Data quality assessment on business partner data in a SAP environment

Niels Weigel (niels.weigel@fazi.de)
Joachim Schmid (joachim.schmid@fazi.de)

FUZZY! Informatik AG,
Eglosheimer Strasse 40, 71636 Ludwigsburg,
GERMANY

Description of the project

In a recent survey by PWC managers cite data quality as a major reason that 88% of CRM, e-business and system migration projects fail to deliver on time and within budget. Even inside existing and long running CRM systems one can not be sure that all information stored in the database are completely “fit for use”. For that reason one of our customers decided to start a project for data quality assessment. During the data quality measurement of the SAP business partner database one task was to examine whether the data correspond to the admitted data quality requirements and to analyse as well how frequently already well-known data quality faults exist. Additionally data errors which are so far not documented are recognized. Based on those errors together with statistical analysis it was verified whether additional valuable information (mobile telephone contained in district …) can be generated.

During that project regularly reviews of the so far obtained results were accomplished together with the IT department of the customer. During these reviews some of the found "unknown remarkableness" could be identified as "actually admitted not-documented remarkableness". The causes for these data errors were quite conscious thereby the IT department, however the meaning for the data quality not so transparent.

Following the data quality measurement a final report could be delivered over the status of the current data quality of the business partner data. The report contains some hints, how the data storage and the business processes, which are connected with the business partner data, can be improved. In addition the impact of a pure data cleansing process on the existing data was represented.

1 Introduction

This project reports on a information quality project. The project was initiated during a product presentation when there where some discussions about the question:

"Will the implementation of a standard address validation and a deduplication tool be the solution of all of our problems?"

The first ones feeling that there are some information quality problems were the employees of the marketing department where the success of some campaigns (e.g. mailings, lead generation, …) was not as big as suggested. During some mailing and marketing campaigns there was a growing unpleasant feeling that something in the business partner data base is wrong. So this department started an internal project for basic analysis of data defects and received some fundamental information about their problems.

This basic analysis of the existing database mainly included terms of contextual IQ like completeness of the data and the amount of information inside a record. Also there was a basic analysis on the content of some data fields. Obviously there were (some) violations of the domain or column integrity because the contents of the data field was not what it should be according to the data model. Situations like "mobile phone number included in name2" or "contact information included in city district name" were quite common, but there was no statistical analysis how often those data defects appear. And if you do not know what other data defects beside those obviously visible defects are included in the data base you can not just implement a data cleansing software tool. A tool for postal validation will populate the city district column/field with correct values. Doing so you will gain column integrity. But what's about the old content of the field. Sure, it was no correct value for the city district column. But are you sure the old value was rubbish, which should only be replaced by a correct value? After a close look to the old values we recognized, that the old values are very valuable for the company. In most of the cases, when the value was wrong, it was because it holds contact information like mobile phone numbers, … instead of the name of a district. You are not sure what information will be deleted or overwritten by an automatic data cleansing program.

Also in some cases for the data custodians it was quite clear why some information are stored in different fields. Due to limitations in the software systems of the data collectors they have to use other data field to capture information because in those systems the data base model was slightly different to the database model of the SAP system. And on the other way to the data consumers the data models of the following external processes were also different from the SAP system so that it was quite easier to store wrong information in wrong fields than to build up a message format on demand for the external process.

So a deeper analysis of the information quality was necessary. The decision for such a data quality offensive was pushed with the support of the management. Starting a data quality assessment workshop for the rating of the business partner data quality with the target to get a clear view and measurement of the actual data quality level.

2 IQ project kick off

The first meeting was set up to clarify the basic requirements and expectations on that IQ project. The different steps have been defined and a project plan has been set up.

2.1 Definition of the IQ team

The IQ team has been set up consisting of employees from the customer from the IT department and the marketing department and of external consultants from FUZZY! Informatik AG.

2.2 Definition of the data production system

The data collectors are the sales persons of the company working and editing the single records in their sales system which is from the point of view of the SAP system a legacy system from which the business partner records are synchronized to the SAP business partner database.

The data custodians responsible for the storage and maintenance of the data is the IT department, mainly responsible for the synchronisation processes from the legacy system to the SAP system and from the SAP system to external systems like the external shipping partners who are responsible for the delivery of the products.

The data consumers who are using the data are the accounting and marketing department, the external shipping partners, and the management for reporting.

2.3 Definition of the data fields

Also the decision on the data fields which will be assessed during the project has been part of the kick off meeting.

The business partner data fields which have been assessed have been:

Name of data field	Description of data field
Name1	Company name part 1
Name2	Company name part 2
Name3	Company name part 3
Cityname	Name of the city, locality
Subcityname	Name of the district of the city
Streetname	Name of the street followed by a building number
POBox	Keyword to indicate a following P.O.Box number
Region	Name of the federal state

3 Data quality assessment

3.1 Definition of quality metrics/rules or criteria

After the decision of the company to start an external information quality project first the data fields of the SAP system which should be monitored have been defined .

3.1.1 Definition of the data base fields and the metrics for the assessment

First of all for the data quality workshop we defined together with the customer the data fields used for the data quality assessment. From the SAP business partner database the standard name fields, address fields and communication fields have been selected.

3.1.2 Subjective metrics

Questionnaires for the IT department for any know defects from further data migrations or data synchronisations between the legacy system and the SAP system.

3.1.3 Objective metrics

Concerning the database integrity rules we had a fix defined database model provided by the SAP system. This means that Domain and Column Integrity, Referential Integrity, Entity Integrity are more or less handled by the SAP system itself.

3.2 Data assessment

The first part of the data assessment was the creation of a meta database with the original content of the records from the business partner database and additional information like:

- statistical values (min, max, length, …)
- formatted strings (e.g. “NNNNN Aaaaaaaaa” for “70123 Stuttgart”)
- type strings (e.g. “{F} {L} {l} {Ct}” for “John Smith Illumination Ltd.”)
- flags (IsMobilePhone, IsDeliveryPoint, …)

Afterwards those additional columns have been used for statistical analysis and validity checks to identify the most common errors inside the business partner database.

3.3 Reasons for data defects

After having identified the main defects in the business partner database the next step was to find out the reasons for those defects. This was an interactive way of discussion and analysis together with the customers different departments. Some of the reasons have been well known due to already existing negative experiences or earlier restrictions, some of the data defects have been so far unknown.

3.4 Changes and their influence

To improve the information quality of the data base the basic process of data cleansing can be one option. But you always have to take into account which influences such a cleansing process can have on the whole data base. You can for example easily clean up the address fields of a data base entry using a postal address validation tool and with

that step you will get a nearly 100 % accuracy and actuality on the address level. But if you address validation tool replaces a mobile number included in the district name by the correct district name you will loose an important information and the quality on your record level will decrease.

So it was quite important to take a close look on the influence of every first-look quality improvement step on all the other data fields or information units.

3.5 Documentation

All the results of the data quality assessment have been documented in a report which can be used by the customer for internal purposes.

4 Conclusion

After the data quality assessment and the final workshop there was a common understanding of the actual data quality level established. The customer had a clearer view on the actual data defects of the business partner data base. Especially they had some good examples about “what” is wrong inside the data base not only the feeling that there is “something” wrong. But what was also a clear output of the assessment and the business process analysis was the fact that simply implementing some address cleansing software components will not solve all the problems. In some parts the information quality will be even worser bader after cleansing the address components because some additional information (e.g. mobile phone number in the district field) will be deleted.

Remarkable is that different groups of persons have a different understanding of fitness for use. And thus a different understanding, which data is of high data quality. The IT department responsible for the SAP system states that data is of high data quality, if each data field contains, what the name of the field is promising. E.g. the field named district should contain a name of a district.

The external shipping partners need to have the mobile phone number printed at the delivery note. But only very few data fields were transmitted from the SAP system to the external shipping partners. The district field is one of them, any phone number fields are not. From the point of view of the shipping partners the data is fit for use, if the data field district contains the mobile phone number. Because then the mobile phone number is printed on the delivery note, and thus available for the drivers.

5 Next Step

The next steps in this information quality project will be the documentation of the IP Map on the site of the customer so that they have a clear view on all the business processes and data flows from the collectors to the custodians to the consumers. This will be essentially for the decision on any data cleansing step inside the business partner data base. Only with the IP Map you can easily track the influence of any changes in the data storage model on all the other components of your information production system.

Beside that process analysis as a fast support for the marketing department to optimise the address data quality of the mailing campaigns the implementation of a batch processing program in the marketing department is started. All addresses for mailing and lead generation purposes will be standardized, validated, and deduplicated for each individual campaign. The data quality requirements of the marketing department are met by that solution, the SAP data base and all processes in the SAP system are not affected by that solution.

After that and with the migration of the SAP system from release 4.5 to release 4.7 the next step will be the online integration of postal address validation and deduplication programs for capturing new addresses into the SAP system. With that milestone all new addresses will be well qualified and the number of new duplicate records will be smaller.

Together with the documentation of the IP Map and the analysis of the influences of a data cleansing process of the SAP business partner data base on all follow up processes an one time cleansing process for the data base will be initiated. With that step an ongoing monitor process of the data base will be implemented to. Using this monitor process the customer has the ability to track the data quality permanently and can always react on any negative tendencies.

The time frame for all the next steps documented here will be within the next four years. This is as always the fact that a real information quality project is not a short time project but a long term change of data and processes inside an organisation on the road to successful business.