

IDENTITY RESOLUTION EXAMPLE

John Johannaber

Hylaine Interview

April 2025

ETL PROCESS & IDENTITY RESOLUTION APPROACH

- **Overview:** To address the ETL scenario, the core of the solution hinges on **identity resolution**, specifically identifying and managing duplicates across and within incoming datasets.

PYTHON-BASED

- **Standardization:** Used smartystreets API to standardize addresses—key for reducing noise and enabling apples-to-apples comparisons.
- **Matching:** Applied fuzzywuzzy for fuzzy string matching on fields like company name and address.
- **Deduplication Logic:**
 - Compared new company lists *against each other* and *against existing CRM data*.
 - Applied thresholds (e.g., fuzz ratio > 80) to flag potential duplicates.
- **Output:** Generated a list grouping records as single entities

PYTHON-BASED

Pros:

- Fast to build and iterate on
- Very little costs
- Great for a POC or short-term/low-volume use cases

Cons:

- Lacks scalability for enterprise-scale data
- Needs ongoing support and maintenance
- Matching accuracy is sensitive to edge cases without additional tuning

ENTERPRISE-GRADE TOOLS

Amperity:

- I've used Amperity before and found it to be **robust, accurate, and enterprise-ready**.
- Pros:
 - Handles large volumes effortlessly
 - Built-in identity resolution logic with customizable matching rules
 - Highly available with no notable downtime in my experience
- Cons:
 - Significant upfront cost and implementation effort
 - Requires team buy-in and IT coordination

Other Vendors (e.g., Neustar, Informatica, Talend)

DECISION FACTORS

	Homegrown (Python)	Amperity/Others
Cost	Low	High
Speed to Deploy	Fast	Slower
Accuracy	Medium	High
Scale	Limited	Enterprise-ready
Support/Monitoring	DIY	Included
Long-term Viability	Good for tactical use	Best for strategic/ongoing needs

USE CASE FIT

- **If this is a one-time ETL event** for a few hundred or thousand records: My Python approach works well.
- **If this becomes a recurring ingestion or customer 360 effort:** A vendor like Amperity is the smarter long-term investment.