



From Catalog Fragmentation to Procurement Visibility: An Agentic Reference Architecture for Puerto Rico's First Government Item Master



Author: Jared Villanueva Valentín
Advisor: Nelliud D. Torres Batista, DBA
Master of Engineering in Computer Engineering
Graduate Project EXPO, Spring 2026

Abstract

Puerto Rico's General Services Administration (ASG) owns the central procurement catalog for the entire Government, yet it holds roughly 300,000 records with severe fragmentation — a single ballpoint pen appears in over 117 textual variants. This work delivers an agentic Big Data architecture: a Databricks Medallion Lakehouse paired with five Microsoft Foundry agents under responsible-AI safeguards. Estimated first-year Azure cost is approximately \$72K, substantially below the \$200K–\$1M manual-curation estimates.

Introduction

In July 2025, ASG issued an RFI for AI solutions to be used across government agencies and municipalities [1]. Ley 73 of 2019 centralizes all procurement under ASG [2], and JEDI is the platform every government purchase has to pass through.

But the structural gap remains: ASG has no authoritative record of what the Government actually buys. Without that record, audits don't tie out and any AI tool built on top of the catalog inherits the same data-quality limitations. This work fixes the data layer first.

Background

The technical core is entity resolution: many catalog records point to the same product under different text. The standard pipeline [3] runs candidate generation, similarity scoring, and survivor selection, with one threshold deciding the false-positive / false-negative trade-off.

Adopting the NIGP taxonomy replaces ASG's legacy 12-category /75-subcategory partition with about 1,957 federally-aligned subcategories, which ties the catalog back to U.S. state and federal procurement systems. The 2025 OECD analysis [4] frames procurement as a high-leverage AI application but warns that public-sector AI has to stay transparent, auditable, and governed.

Problem

A diagnostic on a snapshot of JEDI [5]:

299,986

total records

39,616

with identical descriptions (duplicate candidates)

117+

variants of one blue ballpoint pen

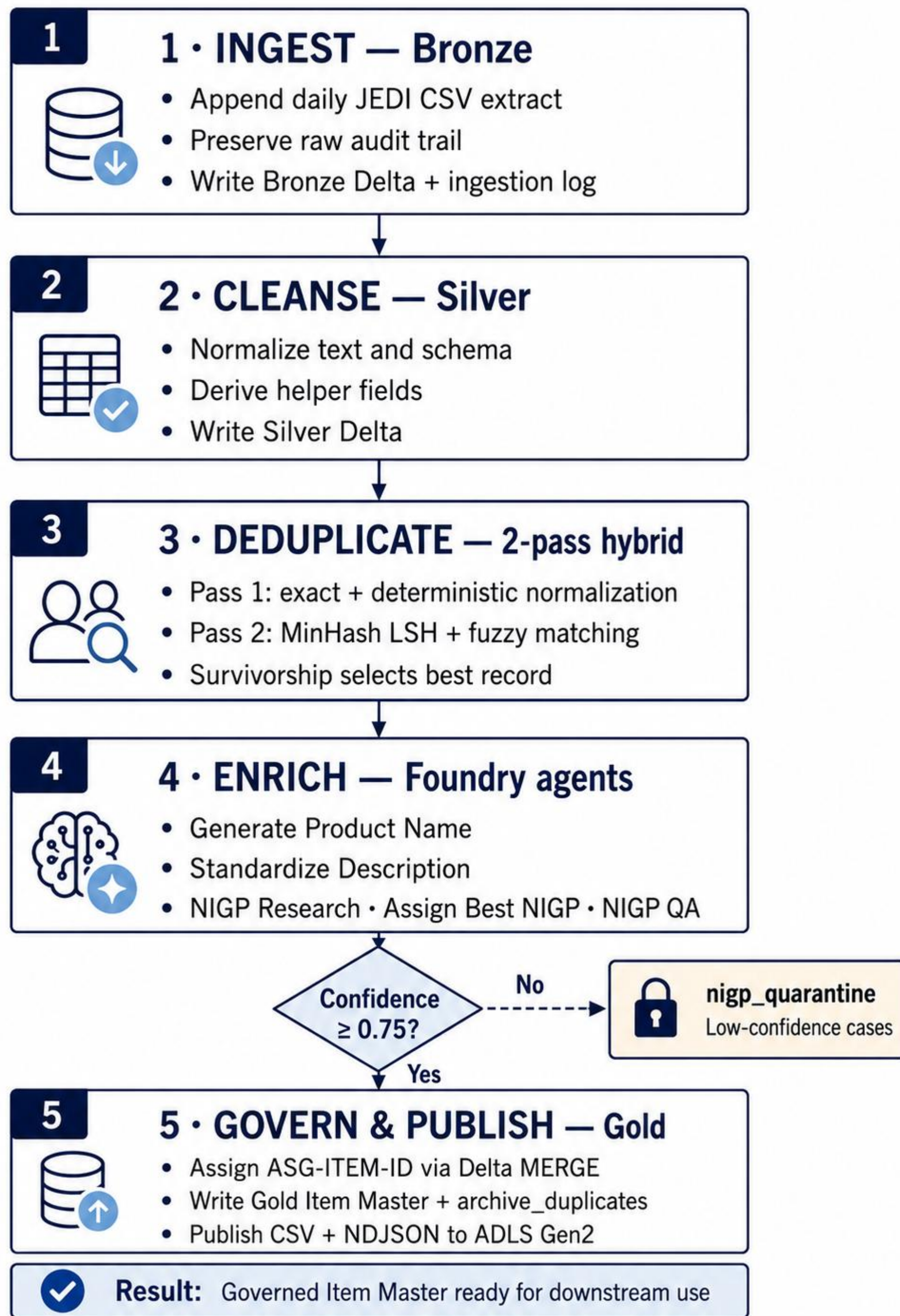
70%

missing brand, model, UoM, or contract data

NIGP (~1,957 sub-categories) replaces the legacy 12 / 75 partition and reconnects the catalog to U.S. federal procurement.

Objectives: (1) unique ASG Item ID, (2) catalog-scale deduplication, (3) AI-assisted NIGP under responsible-AI rules, (4) durable governance.

Methodology



Results and Discussion

Pipeline KPIs vs. target

KPI	Target	Measured
NIGP QA precision	≥ 90%	92% ✓
Generate Product Name compliance	≥ 95%	96% ✓
Pipeline latency (active compute)	≤ 90 min	68–80 min ✓
Deduplication rate	≥ 95%	97% ✓

10 % stratified sample (~30K records) for latency + deduplication; 50-item manual subset for human-review KPIs.

Cost: ~\$72K vs \$200K–\$1M

First-year, all-in. Manual quotes are one-shot deliverables.

NIGP manual quote

\$1,000,000

one-time, hand-curation

Social Glass

\$200K

one-time, hand-curation

As-deployed

\$72K

+ ~\$150/mo infra · ongoing single source of truth

~10×
below either
alternative

ASG-ITEM-ID: a self-describing key

Each segment readable on its own — no external lookup required.

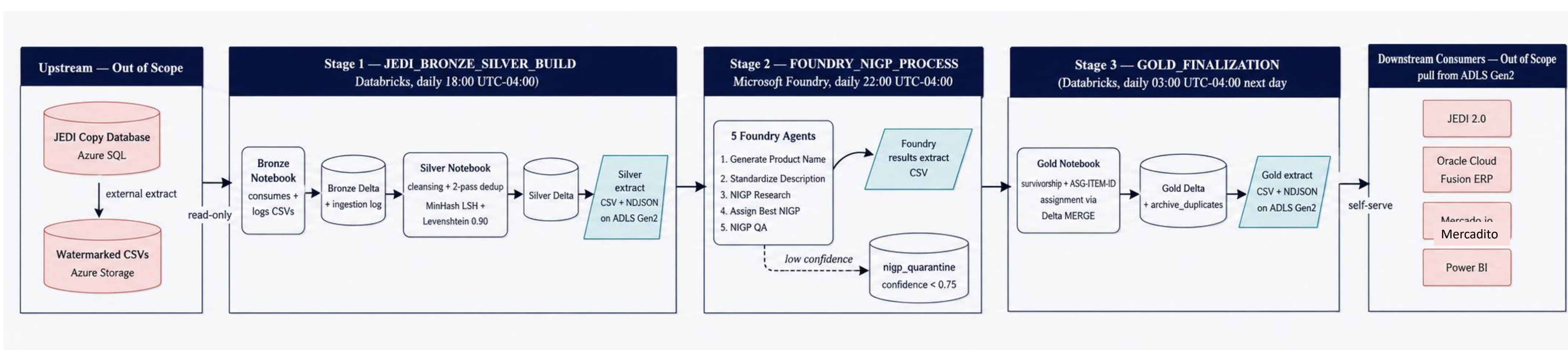


Two simplified variants:
AAA-BB-EEEEEEEEEE for requisitions · AAA for bid solicitations

Delivered — 4 artifacts

- 1 ASG-ITEM-ID accepted by ASG · validated for JEDI + Oracle ERP
- 2 Daily lakehouse pipeline all 4 KPI targets met · ~9h wall-clock · 68–80min compute
- 3 Five Foundry agents deployed under responsible-AI constraints
- 4 Governance framework Platform · Process · Operational

Architecture



Conclusions

ASG's procurement catalog was fragmented to the point that audits, spend analytics, and any AI tool above it all returned unreliable answers. This work delivered the fix: a Databricks Medallion Lakehouse with five Microsoft Foundry agents, publishing a governed Item Master as timestamped extracts on ADLS Gen2.

Four artifacts shipped against the four objectives. Every pipeline KPI hit target, and the first-year cost (~\$72K) is an order of magnitude below manual-curation estimates.

Future Work

Two extensions are designed and scoped to land after the June 2026 operational handover:

- First, a Catalog Steward review-and-reinjection workflow over nigp_quarantine will close the loop on low-confidence agent outputs through audit-logged corrections.
- Second, a requisition-matching mode (NIGP QA Mode 2) will score incoming requisition lines against the Gold Item Master and quarantine weak matches, so new duplicates get caught at the point of request.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Nelliud D. Torres Batista, PhD, for his steady guidance, technical insight, and support throughout this project, and to the Polytechnic University of Puerto Rico Graduate School for its academic support. I am especially grateful to my family for their patience and encouragement throughout this journey. I also extend my appreciation to the Puerto Rico General Services Administration, especially Karla Mercado, Administrator; Joel Fontáñez González, Sub-Administrator; and James Olmeda Rivera, Chief Information Officer, for making this engagement possible and supporting this public-sector data initiative.

References

- [1] WIPR, "ASG anuncia solicitud de propuestas para integrar inteligencia artificial," Jul. 17, 2025.
- [2] Ley Núm. 73 de 2019 — Ley de la Administración de Servicios Generales, P.R.
- [3] A. K. Elmagarmid et al., "Duplicate record detection: A survey," IEEE TKDE, 19(1), 2007.
- [4] OECD, Governing with Artificial Intelligence, OECD Publishing, 2025. doi:10.1787/795de142-en.
- [5] Social Glass, Inc."ASG Item Master: Diagnostic & Problem Definition" Tech. Rep. Deliv. 1, 2025.
- [6] Microsoft, "Medallion Lakehouse Architecture," Microsoft Learn, 2025.
- [7] Microsoft, "Microsoft Foundry Agent Service," Microsoft Learn, 2025.