Automated Data Quality Monitoring

(Research-in-progress)

Lisa Ehrlinger[*†] and Wolfram Wöß[*]
[*]Johannes Kepler University Linz, Austria
[†]Software Competence Center Hagenberg, Austria
{lisa.ehrlinger | wolfram.woess}@jku.at

Abstract: Most existing methodologies agree that the assessment of data quality (DQ) is a cyclic process, which has to be carried out continuously. Nevertheless, the majority of DQ tools allow the evaluation of data sources only at specific points in time, and the automation and scheduling is therefore in the responsibility of the user. In contrast, *automated DQ monitoring* allows the evaluation of applied DQ improvements as well as the comparability between different system states. The reproducibility of DQ assessments is also an important topic for the scientific community in order to review algorithms that improve the DQ of an information system. We are developing a tool for DQ monitoring and our research covers the investigation of suitable DQ metrics for continuous monitoring as well as the development of a standardized approach to storing DQ assessment results over time. In addition, statistical methods to analyze and visualize the resulting time series data are selected and applied.

Keywords: Data Quality, Information Quality, Monitoring, Data Quality Assessment

INTRODUCTION

Although the interest in data quality, from both research and industry, has increased over the last decade, it is still an underestimated topic in operational information systems. From a management perspective, projects that deal with data analytics or data science attract more attention and reward than DQ assurance and data management. Data quality assessment and assurance is however a critical requirement for data analytics, because the quality of data analytics results depends directly on the quality of the underlying data.

The Data Management Association (DAMA) describes data quality management as the analysis, improvement and assurance of data quality, which is not a task that can be carried out once, but includes the ongoing quality assurance when data evolves over time (DAMA 2008). The importance of cyclically carried out quality assurance is also underpinned by most existing methodologies, for example, the Total Data Quality Management (TDQM) methodology developed by the MIT (Wang 1998), AIMQ (Lee et al. 2002), the Total Quality data Management (English 1999), or the management approach by Redman (1997). In addition to the development of methodologies, a number of algorithms has been developed and tools have been implemented to assess and manage the quality of data in information systems. Typical DQ tools support data profiling in various complexity at a specific point in time that can be repeated for subsequent DQ analyses or optionally a continuous verification of data against user-defined rules. However, there is no documentation on how to systematically store data quality assessment (DQA) results over time and reuse them for benchmarking and long-term analyses at later stages.

The main contribution of our research work is the investigation of computational capabilities for the continuous DQ monitoring of real-world information systems. Hence, we aim at creating a preferably generic approach that can be applied for different data models, while special consideration is given to the most widely used relational data. The research work includes the recommendation of suited DQ metrics that are especially meaningful to make statements about the temporal development of DQ as well as the development of a concept for optimized management and storage of DQA results. We evaluate a set of statistical functions in order to analyze the resulting time series data and yield new insights by a suggested visualization. Five research questions summarize the open issues that we identified through literature search and that are addressed by our research work.

- To what extent offer state-of-the-art data quality tools data quality monitoring capabilities?
- Which data quality dimensions and metrics are practically suited for time-series-based monitoring, and which are not?
- What are the requirements and obstacles for storing data quality assessment results as well as metadata, and which technology is most appropriate for the implementation?
- Which statistical functions are most suited to perform time series monitoring?
- Which kind of visualization optimally supports illustration and, thus, understanding and interpretation of data quality analysis over time?

The remainder of this paper is structured as follows: “Data Quality Monitoring and Related Work” describes the different interpretations and applications of the concept of data quality monitoring (DQM). The following section about the “Automated Data Quality Monitoring Application” covers the architecture of the application that we are going to implement and shapes our vision of automated data quality monitoring. The subsequent section “DQM Capabilities in Existing Approaches” presents an ongoing survey about existing data quality tools to analyze their capabilities of continuous DQ monitoring. Finally, a summary of the gained insights and open research questions is provided in the conclusion, along with an outlook on future work.

DATA QUALITY MONITORING AND RELATED WORK

Despite different existing interpretations, the term *data quality* is most frequently described as “fitness for use” (Wang and Strong 1996), referring to the high subjectivity and context-dependency of this topic. *Information quality* is often used as synonym for data quality and even though both terms could be clearly distinguished, because “data” refers to plain facts and “information” describes the extension of those facts with context and semantics, there is no clear consensus about their distinction in literature (Zhu et al. 2014). We use the more common term data quality because the focus is on processing automatically retrievable facts.

The narrower term *data quality monitoring* is mainly used implicitly in literature without an established definition and common understanding. This leads to different interpretations when DQM is mentioned in scientific publications or by companies promoting and describing their DQ tool. In order to clarify the term, we distinguish between three different views on DQM: (1) the methodical view, (2) the data view, and (3) the data quality view, which is promoted in our research work. The methodical view consists of managerial approaches and methodical frameworks that suggest organizational control mechanisms to continuously monitor data quality. A prominent example is the AIMQ (Lee et al. 2002), which was intentionally built to support monitoring DQ improvements over time. Apel et al. (2015) also share this view by declaring DQM as a framework to continuously control the quality of data in an information system, for example by using metrics, reports, or by periodically performing data profiling. In order to explicitly differentiate the proposed work from methodical approaches and to highlight that the monitoring workflow should be executed by an application, the keyword *automated* has been added to the title of this research-in-progress paper.

Another view is transmitted by practitioners and industrial DQ tools, in which DQM is often implemented by continuously checking a set of rules against the data and possibly sending alerts if specific thresholds are exceeded. Although those solutions are often designated as “data quality monitoring”, they actually verify and monitor the data itself, hence categorized as data view in our context. This understanding is a first insight from the ongoing DQ tool investigation described in the remainder of this paper, and also expressed by Gartner (Judah et al. 2016), who describe monitoring in DQ tools as the "deployment of controls to ensure that data (static and streamed) continues to conform to business rules that define data quality for an organization". This type of monitoring can be understood as data monitoring. A practical example for an existing data monitoring system is a documented implementation at the CERN in Geneva, which is used to monitor physical data records in the course of the Compact Muon Solenoid (CMS) experiment (Tuura et al. 2010). The monitoring system performs periodic data quality checks, for instance, if the mean value is within an expected range, and use histograms for visualization.

In contrast to data monitoring, our research objective is to monitor the development of the DQA results over time, which can also be understood as the supervision of data quality. A DQM system continuously collects data quality meta data, for example, metrics for completeness on different aggregation levels. The amount of metadata can then be evaluated in order to derive new insights in the qualitative development of the data in an information system.

Data quality monitoring also differs from *continuous DQ assessment*, where the target is to calculate a single DQ score (per metric) from a set of observations (Naumann and Rolker 2000). An example for continuous DQ assessment is the determination of the availability dimension by investigating a set of attempts to access the information system under observation. In contrast to achieving one final DQ score, the aim of our research work is to compare and analyze a set of historical and current scores for each DQ metric, in order to give evidence of its development over time. A documentation on how to systematically store and reuse the amount of DQA results is vital for managing and mastering the data quality in companies and organizations over a longer period.

AUTOMATED DATA QUALITY MONITORING APPLICATION

In the course of our research work, we are going to implement an application to monitor the data quality of heterogeneous information systems over time. The architecture of this application, illustrated in Figure 1, consists of four components: (1) collection of metadata and computation of DQA results (i.e., DQ assessment), (2) the data quality repository, (3) analysis of the resulting time series data, and (4) a user interface to track the data quality development of the observed system. The following subsections describe each component in more detail.

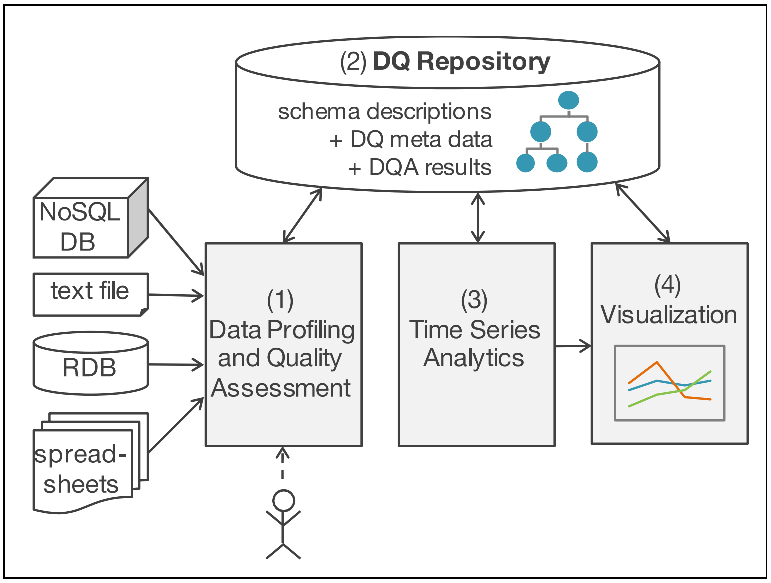

Figure 1: Architecture of Data Quality Monitoring Application

The proposed application architecture in Figure 1 is a proof of concept for our research towards automated data quality monitoring. The main contribution forms the investigation of data quality dimensions and metrics as well as methods for time series analysis and to which extend they are suited for continuous DQ monitoring. This knowledge empowers users of existing data quality tools to extend their DQ assessment manually with functionality for continuous monitoring, for instance, by writing batch functions to collect specific DQ metrics regularly.

Data Profiling and Quality Assessment

The gathering of metadata, referred to as *data profiling* (Naumann 2013), is a prerequisite for calculating specific data quality metrics of an information system. Although a dedicated data profiling application like Metanome[1] could support the metadata extraction process, the focus of component (1) is mainly the calculation and storage of DQA results. Experiences from such tools, especially concerning incremental and continuous data profiling, are a vital starting point for the development of data profiling in this component. Metadata stored in the DQ repository is termed *data quality metadata* (DQMD), because it supports the calculation of DQ metrics or can be directly provided DQ ratings to a user. Storing DQ metadata is also of great advantage for performance-intensive computations like the identification of functional dependencies (Naumann 2013). The data profiling and data quality assessment component (1) enables initial user input of domain-specific information and the specification of DQ metrics including their monitoring frequency.

The DQA component calculates a set of data quality metrics and stores the results in the DQ repository. Data profiling results that are not required for the analysis in the time series component (3) or the visualization component (4) will only be temporarily calculated and not persisted in the DQ repository. Data quality is traditionally perceived as a multidimensional concept that is characterized by a large number of different aspects, so called *dimensions* (Pipino et al. 2002; Wand and Wang 1996). Every dimension can be assigned to one or several *metrics*, which are functions that map the quality dimension to a numerical value. This value allows an interpretation of the fulfillment of a DQ dimension (IEEE 1061-1998). In our approach, we focus on data quality dimensions that refer to the actual values of an information system (i.e., the extension). The quality of the schema (i.e., intension) of an information system is expected to remain stable over short time periods and will only change in the course of redesigns of the observed system. Modifications on the schema-level can be managed by version control.

The major challenge in the development of the DQ collection component is the selection of metrics that are suitable for continuous monitoring, which means, that a repeated calculation is possible and the resulting time series have an informative value. The basis for our investigation are metrics for the most commonly applied DQ dimensions *accuracy*, *completeness*, *consistency*, and *time-related* data quality dimensions (Batini and Scannapieco 2016). We do not include DQ metrics that require a gold standard (i.e., reference data) in the proposed approach, since there exists usually no gold standard in operative information systems and a repeated calculation of such a dimension is therefore not possible.

Hipp et al. (2007) introduced a promising approach for measuring the accuracy by means of rule based outlier detection, where rules are automatically derived from the data itself. Such an approach is well suited for our vision on DQM, since it gives evidence on the development of outlying values over time by solely investigating the data itself. Example metrics for measuring the completeness dimension are simple ratio metrics between the percentage of existing values and missing values (in terms of population or attribute values), or the comprehensive approach for measuring completeness in an integrated information system introduced by Naumann et al. (2004). The comprehensive determination of a set of generic DQ

[1] https://github.com/HPI-Information-Systems/Metanome (August 2017)

metrics to repetitively capture standardized DQ assessment results is one of the major challenges of this research.

DQ Repository

The DQ repository is divided into two components: an ontological description of the assessed IS schema where every schema element can be uniquely identified, as well as a database that stores DQ assessment results over time. The standardized representation of different data models and their schema elements can, for example, be achieved with the *data source description* vocabulary approach (Ehrlinger and Wöß 2015) and information about data quality can be annotated to the respective information system elements using the W3C Data Quality Vocabulary (W3C 2016). The DQA results and DQMD are then stored in a database optimized for time series data, where each record is uniquely identified by a combination of schema element ID, DQ metric or metadata descriptor, and timestamp. We evaluate Apache Cassandra[2] as suitable data store for the repository, because its column based data model offers good support for processing time series data due to the automatic ordering by sequentially writing data to the disk. A similar project is the Metadata Store of the Metadata Management System[3] that aims at storing all kinds of metadata from relational datasets for analysis purposes. In contrast to the Metadata Store, our proposed solution additionally includes a time dimension that requires larger storage if the persisted data is not selected properly. Thus, only metadata that is beneficial for time series DQ metrics will be stored, instead of striving for a greatest possible coverage. Likewise, the content of a monitored information system is used to calculate DQ metrics but is not stored in the DQ Repository since this would lead to an extensive storage requirement, which is not feasible for such a system in practice. The aim of the introduced system is the overall analysis of the development of DQ in an information systemand not a detailed investigation of individual data values at any point in time. One major challenge for DQ storage are steadily changing information systems with therefore changing DQ assessment results. Such modifications on schema-level can be managed by version control.

Time Series Analysis

The stored time series data (comprising DQA results and DQMD) are characterized by a correlation of adjacent points in time, that is, a value x_t at time t depends on its prior values x_{t-1}, x_{t-2}, ... and can be analyzed by exploring plots or applying mathematical models (Shumway 2016). The primary goal of time series analysis in this application is the detection of trends and temporal outliers (i.e., anomalies) to identify changes in the qualitative condition of an information system.

There are several well-researched methods for outlier or anomaly detection in the literature on time-series analysis as well as in the literature on statistical process control, which we are going to evaluate with regard to their appropriateness for DQM. Promising approaches are cumulative sum (CUSUM) methods, exponentially weighted moving average (EWMA), and regression based techniques like autoregressive integrated moving average (ARIMA) models (Chanola et al. 2007; Jones-Farmer et al. 2014). The effectiveness of such methods with respect to its interpretation is closely coupled with a proper visualization in component (4).

An example for an anomaly in data quality monitoring would be the sudden decrease of an attributes completeness rating, which is caused by an increase of inserted NULL values, produced by a faulty measurement device. Causes for such anomalies can be analyzed by applying linear models and decision trees to detect correlations between different quality metrics and DQMD (Breiman 1984). Multivariate time series analysis will be investigated as additional method for correlation detection.

[2] http://cassandra.apache.org (August 2017)

[3] https://github.com/stratosphere/metadata-ms (August 2017)

Visualization

The visualization of DQ results has been researched by Kandel et al. (2012), who highlight the need to automate this step since the determination, which visualization should be generated, may overstrain standard users, especially with a large number of observed data sets and DQ results. Their proposed *data quality bars* can be extended by a time dimension to serve as starting basis for visualizing DQM results. We are currently evaluating Graphite[4], which offers an application programming interface with functionalities to retrieve metrics from a time-series database (e.g., a combination of Cyanite[5] with Apache Cassandra) and to manipulate this data, as well as a web-based dashboard for displaying graph data.

The purpose of component (4) in Figure 1 is twofold. On the one hand, the stored time series data from the DQ Repository can be charted directly, and on the other hand, the results from the time series analysis in component (3) are presented to the user. A promising method for outlier detection via visualization are statistical control charts from the field of statistical process control, which are often applied for quality control in manufacturing. An exemplary application of control charts to analyze the quality of aircraft maintenance data along with comprehensive literature research has been proposed by Jones-Farmer et al. (2014). However, since control charts are primarily suitable for studying one or two attributes there are several opportunities for future research with respect to multivariate control chart methods (Batini and Scannapieco 2016; Jones-Farmer et al. 2014).

DQM CAPABILITIES IN EXISTING APPROACHES

There is lack of scientific studies on data quality assessment tools, especially with respect to monitoring capabilities. The survey by Barateiro and Galhardas (2005) does neither cover state-of-the-art tools nor consider data quality monitoring at all. In order to investigate the current industrial and scientific situation concerning data quality tools, we are currently conducting a survey on DQ monitoring capabilities.

Gartner classifies 17 tools in the Magic Quadrant for Data Quality Tools (Judah et al. 2016) that offer core functionality for data quality assessment, including the two most relevant features for our research: monitoring and metadata management. We use the classification by Gartner is starting basis for our investigation, and extend it with tools that explicitly state that they offer DQM functionality (e.g., Datamartist, DataCleaner) and representative scientific tools.

Based on the architecture of our automated DQM application in Figure 1, we identified four core requirements that must be fulfilled to enable comprehensive data quality monitoring:

- *DQ assessment capabilities*: the tool allows the computation of data quality metrics for the most common DQ dimensions like completeness, accuracy, and consistency. We will specify this requirement in more detail after the determination of appropriate DQ dimensions and metrics for continuous monitoring.
- *Automation*: the calculation of data quality metrics can be scheduled in user-defined time periods.
- *Storage*: the application allows storage and retrieval of DQ assessment results over time.
- *Analysis*: the application supports appropriate visualization and analysis of current and previous DQ assessment results.

[4] http://graphite.readthedocs.io (August 2017)

[5] http://cyanite.io (August 2017)

The extent to which a tool fulfills each requirement is described textually and represented by a threefold rating, which can either be (i) fully supported, (ii) partially supported or there exists a concrete suggestion by the vendor to use a third-party tool, or (iii) not supported.

So far, we investigated the following tools: Talend Open Studio for Data Quality[6], Oracle Enterprise Data Quality[7], Informatica Data-as-a-Service[8] and Cloud Data Quality Radar[9] and Datamartist[10] by nModal Solutions Inc. According to the first insight of this review, current systems seldomly offer pre-defined storage capabilities where the collected DQA results are stored and can be retrieved later-on. Usually, data is verified on-the-fly against specially defined rules or expressions, and optionally alerts can be triggered if those rules are violated. In contrast to such a solution, our research work aims at monitoring DQ on a higher level, for example, the extent to which the completeness of an information system or the proportion of NULL values within a column develops over time. This also enables the comparison of current DQA results with previous results and allows deriving conclusions about the effect of taken DQ measures.

CONCLUSION AND OUTLOOK

The main contribution of our research work is the analysis of monitoring capabilities for data quality assessment results and the development of an automated DQM application that focuses on practical scenarios with typically different and heterogeneous information systems. After completing the investigation of existing DQ tools with respect to their monitoring capabilities, we proceed our research by collecting proposed data quality metrics. We exclude metrics that require a gold standard and instead concentrate on metrics that can be calculated repetitively. We develop new DQ metrics for quality dimensions that lack suitable metrics for continuous DQM. With the results of this investigation, we provide a basis for users of existing data quality tools to select appropriate metrics for scheduled data quality assessment.

In parallel, we implement our proof-of-concept application presented in Figure 1, including the selected DQ metrics, the DQ repository, and a user interface. The gathered metadata that is stored in the DQ repository provides the basis for the development and examination of algorithms for DQMD analysis that allow the comparison of historical with current DQA results. The DQM user interface visualizes results of data quality analysis with different levels of detail to support their interpretation and finally to derive decisions. The interface allows the conduction of a user study in order to illustrate the practical applicability and to verify the effectiveness of the developed algorithms and visualizations.

ACKNOWLEDGEMENTS

The research reported in this paper is supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Science, Research and Economy, and the Province of Upper Austria in the frame of the COMET center SCCH.

[6] https://www.talend.com/download/talend-open-studio (August 2017)

[7] http://www.oracle.com/technetwork/middleware/oedq/overview/index.html (August 2017)

[8] https://www.informatica.com/products/data-quality/data-as-a-service.html (August 2017)

[9] https://www.informatica.com/products/data-quality/cloud-data-quality-radar.html (August 2017)

[10] http://www.datamartist.com/data-quality-monitoring-and-reporting (August 2017)

REFERENCES

Apel, D., Behme, W., Eberlein, R., and Merighi, C. 2015. Datenqualität erfolgreich steuern: Praxislösungen für Business-Intelligence-Projekte, dpunkt.verlag, Heidelberg, Germany.

Barateiro, J., and Galhardas, H. 2005. "A Survey of Data Quality Tools," *Datenbank-Spektrum* (14:5), pp. 15–21.

Batini, C., and Scannapieco, M. 2016. *Data and Information Quality: Concepts, Methodologies and Techniques*, Springer International Publishing.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. 1984. *Classification and Regression Trees*, Chapman & Hall/CRC.

DAMA International 2008. "DAMA-DMBOK Functional Framework," *Technics Publications*.

Ehrlinger, L., and Wöß, W. 2015. "Semi-Automatically Generated Hybrid Ontologies for Information Integration," in *Joint Proceedings of the Posters and Demos Track of 11th International Conference on Semantic Systems – SEMANTiCS2015 and 1st Workshop on Data Science: Methods, Technology and Applications (DSci15)*, A. Filipowska, R. Verborgh, A. Polleres (eds.), Technical University of Aachen (RWTH), pp. 100–104.

English, L. P. 1999. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*, John Wiley & Sons, Inc., New York, USA.

Hipp, J., Müller, M., Hohendorff, J., Naumann, F. 2007. "Rule-Based Measurement of Data Quality in Nominal Data," in *Proceedings of the 12th International Conference on Information Quality*, MIT, Cambridge, pp. 364–378.

IEEE 1061-1998, 1998. "Standard for a Software Quality Metrics Methodology," Institute of Electrical and Electronics Engineers (available at https://standards.ieee.org/findstds/standard/1061-1998.html).

Jones-Farmer, L. A., Ezell, J. D., and Hazen, B. T. 2014. "Applying Control Chart Methods to Enhance Data Quality," *Technometrics* (56:1), pp. 29–41.

Judah, S., Selvage, M. Y., and Jain, A. 2016. "Magic Quadrant for Data Quality Tools," Gartner Research.

Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M., and Heer, J. 2012. "Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, ACM, pp. 547–554.

Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. 2002. "AIMQ: A Methodology for Information Quality Assessment," *Information & Management* (40:2), pp. 133–146.

Naumann, F. 2013. "Data Profiling Revisited," *ACM SIGMOD Record* (42:4), pp. 40–49.

Naumann, F., Freytag, J. C., Leser, U. 2004. "Completeness of Integrated Information Sources," *Journal of Information System* (29:7), pp. 583–615.

Naumann, F., and Rolker, C. 2000. "Assessment Methods for Information Quality Criteria," in *Proceedings of the 5th International Conference on Information Quality*, pp. 148–162.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. 2002. "Data Quality Assessment," *Communications of the ACM* (45:4), pp. 211–218.

Redman, T. C. 1996. *Data Quality for the Information Age*, Artech House, Inc., Norwood, MA.

Shumway, R. H., and Stoffer, D. S. 2016. *Time Series Analysis and Applications*, Free Dog Publishing.

Tuura, L., Meyer, A., Segoni, I., and Della Ricca, G. 2010. "CMS Data Quality Monitoring: Systems and Experiences," *Journal of Physics: Conference Series* (219:7), IOP Publishing.

Wand, Y., and Wang, R. Y. 1996. "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM* (39:11), pp. 86–95.

Wang, R. Y. 1998. "A Product Perspective on Total Data Quality Management," *Communications of the ACM* (41:2), pp. 58–65.

Wang, R. Y., and Strong, D. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management of Information Systems* (12:4), pp. 5–33.

W3C 2016. "Data on the Web Best Practices: Data Quality Vocabulary," A. Riccardo, I. Antoine (eds.), (available at https://www.w3.org/TR/vocab-dqv).

Zhu, H., Madnick, S. E., Lee, Y. W., and Wang, R. Y. 2014. "Data and Information Quality Research: Its Evolution and Future," *Computing Handbook: Information Systems and Information Technology*, H. Topi, A. Tucker (eds.), Chapman & Hall /CRC, pp. 16.1–16.20.