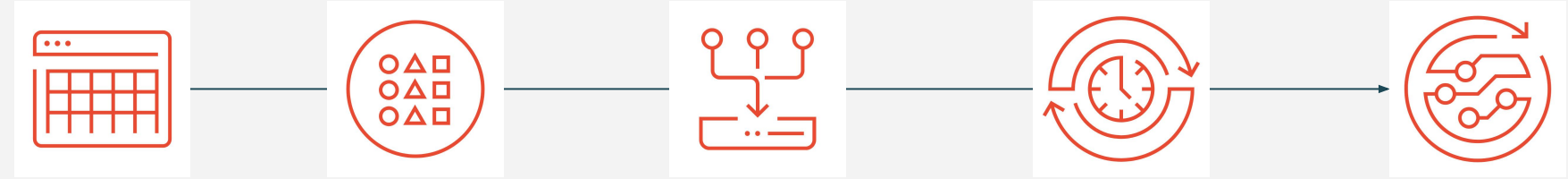
Co-designed with mlflow

- Open model format that supports all ML frameworks
- Feature version and lookup logic hermetically logged with Model

databricks

What is AutoML?

Automated machine learning (**AutoML**) is a fully-automated model development solution seeking to "democratize" machine learning. While the scope of the automation varies, AutoML technologies usually automate the ML process from data to model selection.



select a dataset

automated data prep

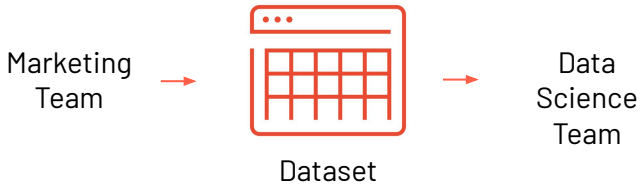
automated feature engineering and selection

automated training and model selection

automated hyperparameter tuning

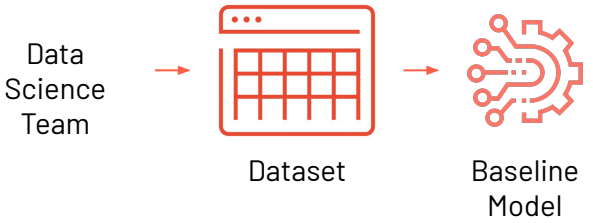
AutoML solves two key pain points for data scientists

Quickly Verify the Predictive Power of a Dataset



"Can this dataset be used to predict customer churn?"

Get a Baseline Model to Guide Project Direction

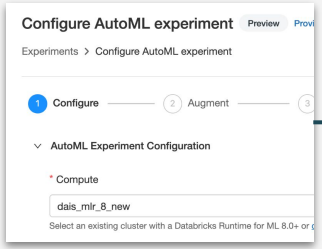


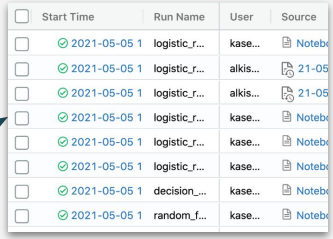
"What direction should I go in for this ML project and what benchmark should I aim to beat?"

Databricks AutoML

A glass-box solution that empowers data teams without taking away control

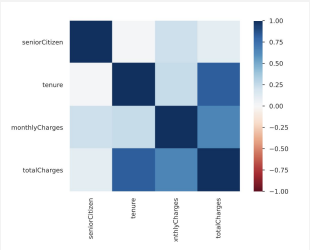
UI and API to start AutoML training





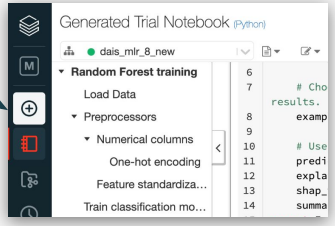
MLflow experiment
Auto-created MLflow Experiment to track models and metrics

→ Easily deploy to Model Registry



Data exploration notebook
Generated notebook with feature summary statistics and distributions

→ Understand and debug data quality and preprocessing



Reproducible trial notebooks
Generated notebooks with source code for every model

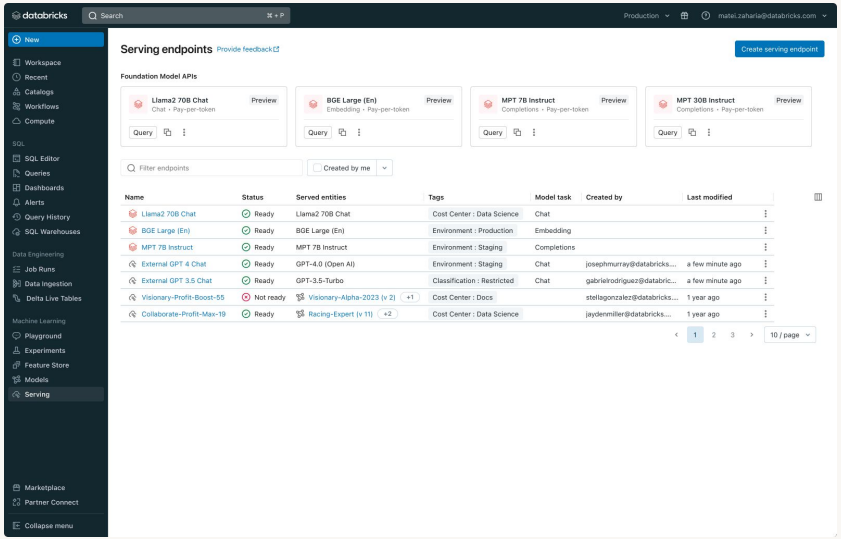
→ Iterate further on models from AutoML, adding your expertise

databricks

Databricks Model Serving

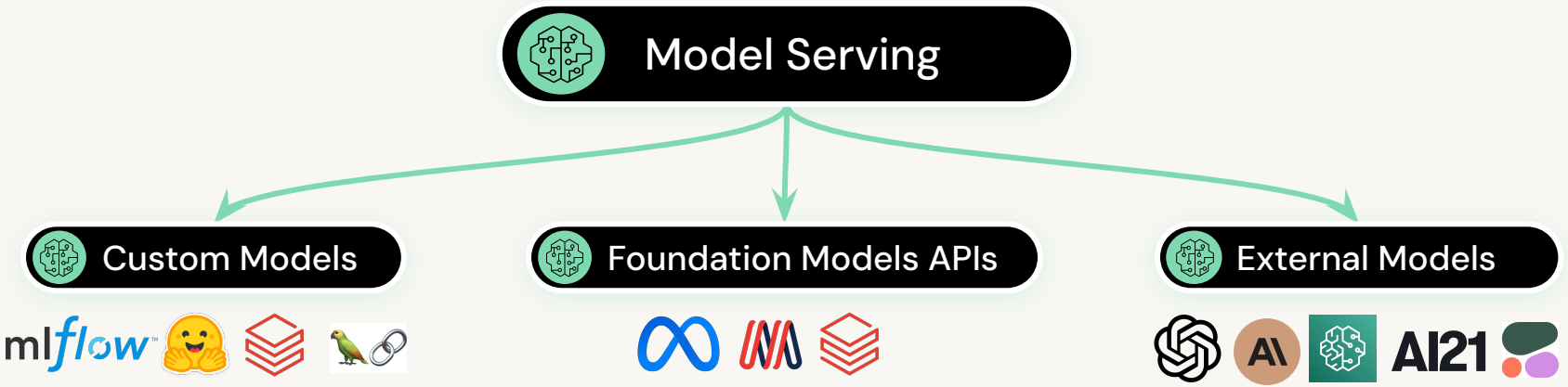
Access, Govern, and Monitor any AI Models with Databricks Model Serving

1. **Build AI apps faster with Unified Model Serving:** Deploy or query any model though a unified interface, including Custom Models, Databricks-managed Foundations Models (like base/fine-tuned Llama3), or externally-hosted models (like OpenAI GPT 3.5)
2. **Reduce TCO with Serverless Serving:** Highly available and scalable serving with LLM-specific optimization that reduces latency and cost
3. **Scale deployments with Data Intelligence Platform:** Automatic feature/vector lookups, monitoring, and unified governance that scales deployments and reduces errors



Databricks Model Serving

Unified UI, API & SDK for managing all types of AI Models



Access any AI Model, be it a fully custom model, an agent/chain model, a Databricks-managed Foundation Model, or a 3rd-party Foundation Model.

Govern and Monitor all Models in one place

Query Models via unified interface (Single API, SDK and UI)

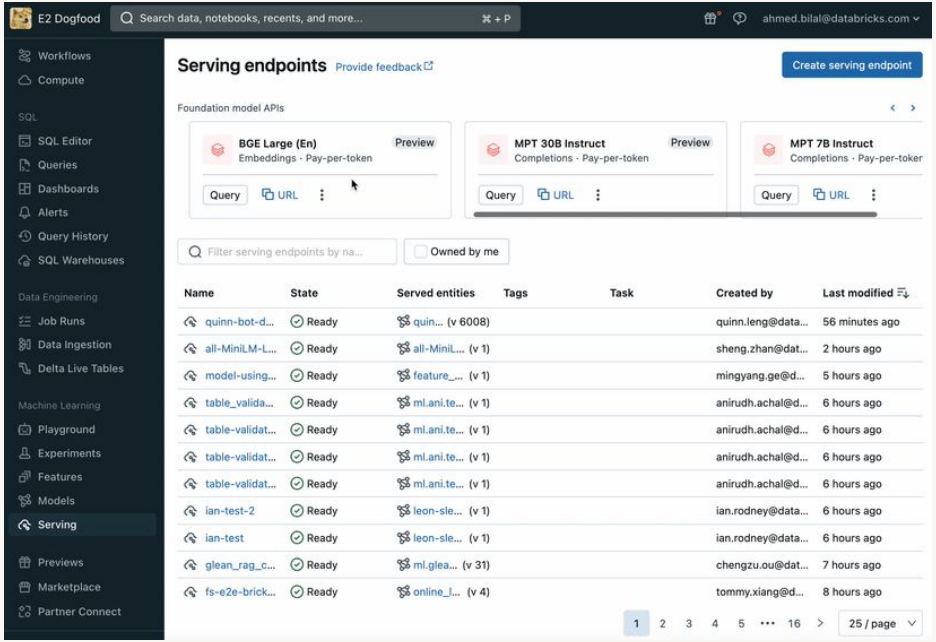
Foundation Model APIs

Fully Featured, Enterprise Grade and Secure API endpoints

Instantly access popular Foundation Models securely within Databricks

Flexible Pricing: Start with pay-per-token pricing or deploy fine-tuned models with performance guarantees on Provisioned Throughput

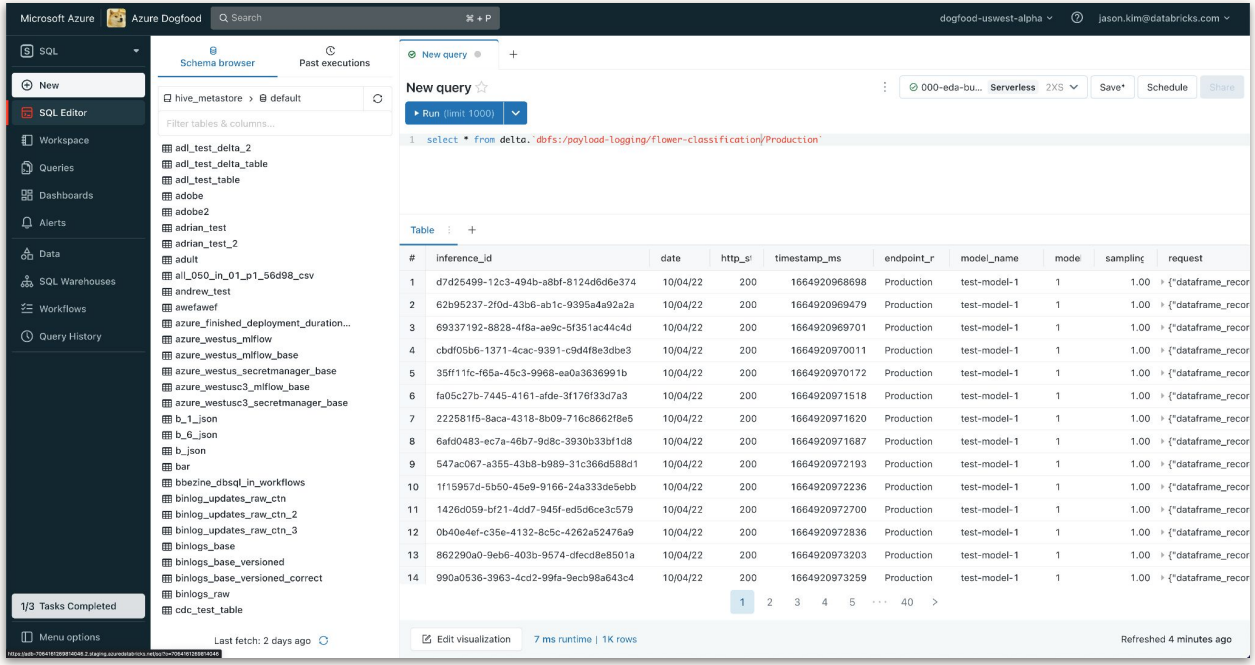
Reduce TCO with latency and throughput optimization



Monitoring with Inference Tables

Automatically capture request/responses in a Lakehouse Delta table

- **Lakehouse support**
  Query inference logs as first-class SQL data and easily integrate with DLT and DBSQL dashboards
- **Monitor model quality**
  Join ground-truth labels and alert on model performance
- **Debug models quickly**
  Investigate errors from raw requests/responses
- **Create training datasets**
  Use data from live endpoint traffic to retrain models
- **Meet compliance requirements**
  Track prediction responses to streamline compliance workflows



databricks

Thank you for downloading this Databricks presentation! Carahsoft serves as the Master Government Aggregator® and Distributor for Databricks, offering expertise in government procurement processes and practices with purchasing available via GSA, SEWP V, ITES-SW and other contract vehicles.

To learn how to take the next step toward acquiring Databricks' solutions, please check out the following resources and information:

For additional resources, please visit carah.io/DatabricksResources

For additional solutions, visit carah.io/DatabricksSolutions

To speak with our team directly, email Databricks@carahsoft.com or reach out at 703-581-6693.

To view our upcoming Databricks events, visit carah.io/DatabricksEvents

For additional Open Source solutions, visit carah.io/OpenSourceSolutions

To purchase, check out the contract vehicles available for procurement at carah.io/DatabricksContracts

carahsoft

For more information, contact Carahsoft or our reseller partners:

Databricks@carahsoft.com | 703-581-6693