Data Quality and Data Trust in IBM Master Data Management System

Przemyslaw Pawluk
supervised by Jarek Gryz

Department of Computer Science and Engineering,
York University,
Toronto ON

Introduction

A good business data model has little value if it lacks accurate, up-to-date customer data. This paper describes how data quality measures are processed and maintained in IBM InfoSphere MDM Server and IBM InfoSphere Information Server. It also introduces a notion of trust, which extends the concept of data quality and allows businesses to consider additional factors, that can influence the decision making process. The solutions presented here utilize existing tools provided by IBM in an innovative way and provide new data structures and algorithms for calculating scores for persistent and transient quality and trust factors.

MDM Environment

MDM provides the technology and processes to manage Master Data in an organization.

Master Data

Master data is typically **high value information** that an organization uses repeatedly across many business processes. It is the data an organization stores about key elements or business entities that define its operation which are:

- Custommers
- Products
- Locations
- etc.

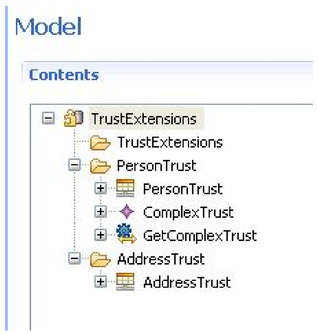

Figure: A model for handling trust factors in MDM

MDM Features

- Consolidate data locked within the native systems and applications
- Manage common data and common data processes independently with functionality for use in business processes
- Trigger business processes that originate from data change
- Provide a single understanding of the domaincustomer, product, account, locationfor the enterprise

Problems with Master Data

- Fragmentation across many applications,
- Inconsistent copies,
- Delays in update propagation

MDM AND Information Server

InfoSphere MDM Server

- Manages Master Data,
- Allows users to define and add quality and trust factors,
- Allows user to modify the data model in a simple way

IBM Information Analyzer

IA is a new module of IBM Information Server for data profiling and analysis.

- Helps in exposing technical and business problems
- Helps the expert in systematic analysis and reporting of results
- Allows experts to focus on the real problem of data quality issues

IBM Information Server

- Addresses the requirements of cooperative effort of experts and data analysts with an integrated software platform
- Provides the full spectrum of tools and technologies required to address data quality issues
- Supplies users and experts with the tooling that allows the detailed analysis of data through:
  - profiling,
  - cleansing,
  - data movement and transformation

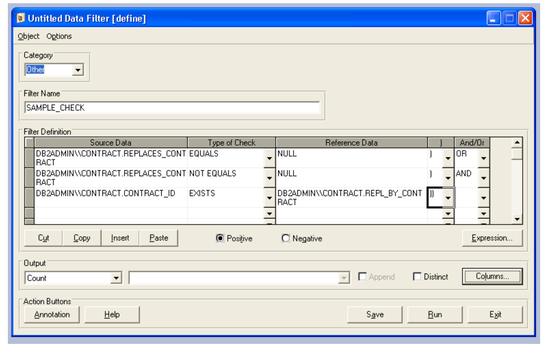

Figure: Sample rule definition in AuditStage

Usage Example – Typical scenario in an insurance industry

- Insurance company stores MD such as:
  - Customer, which can be person or organization
  - Contracts (variety of insurance policies, such as home, life or
- The company keeps information about its employees;
- Users may extend provided data models adding required attributes and actions
- Constraints may be imposed by rule generation
  - Formatting rules describing different formatting issues like length, allowed characters etc.
  - Integrity constraints standard constraints in database systems.
  - Business rules any other rules, for example, dependencies among different fields and values

Data Quality

Data quality is an aggregated value of multiple IQ-criteria such as:

- Accuracy
- Completeness
- Consistency
- Timeliness

Data Trust

- Extension of data quality
- Strongly depends on the user requirements and usage context
- Calculated on different granularity levels (field, row, table)

Examples of trust factors

$$D(key) = max\{0, 100 - 10 * x_{DuplPerctg}\}$$

$$Acc(t_i) = max\{0, 1 - \frac{|t'_i - t_i|}{t'_i}\},$$

Figure: $D(key)$ - duplicate rate for the key, $Acc(t_i)$ - accuracy of some field t_i

Trust Factors

- Data Lineage
- Origination
- Traceability
- Stewardship Status
- Authentication
- Believability
- Reputation
- Reliability

Trust Processing

- Usage of existing tools with minor modifications,
- New data structures to store metadata describing data quality and trust
- Mechanisms for assessing some of the quality and trust factors
- Extension of the existing data with trust/data quality factors

Persistant Factors

- Stored in a database,
- Data extension .
- Off-line acquisition (We use IA and AS to acquire trust scores)

Transient Factors

- Calculated on-line
- Behavioral extension

Conclusion

Measuring data quality and data trust is one of the key aspects of supporting businesses in decision making process or data stewardship. Master Data Management in other hand supports sharing data within and across lines of business. In such case trustworthiness of the shared data is extremely important. Our investigation has resulted in consistent method of gathering and processing quality and trust factors.

Acknowledgements

We would like to thank Paul van Run, Stephanie Hazlewood and Guenter Sauter for valuable discussions and help.
This work has been supported by IBM Center for Advanced Studies and NSERC.

http://www.cse.yorku.ca/~pawluk pawluk@cse.yorku.ca