AI Factory - dbt Documentation

Guide de Transformation et Detection de Fraude

Version: 2.0.0 **Date:** Fevrier 2026

Table des Matieres

1. Introduction

1.1 Objectif

Le projet dbt transforme les donnees brutes (Bronze) en donnees analytiques pretes a l'emploi pour la detection de fraude medicale. L'architecture comprend trois couches:

- **Staging**: Vues 1:1 sur les sources
- **Silver**: Donnees nettoyees avec flags de fraude
- **Gold**: Agregations metier et scoring de risque

1.2 Volumes de Donnees

Couche Silver

Table	Lignes	Description
SILVER.PROVIDER	17,897,600	Provider master enrichi
SILVER.EXCLUDED_PROVIDERS	90,316	Providers exclus LEIE
SILVER.PAYMENTS	16,158,446	Paiements Open Payments
SILVER.PRESCRIPTIONS	1,332,309	Prescriptions Part D
SILVER.FACILITIES	19,221	Etablissements de sante
SILVER.HOSPITAL_SPENDING	64,658	Depenses hospitalieres
TOTAL SILVER	**35,562,550**	

Couche Gold NEW

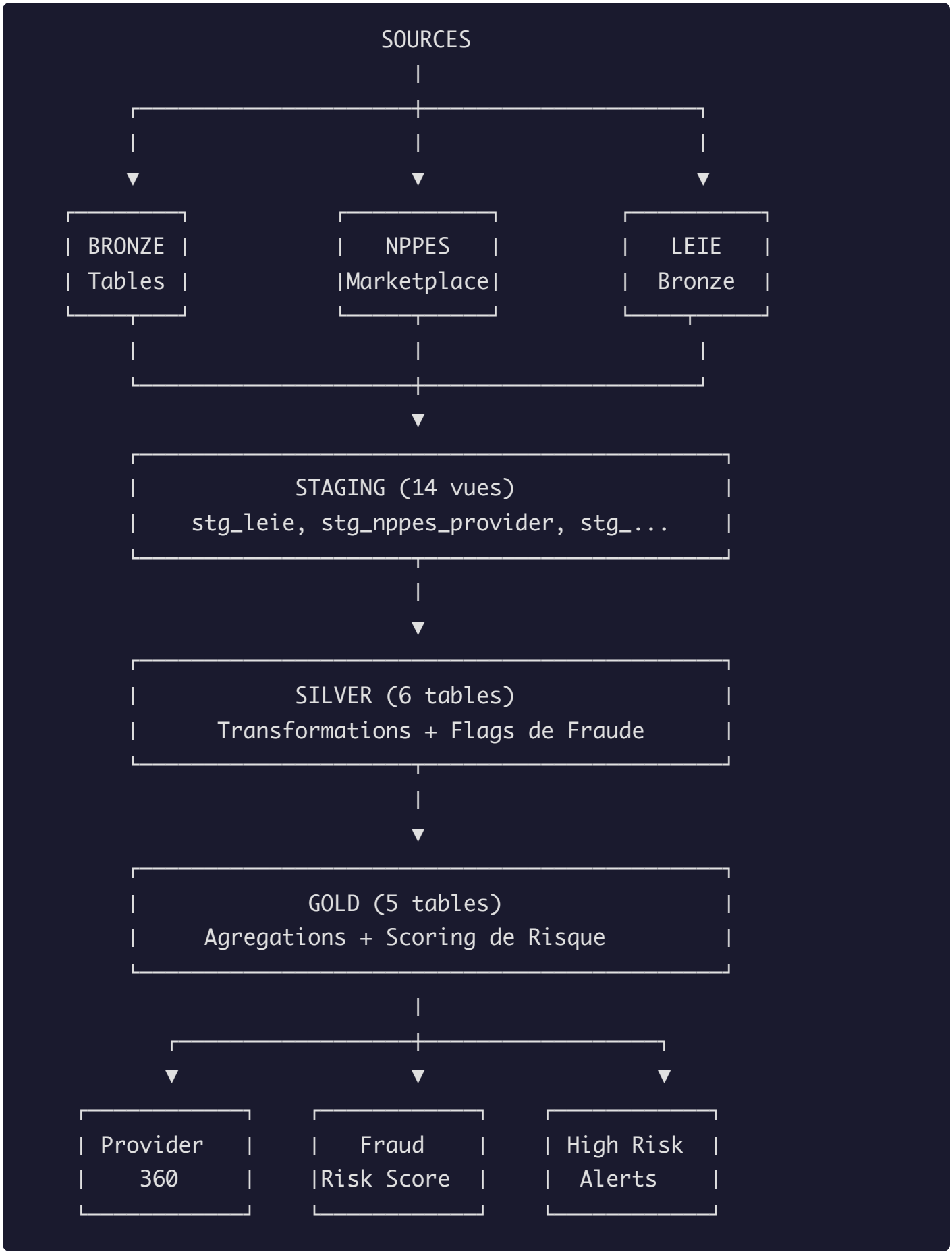
Table	Description
GOLD.PAYMENTS_SUMMARY	Agregations paiements par provider
GOLD.PRESCRIPTIONS_SUMMARY	Metriques prescriptions par provider
GOLD.PROVIDER_360	Vue 360° complete du provider
GOLD.FRAUD_RISK_SCORE	Score de risque fraude (0-100)
GOLD.HIGH_RISK_ALERTS	Alertes a investiguer

2. Architecture dbt

2.1 Structure du Projet

```
dbt/
├── dbt_project.yml          # Configuration projet
├── profiles.yml             # Connexion Snowflake
├── packages.yml             # Dependencies (dbt_utils)
|
├── macros/
|   └── get_custom_schema.sql # Macro pour noms de schemas exacts
|
├── models/
|   ├── staging/             # Vues 1:1 sur Bronze + NPPES
|   |   ├── sources.yml      # Declaration des sources
|   |   ├── schema.yml       # Documentation staging
|   |   └── stg_*.sql        # 14 modeles staging
|   |
|   ├── silver/              # Tables transformees
|   |   ├── schema.yml       # Tests et documentation
|   |   ├── provider.sql
|   |   ├── excluded_providers.sql
|   |   ├── payments.sql
|   |   ├── prescriptions.sql
|   |   ├── facilities.sql
|   |   └── hospital_spending.sql
|   |
|   └── gold/                # Agregations metier NEW
|       ├── payments_summary.sql
|       ├── prescriptions_summary.sql
|       ├── provider_360.sql
|       ├── fraud_risk_score.sql
|       └── high_risk_alerts.sql
|
└── target/                  # Fichiers generes
```

2.2 Flux de Donnees (DAG)



2.3 Materialisation

Couche	Type	Schema	Raison
Staging	VIEW	STAGING	Pas de stockage, lecture directe
Silver	TABLE	SILVER	Performance, donnees transformees
Gold	TABLE	GOLD	Agregations pour dashboards

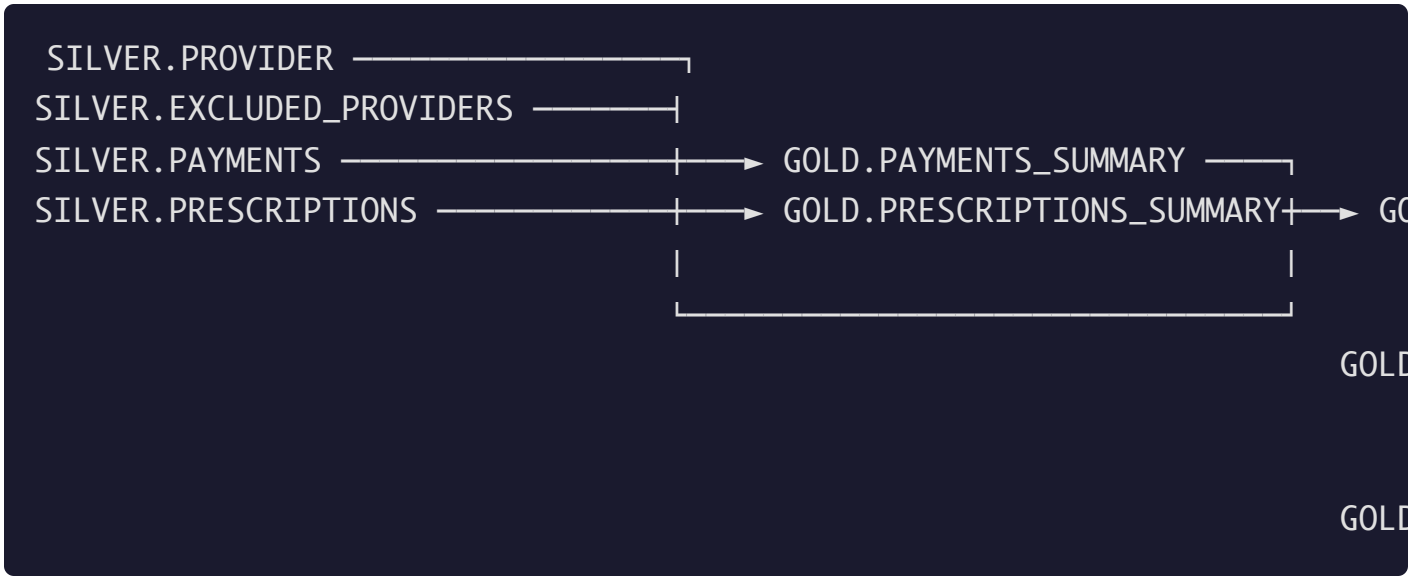
8. Couche Gold NEW

8.1 Objectif

La couche Gold fournit des **tables analytiques pretes a l'emploi** pour:

- Dashboards de detection de fraude
- Scoring de risque par provider
- Alertes automatiques pour investigation
- Vue 360° des providers

8.2 Dependances



8.3 Configuration dbt

```yaml
# dbt_project.yml
models:
  ai_factory:
    staging:
      +schema: STAGING
      +materialized: view
    silver:
      +schema: SILVER
      +materialized: table
    gold:
      +schema: GOLD
      +materialized: table
```

9. Modeles Gold en Detail NEW

9.1 GOLD.PAYMENTS_SUMMARY

Objectif: Agreger les paiements pharma par provider.

Colonnes principales:

Colonne	Type	Description
NPI	VARCHAR	National Provider Identifier
TOTAL_PAYMENTS	NUMBER	Nombre total de paiements
TOTAL_PAYMENT_AMOUNT	NUMBER	Montant total recu ($)
GENERAL_PAYMENT_AMOUNT	NUMBER	Montant paiements generaux
RESEARCH_PAYMENT_AMOUNT	NUMBER	Montant paiements recherche
UNIQUE_PAYERS	NUMBER	Nombre de payeurs distincts
VERY_HIGH_RISK_PAYMENTS	NUMBER	Paiements >= $100K

PCT_HIGH_RISK_PAYMENTS	NUMBER	% paiements a haut risque
RECIPIENT_TIER	VARCHAR	Classification du destinataire

RECIPIENT_TIER - Niveaux:

Tier	Seuil	Description
MEGA_RECIPIENT	>= $1,000,000	Destinataire majeur, audit prioritaire
MAJOR_RECIPIENT	>= $100,000	Destinataire important
SIGNIFICANT_RECIPIENT	>= $10,000	Destinataire significatif
MODERATE_RECIPIENT	>= $1,000	Destinataire modere
MINOR_RECIPIENT	< $1,000	Destinataire mineur

9.2 GOLD.PRESCRIPTIONS_SUMMARY

Objectif: Agreger les prescriptions Part D par provider.

Colonnes principales:

Colonne	Type	Description
NPI	VARCHAR	National Provider Identifier
PRESCRIBER_NAME	VARCHAR	Nom du prescripteur
TOTAL_CLAIMS	NUMBER	Nombre total de prescriptions
TOTAL_COST	NUMBER	Cout total des prescriptions ($)
UNIQUE_DRUGS_PRESCRIBED	NUMBER	Medicaments distincts prescrits
BRAND_CLAIMS	NUMBER	Prescriptions de marque

PCT_BRAND_CLAIMS	NUMBER	% prescriptions de marque
TOTAL_HIGH_RISK_DRUGS	NUMBER	Medicaments a haut risque
PRESCRIBER_VOLUME_TIER	VARCHAR	Classification volume

PRESCRIBER_VOLUME_TIER - Niveaux:

Tier	Seuil (claims)	Description
VERY_HIGH_VOLUME	>= 100,000	Volume tres eleve
HIGH_VOLUME	>= 10,000	Volume eleve
MEDIUM_VOLUME	>= 1,000	Volume moyen
LOW_VOLUME	< 1,000	Volume faible

9.3 GOLD.PROVIDER_360

Objectif: Vue complete a 360° de chaque provider, combinant toutes les sources.

> **Cas d'usage:** Table de reference pour les investigateurs. Permet de voir en un coup d'oeil toutes les informations et metriques d'un provider.

Sources jointes:

- `SILVER.PROVIDER` - Identite et localisation
- `SILVER.EXCLUDED_PROVIDERS` - Statut d'exclusion
- `GOLD.PAYMENTS_SUMMARY` - Metriques paiements
- `GOLD.PRESCRIPTIONS_SUMMARY` - Metriques prescriptions

Colonnes principales:

Categorie	Colonnes
Identite	NPI, ENTITY_TYPE, FULL_NAME, ORGANIZATION_NAME, CREDENTIAL

Localisation	CITY, STATE, ZIP_CODE, PHONE
Specialite	SPECIALTY_CLASSIFICATION, SPECIALTY, PROVIDER_TYPE
Exclusion	IS_EXCLUDED, EXCLUSION_REASON, EXCLUSION_DATE, IS_EXCLUDED_HIGH_RISK
Paiements	TOTAL_PAYMENTS, TOTAL_PAYMENT_AMOUNT, RECIPIENT_TIER
Prescriptions	TOTAL_PRESCRIPTION_CLAIMS, TOTAL_PRESCRIPTION_COST, PCT_BRAND_CLAIMS
Activite	HAS_PHARMA_PAYMENTS, HAS_PRESCRIPTIONS
Financier	TOTAL_FINANCIAL_EXPOSURE

10. Scoring de Fraude NEW

10.1 GOLD.FRAUD_RISK_SCORE

Objectif: Calculer un score de risque composite (0-100) pour chaque provider.

> **Methodologie:** Le score combine plusieurs signaux de risque avec des poids differents. Plus le score est eleve, plus le provider est suspect.

Composantes du Score:

Categorie	Signal	Points
Exclusion (40 pts max)	IS_EXCLUDED = true	+25
	IS_EXCLUDED_HIGH_RISK = true (exclu mais actif)	+15
Paiements (20 pts max)	MEGA_RECIPIENT ou MAJOR_RECIPIENT	+5 a +10

	PCT_HIGH_RISK_PAYMENTS >= 50%	+10
Prescriptions (20 pts max)	PCT_BRAND_CLAIMS >= 80%	+10
	TOTAL_HIGH_RISK_DRUGS >= 10	+10
Anomalies (20 pts max)	HAS_PHARMA_PAYMENTS AND HAS_PRESCRIPTIONS	+10
	NPI inactif mais activite detectee	+10

RISK_TIER - Classification:

Tier	Score	Action recommandee
CRITICAL	>= 70	Investigation immediate
HIGH	50-69	Revue prioritaire
MEDIUM	30-49	Surveillance renforcee
LOW	10-29	Monitoring standard
MINIMAL	< 10	Pas d'action requise

Flags de risque:

- `IS_EXCLUDED` - Provider sur liste LEIE
- `IS_EXCLUDED_HIGH_RISK` - Exclu mais NPI encore actif
- `IS_HIGH_BRAND_PRESCRIBER` - >= 80% prescriptions de marque
- `IS_MAJOR_PAYMENT_RECIPIENT` - >= $100K de paiements pharma

10.2 GOLD.HIGH_RISK_ALERTS

Objectif: Generer des alertes actionnables pour investigation.

> **Priorite:** Chaque alerte represente un cas suspect necessitant une revue humaine. Les alertes sont classees par priorite.

Types d'alertes:

ALERT_TYPE	Priorite	Description
EXCLUDED_STILL_ACTIVE	1 (Critique)	Provider exclu mais NPI encore actif dans NPPES
PRESCRIPTION_BY_EXCLUDED	2 (Critique)	Provider exclu qui prescrit encore
PAYMENT_TO_EXCLUDED	3 (Critique)	Paiements recus par un provider exclu
HIGH_RISK_SCORE	4 (Eleve)	Score de fraude CRITICAL ou HIGH
HIGH_BRAND_PRESCRIBER	5 (Eleve)	Taux de prescriptions de marque >= 80%

Colonnes de l'alerte:

Colonne	Description
ALERT_ID	Identifiant unique de l'alerte
ALERT_TYPE	Type d'alerte
NPI	Provider concerne
PROVIDER_NAME	Nom du provider
RISK_SCORE	Score de risque (0-100)
RISK_TIER	Classification du risque
ALERT_DESCRIPTION	Description detaillee de l'alerte
FINANCIAL_EXPOSURE	Montant financier expose ($)
PRIORITY_RANK	Rang de priorite (1 = plus urgent)

11. Utilisation

11.1 Commandes dbt

```
 # Depuis le container Docker
docker compose run --rm airflow-worker bash -c "cd /opt/airflow/dbt && d

# Executer toutes les couches
dbt run --profiles-dir .

# Executer Silver uniquement
dbt run --select silver --profiles-dir .

# Executer Gold uniquement
dbt run --select gold --profiles-dir .

# Executer un modele specifique
dbt run --select fraud_risk_score --profiles-dir .

# Full refresh (recreer les tables)
dbt run --full-refresh --profiles-dir .
```

11.2 Requetes d'Analyse Gold

Distribution des Risk Tiers:

```
 SELECT RISK_TIER, COUNT(*) as NB_PROVIDERS
FROM AI_FACTORY_DB.GOLD.FRAUD_RISK_SCORE
GROUP BY RISK_TIER
ORDER BY NB_PROVIDERS DESC;
```

Top 10 providers les plus a risque:

```sql
SELECT
    NPI,
    FULL_NAME,
    SPECIALTY,
    FRAUD_RISK_SCORE,
    RISK_TIER,
    TOTAL_FINANCIAL_EXPOSURE
FROM AI_FACTORY_DB.GOLD.FRAUD_RISK_SCORE
WHERE RISK_TIER IN ('CRITICAL', 'HIGH')
ORDER BY FRAUD_RISK_SCORE DESC
LIMIT 10;
```

Alertes par type:

```sql
SELECT
    ALERT_TYPE,
    COUNT(*) as NB_ALERTS,
    SUM(FINANCIAL_EXPOSURE) as TOTAL_EXPOSURE
FROM AI_FACTORY_DB.GOLD.HIGH_RISK_ALERTS
GROUP BY ALERT_TYPE
ORDER BY NB_ALERTS DESC;
```

Alertes prioritaires a traiter:

```sql
SELECT
    PRIORITY_RANK,
    ALERT_TYPE,
    NPI,
    PROVIDER_NAME,
    RISK_SCORE,
    ALERT_DESCRIPTION,
    FINANCIAL_EXPOSURE
FROM AI_FACTORY_DB.GOLD.HIGH_RISK_ALERTS
WHERE PRIORITY_RANK <= 3
```

```
ORDER BY PRIORITY_RANK, RISK_SCORE DESC
LIMIT 20;
```

Vue 360° d'un provider:

```
 SELECT *
FROM AI_FACTORY_DB.GOLD.PROVIDER_360
WHERE NPI = '1234567890';
```

Annexe: Glossaire des Tables Gold

Table	Cle Primaire	Description
GOLD.PAYMENTS_SUMMARY	NPI	1 ligne par provider avec paiements
GOLD.PRESCRIPTIONS_SUMMARY	NPI	1 ligne par provider avec prescriptions
GOLD.PROVIDER_360	NPI	1 ligne par provider (tous)
GOLD.FRAUD_RISK_SCORE	NPI	1 ligne par provider (tous)
GOLD.HIGH_RISK_ALERTS	ALERT_ID	1 ligne par alerte (multiple par provider possible)

Document genere pour AI Factory - dbt v2.0.0 (Silver + Gold)