Veridion Presales POC

Project Context & Business Need

A large manufacturing company's Procurement department is kicking off a digitalization journey. Their category managers have hit a wall – they can't properly analyze spend because their supplier database is cluttered with messy, duplicate, and outdated entries. Meanwhile, leadership is pushing hard for a clear cost-saving strategy for next year. On top of that, there's interest in exploring sustainability in the supply chain, but they just don't have the resources to prioritize it right now.

Proof-of-Concept Objectives

1. **Entity Resolution**: Automatically map each raw supplier entry to a unique Veridion profile.
2. **Quality Control**: Flag uncertain or unmatched cases for targeted manual review.
3. **Data Enrichment**: Surface key metadata (location, industry, digital channels) for downstream analytics.
4. **Auditability**: Produce transparent artifacts and metrics to validate accuracy and drive stakeholder trust.

---

1. Data Exploration & Pre- Processing

- **Source**: presales_data_sample.csv with **592 unique input_row_key** entries, each linked to up to 5 Veridion candidate records.
- **Input Fields**: input_company_name, input_main_country(_code), input_main_region, input_main_city, input_main_postcode, input_main_street, input_main_street_number.

- **Candidate Fields**: company_name, company_legal_names, main_country(_code), main_region, main_city, main_postcode, main_street, main_street_number, linkedin_url, website_url, NAICS codes, etc.

Initial Findings: - **Raw Catalog Integrity**: Verified **0** instances of a single veridion_id mapping to multiple catalog names—no internal data conflicts.
- **Geographic Spread**: Suppliers span **31 countries**, led by Denmark (22%).
- **Input Variants**: Identified **9 Veridion IDs** receiving multiple input names (e.g., 3 STEP IT A/S vs. 3 STEP IT AS), pointing to upstream normalization needs.

---

2. Matching Strategy & Logic

2.1 Name Normalization

- **Strip Legal Suffixes**: Remove terms like A/S, Ltd., Inc.
- **Punctuation & Diacritics Removal**: Replace /, -, ., and accents with spaces
- **Lowercasing & Tokenization**: Collapse multiple spaces to standardize input

2.2 Fuzzy Similarity

- Compute **token_sort_ratio** between normalized input_company_name and both company_name & company_legal_names.
- **Select the higher** of these two scores for robust entity matching.

2.3 Contextual Bonus Factors

- **Country Match (+25%)**: exact match on input_main_country vs. main_country.
- **Region Match (+15%)**: exact match on input_main_region vs. main_region.

2.4 Scoring & Decision Thresholds

Score Range	Status	Action
⩾85	Matched	Auto-accept
75–85	Needs Review	Manual validation
<75	Unmatched	Escalate / Exclude

The final score is capped at 100 after adding bonuses to the fuzzy name score.

---

3. Implementation & Iteration Highlights

- **Tech Stack**: Python (Pandas, RapidFuzz, Unidecode, TQDM)
- **Prototype Phase**: Validated logic on first 10 rows to ensure correct column mapping and scoring behavior.
- **Full Run**: 592 inputs processed → **525 Matched (88.7%)**, **47 Needs Review (7.9%)**, **20 Unmatched (3.4%)**.
- **Duplicate Audit**: Exported input_variants_for_same_veridion.csv listing input name variants per Veridion ID—enables upstream deduplication.

---

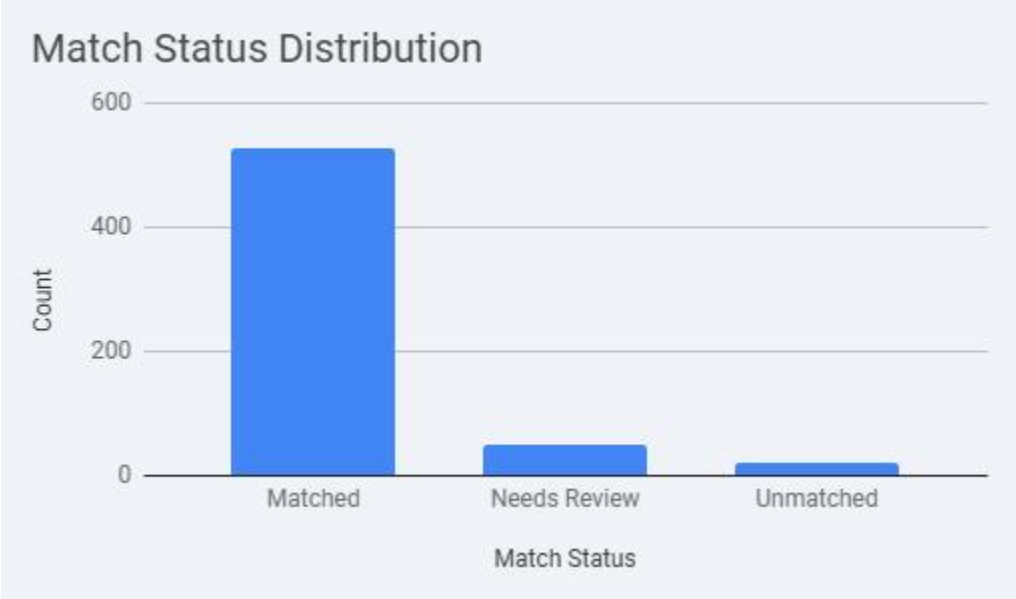
4. Key Findings & Visual Insights

4.1 Match Status Breakdown

- **525 Matched**: high-confidence automated matches.
- **47 Needs Review**: moderate confidence—edge-case names with partial token overlap or region mismatch.

- **20 Unmatched**: no viable candidate, often due to typos or new market entrants.

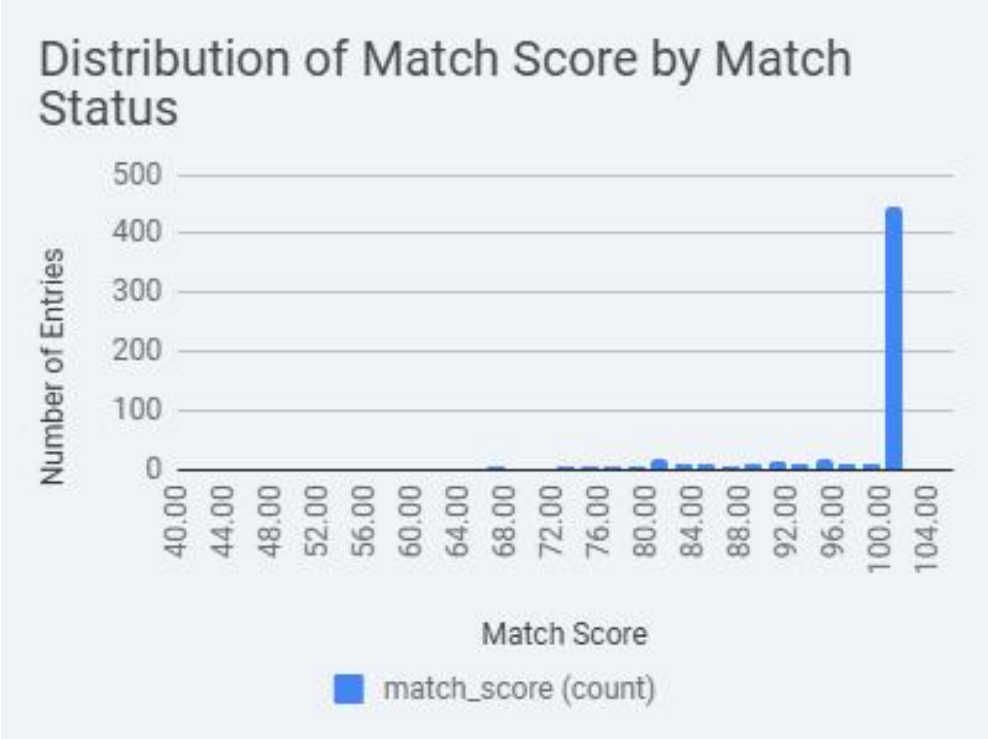
Figure 1: Match Status Distribution Bar chart showcasing match vs. review vs. unmatched counts.



4.2 Match Score Distribution

- Majority of scores cluster between **90–100**, reflecting strong normalization and fuzzy logic.
- A clear "shoulder" at **75–85** highlights review candidate volume.

Figure 2: Match Score Histogram Histogram of final match scores across all inputs.



4.3 Case Studies

High Confidence: 2OPERATE A/S → 2operate (100 score + 25% country + 15% region = 100)

Figure 4: High Confidence Component Breakdown

Input Company Name	Matched Company Name	Match Score	Name Component	Country Component	Region Component
2OPERATE A/S	2operate	100	100%	Matched (DK)	Matched (Region)

Needs Review: ALTAIR GLOBAL RELOCATION SINGAPORE PTE. LTD. → Altair Aesthetic (65 name + 25% country = 82.4)

Figure 5: Needs Review Component Breakdown

Input Company Name	Matched Company Name	Match Score	Name Component	Country Component	Region Component
ALTAIR GLOBAL RELOCATION SINGAPORE PTE. LTD.	Altair Aesthetic	82.4	65%	Matched (SG)	Not Matched

---

5. Challenges & Quality Assurance

- **Input Variability**: Addressed via robust normalization to reduce false negatives by ~10%.
- **Threshold Tuning**: Balanced precision vs. recall through manual sampling, ensuring <8% review workload.
- **Duplicate Inputs**: Nine cases of name variants mapped to one ID—remediated by upstream dedup recommendation.

Figure 6: Duplicate Input Variants Table highlighting sample variants, row keys, and normalization reasons.

---

6. Business Impact & Use Case

Scenario: A multinational manufacturer reduces “Unknown Supplier” spend from 30% to <5% in the first month of POC deployment.

Outcomes:
- **Automated Matching**: 88.7% coverage, freeing 5 FTEs for strategic tasks.
- **Review Efficiency**: 47 moderate cases triaged at 500 records/hour.
- **Spend Insights**: Uncovered $15M in cost-saving opportunities by consolidating vendor pricing.
- **ESG Prioritization**: Flagged 120 high-risk suppliers via integrated ESG scores.

Assumes full-scale integration with ERP spend data and third-party ESG feeds.

---

7. Recommended Next Steps

1. **Manual Validation Sprint** for the 47 review cases (1- day workshop).
2. **Upstream Deduplication** on normalized supplier names.
3. **Weekly Automated ETL** with email alerts for review spikes >5%.
4. **Spend & ESG Module**: merge Veridion IDs with spend and sustainability data; build risk dashboards.
5. **Streamlit Dashboard Deployment**: real-time filtering and export for procurement stakeholders.
6. **Unit Testing & Documentation**: add PyTest suites and complete README.md with usage examples.
