Université Lumière Lyon 2

Institut de la communication

Université Lumière Lyon 2

Cycle de vie et industrialisation d'un projet de machine learning

05/03/2024

Présentateurs :

- Dounya Bourhani
- Paule Naomi Kaldjob
- Albane Nicoullaud
- Fiona Steiner

Professeur : Ricco Rakotomalala
M2 SISE 2023-2024

Plan

Partie 1 : Cycle de vie d'un projet de machine learning

Partie 2 : Plateformes de Gestion de Projets

- Technologies et outils
- MLFlow
- Kubeflow
- Comet.ml
- Airflow
- Comparaison des technologies et outils

Partie 3 : Présentation du TD

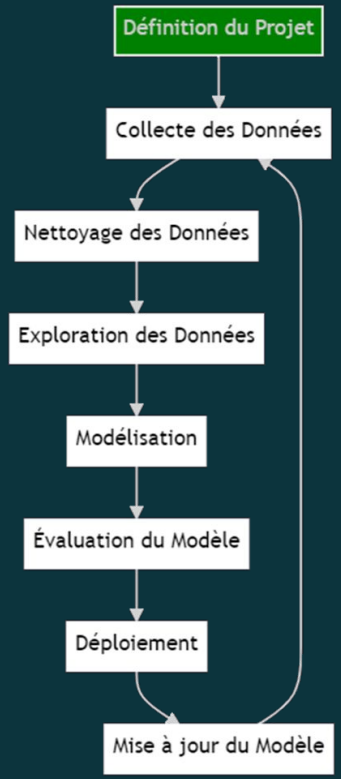
Plan

Partie 1 : Cycle de vie d’un projet de machine learning

Partie 2 : Plateformes de Gestion de Projets

Partie 3 : Présentation du TD

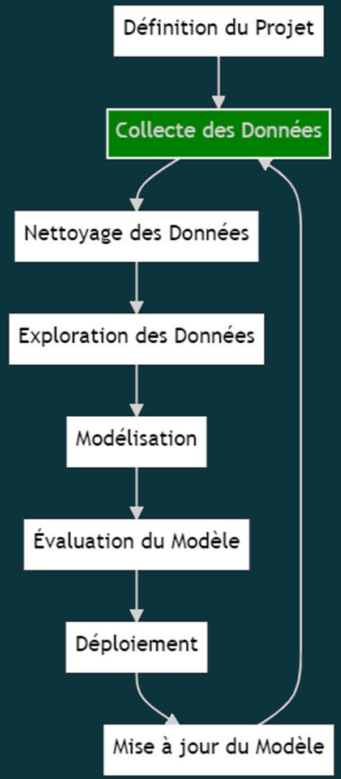
Partie 1 : Introduction au Machine Learning et à son Cycle de Vie



Besoins métiers

- Objectifs du projet
- Contexte du projet
- Contraintes
- Exigences

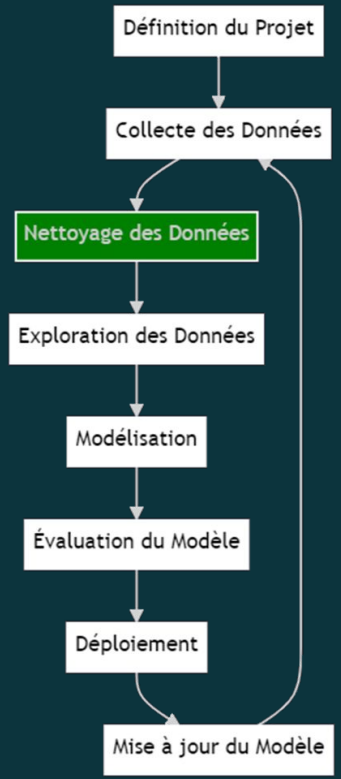
Partie 1 : Introduction au Machine Learning et à son Cycle de Vie



Alimenter modèle de machine learning

- Acquisition des données de diverses sources

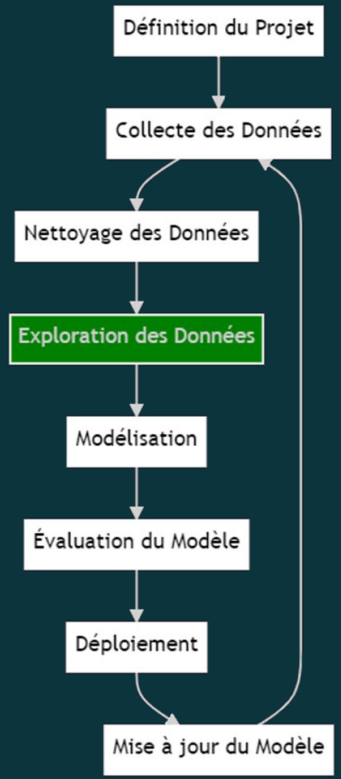
Partie 1 : Introduction au Machine Learning et à son Cycle de Vie



Garantir la qualité des données

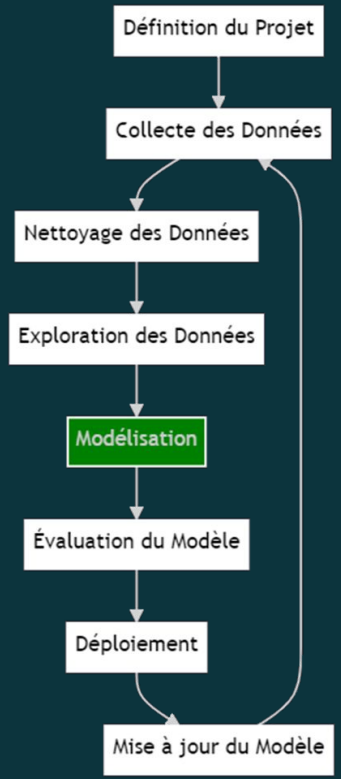
- Traitements des données :
    - Valeurs manquantes
    - Valeurs aberrantes
    - Anomalies

Partie 1 : Introduction au Machine Learning et à son Cycle de Vie



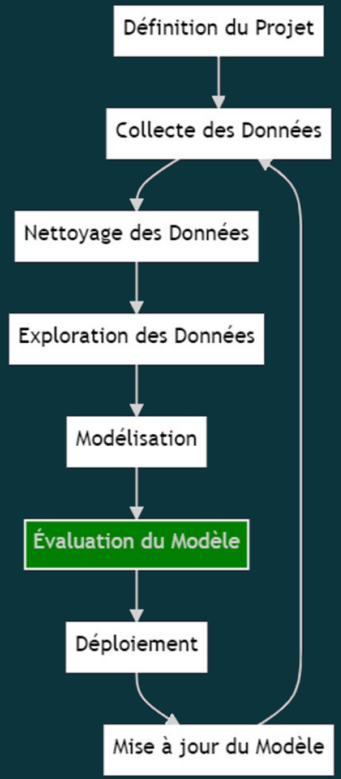
- Analyse préliminaire des données :
  - Statistiques descriptives
  - Segmentation des données

Partie 1 : Introduction au Machine Learning et à son Cycle de Vie



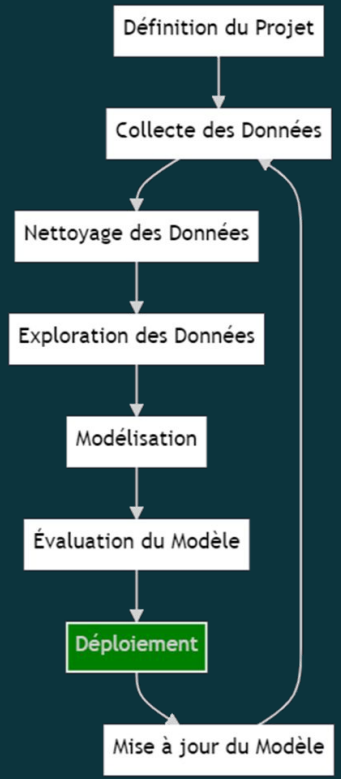
- Conception et entraînement des modèles
- Algorithmes
- Paramètres

Partie 1 : Introduction au Machine Learning et à son Cycle de Vie



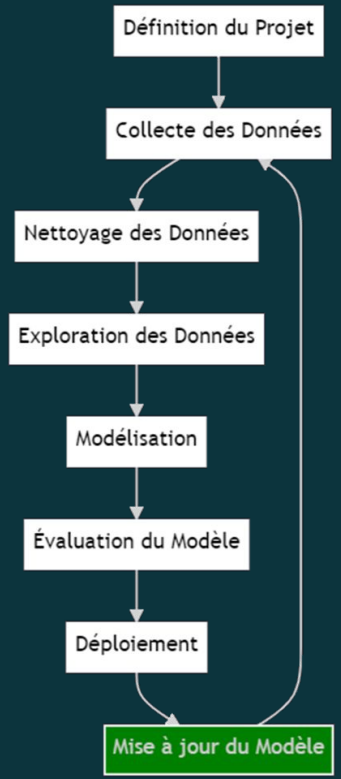
- Test des modèles entraînés
- Métriques appropriées

Partie 1 : Introduction au Machine Learning et à son Cycle de Vie



- Mise en production
- Prédictions

Partie 1 : Introduction au Machine Learning et à son Cycle de Vie



- Surveillance continue
- Mise à jour du modèle selon évolution des données

Plan

Partie 1 : Cycle de vie d’un projet de machine learning

Partie 2 : Plateformes de Gestion de Projets

- **Technologies et outils**

Partie 3 : Présentation du TD

Partie 2 : Plateformes de Gestion de Projets - Technologies et outils

- **Conteneurs (Docker, Kubernetes) :**

  Facilite le déploiement et la gestion d'infrastructure de machine learning

- **Plateformes de Machine Learning (MLflow, Kubeflow, Comet) :**

  Gestion du cycle de vie complet des projets de machine learning

- **Orchestrateurs de Workflow (Airflow, Prefect, Dagster) :**

  Automatisation, planification et gestion des flux de travail

Plan

Partie 1 : Cycle de vie d’un projet de machine learning

Partie 2 : Plateformes de Gestion de Projets

- Technologies et outils
- **MLFlow**

Partie 3 : Présentation du TD

Partie 2 : Plateformes de Gestion de Projets - MLFlow

- Plateforme open source de gestion du cycle de vie des modèles
- Développée par Databricks

- Simplifier le processus de machine learning :
  Expérimentation → Mise en production

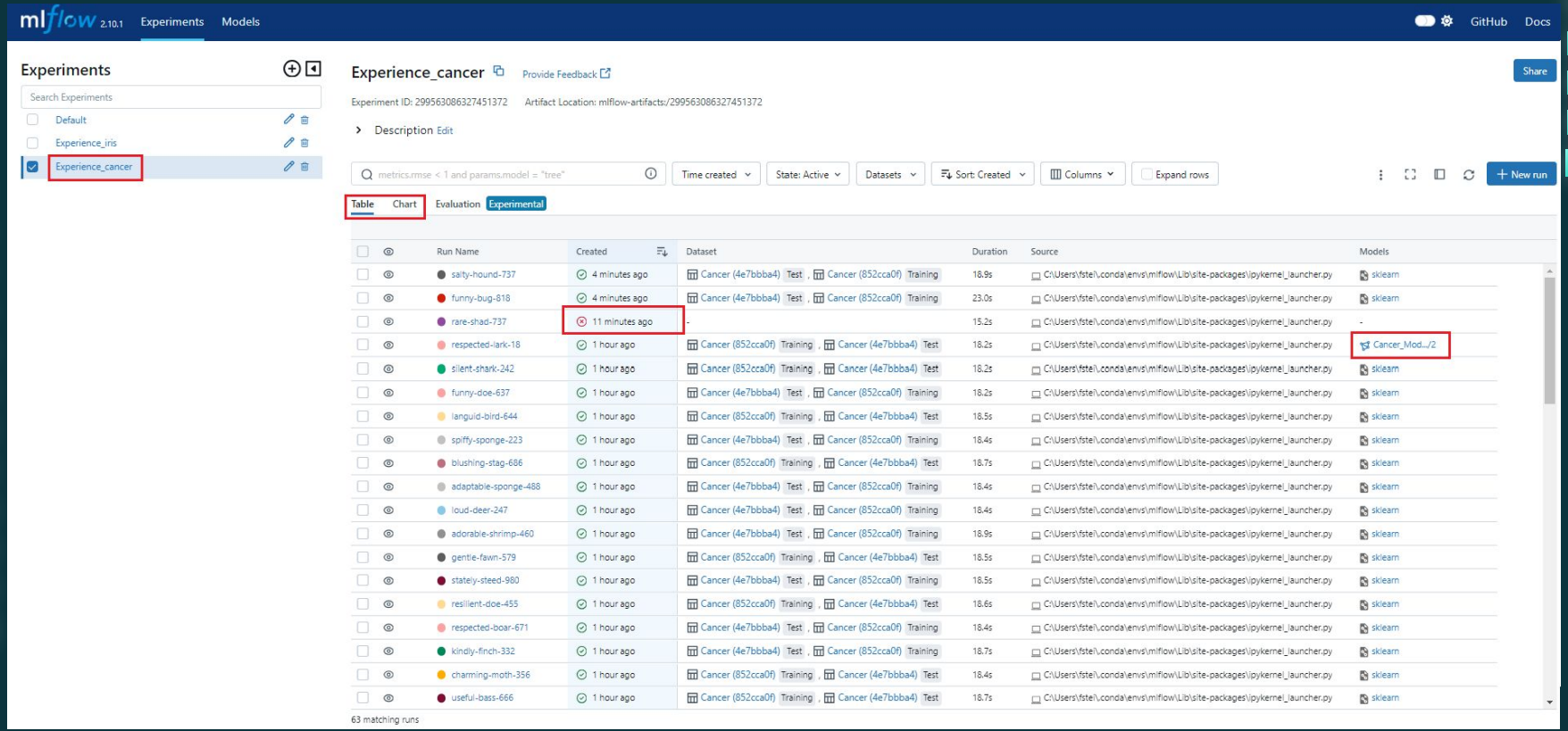
Partie 2 : Plateformes de Gestion de Projets - MLFlow

Fonctionnalités principales :

- Suivi des expériences
- Gestion des projets
- Gestion des modèles
- Registre de modèles

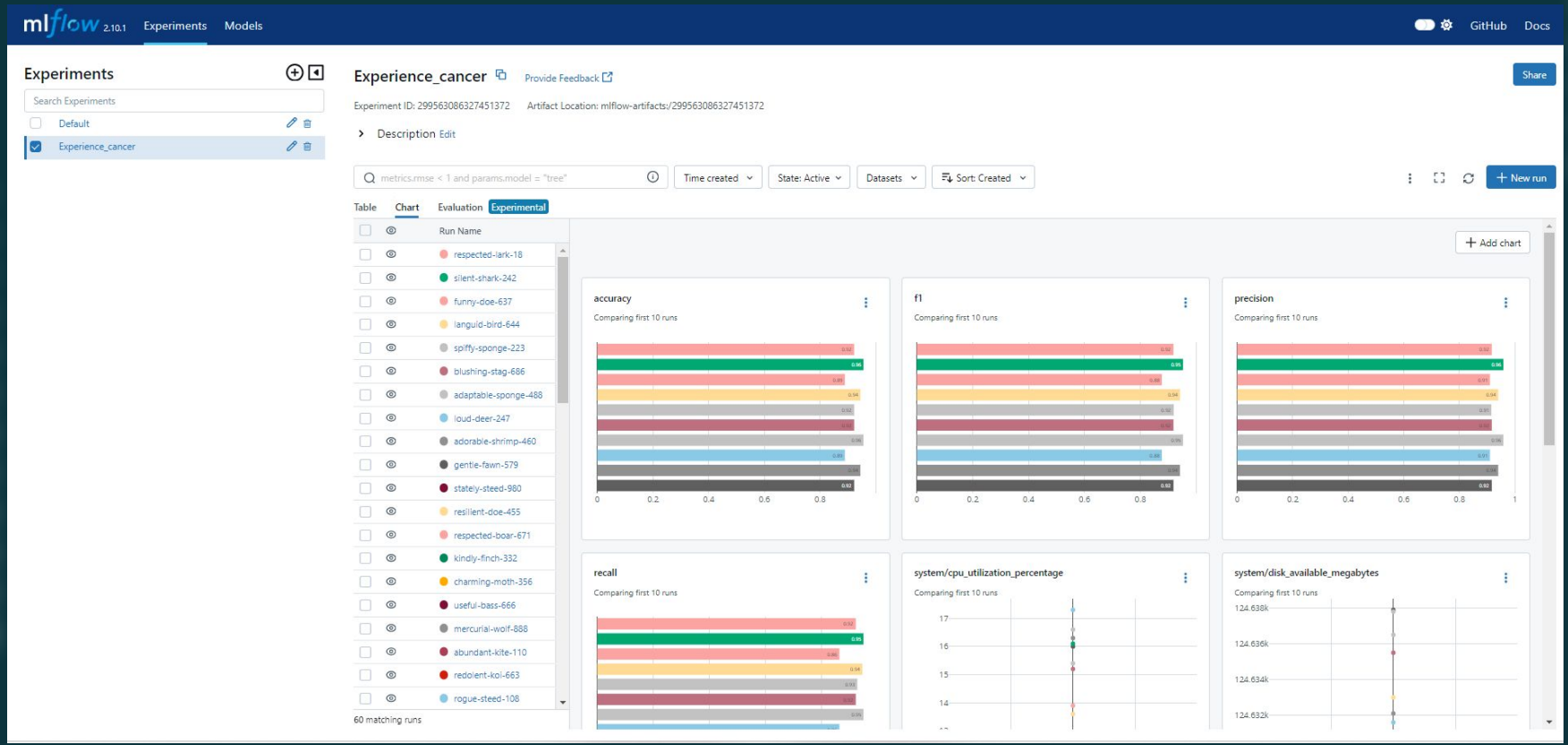
Partie 2 : Plateformes de Gestion de Projets - MLFlow

Interface - Expériences



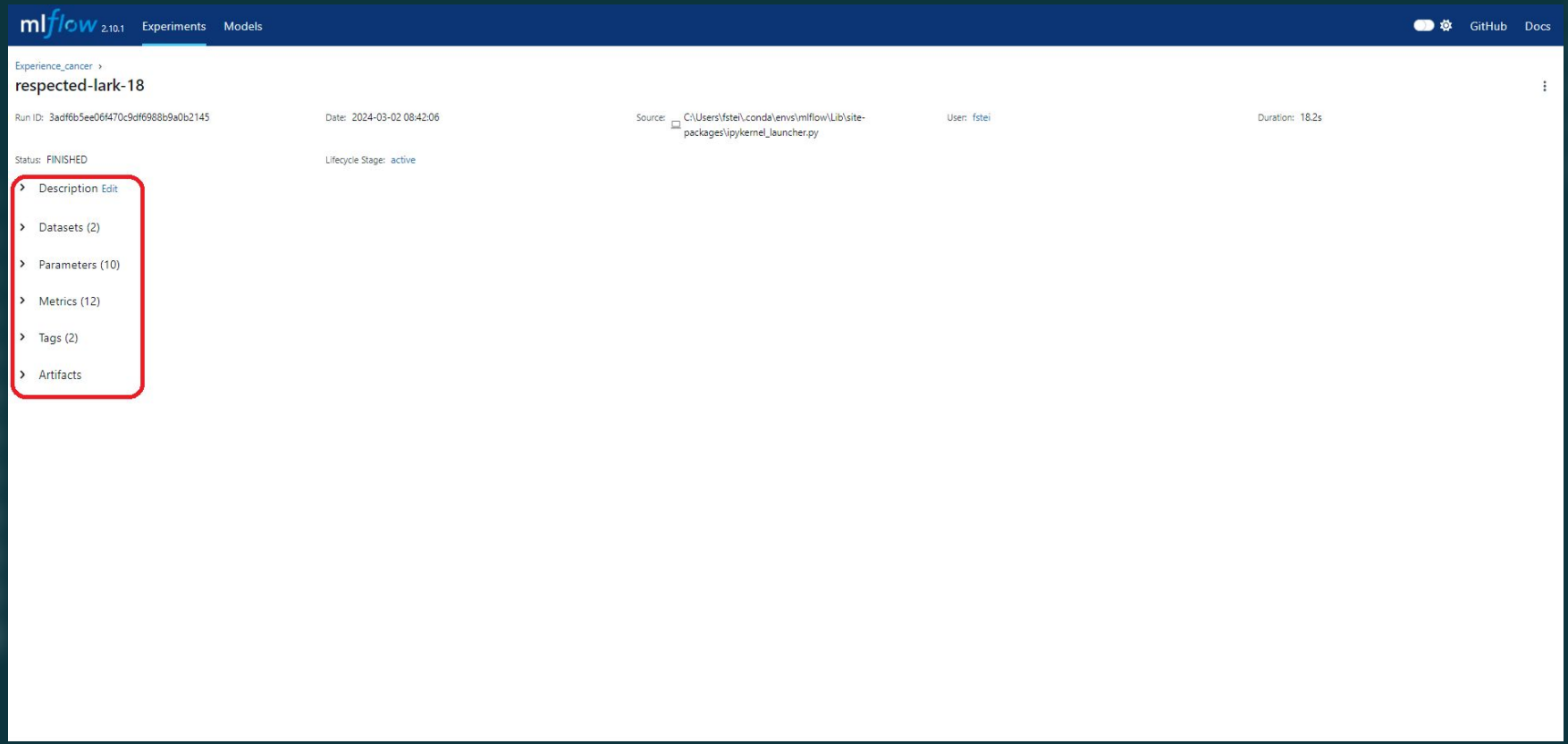
Partie 2 : Plateformes de Gestion de Projets - MLFlow

Interface - Expériences



Partie 2 : Plateformes de Gestion de Projets - MLFlow

Interface - Expériences

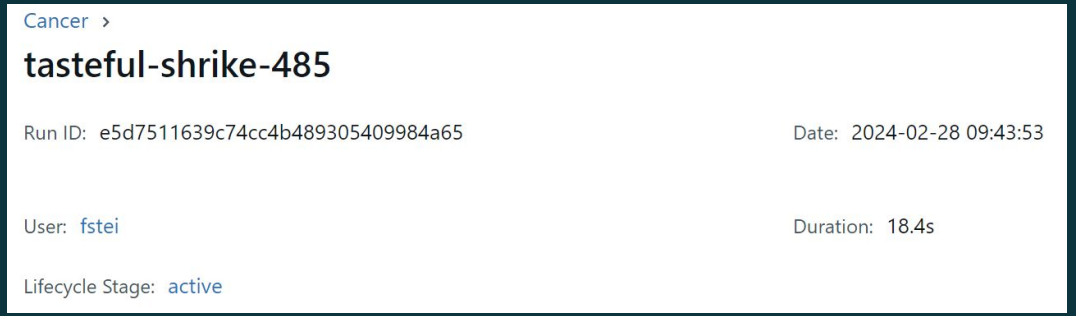


Partie 2 : Plateformes de Gestion de Projets - MLFlow

Interface - Expériences

Pour chaque expérience :

- Run ID
- Date
- User
- Duration
- Description
- Datasets
- Parameters
- Metrics
- Tags
- Artifacts

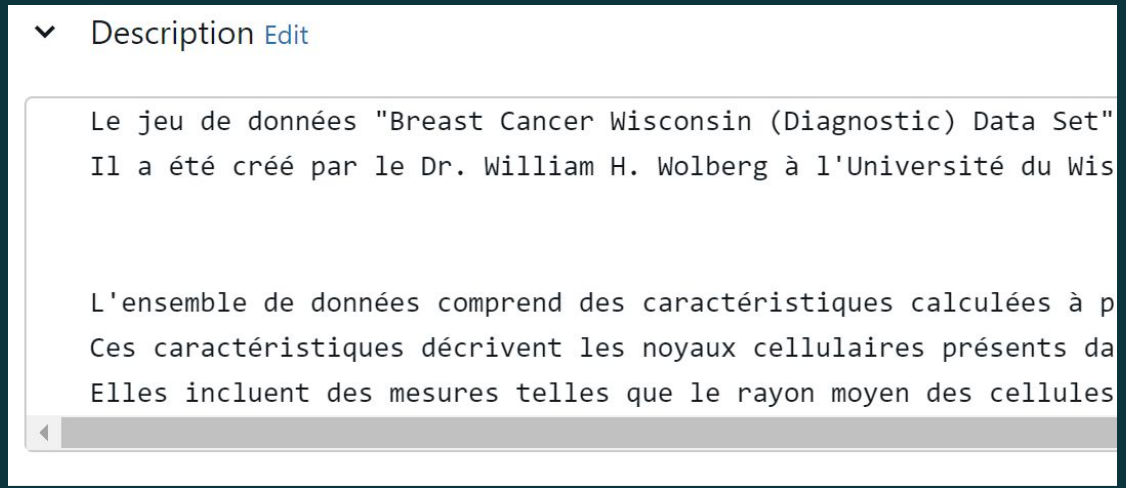


Partie 2 : Plateformes de Gestion de Projets - MLFlow

Interface - Expériences

Pour chaque expérience :

- Run ID
- Date
- User
- Duration
- Description
- Datasets
- Parameters
- Metrics
- Tags
- Artifacts

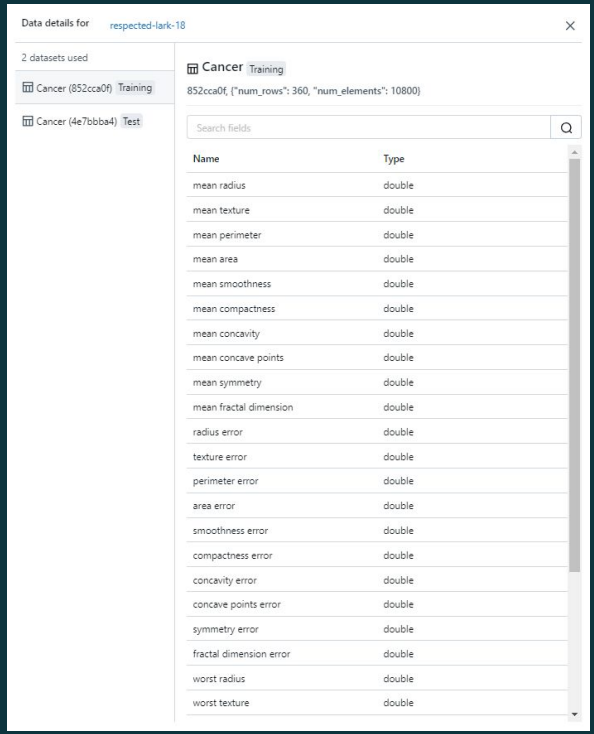


Partie 2 : Plateformes de Gestion de Projets - MLFlow

Interface - Expériences

Pour chaque expérience :

- Run ID
- Date
- User
- Duration
- Description
- **Datasets**
- Parameters
- Metrics
- Tags
- Artifacts

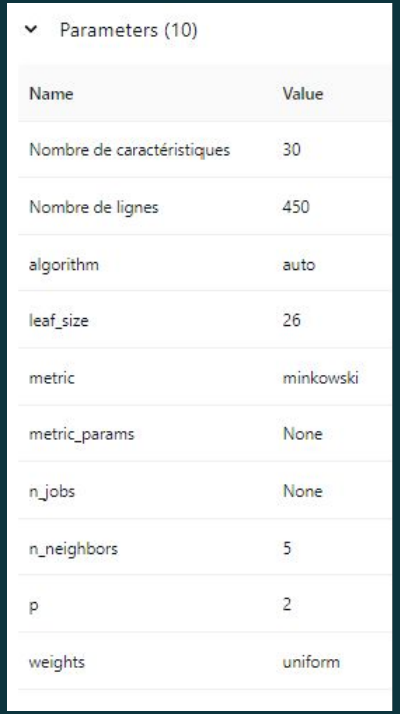


Partie 2 : Plateformes de Gestion de Projets - MLFlow

Interface - Expériences

Pour chaque expérience :

- Run ID
- Date
- User
- Duration
- Description
- Datasets
- Parameters
- Metrics
- Tags
- Artifacts



Parameters (10)

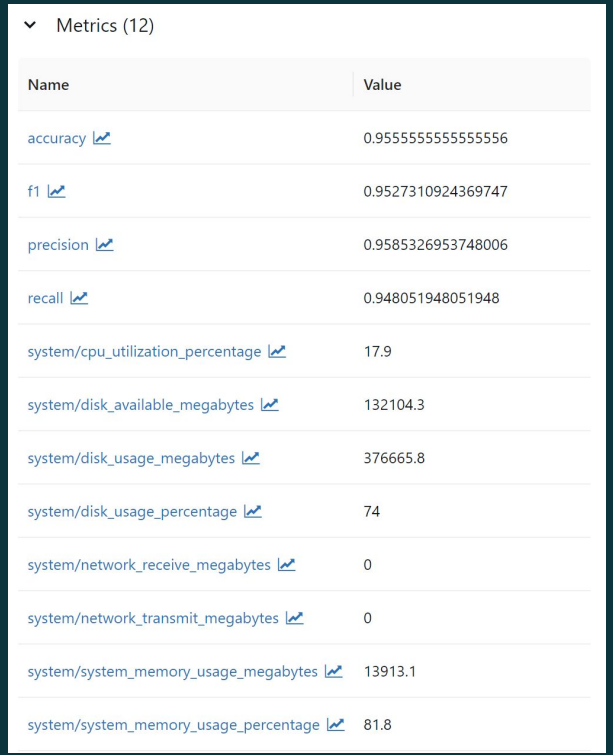
Name	Value
Nombre de caractéristiques	30
Nombre de lignes	450
algorithm	auto
leaf_size	26
metric	minkowski
metric_params	None
n_jobs	None
n_neighbors	5
p	2
weights	uniform

Partie 2 : Plateformes de Gestion de Projets - MLFlow

Interface - Expériences

Pour chaque expérience :

- Run ID
- Date
- User
- Duration
- Description
- Datasets
- Parameters
- Metrics
- Tags
- Artifacts



Metrics (12)

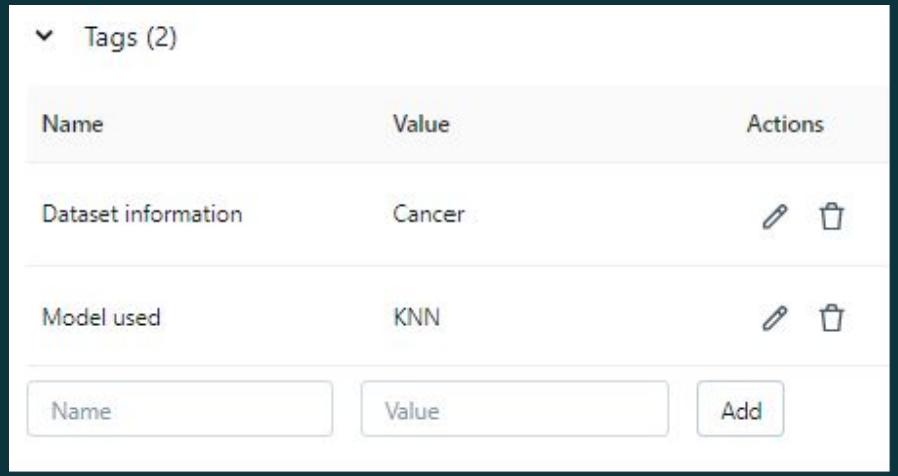
Name	Value
accuracy	0.9555555555555556
f1	0.9527310924369747
precision	0.9585326953748006
recall	0.948051948051948
system/cpu_utilization_percentage	17.9
system/disk_available_megabytes	132104.3
system/disk_usage_megabytes	376665.8
system/disk_usage_percentage	74
system/network_receive_megabytes	0
system/network_transmit_megabytes	0
system/system_memory_usage_megabytes	13913.1
system/system_memory_usage_percentage	81.8

Partie 2 : Plateformes de Gestion de Projets - MLFlow

Interface - Expériences

Pour chaque expérience :

- Run ID
- Date
- User
- Duration
- Description
- Datasets
- Parameters
- Metrics
- Tags
- Artifacts

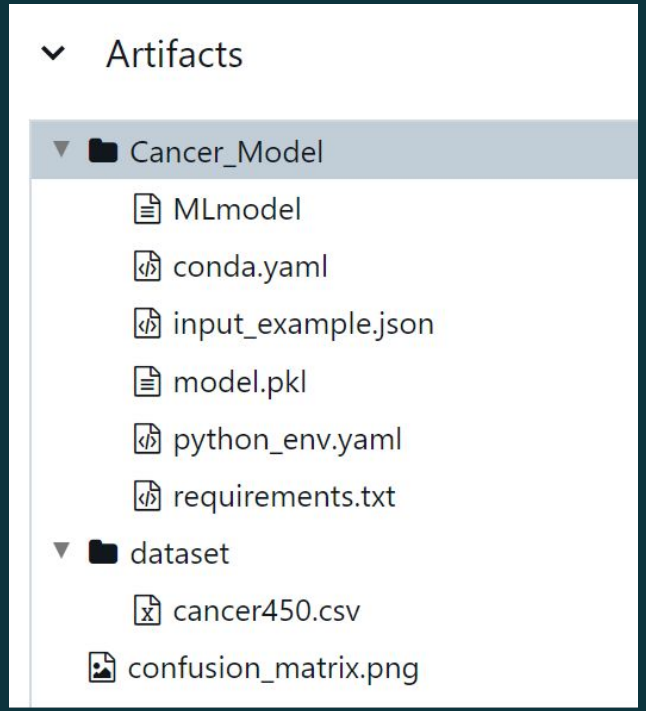


Partie 2 : Plateformes de Gestion de Projets - MLFlow

Interface - Expériences

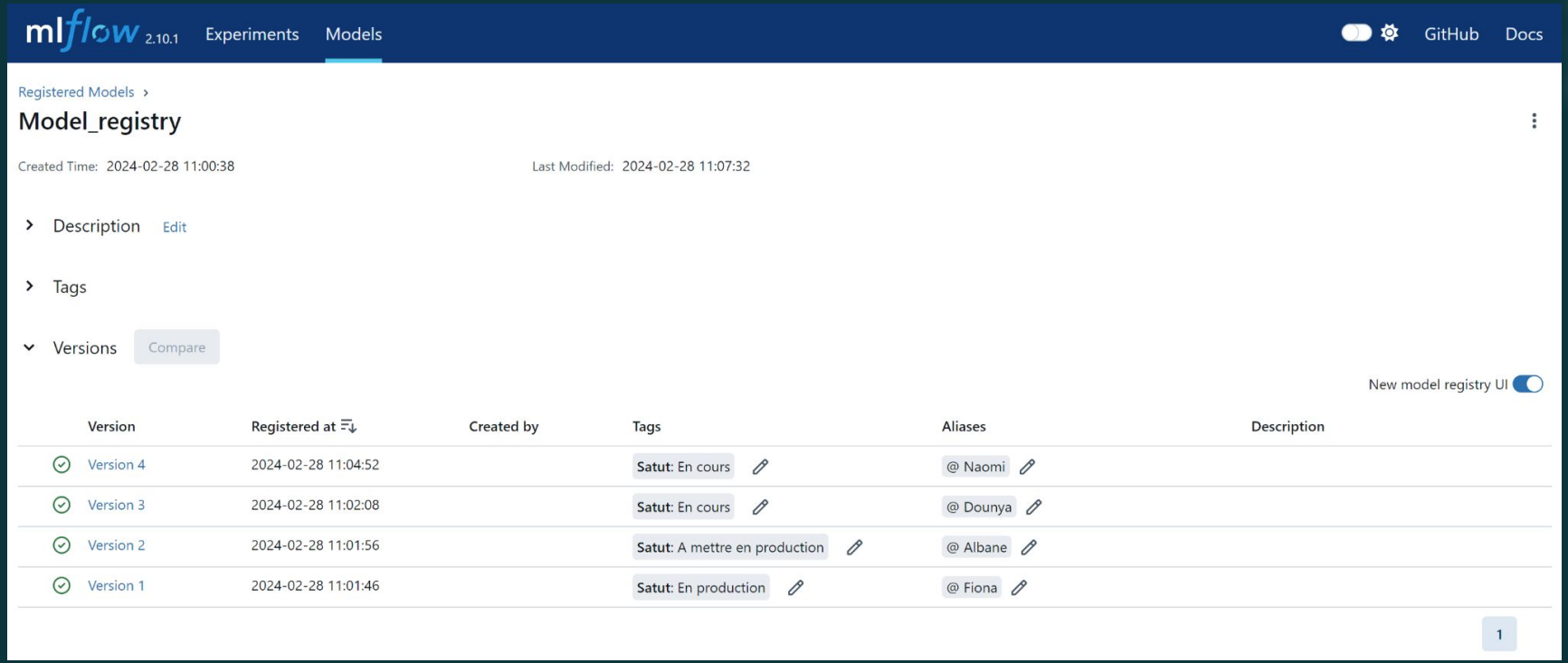
Pour chaque expérience :

- Run ID
- Date
- User
- Duration
- Description
- Datasets
- Parameters
- Metrics
- Tags
- Artifacts



Partie 2 : Plateformes de Gestion de Projets - MLFlow

Interface - Modèles



Partie 2 : Plateformes de Gestion de Projets - MLFlow

Outils de déploiement de MLFlow :

- Local Flask Server
- Local Flask Server with MLServer
- Remote Container Serving
- Kubernetes
- Databricks Model Serving

Plan

Partie 1 : Cycle de vie d’un projet de machine learning

Partie 2 : Plateformes de Gestion de Projets

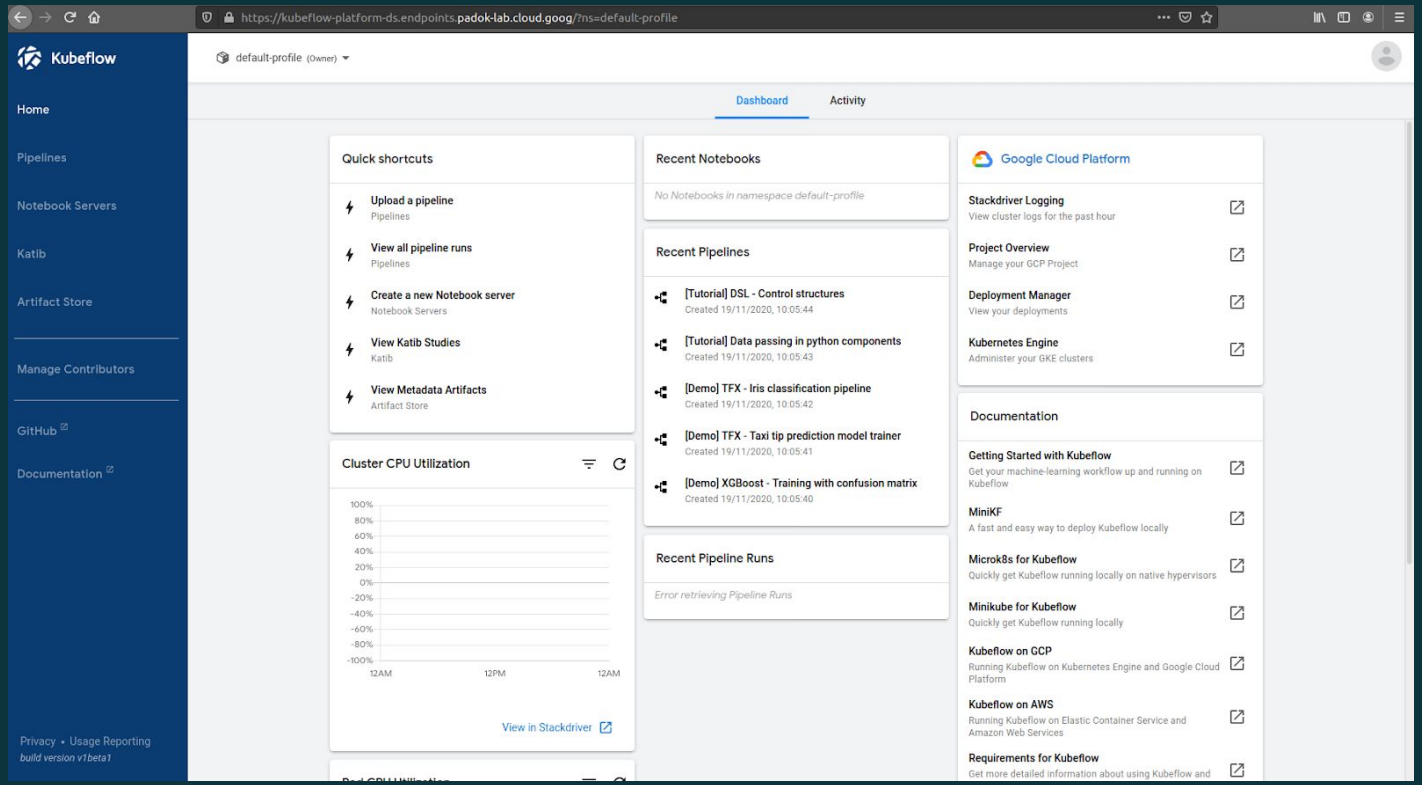
- Technologies et outils
- MLFlow
- **Kubeflow**

Partie 3 : Présentation du TD

Partie 2 : Plateformes de Gestion de Projets - Kubeflow

- Distribution spécifique de Kubernetes conçue pour le machine Learning (ML) et le déploiement de modèles
- Divers composants :
    - Kubeflow Pipeline (KFP)
    - Kubeflow Notebook
    - Kubeflow Central Dashboard
    - Katib
    - Kubeflow Training Operator
    - KServe

Partie 2 : Plateformes de Gestion de Projets - Kubeflow



Plan

Partie 1 : Cycle de vie d’un projet de machine learning

Partie 2 : Plateformes de Gestion de Projets

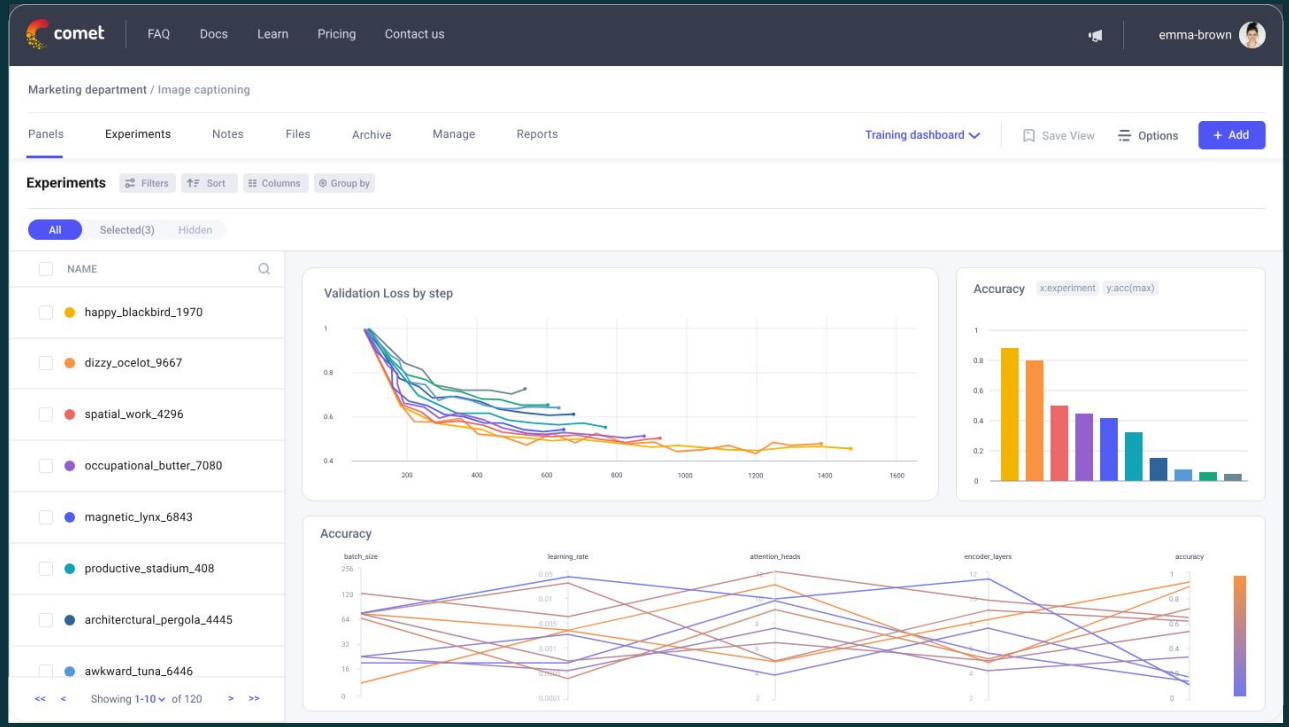
- Technologies et outils
- MLFlow
- Kubeflow
- **Comet.ml**

Partie 3 : Présentation du TD

Partie 2 : Plateformes de Gestion de Projets - Comet

- Plateforme d'apprentissage automatique basée sur le cloud
- Suivre, comparer, expliquer et optimiser leurs modèles tout au long du cycle de vie complet du ML
- Simplifier le processus de machine learning :
  Expérimentation → Mise en production

Partie 2 : Plateformes de Gestion de Projets - Comet



Plan

Partie 1 : Cycle de vie d’un projet de machine learning

Partie 2 : Plateformes de Gestion de Projets

- Technologies et outils
- MLFlow
- Kubeflow
- Comet.ml
- **Airflow**

Partie 3 : Présentation du TD

Partie 2 : Plateformes de Gestion de Projets - Airflow

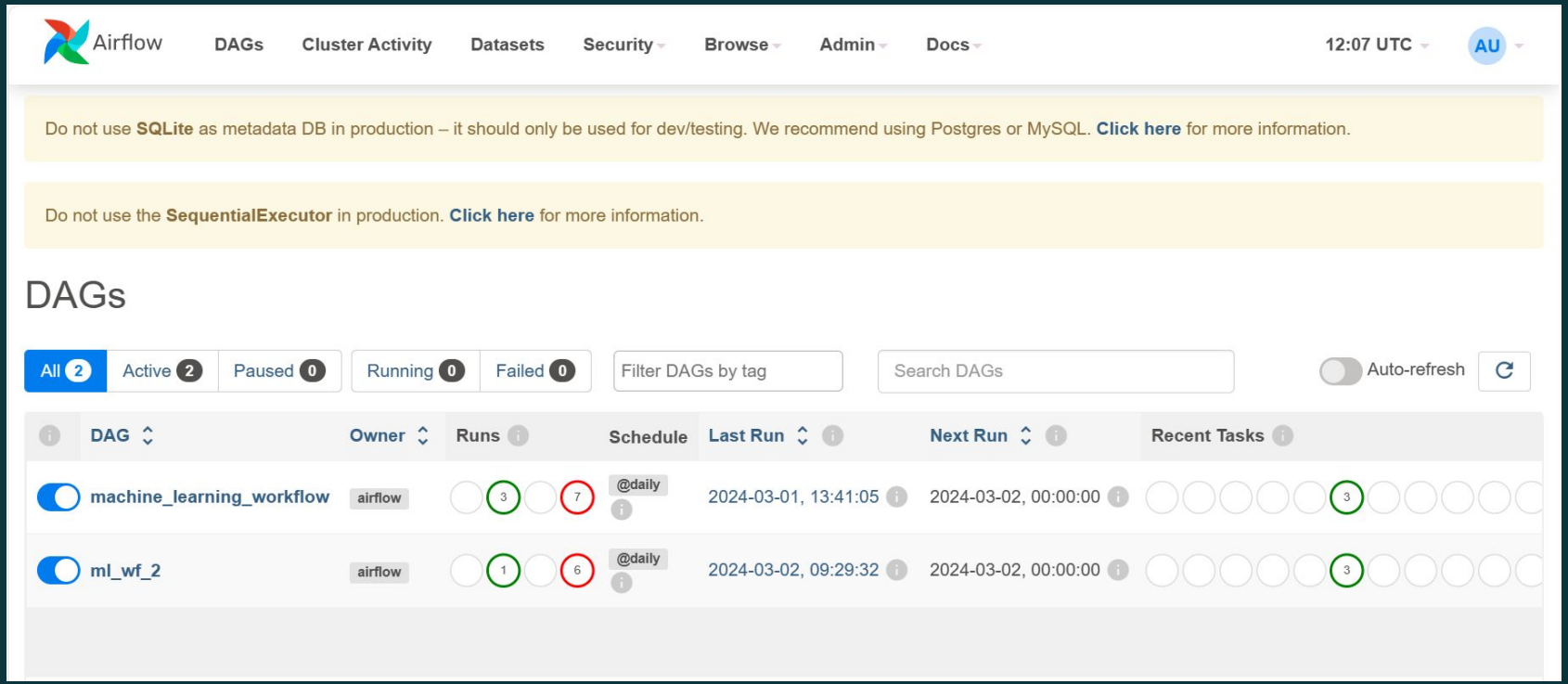
- Plateforme open-source
- Automatiser et orchestrer des workflows complexes
- Développé par Airbnb en 2014
- Installation via pip ou Docker

Partie 2 : Plateformes de Gestion de Projets - Airflow

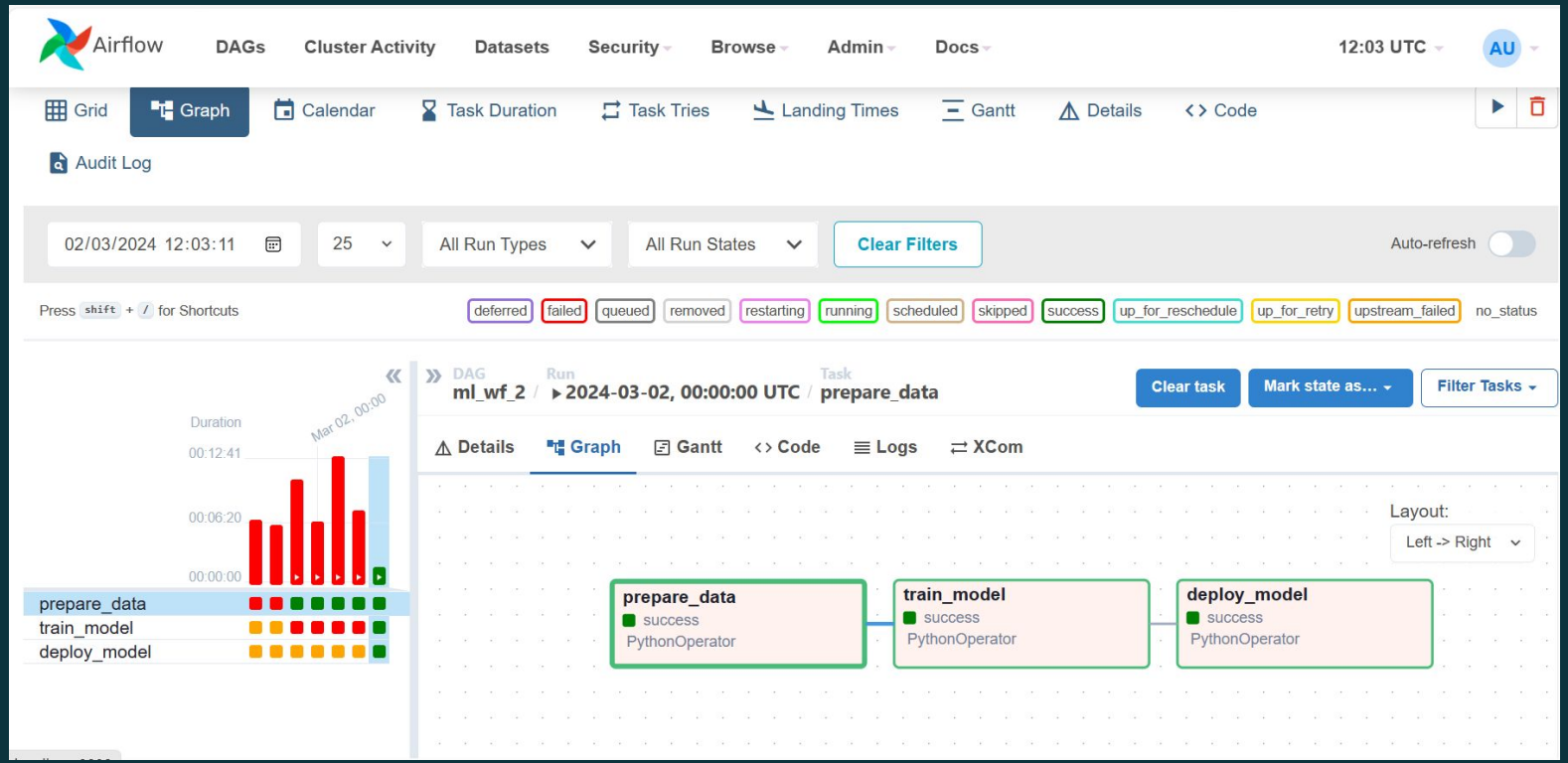
Fonctionnalités principales :

- Graphes acycliques dirigés (DAG)
- Planification avancée
- Connecteurs variés
- Extensibilité
- Reprise sur panne
- Interface Utilisateur Web

Partie 2 : Plateformes de Gestion de Projets - Airflow



Partie 2 : Plateformes de Gestion de Projets - Airflow



Plan

Partie 1 : Cycle de vie d’un projet de machine learning

Partie 2 : Plateformes de Gestion de Projets

- Technologies et outils
- MLFlow
- Kubeflow
- Comet.ml
- Airflow
- **Comparaison des technologies et outils**

Partie 3 : Présentation du TD

Partie 2 : Plateformes de Gestion de Projets

Comparaison des technologies et outils

	MLFlow	Kubeflow	Comet	Airflow	Prefect	Dagster
Objectif	Gestion ML	Orchestration ML	Suivi ML	Orchestration workflows	Orchestration workflows	Orchestration workflows
Facilité d'usage	Facile	Difficile	Facile	Moyenne	Facile	Facile
Intégration	Elevée	Spécifique à Kubernetes	Elevée	Large	Large	Large
Communauté	Croissante	Large	Active	Très active	Croissante	Active
Tarification	Variable selon le fournisseur (AWS, Azure...)	Variable selon le fournisseur (AWS, Azure...)	Basé sur l'utilisation, plans spécifiques au cloud	Variable, dépend des ressources consommées	Plans basés sur l'utilisation et l'échelle	Variable selon le fournisseur et l'usage

Plan

Partie 1 : Cycle de vie d’un projet de machine learning

Partie 2 : Plateformes de Gestion de Projets

Partie 3 : Présentation du TD

Partie 3 : Présentation du TD

1. MLFlow
2. Airflow

Merci pour votre attention.

Avez-vous des questions ?

Annexe

- Documentation MLflow : https://mlflow.org/docs/latest/index.html
- Documentation Airflow : https://airflow.apache.org/docs/
- Installation de MLFlow sur AWS : https://medium.com/ama-tech-blog/mlflow-on-aws-a-step-by-step-setup-guide-8601414dd6ea
- Documentation Comet : https://www.comet.com/docs/v2/
- Documentation Kubeflow : https://www.kubeflow.org/docs/