Kubeflow

Build and manage simple, portable and scalable machine learning workflows

Boris El Gareh

GROUPE CRÉDIT AGRICOLE

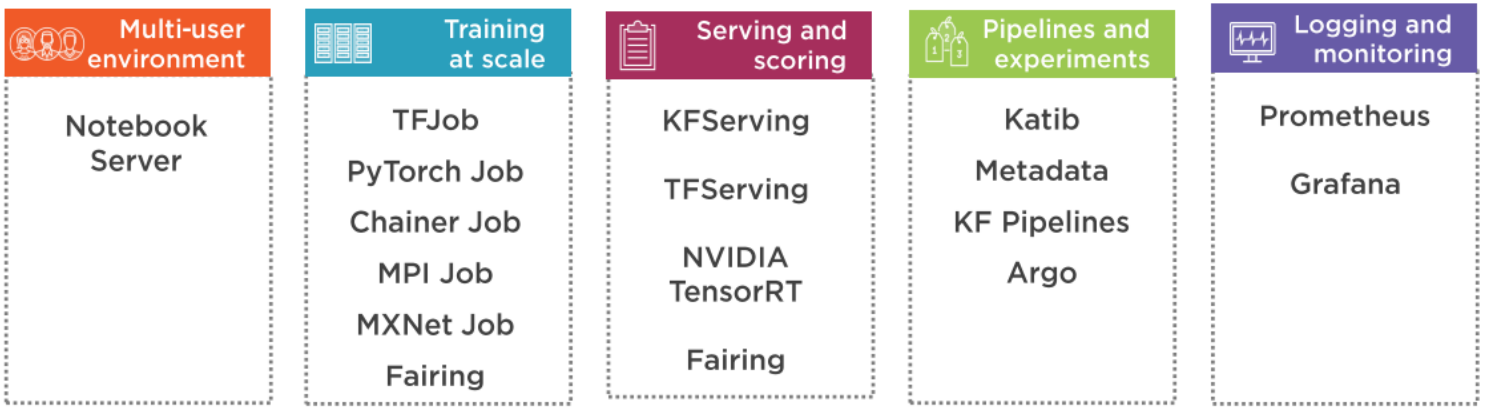
Introduction

- **Deal with scale**
- **Générer, surveiller et analyser**
- **Colaboration**
  - ❑ Travail d'équipe
  - ❑ Environnements différents
- **Tracking des expériences**
  - ❑ Traçabilité
  - ❑ Reproductibilité
  - ❑ Portabilité
- **Déploiement**
- **Gestion des modèles**

Aperçu

- **Open source machine toolkit pour Kubernetes**
- **Kubernetes**
  - ❑ Framework open source de Google
  - ❑ Automatisation du déploiement
  - ❑ Auto-scaling
  - ❑ Orchestration
- **Kubeflow : une adaptation de Kubernetes pour ML**
- **Développé par Google dans un premier temps**

Simple, portable and scalable workflow

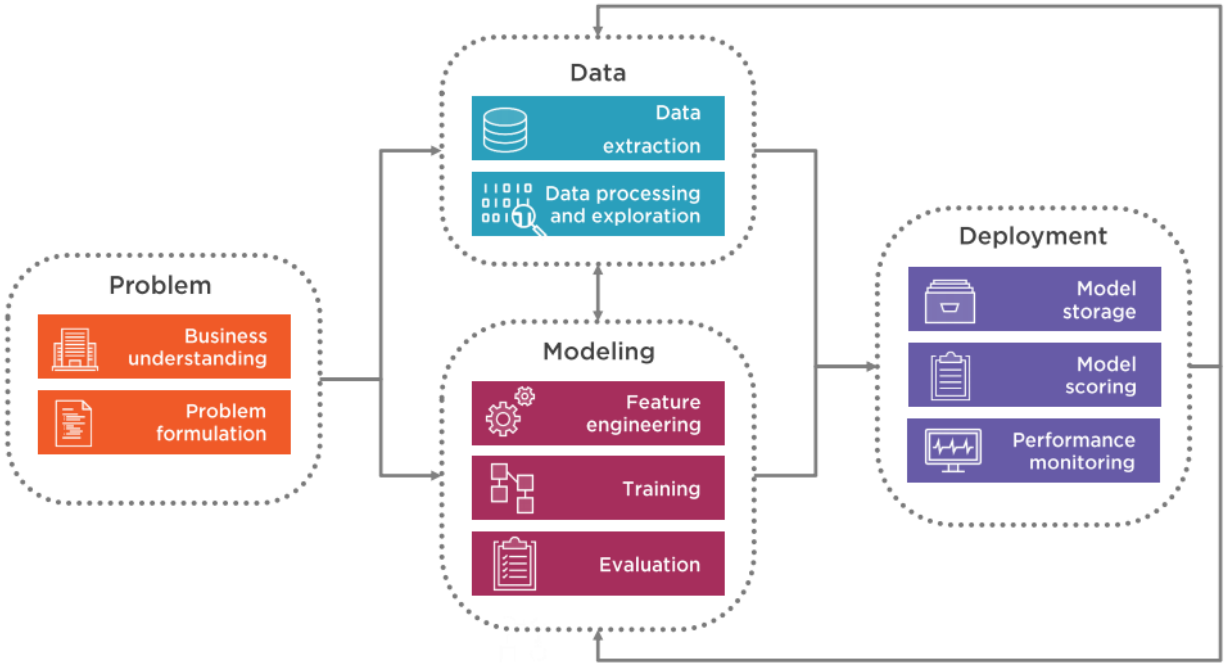


End-to-end ML Workflows avec Kubeflow

- Setting-up de l'environnement Kubeflow
- Construction d'un modèle de ML avec Kubeflow
- Déploiement d'un modèle de ML avec Kubeflow
- Construction d'un pipeline de ML avec Kubeflow Pipeline

- End-to-end image classification system
- Fashion-MNIST dataset
- Open source
- 28x28 pixels images
- Items : vêtements, sac et chaussures
- Réseau de neurones convolutifs (CNN)
  - TensorFlow

Fashion-MNIST ML workflow



Setting up Kubeflow environmenent

Kubeflow Setup

Kubeflow — Machine learning toolkit for Kubernetes

Kubernetes — Open source system that runs everywhere (on-premise, public cloud, hybrid)

Options de déploiement

Public cloud

Google Cloud Platform (GCP)

Amazon Web Services (AWS)

Microsoft Azure

On-premise

On-premise Kubernetes cluster

Private cloud

IBM private cloud

Local

MiniKF

Minikube

MicroK8s

Déploiement sur GCP et GKE

- **UI**
  - ❑ Approche facile
  - ❑ https://deploy.kubeflow.cloud
- **CLI**
  - ❑ Plus de contrôle
  - ❑ Contrôler facilement les versions
  - ❑ Automatisation

Démo setup

- **Prérequis**
  - ❑ Un projet GCP actif
  - ❑ gcloud
    - Google Cloud SDK
  - ❑ Kubectl
    - Interaction avec le cluster K8s
  - ❑ Idéalement travailler avec Linux ou macOS
- **OAuth authentification**
- **Déployer Kubeflow sur GCP**
- **Supprimer Kubeflow sur GCP**
  - ❑ Eviter les mauvaises surprises

Kubeflow Training

Challenges

Experiments tracking

- ❑ Améliorer la productivité
- ❑ Assurer la reproductibilité

Execution

- ❑ Single node
- ❑ Accélérateurs
  - GPU, TPU
- ❑ Multi-node/Multi-worker
  - Entrainement distribué

Environnement de développement

- ❑ Notebook (data scientist)
- ❑ Scripts (ML engineer)

Kubeflow components for training

- **Notebook server**
  - ❑ Interactive multi-user environment
- **Training at scale**
  - ❑ TFJob
  - ❑ PyTorch
  - ❑ MXNet
- **Fairing**
  - ❑ Training jobs depuis le notebook
- **Metadata**
  - ❑ Track model artifacts and metadata
- **Katib**
  - ❑ Hyperparameter tuning

Kubeflow Notebook

- Notebook server
- Utiliser une image pré-existente ou personnalisée
- Authentification et contrôle des accès
- Ajout de volume persistant pour les données ou workspaces
- Configuration des ressources (CPU, RAM)
- Configuration des accélérations (GPU)

Pourquoi utiliser des images personnalisées ?

- **Eviter de setup manuellement l'environnement**
- **Setup des images pour des équipes différentes**
  - ❑ Exploration des données (pandas, matplotlib, etc.)
  - ❑ Machine learning classique (Scikit-learn, etc.)
  - ❑ Deep-learning (Tensorflow, PyTorch, etc.)
- **Team centrale qui gère ces images customs**
  - ❑ Onboarding rapide des data scientists

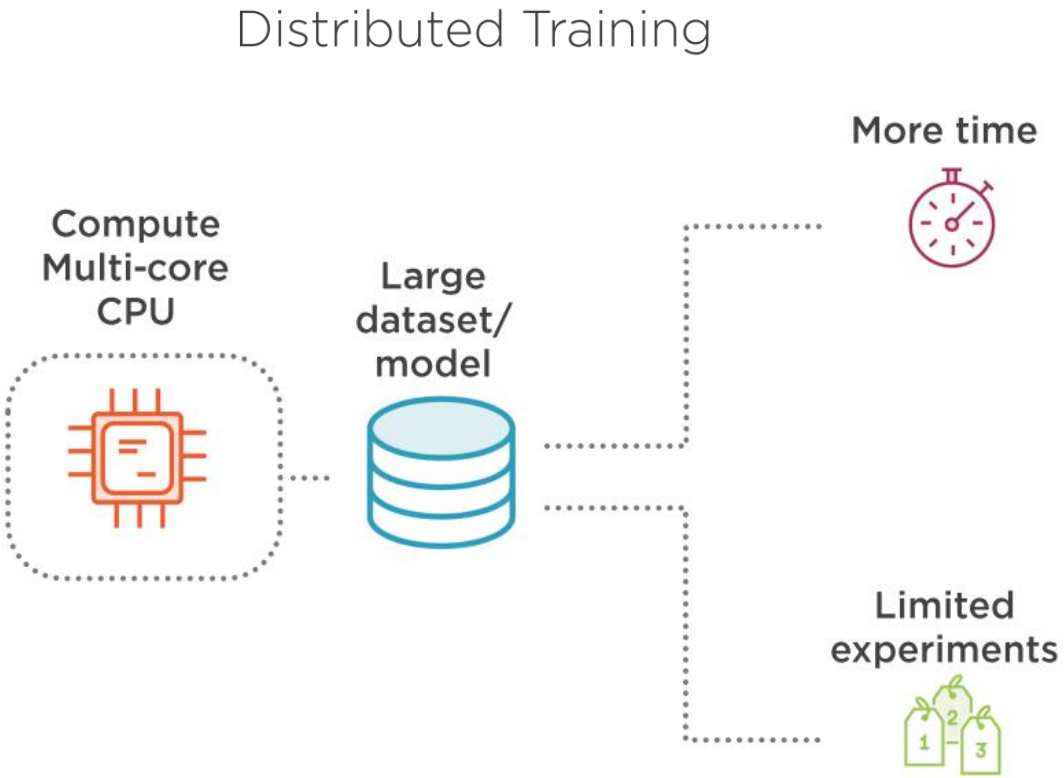
Kubeflow Metadata

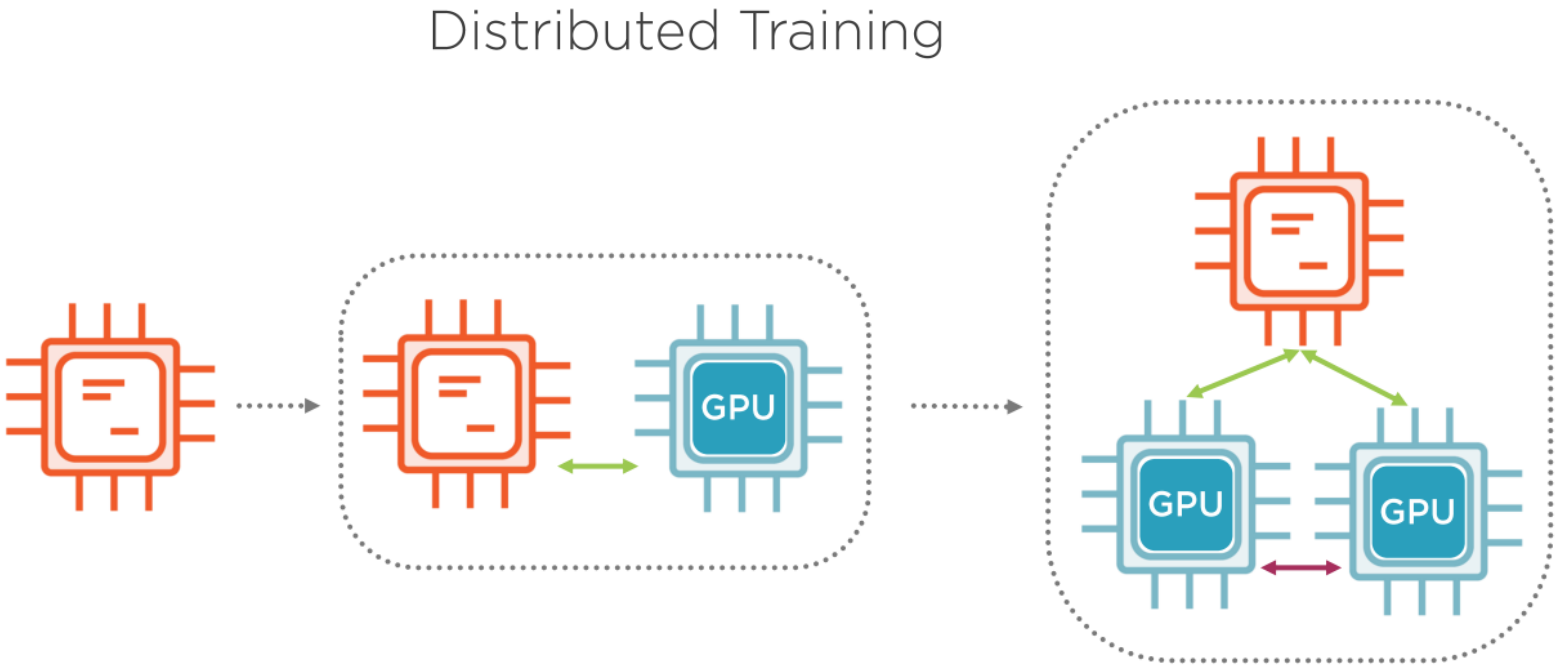
- Track et manage les metadatas
- Backend database pour storer les infos
- API pour requêter et retrouver les infos
- Artifact store dashboard
- Tracking des metadatas
  - ❑ Model
  - ❑ Metric
  - ❑ Dataset

Kubeflow Fairing

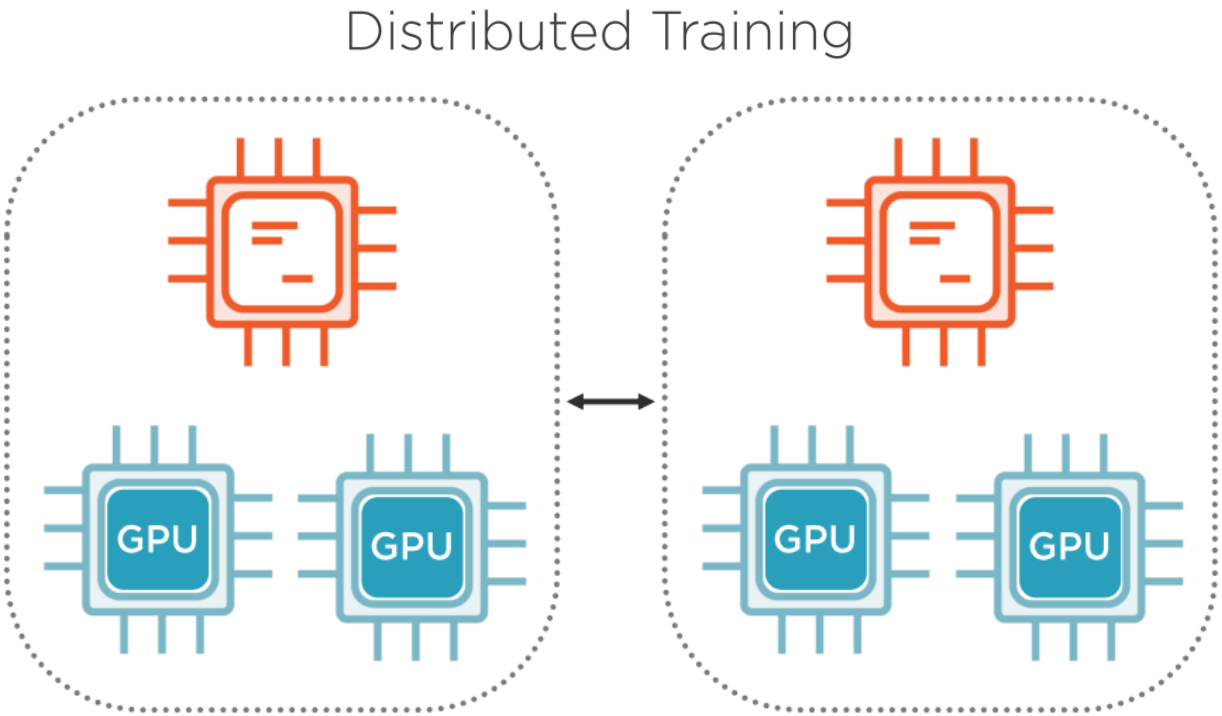
- **Python package to streamline the process**
  - ❑ Build
  - ❑ Train anywhere (local, cloud)
  - ❑ Deploy
- **Couche abstraite**
  - ❑ Executable directement depuis le notebook
  - ❑ Réutilisation des blocks
- **Parfait pour les data scientists**
- **En version beta**

Training at scale

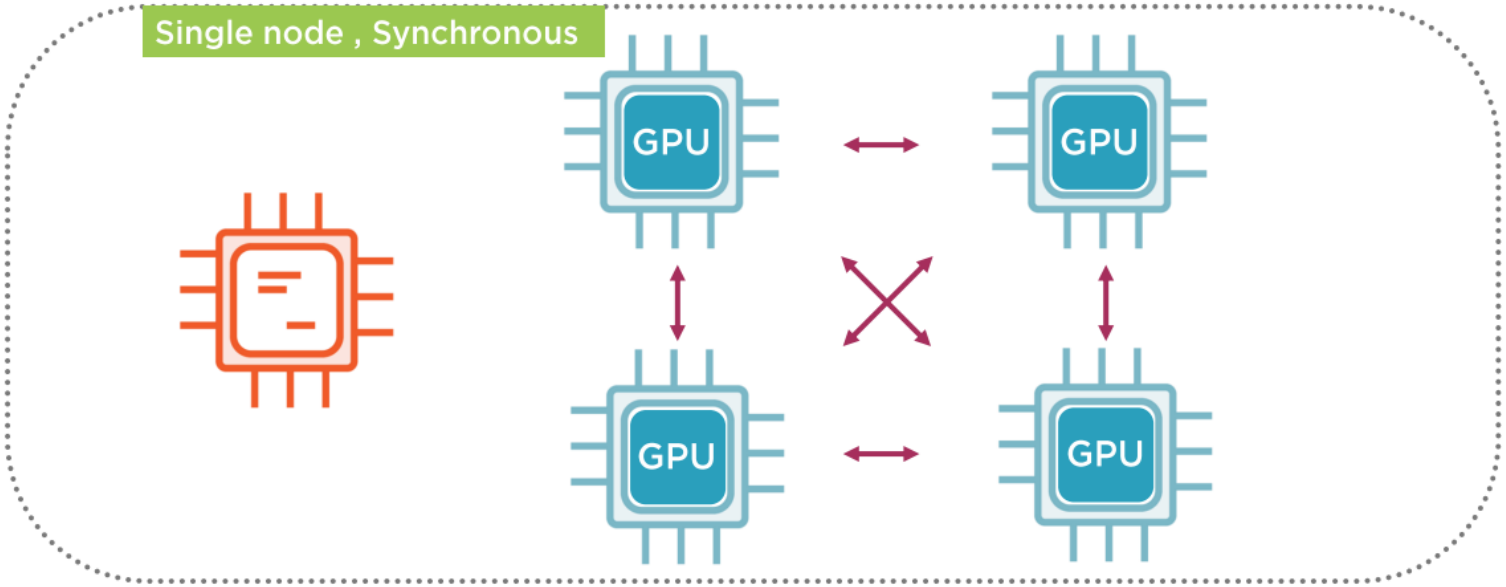



Distributed Training
GPU
GPU
GPU

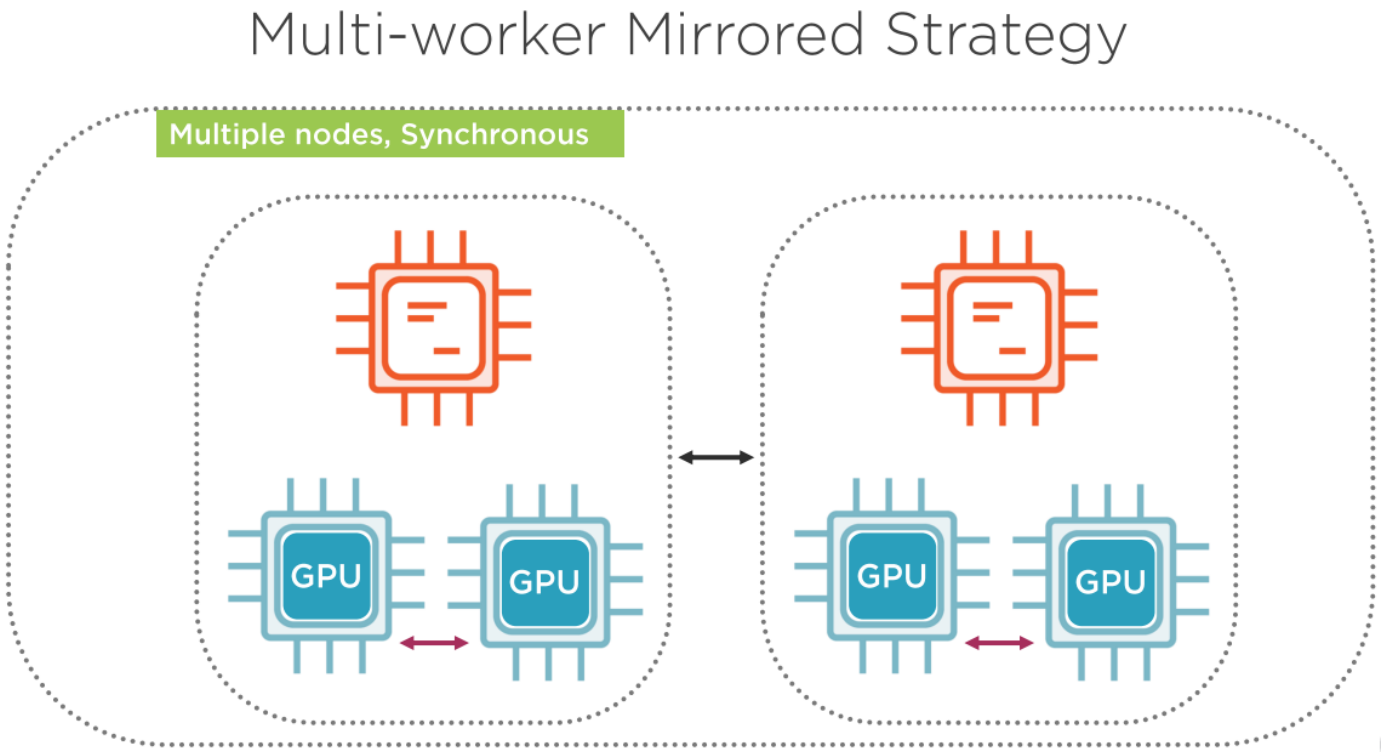
Training at scale



Mirrored Strategy



Training at scale



Hyperparameters

- Paramètres de configuration
  - Learning rate
  - Batch size
- Paramétrés avant processus d'entrainement

Hyperparameter tuning

- Trouver les valeurs optimales qui optimisent la fonction objectif
- Les valeurs optimales améliorent la performance du modèle

- **Inspiré de Google Vizier**
- **Framework agnostic**
- **Plusieurs algorithmes d'optimisation**
  - ❑ Random search
  - ❑ Grid search
  - ❑ Bayesian optimisation
  - ❑ Hyperband

Serving ML Models

Model serving challenges

- Deploiement
- Realease (Canary, A/B test)
- Scaling
- Monitoring
- Explication du modèle

Kubeflow components for serving

- **TensorFlow serving**
  - ❑ Serve TensorFlow models
- **NVIDIA TensorRT**
  - ❑ NVIDIA inference server
- **Seldon core serving**
  - ❑ Support multiple framework
- **KFServing**
  - ❑ High level abstractions for common frameworks

- **Serverless inference on Kebernetes**
- **Supporte des frameworks communs**
  - TensorFlow, XGBoost, Scikit-learn, PyTorch, ONNX, etc.
- **Possibilité de custom**
- **Déploiement**
  - Canary Rollouts
- **Performance monitoring**
  - Prometheus, Grafana, Elasticsearch
- **Amélioration de l'explication des modèles**
- **Auto-scaling et load-balancing**

Kubeflow Pipeline

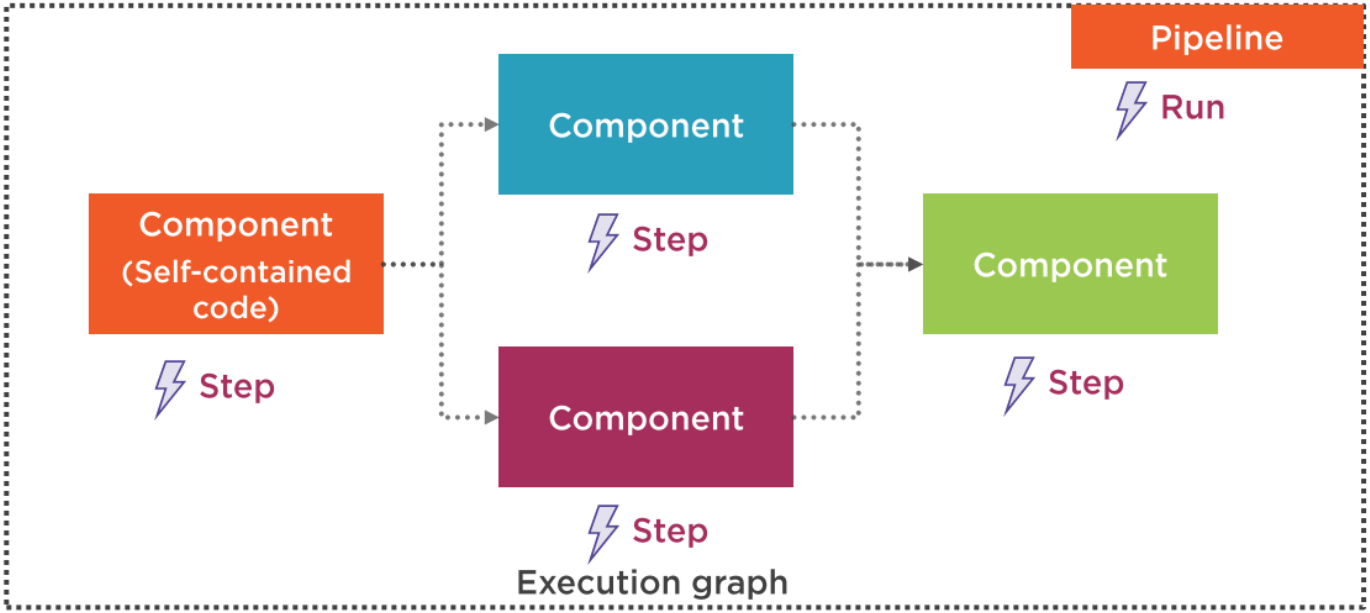
Pourquoi un pipeline ?

- Reproductibilité
- Orchestration de bout-en-bout
- Automatisation
- Expérimentation rapide
- Experiment to production
- Réutilisabilité

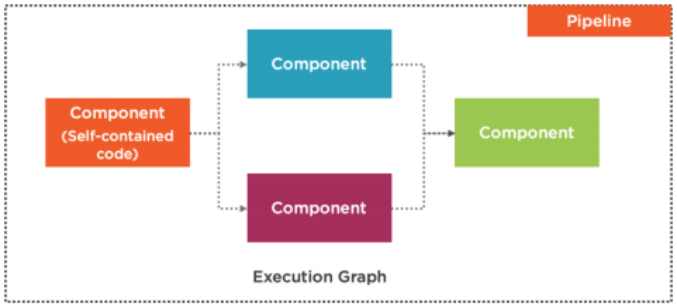
Kubeflow Pipeline

- **Build and deploy portable, scalable, workflows**
- **Expérimentation rapide**
  - ❑ Management et tracking des expériences avec l'UI
- **Orchestration de bout-en-bout**
  - ❑ Utilisation du SDK pour définir et orchestrer le pipeline et ses composants
  - ❑ Multi-step workflow engine
  - ❑ Utilisation d'un notebook pour intéragir avec le SDK

Kubeflow Pipeline Concepts



Kubeflow Pipeline Concepts



Run - 1
Run - 2
Run - 3
Experiment - 1

Run - 4
Run - 5
Run - 6
Experiment - 2

Kubeflow Pipeline Concepts

Pour aller plus loin

Explorer des features avancées de Kubeflow

- **Setup**
  - ❑ Cloud (AWS, Azure), On-premise
  - ❑ Authentification avancée
    - Multi-tenancy component
- **Kustomize**
  - ❑ Customisation des fichiers de conf yaml
- **Katib**
  - ❑ Stratégies d'hyperparameters tuning avancées
- **Pipeline**
  - ❑ Intégration avec des artifacts d'output
  - ❑ CI/CD pipeline