

MLOps avec Databricks : incontournables et leçons apprises.

Histoires d'échecs, de succès et
conseils



databricks



agileDSS

Présentateur - Vincent Fortier

Manager, Field Engineering, Eastern Canada



3.5 ans chez Databricks

3 ans chez Cloudera

3+ années d'amélioration de ma recette de pizza

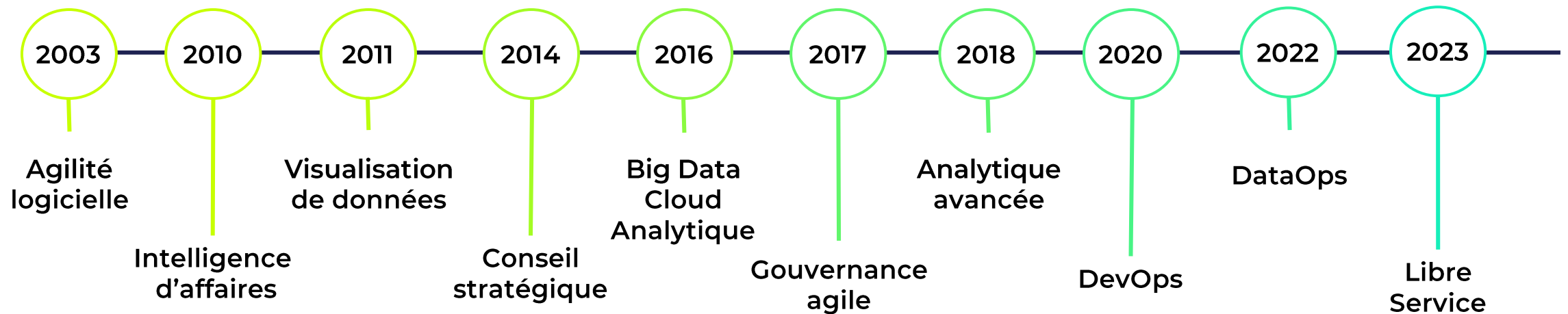
2 chiens saucisse

vincent.fortier@databricks.com



+ agileDSS

Fondé à Montréal en 2003, agileDSS est un collectif d'experts pluridisciplinaires qui vous accompagne dans la mise en place de votre écosystème analytique pour accélérer votre prise de décision.



+ Agenda

1. Introduction
2. Causes des échecs de projets en IA
3. MLOps comme facteur de succès
4. Déployer du MLOps avec Databricks



+ Introduction : chiffres clés

- + En 2021, Gartner Inc. prévoyait que 85 % des implémentations d'IA échoueraient
- + Ils prévoyaient que cette tendance se maintiendrait en 2022

+ Introduction : Définition d'un projet en IA réussi



Projet à valeur ajoutée



Projet qui a un retour sur investissement positif



Projet qui procure un avantage compétitif

+ Cause des échecs de projets en IA

1 Organisation

2 Processus

3 Technologique

1

Organisation

- + Engagement exécutif faible
- + Manque de collaboration entre les équipes

+ Cause des échecs de projets en IA

1

Organisation

2

Processus

3

Technologique

2

Processus

- + Gouvernance de données non mature
- + Absence de surveillance de la performance en production
- + ROI non mesurable

+ Cause des échecs de projets en IA

1

Organisation

2

Processus

3

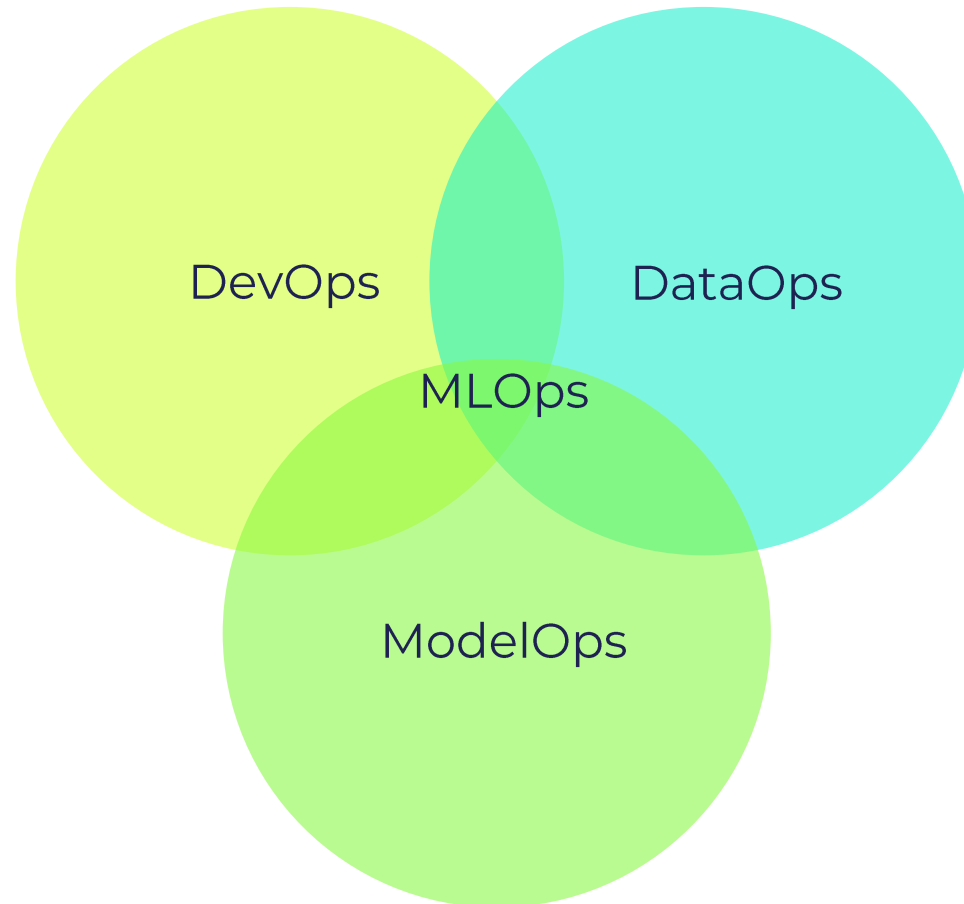
Technologique

3

Technologique

- + Transition du mode expérimental à la production coûteuse
- + Modèles non interprétables (type boîte noire)

+ MLOps comme facteur de succès : définition



+ MLOps comme facteur de succès : Pourquoi?

+ MLOps vous aide à réduire les risques :

- + Risque de conformité : violation des réglementations, de politique corporative
- + Risque technique : modèles de mauvaise qualité, infrastructure fragile

+ MLOps améliore l'efficacité à long terme grâce à l'automatisation :

- + Capturer des erreurs avant le déploiement en production
- + Éviter les tâches manuelles

+ MLOps comme facteur de succès : Principes

P1 : Automatisation du CI/CD

P2 : Orchestration du flux de travail

P3 : Reproductibilité

 P4 : Gestion des versions de données, du code, du modèle

P5 : Collaboration

 P6 : Apprentissage + évaluation continus de modèles

 P7 : Suivi des métadonnées de l'apprentissage auto

P8 : Surveillance continue

 P9 : Boucle de rétroaction

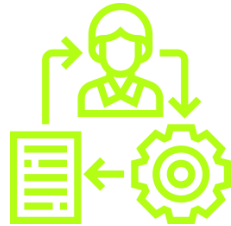
Source : arxiv.org/abs/2205.02302

+ MLOps comme facteur de succès : Intervenants



Parties prenantes

- + Responsable de la valeur ajoutée de la solution IA



Ingénieur de données

- + Bâti des pipelines de données



Scientifique de données

- + Traduit un problème d'affaires en tâche de IA
- + Bâti des modèles



Ingénieur ML

- + Déploie des modèles en production

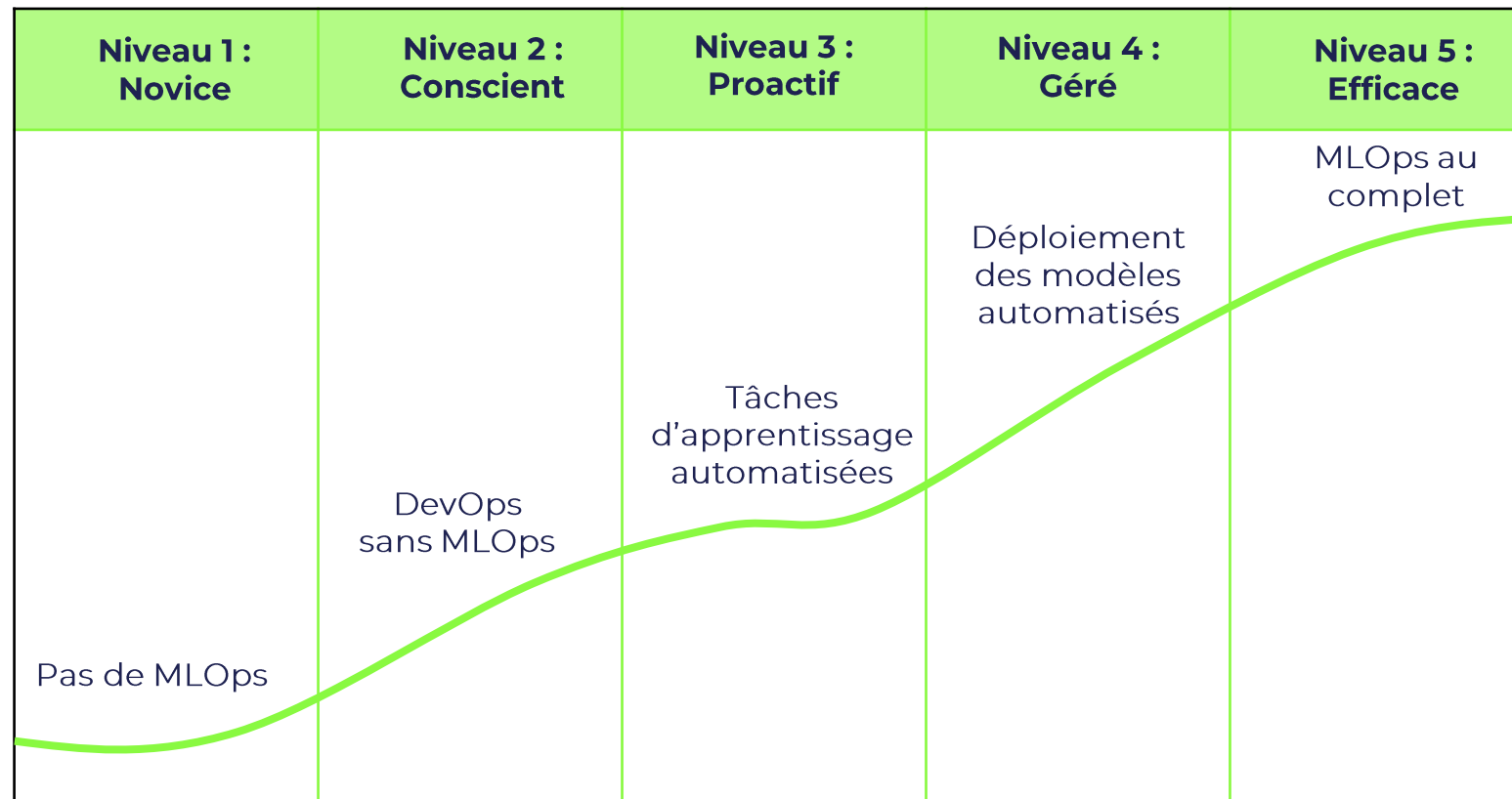


Chef gouvernance de données

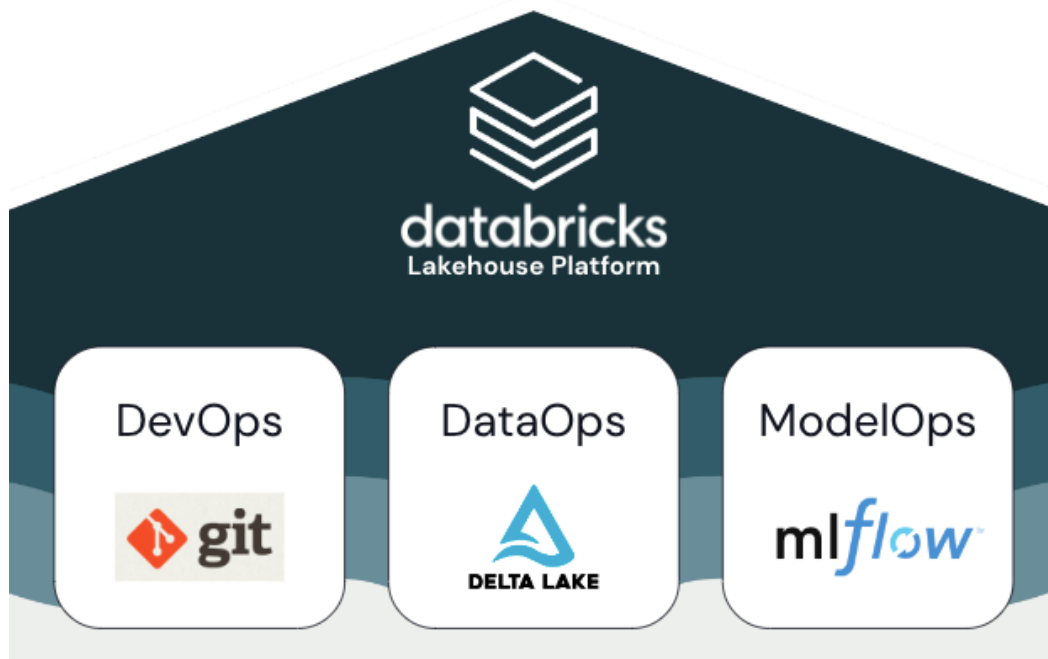
- + Responsable de la gouvernance des données et de la conformité

Source : arxiv.org/abs/2205.02302

+ MLOps: Niveaux de maturité



+ Déployer du MLOps avec Databricks: Pourquoi ?



- + Disponible sur Azure, AWS et Google Cloud
- + Plateforme pour données structurées et non structurées
- + Intégration facile
- + Déploiement du DevOps, DataOps, ModelOps supportés
- + Offre Unity catalog pour la gouvernance de données

Source: https://learn.microsoft.com/en-us/azure/databricks/_static/images/machine-learning/ml-ops-lakehouse.png

+ Déployer du MLOps avec Databricks



Adoptez une approche incrémentale basée sur les niveaux de maturité MLOps



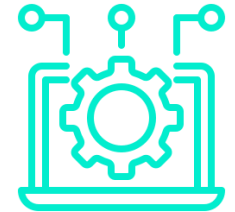
Assurez-vous d'avoir l'expertise nécessaire



Favorisez les bonnes pratiques

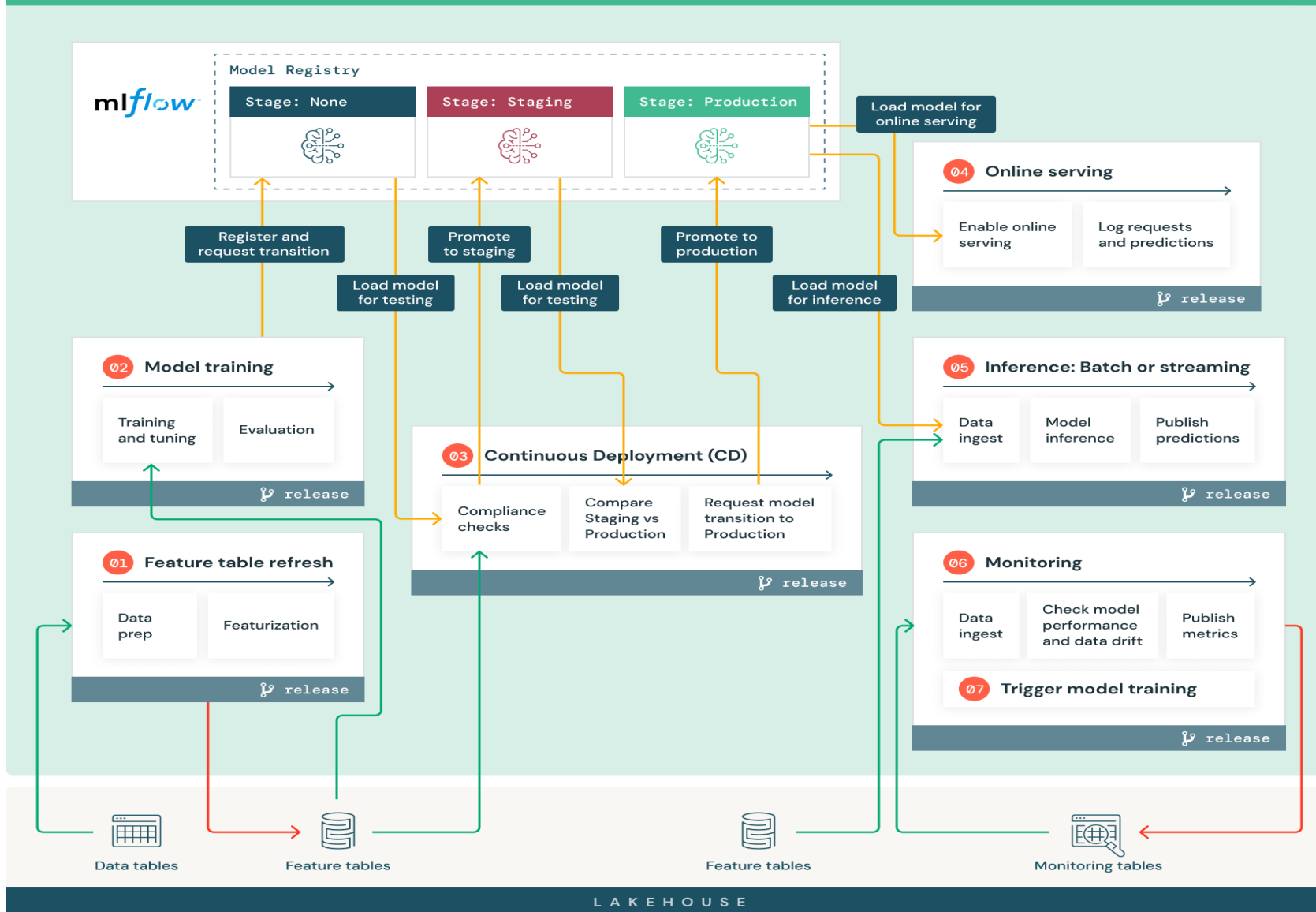


Faites des choix technologiques justifiés



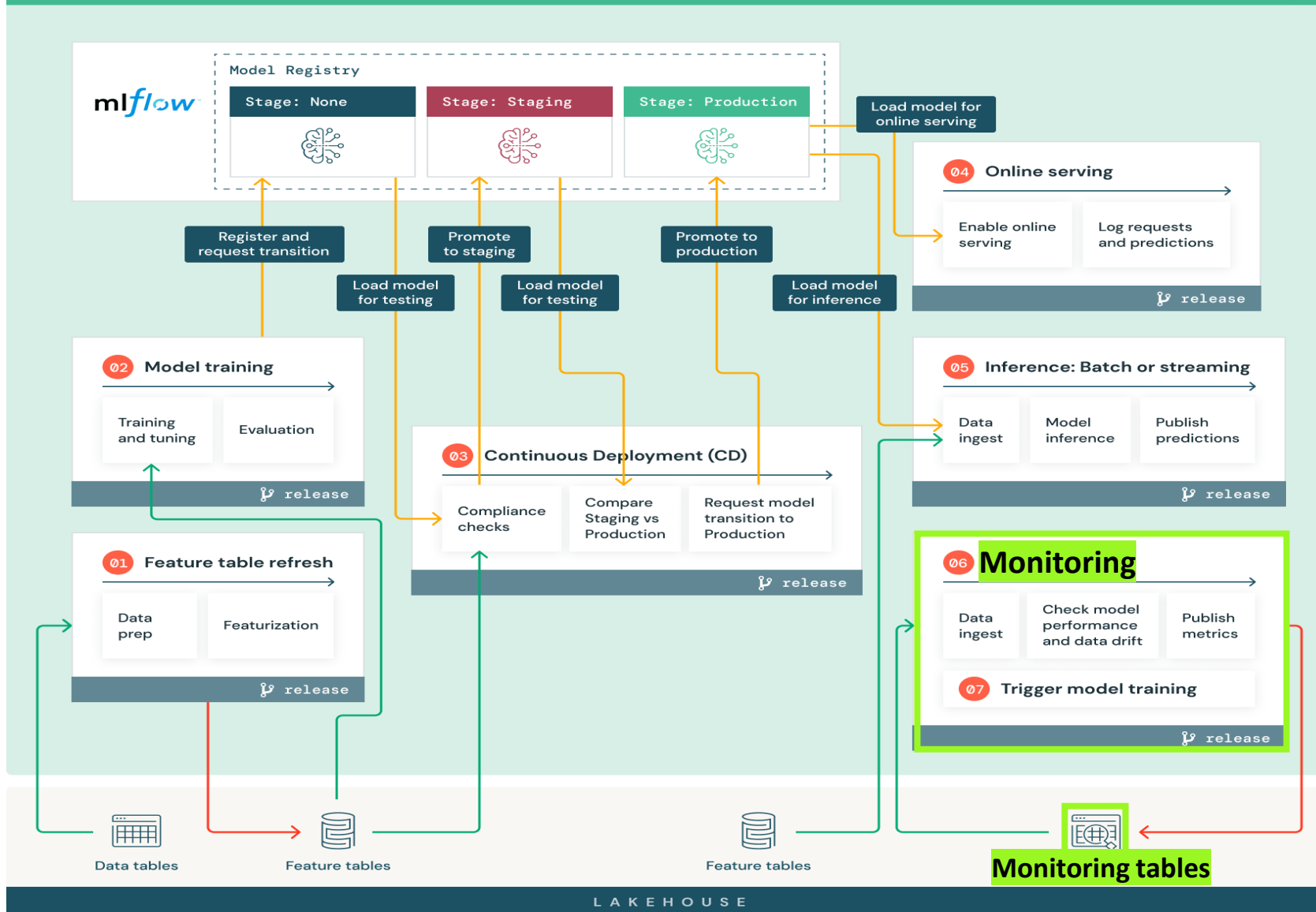
Favorisez l'approche plateforme au détriment de l'approche projet

Production environment



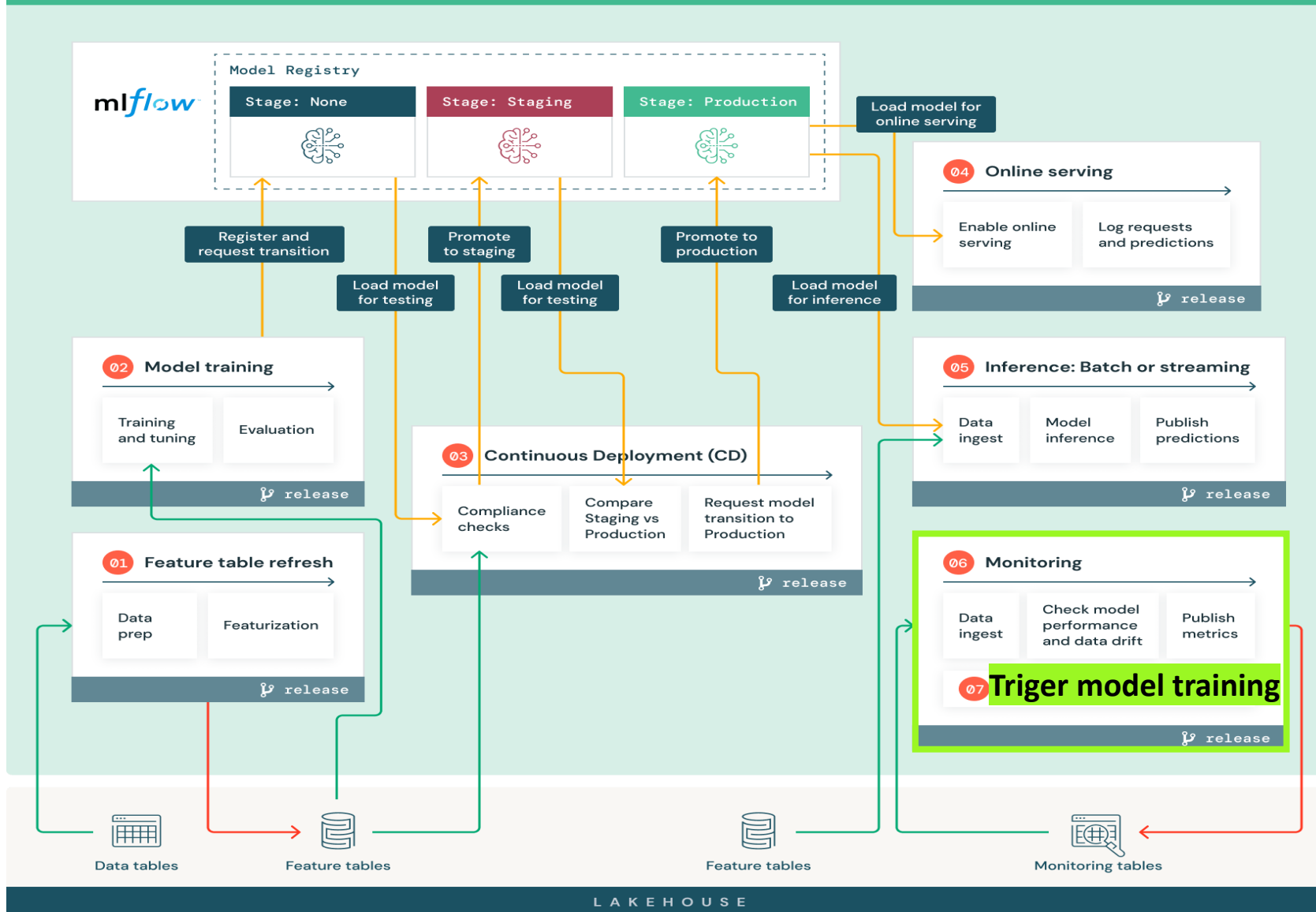
Source https://learn.microsoft.com/en-us/azure/databricks/_static/images/machine-learning/mlops-prod-diagram.png

Production environment



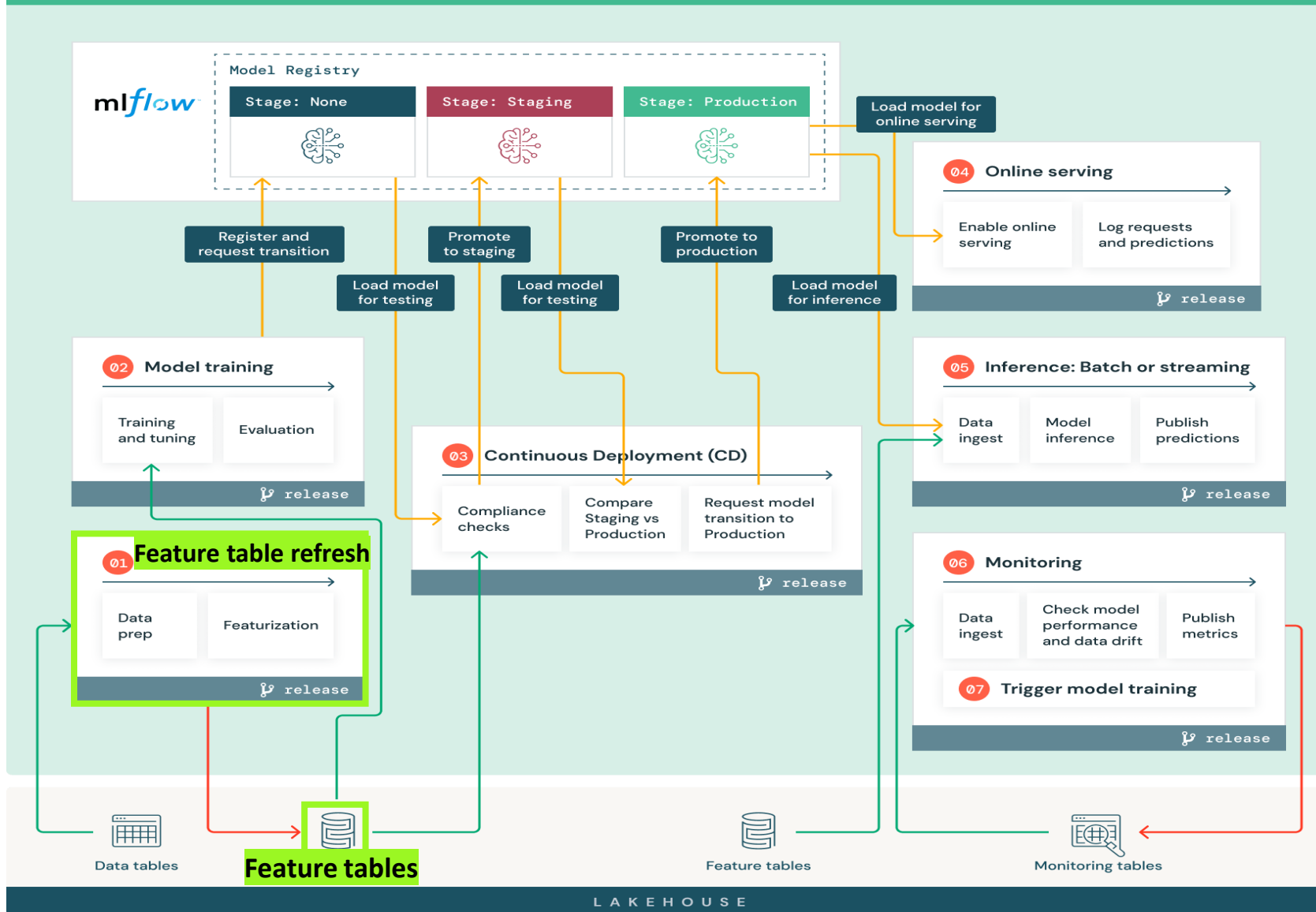
Source https://learn.microsoft.com/en-us/azure/databricks/_static/images/machine-learning/mlops-prod-diagram.png

Production environment



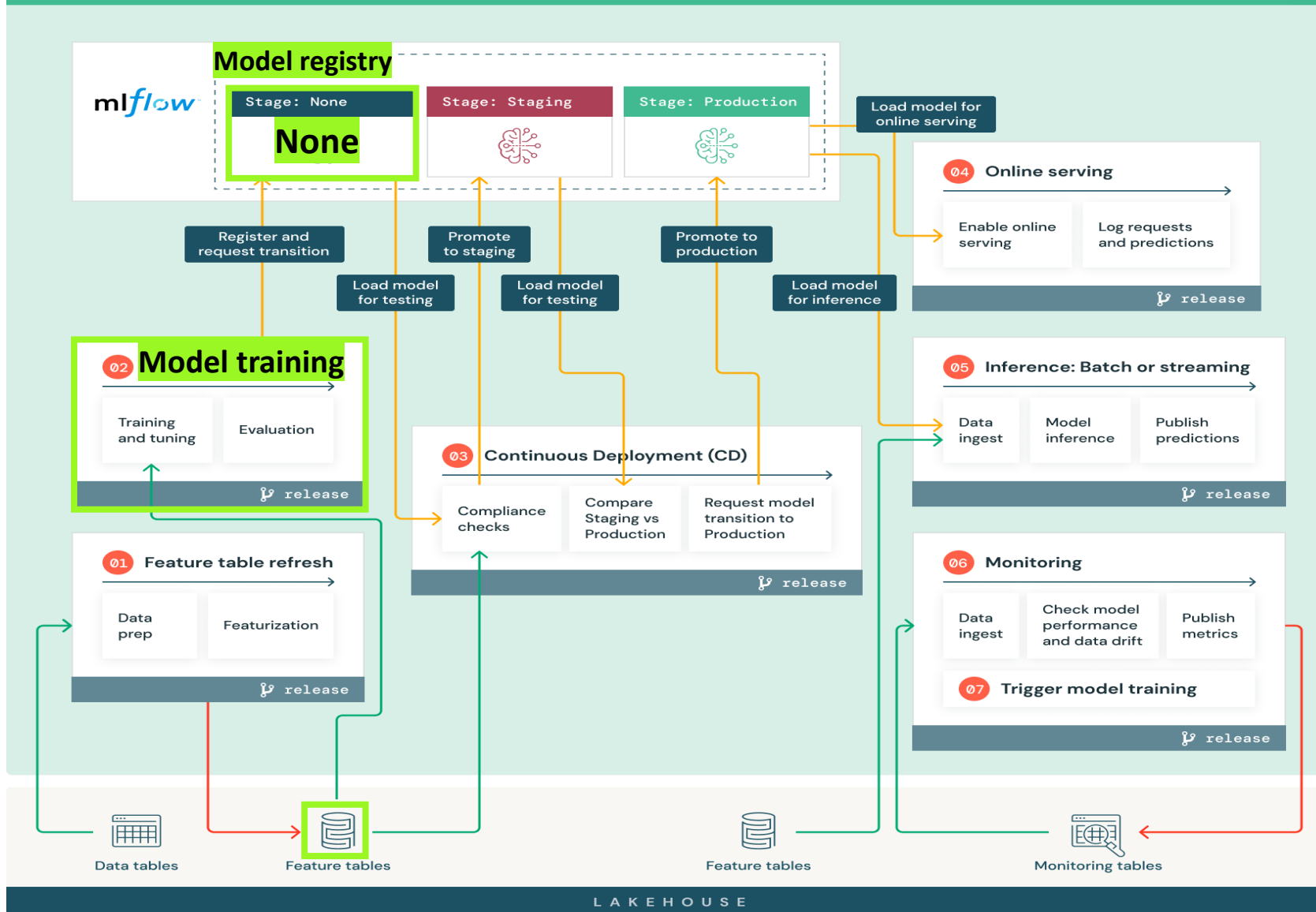
Source https://learn.microsoft.com/en-us/azure/databricks/_static/images/machine-learning/mlops-prod-diagram.png

Production environment



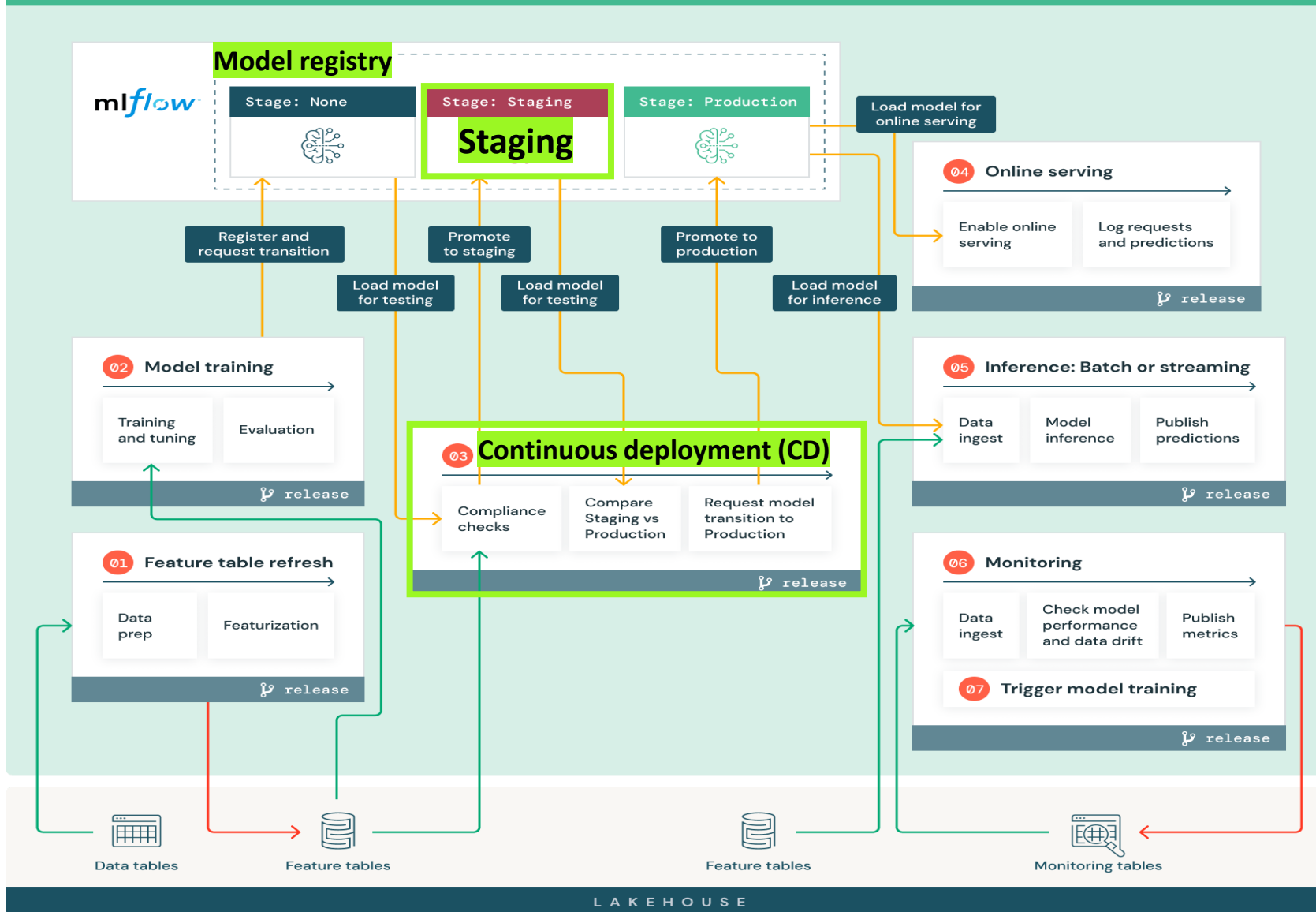
Source https://learn.microsoft.com/en-us/azure/databricks/_static/images/machine-learning/mlops-prod-diagram.png

Production environment



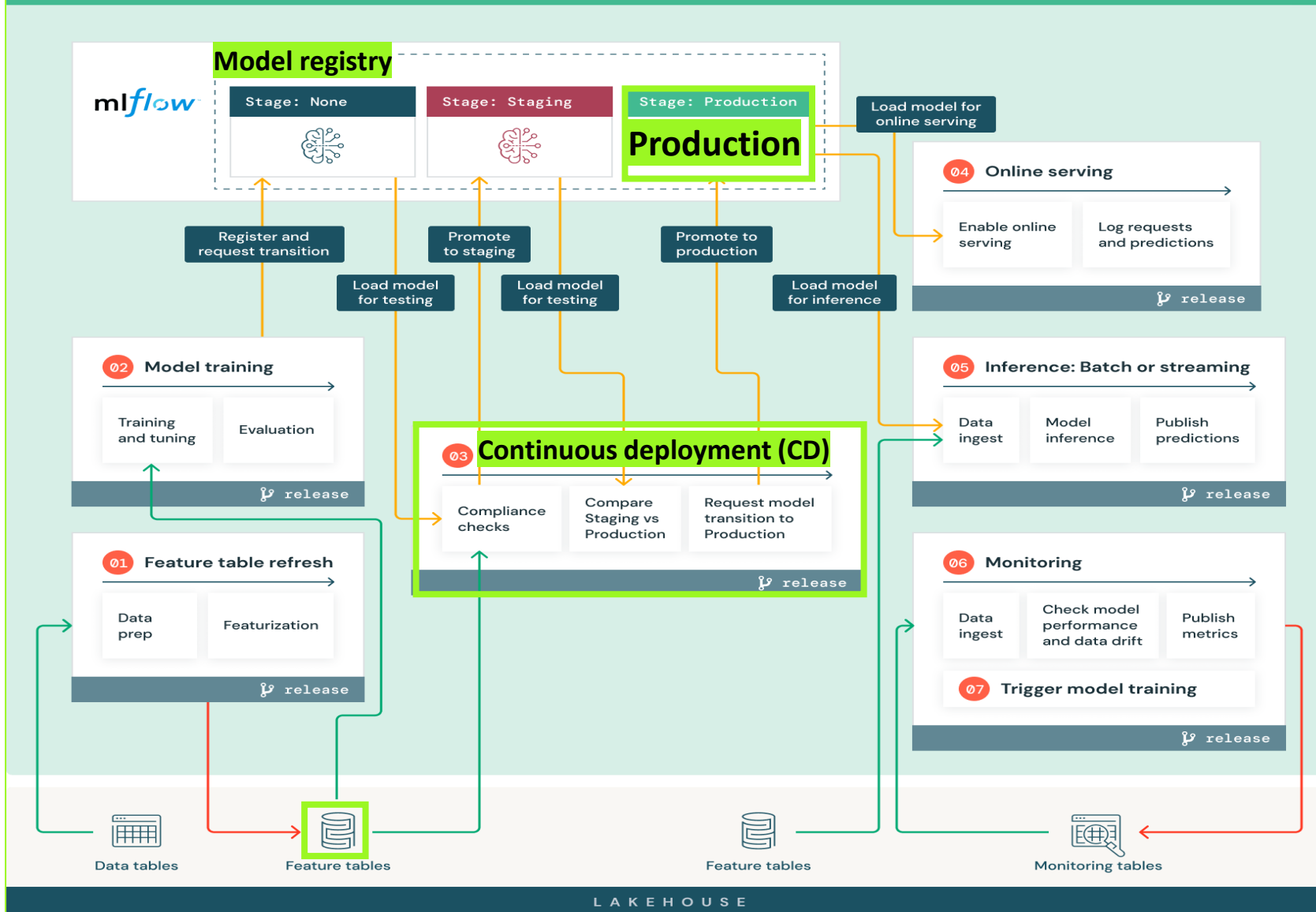
Source https://learn.microsoft.com/en-us/azure/databricks/_static/images/machine-learning/mlops-prod-diagram.png

Production environment



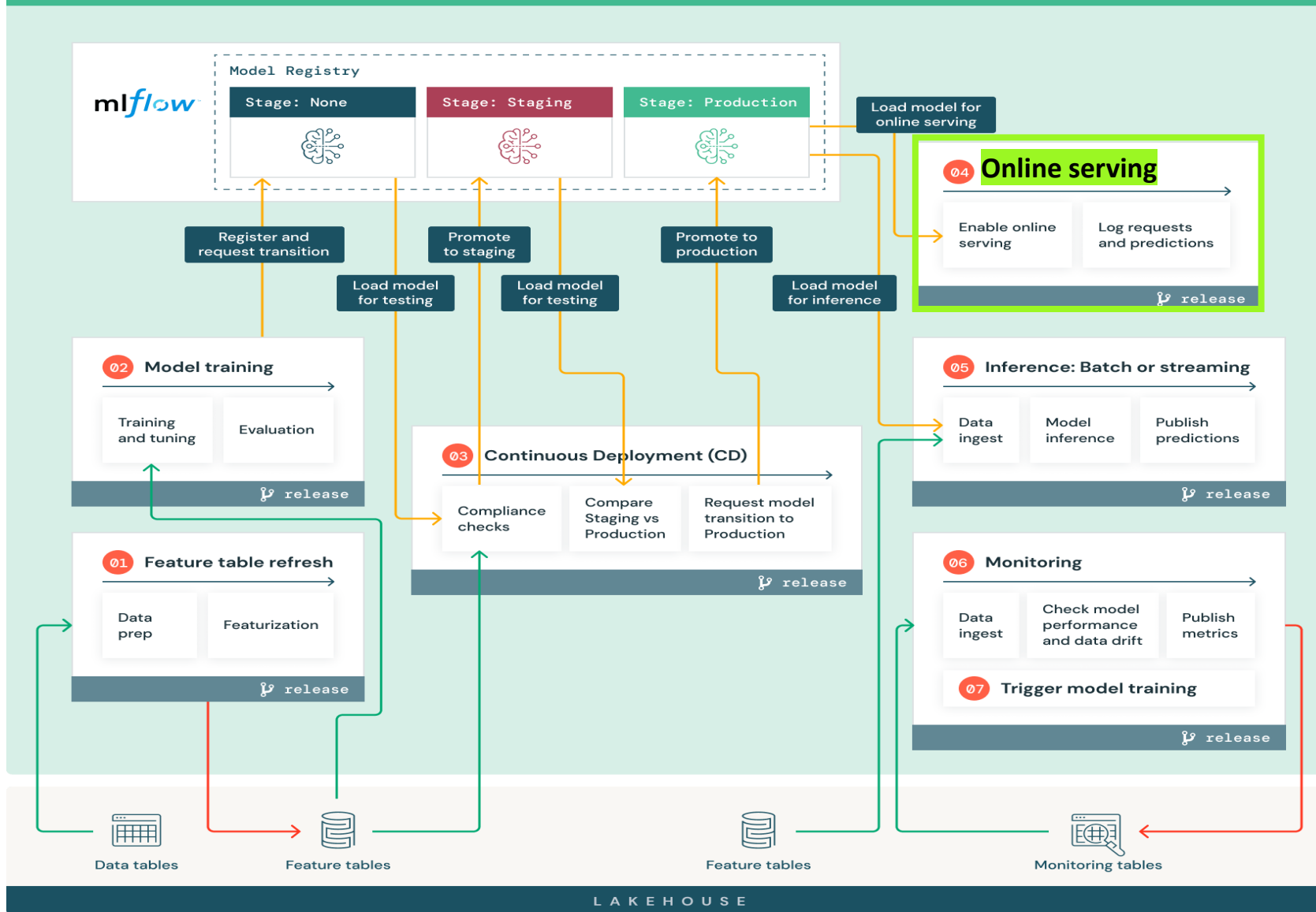
Source https://learn.microsoft.com/en-us/azure/databricks/_static/images/machine-learning/mlops-prod-diagram.png

Production environment



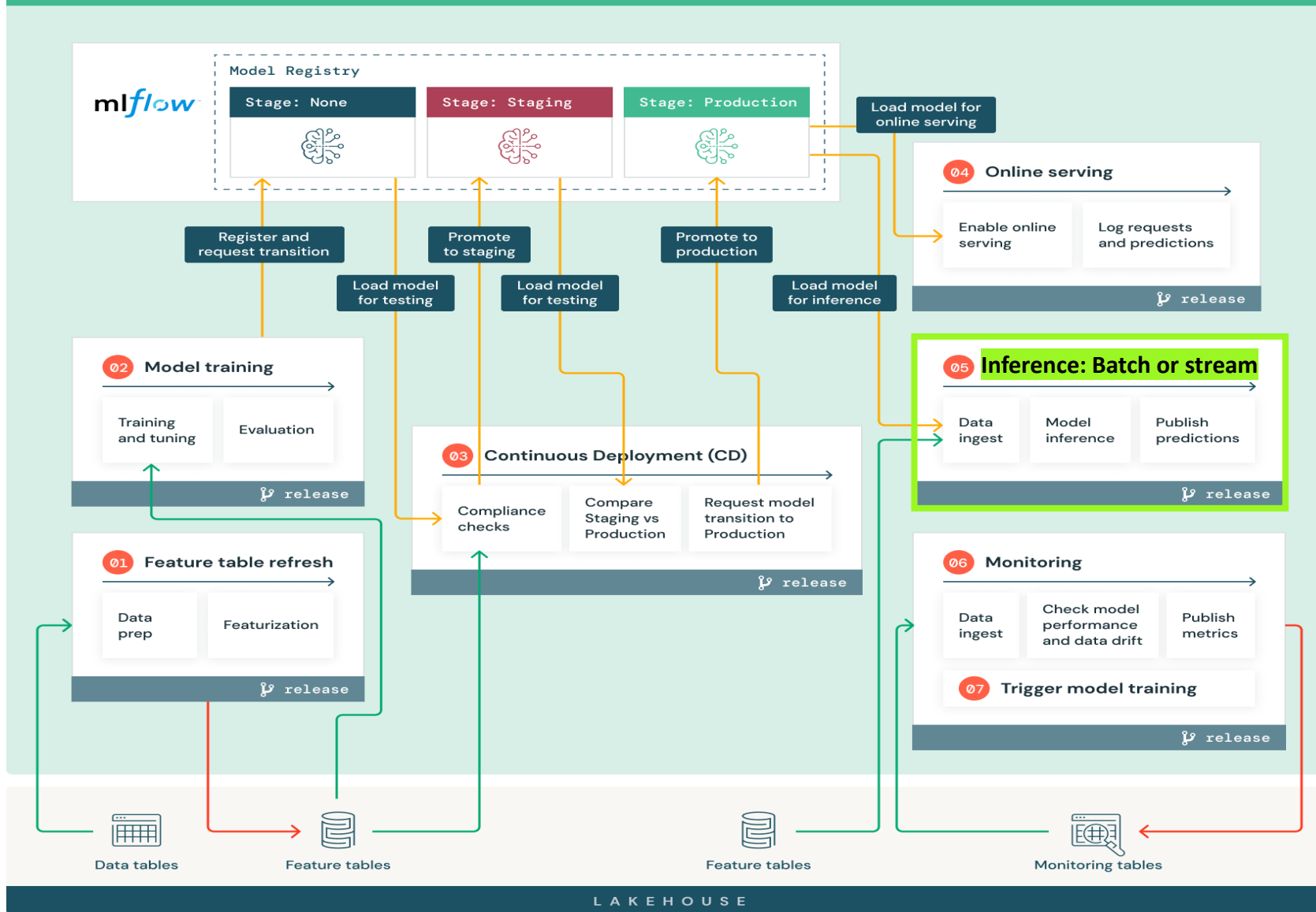
Source https://learn.microsoft.com/en-us/azure/databricks/_static/images/machine-learning/mlops-prod-diagram.png

Production environment



Source https://learn.microsoft.com/en-us/azure/databricks/_static/images/machine-learning/mlops-prod-diagram.png

Production environment



Source https://learn.microsoft.com/en-us/azure/databricks/_static/images/machine-learning/mlops-prod-diagram.png

Gouvernance et MLops



Databricks for Data Science and Machine Learning



Collaborative Multi-Language Notebooks

← Full ML Lifecycle →



Model Training
and Tuning



Model Tracking
and Registry



Model Serving
and
Monitoring



Automation
and
Governance



Open Multi-Cloud Data Lakehouse and Feature Store

Full ML Lifecycle: From Data Ingest to Model Deployment

Data prep designed for

ML



Text



Images



Video / Audio



Tabular

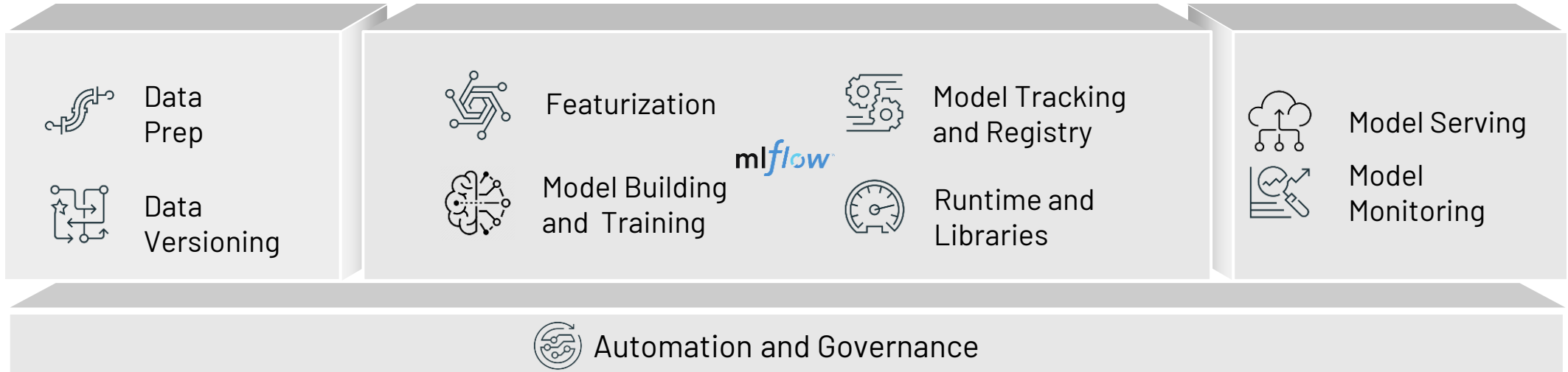
Out-of-the-box environment for all ML frameworks



mlflow

Shop

Deploy anywhere at any scale

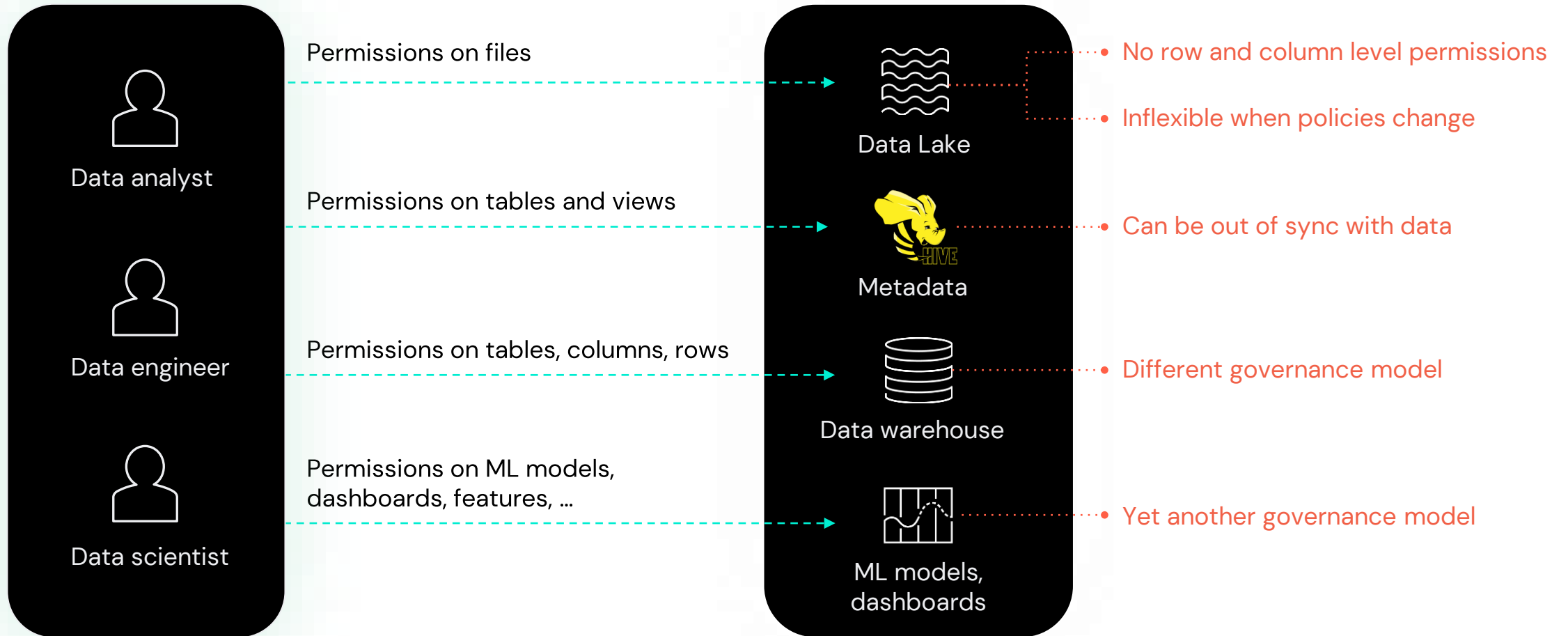


DataOps

ModelOps

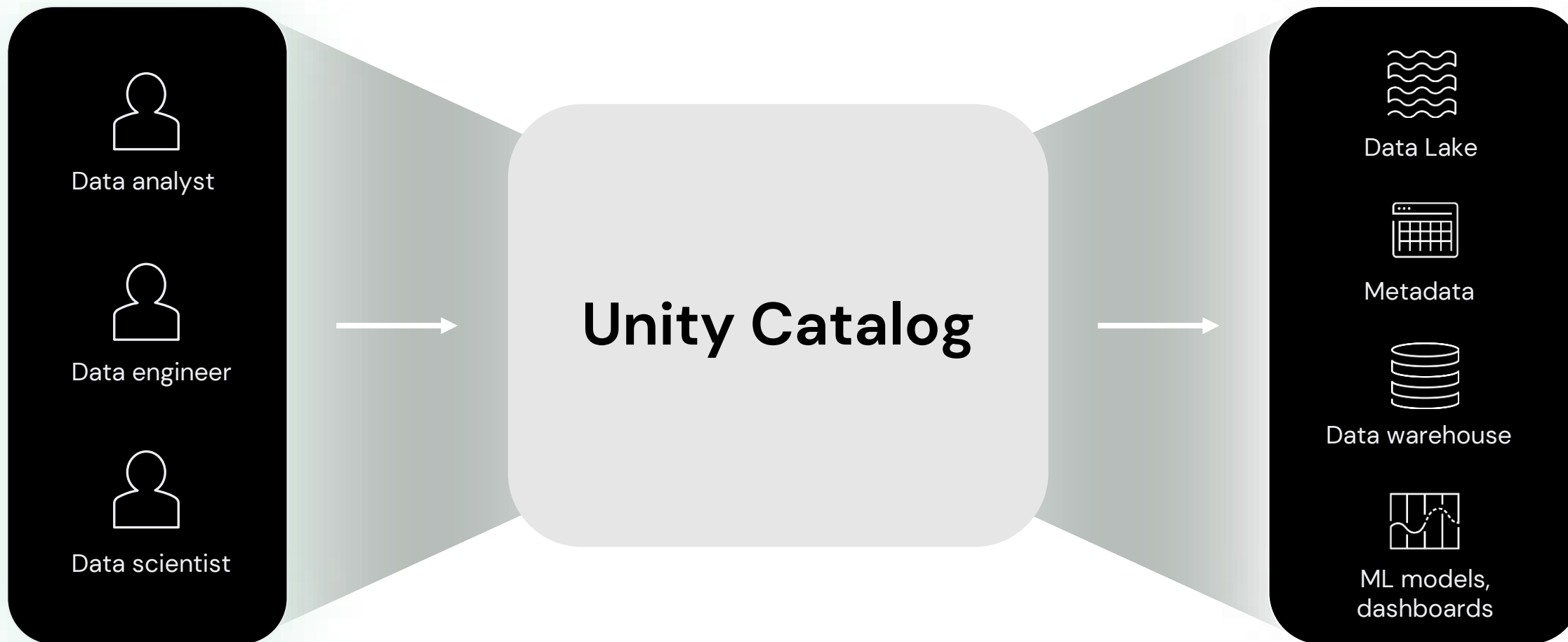
DevOps

Governance for data, analytics and AI is complex



Databricks Unity Catalog

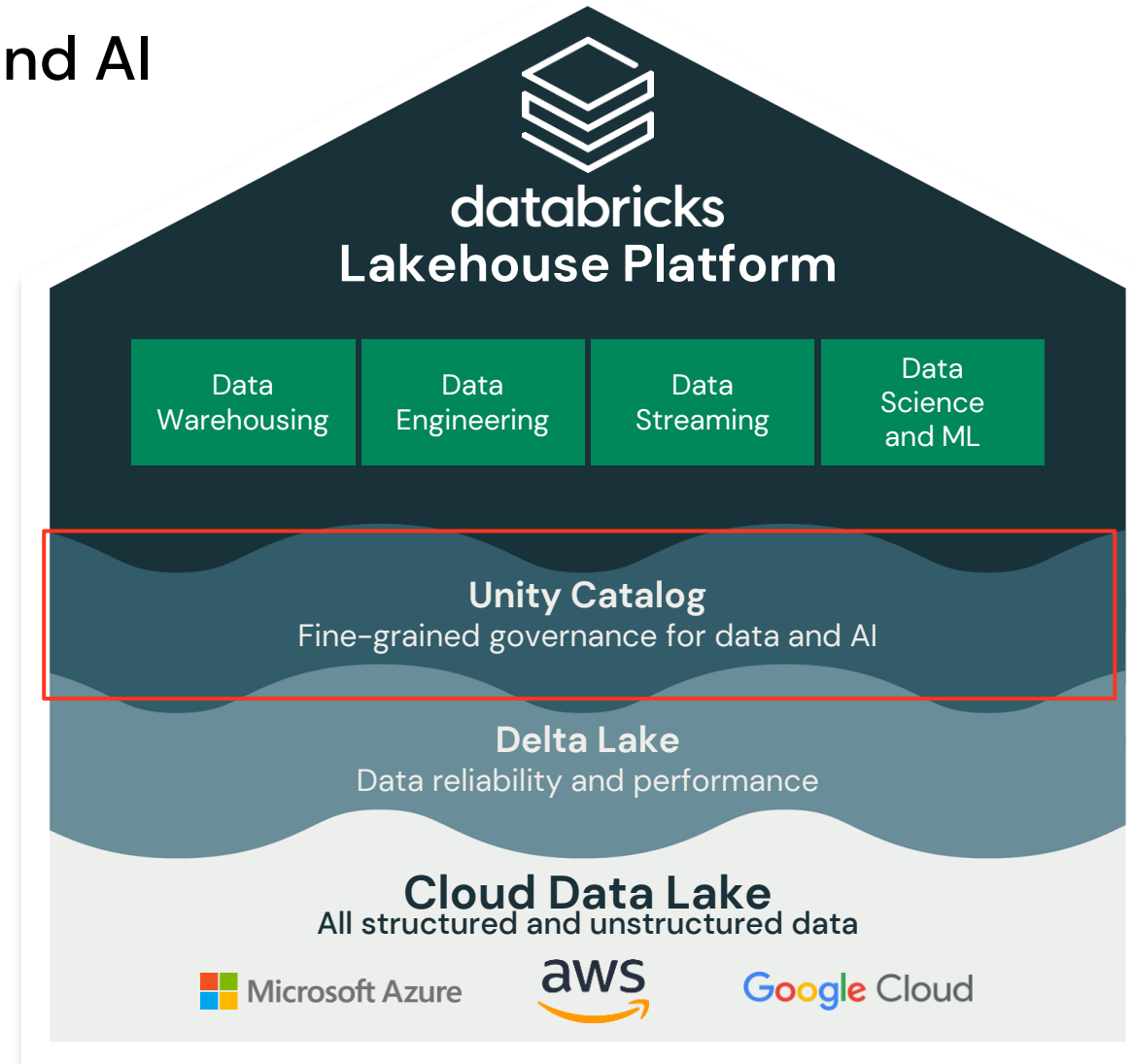
Unified governance for data, analytics and AI



Databricks Unity Catalog

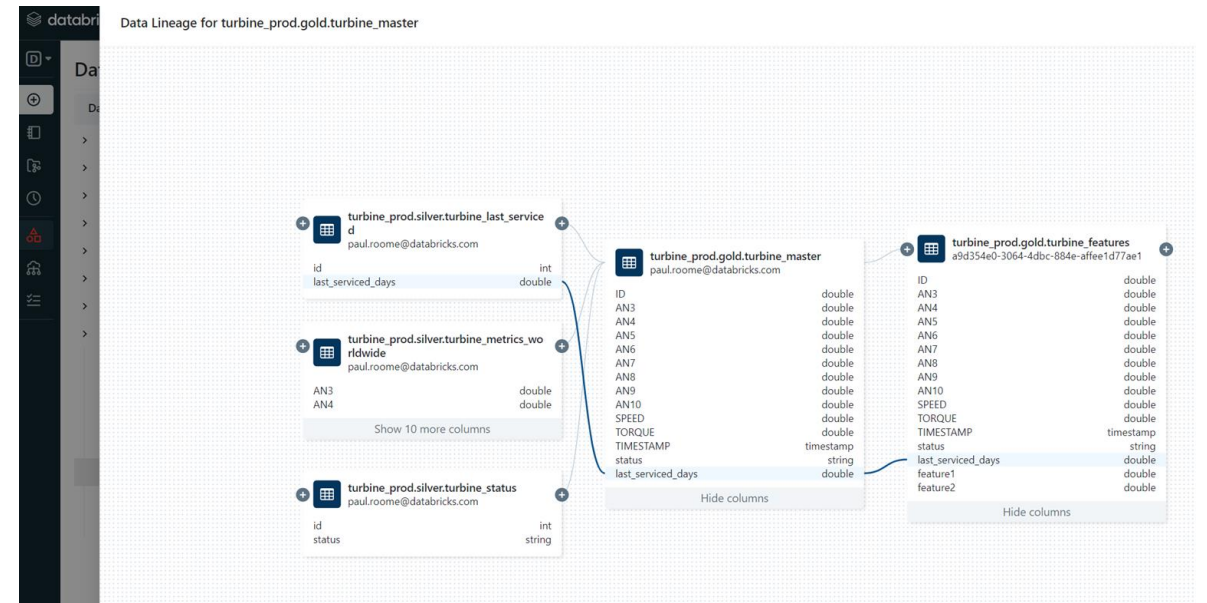
Unified governance for data, analytics and AI

- Map, secure and audit data across clouds
- Open data sharing and collaboration
- Seamless data observability and quality
- Improved productivity and discoverability
- Optimized performance and reduced cost



Seamless data observability and quality

- End-to-end visibility into how data flows in your organization with automated data lineage down to column level
- View lineage across tables, columns, notebooks, workflows, dashboards, models, feature stores
- Captured in real time across all workloads in SQL, Python, Scala, and R



Improved productivity and discoverability

- Quickly find, understand, and reference data from across your data estate with a unified data browsing experience
- Secure by default—leverages common permission model from Unity Catalog
- Auto-generated data insights to understand your data

The screenshot displays the Data Explorer interface with a search overlay and a table insights panel.

Search Overlay:

- Search input: customer
- Catalog: [Dropdown]
- Database: [Dropdown]
- Results:
 - customer (gsethi-uc-demo-tpch.tpch) - Customer Table
 - customer (uc-demo-tpch.tpch) - Customer Table
 - customer (sample_datasets.tpch)
 - customer (ifi.tpch) - Customer Table

Main Interface:

- Left sidebar: Data Explorer navigation tree showing 'main' and 'lineage_test_3'.
- Center pane: Details for 'main.lineage_test_3.delta_target_11' (Delta table).
 - Owner: yash.oz+admin@databricks.com
 - Table Type: MANAGED
 - Table Id: 3d7e357d-0589-4381-9182-d7894bc3a795
 - Columns:

Column	Type	Comment
id	string	
sum	bigint	
count	bigint	
- Right pane: Table Insights for 'main.lineage_test_3.delta_target_11'.
 - Owner: yash.oz+admin@databricks.com
 - Created At: 7/23/2022, 11:13:37 PM
 - Updated At: 7/23/2022, 11:13:37 PM
 - Table Type: MANAGED
 - Metastore Id: 97ab11d1-1206-4044-9806-d2b42922b73e
 - Table Id: 3d7e357d-0589-4381-9182-d7894bc3a795
 - Frequent Users: yash.oz+admin@databricks.com
 - Frequent Queries:

```
select * from main.lineage_test_3.delta_target_11 LIMIT 1000
WITH temporaryTable (averageValue) as
(SELECT max(id)
FROM main.lineage_test_3.delta_source_1)
SELECT id
FROM main.lineage_test_3.delta_target_11, temporaryTable
WHERE id < temporaryTable.averageValue LIMIT 1000
select count from main.lineage_test_3.delta_target_11 LIMIT 1000
select id from main.lineage_test_3.delta_target_11 LIMIT 1000
```

Groupe utilisateurs Montréal / Québec

Rejoignez le groupe de discussion des utilisateurs Databricks de Montréal et Québec.

Le groupe partagera les nouvelles sur Databricks ainsi que les prochaines réunions du groupe utilisateur



Merci de votre attention

Des questions?

407 Rue McGill bureau 500, Montréal

2828 Bd Laurier Suite 700, Québec



databricks



+ Casse-tête ML : Vers le chemin du succès

Phases	Collecte des données	Ingénierie de données	Prétraitement & feature engineering	Entraînement	Surveillance (monitoring)	Intégration de l'application
Tâches / Processus	<ul style="list-style-type: none"> - Identification des sources de données - Ingestion des données 	<ul style="list-style-type: none"> - Modélisation lac de données - Implémentation de pipeline périodique ou continue - Automatisation DataOps 	<ul style="list-style-type: none"> - Imputation des données manquantes - Extraction des traits caractéristiques - Nettoyage des données 	<ul style="list-style-type: none"> - Choix des métriques - Entraînement des modèles - Optimisation des hyperparamètres - Pipeline d'apprentissage - Évaluation & sélection des modèles 	<ul style="list-style-type: none"> - Pipeline de déploiement - API Rest 	<ul style="list-style-type: none"> - Visualisation dashboards de prédiction - Application IA
Rôles	<ul style="list-style-type: none"> - Spécialiste métiers - Architecte de solutions - Ingénieur de données 	<ul style="list-style-type: none"> - Architecte de données - Ingénieur de données - DataOps Ingénieur 	<ul style="list-style-type: none"> - Ingénieur de données - Scientifique de données 	<ul style="list-style-type: none"> - Scientifique de données 	<ul style="list-style-type: none"> - Ingénieur ML - Développeur de rapports 	<ul style="list-style-type: none"> - Spécialiste métiers / gestionnaires - Analyste data - Ingénieur logiciel
Sécurité						