Projet Data Engineering : Pipeline ELT

Johannes AFOUDAH, Mehdi BEN CHEIKH, Ashwin DEVADEVAN, Théo CHANNAROND

Sommaire

Objectifs du Projet

- Mise en place pipeline ELT complet
- Automatisation intégration et transformation
- Faciliter accès et analyse pour métiers
- Réduire temps traitement des données
- Améliorer qualité des données
- Travailler en équipe
- Versionner le projet avec GitHub

Présentation du dataset

🗂️ Source des données

- Fichier CSV : `weatherHistory.csv`
- Taille : ~96 000 lignes
- Période couverte : **2006 à 2016**
- Fréquence : **horaires**
- Lieu : **San Francisco, CA (USA)**

🔍 Contenu du dataset

Chaque ligne représente une **observation météo** à un instant donné.
Le fichier contient notamment :

Colonne	Description
Formatted Date	Date et heure de la mesure (ISO 8601)
Summary	Résumé météo général (ex. "Partly Cloudy")
Precip Type	Type de précipitations (rain, snow)
Temperature (C)	Température en Celsius
Apparent Temperature (C)	Température ressentie
Humidity	Taux d'humidité (0 à 1)
Wind Speed (km/h)	Vitesse du vent
Wind Bearing (degrees)	Direction du vent
Visibility (km)	Distance de visibilité
Cloud Cover	Taux de couverture nuageuse
Pressure (millibars)	Pression atmosphérique
Daily Summary	Résumé météo de la journée

💡 Qualités du dataset

- Données **réelles** et **riches**
- Permet des analyses **temporelles**, **climatiques**, **comparatives**
- Bon point de départ pour de la **modélisation**, **visualisation**, ou **prévision météo**

Comprendre le principe ELT

1

Extract

Extraire les données via API, scripts python, web, etc.

2

Load

Insertion des données transformées dans la base cible (PostgreSQL).

3

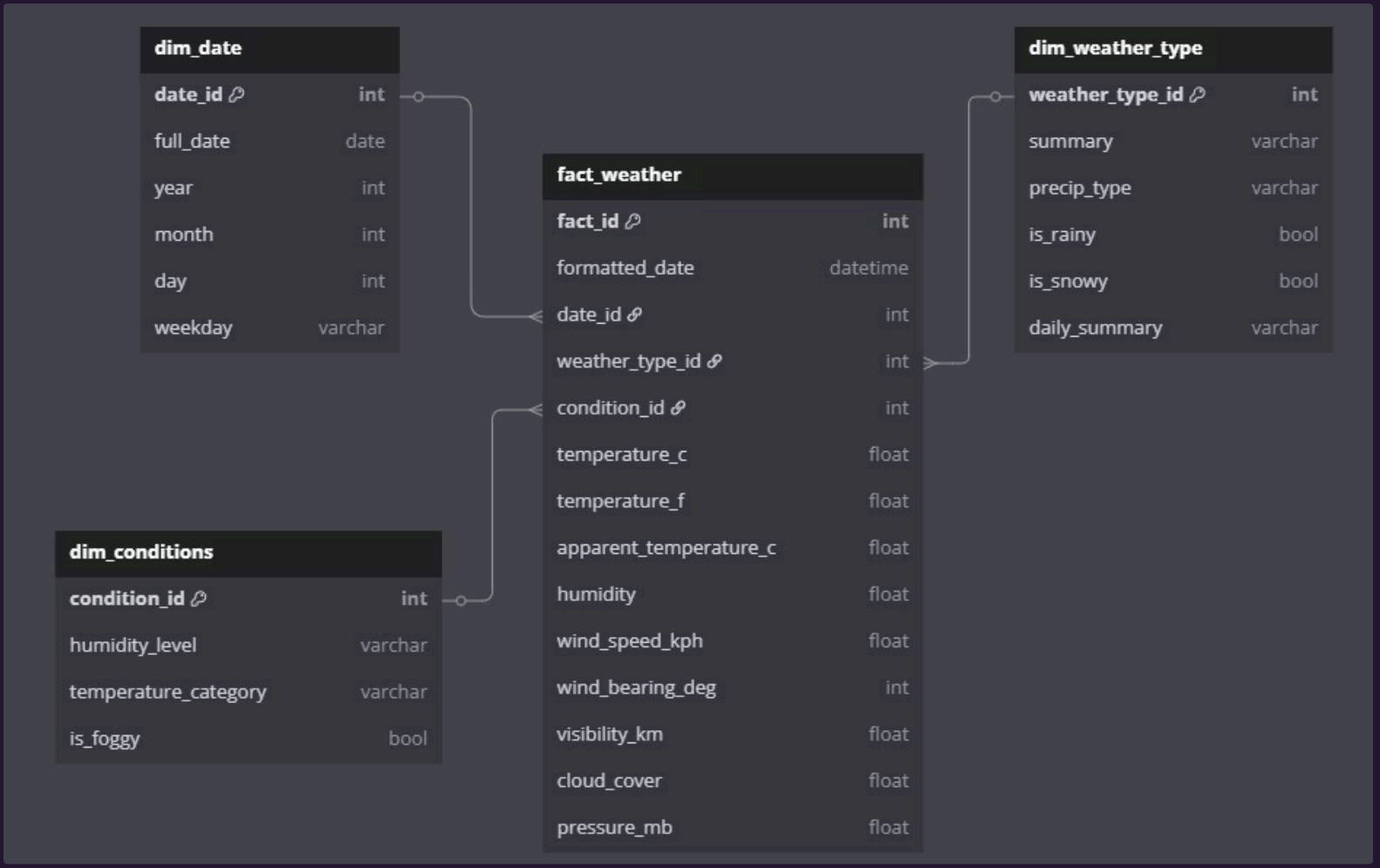
Transform

Transformer les données afin de les nettoyer, les enrichir et les structurer.

Orchestration du Pipeline

Schéma du pipeline :

Schéma en étoile de la base de donnée



Organisation de l'équipe

- Théo CHANNAROND : Gestion du scripts SQL et python pour chargement, nettoyage et enrichissement des données.
- Ashwin DEVADEVAN : Réalisation du schéma en étoile de la base de donnée.
- Mehdi BEN CHEIKH : Réalisation du schéma de la pipeline et de la présentation.
- Johannes AFOUDAH : Gestion du scripts SQL et python pour l'enrichissement des données et séparation du dataset en table de fait et de dimensions.

Difficultés Rencontrées

Difficulté à connecter le script python à PostgreSQL.

Difficulté à faire fonctionner PostgreSQL sur toutes les machines.

Perspectives d'améliorations et conclusion

- **Optimisation SQL :** Refonte des requêtes pour accélérer les traitements.
- **Parallélisation :** Exécution simultanée des tâches pour gagner en rapidité.
- **Monitoring renforcé :** Surveillance continue pour détecter et corriger rapidement.
- **Gestion des erreurs :** Mise en place de procédures automatiques de reprise.

Merci pour votre écoute !