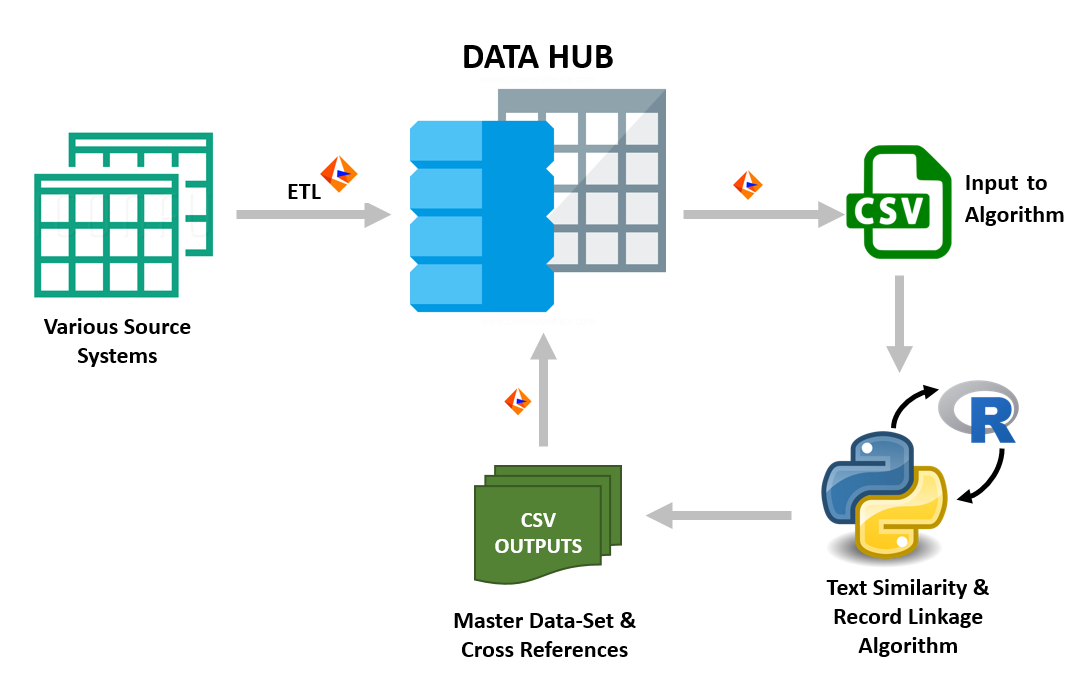
# **ARCHITECTURE**

The incoming clinical sites/ hospitals addresses dataset for this pharmaceutical client was gathered from more than 5 Source systems, with the aim to reconciliate the data flowing from different systems and collating a single source of truth across the board via generating a Master-dataset.

This Extraction-Transformation-Loading (ETL) process is done via Informatica, and the Site address related details are loaded into a single table, which is then leveraged to generate a csv input file for the text similarity & record linkage-algorithm. The algorithm outputs: Master dataset (golden-data set) and the cross references (Source-to-Master record linkages) are then loaded back into the database.



## **DATA COLLATION AND GENERATING INPUT FILE**

* The data gathered from the multiple sources is loaded into the data-hub in the below structure:



Unique Identifier for each-record

Country level batching –Wrapper for loop

Features relevant for similarity score calculation

* The above table is used as a Source to generate the input csv file for the Python Script. The data is filtered on the below two conditions before feeding it to the algorithm script:

1. Data processing is done on Country level batches, and hence Standardized Country Names is a mandatory field across the dataset and cannot be NULL.
2. For a record to be ‘linked’ another near-duplicate record, the sum of similarity score of each feature must cross a defined threshold. Therefore, the incoming records must at-least have a minimum number of ‘NOT NULL’ fields to be able to match the threshold score.