Sanmargar DQS

Data Cleansing and Standardization

Product information and functionality

Sanmargar Team sp. z o.o.

2016-02-10

Sanmargar Team sp. z o.o., ul. Łukowska 1 lok. 133, 04-113 Warsaw, Poland;
office@sanmargar.com, www.sanmargar.com

Table of contents

Functionality **2**
Major aspects of the solution's functionality 2
Technological aspects 2
Key features 3
Benefits and applications **4**
Key benefits 4
Summary of benefits and applications 4

Functionality

Sanmargar DQS (Data Quality Studio) is a solution developed by Sanmargar Team to support validation, standardization and cleansing of customer address and other customer attributes.

Major aspects of the solution's functionality

The Sanmargar DQS solution provides Clients with comprehensive data validation, standardization and correction services. Predefined Sanmargar DQS algorithms make it possible to cleanse postal addresses, first names and surnames, business names, e-mail addresses, telephone numbers, identification information of individuals and businesses. Depending on the data type, cleansing involves, for instance, soft pattern matching techniques, reference data dictionaries and regular expressions.

The high configurability of Sanmargar DQS makes it possible to control cleansing algorithms and result confidence levels, and create non-standard, dedicated solutions for cleansing or classifying other attributes of any data sets, e.g. product names and codes.

A reference data library, comprising Polish address dictionaries and international dictionaries of first names and surnames, is maintained and developed for the purpose of the Sanmargar DQS solution. Reference databases rely not only on official sources (databases administrated by government agencies), but also on our own experiences, eliminating inconsistencies and errors encountered in those databases.

Sanmargar DQS may work with many data sources, databases, and any ETL software. The use of Web Services makes it also possible to validate data entered by users online. These features enable easy integration of Sanmargar DQS with other solutions, including CRM systems, CCF (Central Customer File) systems, e-commerce, etc.

Technological aspects

The current version of the Sanmargar DQS solution comprises:

a) A proprietary Sanmargar module which makes it possible to:

- intelligently parse text data to identify matching patterns based on regular expressions and reference data dictionaries;
- validate identified elements using control algorithms;
- search reference patterns which are the most similar to the identified elements and assess their similarity.

b) A library of reference data dictionaries, comprising Polish postal addresses, international names, surnames, and telephone area codes, adapted in order to enhance the effectiveness of data cleansing algorithms;

c) Predefined, flexible customer data standardization and cleansing processes.

Sanmargar DQS requires a PC architecture computer with an operating system which supports 64-bit Java programs (v1.8). Depending on data volumes, the machine should deliver at least two processing threads and have 4GB of RAM. A typical environment for multi-million volume data migration is four to eight processing threads and 16 to 32 GB of RAM.

Sanmargar DQS uses a PostgreSQL database as a temporary working database. Based on data contained in that database, it is possible to obtain data cleansing statistics. It is also possible to integrate that database with another.

Sanmargar DQS may work with ETL-class solutions (e.g. Pentaho DI, SAP DataIntegration, Talend*).

Key features

- Validation, standardization and correction of postal addresses, first names and surnames, business names, e-mail addresses, telephone numbers, identification information;
- Comprehensive dictionaries of postal addresses, first names and surnames;
- High algorithm configurability;
- Wide range of supported data sources and formats.

Benefits and applications

Key benefits

Sanmargar DQS has been used in several projects involving the migration of data to new ERP/CRM-class systems. Using Sanmargar DQS, we performed the processes of integration, cleansing and deduplication of data obtained from several installations of the legacy systems in order to load it into the target system. Our use of Sanmargar DQS in those processes enhanced the effectiveness of data deduplication many times over.

Sanmargar DQS may be used in a process of customer data quality improvement. The role of Sanmargar DQS in this context is to improve, add and standardize data processed in the main transaction system, enabling the introduction of improved data controls at the time of data input.

The Sanmargar DQS solution may be used as a pre-configured module working with a CCF (Central Customer File) class system, responsible for the cleansing and standardization of address data obtained from a source systems. Thanks to Sanmargar DQS, the integration of data and its transformation into a common data model (CDM) is highly effective and the effectiveness of customer deduplication mechanisms is dramatically improved.

One of the most important benefits of using Sanmargar DQS is gaining knowledge of the quality of any existing customer, supplier, business partner data sets. Sanmargar DQS may also be used to assess the quality of data sets submitted for processing by our business partners. Such knowledge, even without any active correction of errors or filling of gaps, makes it possible to identify data which does not meet the quality criteria set and to take further steps to improve it.

Although the best practice is to prevent the emergence of errors where data is input, this is not always possible or economically reasonable. In such cases, Sanmargar DQS provides a mechanism for active data correction ex-post - involving a one-off or regular service or a pre-configured ongoing solution.

Thanks to Sanmargar DQS, it is possible to reduce operating costs generated by incomplete or incorrect data, including costs of incorrectly addressed or duplicate mail but also legal or business consequences of a failed mail delivery to a customer, partner or debtor. Such costs should also include consequences of an incorrect assessment of exposure to each client where there is duplication of data which cannot be removed without earlier standardization and cleansing.

Summary of benefits and applications

- Integration, cleansing and deduplication of client and supplier data migrated between ERP/CRM systems;
- One-off, regular or on-going improvement of customer data quality;
- Standardization, cleansing and deduplication of data in CCF (Central Customer File) class systems.

Sanmargar Team

Sanmargar Team is a company which focuses on Business Intelligence and Data Integration solutions supporting client management, sales and finance at financial, power, mining and modern universal service sector corporations. Sanmargar Team is also a provider of solutions for reference data management – Metastudio DRM, address data deduplication and cleansing – Sanmargar DQS (Data Quality Studio), solutions supporting centralized management of customer data – Sanmargar CKK (Central Customer File System), and other products supporting data processing.

Contact:

Sanmargar Team sp. z o.o., 04-113 Warsaw, Łukowska 1 / 133, Poland;
phone: +48 22 115 80 71; E-mail: office@sanmargar.com; www.sanmargar.com