

Infrastructure Data Warehouse

Vue d'ensemble

Snowflake est le Data Warehouse central du projet FraudLens. Il stocke et traite les données de santé Medicare/Medicaid selon une architecture Medallion (Bronze → Silver → Gold).

Architecture Snowflake

```
FRAUDLENS_DB (Database)
├── RAW_DATA (Schema)      ← Données brutes depuis S3
├── BRONZE (Schema)        ← Tables Parquet chargées
│   ├── LEIE
│   ├── MEDICARE_HOSPITAL_SPENDING
│   ├── PROVIDER_INFORMATION
│   ├── LONGTERM_CARE_HOSPITAL
│   ├── HOSPICE
│   ├── HOME_HEALTH_CARE
│   ├── MEDICARE_PART_D_PRESCRIBERS
│   ├── OPEN_PAYMENTS_GENERAL
│   ├── OPEN_PAYMENTS_RESEARCH
│   └── OPEN_PAYMENTS_OWNERSHIP
├── STAGING (Schema)      ← Vues dbt sur Bronze
├── SILVER (Schema)        ← Tables transformées
└── GOLD (Schema)         ← Tables analytiques
    ├── PROVIDER_360
    ├── PAYMENTS_SUMMARY
    ├── PRESCRIPTIONS_SUMMARY
    ├── FRAUD_RISK_SCORE
    └── HIGH_RISK_ALERTS
```

Objets Snowflake

Warehouse

Paramètre	Valeur
Nom	FRAUDLENS_WH
Taille	XSMALL
Type	STANDARD
Auto Suspend	300 secondes (5 min)
Auto Resume	TRUE
Initially Suspended	TRUE

```
CREATE OR REPLACE WAREHOUSE FRAUDLENS_WH
  WAREHOUSE_SIZE = 'XSMALL'
  WAREHOUSE_TYPE = 'STANDARD'
  AUTO_SUSPEND = 300
  AUTO_RESUME = TRUE
  INITIALLY_SUSPENDED = TRUE;
```

Database & Schemas

Schema	Description	Usage
RAW_DATA	Données brutes depuis S3	Landing zone initiale
BRONZE	Données après ingestion	Tables Parquet chargées
STAGING	Staging area dbt	Vues sur Bronze
SILVER	Données transformées	Tables nettoyées, jointes
GOLD	Données finales	Tables pour BI et analyse

Intégration S3

Les données sont stockées sur Amazon S3 et chargées dans Snowflake via un Stage externe.

Storage Integration

```
CREATE OR REPLACE STORAGE INTEGRATION S3_INTEGRATION
  TYPE = EXTERNAL_STAGE
  STORAGE_PROVIDER = 'S3'
  ENABLED = TRUE
  STORAGE_AWS_ROLE_ARN = 'arn:aws:iam::XXXX:role/snowflake-role'
  STORAGE_ALLOWED_LOCATIONS = ('s3://ai-factory-bckt/');
```

External Stage

```
CREATE OR REPLACE STAGE BRONZE_S3_STAGE
  STORAGE_INTEGRATION = S3_INTEGRATION
  URL = 's3://ai-factory-bckt/bronze/'
  FILE_FORMAT = PARQUET_FORMAT;
```

File Format

```
CREATE OR REPLACE FILE FORMAT PARQUET_FORMAT
  TYPE = PARQUET
  COMPRESSION = AUTO;
```

Tables Bronze

Les tables Bronze contiennent les données brutes chargées depuis les fichiers Parquet sur S3.

LEIE - Excluded Individuals/Entities

Colonne	Type	Description
NPI	VARCHAR	National Provider Identifier
LASTNAME, FIRSTNAME	VARCHAR	Nom du provider
BUSNAME	VARCHAR	Nom de l'entreprise
SPECIALTY	VARCHAR	Spécialité médicale
EXCLTYPE	VARCHAR	Type d'exclusion
EXCLDATE	VARCHAR	Date d'exclusion
STATE	VARCHAR	État
_LOAD_TIMESTAMP	TIMESTAMP_NTZ	Timestamp de chargement

MEDICARE_HOSPITAL_SPENDING

Colonne	Type	Description
FACILITY_ID	VARCHAR	Identifiant de l'établissement
FACILITY_NAME	VARCHAR	Nom de l'établissement
CLAIM_TYPE	VARCHAR	Type de réclamation
AVG_SPENDING_PER_EPISODE_HOSPITAL	VARCHAR	Dépense moyenne par épisode
STATE	VARCHAR	État

MEDICARE_PART_D_PRESCRIBERS

Colonne	Type	Description
PRSCRBR_NPI	VARCHAR	NPI du prescripteur
PRSCRBR_LAST_ORG_NAME	VARCHAR	Nom du prescripteur
BRND_NAME	VARCHAR	Nom de marque du médicament
GNRC_NAME	VARCHAR	Nom générique
TOT_DRUG_CST	VARCHAR	Coût total
TOT_CLMS	VARCHAR	Nombre de claims

OPEN_PAYMENTS_GENERAL

Colonne	Type	Description
Covered_Recipient_NPI	VARCHAR	NPI du bénéficiaire
Total_Amount_of_Payment_USDollars	VARCHAR	Montant total
Nature_of_Payment_or_Transfer_of_Value	VARCHAR	Nature du paiement
Applicable_Manufacturer_*_Name	VARCHAR	Nom du payeur

NPPES (Snowflake Marketplace)

Source: AFFINE_NPPES_PROVIDER_DATA.REF_DW

Table	Description	Clé
DIM_PROVIDER	Provider master data (7M+ NPIs)	NPI
DIM_PROVIDER_ADDRESS	Adresses des providers	NPI
DIM_PROVIDER_TAXONOMY	Spécialités des providers	NPI
REF_TAXONOMY_CODE	Référentiel taxonomie	TAXONOMY_CODE

Authentication

Méthode: Clé RSA (key-pair authentication)

- Airflow: Clé privée montée dans `/opt/airflow/snowflake/keys/rsa_key.p8`

- GitHub Actions: Secret `SNOWFLAKE_PRIVATE_KEY`
- Streamlit Cloud: Mot de passe dans Secrets