Proof of Concept: Central Data Catalog for Data Governance Client

Students

Benjamin Peter Kern

Fabio Elvedi

Advisor

Prof. Stefan F. Keller

Subject Area

Software

Project Partner

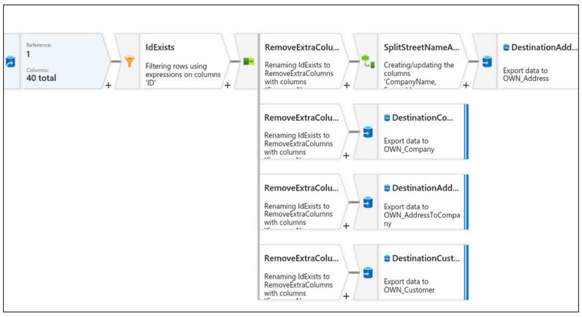
Data Governance Technologies Ltd., Wollerau, SZ

Introduction: DataGovernance Technologies Ltd (DGT) is a Swiss software company focused on automated analysis, classification and management of semi- and unstructured data. Their current solution uses structured data from various sources (CRM, ERP and similar) to establish business context. The data quality of these sources is lacking. Furthermore, data across sources is inconsistent which leads to inaccurate results. This project aims to solve this problem by automating the integration, cleaning and augmentation of these sources.

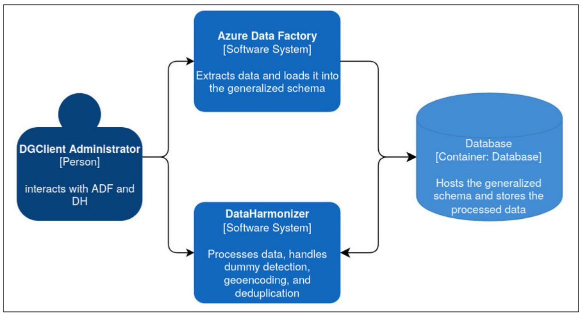
Approach: Since DGT serves many customers, each with different data sources, a modular approach was chosen. The project is separated into two main components: Azure Data Factory (ADF, low-code) and DataHarmonizer (DH, Python). This split provides two advantages: First, the ADF standardizes data into a generalized schema, offering a GUI frontend with simple transformation functions for clients with little technical expertise to integrate new data sources. Second, because all sources are transformed into the generalized schema, the DH can be applied universally, regardless of the customer's original source structure. The DH is responsible for data integration, dummy detection, general deduplication, and data augmentation through geocoding. To ensure real-world applicability and testability, data was generated using ERPNext and SuiteCRM, along with actual CRM and ERP data provided by a customer of DGT.

Result: The DH implemented the generalized schema containing Person, Company and Address. It successfully integrates and enriches data, addressing inconsistencies and enhancing governance. It implements dummy person and company object detection, geocoding, typo detection, and deduplication with high quality. Dummy objects are entries that do not represent an actual, individual identity or a legitimate company. For dummy detection, the model achieves an F1 score of 0.98, which indicates effective minimization of false positives and negatives. The evaluation, based on a manually labeled smaller dataset, indicates that the chosen methods and parameters are suitable. The total score combines a self-trained TF-IDF model, RegEx, heuristics, keyboard smashing detection, a blacklist, and the Zefix API for business name lookup. For person, the heuristic check includes making sure that the name has at least 2 characters, no more than 50 characters, and that the first character is uppercase. These scores are weighted and combined into a total score, optimized by testing 3125 combinations for both person and company detection models. Furthermore, 91% of customer address data was successfully geocoded using OpenStreetMap's Nominatim freeform API. It was determined that only 16% of addresses were unique. Fuzzy matching using the Jaro-Winkler algorithm on person names yielded 44% exact and 10% fuzzy duplicates with similarity > 0.95. For companies, 30% exact duplicates were found. A filter list was used to exclude keywords like "AG" or "GmbH" from matching. About 4% were fuzzy duplicates. The results show high number of duplicates due to the flattened CSV inputs and because they are stored in pairs (there are n*(n-1)/2 ways to pick a pair out of n duplicates). All "must" and "should" requirements were successfully implemented. The only "could" requirement - the customer behavior analysis - was skipped due to time constraints. In the coming weeks DGT will use the DH for a few pilot customers to integrate and clean contextual sources before data mining begins.

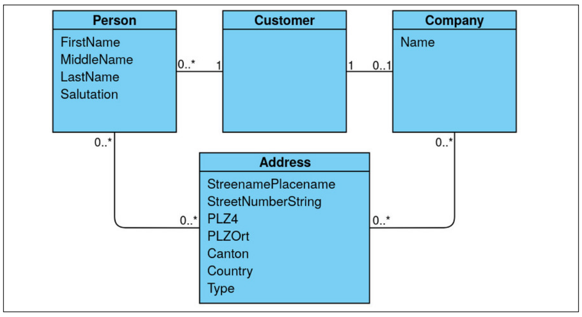
Azure Data Factory showing Data Flow for SuiteCRM. Transforms source data from a CSV to a generalized schema.
Own presentment


Reference:
1
Columns:
40 total
IdExists
Filtering rows using expressions on columns 'ID'
RemoveExtraColu...
Renaming IdExists to RemoveExtraColumns with columns
SplitStreetNameA...
Creating/updating the columns 'CompanyName,
DestinationAdd...
Export data to OWN_Address
RemoveExtraColu...
Renaming IdExists to RemoveExtraColumns with columns
DestinationCo...
Export data to OWN_Company
RemoveExtraColu...
Renaming IdExists to RemoveExtraColumns with columns
DestinationAdd...
Export data to OWN_AddressToCompany
RemoveExtraColu...
Renaming IdExists to RemoveExtraColumns with columns
DestinationCust...
Export data to OWN_Customer

System context diagram showing the overview of all systems and how they interact with each other.
Own presentment


Azure Data Factory
[Software System]
Extracts data and loads it into the generalized schema
DGClient Administrator
[Person]
interacts with ADF and DH
Database
[Container: Database]
Hosts the generalized schema and stores the processed data
DataHarmonizer
[Software System]
Processes data, handles dummy detection, geoencoding, and deduplication

UML class diagram showing the generalized schema into which the Azure Data Factory loads data.
Own presentment


Person
FirstName
MiddleName
LastName
Salutation
Customer
Company
Name
0..*
1
1
0..1
0..*
0..*
Address
StreenamePlacename
StreetNumberString
PLZ4
PLZOrt
Canton
Country
Type
0..*
0..*