

FraudLens

Healthcare Fraud Detection Platform - Documentation Technique

Vue d'ensemble

Pipeline de données complet pour la détection de fraudes dans le secteur de la santé. Ingre et centralise 10+ datasets publics CMS/OIG pour détecter des anomalies dans les paiements Medicare/Medicaid.

Architecture du Pipeline

```
master_dag (Orchestrateur)
|
+--- init_snowflake_environment (Bootstrap)
|     +--- create_warehouse (FRAUDLENS_WH)
|     +--- create_schemas (RAW_DATA, BRONZE, STAGING, SILVER, GOLD)
|     +--- create_s3_stage
|
+--- [Data Downloads - En parallele]
|     +--- leie_download
|     +--- medicare_hospital_spending_download
|     +--- open_payments_download
|     +--- provider_information_download
|     +--- longterm_care_hospital_download
|     +--- hospice_download
|     +--- home_health_care_download
|     +--- medicare_part_d_prescribers_download
|
+--- load_bronze_tables (COPY INTO Snowflake)
|
+--- dbt_transformations
|     +--- dbt run --select staging
|     +--- dbt run --select silver
|     +--- dbt run --select gold
|     +--- dbt test
```

+-- dbt docs generate

Sources de Donnees

1. LEIE (Excluded Individuals / Entities)

Attribut	Valeur
Source	OIG (Office of Inspector General)
Contenu	Liste des individus et entites exclus du programme Medicare/Medicaid
Usage Fraude	Detecter paiements vers des prestataires deja exclus
Cle	NPI, Nom

2. Medicare Hospital Spending by Claim

Attribut	Valeur
Source	CMS Provider Data
Contenu	Paiements Medicare par hopital et type de prestation
Usage Fraude	Identifier depenses anormales par etablissement
Cle	Facility ID

3. Open Payments (2024)

Attribut	Valeur
----------	--------

Source	CMS Open Payments
Contenu	Paielements des laboratoires pharmaceutiques aux medecins
Usage Fraude	Detecter conflits d'interets, kickbacks
Cle	NPI

4. Provider Information (Nursing Home)

Attribut	Valeur
Source	CMS Provider Data
Contenu	Informations detaillees sur les prestataires
Usage Fraude	Enrichissement et validation
Cle	NPI, Provider ID

5. Long-Term Care Hospital

Attribut	Valeur
Source	CMS Provider Data
Contenu	Details des hopitaux de soins prolonges
Usage Fraude	Secteur a risque
Cle	Provider ID

6. Hospice

Attribut	Valeur
----------	--------

Source	CMS Provider Data
Contenu	Informations sur les etablissements de soins palliatifs
Usage Fraude	Secteur a haut risque de fraude Medicare
Cle	Provider ID

7. Home Health Care

Attribut	Valeur
Source	CMS Provider Data
Contenu	Prestataires de soins a domicile avec ZIP Codes
Usage Fraude	Analyses geospatiales
Cle	ZIP Code

8. Medicare Part D Prescribers

Attribut	Valeur
Source	CMS Medicare Part D
Contenu	Prescriptions par medecin
Usage Fraude	Detecter les sur-prescripteurs, pill mills
Cle	NPI

9. NPPES Provider Data

Attribut	Valeur
----------	--------

Source	Snowflake Marketplace (Affine)
Contenu	Referentiel complet de tous les NPI aux Etats-Unis
Usage Fraude	Table de reference maitre pour les jointures
Cle	NPI

Medallion Architecture

L'architecture Medallion organise les donnees en couches progressives de qualite et d'enrichissement.

Schema	Description
BRONZE	Donnees brutes depuis S3 Stage
STAGING	Vues dbt sur Bronze
SILVER	Donnees nettoyees, jointes, enrichies
GOLD	Donnees prêts pour l'analyse et BI

Gold Layer Models

Model	Description
provider_360	Vue 360 complete de chaque provider
payments_summary	Agregations des paiements pharma
prescriptions_summary	Agregations des prescriptions

fraud_risk_score	Score de risque fraude (0-100)
high_risk_alerts	Alertes actionnables

Dashboard Streamlit

Dashboard interactif pour la visualisation des donnees de fraude :

- **Overview** : KPIs executifs, distribution des risques
- **Fraud Alerts** : Liste des alertes avec filtrage et export
- **Provider 360** : Recherche et profil complet de provider
- **Analytics** : Cartes geographiques, analyses de tendances

Lancer le Dashboard

```
cd dashboard
source venv/bin/activate
streamlit run Home.py
```

Stack Technique

Composant	Technologie
Orchestration	Apache Airflow 3.x
Data Warehouse	Snowflake
Data Lake	Amazon S3
Transformations	dbt Core

Dashboard	Streamlit + Plotly
Conteneurisation	Docker + Docker Compose
Traitement	Python 3.12, Pandas, PyArrow

Utilisation

Demarrer l'environnement

```
docker-compose up -d
```

Lancer le pipeline complet

```
docker exec fraudlens-airflow-worker-1 airflow dags trigger master_dag
```

Acceder a l'interface Airflow

Parametre	Valeur
URL	http://localhost:8080
User	airflow
Password	airflow