SUCCESS STORY

« BRINGING DATA FROM DIFFERENT SYSTEMS TOGETHER USING FUZZY MATCHING »

- *Identifying reliable keys in unstructured data*
- *Identifying dublicates via Fuzzy Matching*
- *Creating a unified data model providing high data quality*

Initial Situation

A marketing consultancy was building a Data Warehouse (DWH) System and working on one unified data model which should provide data in high quality as the single source of truth. For analytical purposes, they needed to bring together data from different operational source systems. However, in some cases data that belonged to different departments couldn't be brought together so easily.

In this case their customer names were being typed into a CRM Tool (Hubspot) as well as into their Accounting Tool manually. There was no mapping between objects, shared IDs and company names were typed in freehand into the system (free text fields), therefore containing some spelling differences.
High level reporting was not possible, because the CRM data and accounting data couldn't have been combined.

Actions - Methodology and Technology

A fuzzy matching was used to combine the data from the two different sources. A selection of fuzzy string-matching algorithms was tested, for example Jaro-Winkler Distance, Levenshtein distance, Soundex or cosine similarity. The open-source algorithms can be very efficient and there is a selection to choose from depending on the use case.

Successes and Results

Within only a few iterations of feedback from the company, we could adjust the chosen algorithms and parameters to give the best results for percentage of hits and false positives that was optimal for the client´s needs.

The data from the CRM and Accounting systems were brought together and could have been integrated into one valuable data model.

The matching was integrated into the DWH pipeline with a Google Cloud Function (a serverless solution for GCP to deploy Python code). This solution is incredibly cost efficient and allows the usage of broad variety of algorithms and methods for endless purposes.

What this means for you

Not being able to combine datasets can bring a project to a holt. It can also deprive a company of a possibility to gain important insights or making informed decisions based on data. This can be however easily solved using fuzzy matching algorithms.

Spelling mistakes are one of the most common problems influencing data quality and reducing it substantially. Implementing even a simple fuzzy matching algorithm can minimize or eliminate this problem and can be implemented in a very efficient way using serverless computing on the cloud (Google, Azure, AWS).

Oftentimes where data from two datasets or system needs to be brought together it is thought that it might as well be done manually because it will most likely need to be done once. However, in majority of the cases the problem is recurring, and it is more efficient to automate it.

Are you generally interested in learning more about Data Modelling, Data Fusion or Fuzzy Matching and would like to leverage hidden potential? Would you like to discuss an individual problem with us? Visit us at **www.datanomiq.io** or send us an email to **info@datanomiq.de**.

DATANOMIQ
Applied Data Science

DATANOMIQ is the independent consulting and service partner for business intelligence, process mining and data science. We are opening up the diverse possibilities offered by big data and artificial intelligence in all areas of the value chain. We rely on the best minds and the most comprehensive method and technology portfolio for the use of data for business optimization.