Sujet de TER Master 2 Informatique

Parcours Sciences des données 2025/26

OpenDataHub SQL : la data publique française à portée du SQL

Encadrant : Jérémy GROS (jeremy.gros.dev@gmail.com)

Avec plusieurs années d'expérience dans l'ingénierie data et l'infrastructure, je serai votre coach pour vous accompagner tout au long de ce projet.

Taille attendue du groupe : 3 étudiants

Contexte du projet

Bienvenue dans LE projet de data engineering/ops de cette année !

Data.gouv.fr propose des milliers de datasets publics hétérogènes (CSV, JSON, ZIP, GeoJSON…) sur des sujets variés : transport, météo, finances, santé, mobilité…

Votre mission : transformer ces fichiers bruts en données SQL historisées et fiables, prêtes à être analysées.

C'est un projet avant tout pédagogique : vous serez guidés sur tous les outils que vous n'avez jamais utilisés, et chaque étape sera apprise en pratique, avec mon accompagnement direct sur :

- L'ingestion massive et hétérogène ;
- La mise en place d'un orchestrateur;
- La mise en place d'un lac de données;
- La mise en place d'un entrepôt de données.

Objectifs du projet

- Collecter dynamiquement les datasets via l'API data.gouv.fr ;
- Historiser les fichiers bruts dans un datalake, partitionnés par dataset et date d'extraction ;
- Transformer et modéliser les données (dynamiquement) dans un entrepôt de données ;
- Mettre en place des tableaux de bord pour superviser les datasets synchronisés ;
- Automatiser et superviser l'ensemble : pipelines Airflow, alerting, tests, monitoring et CI/CD ;
- Construire des modèles d'analyse de données (STAR) pour faciliter la réutilisation des données.

Résultats attendus

Pipeline complet : ingestion → historisation → transformation → exposition

- Ingestion dynamique depuis l'API data.gouv.fr ;
- Transformation et modélisation analytique avec tests et documentation ;
- Dashboard de supervision.

Lac et Entrepôt de données :

- Mise en place d'un lac de données avec une stratégie de partitionnement ;
- Mise en place de l'entrepôt de données avec une architecture médaillon.

Architecture solide et documentée :

- Schéma global de la plateforme, de la collecte à l'exposition ;
- Description des choix technologiques ;
- Documentation de la gouvernance et de la qualité des données ;
- Plan de scalabilité et résilience : comment le pipeline peut gérer plus de datasets, augmenter la volumétrie et rester stable.

Code et déploiement :

- Repos Git organisés ;
- Plateforme conteneurisée ;
- Scripts Terraform/Helm pour déploiement de l'infrastructure.

L'objectif : que l'architecture soit robuste, claire et réutilisable, pas seulement un prototype fonctionnel.

Références

- https://www.data.gouv.fr/ (API Data Gouv)
- https://airflow.apache.org/ (Orchestration)
- https://www.getdbt.com/ (Modélisation)
- https://www.snowflake.com/fr/ (Entrepôt de données)
- https://iceberg.apache.org/ (Format de table)
- https://duckdb.org/
- https://k3s.io/ + https://www.rancher.com/products/k3s (Orchestration de conteneurs)
- https://www.docker.com/ (Conteneur)

Possibilité de continuer en Stage : Non (Malheureusement)

Contraintes à respecter

S'amuser, apprendre et LIVRER !

Informations supplémentaires

- Ce projet est conçu pour apprendre en pratique, avec un accompagnement complet sur tous les outils.
- Possibilité d'ajouter une partie IA/ML (optionnel).
- Projet itératif : commencer avec quelques datasets pilotes, puis scaler.
- Ambiance projet de fou : projet complet, le tout guidé par votre mentor pour vous faire monter en compétence sur toute la stack moderne.

Pour exploiter pleinement ce projet, il est recommandé de constituer une équipe avec des personnes orientées data engineering/ops ou souhaitant se spécialiser dans ce domaine. Les profils data science / BI peuvent participer, mais le projet est surtout axé sur la mise en place de pipelines, la transformation et la gestion des données plutôt que sur l'analyse ou le machine learning.

Exemple de roadmap

Phase 1 – Mi-novembre à fin novembre : Prise en main et définition

- Découverte de l'API data.gouv.fr et exploration de quelques datasets pilotes (CSV, JSON);
- Organisation de l'équipe (gestion de projet);
- Définition du périmètre initial;
- Conception de l'architecture;
- Mise en place des dépôts Git;
- Installation et configuration de la plateforme;

Livrable : pipelines simples d'ingestion + stockage brut (local).

Phase 2 – Début décembre à début janvier : Stockage et Transformation

- Automatisation de l'ingestion via Airflow pour les datasets pilotes;
- Stockage dans le lac de données;
- Conception de l'entrepôt de données (médaillons Bronze/Silver/Gold);
- Début de la transformation avec dbt : modèles simples et tests de qualité;
- Mise en place de scripts de supervision simples (logs Airflow + alertes basiques).

Livrable : pipelines ingestion → lac de données → entrepôt de données → transformations dbt avec tests unitaires.

Phase 3 – Début janvier à fin janvier : Entrepôt et architecture analytique

- Ajout d'autres formats de datasets;
- Modélisation analytique STAR pour les datasets pilotes;
- Dashboard minimal pour suivre l'état des pipelines et des datasets.

Livrable : pipelines complet pour quelques types de format de données, dashboard de supervision.

Phase 4 – Début février à mi-février : Mise à l'échelle, finalisation et rédaction

- Ajout d'autres formats de datasets;
- Optimisation des performances et tests de scalabilité;
- Mise en place d'un plan de résilience et recommandations pour production;
- Rédaction du rapport final et préparation de la soutenance.

Livrable : pipelines robuste, architecture complète et prête à évoluer, documentation et rapport final prêts pour la soutenance.