FraudLens DBT Documentation

Data Transformation Layer - Snowflake + dbt

Overview

FraudLens uses dbt (data build tool) to transform raw healthcare data into analytics-ready tables for fraud detection. The project follows the Bronze-Silver-Gold medallion architecture.

3	20+	80+	5
Layers	Models	Tests	Data Sources

Medallion Architecture

Bronze Layer (Staging)

Raw data ingestion with minimal transformation. Source data from NPPES, LEIE, Medicare Part D, Open Payments, and CMS facilities.

STG_*

Silver Layer (Cleaned)

Cleaned, typed, and enriched data. Joins with exclusion lists, calculated metrics, risk flags.

PAYMENTS PRESCRIPTIONS FACILITIES

Gold Layer (Analytics)

Business-ready aggregations and KPIs. Provider 360 profiles, fraud risk scores, and actionable alerts.

PROVIDER_360 FRAUD_RISK_SCORE HIGH_RISK_ALERTS

Performance Optimization

Incremental Materialization

High-volume Silver tables use incremental materialization to process only new data, significantly reducing build times.

Model	Materialization	Unique Key	Strategy
`prescriptions`	INCREMENTAL	PRESCRIPTION_ID	Filter on _LOAD_TIMESTAMP
`hospital_spending`	INCREMENTAL	SPENDING_ID	Filter on _LOAD_TIMESTAMP
`facilities`	INCREMENTAL	FACILITY_ID	Filter on _LOAD_TIMESTAMP
`payments`	TABLE	PAYMENT_ID	Full rebuild (source lacks timestamp)

Snowflake Limitation: Correlated subqueries with MAX() are not supported. We use `run_query` Jinja macro to fetch the max timestamp before SQL execution.

Snowflake Clustering

High-volume tables are clustered on frequently filtered/joined columns to improve query performance.

Model	Cluster Keys	Benefit
`prescriptions`	STATE NPI	Optimizes geographic and provider-based queries
`payments`	STATE NPI PAYMENT_CATEGORY	Optimizes filters by location, provider, and payment type

```
-- Example clustering configuration
{{
    config(
        materialized='incremental',
        unique_key='PRESCRIPTION_ID',
        cluster_by=['STATE', 'NPI']
    )
}}
```

Fraud Risk Scoring Methodology

The `fraud_risk_score` model calculates a composite risk score (0-100) for each provider based on multiple risk signals.

Scoring Components

Risk Category	Max Points	Factors
Exclusion Risks	40	On LEIE list (+25), Still active while excluded (+15)
Payment Risks	20	Mega/Major recipient tier (+5-10), High % risky payments (+10)
Prescription Risks	20	High brand % (+10), High-risk drugs prescribed (+10)
Activity Anomalies	20	Both payments & prescriptions (+10), Inactive NPI with activity (+10)

Risk Tiers

CRITICAL (70-100)

Immediate investigation required. Likely excluded or multiple severe risk factors.

HIGH (50-69)

Priority review. Multiple concerning patterns detected.

MEDIUM (30-49)

Moderate concern. Some risk indicators present.

LOW/MINIMAL (0-29)

Standard monitoring. Few or no risk factors.

Data Quality Testing

Comprehensive dbt tests ensure data integrity across all layers.

Test Types

Schema Tests

`unique`, `not_null`, `accepted_values`, `relationships`

Range Tests

`dbt_utils.accepted_range` for numeric bounds (scores 0-100, amounts >= 0)

Custom SQL Tests

Business logic validation in `tests/` directory

Custom Tests

Test	Purpose
`assert_critical_risk_providers_are_excluded`	Validates exclusion logic for critical-tier providers
`assert_no_negative_financial_amounts`	Ensures all financial values are non-negative
`assert_all_alerts_have_valid_npi`	Verifies referential integrity with provider_360

Test	Purpose
`assert_financial_exposure_calculation`	Validates TOTAL_FINANCIAL_EXPOSURE = payments + prescriptions

Run tests: `dbt test` or `dbt test --select gold` for specific layers.

Key Models Reference

Gold Layer Models

Model	Primary Key	Description
`provider_360`	NPI	Complete provider profile with payments, prescriptions, and exclusion status
`fraud_risk_score`	NPI	Composite fraud risk scoring (0-100 scale)
`high_risk_alerts`	ALERT_ID	Actionable alerts for fraud investigation
`payments_summary`	NPI	Aggregated pharma payment metrics per provider
`prescriptions_summary`	NPI	Aggregated prescription metrics per provider

Alert Types

Alert Type	Priority	Trigger
EXCLUDED_STILL_ACTIVE	1 (Highest)	Provider on LEIE but still in NPPES
PRESCRIPTION_BY_EXCLUDED	2	Excluded provider has Medicare prescriptions
PAYMENT_TO_EXCLUDED	3	Excluded provider receiving pharma payments
HIGH_RISK_SCORE	4	CRITICAL or HIGH risk tier
HIGH_BRAND_PRESCRIBER	5	>80% brand prescriptions (kickback indicator)

Common Commands

```
# Run all models
dbt run

# Run specific layer
dbt run --select silver
dbt run --select gold

# Full refresh (for incremental tables)
dbt run --select prescriptions --full-refresh

# Run tests
dbt test
dbt test --select gold

# Generate documentation
```

```
dbt docs generate
dbt docs serve
```