Building a Successful Data Quality Management Program

WHITE PAPER
# Table of Contents

- **Introduction**
- **DQM within Enterprise Information Management**
  - **What is DQM?**
- **The Data Quality Cycle**
  - **Measurements of Data Quality**
  - **Components of DQM**
  - **Data Discovery**
  - **Data Profiling**
  - **Data Quality Rules**
  - **Data Quality Monitoring**
  - **Data Quality Reporting**
  - **Data Remediation**
- **Key Tenets for Establishing a DQM Program**
- **About Knowledge**
Introduction

Managing the quality of data within an enterprise has become increasingly important due to a number of key business drivers:

- Competitors with better quality information and analytics
- Customers and clients who demand better-informed product and service providers
- Legal and compliance requirements that continue to grow
- Strategic business objectives to increase the ROI of data assets

This paper describes a foundational approach for building a successful Data Quality Management (DQM) Program to address these key business challenges and opportunities. It explains how DQM fits within a larger Enterprise Information Management framework. It also covers all aspects of the DQM cycle, including quantitative and qualitative measures of Data Quality, and includes key tenets for getting your DQM Program started.
DQM within Enterprise Information Management

DQM is one pillar within the broader Enterprise Information Management (EIM) framework. It is typically viewed as being an important Data Governance function owned by the business with roles and responsibilities shared with IT.

What is DQM?

DQM is the management of people, processes, technology, and data within an enterprise, with the objective of improving the measures of Data Quality most important to the organization. The ultimate goal of DQM is not to improve Data Quality for the sake of having high-quality data, but to achieve the desired business outcomes that rely upon high-quality data.

For example, to optimize customer relationship management and increase sales, high-quality data about your current and potential customers is critical to understanding them, communicating with them, and selling to them effectively. Customer data that must be of high quality to achieve desired business outcomes includes:

- Name (Prefix, First, Middle, Last, Suffix)
- Title
- Role
- Company/Organization Name
- Address(es) (Street Address Lines 1-4, City/Locality, State/Province, Country, Postal Code)
- Phone Number(s)
- Email Address(es)
- Contact Preferences

As one financial industry executive recently stated, “How can our customers trust us with their money if we can’t even get their name right?”

In order to improve Data Quality and to achieve the desired business outcomes, the Data Quality Cycle must be understood and continuously managed.
The Data Quality Cycle

Before you can fully understand every component of the Data Quality Cycle, you must first understand the Measure component and how Data Quality is measured.

Measurements of Data Quality

Every organization is unique, but there are a number of quantitative Data Quality measures that are universal:

- **Completeness**: The degree to which all required occurrences of data are populated
- **Uniqueness**: The extent to which all distinct values of a data element appear only once
- **Validity**: The measure of how a data value conforms to its domain value set (i.e., a set of allowable values or range of values)
- **Accuracy**: The degree of conformity of a data element or a data set to an authoritative source that is deemed to be correct or the degree the data correctly represents the truth about a real-world object
Building a Successful DQM Program

- **Integrity**: The degree of conformity to defined data relationship rules (e.g., primary/foreign key referential integrity)
- **Timeliness**: The degree to which data is available when it is required
- **Consistency**: The degree to which a unique piece of data holds the same value across multiple data sets
- **Representation**: The characteristic of Data Quality that addresses the format, pattern, legibility, and usefulness of data for its intended use

In addition to quantitative Data Quality measures, qualitative measures should also be considered. Some examples include:

- **Business Satisfaction Measures**: The increase/decrease in business satisfaction based on surveys
- **Collaboration/Improved Productivity Measures**: Percent of times the Data Governance Council detected and eliminated redundant intra- or inter-departmental projects/initiatives
- **Business Opportunity/Risk Measures**: Business benefit gained due to quality data or business risk realized due to questionable data. Increase in competitive analytics due to data availability and Data Quality improvements
- **Compliance Measures**: Users with access to update/influence the master data are restricted to only those employees who have need and have been approved as part of their job functions

It is very important to establish the measures of Data Quality most important to your organization. This is required to establish a baseline for the quality of your data and to monitor the progress of your DQM initiatives.

The other foundational components of the Data Quality Cycle required to Discover, Profile, Establish Rules, Monitor, Report, Remediate, and continuously improve Data Quality are described in the next section.

**Components of DQM**

Once in place, these key components provide robust, reusable and highly effective DQM capabilities that can be leveraged across the enterprise:

- **Data Discovery**: The process of finding, gathering, organizing and reporting metadata about your data (e.g., files/tables, record/row definitions, field/column definitions, keys)
- **Data Profiling**: The process of analyzing your data in detail, comparing the data to its metadata, calculating data statistics and reporting the measures of quality for the data at a point in time
- **Data Quality Rules**: Based on the business requirements for each Data Quality measure, the business and technical rules that the data must adhere to in order to be considered of high quality
- **Data Quality Monitoring**: The ongoing monitoring of Data Quality, based on the results of executing the Data Quality rules, and the comparison of those results to
defined error thresholds, the creation and storage of Data Quality exceptions and the generation of appropriate notifications

- **Data Quality Reporting**: The reporting, dashboards and scorecards used to report and trend ongoing Data Quality measures and to drill down into detailed Data Quality exceptions
- **Data Remediation**: The ongoing correction of Data Quality exceptions and issues as they are reported

Each of these DQM components is described in greater detail in terms of roles and responsibilities, processes, technologies and business benefits in the sections that follow.

**Data Discovery**

**Roles and Responsibilities**
Data discovery is typically the responsibility of IT. However, tech-savvy business users/managers may also perform data discovery when user-friendly data discovery tools are available.

**Processes**
Data discovery should be an automated process using a robust data discovery tool. The data domains and physical database servers and/or file systems in scope must first be identified, and read-only security permissions to those database servers and/or file systems must be obtained in order to execute the discovery processes.

The discovery tool will gather all of the available metadata and store it in a discovery metadata repository where it can then be queried and analyzed. The metadata captured typically include database schema/file directory names, table/file names and definitions, column/field names and definitions, and any defined database or file relationships (e.g., primary/foreign key relationships).

**Technologies**
A data discovery tool should be utilized. Most, if not all, of the data discovery tools on the market also perform data profiling as well as other DQM functions.

**Business Benefits**
It is nearly impossible to properly manage and leverage data if there is no awareness that the data exists and no understanding of where it resides or how it is defined. Data discovery provides the awareness and understanding as a starting point for managing the data as an asset and obtaining business value from that data.

**Data Profiling**

**Roles and Responsibilities**
Here, again, data profiling is often the responsibility of IT; however, tech-savvy business users may also perform data profiling if the right tool is available.
Processes
Typically, data profiling is performed at the same time as data discovery. First, data discovery gathers all of the metadata related to the data, and then profiling utilizes that metadata to examine the actual data, compare the actual data to the metadata definitions, calculate data statistics, and store the data profiling results as additional metadata in the same repository as the discovery metadata. The profiling results can then be queried, analyzed, and exported by IT and business users.

The measures of Data Quality that should be available from a data profiling tool include completeness and uniqueness, as well as measures of accuracy, integrity, consistency, representation, and validity when compared with the data’s metadata definitions. Some data profiling tools can also compare the actual data to defined data domains and/or business rules to further determine the measures of Data Quality.

Technologies
A data profiling tool should be utilized. Although it is possible to perform some data profiling manually via SQL, visual inspection, etc., this should be avoided because manual methods often rely upon preconceived notions about the data. Data profiling tools have no preconceived notions of the data and deliver the profiling results based on what actually exists in the data.

Business Benefits
Data profiling picks up where data discovery leaves off and provides an even deeper and richer understanding of the actual data. Data profiling results and analysis are a critical input into key data-driven business and IT initiatives: business information and data strategies, data governance, data architecture, application data requirements, master data management, data integration, data warehousing, business intelligence, and advanced analytics.

Data Quality Rules

Roles and Responsibilities
It is typically the responsibility of Data Owners, as part of their Data Governance responsibilities, to define the Data Quality rules for the data domains they own. They should seek input from Data Stewards, Data Architects, and Data Analysts as they define the Data Quality rules. It is the responsibility of IT to design and develop reusable Data Quality rules that can then be implemented, executed, and monitored within batch, real-time and near real-time applications and data integration services across the enterprise.

Processes
Defining Data Quality rules should be driven by business requirements. IT requirements should also be considered (e.g., the quality of metadata required for internal data processing). The data discovery and profiling results are key inputs to be used in defining the Data Quality rules as they provide the metadata definitions of the data which the rules should be based upon.
One of the first steps in defining Data Quality rules is to identify the Critical Data Elements (CDEs) that are most important to the business for each data domain. Data Quality rules should then be developed for each CDE for each Data Quality measure relevant to that CDE (e.g., completeness, validity, accuracy).

**Technologies**
A best practice in the DQM technology space is to leverage a DQM tool that allows for the creation of a reusable Data Quality rule once and then the implementation of that rule across the enterprise on a variety of technology platforms. For example, the same Data Quality rule used to validate address data in a batch ETL process can also be used to prevent the entry of invalid address data in a real-time user application. The reusable rule has a single code base, but the rule can be implemented as a batch ETL component or as a web service that can be called from a real-time application.

**Business Benefits**
The definition, design, and implementation of Data Quality rules are at the heart of a solid DQM program. It is the Data Quality rules that will detect, correct, and prevent poor-quality data and elevate data across the enterprise to data that can be trusted as a valuable asset and used for transacting business and for reporting and analytics to steer the organization in the right direction.

**Data Quality Monitoring**

**Roles and Responsibilities**
It is the responsibility of the Data Owners and Stewards to define their Data Quality monitoring requirements (e.g., where to perform the monitoring, what Data Quality rules to execute, frequency, error thresholds, exceptions, notifications). Once the Data Quality monitoring requirements are defined, it is the responsibility of IT to design, develop, implement, and execute the Data Quality monitoring in production.

**Processes**
As with the Data Quality rules, the Data Quality monitoring should be driven by business requirements and the criticality of the data. The location and frequency of the monitoring, the error thresholds, the criticality of the exceptions, and the required notifications must all be based on business needs. IT must leverage the reusable Data Quality rules as they design, develop, and implement the required Data Quality monitoring components in the production environment.

As the Data Quality monitoring components execute the reusable Data Quality rules, the results are compared to the defined thresholds, and the required exceptions and notifications are generated. The exceptions are stored for Data Quality reporting purposes.

**Technologies**
Robust DQM platforms have Data Quality Monitoring capabilities that should be leveraged.
Business Benefits
As Data Quality is monitored, the business obtains continuous visibility into the quality of data required for key business processes and can react immediately to Data Quality exceptions and issues to correct them.

Data Quality Reporting

Roles and Responsibilities
It is the responsibility of Data Owners and Stewards to define their Data Quality Reporting requirements, and it is the responsibility of IT to design, develop, and implement the required Data Quality reporting components. IT may also be responsible for the design, development, implementation, and execution of any canned Data Quality reports, where required. The execution of ad-hoc, on-demand Data Quality reports, dashboards, or scorecards is the responsibility of the business or IT users.

Processes
A key dependency for Data Quality Reporting is the capture and storage of all Data Quality exceptions. This is performed by the Data Quality Monitoring components as they execute the Data Quality rules.

The detailed exceptions should then be aggregated to provide a count of the number of exceptions by dimensions (e.g., by date/time, source system/application, table/file, Data Quality rule). This aggregated data with exception counts and the underlying detailed exception data allows for the reporting and trending of DQ exceptions over time in dashboards and scorecards and the ability to drill-down into the detailed exceptions.

Data Quality reports, dashboards and scorecards are typically real-time, on-demand capabilities available to all business and IT users who have been granted user privileges. However, there may also be a need for canned Data Quality reporting where IT executes pre-defined reports in batch and distributes the results to all users who have subscribed to the Data Quality reports.

Technologies
A robust DQM platform with Data Quality Reporting capabilities should be leveraged. Those capabilities should include real-time, on-demand reports, dashboards and scorecards as well as batch reporting functionality. The best platforms should have the required data structures to handle the detailed exceptions and the aggregated counts by dimension built into the tool.

Business Benefits
As with Data Quality Monitoring, Data Quality Reporting provides visibility into the quality of critical business data at any point in time. Data Quality Reporting allows business and IT users to determine exactly what Data Quality exceptions are occurring and where those exceptions are originating from as the starting point for remediation efforts. In addition, the dashboards and scorecards with trending over time are particularly useful for understanding whether Data Quality is improving or deteriorating over time as an overall measure of the effectiveness of the DQM efforts.
Data Remediation

Roles and Responsibilities
It is the responsibility of Data Stewards, under the direction of Data Owners, to determine what data requires remediation and the best way to perform the remediation. Data Stewards are also responsible for performing the actual remediation tasks, with input and assistance from IT Data Analysts.

Processes
As Data Quality Monitoring and Reporting highlights data exceptions and the originating source systems or applications, a root cause analysis must be performed by the Data Stewards to determine what is causing the Data Quality exceptions.

Once the root cause(s) are determined, the Data Steward must develop a remediation plan and obtain buy-in and approval from the Data Owner and IT to execute the remediation steps. Remediation involves correcting the invalid data that has already been created and re-executing any business or data processes that have a critical dependency on the quality of the data (e.g., re-executing the creation of a sales report, re-executing a marketing campaign).

Determining how to prevent the Data Quality exceptions in the future and implementing the proper solution, which could involve many aspects of data governance and DQM.

Once the Data Quality Remediation plan is executed, the Data Quality Monitoring and Reporting should reflect improved Data Quality over time.

Technologies
If remediation points to the need for new or modified Data Quality rules, the DQM platform should be leveraged to create or modify the Data Quality rules. Other types of remediation efforts that require data governance or source system/application changes cannot be executed on the DQM platform.

Business Benefits
As poor-quality data is remediated the business benefits from having high quality data for critical business processes and functions. This can have a tremendously positive impact across the enterprise including more efficient processes, more accurate and timely business transactions, lower operating costs, more effective marketing and sales, and increased revenues.
Key Tenets for Establishing a DQM Program

There are a number of key tenets to consider in establishing an effective DQM program:

**Start Small**
Lay out a complete DQM future-state vision, but start small. Do not attempt to “boil the ocean” with your DQM program by trying to implement it all at once. A solid implementation roadmap that starts small and builds the program up over time is the best approach.

**Data Quality is an Ongoing Process**
Establishing DQM is not a one-time event, but an ongoing process that will evolve over time. Effective DQM is often a cultural shift in an organization and takes time. The DQM program should be re-evaluated periodically and modified as required.

**Focus on Data Quality Goals**
Establish the Data Quality measurements important to your organization, understand the Data Quality levels required by the business and set the appropriate Data Quality targets. Do not attempt to quickly fix Data Quality issues without understanding Data Quality measurements, requirements, and targets first.

**Target Data Quality Issues with High Payback**
Target the specific Data Quality pain areas that will provide the highest payback or ROI. As the overall DQM strategy and program is put in place over time, a focus should be placed on tackling the issues of most value to the business.

**Data Owners and Stewards Must Own DQM**
Data Owners and Stewards should take ownership of DQM. Consider tying their compensation to DQM targets to increase their accountability and responsibility for DQM.

**There is More Than One Solution to DQM**
Data typically flows throughout an enterprise from many sources/applications, through various data integration services, and into many target repositories. Data Quality can be managed at any point in the data flow, depending on data governance policies, business requirements, data ownership, system architectures, etc. There are often a number of choices available to manage Data Quality effectively.

Mastering your understanding of Data Quality Measures, the Data Quality Cycle, and keeping the above key tenets in mind will ensure you establish a solid DQM foundation.
About Knowledgent

Knowledgent is a data and analytics firm that helps organizations transform their information into business results through data and analytics innovation. We help clients maximize the value of information by improving their data foundations and advancing their analytics capabilities. We combine data science, computer science, and domain expertise to enable our clients to implement innovative, data-driven business solutions.

Knowledgent operates in the emerging world of big data as well as in the established disciplines of enterprise data warehousing, master data management, and business analysis. We not only have the technical knowledge to deliver game-changing solutions at all phases of development, but also the business acumen to evolve data initiatives from ideation to operationalization, ensuring that organizations realize the full value of their information.

For more information about Knowledgent, visit www.knowledgent.com.