databricks mlflow

Managing the Complete Machine Learning Lifecycle

Successfully building and deploying a machine-learning model is difficult enough to achieve once. What's even harder is enabling other data scientists to reproduce a pipeline, compare the results of different versions, track what's running where, and redeploy and rollback updated models:

Challenges

MYRIAD TOOLS AND FRAMEWORKS

Hundreds of open-source tools cover each phase of the ML lifecycle, from data preparation to model training and deployment. However, unlike traditional software development, where teams aim for standardization, with ML, practitioners usually want to try every available tool to determine whether it can improve results.

HARD-TO-REPRODUCE RESULTS

Given the variety of tools, and without detailed tracking, teams often have trouble getting the same code to work again. Whether a data scientist needs to pass training code to an engineer for use in production or go back to past work to debug a problem, reproducing ML workflow steps is extremely challenging and yet critically important.

HARD-TO-TRACK EXPERIMENTS

The results of ML models depend on the algorithms used, data set quality, and parameter values for the models. Whether practitioners work alone or together, it is difficult to track which parameters, code and data went into each experiment to produce a model.

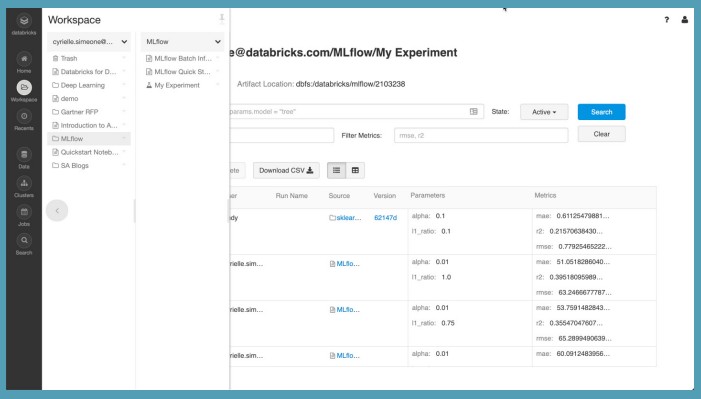
HARD-TO-DEPLOY ML MODELS

Moving a model to production can be challenging due to the variety of deployment tools and environments it needs to run (e.g., REST serving, batch inference, or mobile apps). There is no standard way to move models from any library to any of these tools, thereby creating a new risk with each new deployment.

MLflow is an open source platform unveiled by Databricks for managing the complete machine learning lifecycle. It is open and extensible by design, and platform agnostic for maximum flexibility. With MLflow, data scientists can track and share experiments locally or in the cloud, package and share models across frameworks, and deploy models virtually anywhere.

mlflow

Key Benefits



EXPERIMENT TRACKING

Quickly run experiments with any ML library, framework, or language, locally or in the cloud, and on any platform.

Automatically keep track of parameters, results, code, and data from each experiment, and reproduce runs.

Automatically launch thousands of runs in parallel for hyperparameter tuning and cross-validation, log results, and find the best performers.

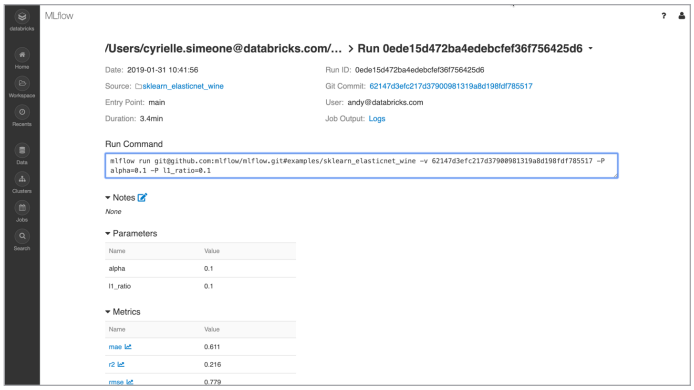
Interactively explore results in one place, and identify best performing models across multiple users.

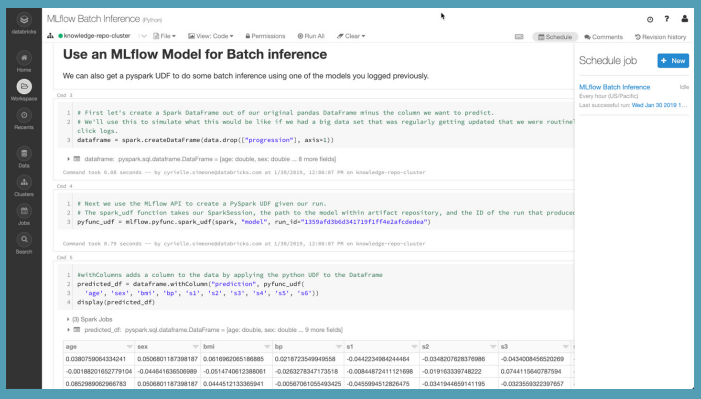
REPRODUCIBLE PROJECTS

Package projects with a standard format that integrates with Git and Anaconda and capture dependencies like libraries, parameters, and data.

Share knowledge and quickly test new ideas by running MLflow projects available on GitHub or other file storage systems.

Quickly launch reproducible runs remotely from your laptop as a Databricks job.





MODELS DEPLOYMENT

Quickly deploy models to any platforms based on your needs, locally or in the cloud, from experimentation to production.

Run batch inference in near real-time and unmatched performance using Apache Spark.™

Expose models via REST APIs with built-in integration with Docker containers, Azure ML, or Amazon SageMaker.

Introducing MLflow Model Registry

MLflow provides a simple way to manage the machine learning lifecycle, from experimentation to production. Now with MLflow Model Registry, practitioners can also use one central place to share ML models, collaborate on moving them from experimentation to online testing and production, integrate with approval and governance workflows, and monitor ML deployments and their performance.

ONE COLLABORATIVE HUB

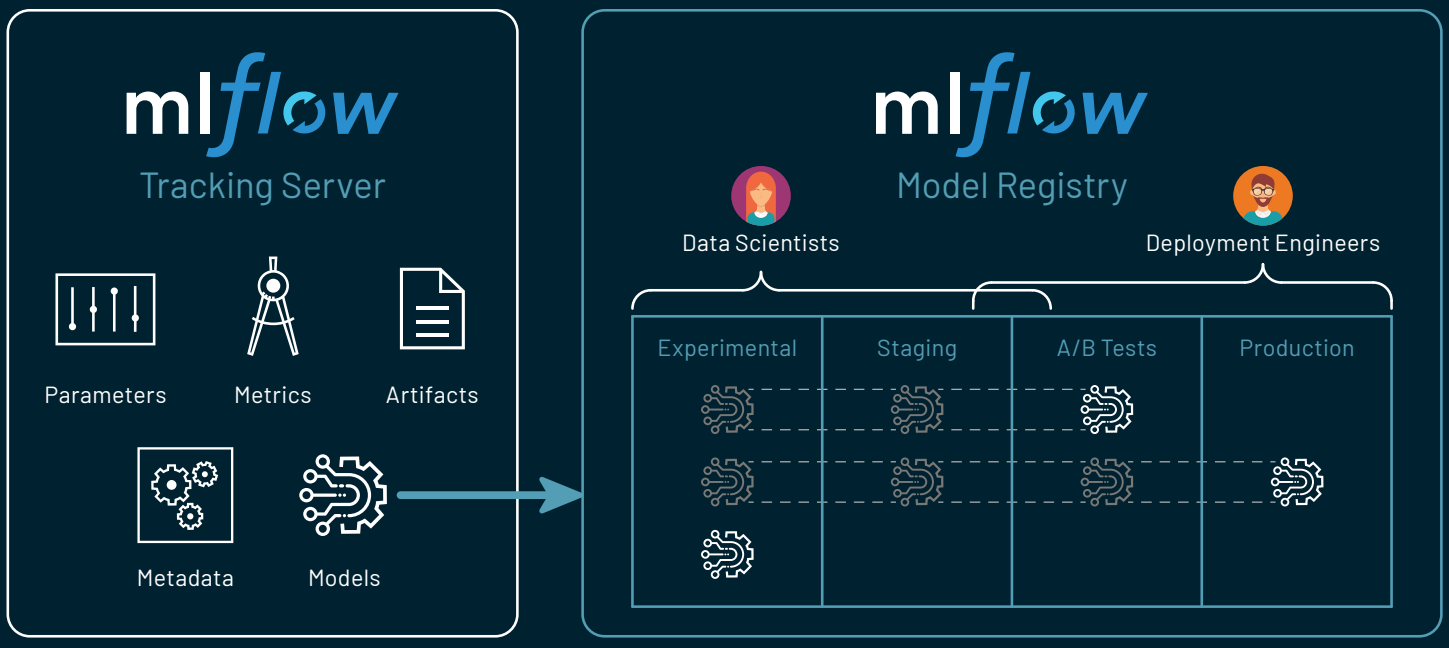
Facilitate the sharing of expertise and knowledge about building and deploying machine learning models by making models more discoverable, and providing collaborative features to jointly improve on common ML tasks.

FLEXIBLE CI/CD PIPELINES

Stay in control of machine learning models by either automatically transitioning a model into production based on predefined conditions, or manually controlling and validating lifecycle stage changes for your models from the experimentation phase, to testing and production.

VISIBILITY AND GOVERNANCE

Large enterprises often have thousands of ML models in the experimentation, testing, and production phases at any point in time. The MLflow Model Registry provides full visibility and enables governance of each by keeping track of model history and managing who can approve changes.



The managed MLflow Model Registry is now available to all Databricks customers in preview.
Contact your sales representative to get started.

Providinging MLflow as-a-service on Databricks

databricks mlflow

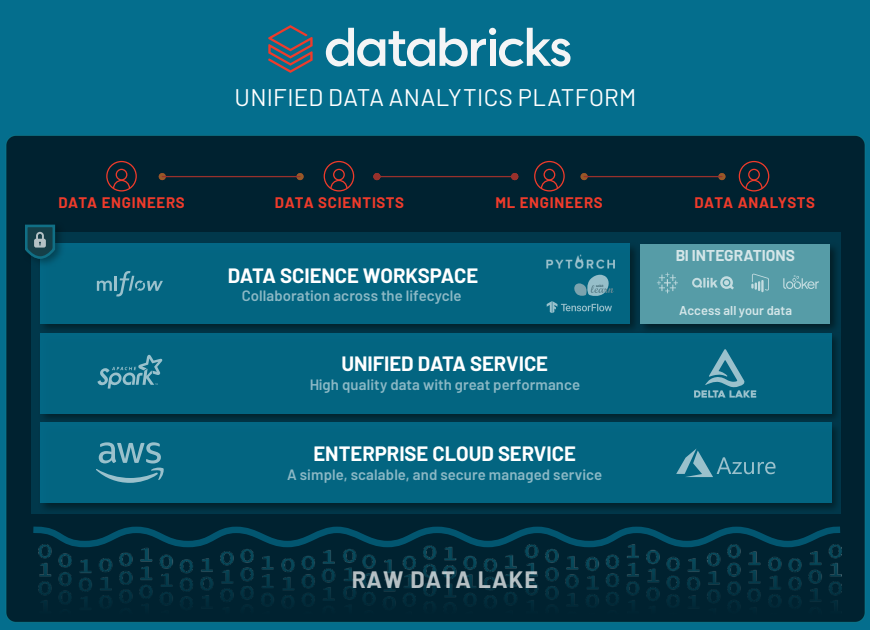
By using managed MLflow on Databricks, practitioners can benefit from out-of-the-box and seamless models tracking, packaging, and deployment capabilities with enterprise reliability, security and scale.

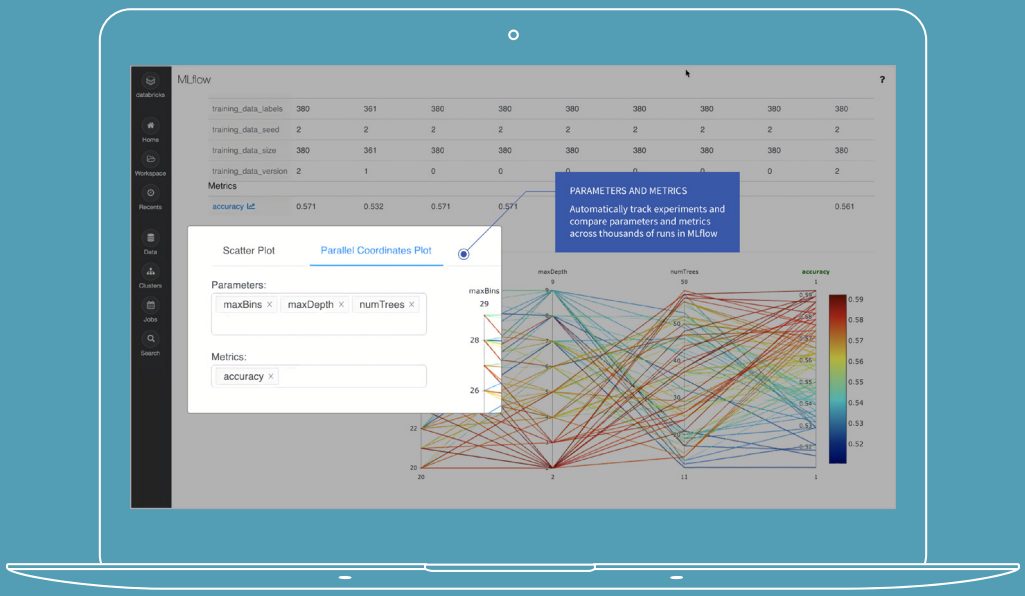
WORKSPACES
Collaboratively track and organize experiments from the Databricks Workspace

DATA VERSIONING
Track large-scale data sets that fed models with Delta Lake time travel

JOBS
Execute runs as Databricks jobs remotely or directly from Databricks notebooks

SECURITY
Take advantage of one common security model for the entire ML lifecycle





databricks

Visit databricks.com/mlflow to get started.