DataScientest • com

Movie Recommandation

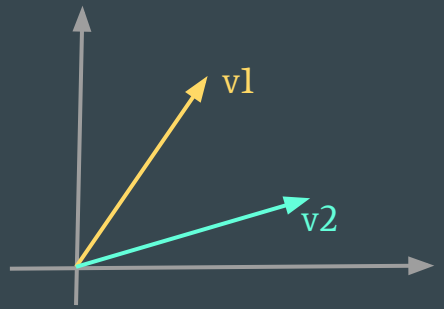
Projet MLOPS - Promotion Janvier 2023

Data et Pre-Process

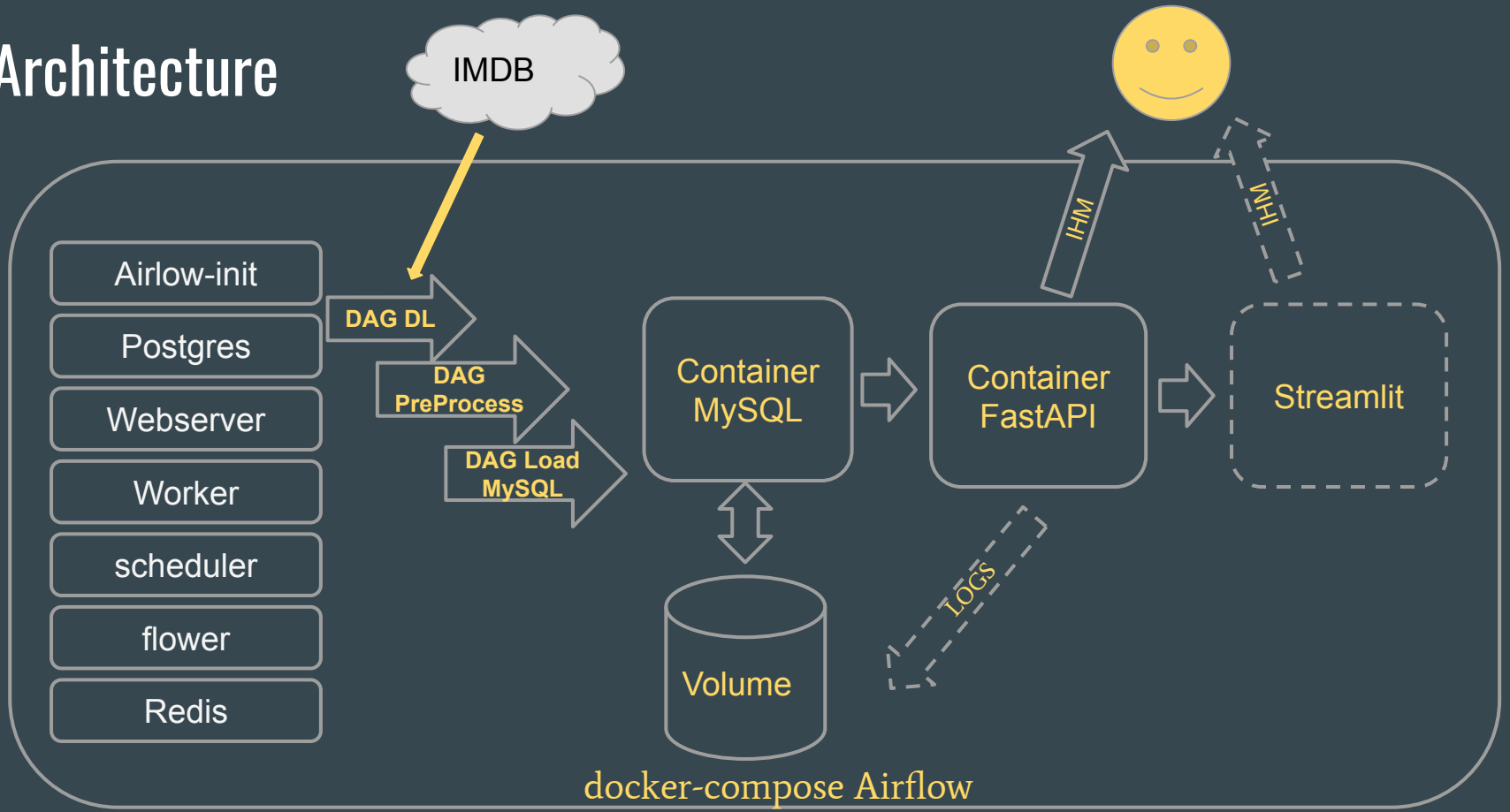
- Deux Sources :
  - IMDB
  - MDL
- Focus sur IMDB, sur le fichier title basics :
  - Fichier brut : 9 672 907 documents vidéo
  - Après pre-process (Films post- années 2010) : 156 090 films
- Preprocess :
  - Sélection des features
  - Nettoyage (Nans et erreurs de saisies)
  - Filtrages pour réduire la taille du dataset
  - Discrétisation (Année et durée)

Modélisation

- Problématique de recommandation :
  - Approche Content-Based vs Collaborative-based
- Modèles Cosine-Similarity et classif. KMeans
  - => le CS nous à semblé plus pertinent et plus rapide à mettre en oeuvre
- Qu'est-ce que Cosine-Similarity ?
  - Calcul matriciel qui renvoie une valeur entre [-1;1]
  - Concaténation de plusieurs features en une seule
  - Tokenisation
  - Calcul du CS sur la totalité de la base
  - Sélection du top 10 : on retourne les Id des films



Architecture



Demo

Points améliorations

- Architecture : dissocier la BDD du Airflow pour gagner en scalabilité
- Preprocess :
  - Ajouter plus de features (Nom des directeurs, acteurs ...) pour gagner en pertinence
  - Ne charger que les nouveaux films
- Calcul :
  - Réaliser un pré-filtrage sur le genre pour baisser le nombre de calculs
  - Pré-calcul
- Finaliser les logs
- Ajouter d'autres modèles + un pipeline CI/CD
- IHM : interface streamlit
- Ajouter une sécurisation par Token