WANT END-TO-END MLOPS? DELTA & DATABRICKS MAKE THIS A REALITY!

WHO ARE WE?

- Tori Tompkins
  - Senior Data Science Consultant
  - Cohost of Data Podcast, Totally Skewed

- Alex Billington
  - Senior Data Science Consultant
  - ML Engineer for 5 years

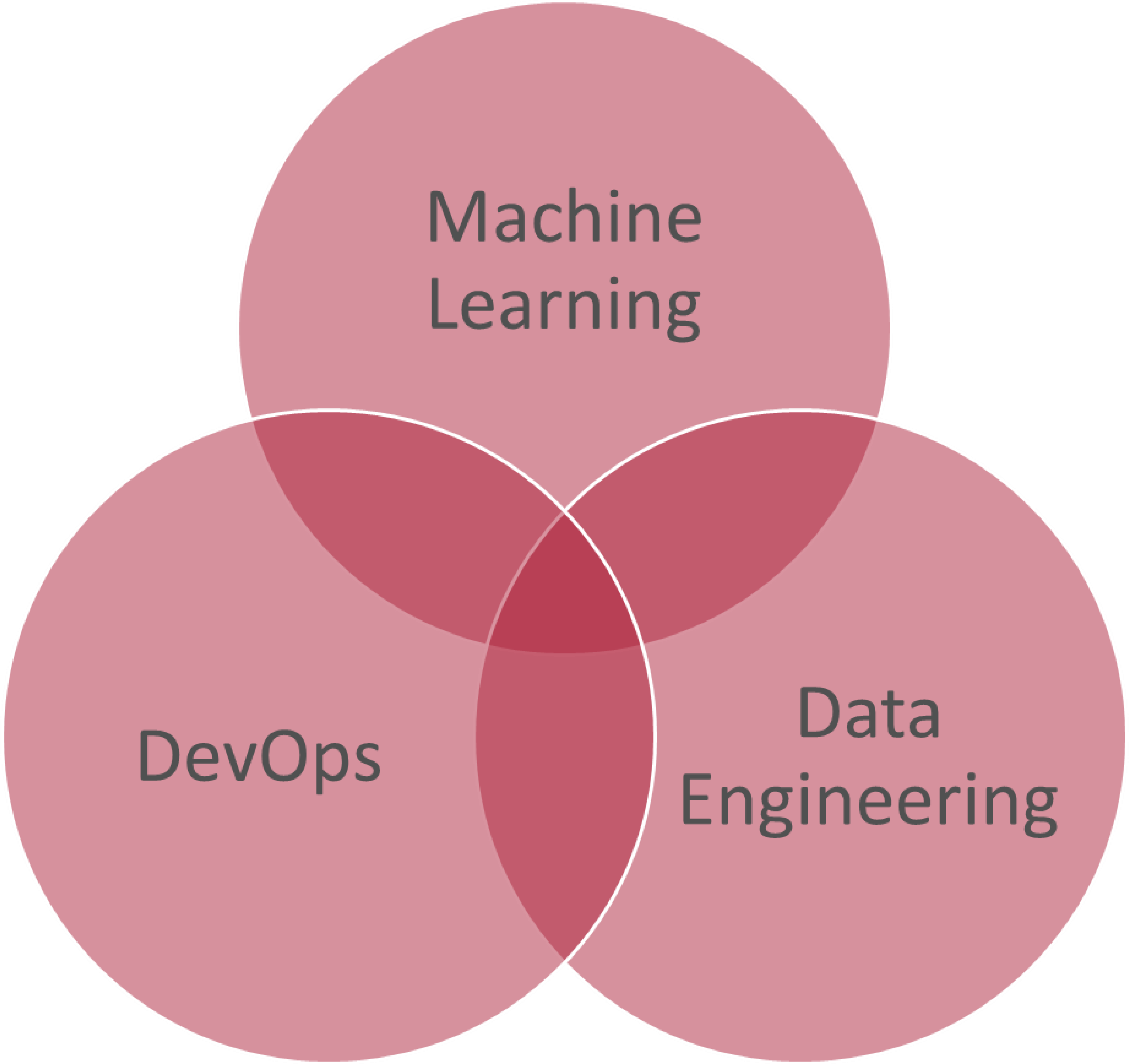
AGENDA

- What is MLOps?
- What is the MLOps lifecycle?
- What is Databricks?
- Databricks features
  - Feature store
  - MLflow
  - Serving Endpoints

FOLLOW ALONG CODE

- https://github.com/ABillington96/MLOps_in_Databricks

WHAT IS MLOPS?

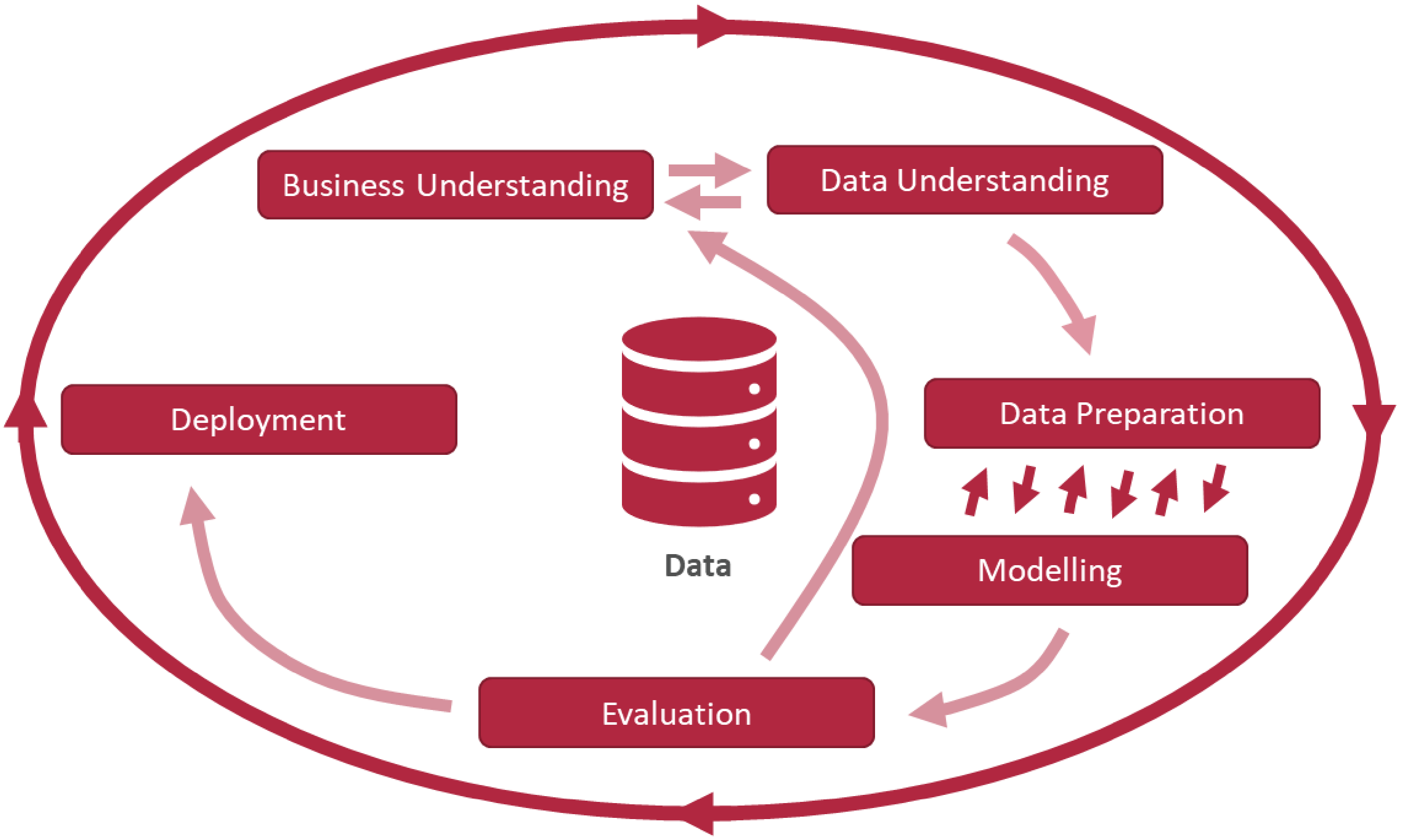


WHY IS MLOPS IMPORTANT?

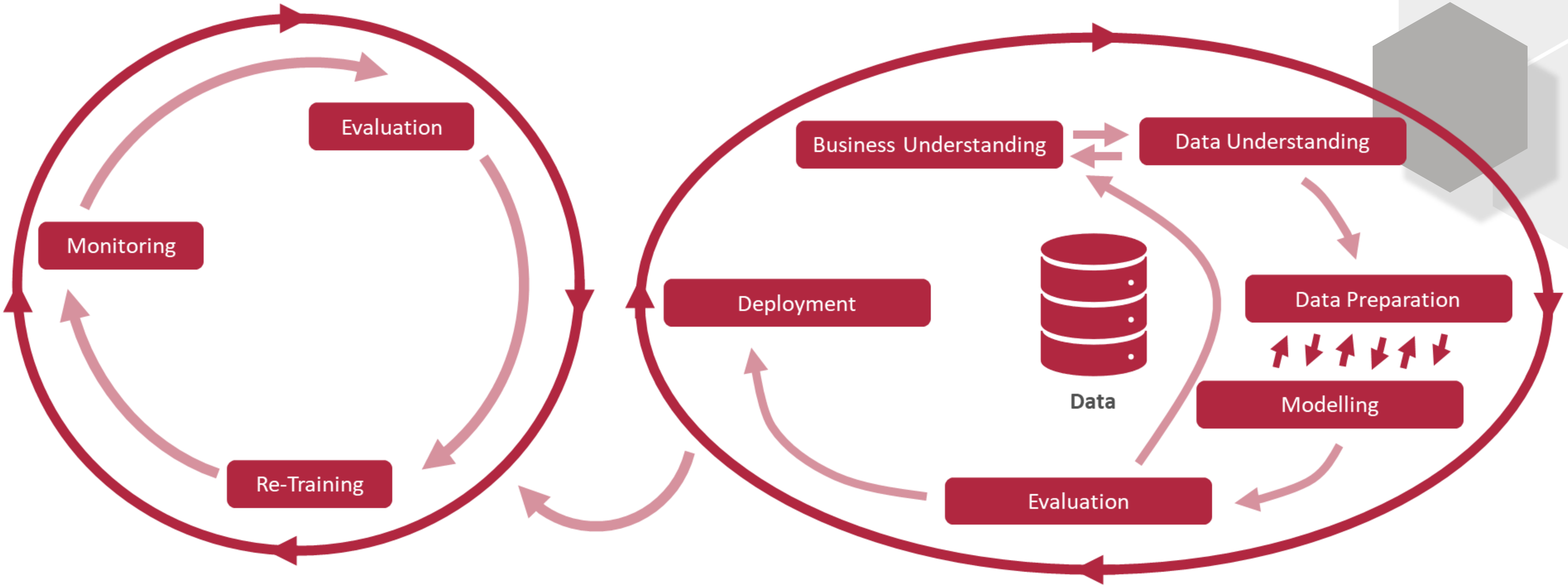
- Scaling
- Trust
- Better Integration
- Compliance
- Reduce Risk and Bias

The MLOps market could grow to around $2 billion by 2025, up from about $185 million in 2020.

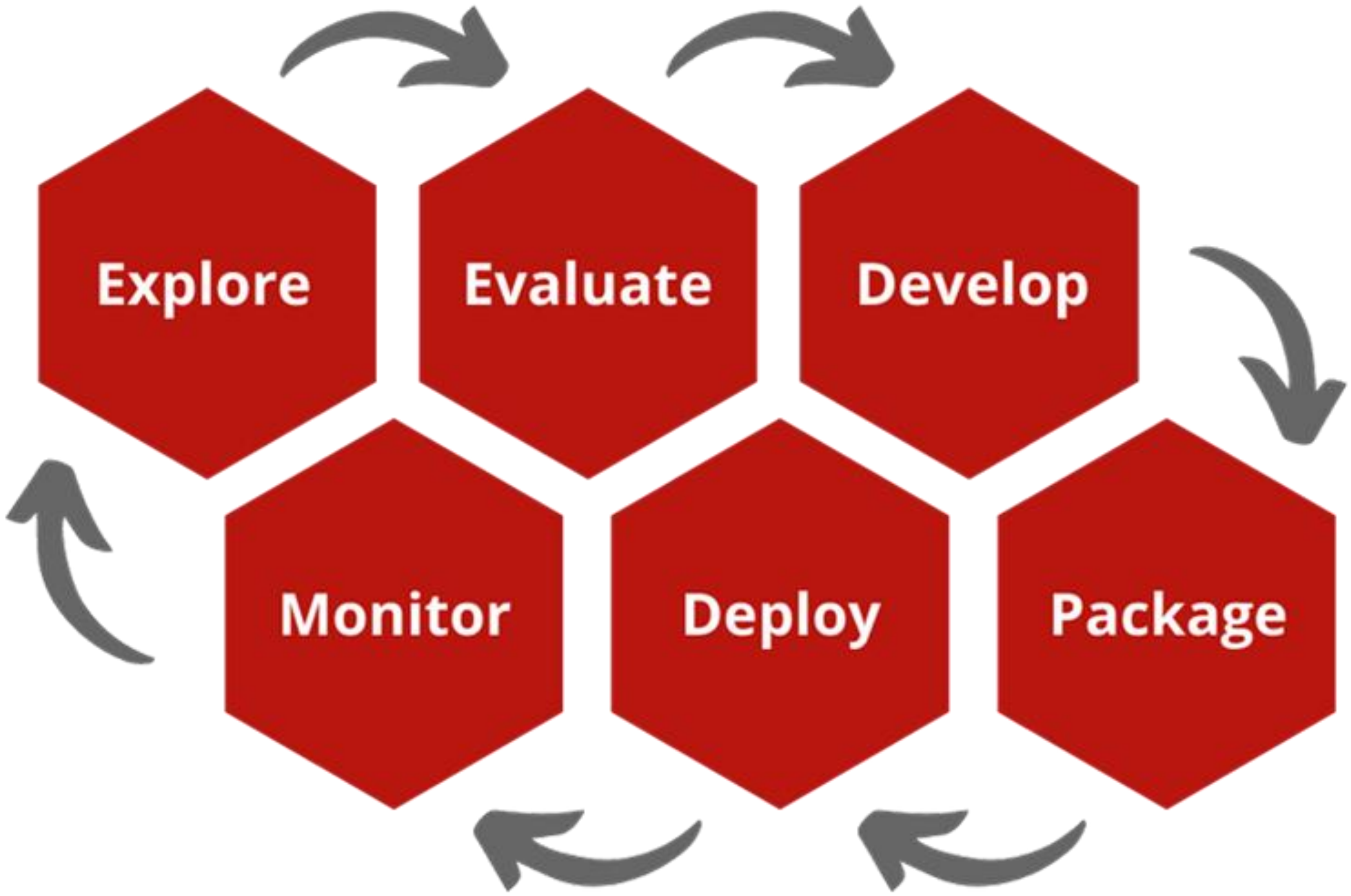
WHY IS MLOPS DIFFERENT FROM DEVOPS?

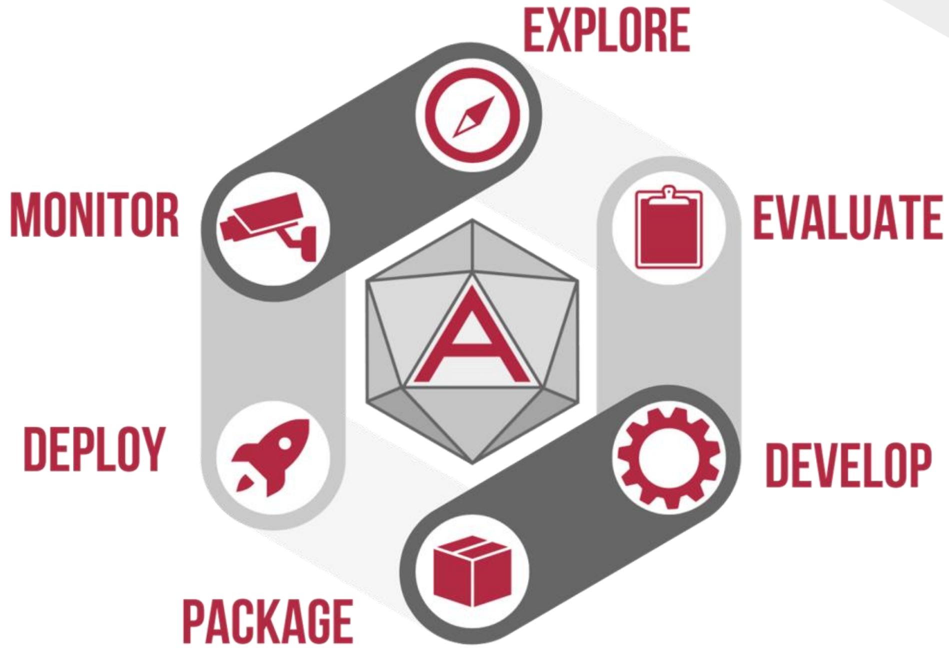


WHY IS MLOPS DIFFERENT FROM DEVOPS?



MLOPS LIFECYCLE





EXPLORE

Explore

- Experimental Phase for Data Scientists (DS) to try models
- MLOps should supply DS the environment they need to achieve this
- This can include leveraging feature stores, model experiment tracking and collaborative workspace

EVALUATE

Evaluate

- Testing phase for models to be evaluated for business value, accuracy and ethics
- MLOps should provide a platform and process to achieve this
- This can include leveraging libraries to analyse results, interpretability, explainability and fairness

DEVELOP

Develop

- Development phase to build robustness into the model inference pipeline
- MLOps should ensure models continuing to the next phase are production-ready
- This can include auditing, unit-testing, data and model versioning

PACKAGE

Package

- Packaging phase to ensure the production environment resembles development
- MLOps should ensure resources are accurate and accessible to prevent unexpected behaviour
- This can include versioning of libraries and frameworks or packaging code into a wheel

DEPLOY

Deploy

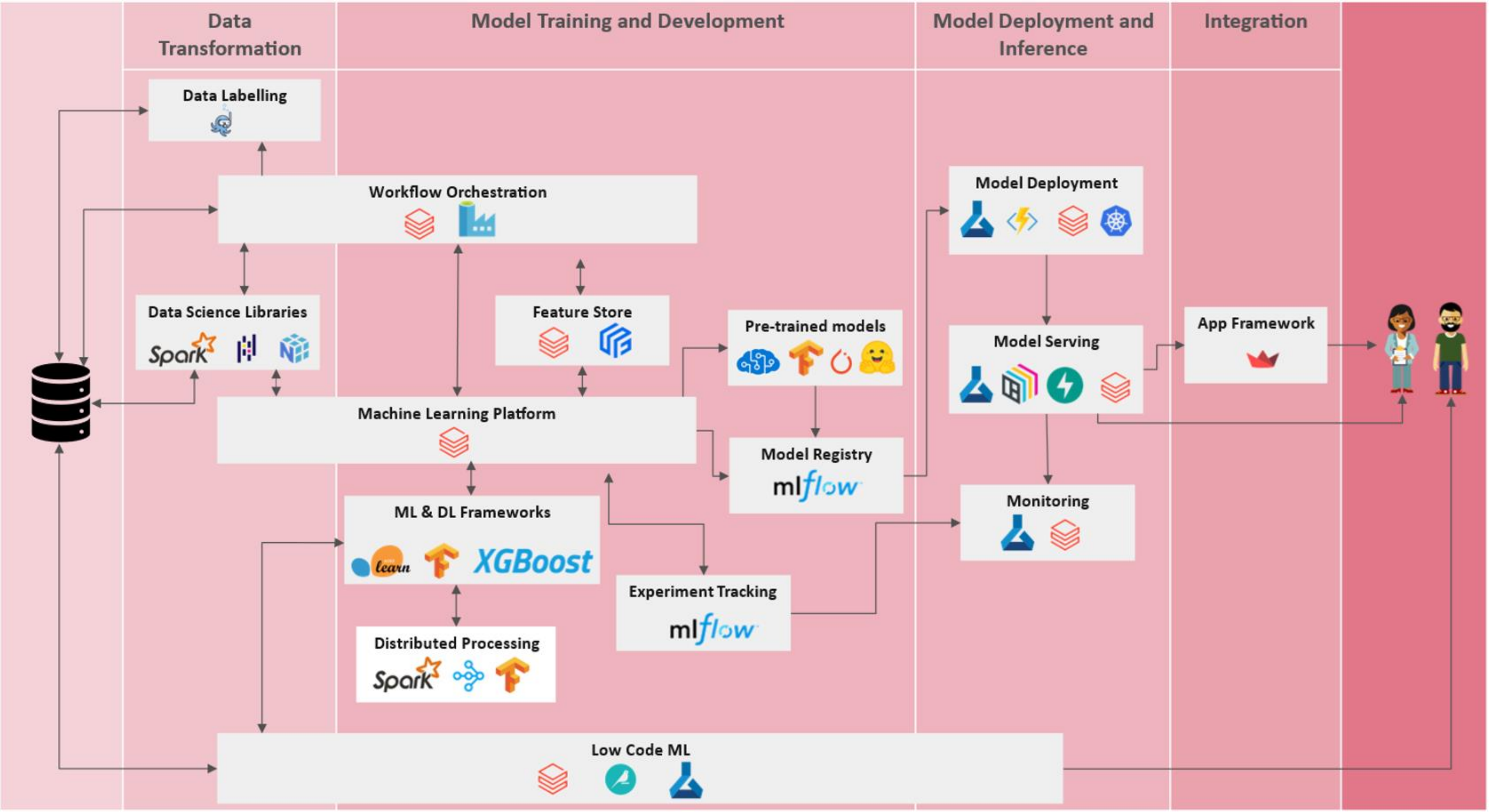
- Deployment phase to serve the model to user base
- MLOps should deploy the model according to business requirements
- This can include considering model complexity, size, importance, need for autoscaling or A/B testing etc

MONITOR

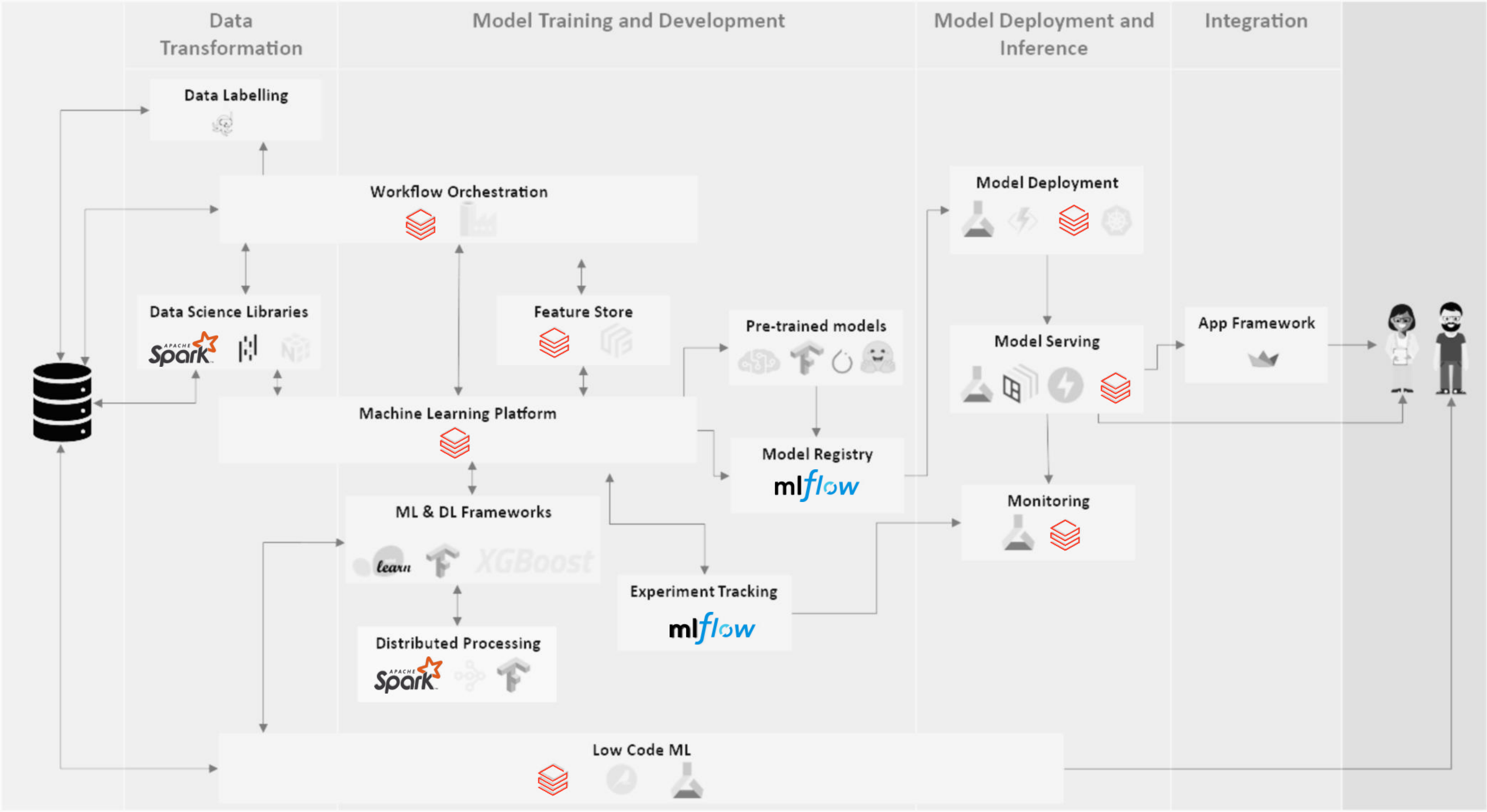
Monitor

- Monitoring phase to ensure models remain accessible and accurate
- MLOps should monitor models for performance and data drift
- This can include tracking input and output data, performing drift analysis and automatic alerts

MLOPS TOOLS


Data Transformation
Model Training and Development
Model Deployment and Inference
Integration
Data Labelling
Workflow Orchestration
Data Science Libraries
Spark
Feature Store
Pre-trained models
Machine Learning Platform
Model Registry
mlflow
ML & DL Frameworks
learn
XGBoost
Distributed Processing
Spark
Experiment Tracking
mlflow
Model Deployment
Model Serving
Monitoring
App Framework
Low Code ML

MLOPS TOOLS


Data Transformation
Model Training and Development
Model Deployment and Inference
Integration
Data Labelling
Workflow Orchestration
Model Deployment
Data Science Libraries
Feature Store
Pre-trained models
Model Serving
App Framework
Machine Learning Platform
Model Registry
mlflow
Monitoring
ML & DL Frameworks
learn
XGBoost
Experiment Tracking
mlflow
Distributed Processing
Spark
Low Code ML

WHAT IS DATABRICKS?

databricks

- Platform built on Spark
- Optimised for distributed computing
- Popular for Data Engineering, Analytics and Science workloads

WHAT IS DELTA LAKE?

DELTA LAKE

- **Open-source storage layer designed to run on top of an existing data lake**

- ACID transactions
- Scalable metadata handling
- Unified batch & streaming
- Schema enforcement
- Time travel (data versioning)
- Upserts and deletes
- Optimised file management

DELTA LAKE

A delta lake, an evolution of data storage, preserves the integrity of your original data without sacrificing the performance and agility required for real-time analytics, artificial intelligence (AI), and machine learning (ML) applications.

DATA LAKE

A data lake is a massive accumulation of raw data in multiple formats. The sheer volume and variety of information in a data lake can make analysis cumbersome and, without auditing or governance, the quality and consistency of the data can be unreliable.

A data lakehouse combines the flexibility and scalability of a data lake with the structure and management features of a data warehouse in a simple, open platform.

DATA LAKEHOUSE

A data warehouse gathers information from multiple sources, then reformats and organizes it into a large, consolidated volume of structured data that's optimized for analysis and reporting. Proprietary software and an inability to store unstructured data can limit its usefulness.

DATA WAREHOUSE

DATABRICKS FOR MLOPS

- Extremely powerful Spark engine
- Supports R, SQL and Python
- Collaborative workspaces
- Powerful machine learning libraries
- Feature Store
- Serving Endpoints
- AutoML
- Git integration
- Fully managed version of MLflow
- NEW Dolly LLM

THE EXPLORE PHASE - MLFLOW

WHY MLFLOW?

Explore

- Which dataset was used?
- Which ML algorithm did we use to train the model?
- What hyperparameter values were used?
- What were the performance like?
- How can we convey all these info to our team easily?

Reproducible

Traceable

MLFLOW - COMPONENTS

Explore

mlflow Projects

Packaging format for reproducible runs on any compute platform

mlflow Tracking

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

mlflow Models

General model format that standardizes deployment options

mlflow Model Registry

Centralized and collaborative model lifecycle management

DATABRICKS MLFLOW

Experiment Tracking	Open Source MLflow	Managed MLflow on Databricks
MLflow tracking API	✓	✓
MLflow tracking server	Self-hosted	Fully managed
Notebooks integration	✗	✓
Workspace integration	✗	✓
Reproducible Projects		
MLflow Projects	✓	✓
Git and Conda integration	✓	✓
Scalable cloud/clusters for project runs	✗	✓
Model Management		
MLflow Model Registry	✓	✓
Model versioning	✓	✓
ACL-based stage transition	✗	✓
CI/CD workflow integrations	✓	✓

Explore

DATABRICKS MLFLOW

Flexible Deployment		
MLflow Models	✓	✓
Built-in batch inference	✗	✓
Built-in streaming analytics	✗	✓
Security and Management		
High availability	✗	✓
Automated updates	✗	✓
Role-based access control	✗	✓

Integrates with Feature Store

Facilitates Databricks Serving Endpoints

Explore

MLFLOW DEMO

Explore

THE EVALUATE PHASE – MLFLOW (AGAIN)

EVALUATION

- mlflow.evaluate
- Built in metrics
  - **Regressor models:** example_count, mean_absolute_error, mean_squared_error, root_mean_squared_error, sum_on_target, mean_on_target, r2_score, max_error, mean_absolute_percentage_error.
  - **Binary classifiers:** true_negatives, false_positives, false_negatives, true_positives, recall, precision, f1_score, accuracy_score, example_count, log_loss, roc_auc, precision_recall_auc.
  - **Multiclass classifiers:** accuracy_score, example_count, f1_score_micro, f1_score_macro, log_loss
- Custom metrics
- Explainability metrics
  - SHAP
- Fairness metrics
  - Fairlearn

Evaluate

THE DEVELOP PHASE – FEATURE STORE

DATABRICKS FEATURE STORE

Develop

A "**feature**" refers to the entire column in the dataset

Transaction_id	in_foreign_country	size_compared_to_avg_transaction	fraud?
7485	False	0.8x	False
46854	True	21.2x	True
3521	True	$\boxed{1.1x}$	False

A "**feature value**" refers to a single value of a feature column

FEATURE STORE CONCEPTS

Develop

- Registry – Central interface for managing feature definitions + metadata
- Offline Store – Store large volumes of feature data used train and test
- Online Store – Low latency database for real-time inference

WHY DO YOU NEED A FEATURE STORE?

- Easily reuse new features
- Easily explore features
- Data pipelines can be shared across both training and serving
- Constant features across teams
- Provides lineage
- Feature tracking

Develop

ADVANCING ANALYTICS

DATABRICKS FEATURE STORE

Develop

- Backed by Delta Lake
- Feature Store Library
- Feature Store UI
- Offline Store
- Online Store Compatability
- Training Set Functionality

DATABRICKS FEATURE STORE DEMO

Develop

ADVANCING
ANALYTICS

THE PACKAGE PHASE – CI/CD

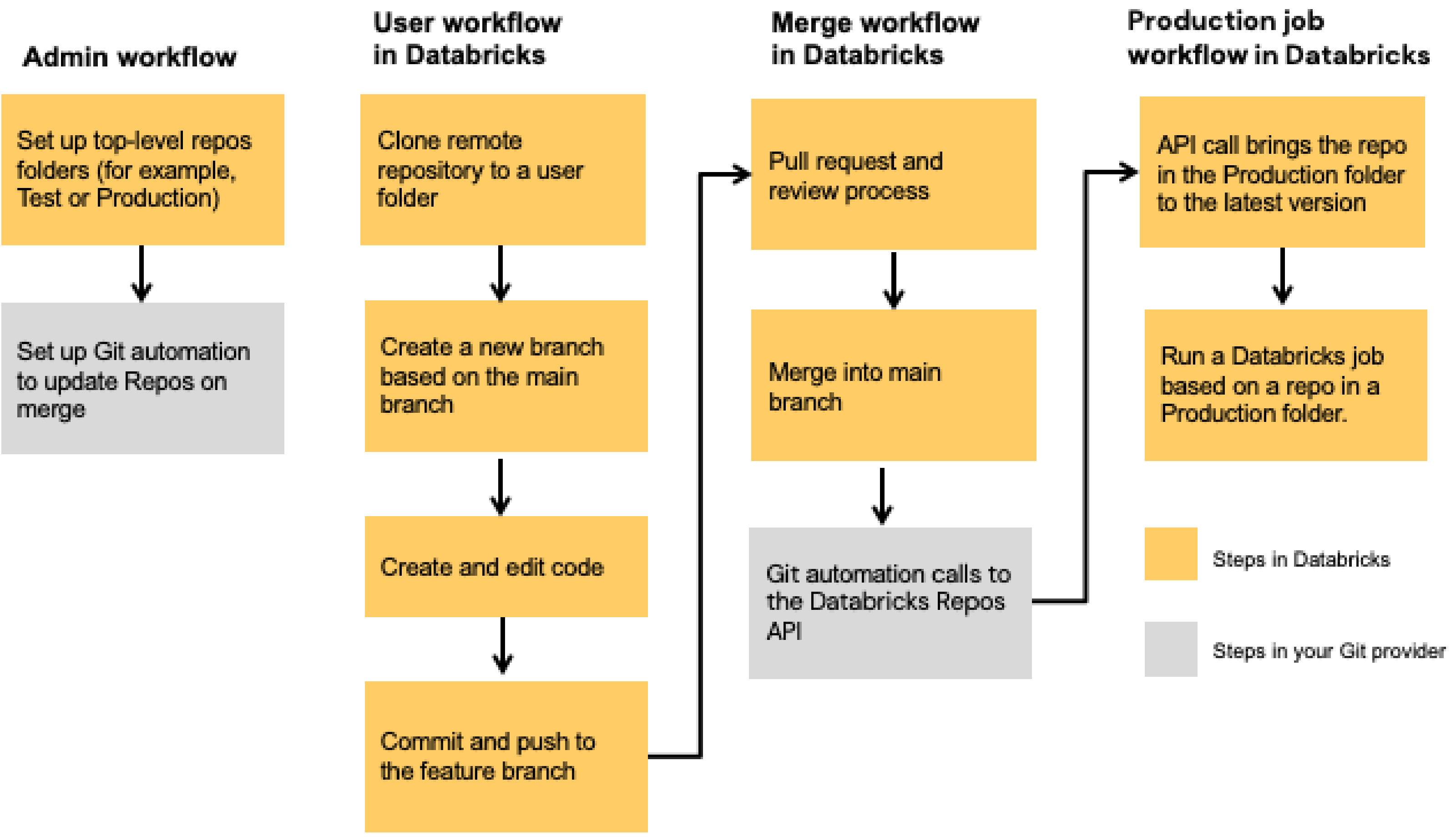
DEVOPS

Bitbucket

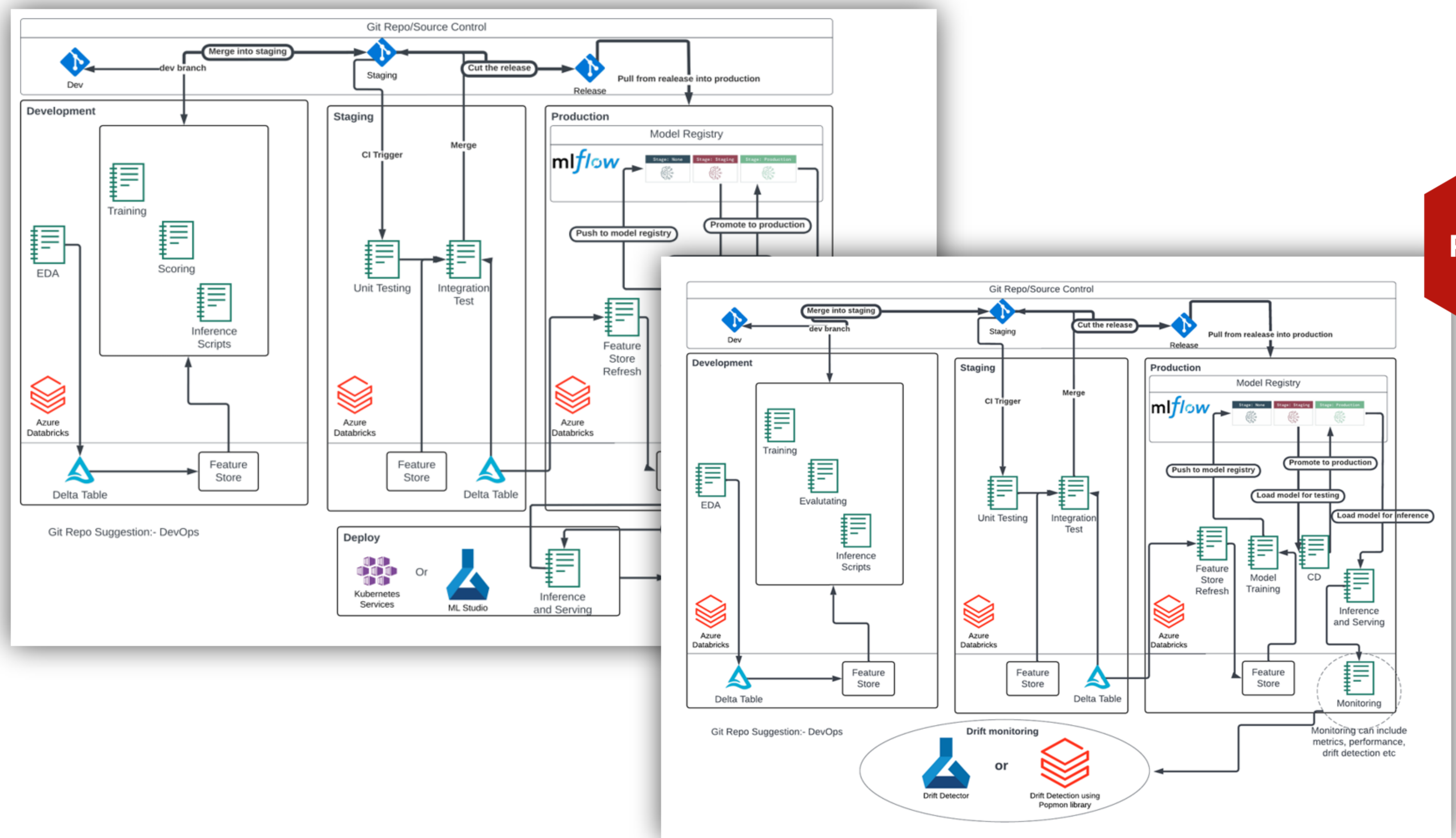
GitLab

AWS CodeBuild

Package

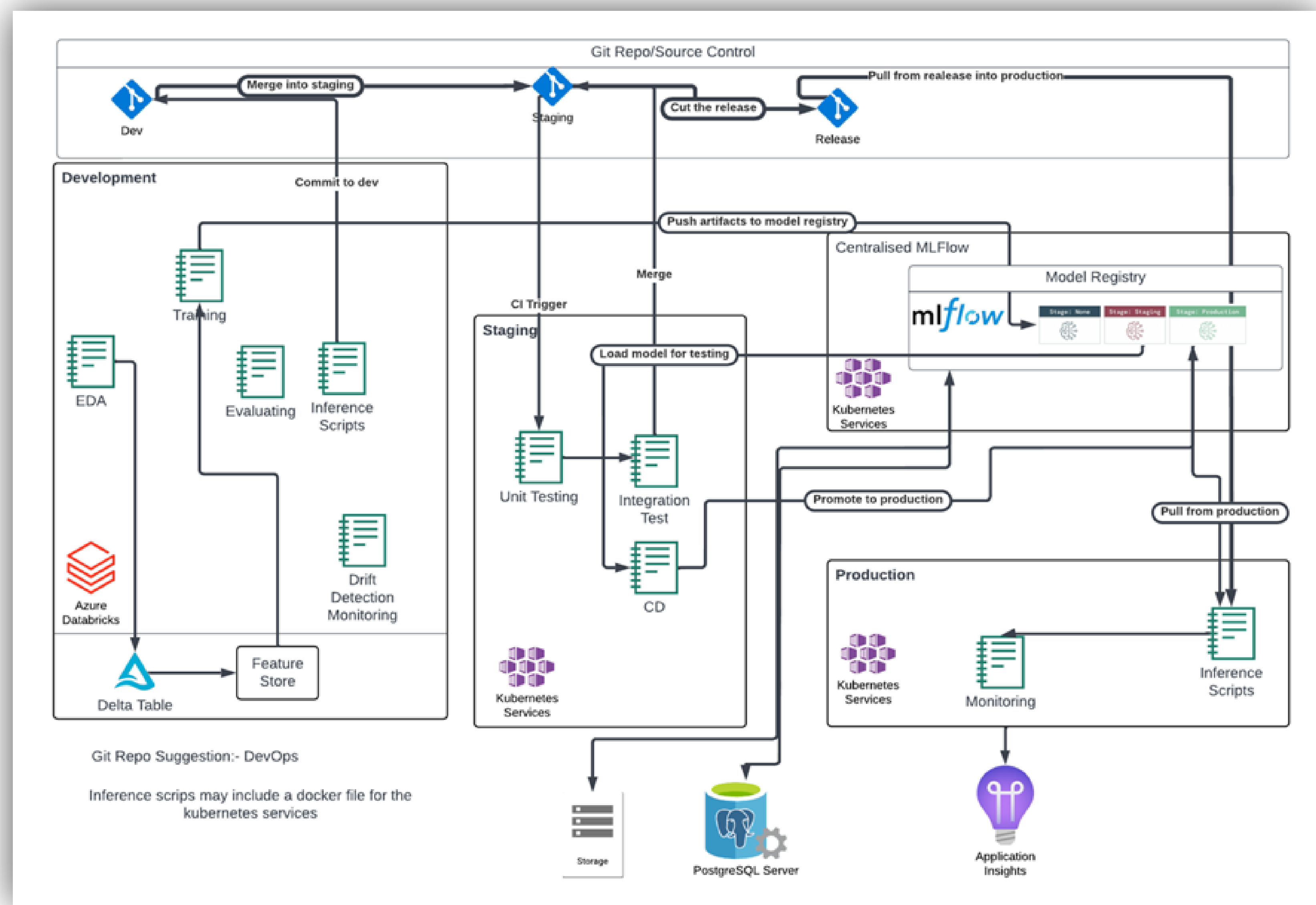


CODE-FIRST DEPLOYMENT



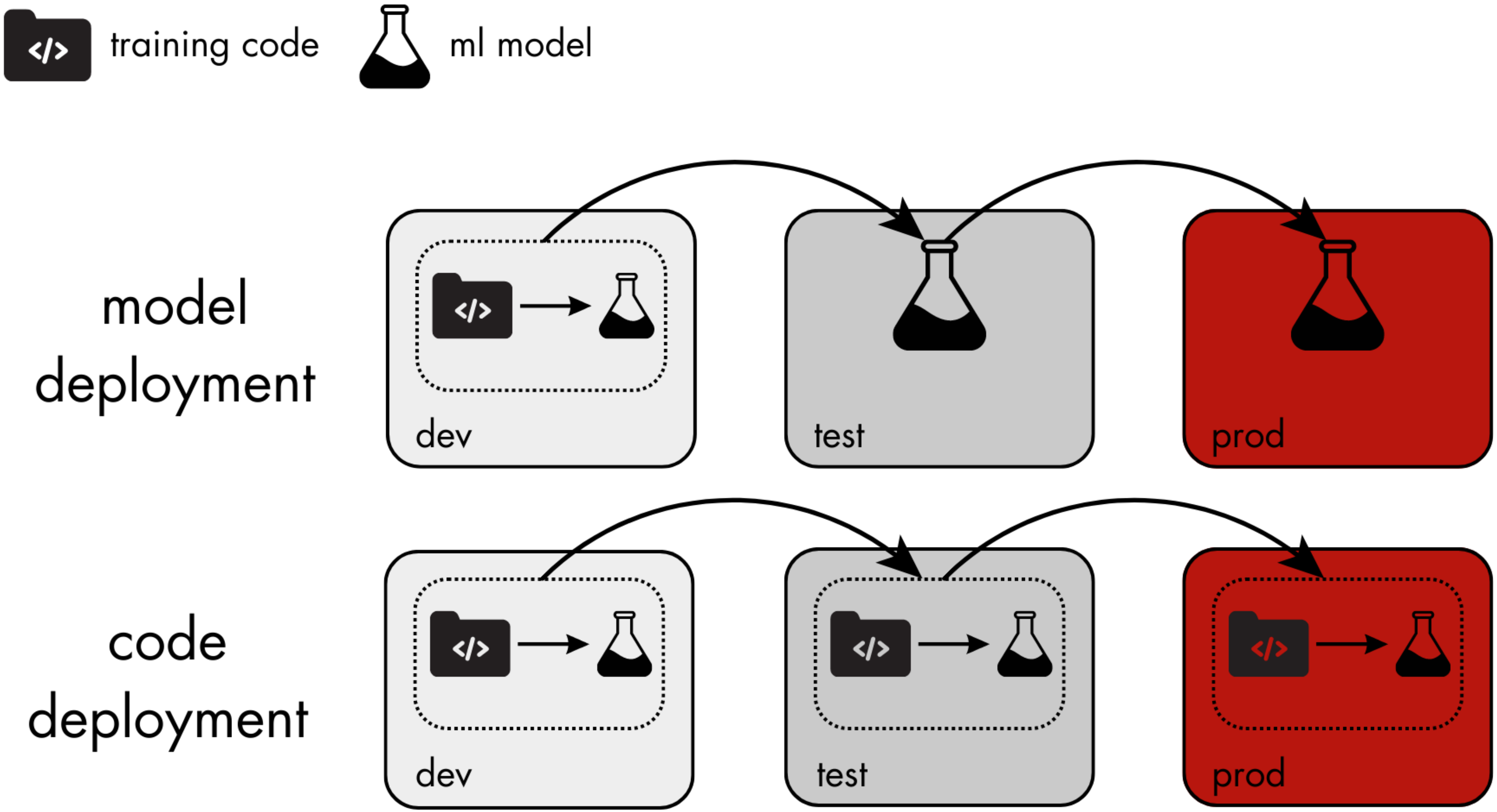
Package

MODEL-FIRST DEPLOYMENT



Package

CODE-FIRST VS MODEL-FIRST



Package

THE DEPLOY PHASE – SERVING ENDPOINTS

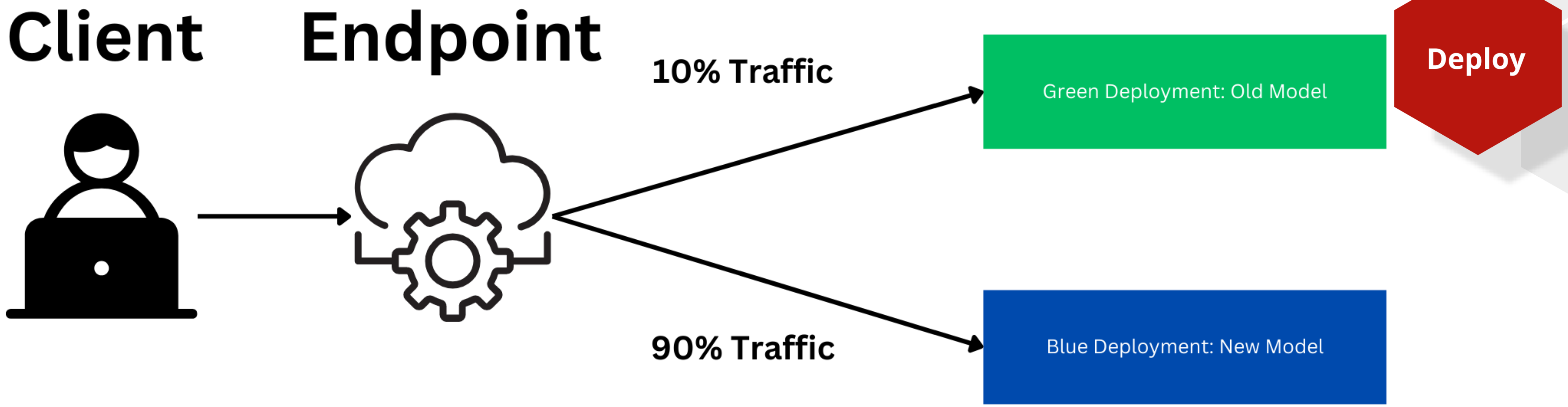
MODEL SERVING

Deploy

- Real time scoring provides near real-time results on small records
  - Fraud Detection
  - Predictive Maintenance
  - Recommender systems

- Batch scoring provides results for a large volume of records in a single operation
  - Churn Prediction
  - Credit Risk Analysis
  - Forecasting

GREEN BLUE DEPLOYMENT



DATABRICKS SERVING/SERVERLESS ENDPOINTS

Deploy

Serverless doesn’t mean no servers!

The cloud providers take care of underlying infrastructure

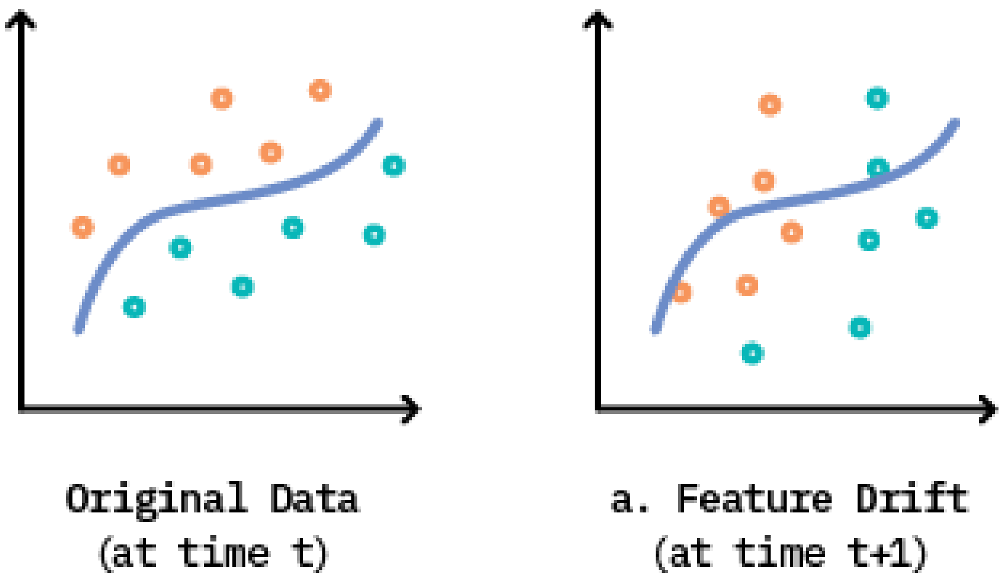
- Scalable
- Cost-effective
- Flexible
- Easy to deploy & manage

DATABRICKS SERVING ENDPOINT DEMO

Deploy

THE MONITOR PHASE – MODEL DRIFT

MODEL DRIFT



- Data drift
- Label drift
- Prediction drift
- Concept drift

Monitor

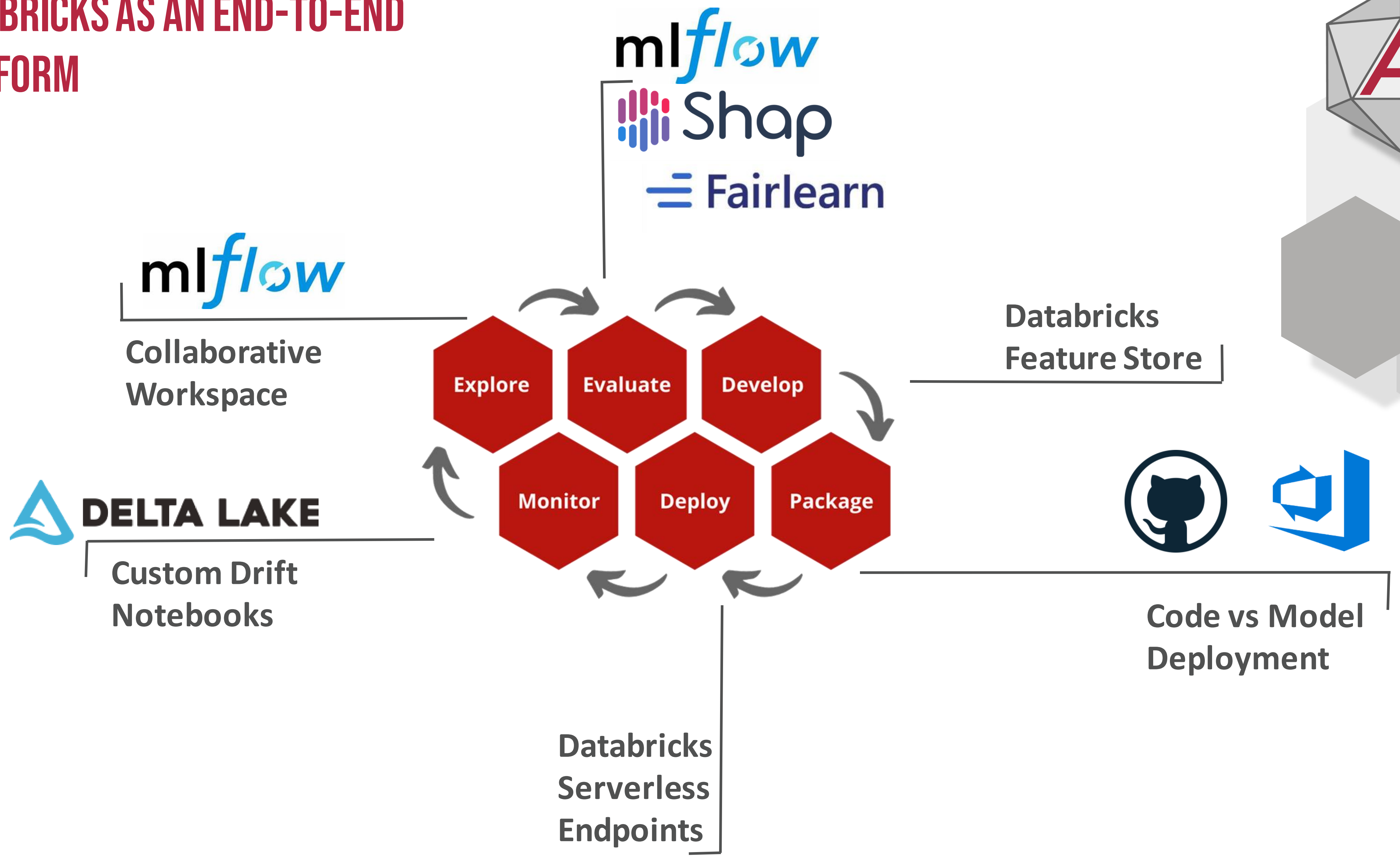
DRIFT TRACKING IN DATABRICKS

Monitor

- No current explicit support
- Log data with Delta
- Custom notebooks:
  - Summary statistics
  - Correlation statistics
  - Kullback-Liebler Divergence/ Jenson-Shannon Divergence
  - Chi-Squared test

END-TO-END MLOPS – DATABRICKS?

DATABRICKS AS AN END-TO-END PLATFORM

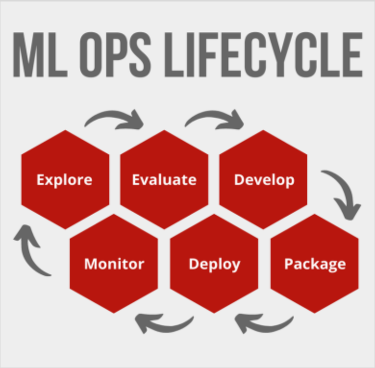


LEARN MORE

FEATURE STORES, MOVIES, ICECREAM AND S'MORE

DESIGNING THE RIGHT ML PLATFORM WITH MLOPS

FEATURE STORES AND WHY YOU NEED THEM



WHAT IS MLOPS AND WHY DO YOU NEED IT?

HOW TO FIX DIFFERENT TYPES OF MODEL DRIFT

QUESTIONS?

ADVANCING
ANALYTICS