MLOps Virtual Event: Building Machine Learning Platforms

Matei Zaharia

Chief Technologist, Databricks

@matei_zaharia

databricks

A Common Story

ginablaber
@ginablaber

The story of enterprise Machine Learning: “It took me 3 weeks to develop the model. It’s been >11 months, and it’s still not deployed.” @DineshNirmalIBM #StrataData #strataconf

10:19 AM · Mar 7, 2018 · TweetDeck

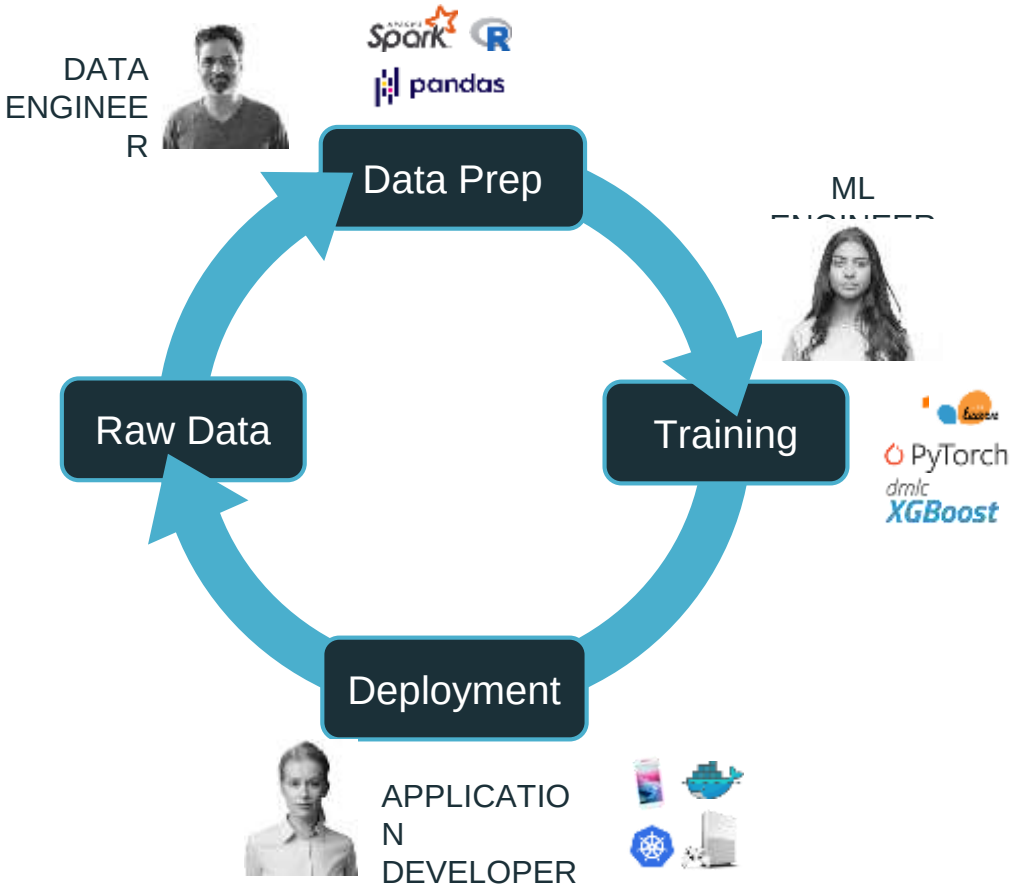
Even After Deploying, Operating ML is Complex!

- Monitoring performance of the model
- Data drift
- Governance and security

Many ML teams spend >50% of their time maintaining existing models

Why is ML Hard to Operationalize?

- Dependence on data
- Multiple, application-specific ways to evaluate performance
- Many teams and systems involved



Response: ML Platforms

- Software platforms to manage ML applications, from development to production
- Most companies that use ML at scale are building one
- Tech examples: Facebook FBLearner, Google TFX, Uber Michelangelo

Common Components in an ML Platform

- Data management, in development and at scoring time
  - Data transformation, quality monitoring, data versioning
  - Feature stores

- Model management
  - Packaging, review, quality assurance, versioning

- Code and deployment management
  - Reproducibility, deployment, monitoring, experimentation

Our Approach at Databricks

Azure

- Every team's requirements will be different, and will change over time

- Provide a **general** platform that is **easy to integrate** with diverse tools

databricks

Data science & ML workspace

mlflow

Open source machine learning platform

python R Java

DELTA LAKE

Transactional, versioned data lake storage

Amazon S3 Azure STORAGE HDFS

In This Webinar

- How we and other organizations handle the different components of a machine learning platform
- Demos and experience from 4 different companies

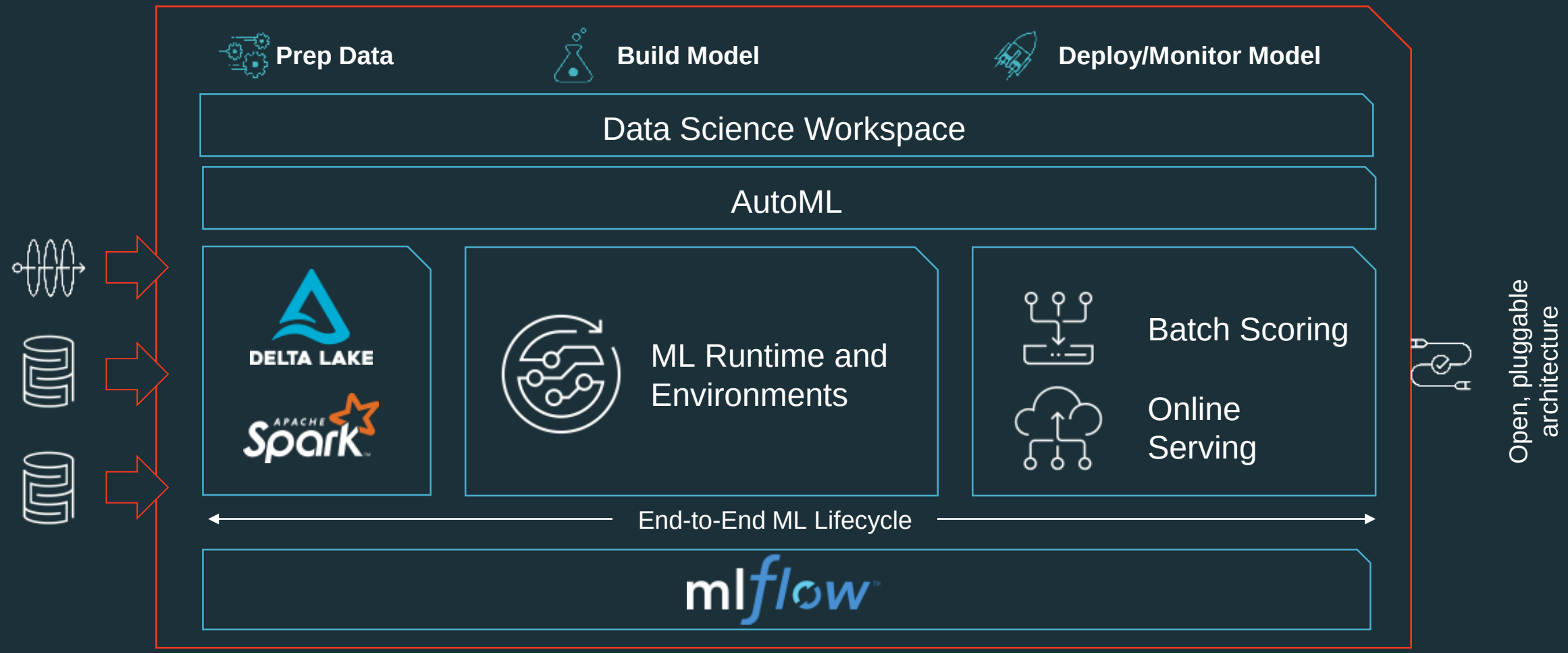
End-to-End Data Science and Machine Learning on Databricks

Clemens Mewald

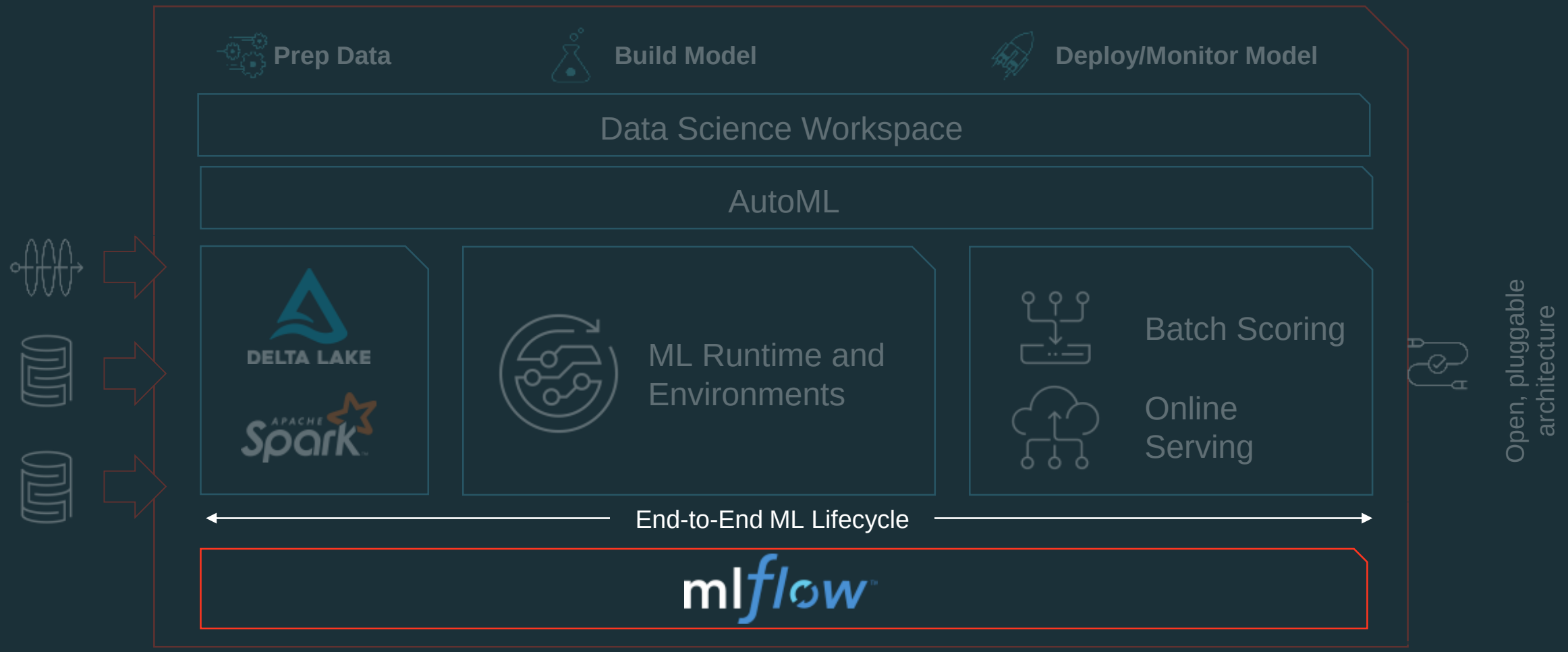
Director of Product Management, Databricks

databricks

End-to-End Data Science and ML on



End-to-End Data Science and ML on databricks



mlflow Components

mlflow

Projects

Packaging format for reproducible runs on any compute platform

mlflow Components

mlflow Projects

Packaging format for reproducible runs on any compute platform

mlflow Models

General model format that standardizes deployment options

mlflow Components

mlflow Projects

Packaging format for reproducible runs on any compute platform

mlflow Models

General model format that standardizes deployment options

mlflow Tracking

Record and query experiments: code, metrics, parameters, artifacts, models

mlflow Components

mlflow Projects

Packaging format for reproducible runs on any compute platform

mlflow Models

General model format that standardizes deployment options

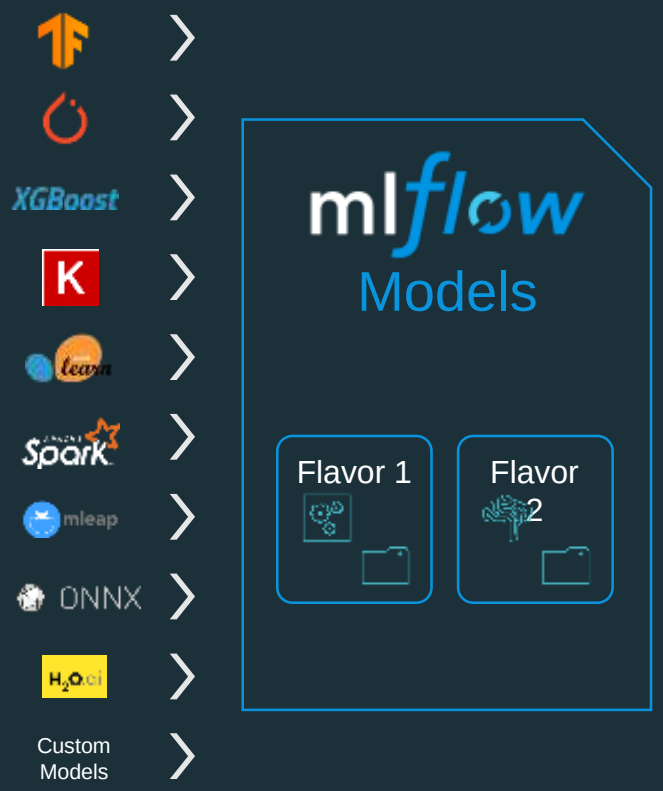
mlflow Tracking

Record and query experiments: code, metrics, parameters, artifacts, models

mlflow Model Registry

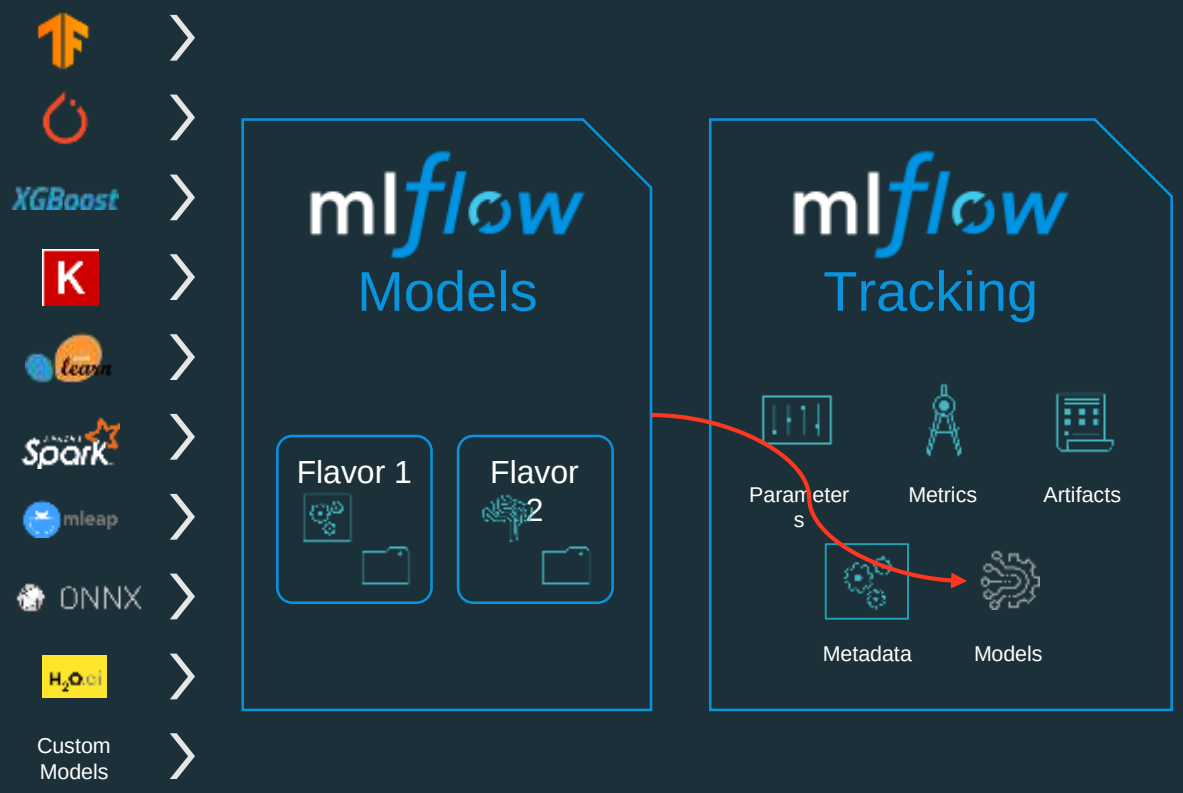
Centralized and collaborative model lifecycle management

mlflow Model Lifecycle

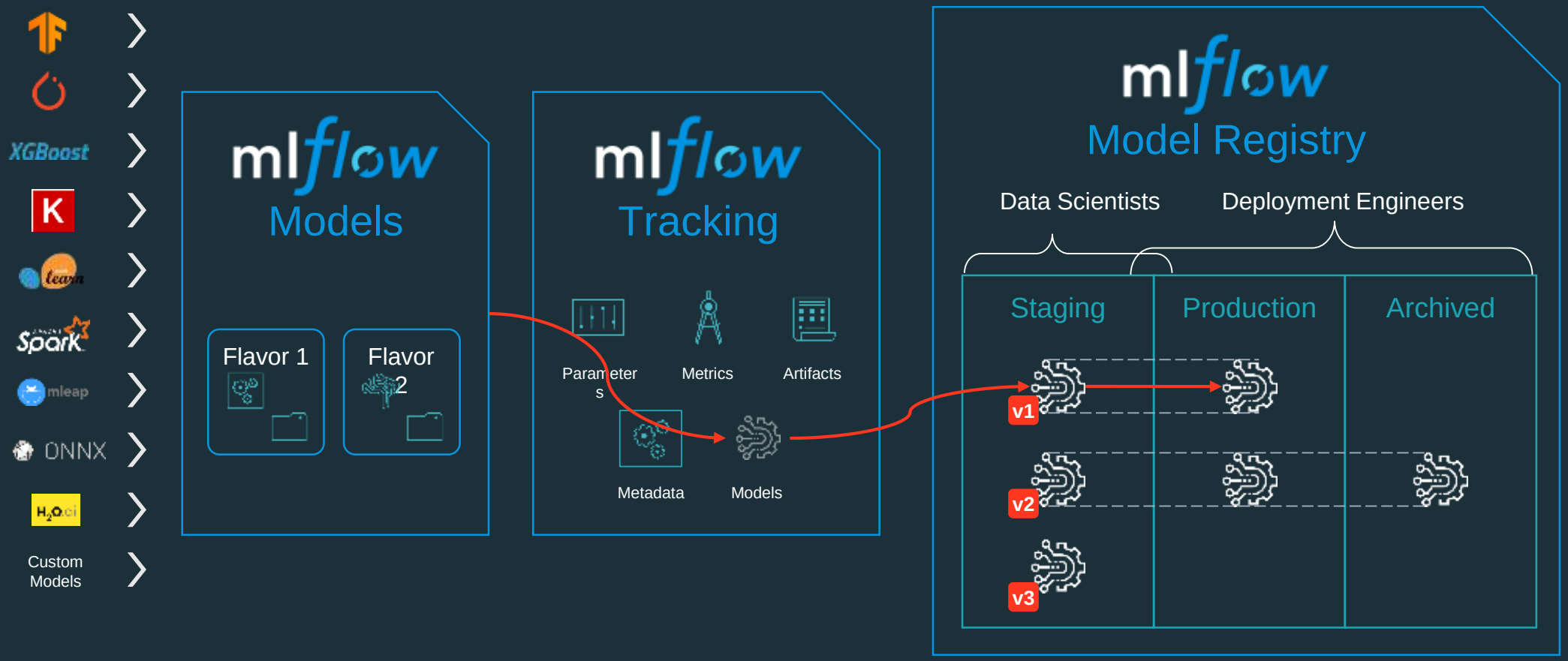


databricks

mlflow Model Lifecycle

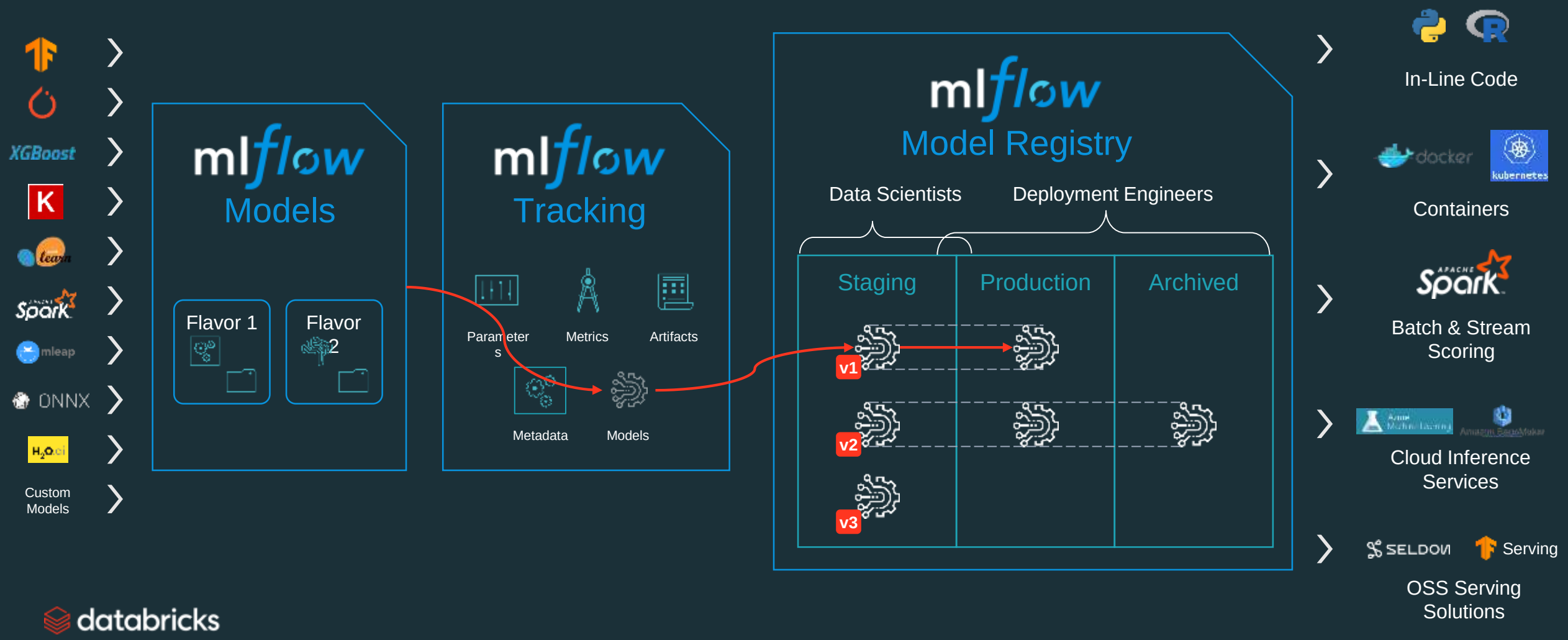


mlflow Model Lifecycle



databricks

mlflow Model Lifecycle



databricks

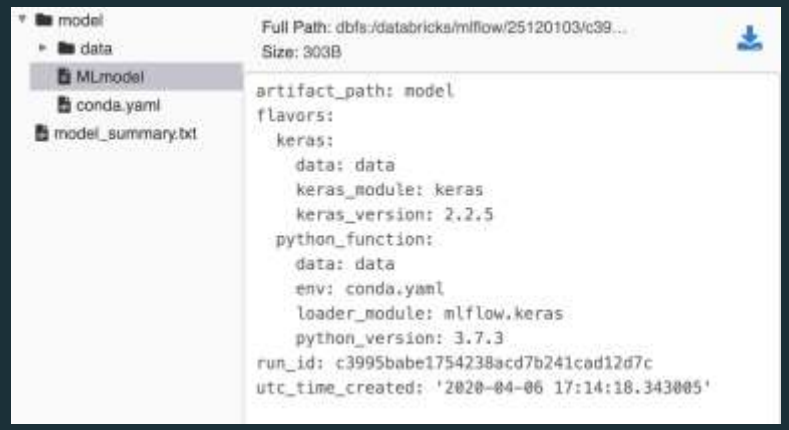
mlflow Auto-Logging

Auto-logging for ML Frameworks: A single line of code logs parameters, metrics, and artifacts.

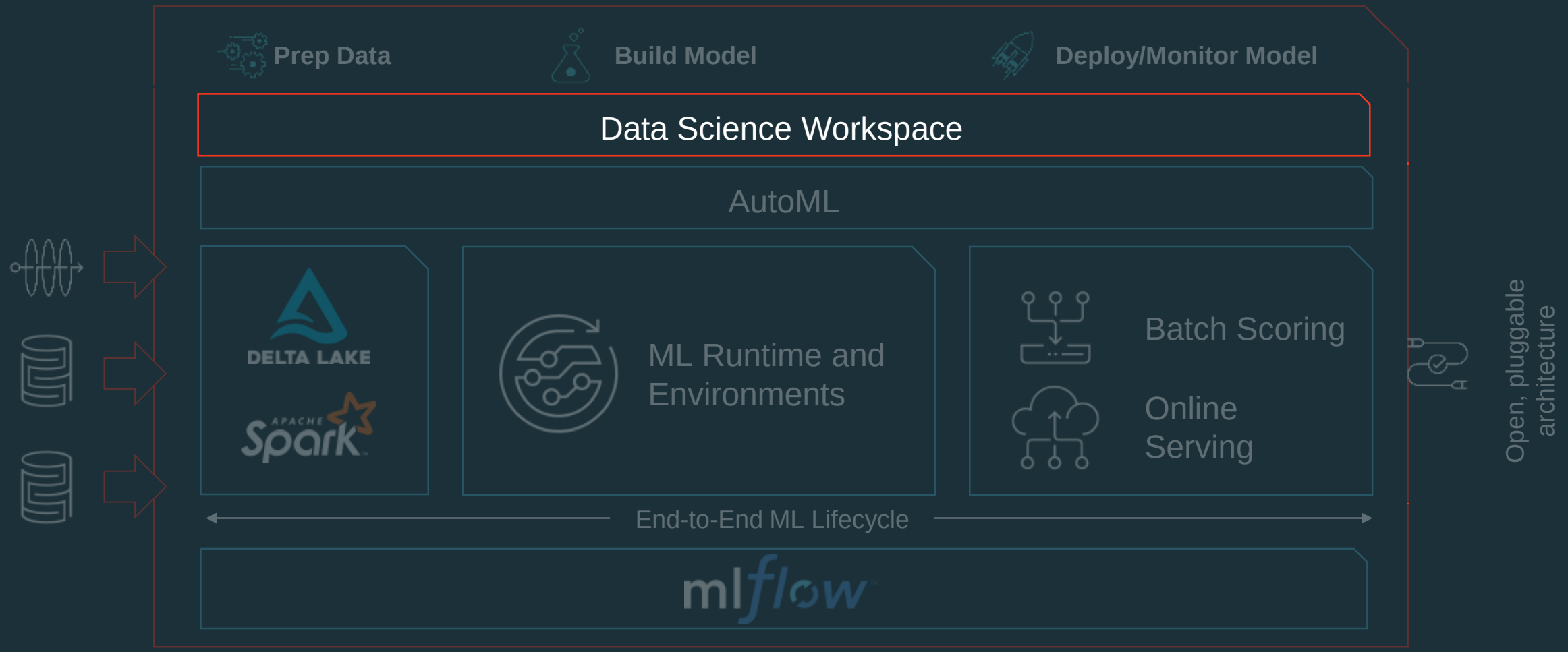
```
mlflow.keras.autolog() # or: mlflow.tensorflow.autolog()
```

Parameters and (a time series of) metrics

Artifacts (including model)



End-to-End Data Science and ML on databricks



Databricks Notebooks

Provide a collaborative environment for Unified Data Analytics

Multi-Language

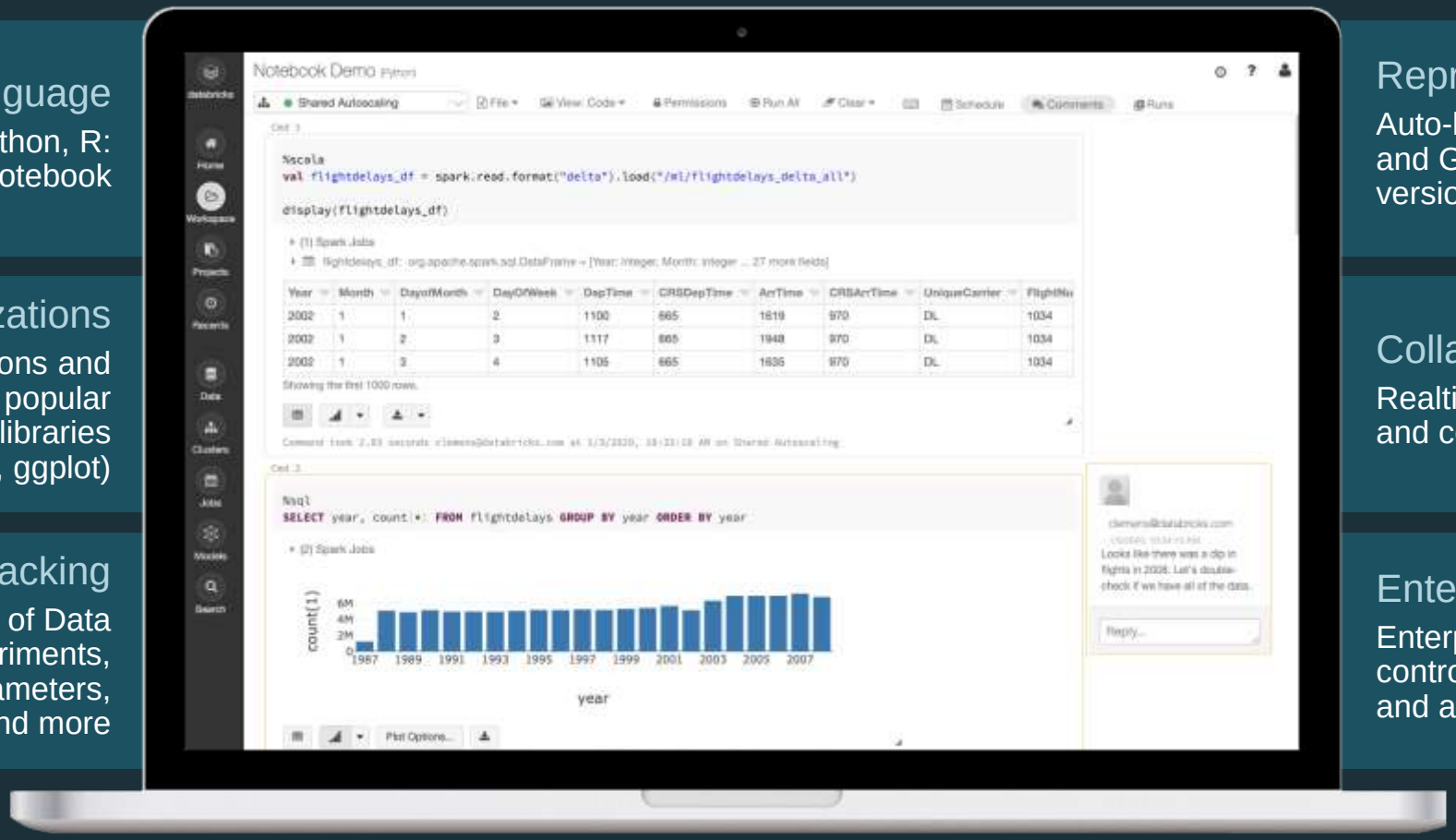
Scala, SQL, Python, R: All in one notebook

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 for version control

Collaborative

Realtime co-editing and commenting

Enterprise Ready

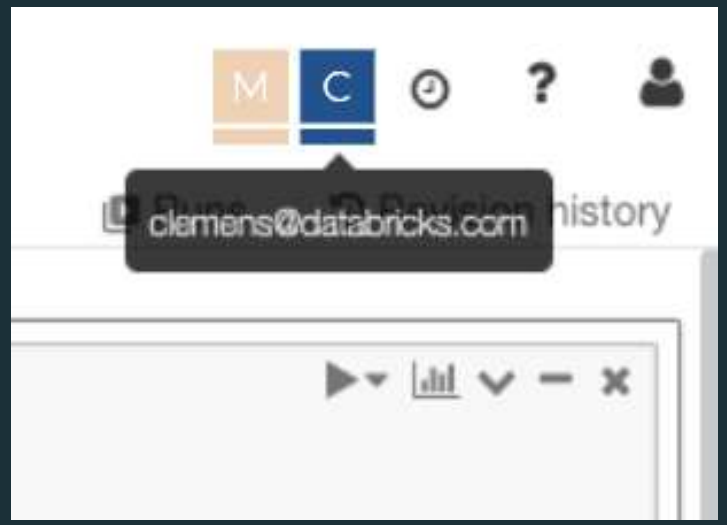
Enterprise grade access controls, identity pass-through, and auditability

databricks

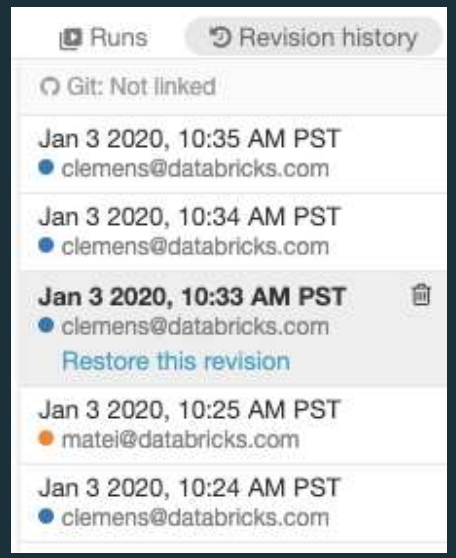
Databricks Notebooks for Collaborative Data Science

Data Engineers, Data Scientists, ML Engineers, and Data Analysts can all collaborate in one shared environment using modern collaboration patterns.

Co-Presence / Co-Editing

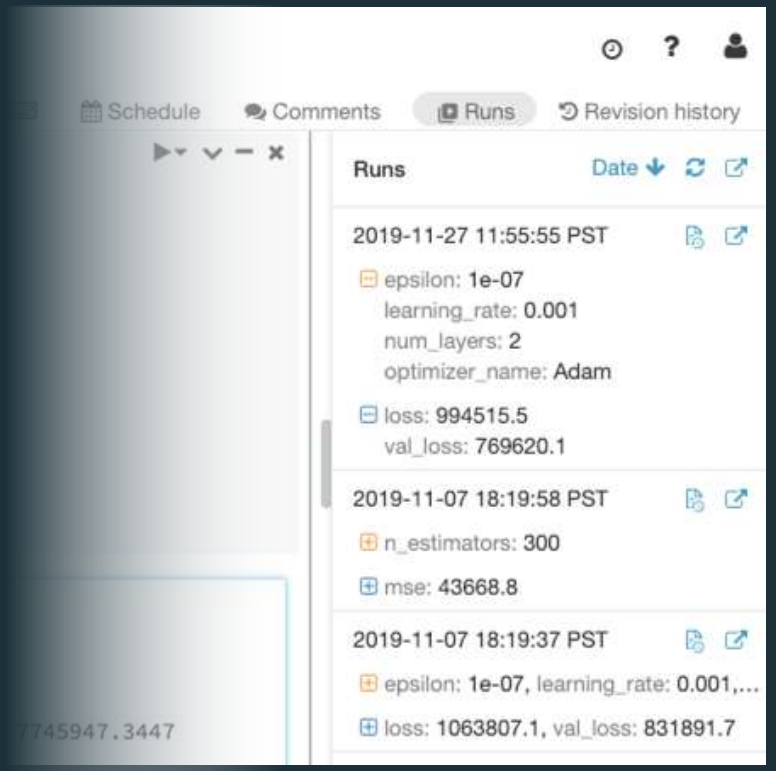


Versioning



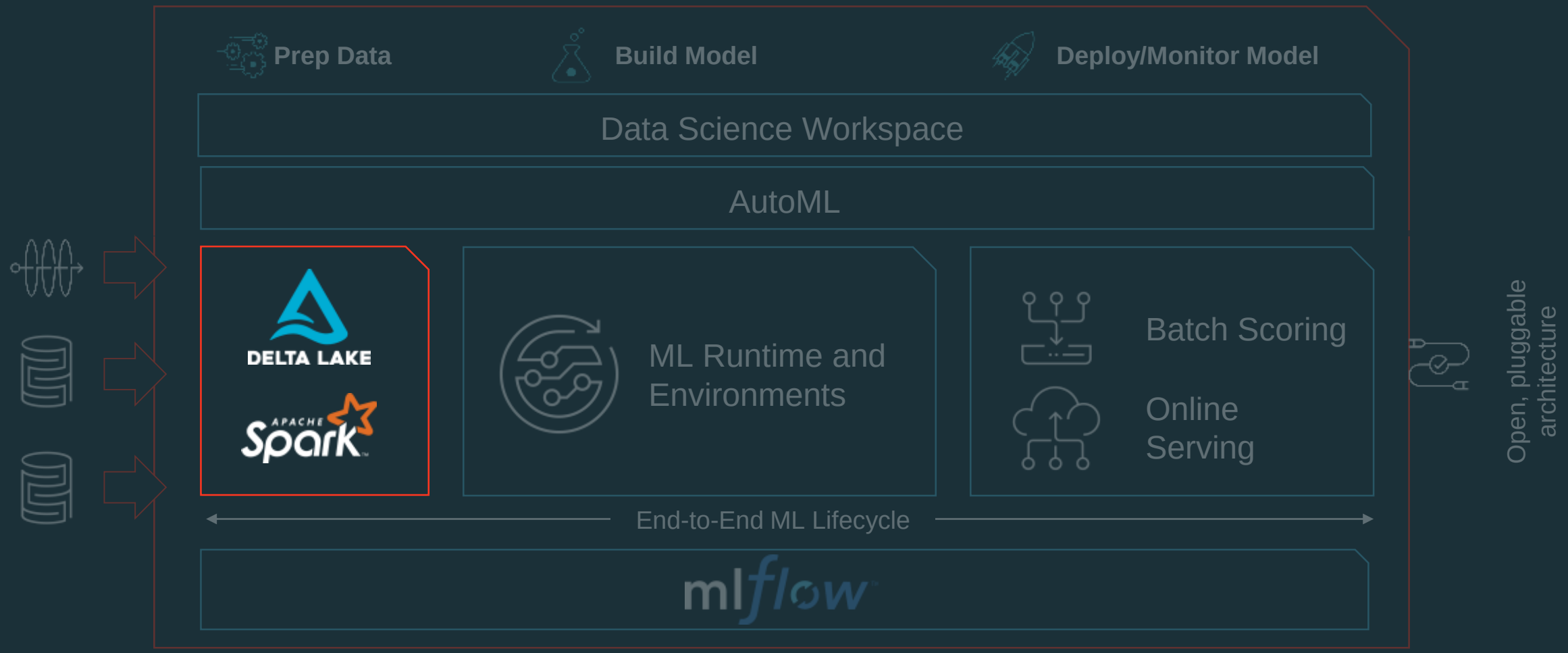
Commenting

mlflow Integration with Databricks Notebooks



- Runs Sidebar integrated with MLflow Tracking
- Track runs, sort by metrics and parameters
- Linked to revision history of the notebook

End-to-End Data Science and ML on databricks



for Data Science and ML

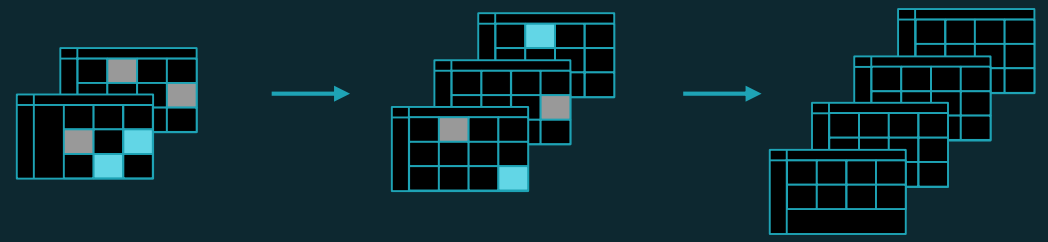
Files

Streaming

Batch

3rd Party Data Marketplace

DELTA LAKE™



Ingestion Tables → Data Catalog → Feature Store

Your Existing Data Lake

Azure Data Lake Storage

hadoop

Amazon S3

ML Runtime

- Schema enforced high quality data
- Optimized performance

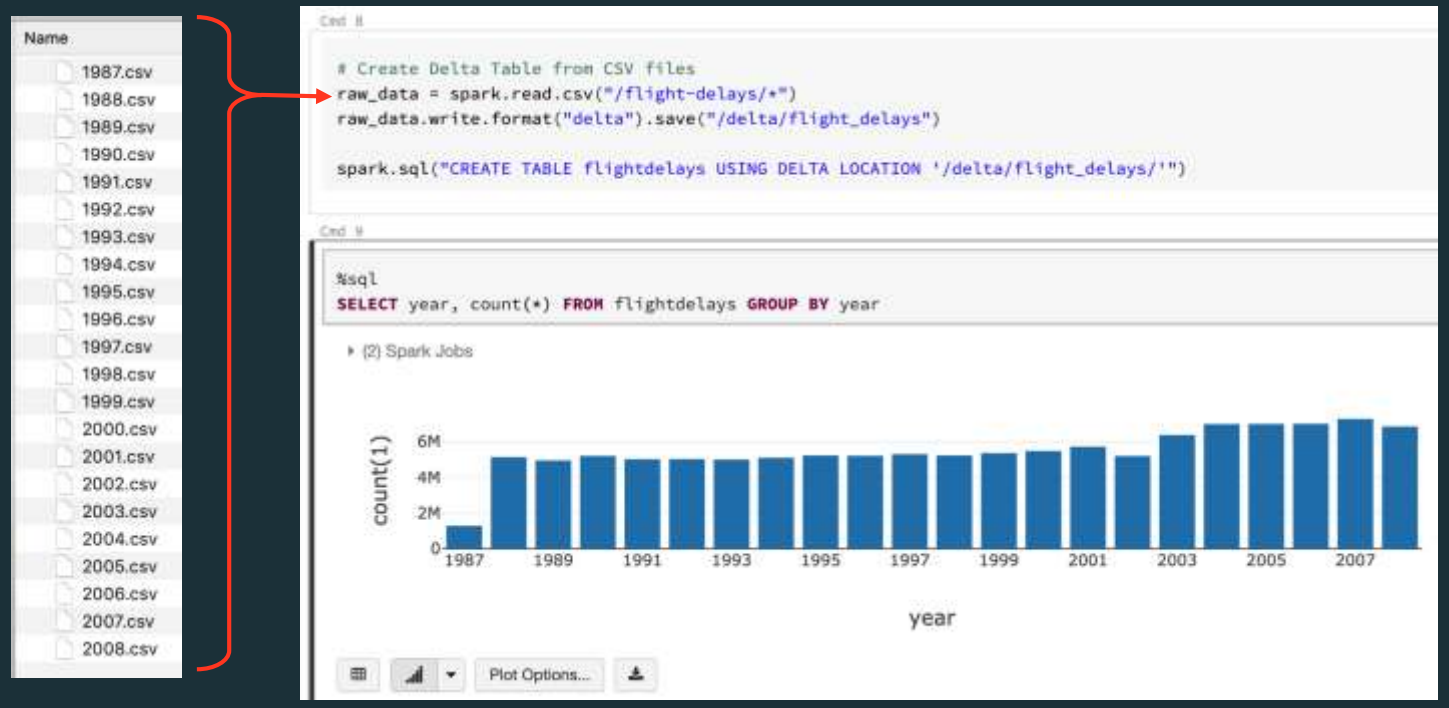
mlflow™

- Full data lineage / governance
- Reproducibility through time travel

databricks

for Data Science and ML

Ingest data and visualize data distribution



for Data Science and ML

Data versioning and time travel

Cmd 10

```
%sql
DESCRIBE HISTORY flightdelays
```

(3) Spark Jobs

version	timestamp	userId	userName	operation	operationParameters	job	notebook	clusterId	readVersion	isolati
7	2019-10-08T16:47:22.000+0000	101543	clemens.mewald@databricks.com	MERGE	{"predicate":"((k.`Origin` = f.`Dest`) AND (k.`Dest` = f.`Origin`))"}	null	{"notebookId":"21190735"}	0304-201045-hoary804	6	WriteS
6	2019-10-08T16:44:16.000+0000	101543	clemens.mewald@databricks.com	MERGE	{"predicate":"((k.`Origin` = f.`Origin`) AND (k.`Dest` = f.`Dest`))"}	null	{"notebookId":"21190735"}	0304-201045-hoary804	5	WriteS

Command took 6.04 seconds -- by clemens.mewald@databricks.com at 11/10/2019, 2:19:12 PM on Shared Autoscaling

for Data Science and ML

Data versioning and time travel

Cmd 10

```
%sql
DESCRIBE HISTORY flightdelays
```

(3) Spark Jobs

version	timestamp	userId	userName	operation	operationParameters	job	notebook	clusterId	readVersion	isolati
7	2019-10-08T16:47:22.000+0000	101543	clemens.mewald@databricks.com	MERGE	{"predicate":"((k.`Origin` = f.`Dest`) AND (k.`Dest` = f.`Origin`))"}	null	{"notebookId":"21190735"}	0304-201045-hoary804	6	WriteS
6	2019-10-08T16:44:16.000+0000	101543	clemens.mewald@databricks.com	MERGE	{"predicate":"((k.`Origin` = f.`Origin`) AND (k.`Dest` = f.`Dest`))"}	null	{"notebookId":"21190735"}	0304-201045-hoary804	5	WriteS

Command took 6.84 seconds -- by clemens.mewald@databricks.com at 11/10/2019, 2:19:12 PM on Shared Autoscaling

Cmd 11

```
%sql
SELECT * FROM flightdelays VERSION AS OF 6
```

(4) Spark Jobs

Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime
2006	11	1	3	1900	1130	2140	1262	EV	4608	N902EV	160	132
2006	11	2	4	1903	1130	2056	1262	EV	4608	N920EV	113	132

mlflow Integration with Delta

Auto-Logging for any Spark Datasource

Cmd 1

```
with mlflow.start_run():
  mlflow.spark.autolog()

  df = spark.read.table('flightdelays')
  display(df)
```

(2) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [Year: integer, Month: integer ... 27 more fields]

Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTi
2004	5	1	6	1746	675	855	300
2004	5	2	7	1745	675	835	300
2004	5	3	1	1748	675	837	300
2004	5	4	2	1752	675	837	300
2004	5	5	3	1747	675		
2004	5	6	4	1749	675		
2004	5	7	5	1747	675		
2004	5	8	6	1749	675		

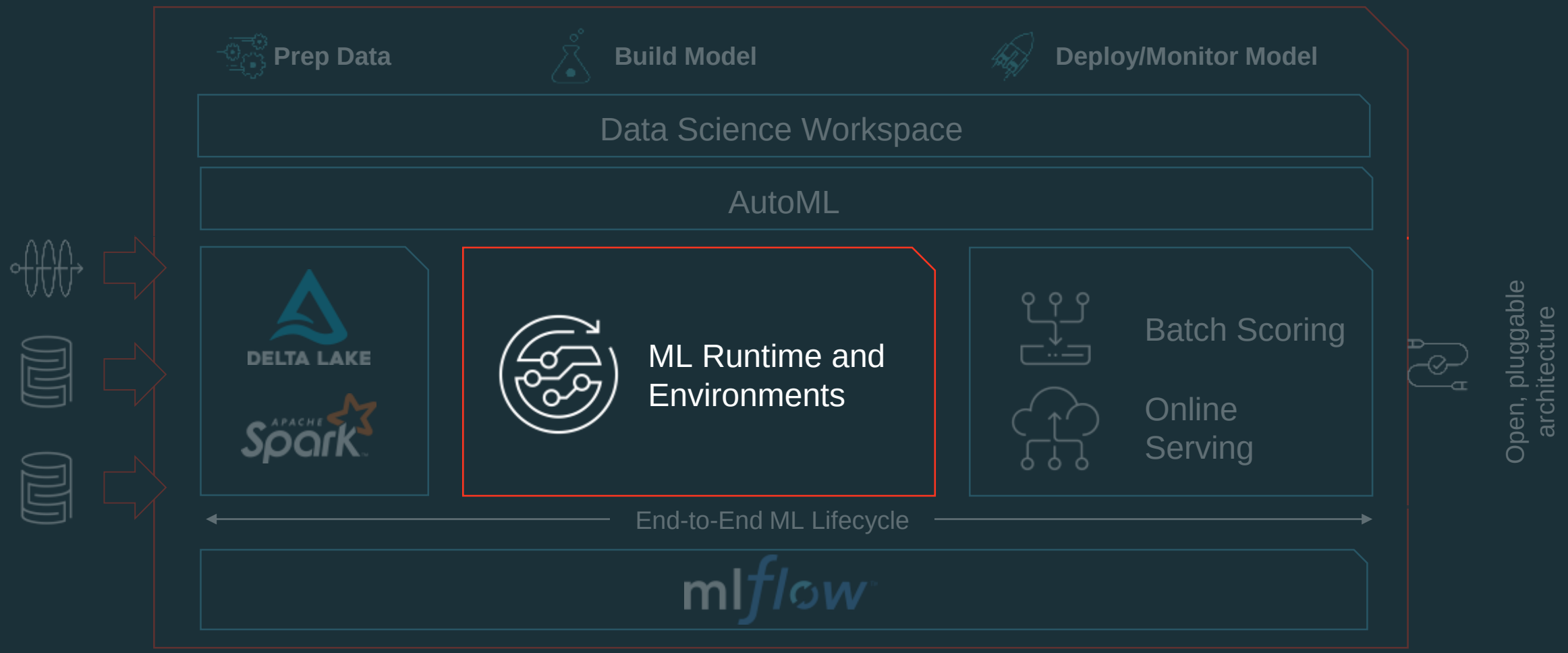
Showing the first 8809 rows.

Command took 2.80 seconds -- by clemens.mewald@databricks.com at 4/2

Tags

Name	Value	Actions
sparkDatasourceInfo	path=dbfs:/ml/flightdelays_delta_all/_delta_log/00000000000000000000.json,format=json	

End-to-End Data Science and ML on databricks



Machine Learning Runtime

Packages and optimizes most common ML Frameworks

XGBoost

...

Machine Learning Runtime

Packages and optimizes most common ML Frameworks

XGBoost

...

Built-in Optimization for Distributed Deep Learning

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

Machine Learning Runtime

Packages and optimizes most common ML Frameworks

XGBoost

...

Built-in Optimization for Distributed Deep Learning

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

Built-In AutoML and Experiment Tracking

mlflow

AutoML and Tracking / Visualizations with MLflow

Machine Learning Runtime

Packages and optimizes most common ML Frameworks

XGBoost

...

Built-in Optimization for Distributed Deep Learning

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

Built-In AutoML and Experiment Tracking

mlflow

AutoML and Tracking / Visualizations with MLflow

Customized Environments using Conda

requirements.txt
conda.yaml

Customization

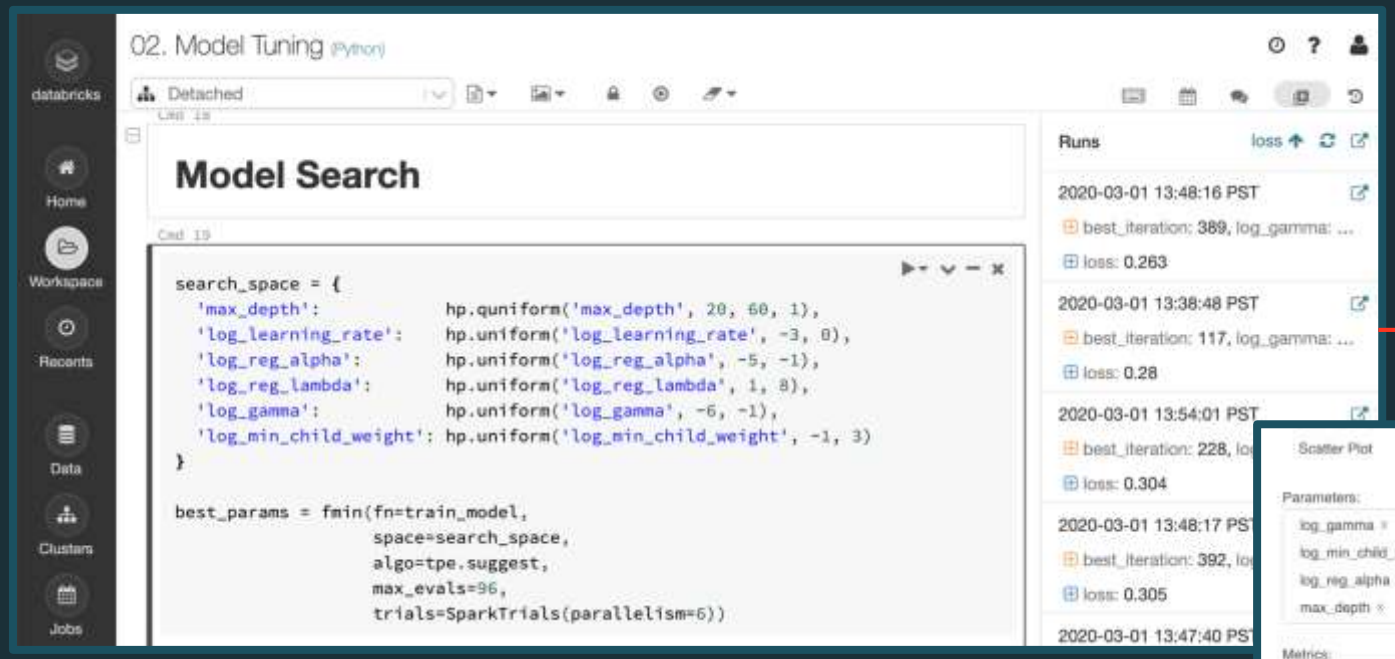
Pre-configured Environment

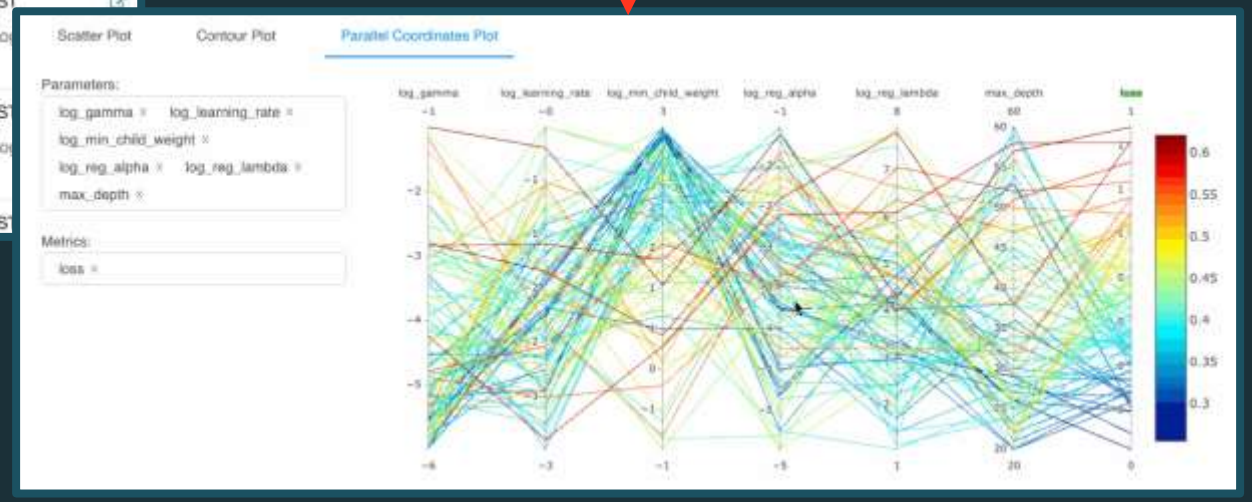
Machine Learning

Conda-Based

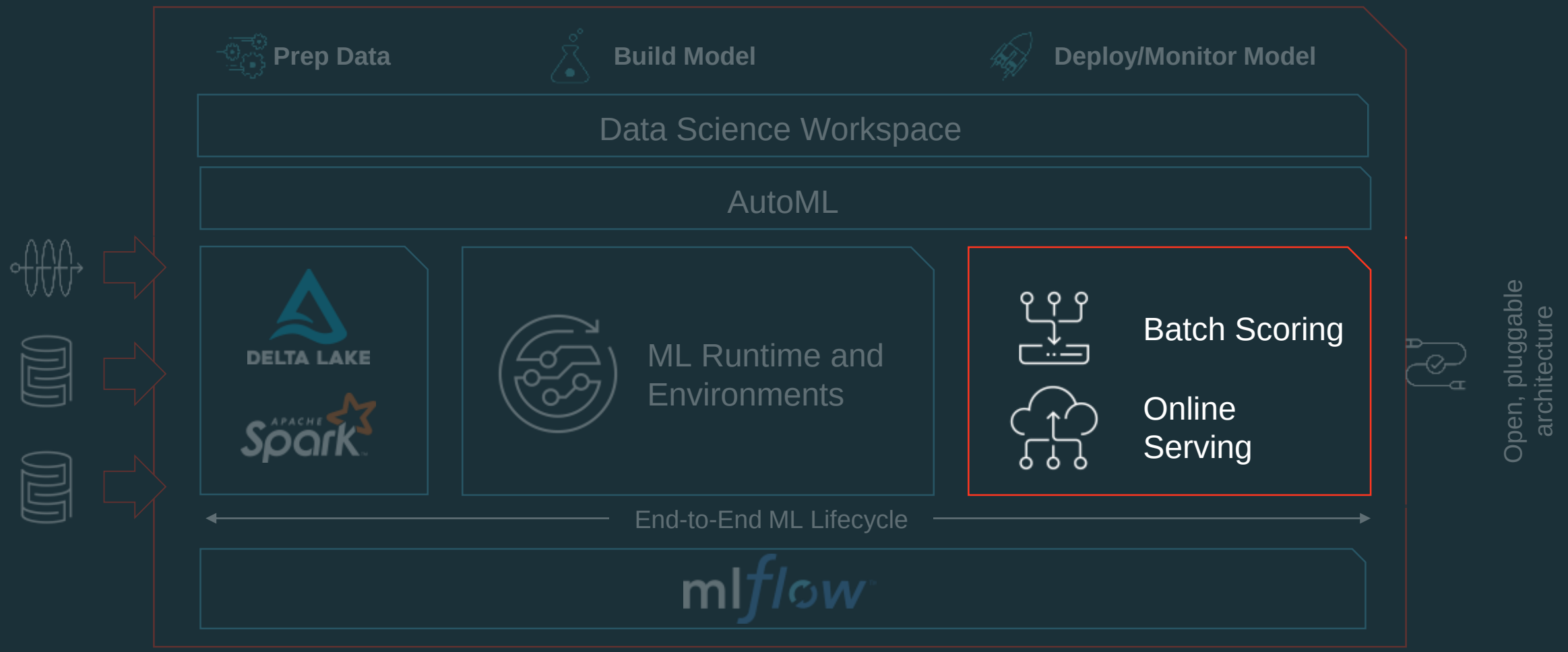
mlflow Integration with ML Runtime

Hyperopt autlogging to MLflow

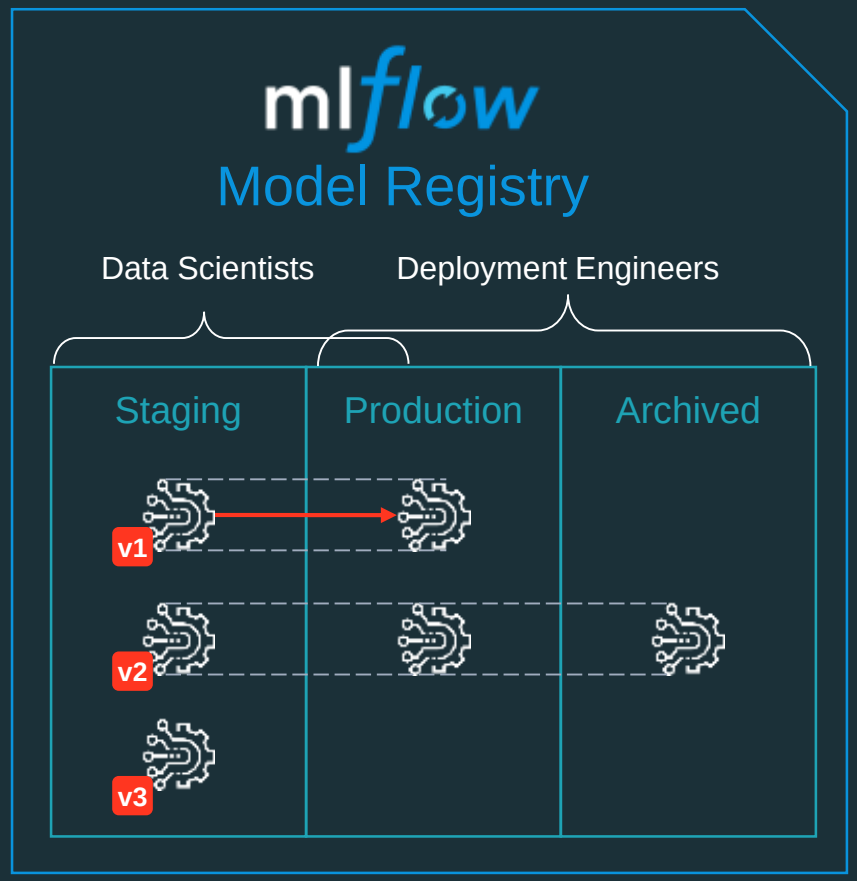




End-to-End Data Science and ML on databricks



mlflow Model Deployment



In-Line Code

Containers

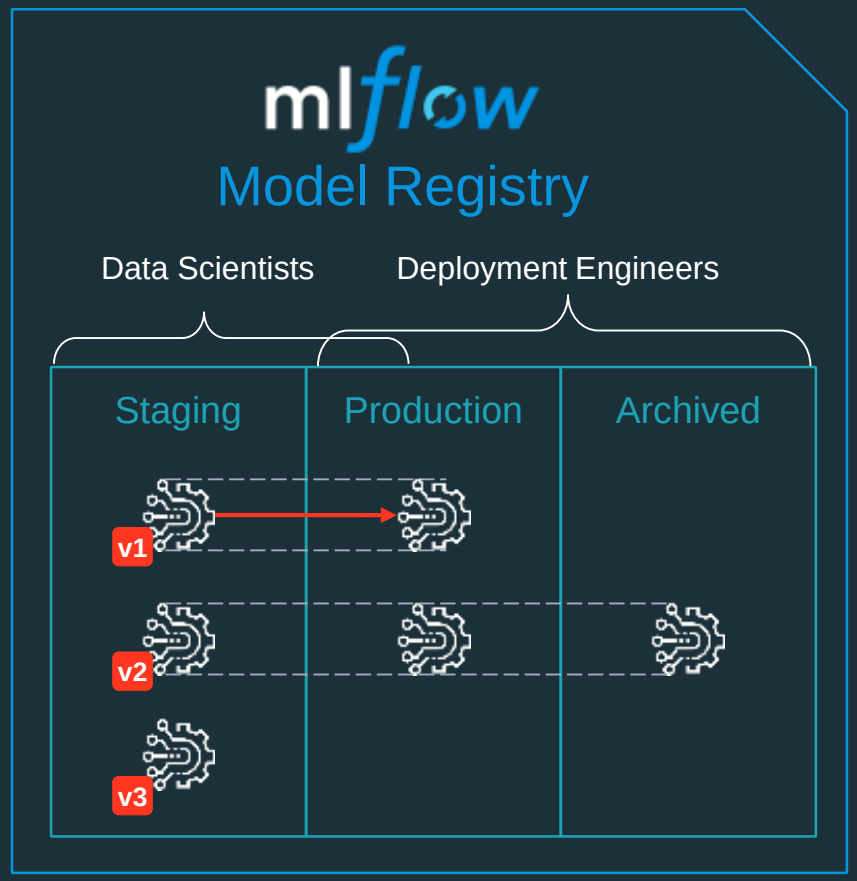
Batch & Stream Scoring

Cloud Inference Services

OSS Serving Solutions

databricks

mlflow Model Deployment



In-Line Code

Containers

Batch & Stream Scoring

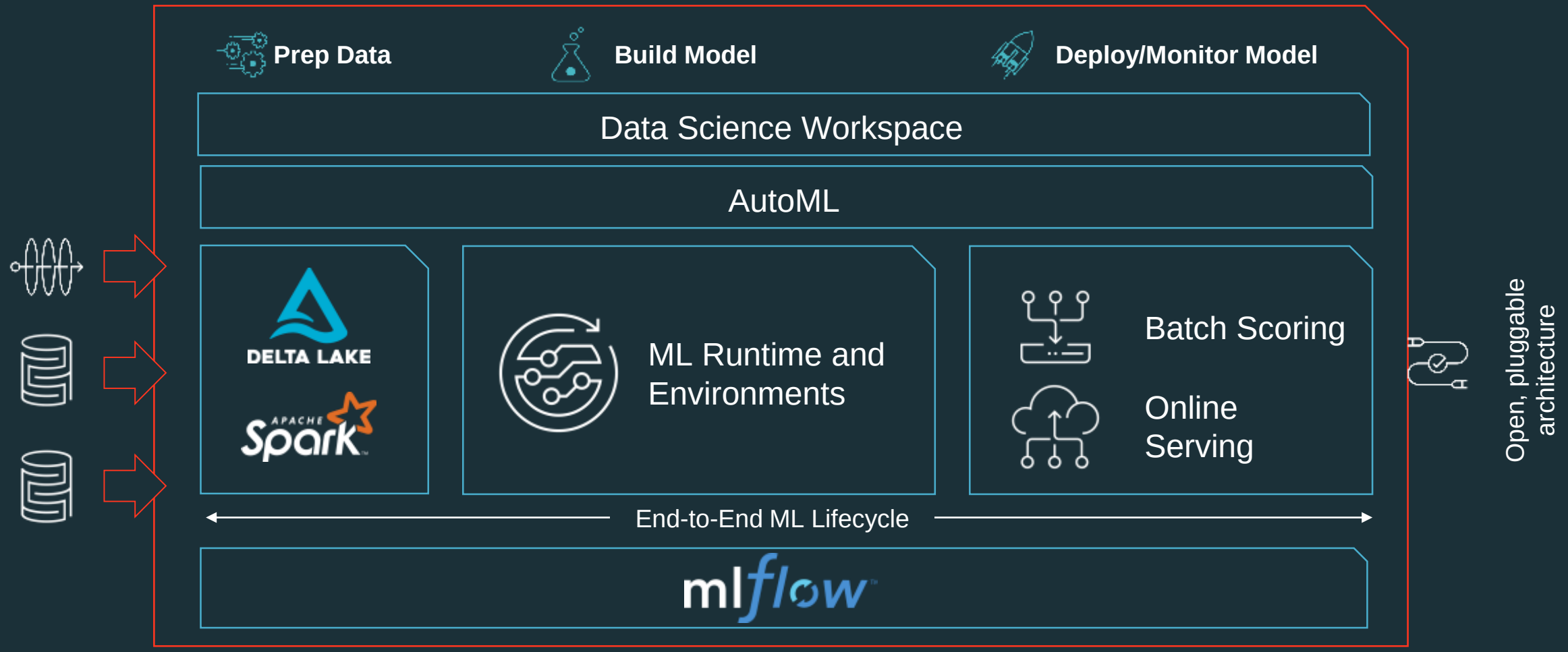
Cloud Inference Services

OSS Serving Solutions

```
model_udf =
    mlflow.pyfunc.spark_udf(
        spark,
        model_uri='models:/forecast/production')
```

databricks

In summary, databricks accelerates the full ML Lifecycle



databricks