WHITEPAPER

EMERGERE TECHNOLOGIES
TECHNOLOGY ONWARDS AND UPWARDS

PRODUCT MASTER DATA AND THE UNTAPPED MARGIN OPPORTUNITY IN US HEALTHCARE SUPPLY CHAINS

The Executive Case for SKU Goldenization in US Healthcare

EXECUTIVE SUMMARY

US healthcare systems are operating billion-dollar supply chains on product data which cannot be entrusted fully. This is largely because of the duplicate SKUs, inconsistent supplier descriptions and disconnected systems which are more of a norm. The financial and operational consequences of such a norm are obvious and measurable: eroded GPO contract savings, charge capture leakage, procurement errors and regulatory exposure that only compounds with every new supplier added to the network.

This whitepaper makes the executive case for treating product master data as strategic infrastructure and explains what SKU remediation and Goldenization mean in practice. It also quantifies the cost of inaction and presents a framework that leading healthcare systems are using to establish a single, authoritative product across the entire enterprise.

For supply chain leaders, CFOs and CIOs who have already invested in ERP Modernization, contract management platforms and analytics, this is the missing foundation that determines whether those investments deliver.

> *"With regard to Healthcare, Product data is more than a data problem. It is a revenue problem, compliance problem, and increasingly a patient safety problem."*

WHAT BAD PRODUCT DATA COSTS THE US HEALTHCARE

The average large US integrated delivery network (IDN) manages between 300,000 and 1.5 million active SKUs across its supply chain systems. Across those catalogues, industry data consistently shows that 20 to 40 percent of records contain material errors like wrong descriptions, mismatched unit-of-measure, missing clinical attributes, or outright duplicates that exist because no one has ever reconciled supplier feeds against a single master. The consequences of these errors show up in three places that matter to the C-suite.

Financial Leakage

GPO contract compliance depends on purchasing the right item, at the right price, through the right channel. When SKU data is inconsistent, purchases flow outside contracted tiers generating off-contract spend that erodes savings the procurement team negotiated. Charge capture failures, where the item used does not match the item billed, represent a direct revenue loss that is almost entirely attributable to data quality. Conservative industry estimates place preventable supply chain leakage at 3 to 8 percent of total supply spend annually.

20–40%
of SKU records contain material errors

3–8%
of supply spend lost to data leakage

$1M+
average annual charge capture loss per facility

Operational Friction

Clinicians and procurement staff lose measurable time resolving product ambiguity, verifying whether two SKUs are the same item, confirming substitution equivalence during a shortage, or correcting a purchase order flagged by the ERP. Every hour spent on manual data reconciliation is an hour lost on higher-value work. In regulated industries where the labour environment is already under pressure, this is not a tolerable overhead.

Regulatory and Compliance Risk

The FDA's Unique Device Identification (UDI) requirements mandate that implantable and Class II/III medical devices be tracked with precision through the supply chain. Bad SKU data creates gaps in UDI compliance, audit trails, and recall readiness. For health systems operating 340B programmes, product misidentification can compromise programme eligibility, often triggering audits, clawbacks, and potential exclusion from the programme entirely.

FROM DATA LIABILITY TO STRATEGIC ASSET

Most health systems have not treated product master data as a strategic asset because it has never been framed that way. It lives in the domain of IT, supply chain operations, or the vendor master team and very rarely appears on a board agenda until something goes wrong.

Of late, we notice a change in that framing. The same forces driving digital transformation in healthcare: value-based purchasing, supply chain resilience after COVID-era shortages, and the proliferation of AI-driven analytics, all depend on clean, structured and trusted product data as their input. Unreliable SKU data not only weakens analytical accuracy but also causes AI-powered procurement systems to learn and replicate erroneous patterns at scale.

> *"SKU remediation and goldenization are the foundational work that makes the rest of the Data and AI investment defensible."*

What Remediation Means in Practice

SKU remediation is the systematic process of identifying and correcting errors across a product catalogue. It covers duplicate detection, attribute standardisation, unit-of-measure normalisation, supplier description reconciliation, and the application of validation rules that prevent bad data from re-entering the system. It is a discipline applied continuously across the data lifecycle.

What Goldenization Means in Practice

Goldenization takes remediated data one step further. Where the same product exists across multiple source systems :an ERP, a supplier portal, a GPO catalogue, and a clinical inventory system , goldenization reconciles all representations and produces a single master record: the golden record. This record becomes the authoritative reference for purchasing, compliance, reporting, and analytics. Every downstream system reads from it and every supplier feed is validated against it.

WITHOUT GOLDENIZATION

- Same product listed under 4 different names
- Purchasing split across non-contracted tiers
- Audit trail incomplete for UDI compliance
- Analytics producing contradictory outputs

WITH A GOLDEN RECORD

- One record, one description, one source of truth
- GPO contract compliance automatically enforced
- Full UDI traceability from order to bedside
- Analytics operating on verified, consistent data

SKU MODERNIZATION FOR A US HEALTHCARE PROVIDER

How AI-driven product master data remediation converts a fragmented SKU catalog into a trusted margin and compliance asset.

264K+
SKUs Enriched

22
Attributes Extracted

12.6%
Duplicates Removed

Across the US healthcare supply chain ecosystem, where every product movement directly impacts patient care, operational continuity, and financial performance, data accuracy has become a critical business necessity. For a leading healthcare solutions provider supporting hospitals and integrated health networks nationwide, years of fragmented systems, siloed source data, and ungoverned catalog expansion had quietly evolved into a costly operational challenge. Hidden within the item master driving purchase orders, contract matching, inventory workflows, and compliance processes was a growing crisis of duplicate records, missing clinical attributes, inconsistent classifications, and disconnected data structures that limited visibility and slowed decision-making across the supply chain.

What initially appeared to be a data management issue was, in reality, a systemic barrier to efficiency, scalability, and operational intelligence. This is the story of how they chose to confront it.

KEY CHALLENGES:

PHANTOM SKUs & DUPLICATE RECORDS

Location codes, abbreviation conventions, and migration artifacts routinely multiply a single product into 8–12 discrete records across SAP ECC, PSO, and legacy Lawson environments. This results in the spend getting invisibly fragmented across phantom line items, volume thresholds for GPO tier commitments becoming reliably unmet, and true consolidation almost structurally out of reach.

Missing Clinical & Regulatory Attributes

Across 264K+ medical product records, surface-level fields fill at 100% but the attributes that govern patient safety and compliance tell a different story. Sterility status reads UNKNOWN for 67.8% of records. Latex status is unpopulated for 69.4%. FDA device classification required for 510(k) reporting is absent from source fields entirely, requiring LLM extraction. AAMI fluid-resistance levels and ASTM standards fill at under 5%

GPO Contract Leakage

When one SKU exists as 8 fragmented variants across facility codes, the system reports 8 small purchases instead of one consolidated volume. Tier thresholds are never crossed, rebates go unclaimed, and contracted pricing stays dormant. Conservative estimates place resulting leakage at 2–5% of addressable supply spend —for a mid-size IDN, millions of dollars sitting uncaptured in contracts that were already negotiated.

Multi-System Fragmentation

The average US health system runs product data across 3 to 6 source systems simultaneously viz SAP ECC, a BW analytics layer, specialty distribution platforms like PSO or MARA, and legacy ERPs from prior acquisitions. Each applies its own naming conventions, classifications, and attribute vocabularies. As a result, cross-system spend analysis defaults to quarterly manual reconciliation, category managers cannot see consolidated supplier spend, and value analysis committees cannot identify substitution candidates. The data exists but is simply trapped in incompatible formats with no golden record to unify it.

OUR APPROACH

Six Step Remediation Process and how it works

We adopted a deliberate six step remediation framework: understand the data, enrich what is missing, standardize classifications, eliminate duplicates, validate quality, and publish a governed single source of truth. Every stage builds on the previous one because sustainable data transformation cannot afford shortcuts.

Step 1 — Source Profiling & Unified Staging

Before any enrichment begins, the data is understood. Product records from SAP ECC (KNA1/MARA) and PSO are loaded into a unified Snowflake staging environment, and a column-level completeness scan runs across all 22 target attributes. This diagnostic determines which fields can be populated from source data, which require LLM extraction, and which need external reference lookups via NPPES or GUDID. On the 264K+ records in this engagement, only operative noun and product description exceeded a 90% fill rate. Every other attribute including those governing patient safety and regulatory compliance entered the pipeline critically incomplete.

Step 2 — LLM Attribute Extraction

With the gap map established, Claude Sonnet extracts the missing attributes directly from raw description fields' structured outputs across five categories: product identity, physical attributes, safety flags, regulatory attributes, and PPE standards. A critical design principle governs this step: the model extracts only what is explicitly stated in the source text, never what can be inferred. This preserves audit defensibility as every extracted value can be traced back to the language that produced it. Benchmarking against Llama 3 confirmed near-zero enum typo rates for Sonnet across all 22 attributes, a meaningful distinction when downstream systems depend on exact field values for contract matching and formulary governance.

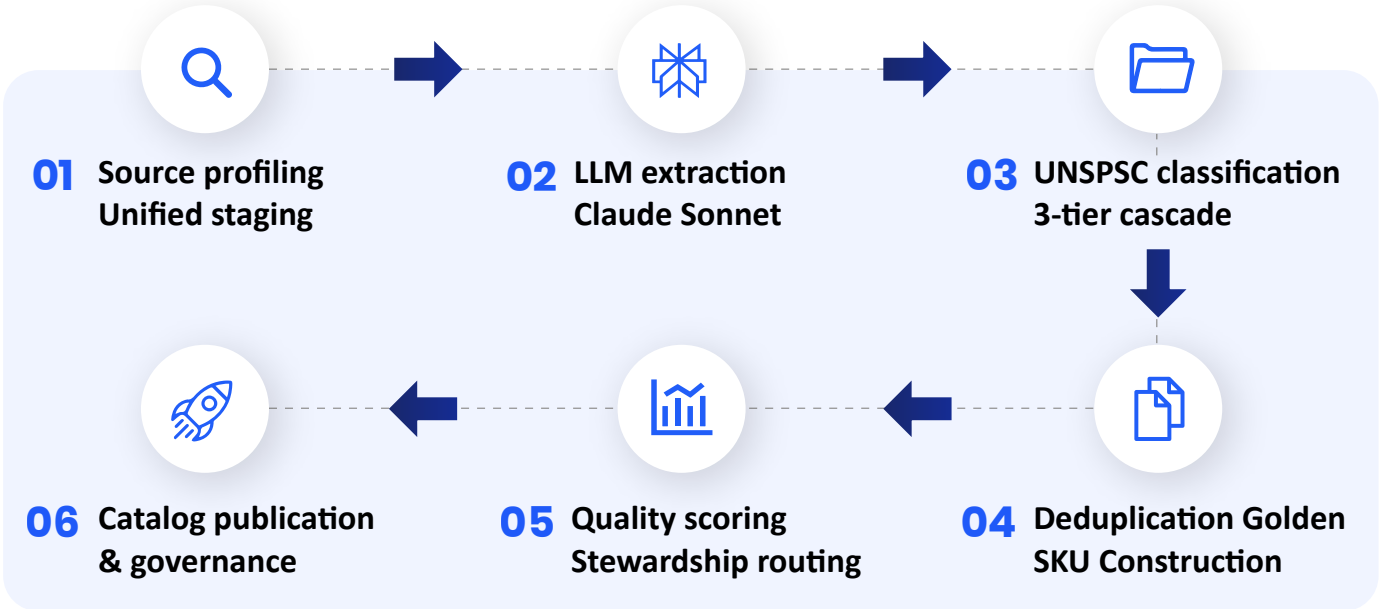

Fig:1 Six step SKU Remediation Framework

Step 3 — UNSPSC Classification via Three-Tier Cascade

Next, each product moves through a cascading three-step narrowing process , the segment, then family, then class , with each tier informed by the enriched attributes produced in Step 2. The cascade architecture prevents the category drift and misclassification errors that single-pass approaches consistently produce at scale. The output is a validated, consistent UNSPSC code for all 264K+ products enabling GPO benchmarking, spend analytics, and category-level reporting that were structurally impossible against the uncategorized source catalog.

Step 4 — Deduplication & Golden SKU Construction

With records enriched and classified, deduplication runs across three tiers designed to catch fundamentally different types of redundancy. Tier 1 resolves exact GTIN and NDC matches which are the clearest cases. Tier 2, the workhorse of the process, applies LLM-normalized name matching combined with unit-of-measure hashing to collapse the distribution-center phantom variants that fragment spend across facility codes. Tier 3 deploys Cortex EMBED semantic similarity to catch rebrands, abbreviation mismatches, and supplier naming conventions that defeat deterministic matching entirely. Every duplicate cluster resolves to a single golden record, with full lineage preserved back to all contributing source variants. Throughout the framework, data is only consolidated without discarding.

Step 5 — Quality Scoring & Stewardship Routing

Every golden SKU that emerges from deduplication receives a composite quality score evaluated across three dimensions: attribute completeness, source confidence, and match tier. This score determines what happens next. High-confidence records are auto-approved and flow directly to publication. Borderline records route to a human steward queue, accompanied by the LLM's rationale for the decisions made giving reviewers context. Records falling below threshold are held and directed toward sourcing intervention or external reference enrichment. The result is a tiered workflow that applies human attention precisely where it adds value, rather than uniformly across all records.

Step 6 — Golden Catalog Publication & Governance

The remediated catalog is published as three distinct datasets. The golden master contains one row per canonical product which is fully enriched, classified, and scored. The variants table maps every source record to its golden parent, preserving the lineage that procurement, finance, and compliance teams need for reconciliation. The audit log captures every enrichment decision, every match resolution, and every classification assignment, time-stamped and traceable. All three are published to Snowflake and integrated directly with GPO contract matching, ERP write-back pipelines, and value analysis tooling, so the clean data gets to work.

TECHNOLOGY ARCHITECTURE : HOW THE PIPELINE IS BUILT AND WHY IT IS BUILT THIS WAY

The six-stage remediation process described above does not run on a separate platform, a third-party orchestration layer, or an external cloud environment. It runs entirely inside Snowflake which shapes everything from data security to audit readiness to the speed at which enriched records reach downstream systems.

Snowflake as the Execution Layer

Snowflake serves as the native execution layer for the entire enrichment pipeline including source profiling, attribute extraction, UNSPSC classification, deduplication, quality scoring, and golden catalog publication using Cortex LLM functions. Product data never leaves the customer's Snowflake environment, eliminating external API dependencies, reducing latency, and maintaining HIPAA and SOC 2 compliance through existing IAM, audit logging, and residency controls.

Model Selection: Why Claude Sonnet

For healthcare SKU enrichment, model precision is critical. The pipeline extracts 22 structured attributes including AAMI levels, ASTM codes, and FDA device classes where even minor enum inconsistencies can break downstream validation and compliance workflows. While Llama 3 introduced 3–4% enum corruption during testing, Claude Sonnet achieved near-zero error rates across 264K+ records, making it the recommended production model.

Source Systems in Scope

The pipeline ingests from four source systems, each contributing a distinct layer of product intelligence to the unified staging environment.

01 **SAP ECC** as the primary product system of record

02 **MARA** for packaging and physical attributes

03 **Snowflake Cortex** as the enrichment and publication layer

04 **GUDID/NPPES** for FDA and GTIN validation

Together, these systems create a unified, governed golden product record ready for ERP write-back, GPO contract matching, and value analysis.

Scaling the Framework Forward

Proven on a live catalog of over 264K+ records, this framework is designed to scale seamlessly to millions of SKUs through batched, parallel processing. While this implementation leveraged Snowflake Cortex to align with the client's existing environment and compliance requirements, the underlying remediation framework is platform-agnostic and can be deployed across any modern cloud data platforms with LLM orchestration capabilities.

BUSINESS VALUE THROUGH SKU MODERNIZATION

The results of our SKU modernization against a live item master of 264K+ medical-surgical product records moved quickly from data quality metrics to business value. Here is what the engagement surfaced.

GPO Contract Compliance · $8M–$20M in recoverable margin

The POC identified that phantom SKU fragmentation was splitting single-product purchases across 8–12 discrete line items, preventing the health system from ever crossing GPO volume thresholds. Consolidating those variants into golden records immediately unlocked tier eligibility that had been sitting dormant in contracts already signed. On a $400M supply base, the compliance uplift translates to $8M–$20M annually with no renegotiation required.

Catalog Maintenance Cost · ~80% reduction in manual effort

The enrichment pipeline replaced the work that 3–8 FTEs were performing manually — classifying, correcting, and deduplicating records on a continuous basis. The POC demonstrated that LLM-driven enrichment at scale does not just speed this work up but also eliminates the need for it as a standing function, freeing those resources for value analysis, formulary rationalization, and supplier management.

Clinical Safety Risk · A safety flag layer that did not exist before

Across the 264K+ records in scope, the POC found sterility status unknown for 67.8% of products and latex flags absent for 69.4%. By the end of the engagement, every record carried a validated safety attribute set enriched from product descriptions and cross-referenced against GUDID device identifiers, giving the clinical and supply chain teams a compliance baseline they could not previously demonstrate.

Spend Analytics Maturity · Real-time category visibility, live

With a fully classified golden catalog published to Snowflake post-POC, category managers gained the ability to run real-time cross-supplier, cross-facility spend analysis for the first time. Substitution candidates that previously required weeks of manual taxonomy work became immediately visible. The product master stopped being a maintenance burden and started functioning as a strategic decision-making asset.

The Aggregate Picture

Across GPO compliance recovery, operational cost reduction, clinical risk mitigation, and analytics capability, the POC demonstrated that the financial return from SKU modernization is very structural. The margin was always there. It was buried in data that had never been governed. Our remediation framework brought it to the surface.

FIVE RECOMMENDATIONS FOR HEALTHCARE PROVIDERS

01 Audit Your Product Master Before Your Next GPO Cycle

Commission a column-level completeness scan across all source systems before your next GPO contract review. Quantify the percentage of SKUs missing manufacturer, UOM, FDA class, and safety flags. Map the fill-rate gap to estimated contract leakage. This audit is the business case that converts a data quality initiative into a margin recovery initiative with a named dollar figure attached.

02 Treat Safety Flag Enrichment as a Clinical Governance Requirement

Sterility, latex, and disposable flags are patient safety controls, not convenience attributes. Regulatory frameworks including The Joint Commission standards and CMS Conditions of Participation require that formulary items be traceable to clinical and safety specifications. Position attribute enrichment as a clinical governance initiative and it will receive the executive sponsorship and resourcing it warrants.

03 Implement UNSPSC Classification Across Your Full Catalog

UNSPSC is the prerequisite for every downstream spend analytics capability: category benchmarking, GPO utilization reporting, spend cube construction, and supplier consolidation modeling. A one-time AI-driven classification pass on your full item master validated through a cascading three-tier narrowing architecture that delivers a classified catalog that supports 5–10 years of category management work without re-classification.

04 Build a Golden SKU Master as a Shared Enterprise Service

A single, de-duplicated, enriched golden catalog published to your enterprise data warehouse and integrated with your ERP write-back layer eliminates the redundant cleansing work that supply chain, clinical informatics, finance, and value analysis teams currently perform independently. Govern the golden master through a stewardship function, and it compounds in quality and coverage with each procurement cycle. Treat it as infrastructure, not a project.

05 Start with a Time-Boxed Proof of Concept on a High-Spend Category

Select a high-spend, high-duplication category — surgical supplies, PPE, implants and run the full remediation pipeline on 10,000–50,000 SKUs. A focused POC delivers quantified outcomes within 4–6 weeks: duplicate rate, enrichment coverage by attribute, UNSPSC classification accuracy, and a GPO contract compliance estimate. These metrics are sufficient to fund a full enterprise deployment and create organizational alignment across supply chain, finance, and clinical leadership.

WHERE THE DATA ENDS AND THE STRATEGY BEGINS

The problem this whitepaper has described is not a new one. Health systems have known for years that their item masters were degraded ,that duplicates were multiplying, that clinical attributes were missing, that spend analytics were running on a foundation too unstable to support the decisions being made on top of it. What has been missing is a method precise enough to fix it at scale, and a technology architecture secure enough to run inside the environments healthcare organizations are actually permitted to use. This SKU modernization framework delivers both.

The Conversation Worth Having

If your organization is purchasing $200M or more in medical-surgical supply annually and has not conducted a structured item master diagnostic in the last eighteen months, the numbers in this whitepaper are likely conservative estimates of your own catalog. We offer a no-commitment item master diagnostic that profiles your source systems, quantifies duplicate density, measures attribute completeness across clinical and regulatory fields, and produces a recovery estimate specific to your supply base. The diagnostic runs in your Snowflake environment. Your data does not move. The output is a clear, quantified picture of what SKU modernization would return before any remediation work begins. The data already exists inside your organization. The contracts are already signed. The margin is already there.

Let us show you exactly where it is.