Building the Cost of Risk Engine: From Raw Data to Trusted Intelligence

A Production-Hardened Data Pipeline for Accurate Portfolio Risk Assessment.

Breno Ruy, Analytics Engineer
December 4, 2025

What is our true Cost of Risk?

Flying Blind: Business stakeholders lacked a trusted, scalable metric for unit economics.

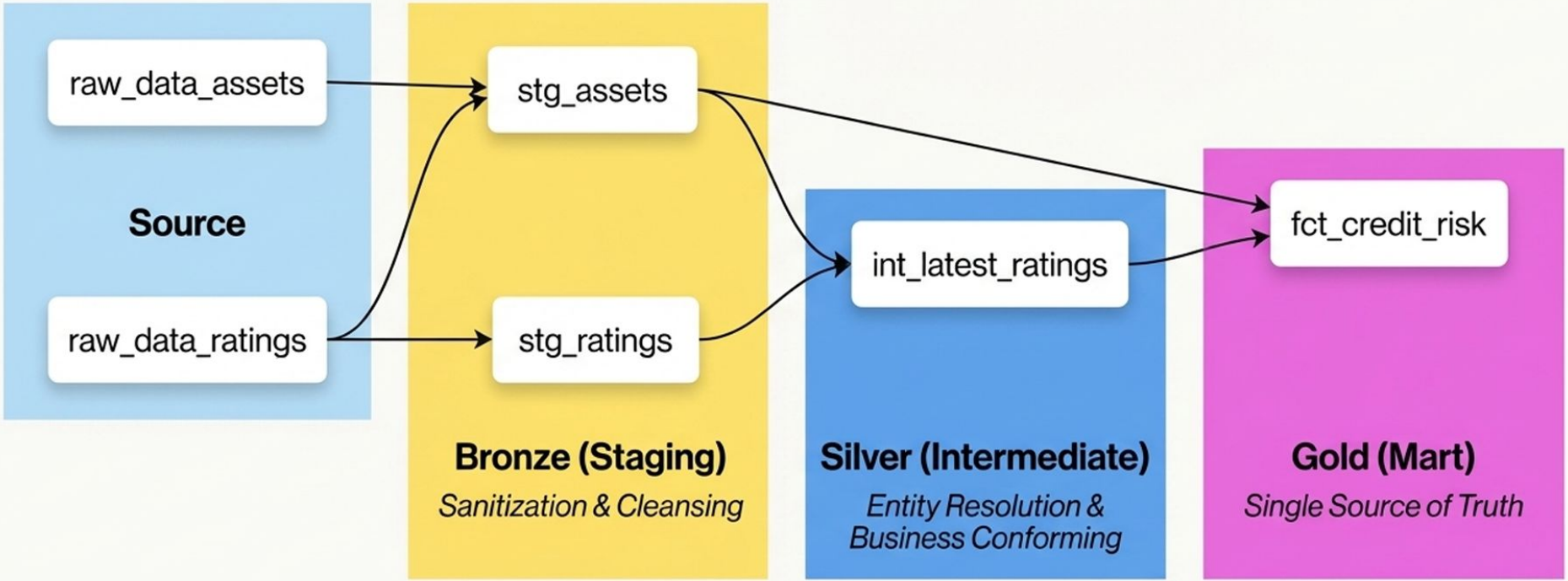
Data Chaos: Raw data sources were riddled with inconsistencies, artifacts, and missing identifiers.

The Goal: Build a ‘Trust Factory’ – an end-to-end data pipeline to deliver a single source of truth for Cost of Risk.

The Solution: A Modern Stack for Scalable Analytics

An auditable, modular, and scalable architecture designed for financial rigor.

Architecture Deep Dive: From Raw Logs to Financial Ledger



The First Checkpoint: Staging & Defensive Engineering

Raw data is guilty until proven innocent.

Ghost Loans

Challenge: ~450 records with `face_value = 0`.

Impact: Skewed Loan Count and Average Ticket Size.

Solution: Hard filter `where face_value > 0` in `stg_assets`.

Data Types

Challenge: Timestamps and numeric values stored as strings.

Impact: Downstream join failures and incorrect calculations.

Solution: Explicitly cast types (e.g., `created_at` to `TIMESTAMP`) at the earliest stage.

Status Normalization

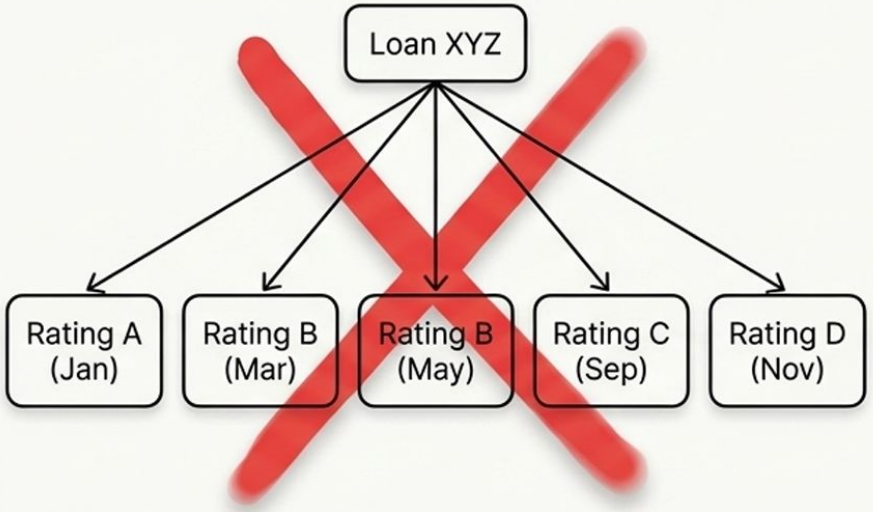
Challenge: `Repaid` and `Settled` used interchangeably.

Impact: Inconsistent reporting of paid-off loans.

Solution: Standardized both to a single 'Settled' status in the staging model.

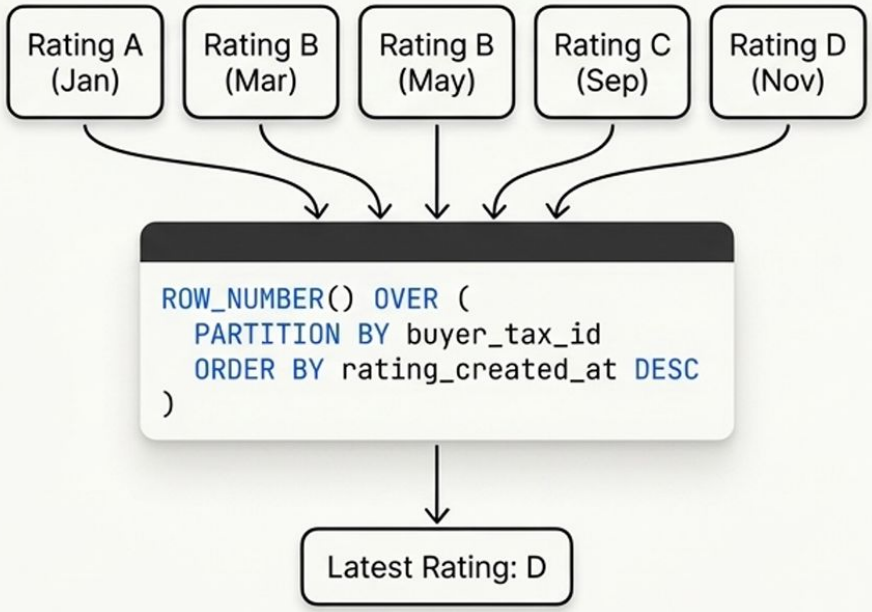
The Resolver Engine: Calculating the Current Risk Profile

The Problem



A simple join would create a 5x row explosion.

The Solution

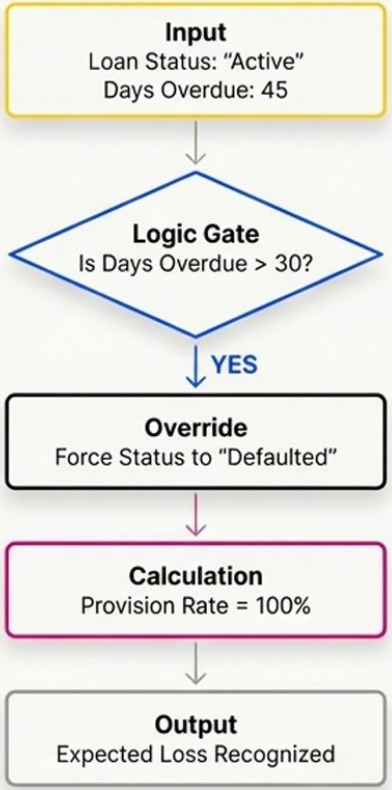


Window function isolates the most recent rating, ensuring a 1-to-1 relationship with assets.

The Breakthrough: The “Implicit Default” Logic

R$ 2.38 Million

In previously under-reported risk, identified and corrected.



Defending Design Choices: Identity & Versioning

The Challenge

1. **asset_id** was missing from the source.
2. Loans are renegotiated (due dates change), creating new versions.

created_at: 2024-01-15	due_date: 2025-10-15
buyer_tax_id: 12345678901	
face_value: R$ 100,000	

created_at: 2024-01-15	due_date: 2025-11-15
buyer_tax_id: 12345678901	
face_value: R$ 100,000	

The Solution - The 'Ultimate' Surrogate Key

```
to_hex(md5(concat(
  created_at,
  buyer_tax_id,
  face_value,
  due_date, ...
)))
```

- ✓ **Deterministic:** Ensures the same loan always gets the same ID, even on a full reload. Critical for idempotency.
- ✓ **Preserves History:** Including due_date and face_value in the hash means a renegotiation generates a *new*, unique ID.
- ✓ **Audit-Proof:** The old loan version is preserved as a settled record, while the new one becomes the active liability. This prevented over 22,000 key collisions.

Defending Design Choices: History & Scalability

Handling History

Our model prioritizes an immediate, real-time view of portfolio risk. If a client's rating drops today, our expected loss spikes today. This provides an early warning system for the Risk team.

Built for Scale

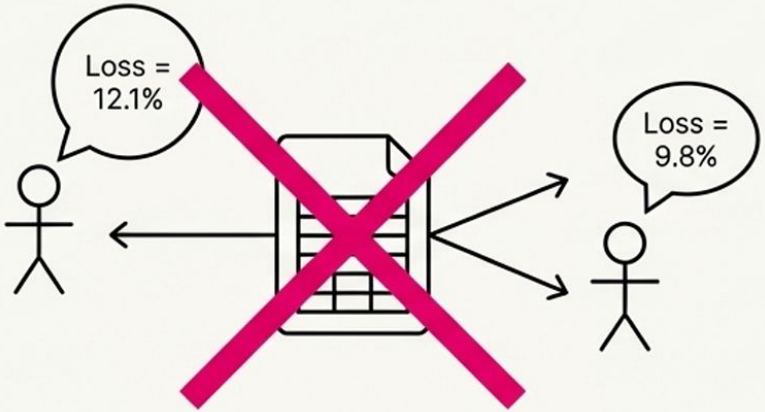
Serverless Compute: BigQuery scales automatically. We aren't managing clusters; we're just running SQL.

Modularity: The Bronze/Silver/Gold structure allows us to add new sources or change logic in one layer without breaking others. Maintenance is simplified.

Table Materialization: `fct_credit_risk` is materialized as a table, not a view. This guarantees fast query performance in our BI tool and creates a frozen, auditable snapshot for financial reporting.

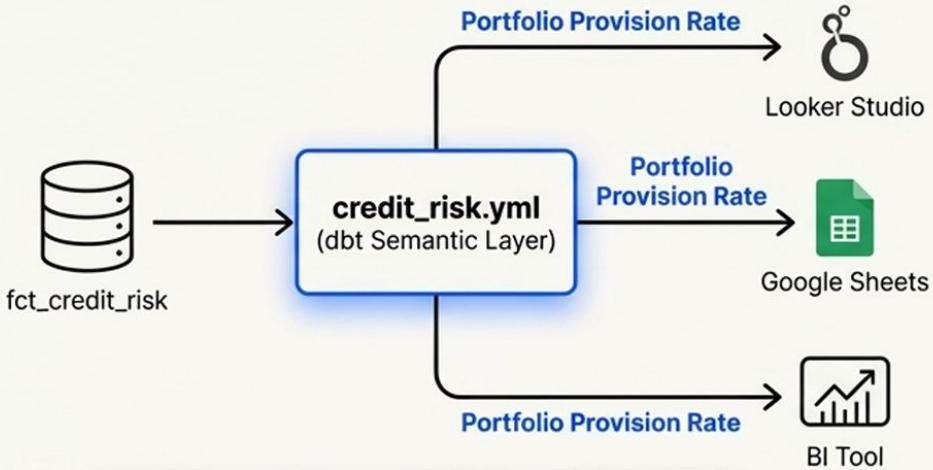
The Semantic Layer: Ending the ‘Excel Problem’

The Challenge



The 'Excel Problem': Two analysts, two different definitions of 'Loss'.

The Solution




```
metrics:
  - name: portfolio_provision_rate
    type: derived
    expression: total_expected_loss / total_exposure
```

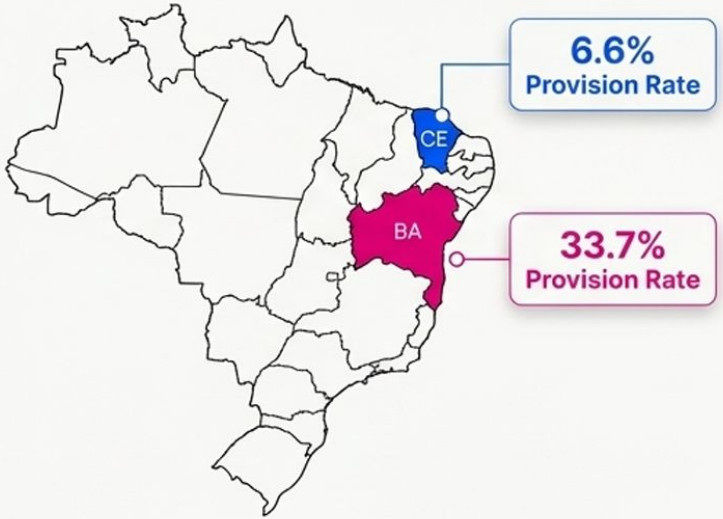
Define once, use everywhere. Stakeholders can slice by any dimension (State, Cohort) and get a mathematically correct ratio.

From Data to Decisions: A Broken Model & Geographic Toxicity

The Credit Model is “Broken”

Our “best” clients are defaulting at nearly the same rate as our “worst.” We are mispricing risk for A-rated clients.

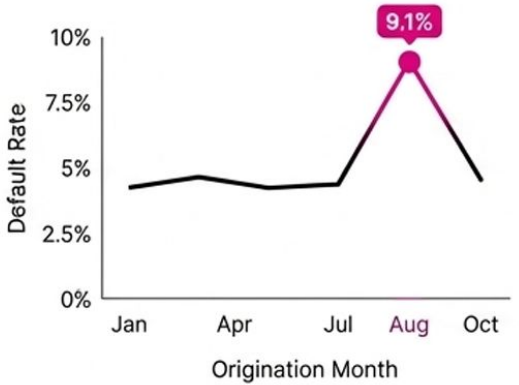
The Bahia Problem



Bahia is a toxic market, losing R$0.33 for every R$1.00 lent. Underwriting criteria needs immediate review.

Tactical Opportunities: The ‘August Anomaly’ & ‘Weekend Effect’

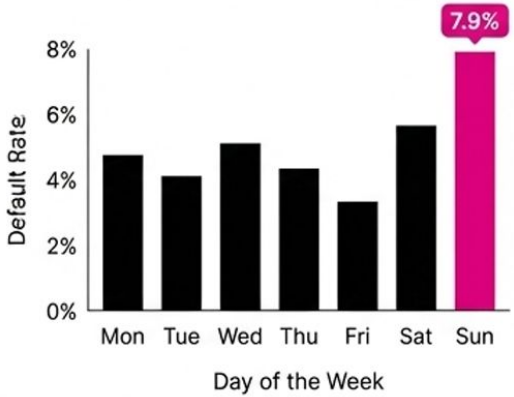
The August Anomaly



August 2025 cohort default rate spiked to **9.1%**, compared to 7.8% in July and 4.5% in October.

Action: Post-mortem on August marketing and credit policies.

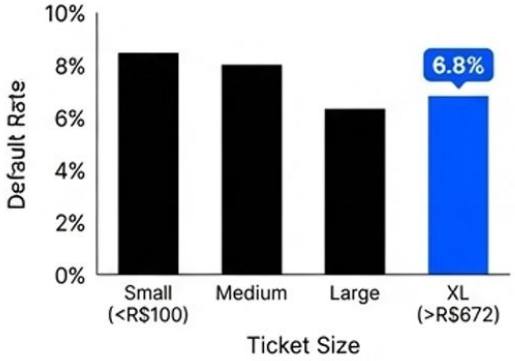
The Weekend Effect



Loans originated on **Sunday are the riskiest (7.9% default).**

Action: Consider automated blocks or manual reviews for weekend applications.

The “Whale” Paradox



Largest loans (>R$672) are our safest (6.8% default).

Action: Potential to increase credit limits for high-performing borrowers.

Achievements & The Road Ahead

Key Achievements

- **Trusted Engine:** Delivered a production-hardened Cost of Risk engine.
- **Unlocked Value:** Uncovered **R$ 2.38M** in hidden risk.
- **Actionable Insights:** Identified critical flaws in credit modeling and geographic strategy.

Recommended Next Steps

- **Automate Ingestion:** Replace manual CSV uploads with an automated connector like Airbyte or Fivetran.
- **Implement Alerting:** Configure dbt Source Freshness tests to alert if source data is not updated within 24 hours.
- **Capture Full History:** Implement **dbt Snapshots** on the ratings table to enable point-in-time analysis and track rating migrations over time.