International Telecommunication Union

ITU-T Technical Specification

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(19 July 2019)

ITU-T Focus Group on Data Processing and Management to support IoT and Smart Cities & Communities

Technical Specification D4.4

Framework to support data quality management in IoT

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The procedures for establishment of focus groups are defined in Recommendation ITU-T A.7. ITU-T Study Group 20 set up the ITU-T Focus Group on Data Processing and Management to support IoT and Smart Cities & Communities (FG-DPM) at its meeting in March 2017. ITU-T Study Group 20 is the parent group of FG-DPM.

Deliverables of focus groups can take the form of technical reports, specifications, etc., and aim to provide material for consideration by the parent group in its standardization activities. Deliverables of focus groups are not ITU-T Recommendations.

Technical Specification D4.4

Framework to support data quality management in IoT

Summary

This Technical Specification specifies a framework to support data quality management in IoT for the provision of quality data and services. The relevant requirements and technologies that support the data quality management are defined in this Technical Specification.

Acknowledgements

This Technical Specification was researched and principally authored by Ilyoung Chong (Hankuk Univeristy of Foreign Studies), Muhammad Aslam Jarwar (Hankuk Univeristy of Foreign Studies), Nathalie Feingold (NPBA) and Xiaomi An (RUC) under the chairmanship of Gyu Myoung Lee (Korea, Rep.of).

Additional information and materials relating to this Technical Specification can be found at: www.itu.int/go/tfgdpm. If you would like to provide any additional information, please contact Denis Andreev at tsbfgdpm@itu.int.

Keywords

Data quality; Data quality classification; Data quality management; IoT

Technical Specification D4.4

Framework to support data quality management in IoT

Table of Contents

1. Scope 6
2. References 6
3. Definitions 6
3.1 Terms defined elsewhere 6
3.2 Terms defined in this Technical Specification 6
4. Abbreviations and acronyms 7
5. Conventions 7
6. Overview of data quality 7
6.1 Concept of data quality 7
6.2 Data quality classification 7
6.2.1 Intrinsic data quality 8
6.2.2 Contextual data quality 8
6.2.3 Representational data quality 8
6.2.4 Accessibility data quality 9
7. Data quality management in IoT 9
7.1 Overview of data quality management 9
7.2 Data quality management characteristics 11
8. Requirements to support data quality management in IoT 11
8.1 General requirements of data quality management 11
8.2 Requirements of data acquisition 12
8.3 Requirements of data quality assessment 12
8.4 Requirements of data quality evaluation 13
8.5 Requirements of data quality improvement. 13
8.6 Requirements of data quality ranking 14
8.7 Setup process to support data quality monitoring 14
9. Functional model to support data quality management 15
9.1 Data acquisition capability 15
9.2 Data quality assessment capability 15
9.3 Data quality evaluation capability 16
9.4 Data quality improvement capability 17
9.5 Data quality ranking capability 18
9.6 Setup process to support data quality monitoring 19

Appendix A: Intelligent data quality management using machine learning and deep learning 20

Appendix B: Definition of business goal for data quality management 22

Bibliography 23

Technical Specification D4.4

Framework to support data quality management in IoT

1. Scope

This technical Specification identifies the followings to provide data quality management in IoT. This Technical Specification addresses a framework for data quality management. The scope of this Technical Specification covers several key requirements with respect to data quality management in IoT and many important elements to fulfil these requirements. Specifically, it covers the following:

- Overview of data quality management;
- Data quality management in IoT;
- Requirements of data quality management in IoT;
- Functional model to support data quality management.

2. References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Technical Specification. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Technical Specification are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Technical Specification does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T Y.2060] Recommendation ITU-T Y.2060 (2013), *Overview of the Internet of things*

[ITU-T Y.3600] Recommendation (07/2016), *Big data standardization Roadmap*

[ITU-T Technical Paper] Technical Paper (11/2016), *Analysis of Digital Data Technologies Toward Future Data Eco-Society*

3. Definitions

This Technical Specification uses the following terms defined elsewhere:

3.1 Terms defined elsewhere

3.1.1 Internet of Things (IoT) [ITU-T Y.2060]: A global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies.

3.2 Terms defined in this Technical Specification

3.2.1 Critical data: The data to be required for the successful completion of a task in an application or service within a specific business context. It is characterized by defined attributes and priorities.

3.2.2 Data quality: The degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions.

3.2.3 Non-critical data: Low business value and impact data.

3.2.4 Poor Data: The data classified with low quality during the data quality assessment process.

4. Abbreviations and acronyms

This technical Specification uses the following abbreviations and acronyms:

AI Artificial Intelligence

DL Deep Learning

DQM Data Quality Management

IoT Internet of Things

ML Machine Learning

5. Conventions

None

6. Overview of data quality

6.1 Concept of data quality

The data quality will be attributed to a quality supporting business processes, analysis techniques, and it will be raised as to whether a data quality of the existing data would be worthwhile or plausible. Thus, data quality is necessarily architected, measured and assessed with consideration, and the possible risks in deploying data business will be overcome well.

For data that is collected through sensing to be fit for use, it should possess certain features so that it can satisfy a set of system requirements.

There is increasing awareness of the criticality of data to making informed decisions and how inaccurate data can lead to disastrous consequences. The challenge lies in ensuring that enterprises collect relevant data for their business, manage or govern that data in a meaningful and sustainable way, ensure high quality for critical data, and analyse the high quality data to accomplish stated business objectives.

6.2 Data quality classification

Based on the concept of data quality, the data quality classification have been identified as follows (Figure):

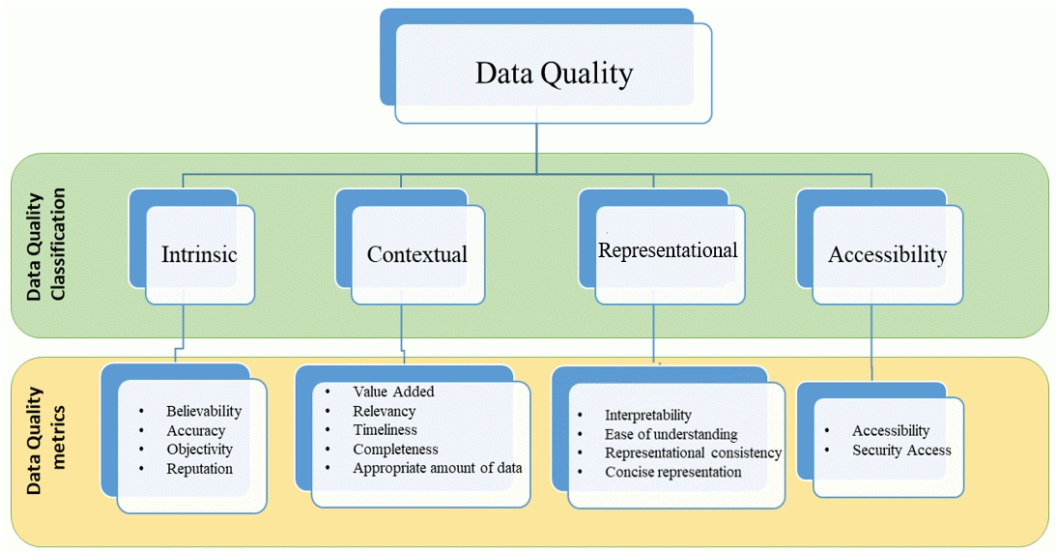

Figure 6-1. Classification of data quality

6.2.1 Intrinsic data quality

The intrinsic category of data quality that is related to the actual values of data regardless of the context or data components (e.g., accuracy).

Following are the data quality metrics of this category:

- Believability: The believability is the extent to which data are accepted or regarded as true, real, and credible;
- Accuracy: The degree to which data correctly describes the "real world" object or event being described;
- Objectivity: The extent to which data is unbiased, unprejudiced, and impartial;
- Reputation: The extent to which data is highly regarded in terms of its source or content.

6.2.2 Contextual data quality

The contextual category of data quality that is related with respect to other data components and within a certain context (e.g., completeness, timeliness, and consistency).

Following are the metrics of contextual data quality:

- Value added: The extent to which data is beneficial and provides advantages from its use;
- Relevancy: The extent to which data is applicable and helpful for the task at hand;
- Timeliness: The degree to which data represent reality from the required point in time;
- Completeness: The proportion of stored data against the potential of "100% complete";
- Appropriate amount of data: This means that the sufficient amount of data is available for use to compute a result of data items.

6.2.3 Representational data quality

The representational category captures aspects related to the design of the data.

Following are the metrics of representational data quality:

- Interpretability: The extent to which data is in appropriate languages, symbols, units, and the definitions are clear;

- Ease of understanding: The ease of understanding means the extent to which data is easily comprehended;
- Representational consistency: The extent to which a whole dataset is presented in the data structure and format;
- Concise representation: The extent to which data is compactly represented.

6.2.4 Accessibility data quality

This category deals with the data quality aspects related to data infrastructure such as accessibility, security access, data retrieval.

Following are the metrics of representational data quality;

- Accessibility: The extent to which data is available, or easily and quickly retrievable;
- Security access: The security access metrics measures that the data is safeguarded from unauthorized access and preventing data loss.

7. Data quality management in IoT

7.1 Overview of data quality management

Due to the advanced information and communication technologies, a very huge amount of data is generated in the operation, production, and management of the IoT business . However, the quality and reliability of that data are questionable. This newly collected data from out sources and already available data in the organization should be pre-processed in order to check the data completeness, accuracy, and consistency for the improved and high quality decision making services. Data quality management is an administration type that incorporates the role establishment, role deployment, policies, responsibilities and processes with regard to the acquisition, maintenance, disposition and distribution of data. Data quality management initiative promotes a strong partnership between technology groups and the business. Building and controlling the entire data environment, that is, architecture, systems, technical establishments, and databases will be promoted through the overall environments to acquire, maintain, disseminate and dispose of an organization's electronic data assets. Further following are the key motivations for the data quality management in the IoT .

- Data quality management is the set of process and activities of combing all the entities in the organization including people, technology and culture to promote the common goal of data quality provision to the IoT data consumers;
- Data quality management should be an open system where data quality professionals need to interact with data consumers and other stakeholders of data freely in order to identify the information need and data quality requirements;
- Data quality management is also crucial when it comes to choosing one dataset for business IoT applications over the other available datasets;
- The data quality will be attributed to a quality supporting business processes, analysis techniques, and it will be raised as to whether a data quality of the existing data would be worthwhile or plausible;
- Information in data application can be derived from the processing of data to give meaning and sense to it with multiple types of metadata;

- To achieve data quality for IoT applications, it should be necessary to collect and pre-process the data, perform the quality assessment of existing data, evaluate the results and improve it, if necessary, perform the ranking and monitoring of data for the sustainable data quality management. The overview of various data quality management capabilities have been illustrated in Figure and the detail of these components will be presented in chapter 9 functional model of data quality management.

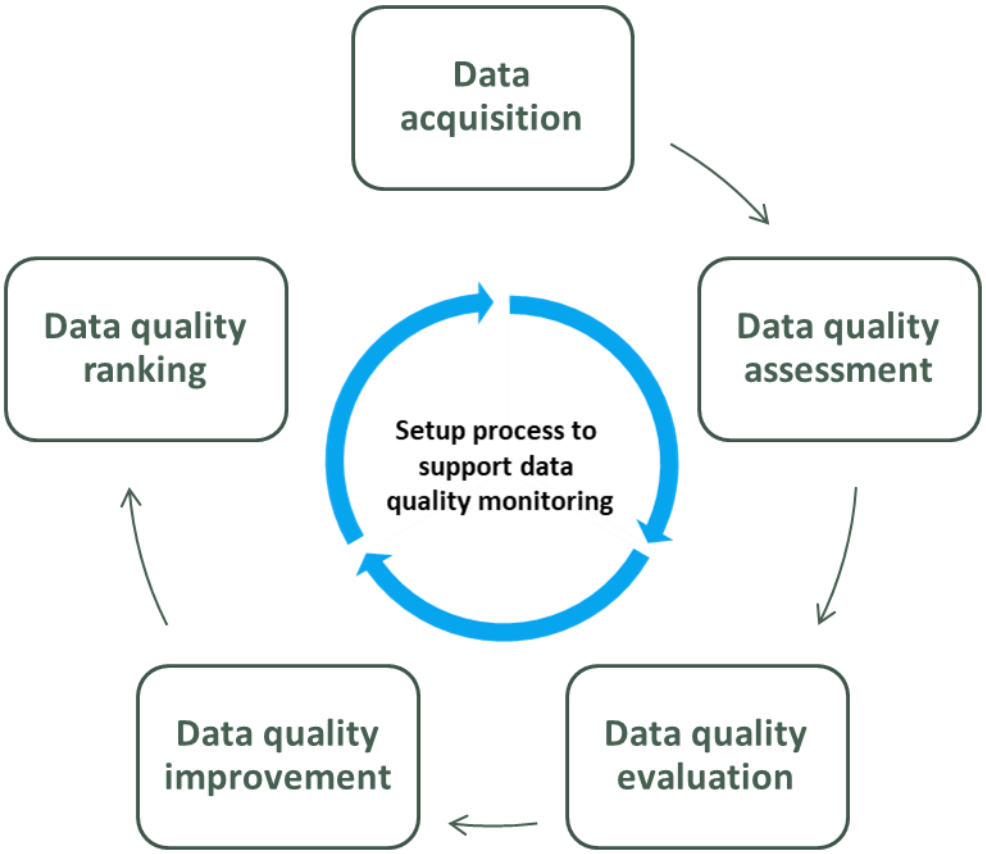

Figure 7-1. Overview of data quality management

- Data acquisition: The data acquisition supports internal and external data with consistency, storage efficiency, retrieval efficiency and security efficiency.
- Data quality assessment: To assess the existing data quality, common data quality metrics are applied;
- data quality evaluation: The evaluation of data quality management procedures is mandatory because it also focuses on the economic advantages to choose data quality improvement solutions. To ensure the suitable level of data quality evaluation, the following functionalities should be provided by the data quality management framework;
- Data quality improvement: To improve the data quality a design data quality plan is applied for the data quality improvement;
- Data quality ranking: Rank the data as per achieved data quality level, the ranking of data for further use in the business services;

– Support process to monitor data quality: To support the ongoing data quality monitoring and management.

7.2 Data quality management characteristics

Based on the data quality and data quality management overview, the data quality management framework should have the following characteristics.

- The quality data should be managed in a way that it should be accurate, available, consistent, confident, integrated, relevant, reusable, current, and complete;
- The quality of data should be measured with relevant data quality metrics;
- The data quality management framework contains data provenance information;
- The management system contains methodologies to maintain the metadata;
- Data quality management has the capability to manage reference data;
- The system provides support to improve the data quality if the received data has poor quality;
- A central data modeling system is being provided in the data quality management framework;
- The ranking of the data with respect to its data quality indicators is supported.

8. Requirements to support data quality management in IoT

To support the business goal of data quality management in the IoT, many aspects are required to be considered. Following sections highlight some general requirements to support data business goals of quality management and some specific requirements with respect to each business process in the data quality management framework.

8.1 General requirements of data quality management

To receive the appropriate level of data quality following general aspects are required:

– The system should be scalable in order to handle growing large volume of IoT data;

– The data quality management should be available during the processing of data;

– Define the set of validations that need to be developed to measure the quality of data;

– The data quality management process should be interoperable in order to handle data received from heterogeneous sources;

– For the ongoing improvement in the data quality, the system should be able to monitor data quality with some length of period;

– privacy of users must be considered during the processing of data in order to estimate and improve the data quality;

– Privacy protection during the data collection process;

– In order to increase the data quality completeness aspects, handling of missing data with sophisticated techniques should be provided;

– To increase the accuracy and confidence in the IoT data, the data quality management system should be able to detect and correct the false and suspect data;

– Ability to produce ad hoc reports indicating gaps in data for any attributes selected

– Ability to identify the number of data points that are not populated for each required attribute field;

– It is required that the data collection scheduler should balance the data collection interval in order to increase efficiency (for example: the battery life of IoT sensing objects.);

– The data quality management framework should support time alignment in the collection of data in the case multi-sources data collection request for the IoT service provision;

– To estimate and improve the quality of data efficiently, the data collection methodologies in the data quality framework should support the collection of metadata along with the data;

– The data quality management framework should support the coordinated workflow mechanism in order to handle the collection of duplicated data.

8.2 Requirements of data acquisition

To acquire the IoT data for business needs, the data quality management framework should support the following functionalities:

– Ensure that no extra data should be collected with respect to the required policy. The collection of extra data reduces customers privacy and introduces the data leakage issues of data quality;

– Supports of acquiring and validation of data from external sources;

– Supports of acquiring and validation of data from internal sources;

– Support data masking during the data acquiring phase in order to handle privacy aspects.

8.3 Requirements of data quality assessment

To measure the quality of data with optimal precision, data quality management framework should support the following functionalities:

– Due to the large volume of IoT data, it is difficult to estimate the quality of data within a reasonable amount of time, therefore the data quality management framework should be efficient and scalable to support efficient data quality assessment;

– As there is no unified standard to estimate the quality of data for various IoT applications, therefore the framework should support to measure the quality of data with many data quality metrics;

– The framework should support to estimate data quality with common metrics initially;

– The data quality management framework should support a basic set of data quality measurement methodologies;

– There should be a support to add new data quality indicator and measurement methodology in the framework;

– The measurement of data quality should be considered according to various interest groups;

– There should be the provision of update in the measurement methodology of existing data quality indicators;

- Data quality assessment functions should support individual quality measurement for each data source;
- Data quality assessment functions should support aggregated measurement for data received from multiple sources;
- Periodic data quality assessment capability should be supported for non-critical data;
- Continuous data quality assessment functions should be supported for critical data.

8.4 Requirements of data quality evaluation

The evaluation of data quality management procedures is mandatory because it also focuses on the economic advantages to choose data quality improvement solutions. To ensure the suitable level of data quality evaluation, the following functionalities should be provided by the data quality management framework:

- The system should be able to locate critical areas of data which affects the quality of data;
- The system should be able to optimize the time used in the data quality assessment;
- The system should also check the data reputation after its usage in the IoT applications so that other services can get the benefit;
- The system should be able to evaluate the direct and indirect cost of the data quality process;
- The system should check that the results of data quality assessment are up to standard for the IoT applications and services;
- The system should recommend improving the grey areas of data which weaken the level of data quality;
- The system should support to evaluate the business rules from time to time defined against the various data sources.

8.5 Requirements of data quality improvement.

To ensure to choose suitable methodologies to improve data quality, the following functionalities should be provided by the data quality management framework:

- The system should be able to perform data various types of data interpolation in order to handle missing data;
- The system should support to detect and correct data outliers in the streaming data;
- The system should be able to rectify and improve the data effected by malwares attacks.
- It is required that the framework should support data deduplication;
- Data should be transformed in the encrypted format in order to increase security and confidence of data;
- The system should support the common data representation format in order to increase the data interpretability;
- The data quality management model should have high power infrastructure in order to support data availability in the peak hours and emergency situation;

– The system should have a predefine data threshold in order to validate data accuracy;

– The data quality management framework should support a proactive approach to improve the data quality if possible.

8.6 Requirements of data quality ranking

To ensure the suitable level of data quality ranking, the following functionalities should be provided by the data quality management framework:

– In consideration of fitness for use for the task at hand, the system should be able to assign weights to the preferred data quality metrics for measuring the ranking of data;

– In the IoT services sometimes the data of individual source is used and another time data from multiple sources are used collectively. Therefore, the data quality management framework should have the provision of individual aggregative data quality ranking;

– For ongoing improvement and provision of services, the quality of data should be ranked in different time slots;

– The data quality management framework should support estimated or predicted data quality ranking if the actual data quality ranking is under processing, in order to support to predict IoT service behavior modeling;

– The system should provide the ranking of the quality of data in terms of its business usage and technical ranking;

– For the sustainable data quality management, the system should support to monitor and improve the data at the source level.

8.7 Setup process to support data quality monitoring

To support the ongoing data quality monitoring, the following functionalities should be provided by the data quality management framework:

– The system should support periodic monitoring, that provides feedback on the data quality management process and it enables dynamic tuning;

– The system should be able to generate alerts when the level of data quality decreases to a certain threshold;

– The data quality management framework should have the capability of periodical reporting of quality;

– For the sustainable data quality management, the system should support to monitor the quality of data at the source level;

– The data quality management framework should support multiple schedulers to monitor ongoing data quality;

– The framework should support appropriate visualization data quality monitoring results;

– The framework should support tracking levels of data quality over time for monitoring the ongoing process and improvement.

9. Functional model to support data quality management

The functional model of data quality management provides mechanisms to support the management of IoT data. The functional capability of this model is distributed as follow.

9.1 Data acquisition capability

Data acquisition is the processes for bringing data that has been created by a source outside the organization, into the organization, for business use. The data acquisition supports internal and external data with consistency, storage efficiency, retrieval efficiency and security efficiency. Figure 9-1 shows the function model of data acquisition capability with respect to data quality in this capability.

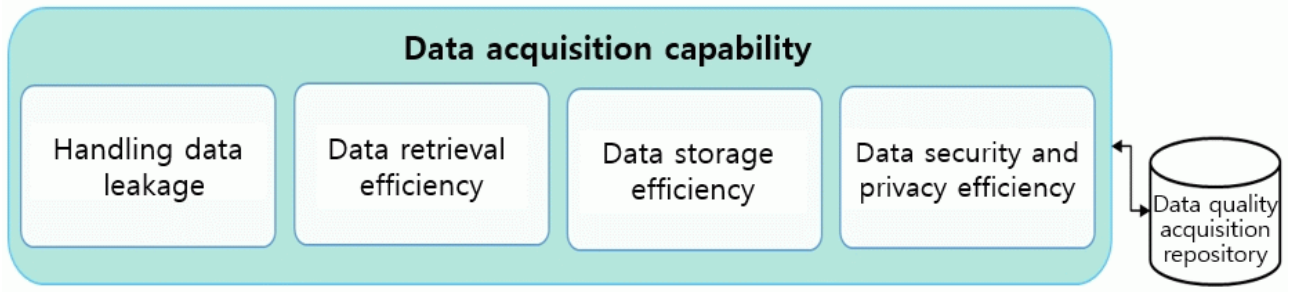

Figure 9-1. Functional model of Data acquisition capability

– Handling data leakage function: Data leakage issue occurs when the IoT service or application acquires more data than the necessary requirements. This function handles these issues when data acquisition has been performed in the data quality management platform;

– Data retrieval efficiency function: This function provides the capability to acquire data from the repository within an appropriate time. Due to the application of this function, the data availability and timeliness aspects of data could be improved;

– Data storage efficiency function: This function supports the mechanism to store the acquired data within an appropriate time and consistent format.;

– Data security and privacy function.

9.2 Data quality assessment capability

The capability offers the methodologies to estimate the data quality of the received and existing IoT data. The quality of data is measured with various data quality metrics. Figure 9-2 shows the functions of data quality assessment capability.

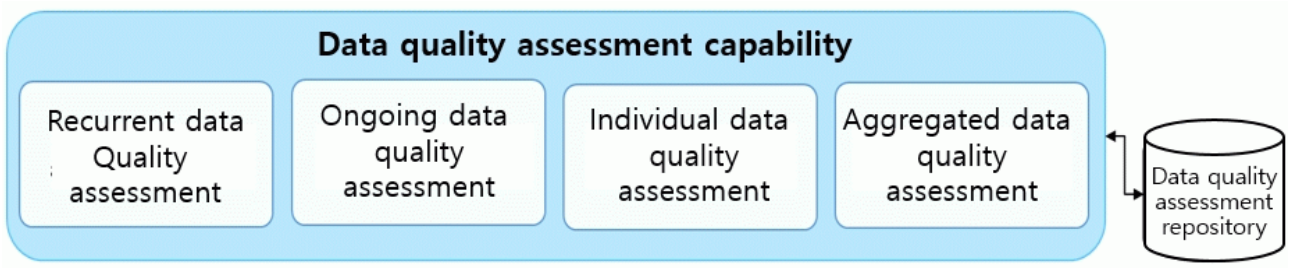

Figure 9-2. Function model of data quality assessment capability

– Recurrent data quality assessment function: The recurrent data quality assessment function supports the measurement of IoT data quality after each specific interval. This function checks the quality of data with respect to various aspects as defined in the function template;

– Ongoing data quality assessment: This function checks the quality of data received from IoT objects periodly. The main capability of this function is to ensure the validation of the quality of data received from critical IoT data sources. This function also supports priority assessment among the critical data categories;

– Individual data quality assessment: This function provides the mechanism to check and measure the quality of data of each source separately. Another capability of this function is that the individual data quality assessment function can be applied standalone or it can be used collectively with recurrent and ongoing data quality assessment function;

– Aggregated data quality assessment: In order to use data from various IoT data sources in a single service, then the IoT service required the aggregated data quality level for all of the data used in the service. The aggregated data quality checking function provides this capability. In aggregated data quality checking, good or poor quality affects the overall checking results.

9.3 Data quality evaluation capability

The details of data quality evaluation capability in accordance with data quality classification and characteristics to support data quality management is presented here. Figure 9-3 shows the functional model of data quality evaluation capability.

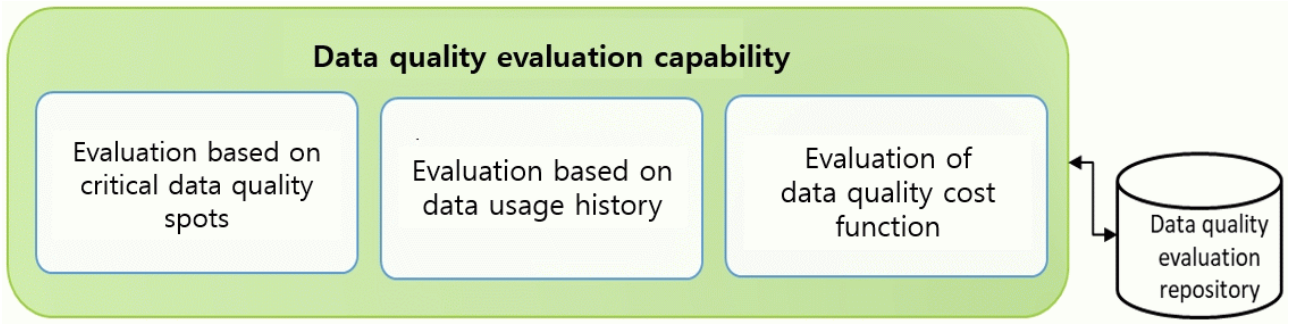

Figure 9-3. Functional model of data quality evaluation capability

– Evaluation based on critical data quality spots: The critical data quality spots evaluation function evaluates areas which hamper the overall quality of a dataset and analyses the root causes in order to avoid issues in future data;

– Evaluation based on data usage history: This function enables the capability of data quality management framework to evaluate the quality of data based on its usage in the IoT services. The feedback from the users could be the input in order to make the evaluation of the quality of the data;

– Evaluation of data quality cost: The data quality cost evaluation function supports the evaluation of the process of data quality. The main focus of this function is to measure the people and computational time, and system memory by considering data quality aspects, because the data quality assessment process requires high computation power and large memory due to the large volume of IoT data.

9.4 Data quality improvement capability

The data quality improvement capability to enhance the data spots whose data quality has been estimated and evaluated as poor. Figure 9-4 shows the functional model of data quality improvement capability.

– Data quality constraints validation: The data quality constraints validation function enables to detect missing interrelations in the IoT data. For the identification of missing interrelations in the data, this function uses the reference of dependency and integrity constraints. Due to the functional capability of this function, the data quality aspects towards data consistency and data representation could be improved in the data quality management framework;

– Data outliers: This outliers function includes identifying those values from a dataset or a single data source which are not coming in a certain range. As the data outliers is a major data quality issue in the IoT environment, by analyzing and fixing data outliers, the quality of data in terms of data accuracy, and data consistence could be enhanced;

– Data interpolation: The interpolation function is an estimation of a value within two known values in a sequence of values. The data interpolation functions enable the various mechanism to improve the quality of data in terms of data completeness and accuracy. This function applies various methodologies to estimate the missing data in the IoT data streams;

– Data deduplication: The deduplication is the process of data cleaning. The data deduplication function provides the mechanism to reduce a large volume of duplicate data by detecting the same copy or instance of data for similar real-world events. The data pointers are supported by this function which is referring to the unique copy of data. This function enhances the data availability aspects of data quality management and reduces the size of data storage and data management efforts;

– Data representation: The data representation function enables to improve the quality of data in terms of data interoperability. This function provides the mechanism to translate and transform the data in the common representation and storage format. The advantages of this function are the improvement of data consistency, data availability, and data accuracy.

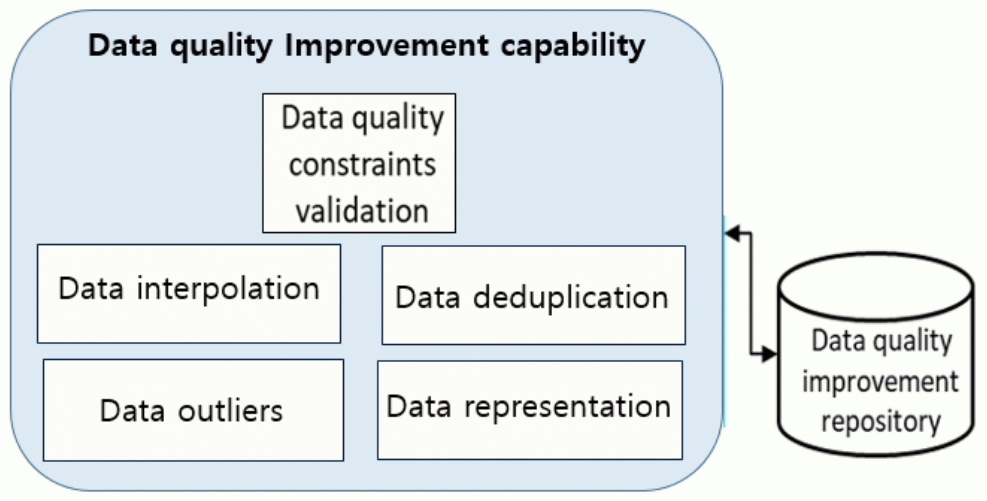

Figure 9-4. Functional model of data quality improvement capability

9.5 Data quality ranking capability

To ensure the suitable level of data quality ranking, three functions of data quality ranking capability are needed to enhance the business value of the data in the business applications. Figure 9-5 shows the functional model of data quality ranking capability.

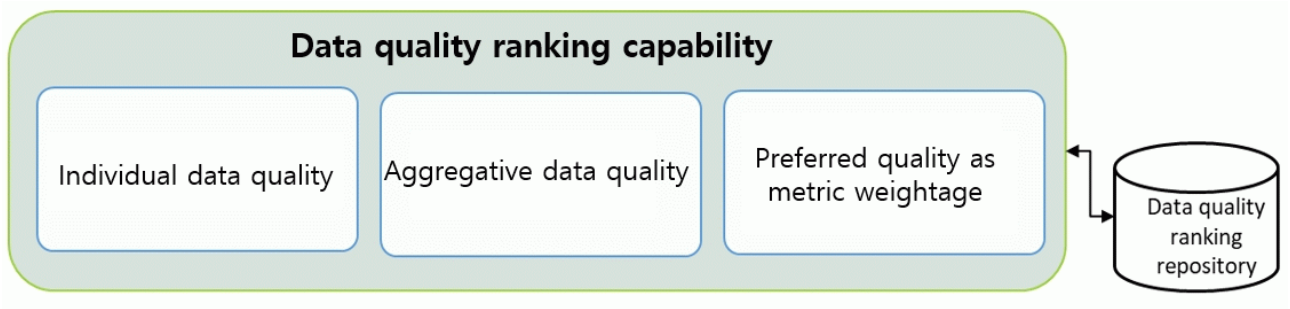

Figure 9-5. Functional model of data quality ranking capability

– Individual data quality ranking: The individual data quality ranking function enables to rank the data separately as per the data quality checking results with respect to each data source and data quality aspects such as data accuracy, data completeness, data consistency, and data availability. The applications of this function support to use the data individually with respect to the IoT object real-world sensing capability;

– Aggregative data quality ranking: The functional capability of the aggregative ranking function supports the ranking of data quality collectively of all the data received from multiple sources. Further, this function enables business-oriented data quality ranking such as location, time, and also data quality ranking'

– Preferred ranking as metric weightage: To create and update application specific data quality ranking templates, this function enables to assign data quality weightage to preferred data quality metrics, so that the data could be categorized and used in the specific context.

9.6 Setup process to support data quality monitoring

To support the ongoing data quality monitoring for sustainable data quality, this function group provides many functions to monitor data quality at each stage in the data quality management framework. Figure 9-6 shows the data quality monitoring functions in this group.

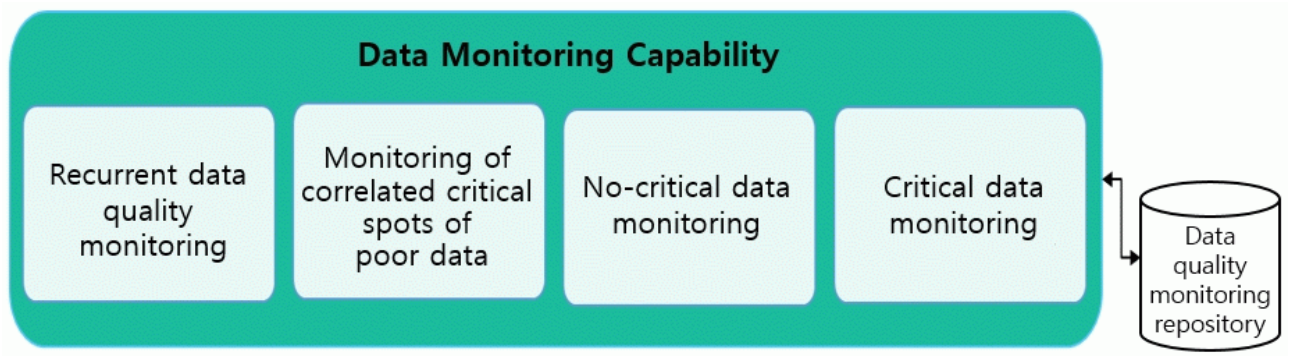

Figure 9-6. Data quality monitoring capability

– Recurrent data quality: The recurrent data quality monitoring function ensures the monitoring of various aspects of data quality periodically. This function supports various components which create many schedules for the checking of the entire process of data quality management systems. Choice of data attributes to monitor, and how to monitor them, is one of the key decisions of the design phase of the data quality management;

– Monitoring correlated critical spots of poor data: The monitoring correlated critical spots of poor data enables to monitor the quality of data and related data whose quality estimated as weak in the data quality evaluation and ranking function. This function generates reports which show the trends of correlated critical data quality improvement or deterioration from time to time.

– Critical data monitoring: The critical data monitoring function supports an important issue in the data quality management. The critical data are more important than the rest of the data in the data quality management framework such as IoT data to monitor patient, surveillance data, etc. The monitoring mechanism in this function focuses on critical data defined as per IoT applications.

– Non-critical data monitoring: This type of data has also business value and impact but not as serious as critical data. Similarly, as monitoring of critical data quality function, the mechanism in this function focuses on non-critical data monitoring. Difference between critical and non-critical data could be defined in the data quality monitoring templates stored into the data quality monitoring templates repository.

Appendix A: Intelligent data quality management using machine learning and deep learning

A.1 Machine Learning based data quality management

The identification and correction of measurement errors often involves labor intensive case-by-case evaluations by statisticians. Machine learning will increase the efficiency and effectiveness of these evaluations. It proceeds in two steps: in the first step, a supervised learning algorithm exploits data on decisions to flag data points as erroneous to approximate the results of the human decision making process, and in the second step, the algorithm applies the first-step knowledge to predict the probability of measurement errors for newly reported data points.

Further improving and maintaining high data quality is a central goal of official statistics. In the field of data quality management (DQM), the collection of data on human decisions in the DQM process creates an opportunity to increase the efficiency and effectiveness of DQM with machine learning (ML). It is necessary to predict measurement errors on the basis of data on human decisions to flag data points as erroneous. These predicted probabilities of measurement errors facilitate the work of statisticians and form the basis for a ML based approach to automate their checks.

The main focus of an application of ML to DQM is in both applications;

- prediction of measurement errors;
- help to overcome data gaps.

To support DQM in both applications, the ML algorithms predict if a human decision maker would flag data points. In the application to data gaps, the algorithms predict missing values.

The ML yields accurate out-of-sample predictions and increases the efficiency of DQM. The potential of ML for official statistics is not limited to the prediction of measurement errors. Another important problem that ML can help to overcome is missing data. Out-of-sample predictions of missing values with ML algorithms can help to close data gaps in a wide range of datasets.

The use of ML based matching algorithms enables the service platform to ingest data for standardization at scale. Typical use cases include matching specific records or data sets to a common standard and transforming data to this standard, allowing for the creation of relationships and accurate links between base data and derived data. This workflow is particularly important in ML and fraud detection scenarios, where a high volume of customer's due diligence and transaction data from disparate systems necessitates extensive standardization to set flags and generate meaningful derived data.

Standardization simplifies deduplication issues and accuracy-related data quality problems. The flexibility of ML provides that changes to these metrics, it can be applied across the entire data set in a cost-effective way, reducing the overhead of moving to a new standard.

A.2 Deep Learning based DQM

Deep Learning (DL) in Artificial Intelligence (AI) might help us discover where the master data is kept. DL might be able to "spot" where the most frequently referenced data reside.

In fact, there are two other tasks in the DL applied to DQM that are much different and we don't need, and cannot use, DL.

The first task is the identification of the master data and the second concerns the enforcement of the policies that sustain it. The former steps should take no more than an hour with the right business people in the room; simply ask the business users such things as:

- What is the most important data (at a conceptual level) that is needed to make business process A work as planned?
- How much of this data is needed also to make business process B work as planned?
- How much less data can you use to make business process C work as planned?

The second task is at the other extreme; the enforcement of policy. It is the work of policy

enforcement that sustains the level of data quality and the effectiveness of the workflows executed to meet that data quality and business process.

Appendix B: Definition of business goal for data quality management

There is a strong correlation between the business processes and data quality. Thus, the design of business goal for data quality management should take place before any initiative associated with data-driven decision making and should continue during all the business process. Figure B-1 shows the process to perform the design of business goal in the data quality management.

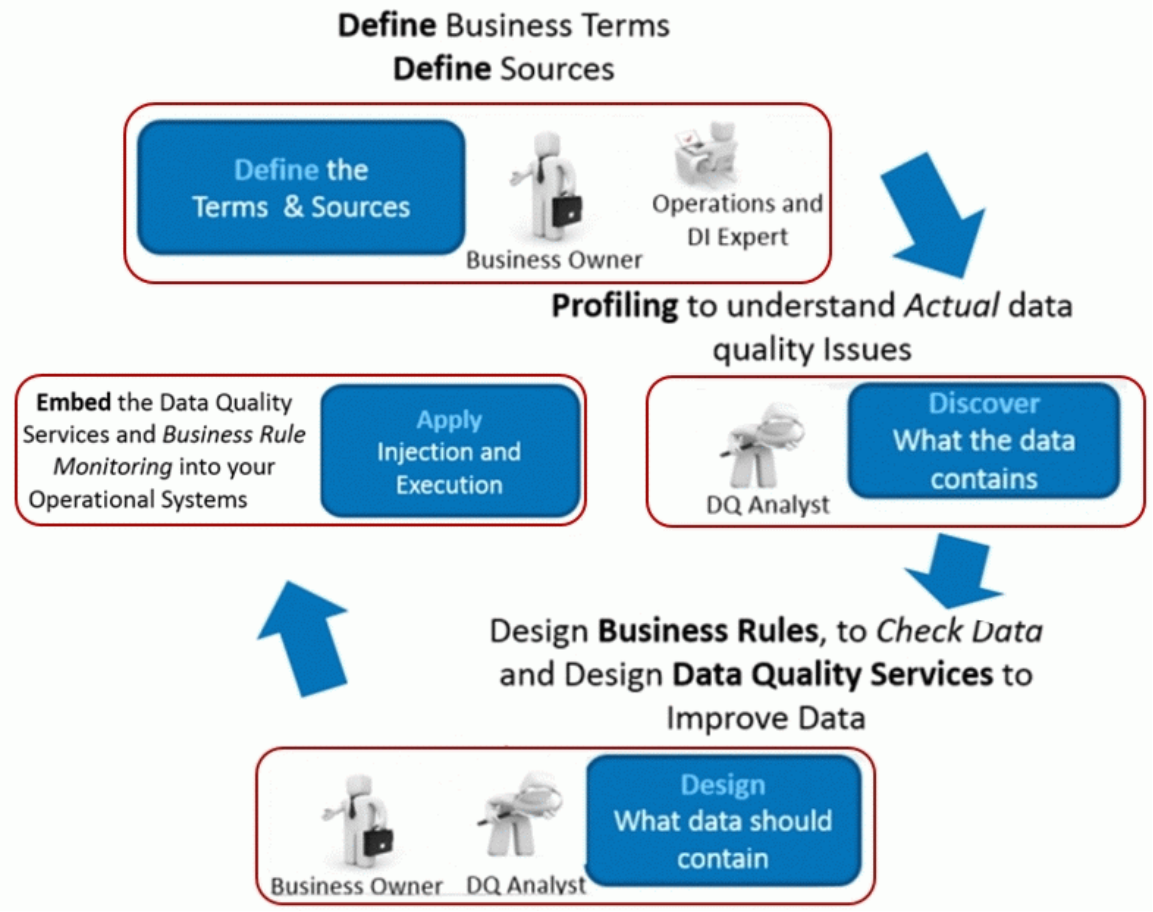

Figure B-1. Definition process of business goal for data quality management

Bibliography

[b-1] H. S. P. Panahy , F. Sidi, S. L. Affendey, A. M. Jabar, H. Ibrahim and A. Mustapha, "A Framework to Construct Data Quality Dimensions Relationships," *Indian Journal of Science and Technology,* vol. 6, no. 5, May 2013.

[b-2] A. Immonen, P. Pääkkönen and E. Ovaska, "Evaluating the quality of social media data in big data architecture," *IEEE Access,* vol. 3, pp. 2028-2043, 2015.

[b-3] C. Batini, . D. Barone , . M. Mastrella , . A. Maurino and C. Ruffini, "A FRAMEWORK AND A METHODOLOGY FOR DATA QUALITY ASSESSMENT AND MONITORING," in *12th International Conference on Information Quality*, Cambridge, MA, 2007.

[b-4] W. Y. Lee, D. M. Strong, B. K. Kahn and R. Wang, "AIMQ: A Methodology for Information Quality Assessment," *Information & Management,* vol. 40, p. 133–146, 2002.

[b-5] B. Heinrich, M. Kaiser and M. Klier, "How to measure data quality? A metric-based approach," in *Twenty Eighth International Conference on Information Systems*, Montreal, 2007.

[b-6] H. Huang, B. Stvilia, C. Jörgensen and W. H. Bass, "Prioritization of data quality dimensions and skills requirements in genome annotation work," *Journal of the American Society for Information Science and Technology,* vol. 63, no. 1, pp. 195-207, January 2012.

[b-7] C. Batini, M. Palmonari and G. Viscusi, "Opening the Closed World: A Survey of Information Quality Research in the Wild," in *The Philosophy of Information Quality*, vol. 358, L. FLORIDI, Ed., Oxford, Springer International Publishing, 2014, pp. 43-73.

[b-8] N. Laranjeiro, S. N. Soydemir and J. Bernardino, "A Survey on Data Quality: Classifying Poor Data," in *21st Pacific Rim International Symposium on Dependable Computing*, Zhangjiajie, 2015.

[b-9] C. Batini, C. Cappiello, C. Francalanci and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM computing surveys (CSUR),* vol. 41, no. 3, p. 16, 2009.

[b-10] L. Liu and L. N. Chi, "Evolution Data Quality: A theory-specific view," in *Seventh International Conference on Information Quality (ICIQ-02)*, Boston, 2002.

[b-11] F. Sidi, H. S. P. Panahy, S. L. Affendey, A. M. Jabar, H. Ibrahim and A. Mustapha, "Data Quality:A Survey of Data Quality Dimensions," in *International Conference on Information Retrieval & Knowledge Management*, Kuala Lumpur, 2012.

[b-13] W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim and D. Lee, "A Taxonomy of Dirty Data," *Data Mining and Knowledge Discovery,* vol. 7, no. 1, p. 81–99, January 2003.

[b-14] P. Oliveira, F. Rodrigues and P. R. Henriques, "A formal definition of data quality problems," in *Proceedings of International Conference on Information Quality (MIT IQ Conference)*, Cambridge, MA, 2005.

[b-15] N. Askham, D. Cook, M. Doyle, H. Fereday, M. Gibson, U. Landbeck, R. .. Lee, C. Maynard, G. Palmer and J. Schwarzenbach, "The six primary dimensions for data quality assessment," DAMA UK Working Group, United Kingdom, 2013.

[b-16] N. Laranjeiro, S. Soydemir and J. Bernardino, "A survey on data quality: classifying poor data," in *IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)*, 2015 .

[b-17] D. Loshin, The Practitioner's Guide to Data Quality Improvement, vol. A volume in MK Series on Business Intelligence, J. Niles, Ed., Boston: Elsevier, 2011, p. 129–146.

[b-18] T. Gschwandtner, J. Gärtner, W. Aigner and S. Miksch, "A Taxonomy of Dirty Time-Oriented Data," in *5 International Cross-Domain Conference and Workshop on Availability, Reliability, and Security*, Prague, 2012.

[b-19] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, p. 5–33, 1 March 1996.

[b-20] L. L. Y. W. L. a. R. Y. W. ipino, "Data quality assessment," *Communications of the ACM*, 2002.

_____