Rapport de projet – Infrastructure Data sur le Cloud

# 1. Contexte et objectifs

Ce projet vise à mettre en œuvre une infrastructure de gestion de données moderne dans le Cloud, incluant un Data Lake, un Data Warehouse, un pipeline ETL, ainsi que des visualisations interactives. L’objectif est de collecter et analyser des données d’annonces d’emploi dans le domaine de la Data Science.

# 2. Tâches réalisées

- Création d’un repository GitHub pour centraliser le projet.  
- Construction d’un pipeline ETL en Python pour la collecte et la transformation des données.  
- Génération manuelle d’un fichier data.csv simulant des données d’annonces (en raison de difficultés avec les APIs).  
- Nettoyage et structuration des données dans un notebook Jupyter (exploration.ipynb).  
- Stockage des données dans un fichier CSV, simulant un Data Lake.  
- Création d’un tableau de bord interactif avec Streamlit (app.py).  
- Génération de visualisations interactives à partir des données nettoyées.  
- Rédaction d’un README complet et d’un fichier requirements.txt.

# 3. Difficultés rencontrées

- L’API France Travail (anciennement Pôle Emploi) nécessite une autorisation d’accès complexe. En conséquence, j’ai dû générer des données simulées réalistes manuellement dans un fichier CSV.  
- Quelques problèmes d’environnement avec Python et Streamlit sur Windows, résolus via PowerShell et Git Bash.  
- Difficulté à déployer l’application via Streamlit Cloud à cause d’un bug temporaire (erreur 1101).

# 4. Solutions mises en place

- Utilisation de données simulées réalistes pour respecter les objectifs de collecte, nettoyage et visualisation.  
- Tests et exécution du dashboard en local avec Streamlit.  
- Documentation de toutes les étapes dans un README clair et structuré.

# 5. Conclusion

Ce projet m’a permis de comprendre les principaux composants d’une infrastructure data moderne : collecte, transformation, stockage et visualisation. Malgré certaines contraintes techniques, le résultat final respecte les objectifs fixés, et démontre une bonne maîtrise des outils de la chaîne de données.