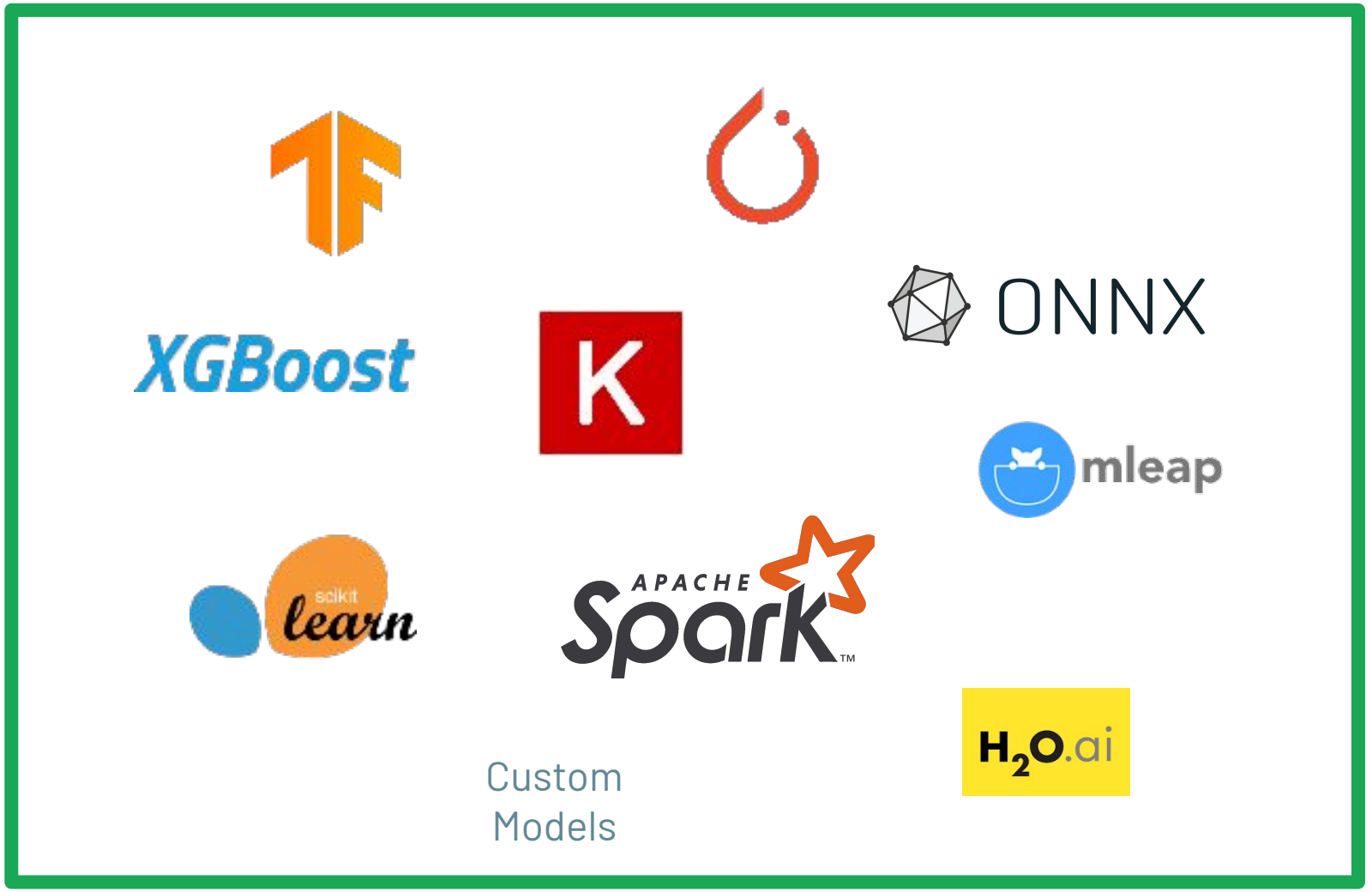
MSFT- ADB Radio Series Part III

John O'Dwyer - Developer Advocate

Nikhil Gupta - Sr. Solution Architect

databricks

Machine Learning and AI



Hardest Part of AI isn't AI, it's Big Data

"Hidden Technical Debt in Machine Learning Systems," Google NIPS 2015

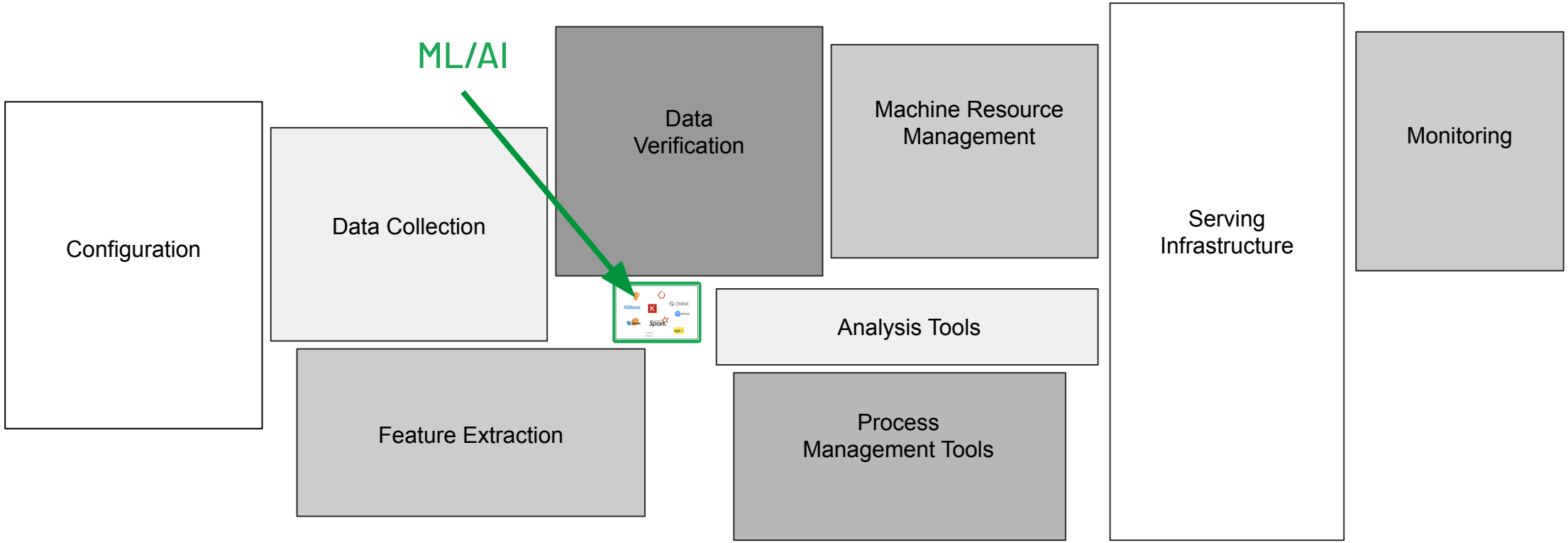
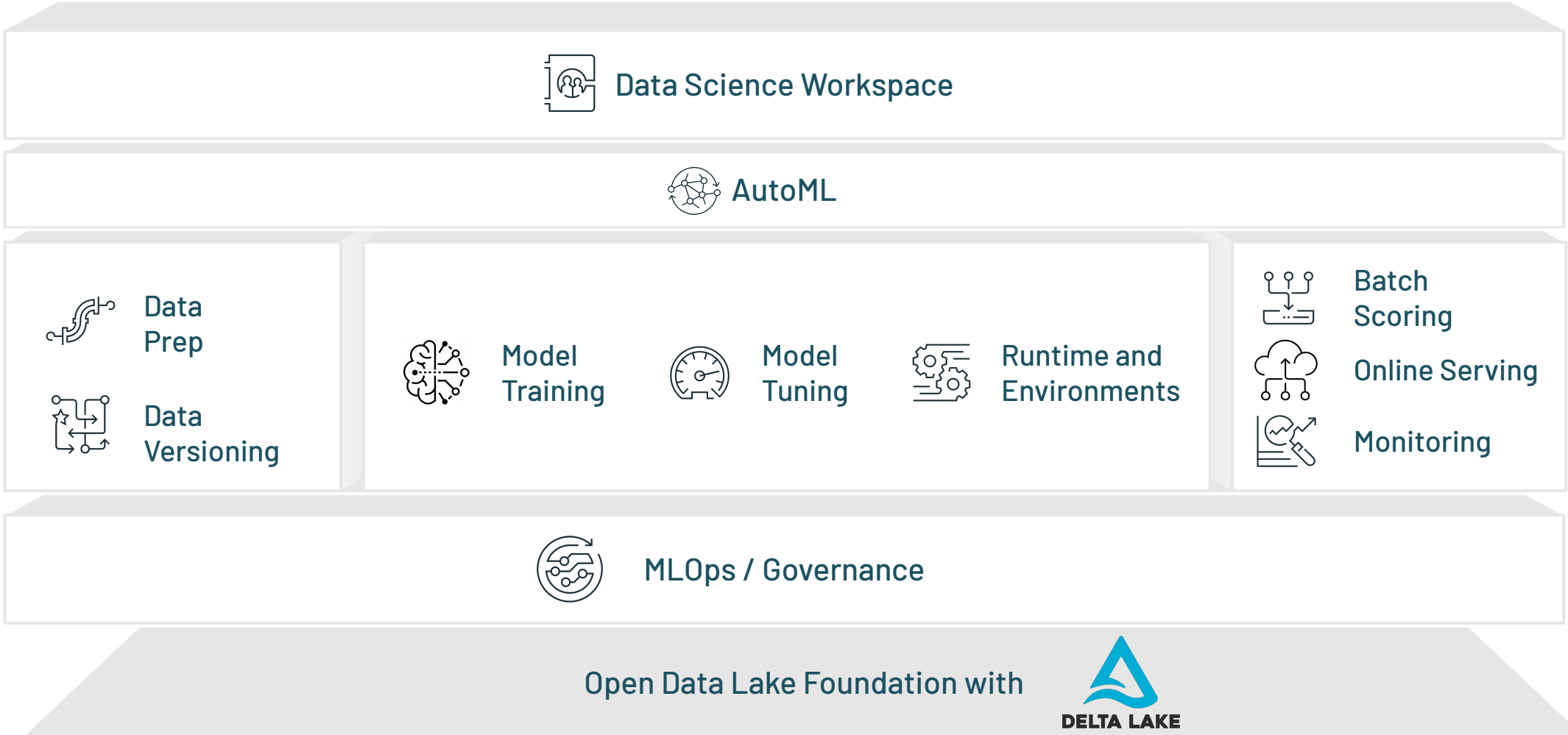


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.

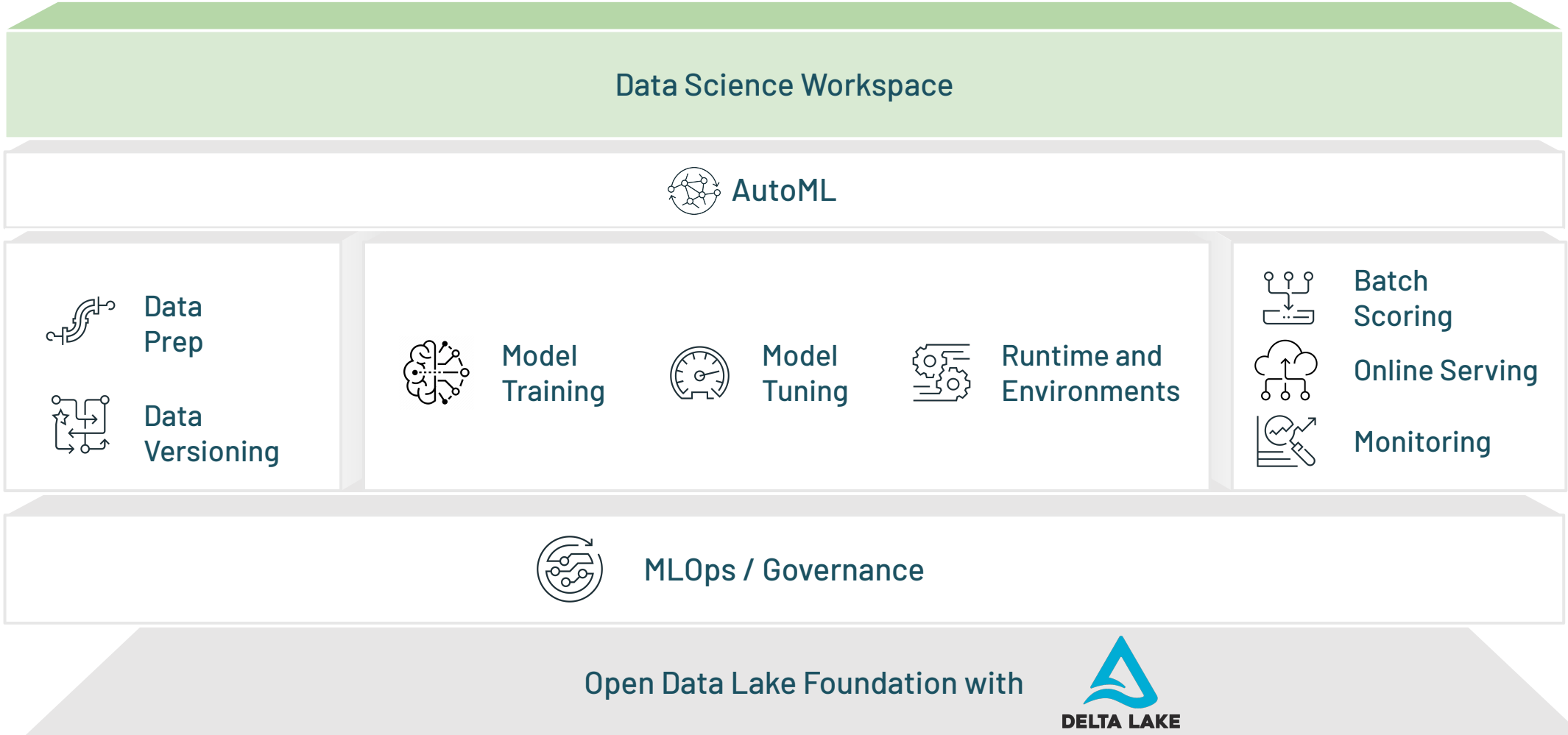
Databricks Machine Learning

A data-native and collaborative solution for the full ML lifecycle



databricks

Data Science Workspace



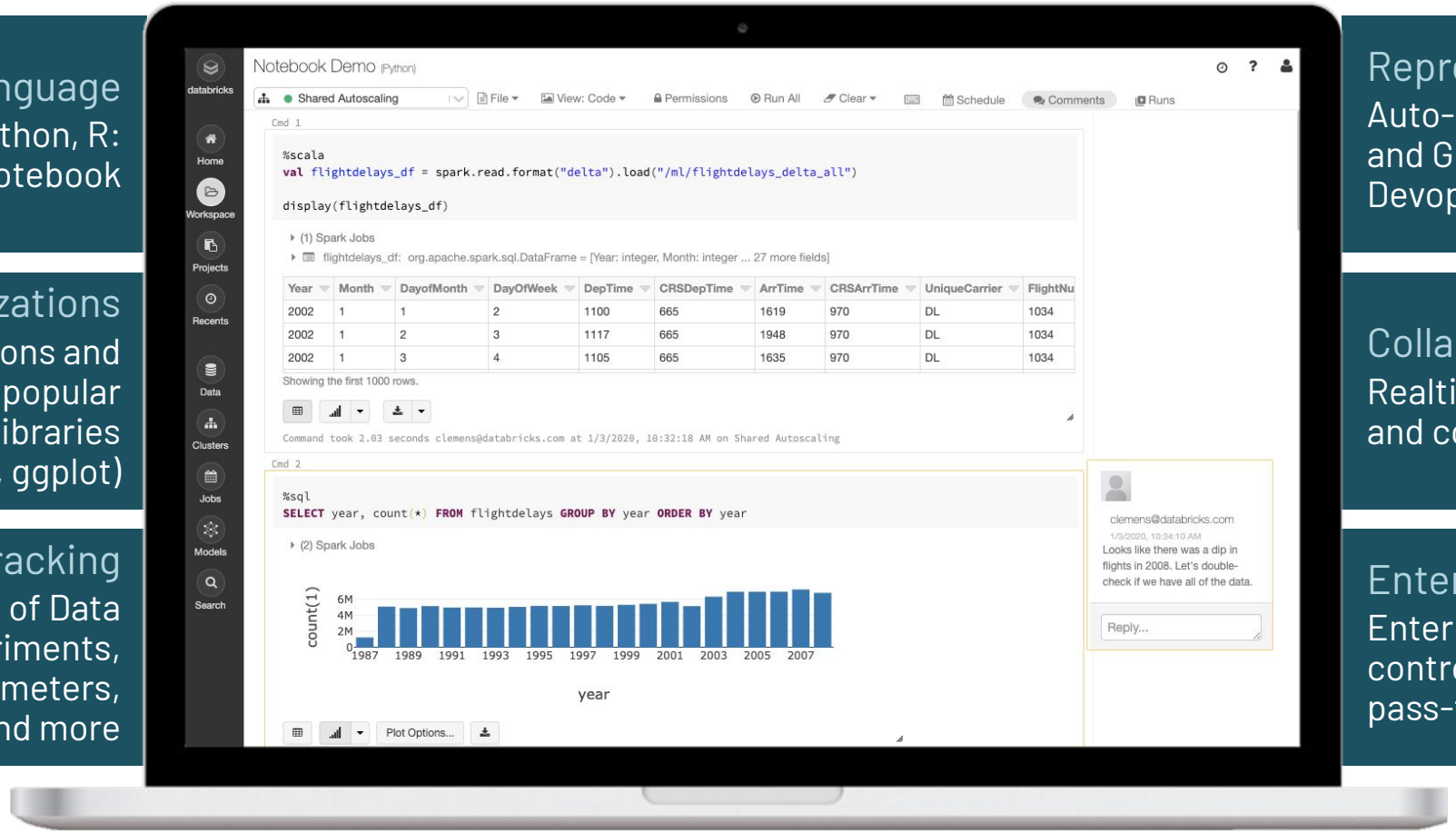
Databricks Notebooks

Provide a collaborative environment for Unified 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 with Azure Devops for version control

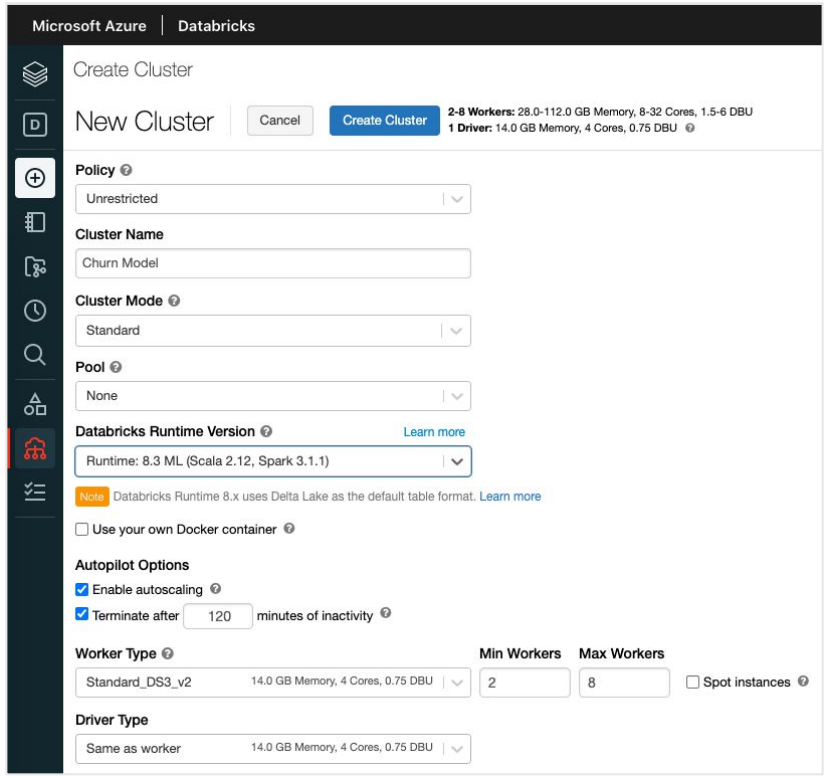
Collaborative
Realtime co-editing and commenting

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

databricks

Hassle-free Compute Infrastructure

Clusters for data processing and model creation

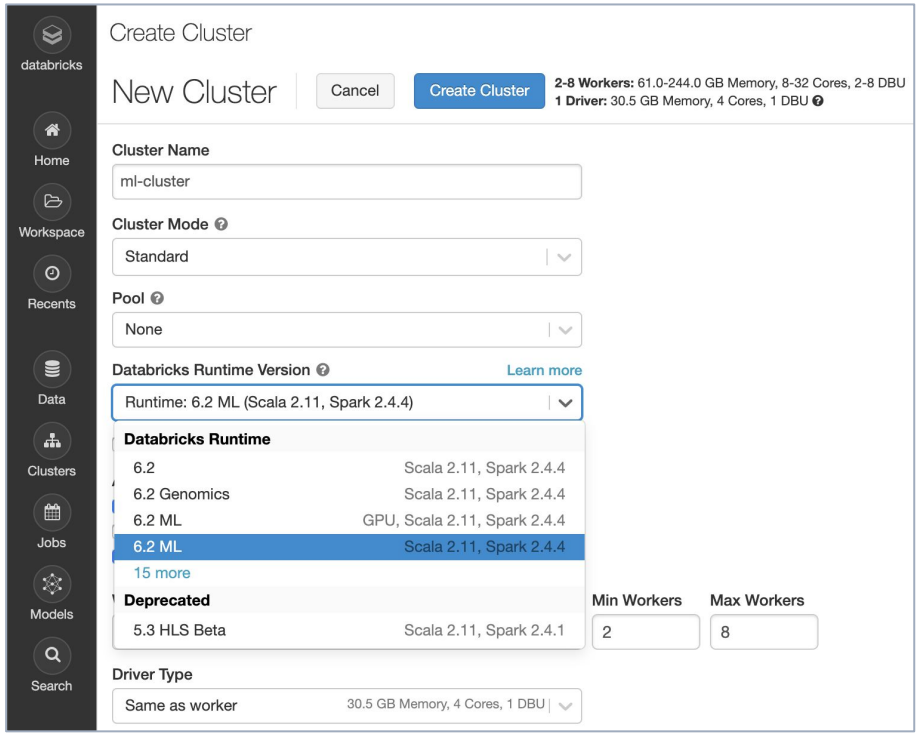


- Compatible with all supported instance/VM types
- User-friendly cluster management
- Optimized Spark performance
- Cost-optimized autoscaling
- GPU support
- Backward compatibility with automatic upgrades
- Data caching

Configure an optimized cluster with just a few clicks.

Databricks ML Runtime

Infrastructure Optimized for ML



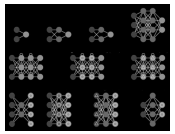
Packaged most common ML Tools




Simplified Distributing ML/DL Libraries

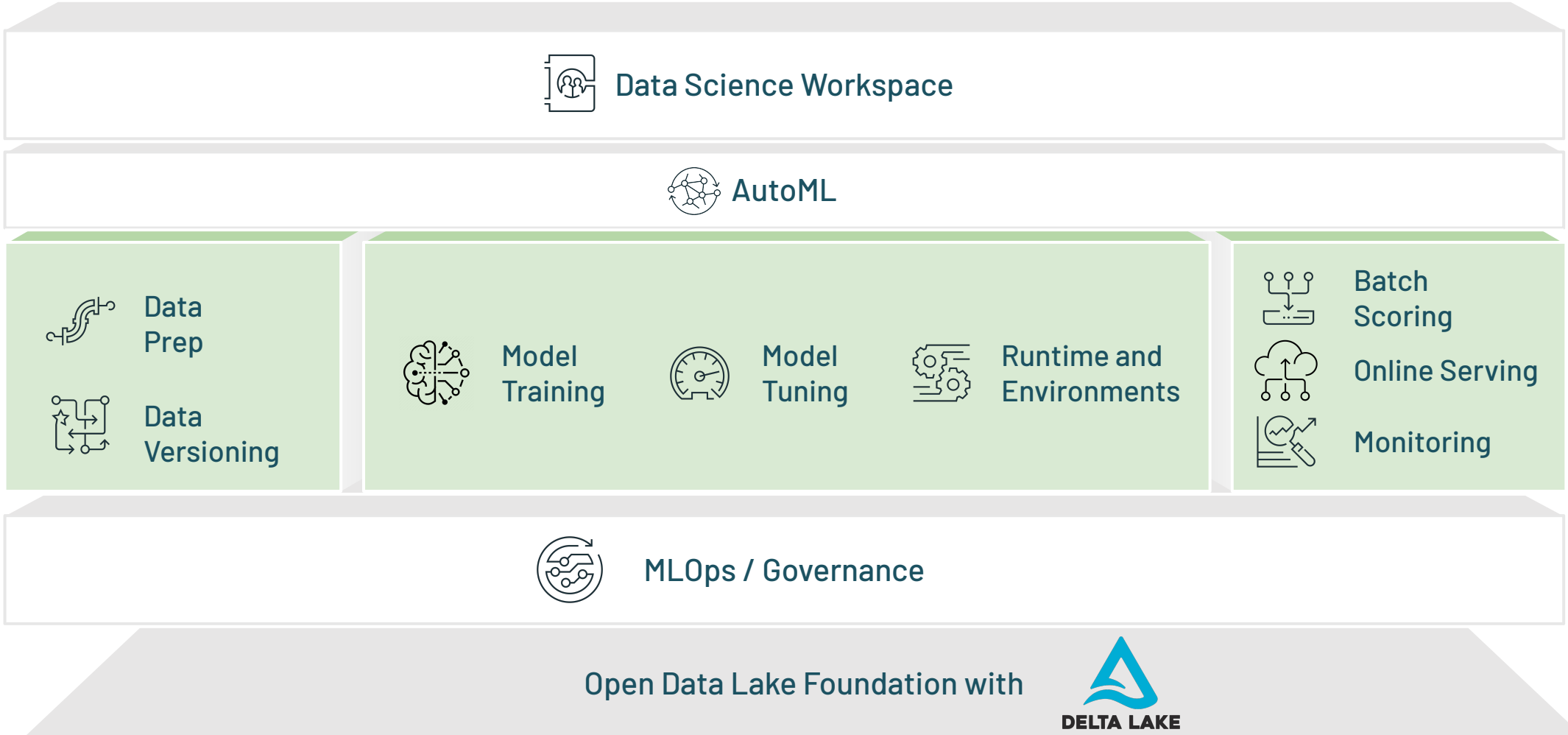
Distribute and Scale any Single-Machine ML Code to 1,000's of machines.

Built-In AutoML and MLflow

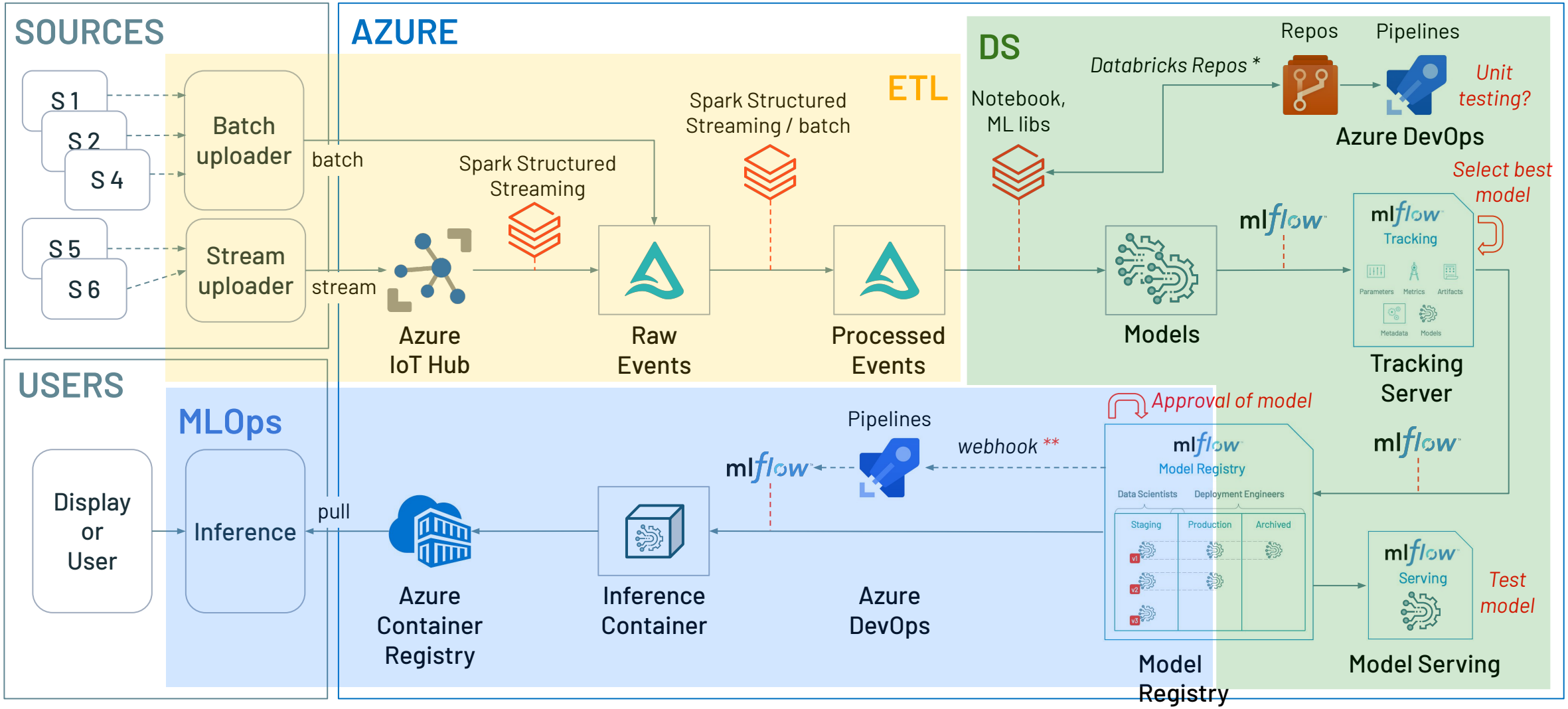


AutoML and Tracking / Visualizations with MLflow

MLflow - From data to model serving end to end



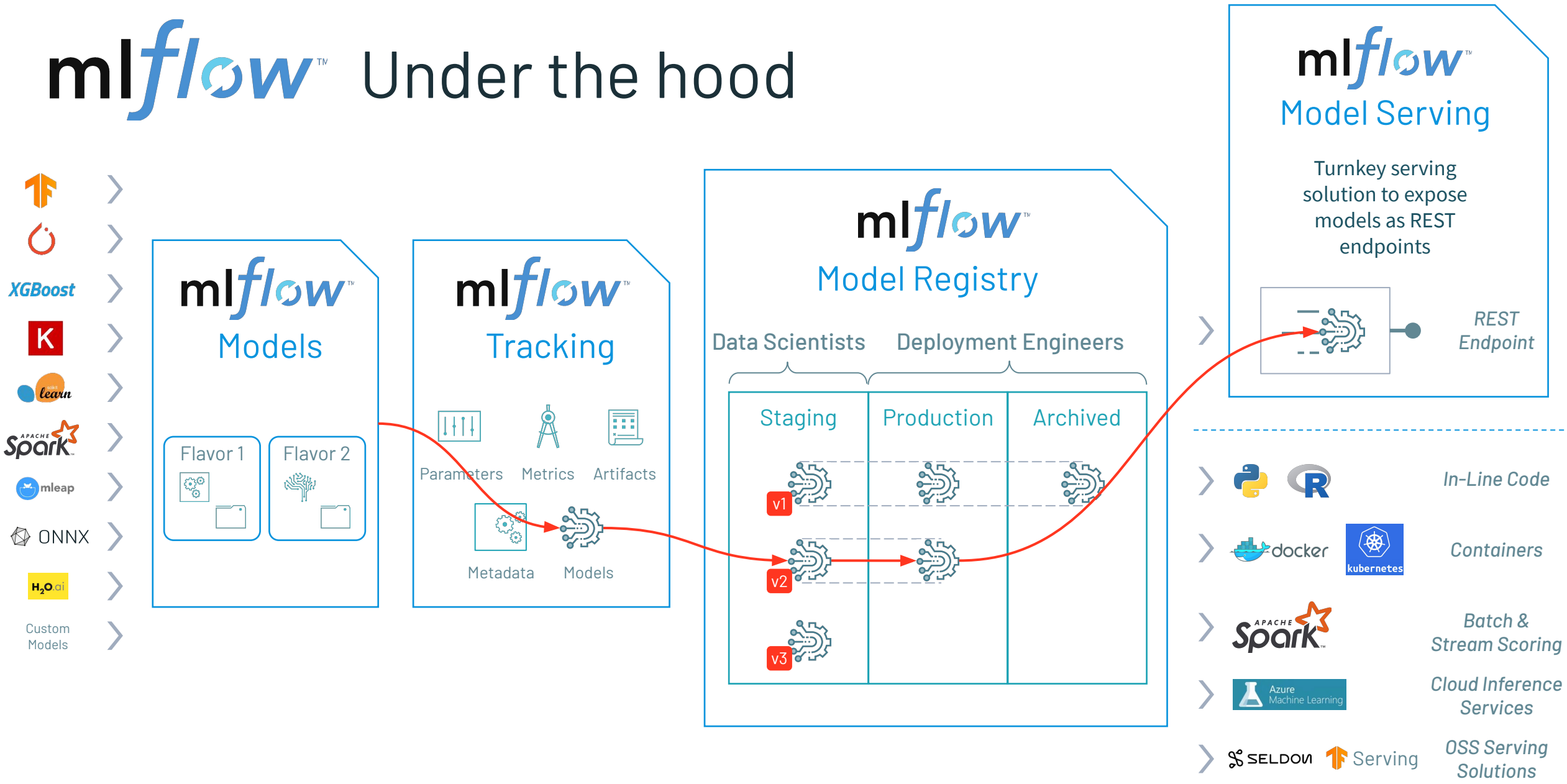
The Big Picture end to end



databricks

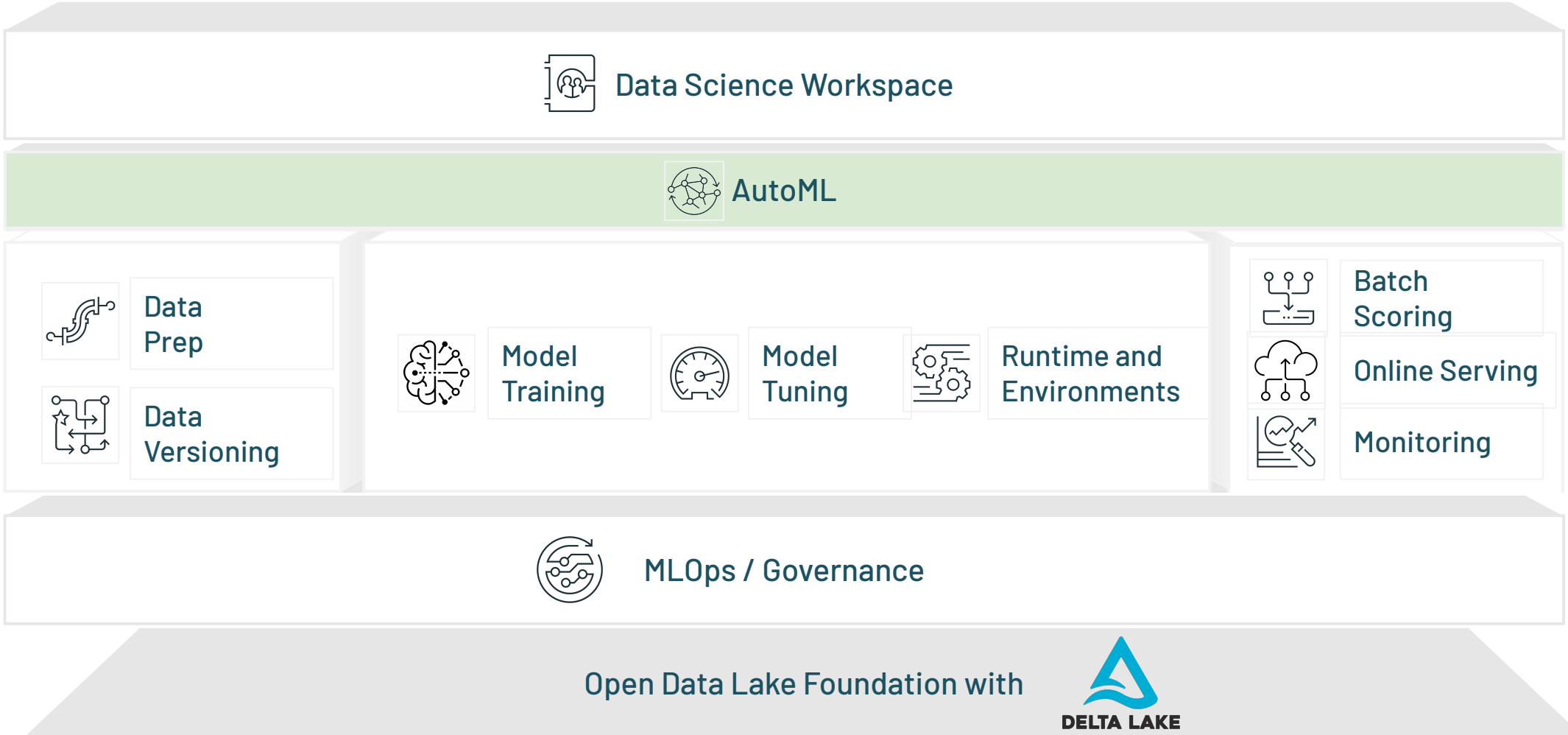
* *In public preview* ** *Currently in private preview*

mlflow Under the hood

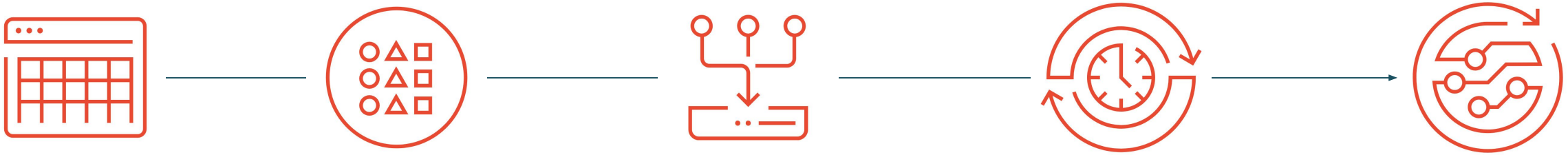


databricks

AutoML - Glass box approach to Machine Learning



Automating the ML workflow



select a dataset

automated data prep

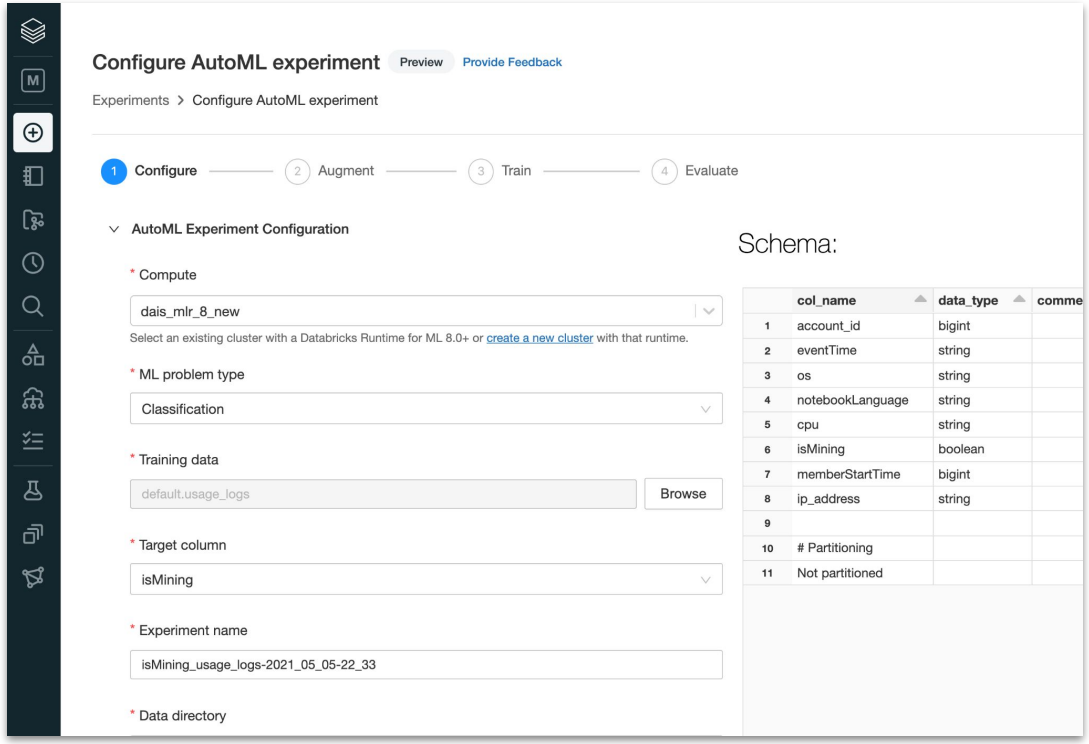
automated feature engineering and selection

automated training and model selection

automated hyperparameter tuning

AutoML in the UI or using the API

UI in Databricks Machine Learning




```
databricks.automl.classify(df,
    target_col='label',
    timeout_minutes=60)
```

AutoML output

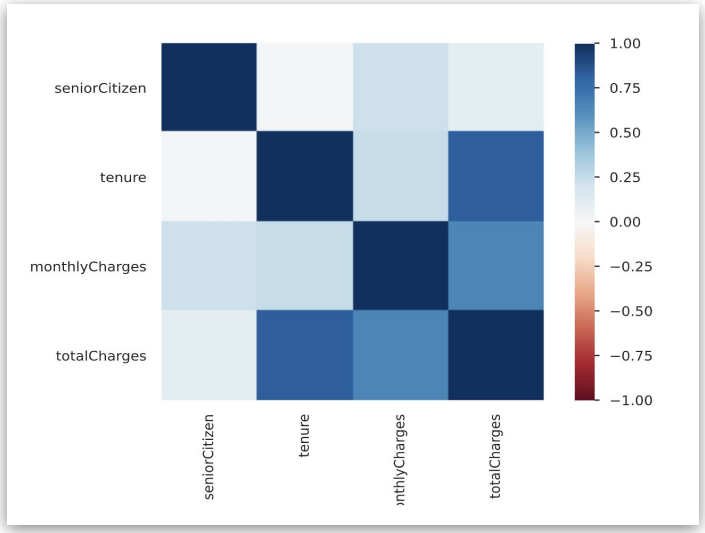
MLflow experiment

Auto-created MLflow Experiment to track models and metrics

Start Time	Run Name	User	Source
2021-05-05 1	logistic_r...	kase...	Noteb
2021-05-05 1	logistic_r...	alkis...	21-05
2021-05-05 1	logistic_r...	alkis...	21-05
2021-05-05 1	logistic_r...	kase...	Noteb
2021-05-05 1	logistic_r...	kase...	Noteb
2021-05-05 1	logistic_r...	kase...	Noteb
2021-05-05 1	decision_...	kase...	Noteb
2021-05-05 1	random_f...	kase...	Noteb

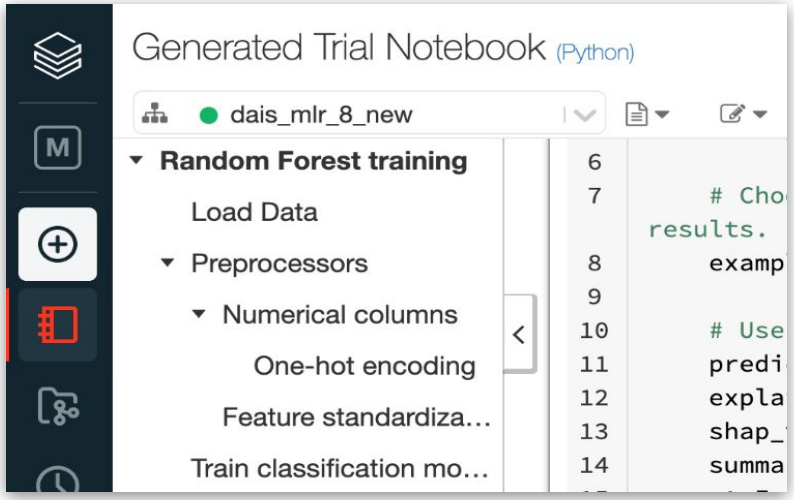
Data exploration notebook

Generated notebook with feature summary statistics and distributions



Reproducible trial notebooks

Generated notebooks with source code for every model



databricks

AutoML Features

Problem Types	Models / Tuning	Features	Tracking	Evaluation	Deployment
Classification	scikit learn	Numerical	mlflow	Shap	Batch Scoring
Regression	XGBoost	Categorical	Metrics, Parameters, Artifacts, Models	mlflow	Model Serving
	HYPEROPT	Timestamp			

Thank you