Get Started with Machine Learning on Databricks

We will start at 3 minutes past the starting time...

Meet your instructor

Venkita Krishnan Mani, Technical Instructor

linkedin.com/in/venkitakrishnan

Now

- Technical Instructor

Then

- Big Data & Spark Consultant
  - JPMC / Citi Bank / Deutsche Bank / Zarantec
- Hadoop Practice Lead –
  - Nichetek / Collabera India / CavalierIT

Interests

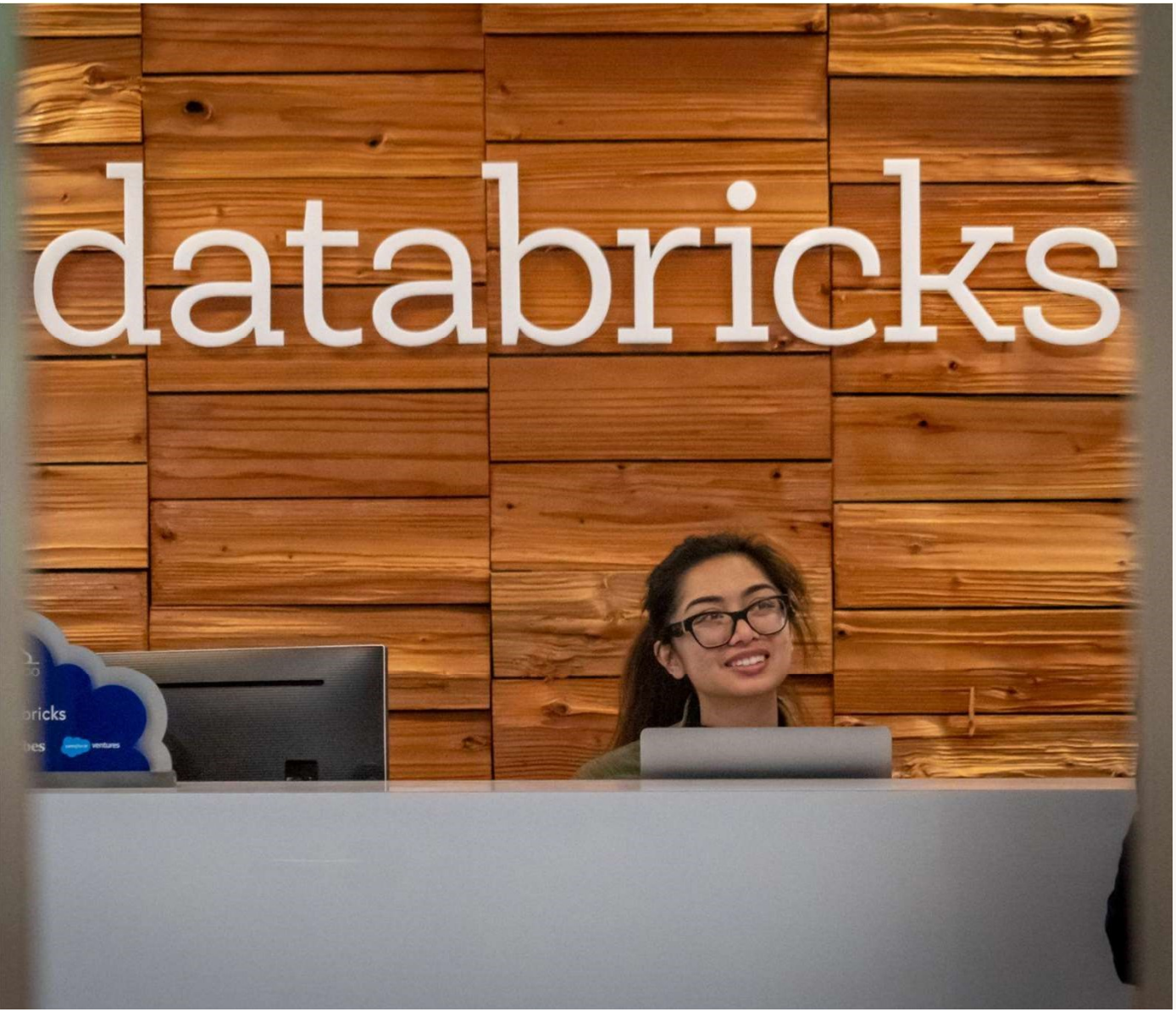
- Consulting / Teaching & Mentoring

Get Started with Machine Learning on Databricks

Session topics

- Databricks Lakehouse Platform overview
- Databricks Machine Learning feature dive/demos:
  - UI tour
  - Creating a table with Feature Store
  - Developing a baseline model with AutoML
  - Managing the ML lifecycle using Model Registry
  - Deploying a model for batch inference
  - Scheduling a model refresh with with Databricks Workflows

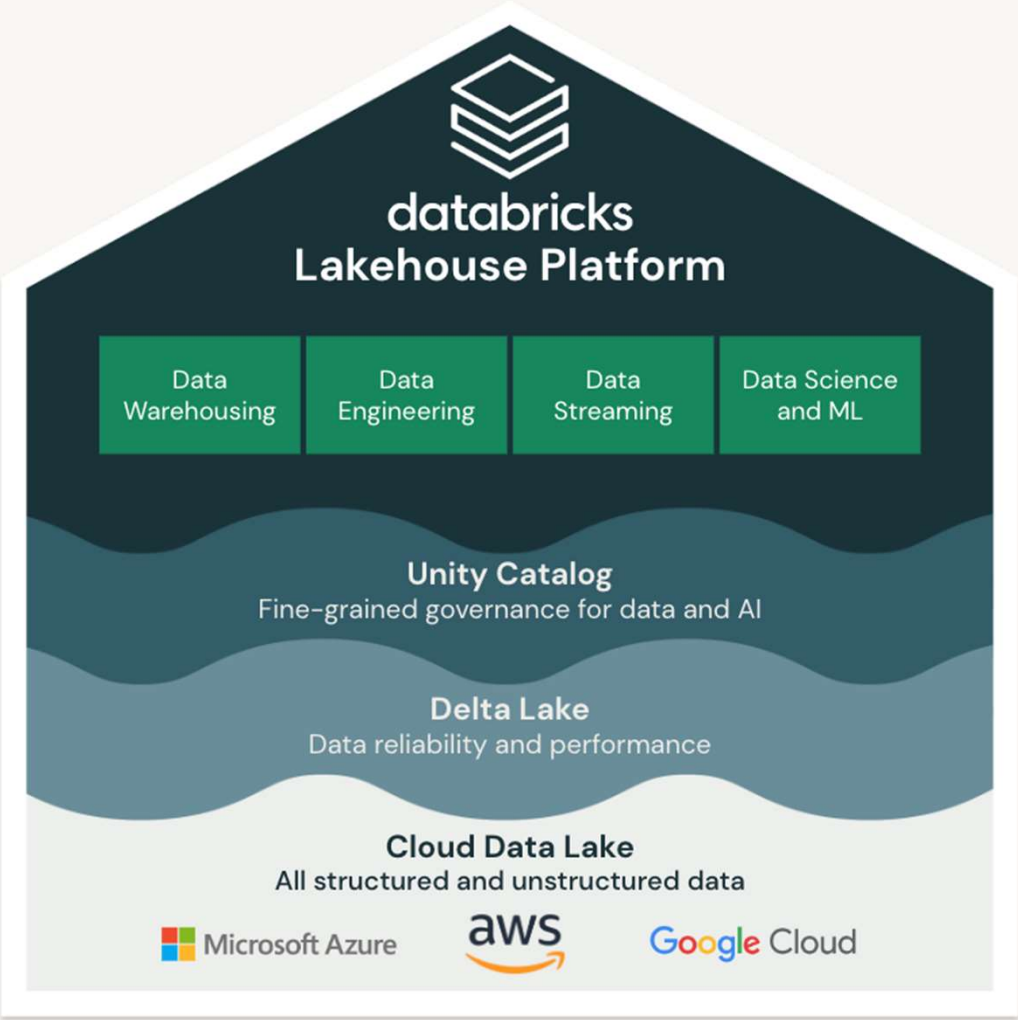
Databricks Lakehouse Platform Overview



APACHE Spark™

DELTA LAKE

mlflow™



Databricks Lakehouse Platform

Simple

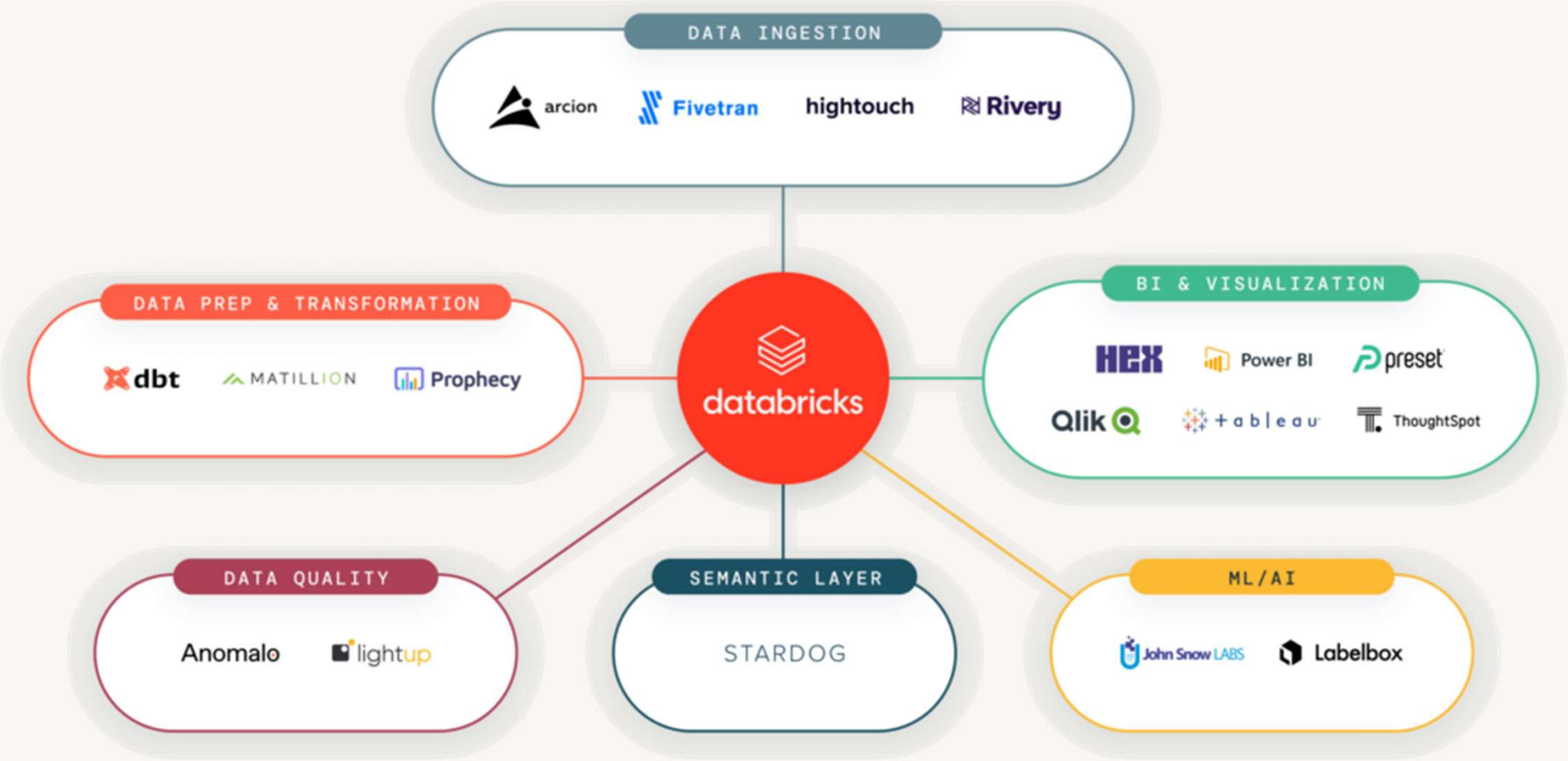
Unify your data warehousing and AI use cases on a single platform

Open

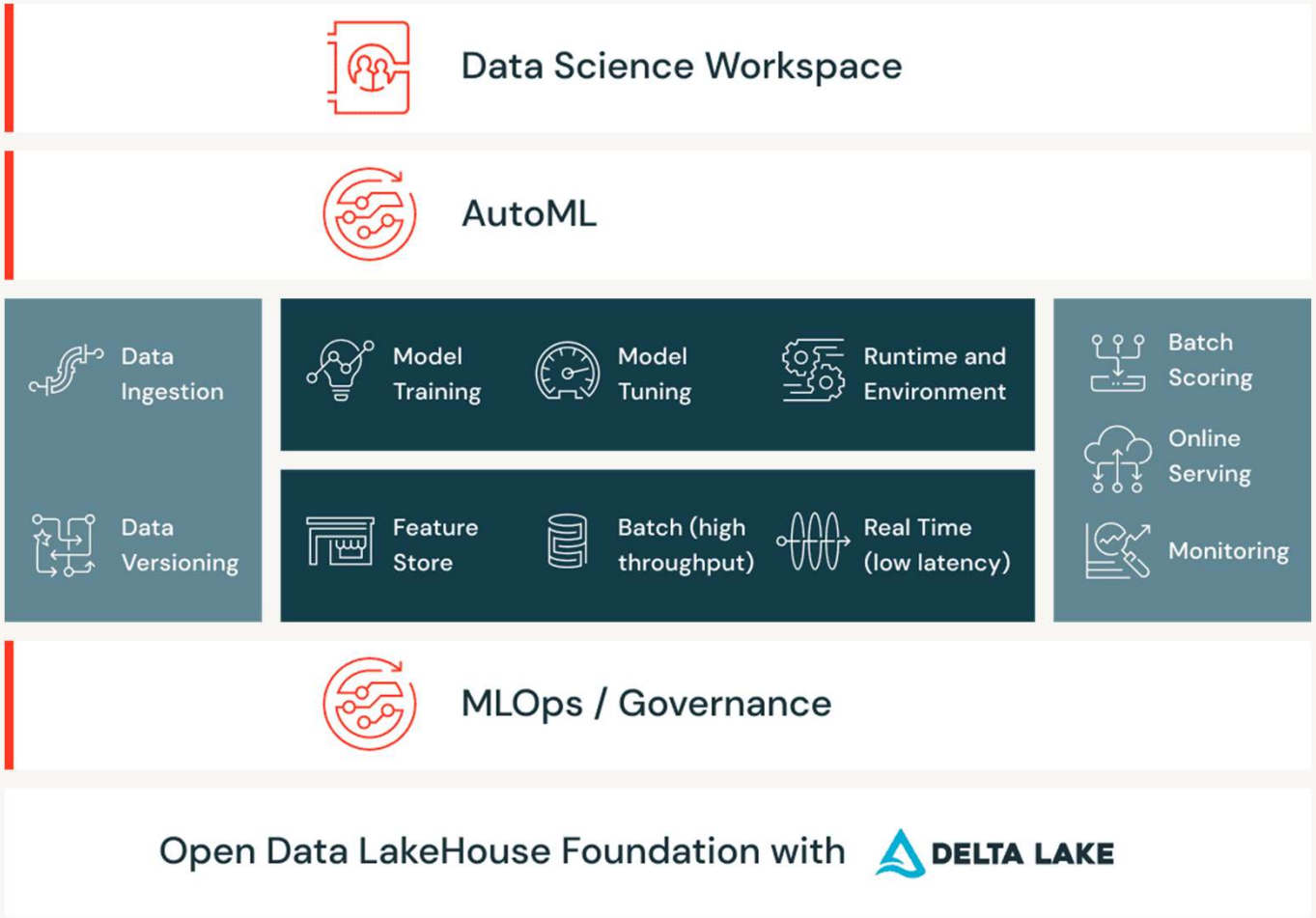
Built on open source and open standards

Multicloud

One consistent data platform across clouds



Introduction to Databricks Machine Learning



Benefits of using Databricks ML

- Production machine learning depends on code and data.

Benefits of using Databricks ML

- Production machine learning depends on code and data.
- Supports git repositories for version control.

Benefits of using Databricks ML

- Production machine learning depends on code and data.
- Supports git repositories for version control.
- Full team collaboration support.

Benefits of using Databricks ML

- Production machine learning depends on code and data.
- Supports git repositories for version control.
- Full team collaboration support.
- Development and production workflow support.

Databricks Machine Learning Runtime

- Optimized and preconfigured ML Frameworks
- Turnkey distributed ML
- Built-in AutoML
- GPU support out of the box

Built-in **ML Frameworks** and **model explainability**

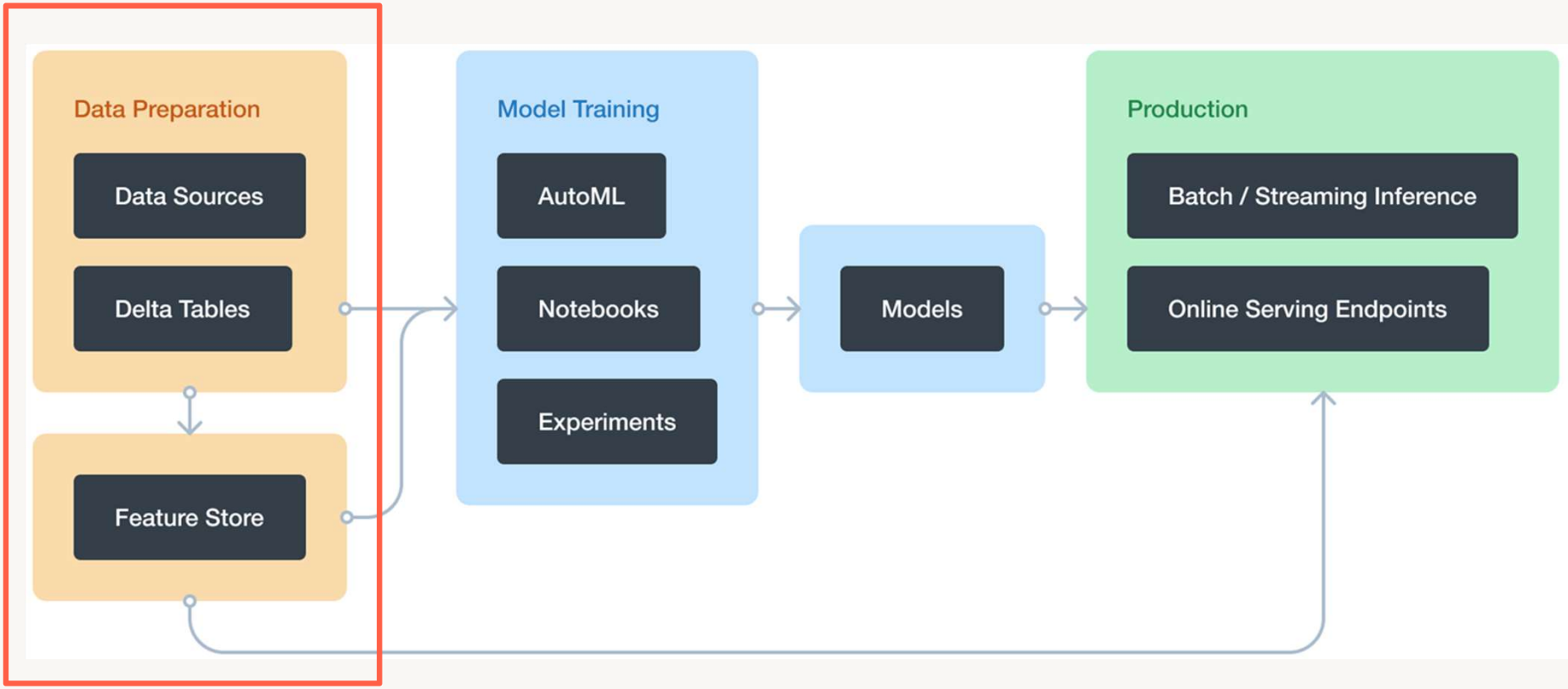
Built-in support for **distributed training**

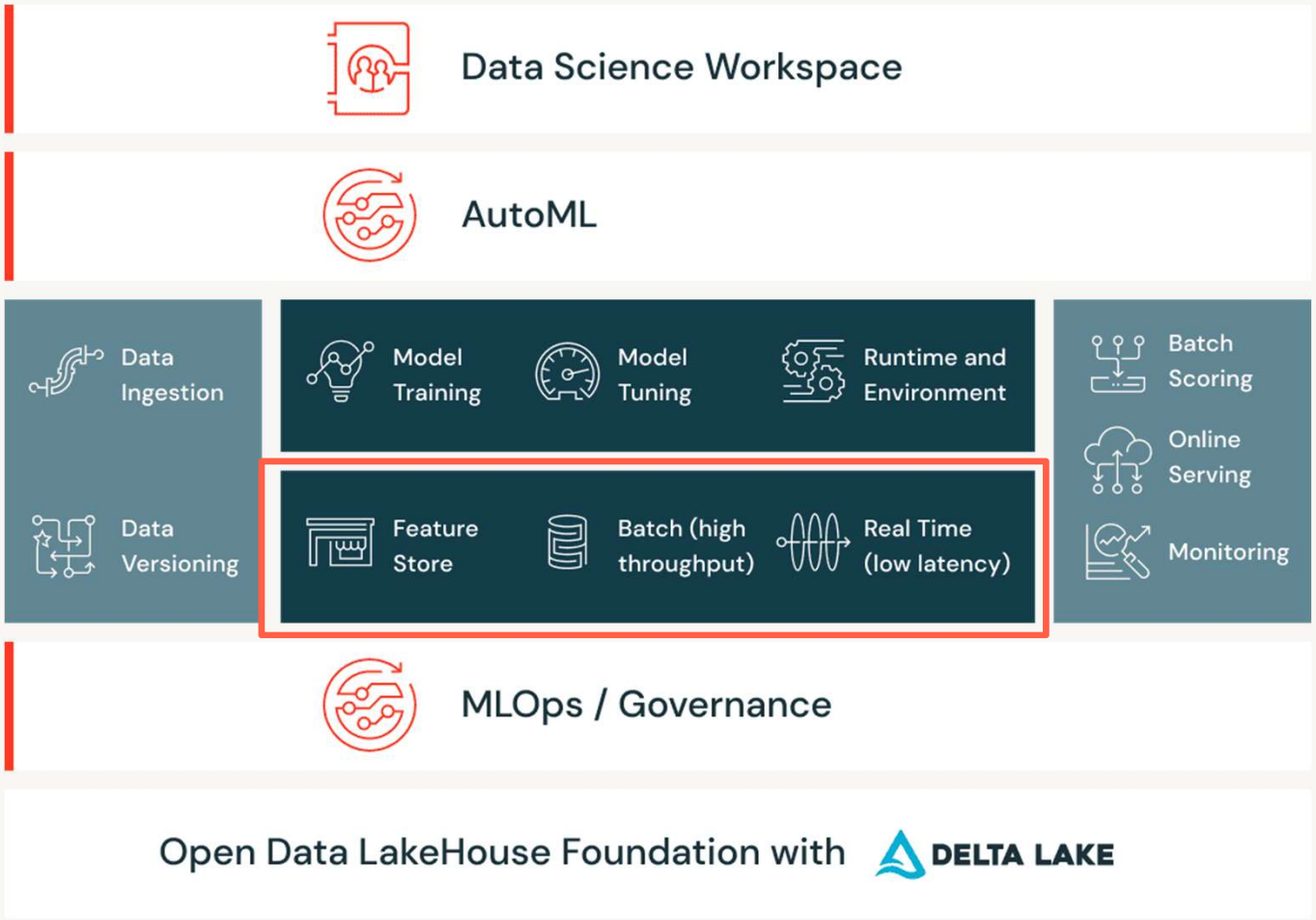
Built-in support for **AutoML** and **hyperparameter tuning**

Built-in support for **hardware accelerators**

Tour of the Databricks ML user-interface

Databricks ML Data Preparation with Feature Store





Benefits of the Feature Store

- Organization wide access and availability.

Benefits of the Feature Store

- Organization wide access and availability.
- Feature Registry for tracking and reuse.

Benefits of the Feature Store

- Organization wide access and availability.
- Feature Registry for tracking and reuse.
- Feature consistency between training and inference.

Delta Lake and MLflow

DELTA LAKE

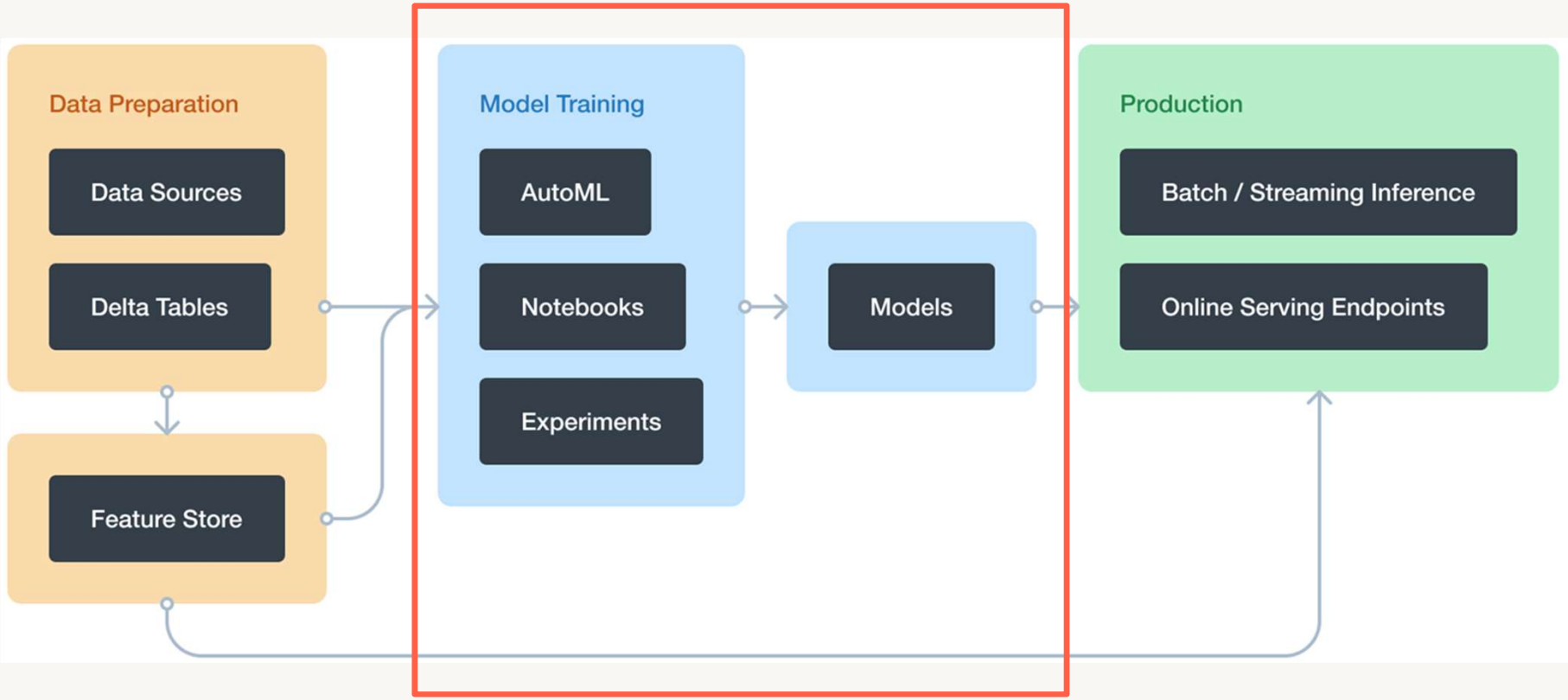
- Open storage format
- Built-in data versioning and governance
- Native access through PySpark, SQL, etc.

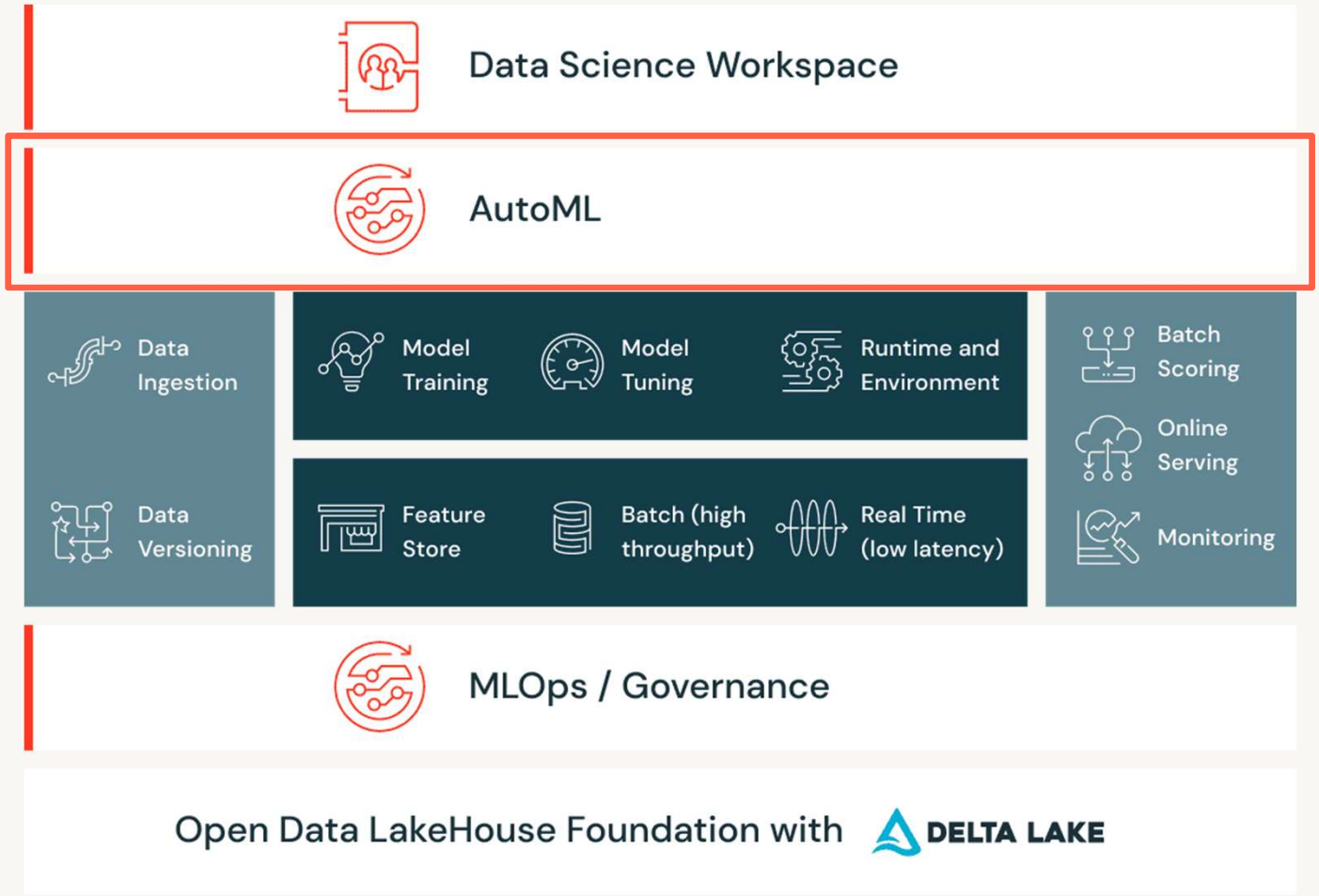
mlflow™

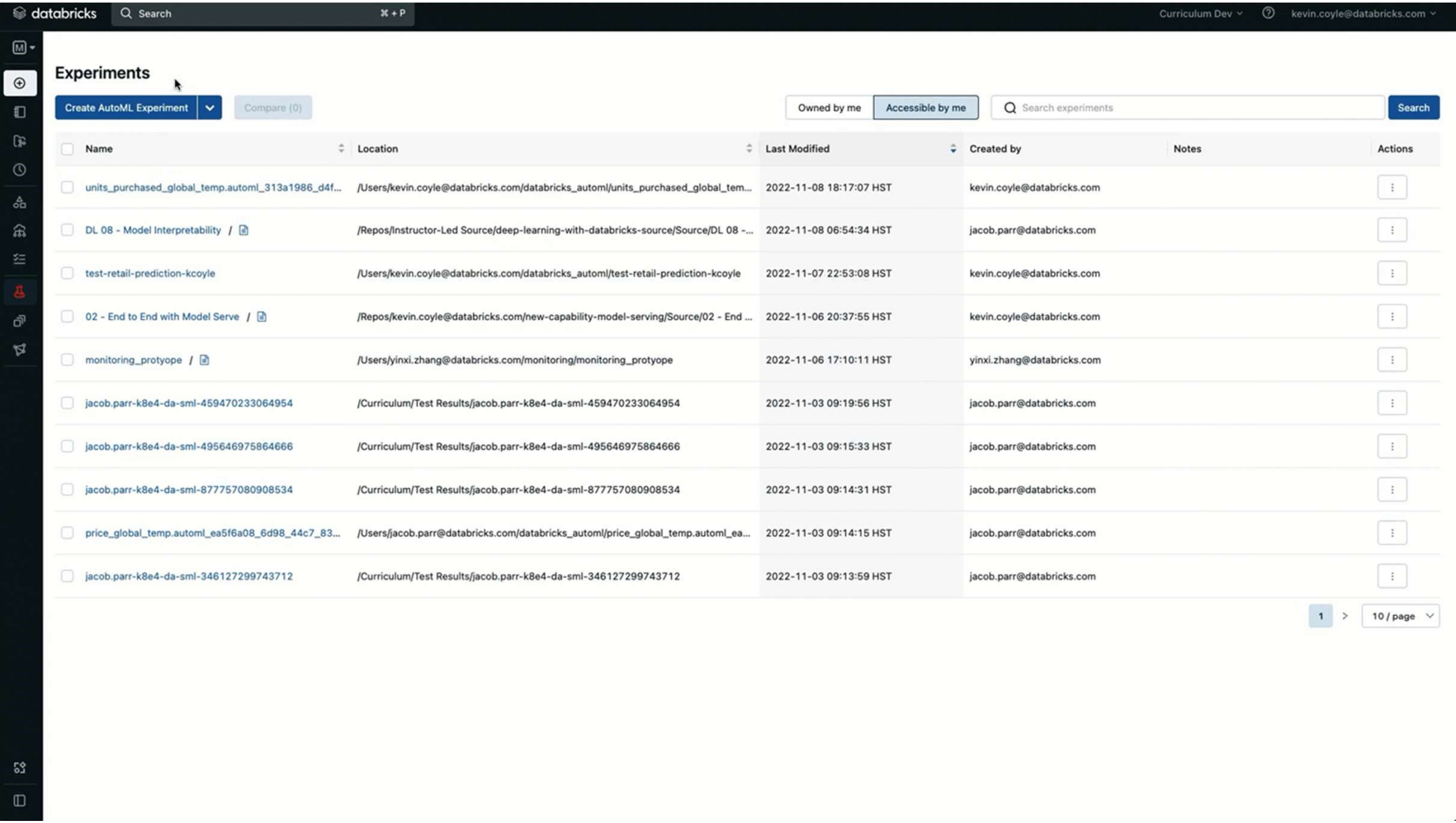
- Tracking and auto-logging
- Model Registry
- Model Serving

Demo: Use Feature Store

Model Training with Databricks ML






databricks
Search
⌘ + P
Curriculum Dev
kevin.coyle@databricks.com
Experiments
Create AutoML Experiment
Compare (0)
Owned by me
Accessible by me
Search experiments
Search
Name
Location
Last Modified
Created by
Notes
Actions
units_purchased_global_temp.automl_313a1986_d4f...
/Users/kevin.coyle@databricks.com/databricks_automl/units_purchased_global_tem...
2022-11-08 18:17:07 HST
kevin.coyle@databricks.com
DL 08 - Model Interpretability
/Repos/Instructor-Led Source/deep-learning-with-databricks-source/Source/DL 08 -...
2022-11-08 06:54:34 HST
jacob.parr@databricks.com
test-retail-prediction-kcoyle
/Users/kevin.coyle@databricks.com/databricks_automl/test-retail-prediction-kcoyle
2022-11-07 22:53:08 HST
kevin.coyle@databricks.com
02 - End to End with Model Serve
/Repos/kevin.coyle@databricks.com/new-capability-model-serving/Source/02 - End ...
2022-11-06 20:37:55 HST
kevin.coyle@databricks.com
monitoring_protyope
/Users/yinxi.zhang@databricks.com/monitoring/monitoring_protyope
2022-11-06 17:10:11 HST
yinxi.zhang@databricks.com
jacob.parr-k8e4-da-sml-459470233064954
/Curriculum/Test Results/jacob.parr-k8e4-da-sml-459470233064954
2022-11-03 09:19:56 HST
jacob.parr@databricks.com
jacob.parr-k8e4-da-sml-495646975864666
/Curriculum/Test Results/jacob.parr-k8e4-da-sml-495646975864666
2022-11-03 09:15:33 HST
jacob.parr@databricks.com
jacob.parr-k8e4-da-sml-877757080908534
/Curriculum/Test Results/jacob.parr-k8e4-da-sml-877757080908534
2022-11-03 09:14:31 HST
jacob.parr@databricks.com
price_global_temp.automl_ea5f6a08_6d98_44c7_83...
/Users/jacob.parr@databricks.com/databricks_automl/price_global_temp.automl_ea...
2022-11-03 09:14:15 HST
jacob.parr@databricks.com
jacob.parr-k8e4-da-sml-346127299743712
/Curriculum/Test Results/jacob.parr-k8e4-da-sml-346127299743712
2022-11-03 09:13:59 HST
jacob.parr@databricks.com
1
10 / page

Required parameters:

- `dataset` - Input Spark or pandas DataFrame that contains training features and targets. If using a Spark DataFrame, it will convert it to a Pandas DataFrame under the hood by calling .toPandas() - just be careful you don't OOM!
- `target_col` - Column name of the target labels

We will also specify these optional parameters:

- `primary_metric` - Primary metric to select the best model. Each trial will compute several metrics, but this one determines which model is selected from all the trials. One of `r2` (default, R squared), `mse` (mean squared error), `rmse` (root mean squared error), `mae` (mean absolute error) for regression problems.
- `timeout_minutes` - The maximum time to wait for the AutoML trials to complete. `timeout_minutes=None` will run the trials without any timeout restrictions
- `max_trials` - The maximum number of trials to run. When `max_trials=None`, maximum number of trials will run to completion.

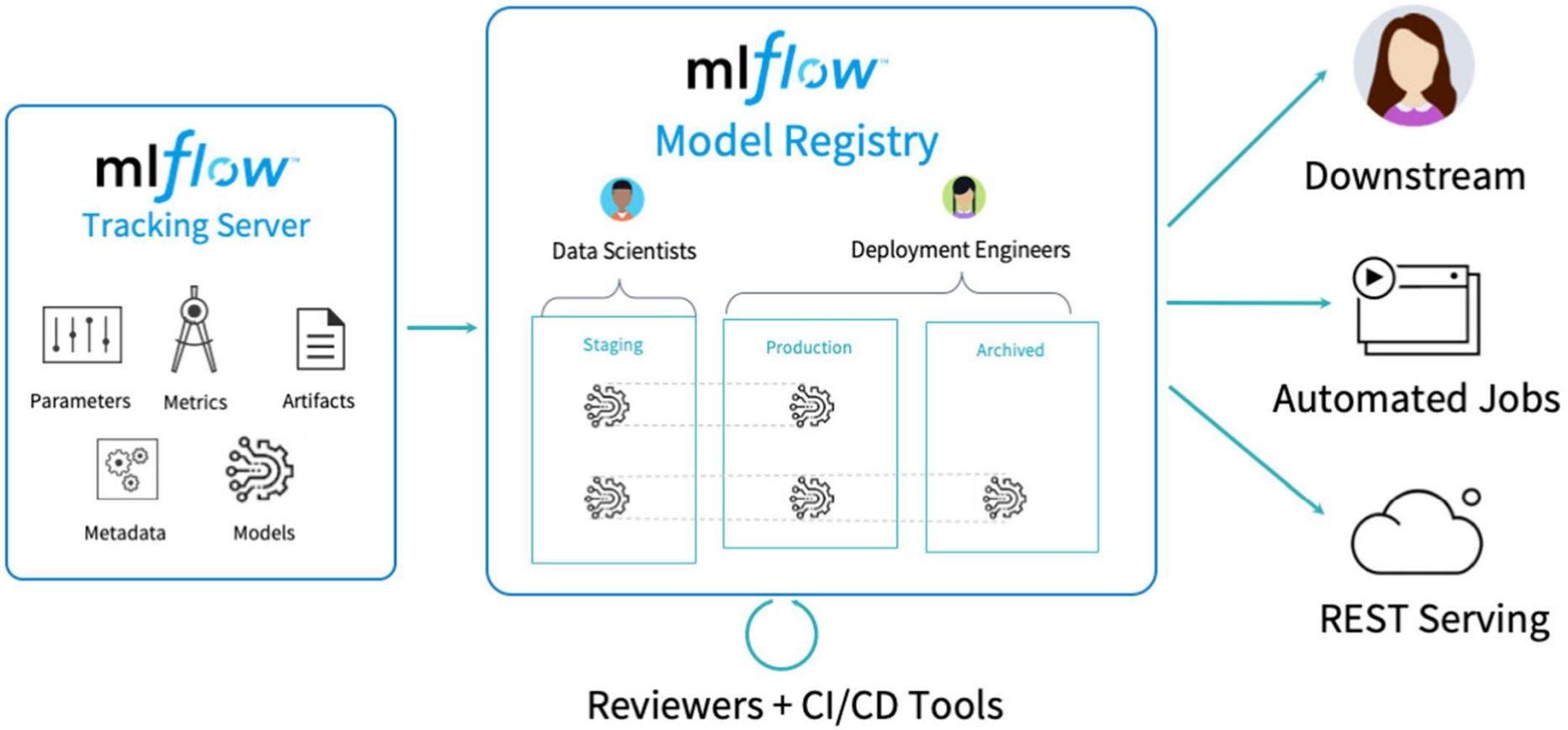
```
from databricks import automl

summary = automl.regress(train_df, target_col="price", primary_metric="rmse", timeout_minutes=5, max_trials=10)
```

Demo: Using AutoML

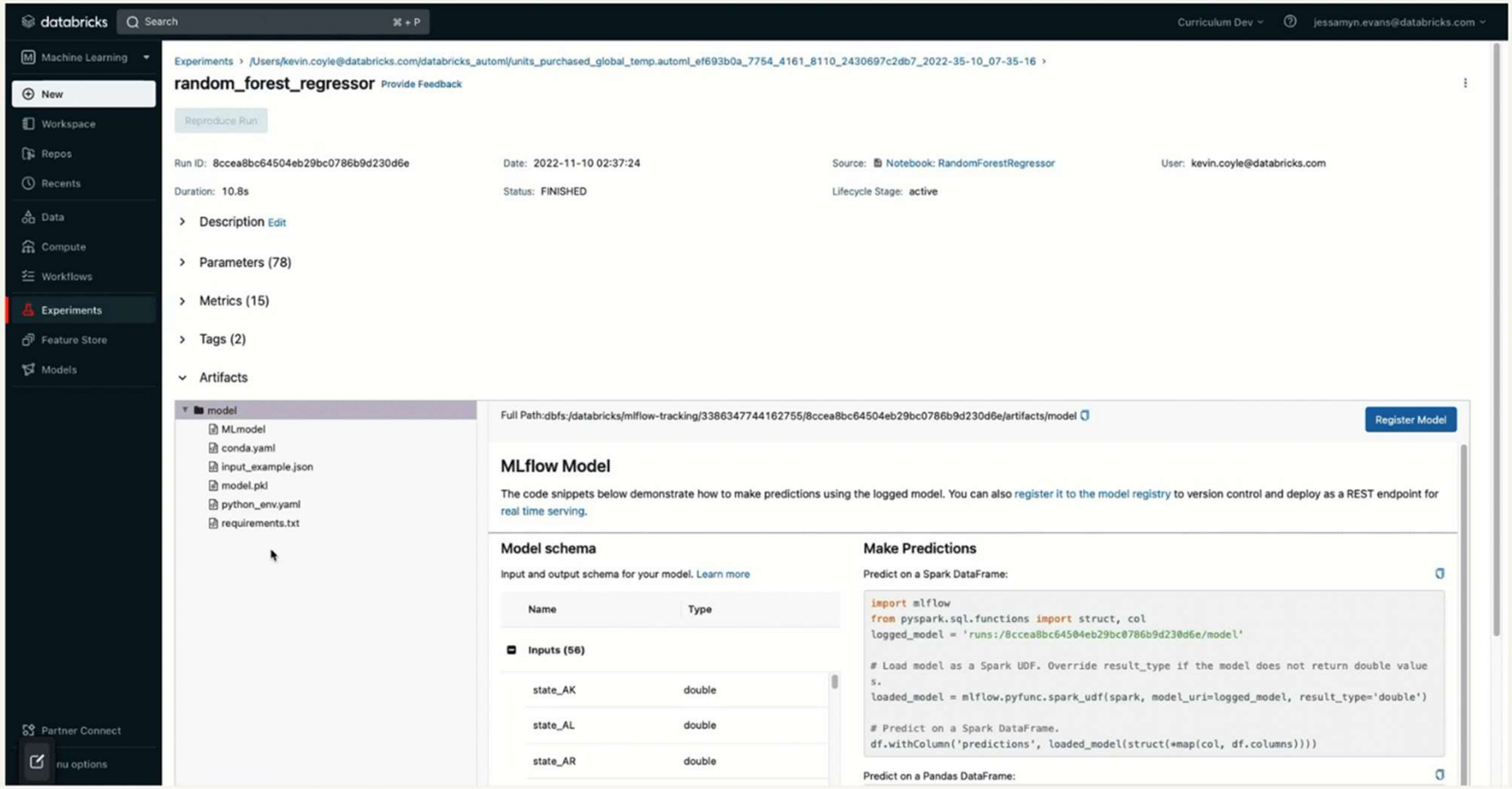
What is MLflow?

mlflow™

An open-source platform for managing the end-to-end machine learning lifecycle.



MLflow Tracking and Autologging



MLflow Models

MLflow Model

The code snippets below demonstrate how to make predictions using the logged model. You can also register it to the model registry to version control and deploy as a REST endpoint for real time serving.

Model schema

Input and output schema for your model. Learn more

Name	Type
Inputs (56)	
state_AK	double
state_AL	double
state_AR	double
state_AZ	double
state_CA	double
Outputs (1)	
-	Tensor (dtype: float64, shape:

Make Predictions

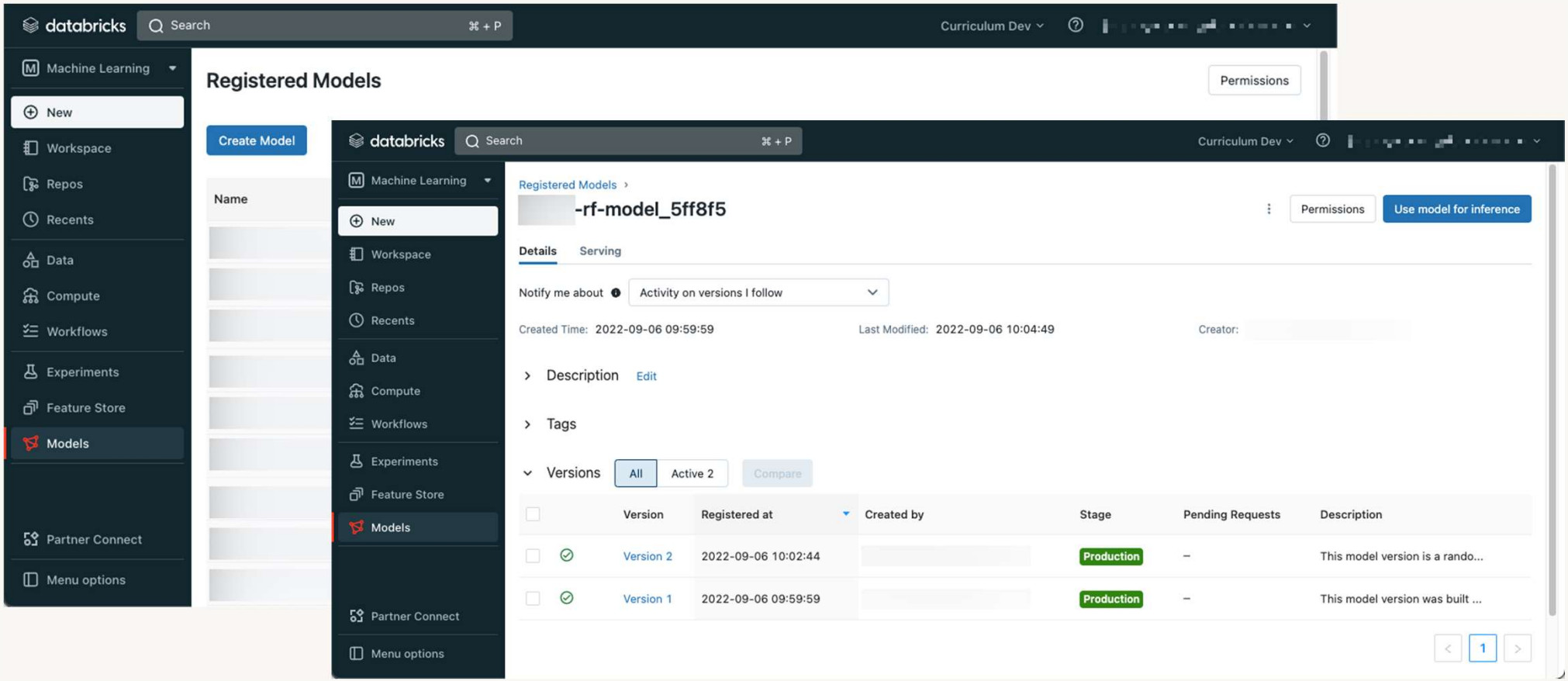
Predict on a Spark DataFrame:

```
import mlflow
from pyspark.sql.functions import struct, col
logged_model = 'runs:/8ccea8bc64504eb29bc0786b9d230d6e/model'

# Load model as a Spark UDF. Override result_type if the model does not return double value
s.
loaded_model = mlflow.pyfunc.spark_udf(spark, model_uri=logged_model, result_type='double')

# Predict on a Spark DataFrame.
df.withColumn('predictions', loaded_model(struct(*map(col, df.columns))))
```

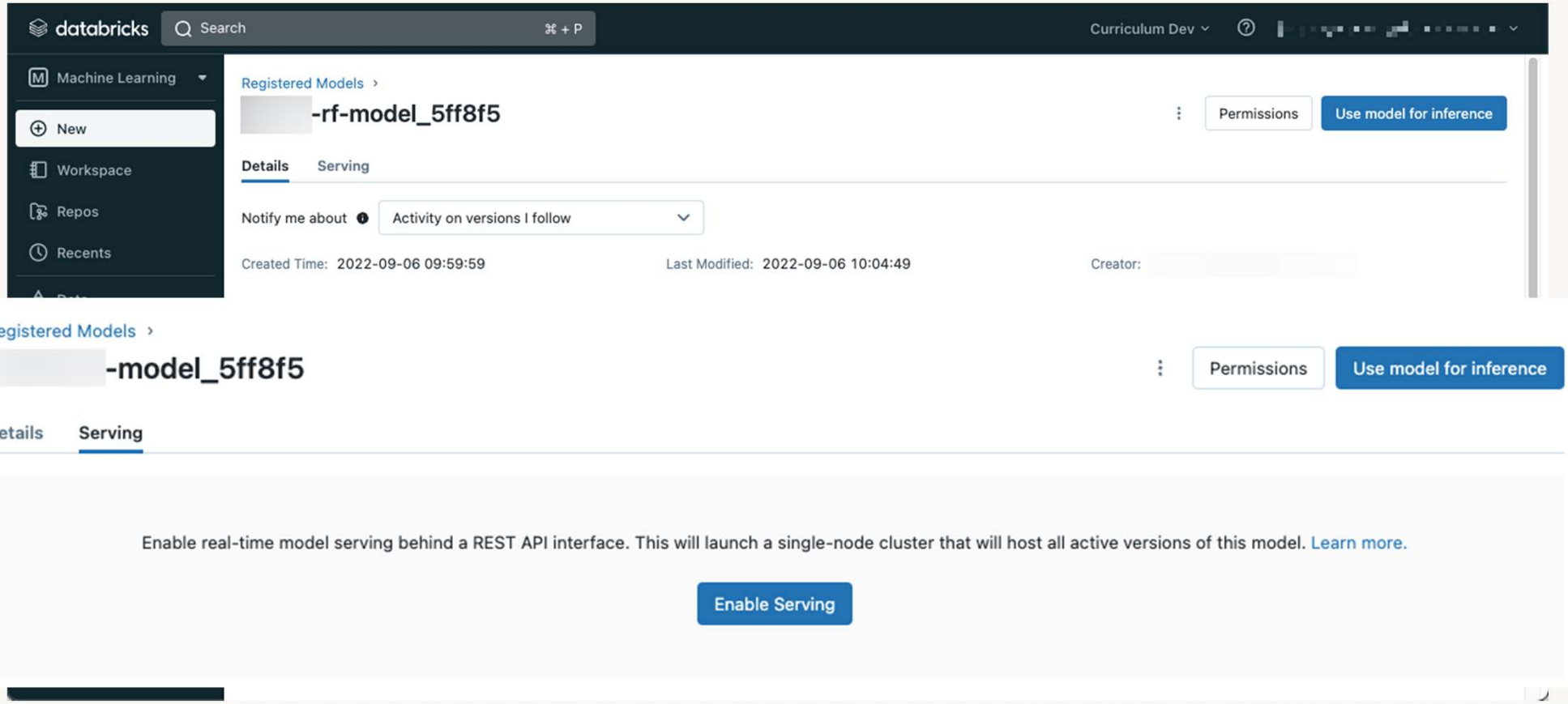
Predict on a Pandas DataFrame:

```
import mlflow
logged_model = 'runs:/8ccea8bc64504eb29bc0786b9d230d6e/model'

# Load model as a PyFuncModel.
loaded_model = mlflow.pyfunc.load_model(logged_model)

# Predict on a Pandas DataFrame.
import pandas as pd
loaded_model.predict(pd.DataFrame(data))
```

MLflow Model Registry



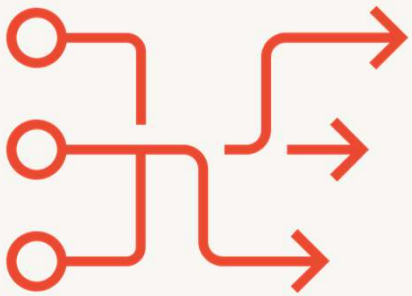
MLflow Model Serving



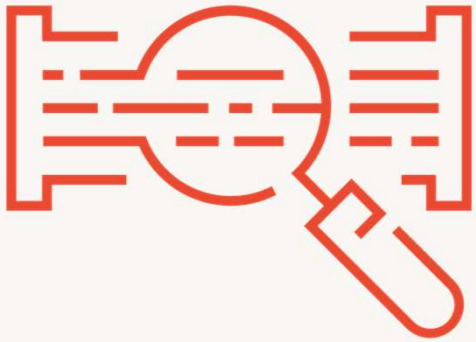
Deploy a model for batch inference

Deploying Models to Production

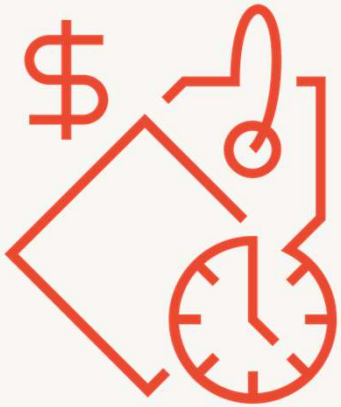
Batch



Streaming



Real-time serving



Questions?

To learn more about Databricks Machine Learning, check out the following resources:

Databricks Academy
Databricks Community