Getting Started with SAP BusinessObjects Data Quality

Data Quality, part of SAP’s Data Services, is a powerful tool for cleansing and matching customers, businesses, postal addresses, and much more. However, its wide range of options and capabilities can be intimidating for new users. In this series of posts, I will focus on the key transforms and settings required to get your first cleanse up and running.

What is Data Cleansing?

Data Cleansing in the context of SAP Data Quality covers a range of activities including cleansing, standardization, and supplementation. Using dictionaries and reference files, it not only repairs and reformats, but in some cases also completes and fills missing data.

Cleansing broadly refers to Data Quality’s ability to identify and flag incorrect or suspicious data. This will catch invalid phone numbers, non-existent postal addresses, improperly formatted email addresses, and incorrect dates, as well as common misspellings of names and businesses.

Standardization gives you the ability to correct or reformat your data. For example, Data Quality can provide a list of suggested first names, correct the spelling of a city, or standardize phone number and address punctuation.

Supplementation is most notable with postal data, where partial records often contain enough information to determine the missing field, such as zip code uniquely identifying city and state.

Installation Tips

Two aspects of Data Quality requiring special attention during installation are the Cleansing Package and the Reference Data files.

The Cleansing Package is included in the overall Data Services install, but will not be available unless you have first downloaded the correct PERSON_FIRM.zip file from the SAP Service Marketplace and placed it in the proper installation directory. Once this is complete, “Cleansing Package” will be available as a feature during Data Services installation.

After installation, the Reference Files must be downloaded separately from SAP Service Marketplace, unzipped, and placed in \SAP BusinessObjects\Data Services\DataQuality\reference_data\.

Choosing the Appropriate Transform

Data Quality offers an intimidating selection of transforms, but choosing the right one is simple once you’ve identified the data you need to cleanse.

Postal Data

For postal or address data, the Address Cleanse transforms are your best choice. These transforms are divided into three categories, but one is sufficient for most addresses within the USA:
USA_Regulatory_Address_Cleanse. Of the others, Geocoder is useful when assigning geocode data without additional cleansing, and Global_Address_Cleanse is similar to the USA cleanse, but with country-specific transforms to accommodate foreign addresses. For simplicity, we are limiting our scope to just the USA.

Within the USA_Regulatory_Address_Cleanse section, you may see several transforms pre-configured for specific types of cleansing such as NCOA, EWS, and SuiteLink. We will focus on the base transform, USARegulatory_AddressCleanse.

Non-Postal Data
Nearly all non-address data is handled by the Data_Cleanse transforms, including first and last names, titles, firms, dates, email addresses, and phone numbers. Again, in this section you may see several transforms pre-configured to accommodate different languages, but we will limit our scope to the EnglishNorthAmerica_DataCleanse.

In the next post, we will examine how to configure these transforms to best get you up and running quickly with Data Quality.

Bruce Labbate
Business Intelligence Consultant
Decision First Technologies
bruce.labbate@decisionfirst.com

Bruce Labbate is a business intelligence consultant specializing in data warehousing, data quality, and ETL development. Bruce delivers customized SAP BusinessObjects Data Services solutions for customers across all industries. With Decision First Technologies, Brett utilizes Data Integrator, Data Quality, Information Design Tool, and a variety of database technologies, including SQL Server, DB2, Oracle, Netezza, and HANA.