Telecom Valley | SophiaConf

100% OPEN SOURCE - Conférences et Workshops

MARDI 29 JUIN - CONFÉRENCE

Déploiement sur le cloud d'approches non supervisées d'IA pour détecter les corruptions de données

Sébastien MARTI
& Rémy LARROYE

orange™

Sébastien MARTI
DataScientist

Rémy LARROYE
DataEngineer / Alternant

orange™

26 Pays

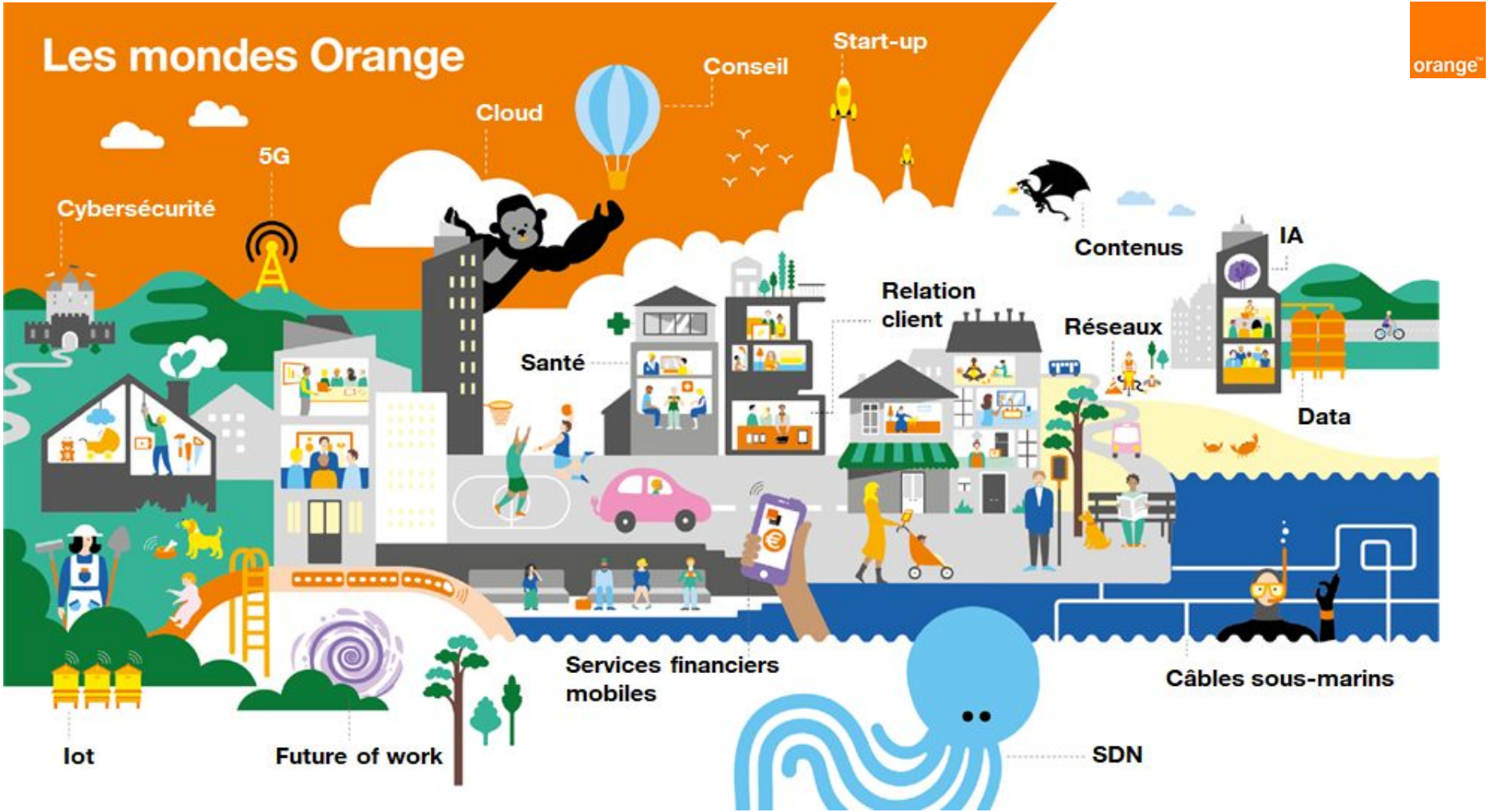
266 000 000 Clients

42 200 000 000 € (CA 2019)

40 Thèses / an

600 Alternants / an en région Sud-Est

1865 Offres : **https://orange.jobs/**


Les mondes Orange
orange
Start-up
Conseil
Cloud
5G
Cybersécurité
Contenus
IA
Relation client
Réseaux
Santé
Data
Services financiers mobiles
Câbles sous-marins
Iot
Future of work
SDN

Nos enjeux

Fiabiliser les données

- Plus de projets pilotés par la donnée
- Plus d’automatisations
- Des modèles sensibles aux bruits

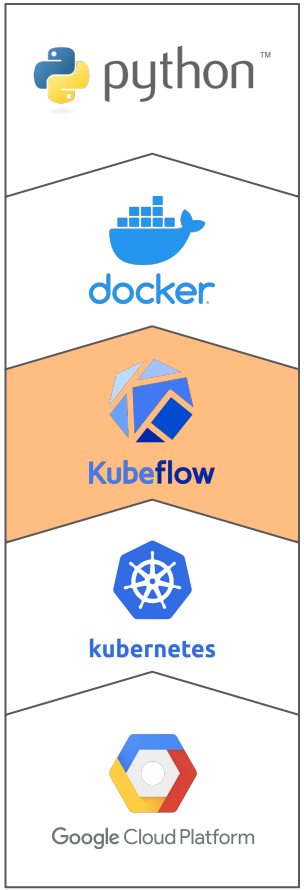
Embrasser le cloud

- Réduction des coûts
- Scalabilité, disponibilité des ressources
- Développement plus agile

Le cloud pour nous

Plateforme basé sur Kubeflow :

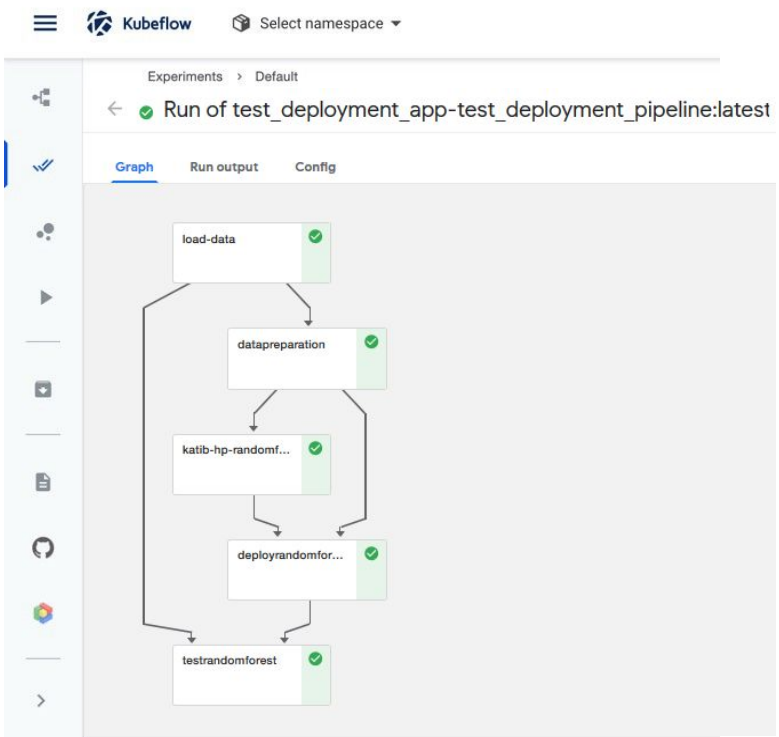
- ➔ Open source, fait pour le cloud
- ➔ Pas d'adhérence avec un cloud provider
- ➔ Gestion de l'ensemble d'un projet de l'exploration à la production



Nos objectifs avec Kubeflow

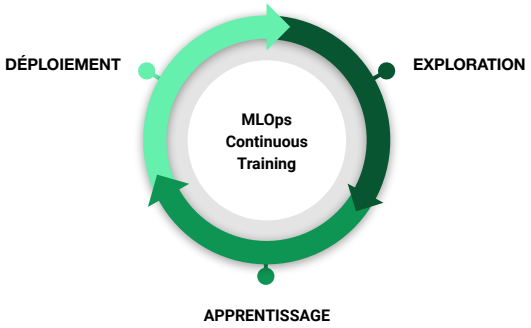
Plateforme basé sur Kubeflow :

- ➔ Réutilisation de composants
- ➔ Automatisation
- ➔ Standardisation



Kubeflow pipeline

Programme



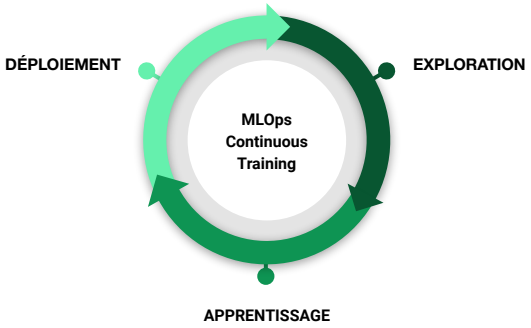
Exploration

Apprentissage

Déploiement

Perspectives

Programme



Exploration

Les algos étudiés
Mise à disposition d’espace d’exploration

Apprentissage

Déploiement

Perspectives

Exploration

Monitorer les qualités des données

Plusieurs aspects de la qualité des données :

- ➔ Conformité (respect du format)
- ➔ Unicité (valeurs dupliquées)
- ➔ Complétude (données vides)
- ➔ Intégrité (fréquences des valeurs)
- ➔ Cohérence (corrélations)

Exploration

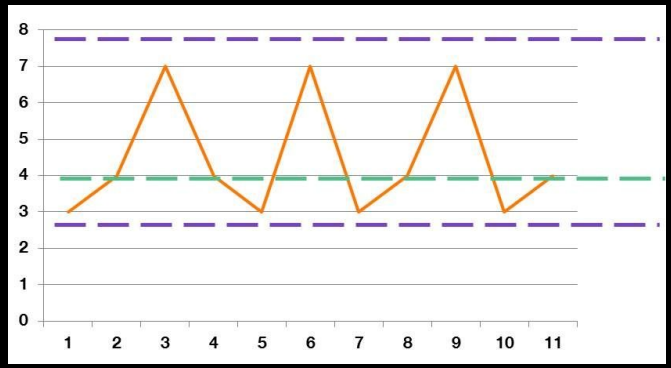
Monitorer les qualités des données

Plusieurs aspects de la qualité des données :

- ➔ Conformité (respect du format)
- ➔ Unicité (valeurs dupliquées)
- ➔ Complétude (données vides)
- ➔ Intégrité (fréquences des valeurs)
- ➔ Cohérence (corrélations)

} Control-chart

Control-Chart



Démarrage à froid :
N échantillons tirés aléatoirement

Exploration

Monitorer les qualités des données

Plusieurs aspects de la qualité des données :

- ➔ Conformité (respect du format)
- ➔ Unicité (valeurs dupliquées)
- ➔ Complétude (données vides)

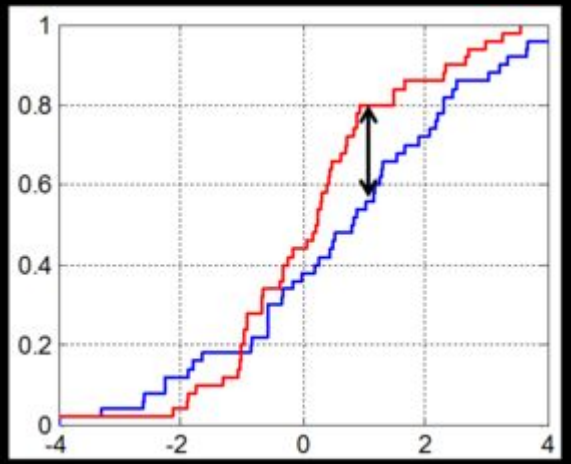
} Control-chart

- ➔ Intégrité (fréquences des valeurs)

} KS-test

- ➔ Cohérence (corrélations)

KS-Test



Exploration

Monitorer les qualités des données

Plusieurs aspects de la qualité des données :

- ➔ Conformité (respect du format)
- ➔ Unicité (valeurs dupliquées)
- ➔ Complétude (données vides)
- ➔ Intégrité (fréquences des valeurs)
- ➔ Cohérence (corrélations)

Control-chart

KS-test

Autoencodeur

Isolation Forest

Isolation Forest

Multiples arbres d'isolation pour identifier la fréquence des valeurs

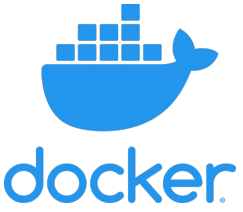
Autoencodeur

Compression/Décompression de données

Erreur de reconstruction

Exploration

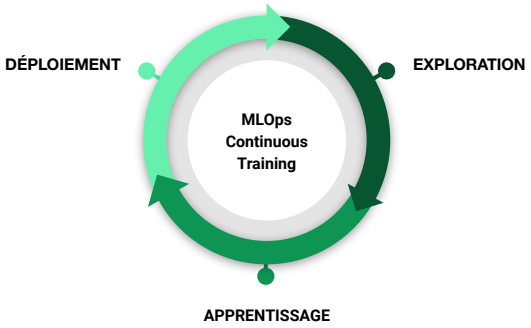
Mise à disposition d’espace d’exploration



De nombreux avantages à travailler sur le cloud :

- ➔ Espace de travail déjà configuré
- ➔ Accès aux ressources
- ➔ Possibilité travail collaborative
- ➔ Mise à disposition dans la minute

Programme



Exploration

Apprentissage

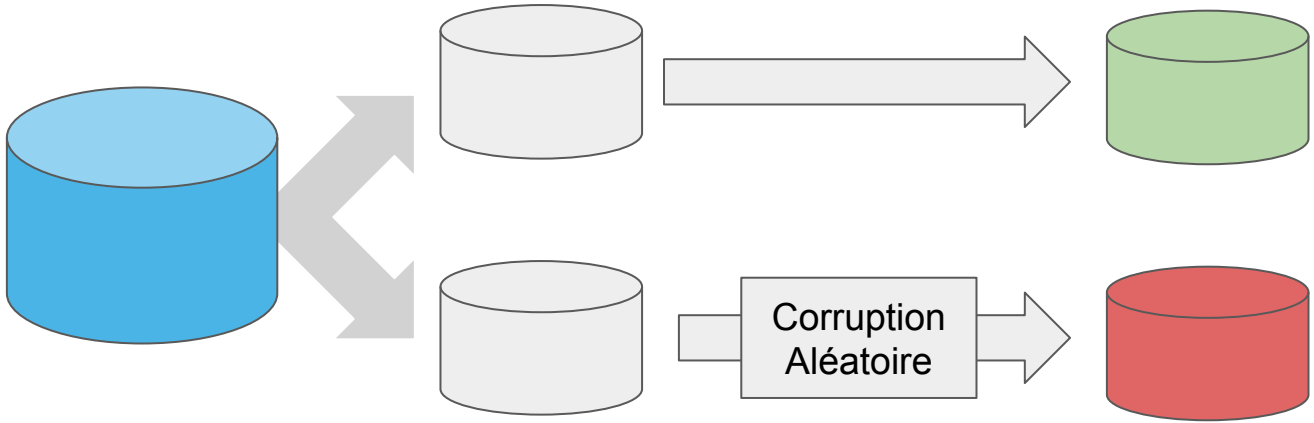
Comment mesurer la performance ?
Entraînements à grande échelle de modèles

Déploiement

Perspectives

Apprentissage

Mesurer la performance



Apprentissage

Entraînement des modèles

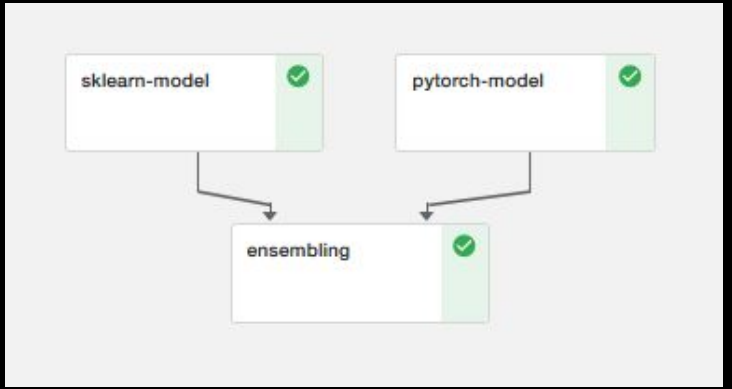
- ➔ Entrainement distribué
- ➔ Facilitation par les pipelines
- ➔ Recherche des meilleurs hyper-paramètres

Apprentissage

Entraînement des modèles

- ➔ Entrainement distribué
- ➔ **Facilitation par les pipelines**
- ➔ Recherche des meilleurs hyper-paramètres

Kubeflow pipeline

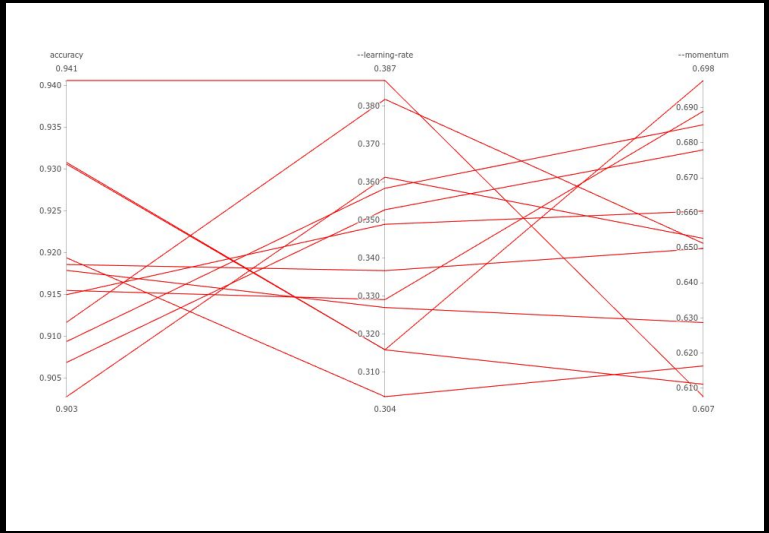


Apprentissage

Entraînement des modèles

- ➔ Entrainement distribué
- ➔ Facilitation par les pipelines
- ➔ **Recherche des meilleurs hyper-paramètres**

Katib

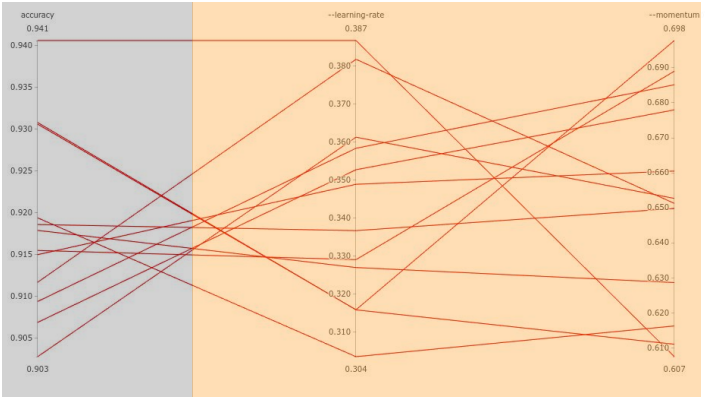


Apprentissage

Entraînement des modèles

- ➔ Entrainement distribué
- ➔ Facilitation par les pipelines
- ➔ **Recherche des meilleurs hyper-paramètres**

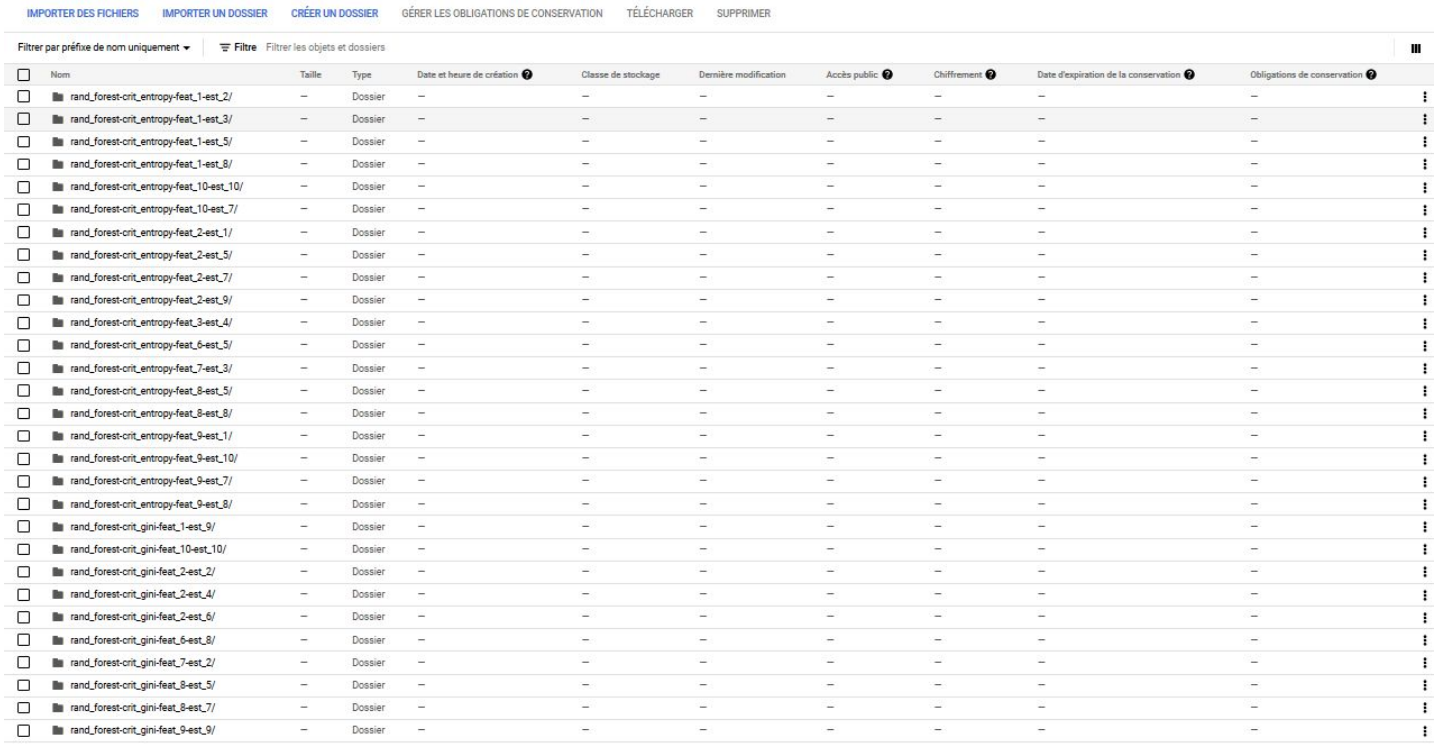
Katib



Cible Hyper-paramètres

Apprentissage

Le besoin de garder des traces



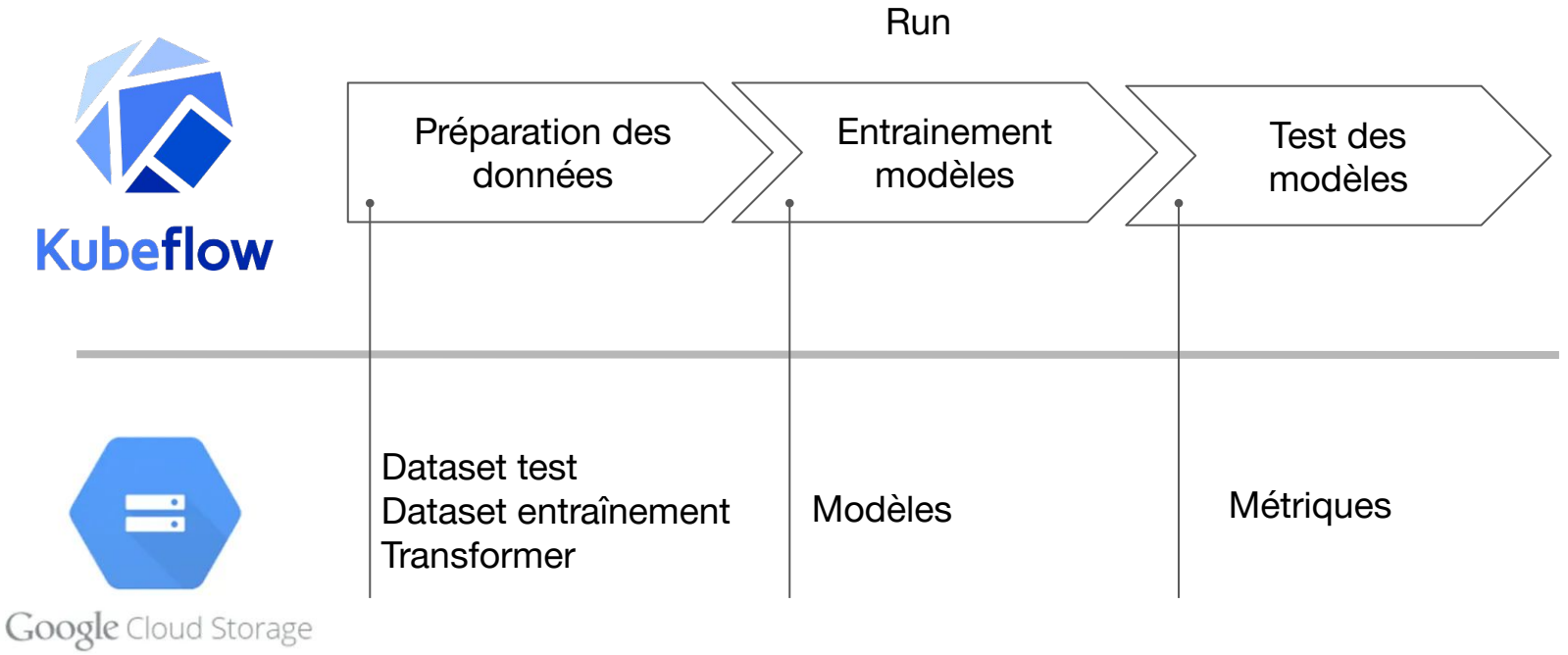
IMPORTER DES FICHIERS | IMPORTER UN DOSSIER | CRÉER UN DOSSIER | GÉRER LES OBLIGATIONS DE CONSERVATION | TÉLÉCHARGER | SUPPRIMER

Filtrer par préfixe de nom uniquement | Filtre | Filtrer les objets et dossiers

Nom	Taille	Type	Date et heure de création	Classe de stockage	Dernière modification	Accès public	Chiffrement	Date d'expiration de la conservation	Obligations de conservation
rand_forest-crit_entropy-feat_1-est_2/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_1-est_3/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_1-est_5/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_1-est_8/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_10-est_10/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_10-est_7/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_2-est_1/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_2-est_5/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_2-est_7/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_2-est_9/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_3-est_4/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_6-est_5/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_7-est_3/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_8-est_5/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_8-est_8/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_9-est_1/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_9-est_10/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_9-est_7/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_entropy-feat_9-est_8/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_gini-feat_1-est_9/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_gini-feat_10-est_10/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_gini-feat_2-est_2/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_gini-feat_2-est_4/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_gini-feat_2-est_6/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_gini-feat_6-est_8/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_gini-feat_7-est_2/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_gini-feat_8-est_5/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_gini-feat_8-est_7/	–	Dossier	–	–	–	–	–	–	–
rand_forest-crit_gini-feat_9-est_9/	–	Dossier	–	–	–	–	–	–	–

Apprentissage

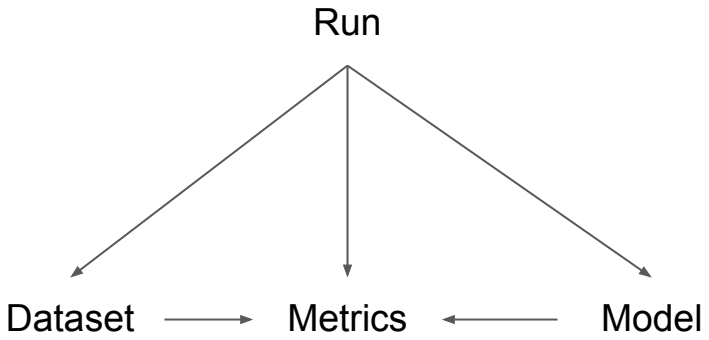
Le besoin de garder des traces



Apprentissage

Le besoin de garder des traces

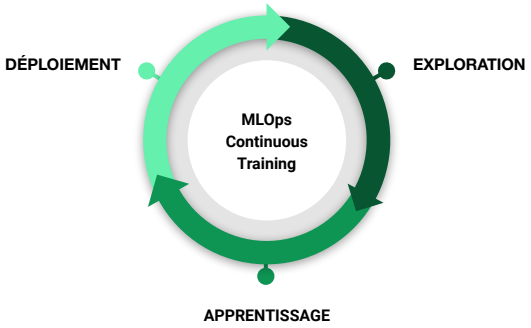
google/**ml-metadata**



Model

```
name="MNIST",
description="model to recognize handwritten digits",
owner="someone@kubeflow.org",
uri="gcs://my-bucket/mnist",
model_type="neural network",
training_framework={
       "name": "tensorflow",
       "version": "v1.0"},
hyperparameters={
       "learning_rate": 0.5,
       "layers": [10, 3, 1],
       "early_stop": True
},
version="v0.0.1",
labels={"mylabel": "l1"}))
```

Programme



Exploration

Apprentissage

Déploiement

Paramétrage à chaud
La puissance du cloud, simplement

Perspectives

Déploiement

Régler la sensibilité à chaud

Paramétrage à chaud en fonction de la fréquence maximale d’alerting désirée

Modulation de la sensibilité par aspects de la qualité des données

Approche hybride

Déploiement

La puissance du cloud, simplement

Avantages de KFServing :

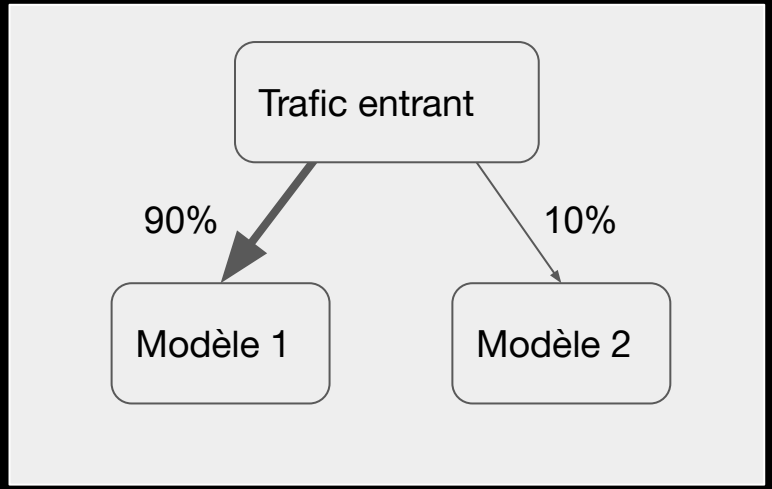
- ➔ Automatisable
- ➔ Déploiement en mode canary
- ➔ Élasticité de l’infrastructure

Déploiement

La puissance du cloud, simplement

Avantages de KFServing :

- ➔ Automatisable
- ➔ **Déploiement en mode canary**
- ➔ Élasticité de l’infrastructure



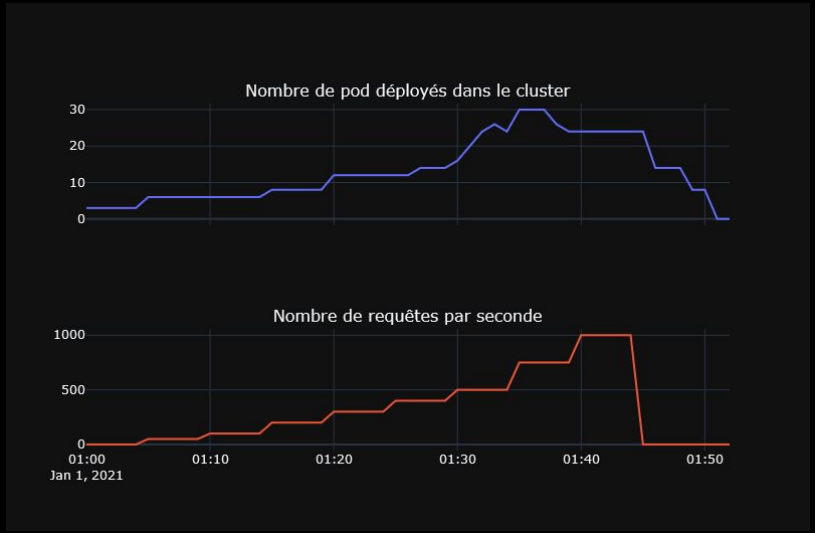
Déploiement

La puissance du cloud, simplement

Avantages de KFServing :

- ➔ Automatisable
- ➔ Déploiement en mode canary
- ➔ **Élasticité de l'infrastructure**

Test KFServing



Perspectives

Embrasser le cloud

- Impact environmental
- Augmentation du nombre de projet
- Première mise en production

Fiabiliser les données

- Stratégies de réapprentissage
- Opportunité sur les Autoencodeurs
- Intégration aux chaînes MLOps

Merci !

#lifeatorange sur orange.jobs