Rapport Technique — JobTech Data Lake & API

Mamadou Diarrassouba
Kouakou Adahe
Tedongmo Gangnimaze Nathanael

Challenge : Cartographie du marché de l'emploi Tech en Europe

1. Contexte et Objectif

Ce projet s'inscrit dans le cadre du challenge européen Talent Insight, visant à cartographier l'offre et la demande d'emplois Tech en Europe.

Objectifs principaux :

- Centraliser et structurer les données issues de multiples sources
- Mettre en place une chaîne de traitement automatisée (pipeline)
- Fournir une API REST pour l'accès à des indicateurs clés (salaires, technologies, diversité, tendances)

2. Sources de Données

Le projet s'appuie sur **6 principales sources de données** :

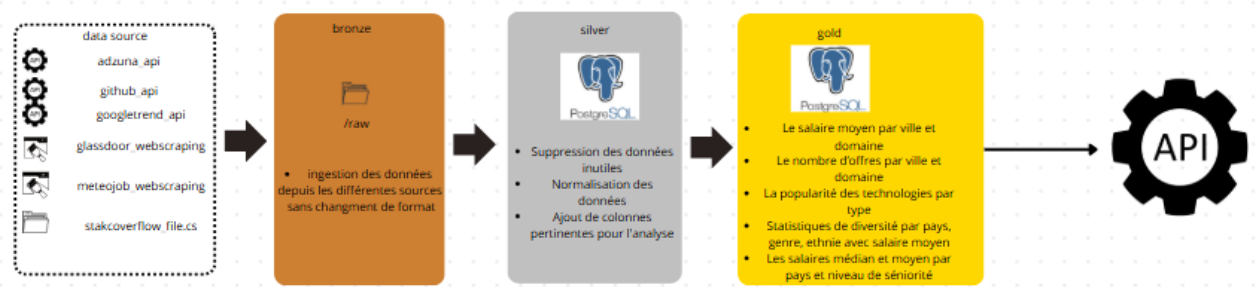
- **Sites d'emploi** : Meteojob, Adzuna, Glassdoor
- **Stack Overflow Developer Survey** : données sur les développeurs (profil, techno, salaires, diversité)
- **GitHub** : projets publics, langages, géolocalisation des repos

- **Google Trends** : popularité des technologies par pays et dans le temps

3. Architecture Pipeline de Données

L'architecture suit une logique par **couches** (Data Lake → SilverDB → GoldDB → API) :

1. **Datalake brut** (raw/) : stockage initial des fichiers JSON/CSV
2. **SilverDB** : base PostgreSQL contenant les données nettoyées
3. **GoldDB** : base PostgreSQL contenant les **datamarts**
4. **API Django REST** : couche d'exposition aux consommateurs



4. Étapes du Pipeline de Traitement

L'ensemble de la pipeline est automatisé via un script `main.py` :

1. **Scraping hebdomadaire** des sources (emplois, GitHub, StackOverflow, Trends)
2. **Nettoyage & transformation** : typage, traitement des salaires, séparation des colonnes multi-valeurs…
3. **Chargement dans SilverDB** : staging de données propres
4. **Construction des datamarts** dans GoldDB via agrégation
5. **Démarrage de l'API** Django pour consultation en temps réel

5. Datamarts construits

Nom du datamart	Description
`datamart_rh_market`	Salaires moyens et volume d'offres par ville/domaine

`datamart_tech_popularity`	Popularité des technos (langages, outils, bases de données…)
`datamart_diversity_conditions`	Analyse par pays de la diversité (genre, accessibilité, remote)
`datamart_salary_by_country_seniority`	Salaire moyen par pays et niveau de séniorité

6. Endpoints exposés (API Django REST)

Tous les endpoints sont disponibles sous `/api/` :

- **RH & Marché de l'emploi**
  - `/api/rh/salaire/?ville=Paris&domaine=Data`
  - `/api/rh/offres/?ville=Lyon&domaine=DevOps`
- **Popularité des technologies**
  - `/api/tech/popularity/?type=Language`
- **Diversité et conditions de travail**
  - `/api/diversity/?country=Germany`
- **Salaire par pays et séniorité**
  - `/api/salary/by-country-seniority/?country=France`

documentation api

Data Lake Challenge
https://documenter.getpostman.com/view/15408096/2sB34bMQ2Y