



Airline Project

Modèle de prédiction de retards de vols



Équipe de travail

Datascientest - Parcours Data engineer



Mehdi FEKIH



Sirine DHOUIB



Ayoub RABEH



Maxime ROUX



CONTENU

01

Introduction &
contexte

02

Récolte &
Stockage de
données

03

ML process

04

Déploiement

05

Simulation

06

Conclusion

Introduction & contexte

Contexte:

Projet “fil rouge”: Prédiction du retard de vols
cursus Datascientest “Data Engineer”
Bootcamp 3 mois



Environnement et collaboration:

- Github - https://github.com/maxroux/mai24_bde_airlines/
- Points d'étapes hebdomadaires, mentor DS



Réalisations:

- Import des données
- Implémentation de modèles ML
- Déploiement et Monitoring



Récolte & Stockage des données

01

Choix des données: **Lufthansa API**

- Données de référence (statiques): Countries, Cities, Airports, Airlines, Aircraft
⇒ **Plusieurs sources "Reference data"**
- Données de vols (dynamiques): aéroport départ/arrivée, horaires, statut, etc.
⇒ **1 seule source "Flight Status"**

02

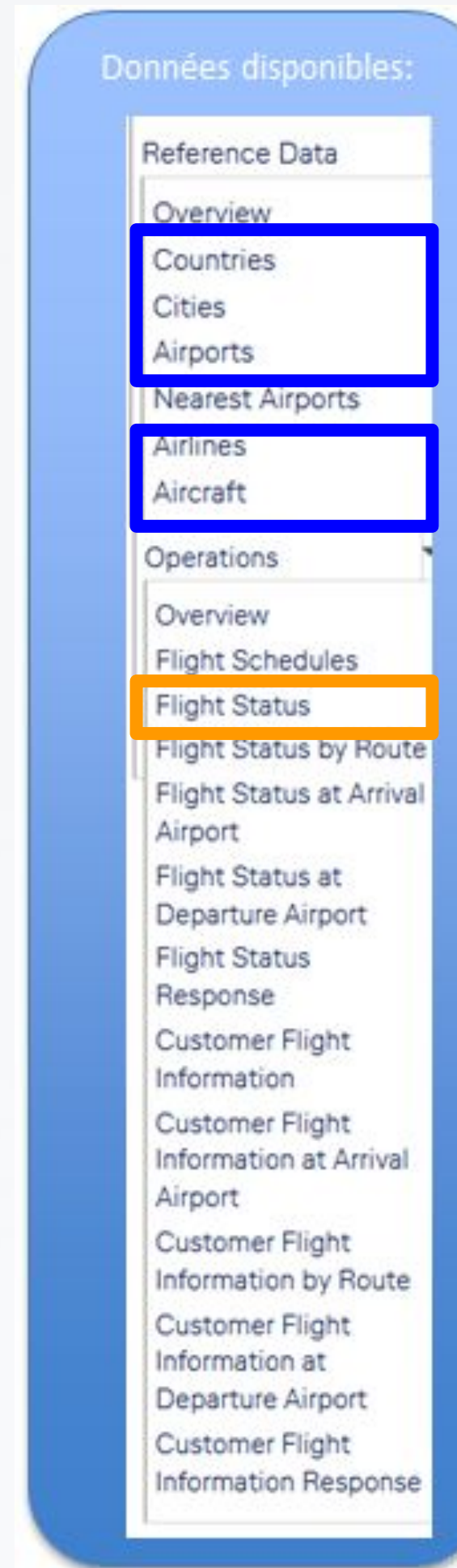
Récolte des données:

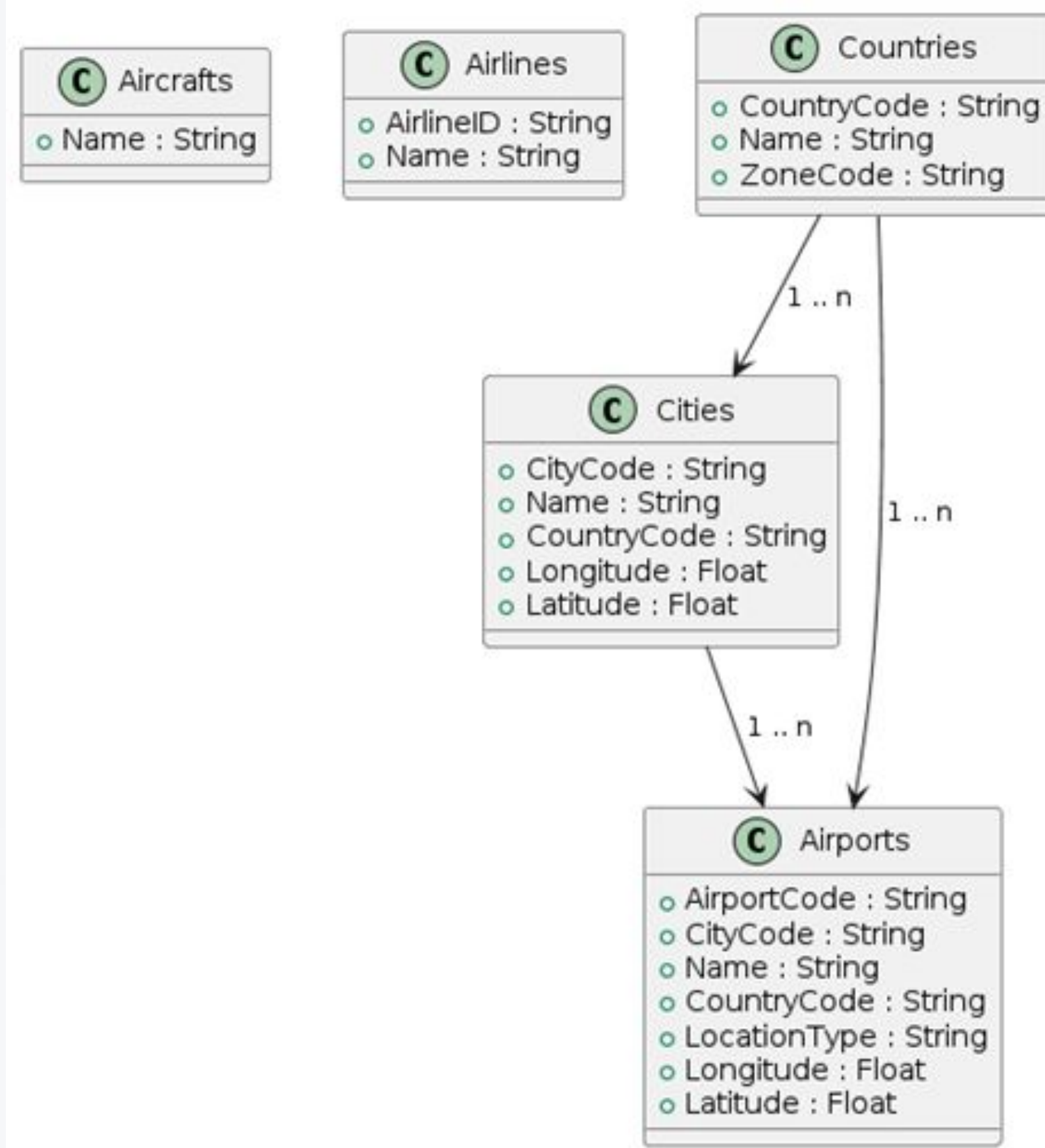
- Connexion à l'API
- Boucle avec limit / offset pour récolter toutes les données
- Planification journalière (manuelle, puis avec Airflow) pour les données de vol

03

Stockage des données:

- Données de référence → PostgreSQL (SGBD relationnelle)
- Données de vols → MongoDB (BDD NoSQL)





Machine Learning

Chargement

Mongodb pour charger les données.

- Horaires des vols départ/arrivé.
- Aéroports.
- Compagnie Aérienne.
- Status du vol.

Prétraitement

- Sélection des variables
- Suppression des duplicates
- Traitement des valeurs manquantes
- Encodages des données :
 - One Hot encoding
 - Frequency encoding
 - Standard Scales

Entrainement

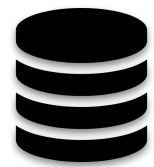
- Forêts aléatoires
- Régression linéaire
- Support Vector Machines
- LGBRM Regressor

Evaluation

- MAE (Erreur Absolue Moyenne).
- MSE (Erreur Quadratique Moyenne).
- RMSE (Racine carré de MSE).
- R^2

Déploiement : stack technique

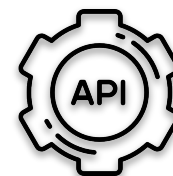
Base de données



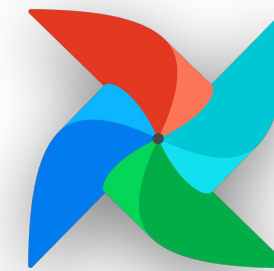
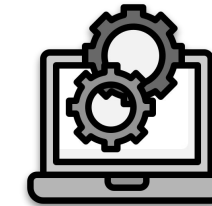
Versioning ML



API



Orchestration



Monitoring





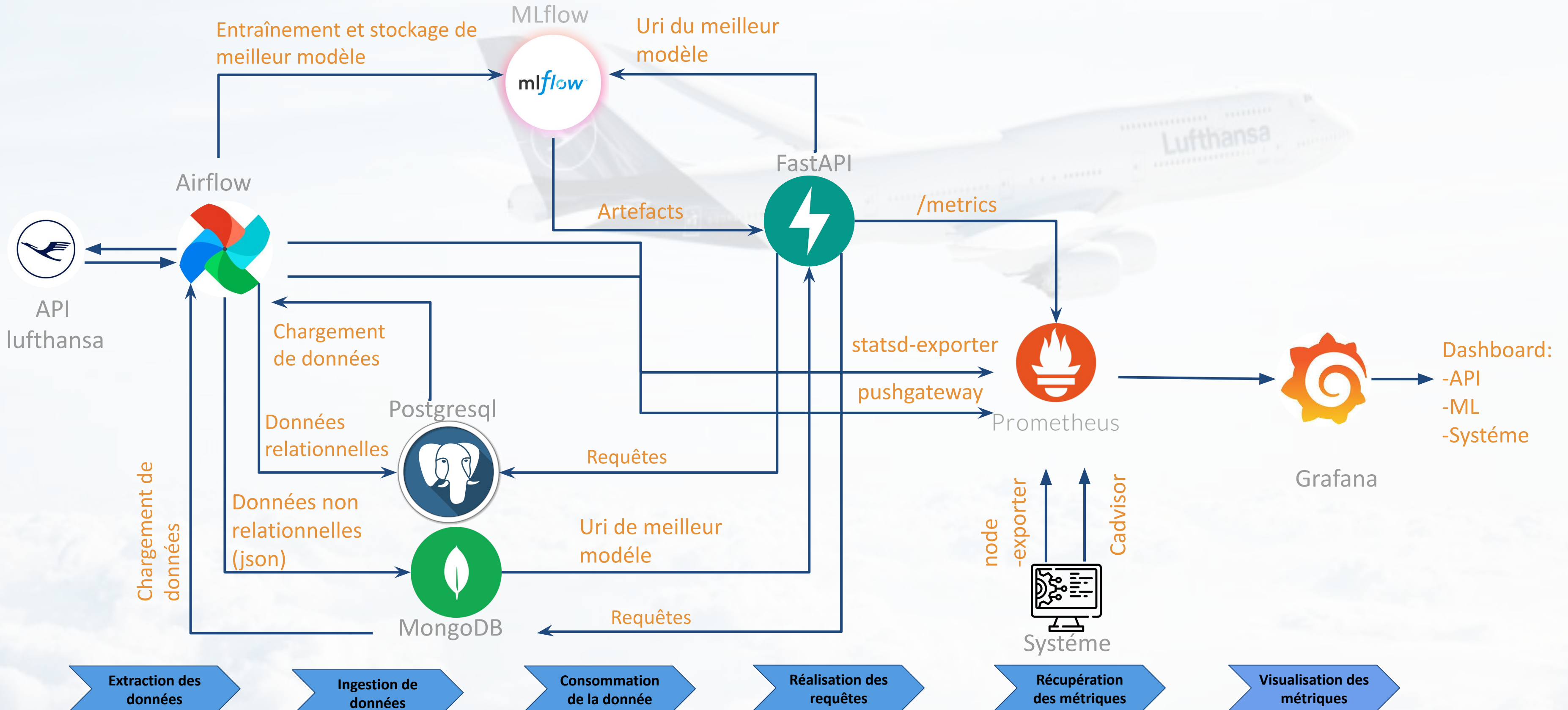
Orchestration

Architecture des données

Versioning de modèle

Déploiement

Monitoring



Démonstration du déploiement

AIRFLOW : <http://airlineproject.duckdns.org:8085/>

MLFLOW : <http://airlineproject.duckdns.org:5001/>

API : <http://airlineproject.duckdns.org:8002/>

PROMETHEUS : <http://airlineproject.duckdns.org:9090/>

GRAFANA : <http://airlineproject.duckdns.org:3001/>

DASHBOARD : <http://airlineproject.duckdns.org:8050/>



Conclusion

objectif 1



Récolte et
consommation de
données

Base de données riche

objectif 2



Déploiement

Un accès opérationnel

Objectif 3



Automatisation et
monitoring

Une surveillance continue

Améliorations possibles:

Données en entrée: test et ajout éventuel de paramètres additionnels (météo, saisonnalité, etc.)

Ergonomie: Intégration du dataset “Flight Schedule” pour saisir des paramètres de prédiction de vols futurs

...