AI-DRIVEN COMPANY DATA ENRICHMENT SYSTEM – SOUTH AFRICA SMARTPHONE FINANCING DATASET [Sorvanis - Data and AI Architect internship Assignment]

System Architecture

The system follows a modular data-enrichment pipeline with four layers:

(1) Data Ingestion – raw CSV data is loaded into the pipeline;

(2) Entity Resolution – each company name and website is normalized to ensure consistency;

(3) Web Data Retrieval – official websites, fintech directories, and Wikipedia pages are queried using an automated crawler or API; and

(4) AI-driven Attribute Extraction – extracted unstructured text is processed using NLP models (LLMs, NER, and pattern matchers) to populate missing fields such as headquarters, founded year, and business model.

AI Extraction Approach

AI models leverage Named Entity Recognition (NER) and Large Language Models (LLMs) to detect key entities from company descriptions.

A hybrid approach is used:

LLMs interpret unstructured web text to infer business models, financing types, and eligibility terms; rule-based regex patterns extract structured values like years and cities. Data augmentation ensures robustness when limited online information is available.

Confidence & Quality

Each extracted attribute is assigned a confidence score (0–1) based on AI model certainty, source reliability, and cross-source consistency. If multiple sources disagree, weighted voting favors official or verified domains. Low-confidence fields are flagged for human review. A continuous learning loop updates extraction weights using feedback from verified data.

Schema Materialisation

Extracted fields are validated through type-checking and normalization. City names are matched against a predefined South African locality list; years are constrained between 1900 and the current year. The final enriched schema is stored in tabular format (CSV or relational DB) with versioning, allowing traceability and reproducibility.

Scalability & Ethics

The pipeline scales horizontally via distributed crawling and API-based enrichment. Batch processing and async retrieval ensure performance for thousands of companies. Ethical compliance is maintained by sourcing only publicly available data and avoiding any personally identifiable information (PII). Data integrity is safeguarded through source provenance tracking and transparent audit logs.

BY

MODEKURTI SRIHARSHA 22eg105d31@anurag.edu.in