Università degli Studi di Roma Tre

Ingegneria dei Dati

Homework 5

# Data Integration

**Antonio Lanza, Daniel Luca**

**546969, 546357**

**Project repository**:
https://github.com/AntonioSouls/IDD_HOMEWORK5_Data-Integration.git

# 1. Introduzione

Nell'era della Data Science, l'integrazione dei dati è un elemento cruciale per garantire analisi accurate e decisioni informate. Le informazioni aziendali provengono spesso da fonti eterogenee, con formati e strutture differenti, rendendo difficile il loro utilizzo in modo efficace. Un sistema di Data Integration ben progettato consente di unificare e armonizzare questi dati, migliorando la qualità e l'affidabilità delle informazioni a disposizione delle aziende.

Il processo di integrazione dei dati si compone di tre fasi fondamentali:

1. **Allineamento degli Schemi (Schema Alignment):** consiste nell'identificare e mappare i campi comuni tra le diverse fonti di dati, creando uno schema mediato coerente;
2. **Collegamento dei Record (Record Linkage):** prevede il confronto e l'associazione di record che rappresentano la stessa entità ma provengono da database differenti;
3. **Fusione dei Dati (Data Fusion):** integra i record corrispondenti in un'unica rappresentazione consolidata, eliminando duplicati e risolvendo eventuali conflitti nei valori;

In questo studio ci siamo concentrati sulle prime due fasi, sperimentando e confrontando diverse tecniche, sia tradizionali che innovative.

# 2. Schema Alignement

Per questa fase ci siamo affidati all' utilizzo di un LLM (*Gemini-Flash-Thinking*) al quale abbiamo inviato le tabelle sorgenti e questo ha provveduto a fornirci una descrizione accurata di ogni singolo attributo per ogni tabella sorgente. Dopodiché, attraverso queste descrizioni generate, abbiamo chiesto al modello di mappare sotto uno stesso attributo tutti gli attributi delle sorgenti che avessero stessa descrizione. Dopo un ritocco manuale delle risposte che il modello ha fornito in questo scenario, siamo riusciti ad ottenere un ottimo **Schema Mapping** dal quale siamo partiti per poi costruirci l'intero **Schema Mediato** e popolarlo attraverso un semplice script Python

# 3. Record Linkage

In questa fase, invece, ci si concentra sul trovare tutti i record dello **Schema Mediato** che si riferiscono alla stessa entità. Per fare ciò, si sono dovuti elaborare i seguenti due step.

## 3.1 Blocking

Molto utile per evitare i confronti di un record con tutti gli altri della tabella, ma piuttosto si raggruppano in vari blocchi i record candidati ad essere simili per poi eseguire il **Pairwise Matching** solo sugli elementi in uno stesso blocco invece che su tutta la tabella, risparmiando tempo computazionale.
Si è deciso di effettuare due strategie di **Blocking**:

- **LOCALITY SENSITIVE HASHING** = Tokenizzare i nomi delle aziende per parole e fare un Hashing dei token così costituiti. Dopodiché, inserire in uno stesso blocco quelle aziende che condividevano lo stesso Hash;
- **TRI-GRAM BLOCKING** = Tokenizzare i nomi delle aziende per trigrammi e fare un Hashing dei token così costituiti. Dopodiché, inserire in uno stesso blocco quelle aziende che condividevano lo stesso Hash;

## 3.2 Pairwise Matching

E' la fase in cui si confrontano tutte le possibili coppie di nomi all'interno di un blocco. Anche in questo caso, tale operazione, si è deciso di effettuarla in due modi diversi:

- **RECORD LINKAGE TOOLKIT** = E' una libreria che mette a disposizione strumenti per fare il Pairwise Matching calcolando la distanza tra coppie secondo la metrica *Jaro-Winkler*;
- **DITTO** = E' una rete neurale che abbiamo addestrato al Pairwise Matching attraverso un Training Set costruito manualmente e poi eseguito sui dati delle sorgenti per fargli effettuare il matching;

## 4. Metriche Prestazionali

Le metriche ottenute sono espresse nelle seguenti immagini:

| | | |
|---|---|---|
| LOCALITY SENSITIVE HASHING | Jaro – Winkler | 3 min |
| | Ditto | 17 min |
| TRI-GRAM HASHING | Jaro – Winkler | 3,14 min |
| | Ditto | 18 min |

Figura 1: Tempi di esecuzione delle varie strategie applicate

| | | PRECISION | RECALL | F-MEASURE |
|---|---|---|---|---|
| LOCALITY SENSITIVE HASHING | Jaro – Winkler | 1 | 0,74 | 0,85 |
| | Ditto | 1 | 0,84 | 0,91 |
| TRI-GRAM HASHING | Jaro – Winkler | 1 | 0,76 | 0,86 |
| | Ditto | 1 | 0,87 | 0,93 |

Figura 2: Performance delle varie strategie applicate