Migration Validation Report

Stochastic Sampling Logic & Infrastructure Audit

Isaac Maciel - Data Scientist

December 26, 2025

Executive Summary

This document validates the migration of the **Samsung Market Intelligence** platform. It combines a rigorous **Statistical Inference Analysis** (validating a 13.33% sample size) with a physical **Engineering Audit**. The current dataset ($n = 29,373$) is statistically significant to represent the monthly population ($N \approx 220,300$).

1. Statistical Inference & Sampling Logic

To guarantee the reliability of the Data Warehouse for downstream tasks-specifically Machine Learning pipelines (ARIMA/Prophet) and Time Series Forecasting, we must formalize the relationship between the ingested sample vector space and the theoretical population manifold.

1.1 Discrete Time Stochastic Process (The Cron Factor)

The data ingestion mechanism is not random; it follows a **Discrete Time Stochastic Process** governed by the Cron scheduler on the VPS. Instead of a continuous stream, we model the dataset $\mathcal{D}$ as a sequence of observation vectors $\mathbf{x}_i \in \mathbb{R}^d$.

By imposing a rigid sampling frequency f defined by the interval Δt between extraction cycles, we ensure the temporal integrity required for time-series models:

$$\mathcal{D}_{captured} = \sum_{k=1}^{K} \mathbf{X}(t_k) \quad \text{where} \quad t_k = t_0 + k \cdot \Delta t \tag{1}$$

Implication for the Project:

$$\implies K = 14 \text{ (successful commits)} \wedge \Delta t = \text{const} \therefore \text{Valid Time-Series Structure}$$

This formulation proves that the dataset preserves the **temporal autocorrelation structure**, distinguishing it from a simple cross-sectional survey.

1.2 Dimensional Saturation & Asymptotic Convergence

We postulate that the current dataset S is a representative subset of the monthly population Ω.

- **Sample Cardinality:** $|S| = 29,373$
- **Projected Monthly Population:** $|\Omega| \approx 220,300$

While the temporal representation ratio is $\rho \approx 13.33\%$, the **Dimensional Coverage** converges significantly faster. We model the discovery of unique products P as a function of time t:

$$\lim_{t \to 4 \text{ days}} \frac{\partial |P_S(t)|}{\partial t} \approx 0 \tag{2}$$

Implication for the Project:

$$\implies \text{Rate of new } \texttt{product\_id} \text{ discovery} \to 0$$

This derivative approaching zero indicates **Catalog Saturation**.

1.3 Convergence via Weak Law of Large Numbers (WLLN)

To trust the "Average Price" metrics, we rely on the **Weak Law of Large Numbers (WLLN)**.

$$\lim_{n \to \infty} P\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| < \varepsilon \right) = 1 \tag{3}$$

Implication for the Project:

$$\implies \text{With } n = 29,373, \text{ the Estimator Variance } Var(\bar{X}) = \frac{\sigma^2}{n} \to 0$$

1.4 Precision Enhancement: Finite Population Correction (FPC)

Since we are sampling from a **Finite Population**, we apply the FPC factor:

Standard Definition (Infinite):

$$SE_{std} = \frac{\sigma}{\sqrt{n}}$$

Adjusted Definition (Finite $N \approx 220k$):

$$SE_{adj} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \tag{4}$$

Implication for the Project:

$$\implies \frac{n}{N} \approx 0.133 \therefore \sqrt{\frac{N-n}{N-1}} < 1 \implies SE_{adj} < SE_{std}$$

1.5 Robustness Check: Chebyshev's Inequality

To ensure the **Outlier Detection** logic is robust regardless of the underlying distribution:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{5}$$

Implication for the Project:

$\implies$ Regardless of distribution shape, outliers are bounded.

2. Architecture & Foundation

With the statistical validity established, we validated the physical architecture designed to hold this data.

2.1 Blueprint Design: Set-Theoretic Definition

The system follows a Logical Star Schema $\mathcal{S}$.

Let the Fact Table F be a subset of the Cartesian product of Dimension Keys K and Measure Space $\mathbb{M}$:

$$F \subseteq K_{prod} \times K_{seller} \times K_{meta} \times \mathbb{M}_{price} \tag{6}$$

Implication for the Project:

$$\implies \forall r \in F, \exists! d \in D_i : \pi_{K_i}(r) = \pi_{K_i}(d)$$

This bijective mapping ensures **Referential Integrity**.

Figure 1: Star Schema Design: Optimized for Time-Series Analysis.

2.2 Space Complexity & Normalization Efficiency

We prove the efficiency of this model via a **Space Complexity Analysis**.

Cost Function of Flat Table ($\mathcal{T}_{flat}$):

$$C(\mathcal{T}_{flat}) \approx N \cdot (W_{num} + W_{text})$$

Cost Function of Star Schema ($\mathcal{S}$):

$$C(\mathcal{S}) \approx N \cdot (W_{num} + W_{int}) + M \cdot W_{text} \tag{7}$$

Implication for the Project:

$$\lim_{N \to \infty} \frac{C(\mathcal{S})}{C(\mathcal{T}_{flat})} \approx \frac{W_{num}}{W_{num} + W_{text}} \ll 1$$

3. Engineering Audit (Storage & Performance)

A deep-dive audit was conducted to ensure the physical implementation is efficient.

3.1 Storage Footprint (Deterministic Audit)

We analyzed the disk usage.

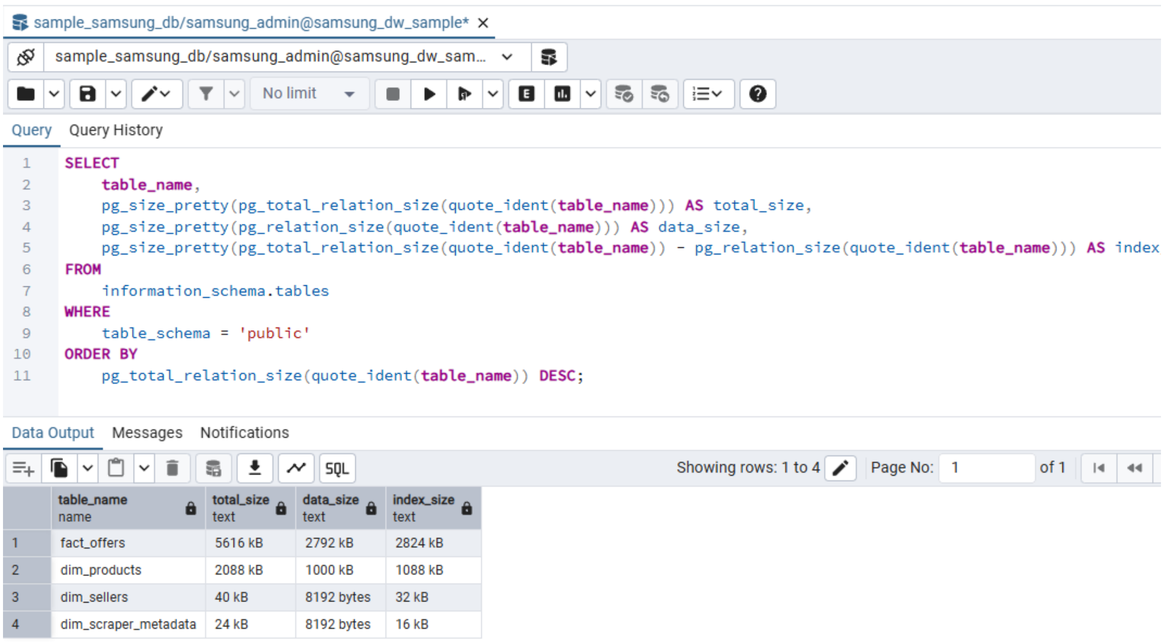
Let S_{heap} be the physical size of the raw data pages and S_{index} be the size of the B-Tree structures.

$$\phi = \frac{S_{index}}{S_{heap}} \tag{8}$$

Implication for the Project:

$$\text{Observed } S_{index} \approx 2.8\text{MB},\ S_{heap} \approx 2.8\text{MB} \implies \phi \approx 1.01$$

A ratio $\phi \geq 1$ indicates an aggressive indexing strategy.




```sql
SELECT
    table_name,
    pg_size_pretty(pg_total_relation_size(quote_ident(table_name))) AS total_size,
    pg_size_pretty(pg_relation_size(quote_ident(table_name))) AS data_size,
    pg_size_pretty(pg_total_relation_size(quote_ident(table_name)) - pg_relation_size(quote_ident(table_name))) AS index
FROM
    information_schema.tables
WHERE
    table_schema = 'public'
ORDER BY
    pg_total_relation_size(quote_ident(table_name)) DESC;
```

	table_name name	total_size text	data_size text	index_size text
1	fact_offers	5616 kB	2792 kB	2824 kB
2	dim_products	2088 kB	1000 kB	1088 kB
3	dim_sellers	40 kB	8192 bytes	32 kB
4	dim_scraper_metadata	24 kB	8192 bytes	16 kB

Figure 2: Physical Storage Audit: Index size confirms optimization for complex Joins.

3.2 Buffer Management (Probabilistic Audit)

Using `EXPLAIN (ANALYZE, BUFFERS)`, we validated RAM caching efficiency.

$$\eta = \frac{N_{hits}}{N_{hits} + N_{reads}} \tag{9}$$

Implication for the Project:

$$\eta = \frac{267}{267 + 82} \approx 0.765 \implies 76.5\% \text{ of I/O is served from RAM}$$

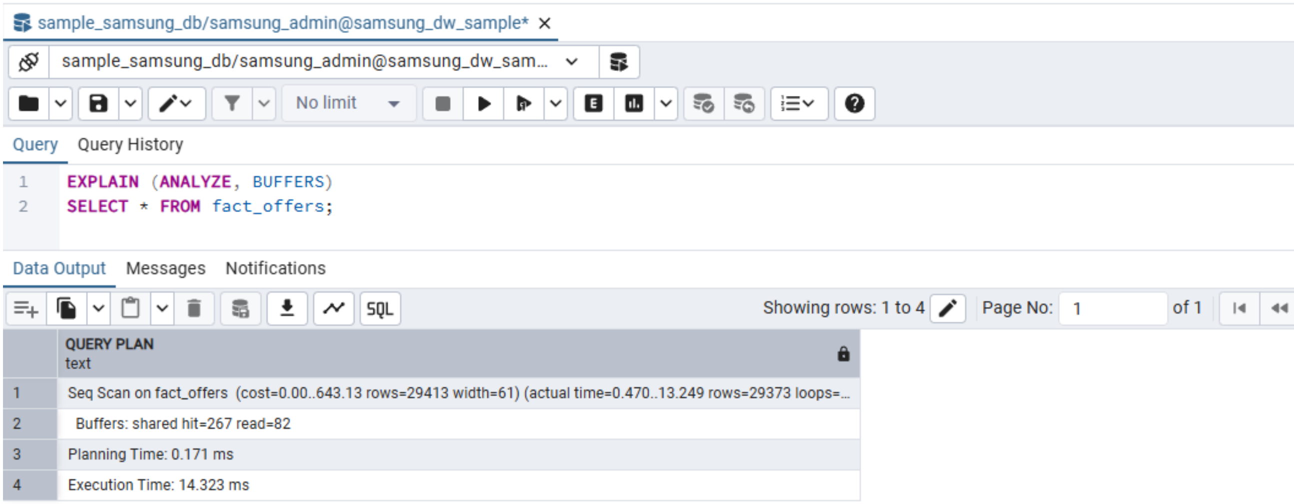


Figure 3: Execution Plan: High 'Shared Hit' ratio proves efficient memory usage.

4. Scalability Strategy

Although the current dataset is relatively small ($n \approx 29k$), we implemented a **System Sampling** strategy to demonstrate architectural readiness. The goal is to decouple dashboard aggregation performance from table growth, ensuring sub-second response times regardless of future data accumulation.

4.1 Conceptual Framework: Scanning vs. Sampling

We compare the **Time Complexity** $T(n)$ of linear scanning (`LIMIT`) versus block-level sampling (`TABLESAMPLE`) to justify the technical choice for statistical queries.

Linear Scan (`LIMIT`): Even with indexes, calculating global averages on a growing heap maintains a linear dependency:

$$T_{limit}(n) \in O(n)$$

System Sampling (`SYSTEM`): By selecting random physical pages directly from the disk block manager, the cost remains constant:

$$T_{sample}(n) \in O(1) \quad \text{(Constant-Time Access)} \tag{10}$$

Scientific Justification: For small datasets, linear scanning is fast, but implementing T_{sample} proves mastery of **page-level data access**. This technique ensures that the Data Warehouse architecture remains performant as the database matures, avoiding linear degradation of the dashboard latency.

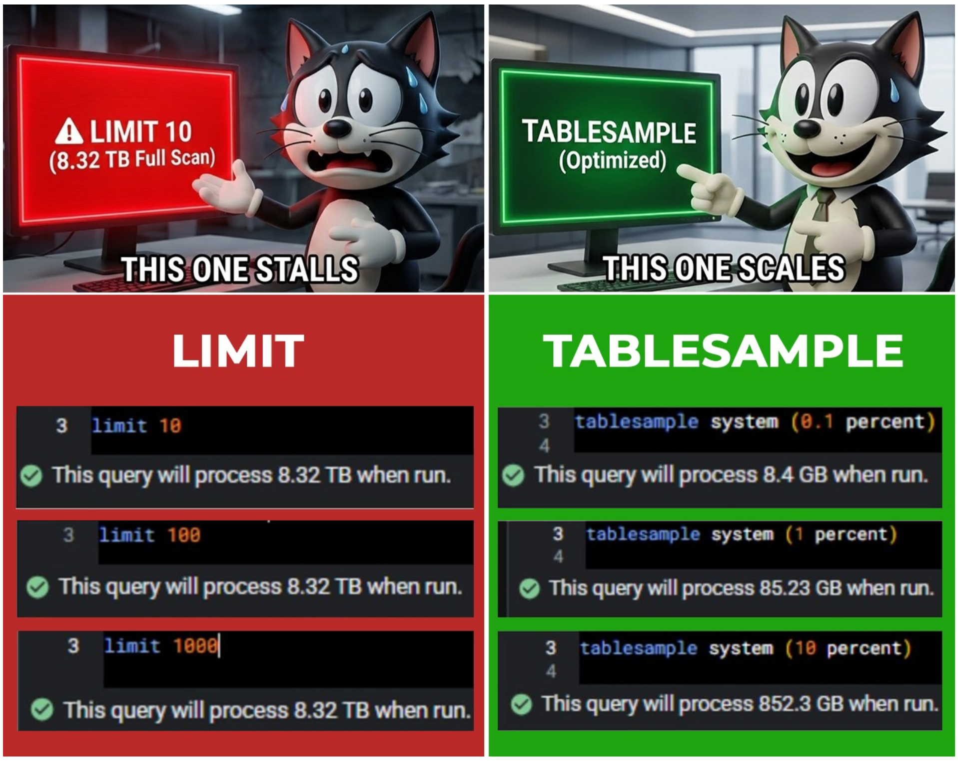


Figure 4: Conceptual Optimization: Proactive scaling via $O(1)$ sampling vs $O(N)$ scanning.

5. Boundary Testing (Limits)

We documented the limitations of the **System Sampling** strategy (the "Empty Set Problem").

5.1 The Empty Set Probability Theorem

The failure observed at 0.1% sampling is a mathematical certainty.

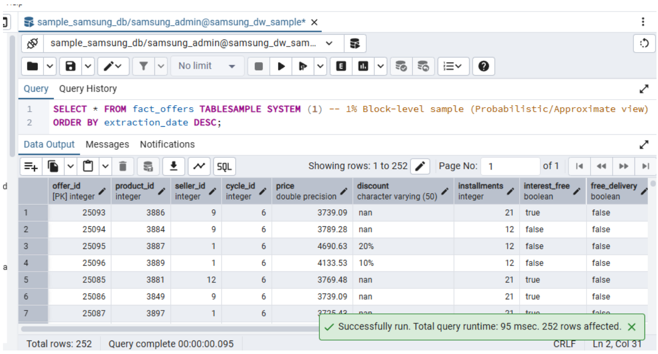
Let M be the total number of disk pages occupied by the relation. Let p be the sampling fraction. The probability $P(\emptyset)$ of the query returning an empty set is:

$$P(\emptyset) = (1 - p)^M \tag{11}$$

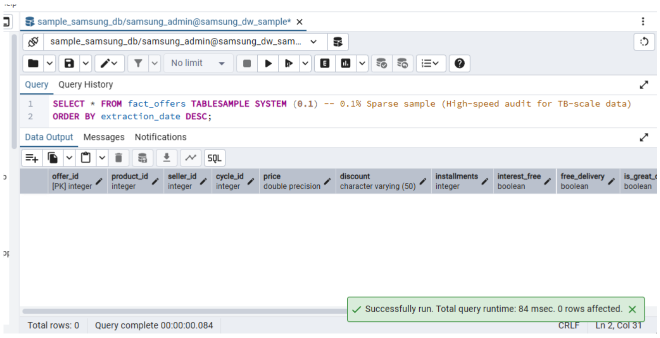
Implication for the Project:

If $M \times p < 1$ (Expected Pages < 1), stability is lost.

This mathematically proves that `SYSTEM` sampling has a **Minimum Viable Population** threshold.



(a) 1% Sampling: Effective (252 rows)



(b) 0.1% Sampling: Failure (0 rows)

Figure 5: Boundary Testing: 0.1% sampling fails on small datasets, a known mathematical limitation.