ACTIAN™
a division of HCLSoftware

zeenea

SOLUTION GUIDE

A Guide to Data Quality Management

for data-driven organizations

Table of Contents

3 Introduction

3 What is Data Quality?

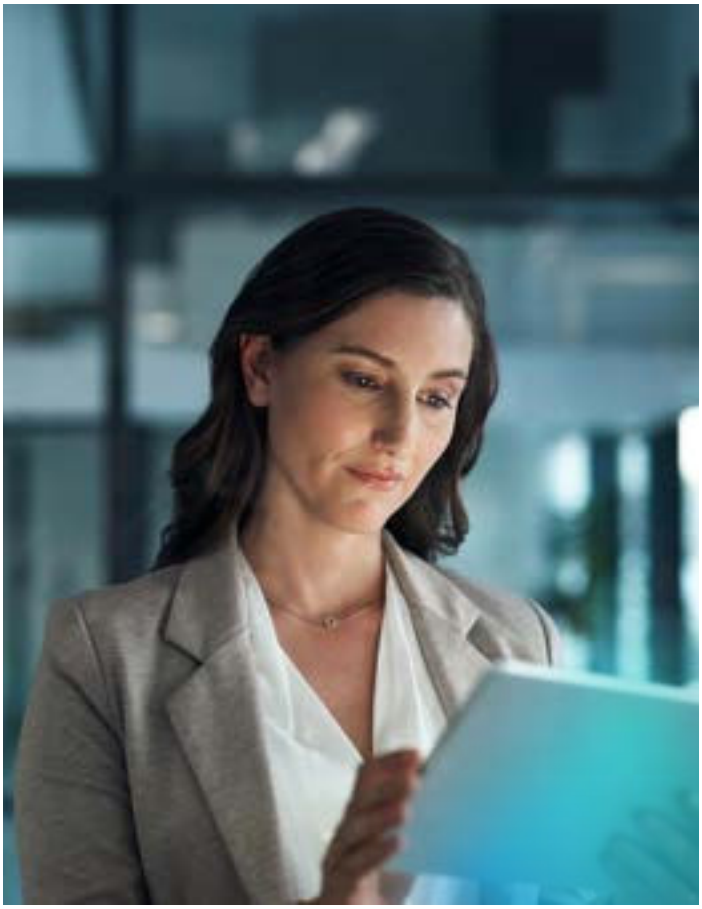
4 The nine dimensions of Data Quality

9 Data Quality Management

12 The features of a Data Quality Management tool.

14 Data catalogs and Data Quality Management

16 Take away



ACTIAN™ a division of HCLSoftware | zeenea

Introduction

If we were to take some data professionals at their word, improving Data Quality is the panacea to all our business woes and should therefore be the top priority. In our opinion, this should be nuanced: Data Quality is a means amongst others to limit the uncertainties of meeting corporate objectives.

> Data Quality usually refers to a company's ability to ensure the longevity of its data.

At Zeenea (a data catalog provider), we believe Data Quality is ensured through the 9 following dimensions - completeness, accuracy, validity, uniqueness, consistency, timeliness, traceability, clarity, and availability - all essential to extract value to your company. We will detail these dimensions with the help of a simple example in part one.

We will then elaborate on how Data Quality Management is an important challenge for organizations seeking to extract maximum value from their data.

We will also draw parallels between these different Data Quality dimensions and the different risk management phases to overcome - identification, analysis, evaluation, and processing. This will enable you to hone your risk management reflexes by tying in Data Quality improvement processing to a company objective (and evaluating the ROI on each quality dimension).

Once we have established the main features of an enterprise Data Quality Management tool, we will detail how a data catalog - though not a Data Quality tool- can contribute towards Data Quality improvement (through the clarity, availability, and traceability dimensions mentioned above).

What is Data Quality?

The definition of DAMA

Asking Data Analysts or Data Engineers for a definition of Data Quality will provide you with very different answers - even within the same company, amongst similar profiles. Some, for example, will focus on the unity of data, while others will preferto reference standardization. You may yourself have your own interpretation.

The ISO 9000-2015 norm defines quality as "the capacity of an ensemble of intrinsic characteristics to satisfy requirements". DAMA International (The Global Data Management Community) – a leading international association involving both business and technical data management professionals - adapts this definition to a data context:

> Data Quality is the degree to which the data dimensions meet requirements.

Data Quality is also often described as a "fitness for use", meaning it meets certain standards for usage and expected objectives.

The dimensional approach to Data Quality

From an operational perspective, Data Quality translates into what we call **Data Quality dimensions**, in which each dimension relates to a specific aspect of quality. The 4 dimensions most often used are generally completeness, accuracy, validity, and availability - which we will detail further down.

In literature, there are many dimensions and different criteria to describe Data Quality. There isn't however any consensus on what these dimensions actually are. For example, DAMA enumerates sixty dimensions – when most Data Quality Management (DQM) software vendors usually offer up five or six.

The nine dimensions of Data Quality

At Actian Zeenea, we believe that the ideal compromise is to take into account nine Data Quality dimensions: completeness, accuracy, validity, uniqueness, consistency, timeliness, traceability, clarity, and availability.

We will illustrate these nine dimensions and the different concepts we refer to in this publication with a straightforward example.

> Arthur is in charge of sending marketing campaigns to clients and prospects to present his company's latest offers. He encounters, however, certain difficulties:
>
> - Arthur sometimes sends communications to the same people several times,
> - The emails provided in his CRM are often invalid,
> - Prospects and clients do not always receive the right content,
> - Some information pertaining to the prospects are obsolete,
> - Some clients receive emails with erroneous gender qualifications,
> - There are two addresses for clients/prospects but it's difficult to understand what they relate to,
> - He doesn't know the origin of some of the data he is using or how he can access their source.

Below is the data Arthur has at hand for his sales efforts. We shall use them to illustrate each of the nine dimensions of Data Quality:

Client Type	Gender	Title	First Name	Last Name	Street Address 1	Street Address 2	Email	Last Update (min)
Prospect	Male	Mr.	John	Kelly	384 Broad St.	300 Great Swamp Rd.	jkelly@wallmart.com	54
Customer	Female	Mrs.	Lisa	Smith	123 Summer Rd.	345 Sewgen Dr.	l.smith@amazon.com	2
Prospect	Female	Mrs.	Anna	Lincoln	7262 18th St.	460 Newcomb Dr.	annalincoln@apple	498000
Csustomer	Male	Mr.	Patrico	Brown	88 Rio Grand Ave.	23 Tasaga Ave.	pbrown@cvshealth.com	40
Customer	Female	Mr.	Lisa	Smith	123 Summer Rd.	345 Sewgen Dr.	l.smith@amazon.com	2
Prospect	Female	Mr.	Aria	Duz	5 Baker's Lane	1248 Broadway	aria.duz@mckesson.com	120
Customer	Male	Mrs.	Lino	Rodriguez	1230 Hoes Lane	283 Davidson Dr.	lino.rodriguez@exxonmobil.com	47
Customer	Male	Mr.	Richard	Isantengse	Mulberry 71 St.	987 First Ave.		12

Completeness

Is the data complete? Is there information missing? The objective of this dimension is to identify the empty, null, or missing data.

> In this example, Arthur notices that there are missing email addresses:

Client Type	Gender	Title	First Name	Last Name	Street Address 1	Street Address 2	Email	Last Update (min)
Customer	Female	Mr.	Lisa	Smith	123 Summer Rd.	345 Sewgen Dr.	l.smith@amazon.com	2
Prospect	Female	Mr.	Aria	Duz	5 Baker's Lane	1248 Broadway	aria.duz@mckesson.com	120
Customer	Male	Mrs.	Lino	Rodriguez	1230 Hoes Lane	283 Davidson Dr.	lino.rodriguez@exxonmobil.com	47
Customer	Male	Mr.	Richard	Isantengse	Mulberry 71 St.	987 First Ave.		12

> It could be helpful here for Arthur to use postal address verification services.

Validity

Does the data conform with the syntax of its definition? The purpose of this dimension is to ensure that the data conforms to a model of a particular rule.

> Arthur noticed that he regularly gets bounced emails. Another problem is that certain prospects/clients do not receive the right content because they haven't been accurately qualified. For example, the email address annalincoln@apple isn't in the correct format and the Client Type Csutomer isn't correct.
>
> To solve this issue, he could for example make sure that the Client Type values are part of a list of reference values (Customer or Prospect) and that email addresses conform to a specific format.

Client Type	Gender	Title	First Name	Last Name	Street Address 1	Street Address 2	Email	Last Update (min)
Customer	Female	Mrs.	Lisa	Smith	123 Summer Rd.	345 Sewgen Dr.	l.smith@amazon.com	2
Prospect	Female	Mrs.	Anna	Lincoln	7262 18th St.	460 Newcomb Dr.	annalincoln@apple	498000
Csustomer	Male	Mr.	Patrico	Brown	88 Rio Grand Ave.	23 Tasaga Ave.	pbrown@cvshealth.com	40
Customer	Female	Mr.	Lisa	Smith	123 Summer Rd.	345 Sewgen Dr.	l.smith@amazon.com	2

Validity vs. Accuracy

It is sometimes difficult to distinguish between those 2 dimensions. Validity relates to the coherence of the data in relation to a defined domain. Accuracy relates to the data describing an object or event from the real world. We could, for example, run a validity test on a postal code that is valid, while the accuracy test shows the address to be wrong - the postal code can be considered valid when a given city in the "real world" does not exist.

Consistency

Are the different values of the same record in conformity with a given rule? The aim is to ensure the coherence of the data between several columns

> Arthur noticed that some of his male clients complain about receiving emails in which they are referred to as Miss. There does appear to be an incoherence between the Gender and Title columns for Lino Rodriguez.
>
> To solve these types of problems, it is possible to create a logical rule that ensures that when the id Gender is Male, the title should be Mr.

Client Type	Gender	Title	First Name	Last Name	Street Address 1	Street Address 2	Email	Last Update (min)
Customer	Female	Mr.	Lisa	Smith	123 Summer Rd.	345 Sewgen Dr.	l.smith@amazon.com	2
Prospect	Female	Mr.	Aria	Duz	5 Baker's Lane	1248 Broadway	aria.duz@mckesson.com	120
Customer	Male	Mrs.	Lino	Rodriguez	1230 Hoes Lane	283 Davidson Dr.	lino.rodriguez@exxonmobil.com	47
Customer	Male	Mr.	Richard	Isantengse	Mulberry 71 St.	987 First Ave.		12

Timeliness

Is the time lapse between the creation of the data and its availability appropriate? The aim is to ensure the data is accessible in as short a time as possible.

> Arthur noticed that certain information on prospects is not always up to date because the data is too old. As a company rule, data on a prospect that is older than 6 months cannot be used.
>
> He could solve this problem by creating a rule that identifies and excludes data that is too old. An alternative would be to harness this same information in another system that contains fresher data.

Client Type	Gender	Title	First Name	Last Name	Street Address 1	Street Address 2	Email	Last Update (min)
Prospect	Male	Mr.	John	Kelly	384 Broad St.	300 Great Swamp Rd.	jkelly@wallmart.com	54
Customer	Female	Mrs.	Lisa	Smith	123 Summer Rd.	345 Sewgen Dr.	l.smith@amazon.com	2
Prospect	Female	Mrs.	Anna	Lincoln	7262 18th St.	460 Newcomb Dr.	annalincoln@apple	498000
Csustomer	Male	Mr.	Patrico	Brown	88 Rio Grand Ave.	23 Tasaga Ave.	pbrown@cvshealth.com	40
Customer	Female	Mr.	Lisa	Smith	123 Summer Rd.	345 Sewgen Dr.	l.smith@amazon.com	2

Uniqueness

Are there duplicate records? The aim is to ensure the data is not duplicated.

> Arthur noticed that certain information on prospects is not always up to date because the data is too old. As a company rule, data on a prospect that is older than 6 months cannot be used.

Client Type	Gender	Title	First Name	Last Name	Street Address 1	Street Address 2	Email	Last Update (min)
Prospect	Male	Mr.	John	Kelly	384 Broad St.	300 Great Swamp Rd.	jkelly@wallmart.com	54
Customer	Female	Mrs.	Lisa	Smith	123 Summer Rd.	345 Sewgen Dr.	l.smith@amazon.com	2
Prospect	Female	Mrs.	Anna	Lincoln	7262 18th St.	460 Newcomb Dr.	annalincoln@apple	498000
Csustomer	Male	Mr.	Patrico	Brown	88 Rio Grand Ave.	23 Tasaga Ave.	pbrown@cvshealth.com	40
Customer	Female	Mr.	Lisa	Smith	123 Summer Rd.	345 Sewgen Dr.	l.smith@amazon.com	2

Clarity

Is understanding the metadata easy for the data consumer? The aim here is to understand the significance of the data and avoid interpretations.

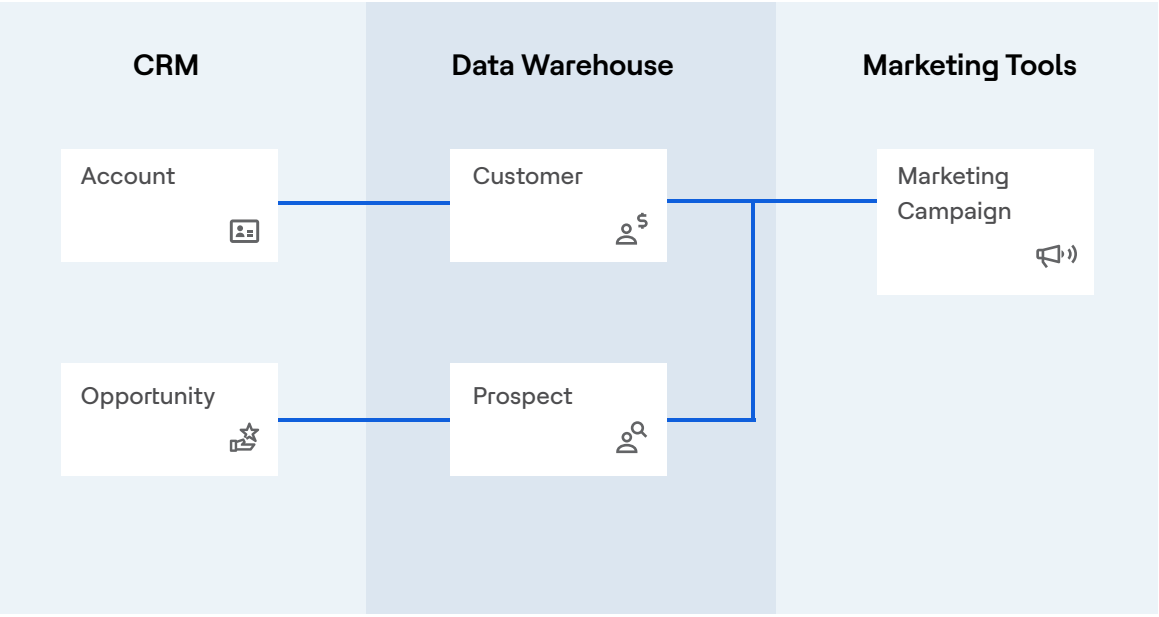
> Arthur has doubts about the two addresses given as it is not easy to understand what they represent. The names Street Address 1 and Street Address 2 are subject to interpretation and should be modified, if possible. Renaming within a database is often a complicated operation and should be correctly documented with at least one description:

Client Type	Gender	Title	First Name	Last Name	Street Address 1 (Billing Address)	Street Address 2 (Billing Address)	Email	Last Update (min)

Traceability

Is it possible to obtain traceability from data? The aim is to get to the origin of the data, along with any transformations it may have gone through

> Arthur doesn't really know where the data comes from or where he can access thedata sources. It would have been quite useful for him to know this as it would have ensured the problem was fixed at the source. He would have needed to know that the data he is using with his marketing tool originates from the data of the company data warehouse, itself sourced from the CRM tool.



ACTIAN a division of HCLSoftware | zeenea

Availability

How can the data be consulted or retrieved by the user? The aim is to facilitate access to the data.

Arthur doesn't know how to easily access the source data. Staying with the previous schema, he wants to effortlessly access data from the data warehouse or the CRM tool. In some cases, Arthur will need to make a formal request to access this information directly

Data Quality Management

The challenges of Data Quality for organization

Initiatives for improving the quality of the data are usually put in place by organizations to **meet the conformity requirements and risk reduction**. They are indispensable for reliable decision-making. There are unfortunately **many stumbling blocks that can hinder Data Quality improvement initiatives**. Below are some examples:

- The exponential growth of the volume, speed, and variety of the data make the environment more complex and uncertain;
- Increasing pressure from conformity regulations such as GDPR, BCBS 239, or HIPAA;
- Teams are increasingly decentralized, and each have their own domain of expertise
- IT and data teams are snowed under and don't have time to solve Data Quality issues;
- The data aggregation processesare complex and long;
- It can be difficult to standardize data between different sources;
- Change audits among systems are complex;
- Governance policies are difficult to implement

Having said that, there are also numerous opportunities to grab. High-quality data enables organizations to facilitate innovation with artificial intelligence and ensure a more personalized customer experience. Assuming there is enough quality data. Gartner has actually forecasted that until 2022, 85% of AI projects will produce erroneous data as a result of bias in the data, algorithms, or from teams in charge of data management

Reducing the level of risk by improving the quality of the data

Poor Data Quality should be seen as a risk and quality improvement software as a possible solution to reduce this level of risk

Processing a quality issue:

If we accept the notion above, any quality issue should be addressed in several phases: Data Quality improvement initiatives. Below are some examples:

1. **Risk identification**: this phase consists in seeking out, recognizing, and describing the risks that can help/prevent the organization from reaching its objectives – in part because of a lack of Data Quality

> if we take Arthur's example, poor Data Quality is a risk that negatively impacts its marketing objective of ensuring a 30% rise in leads generated by the end of the year.

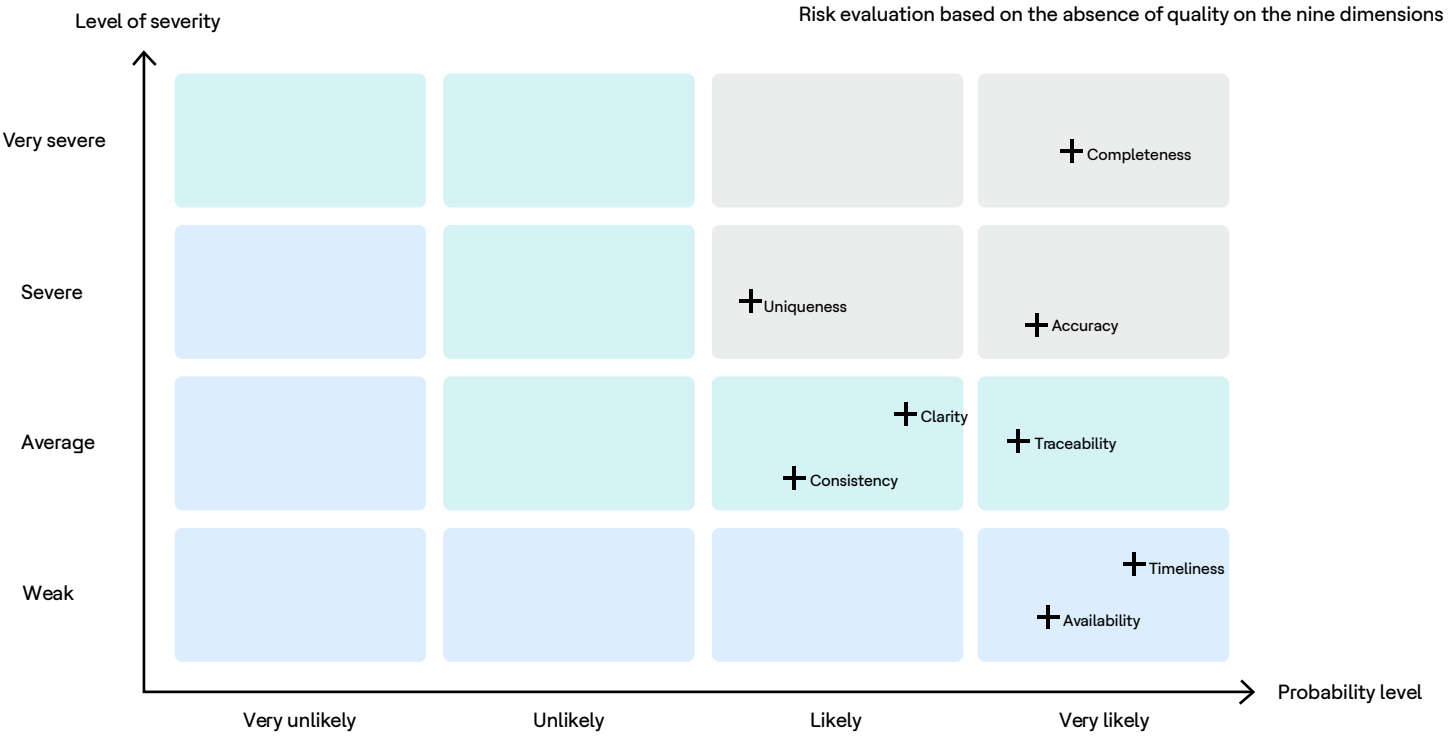
2. **Risk Analysis:** the aim of this phase is to understand the nature of the risk and its characteristics. It includes factors for event similarities and their consequences, the nature, and importance of these consequences, etc.

   Here, we should seek to identify what has caused the poor quality of the marketing data. We could cite for example:

- ✕ A poor user experience of the source system leading to typing errors;
- ✕ A lack of verification of the completeness, accuracy, validity, uniqueness, consistency, or timeliness ofthe data;
- ✕ A lack of simple means to ensurethe traceability, the clarity, and availability of the data; The data aggregation processes are complex and long;
- ✕ The absence of a governance process and the implication of business teams.

> The consequences for Arthur are immediate and mostly financial. For instance, the lack of completeness on emails will reduce the efficiency ofthe marketing campaign. Similar consequences could be seen in all nine dimensions of Data Quality.

3. **Risk evaluation:** the purpose of this phase is to compare the results of the risk analysis with the established risk criteria. It helps establish whether further action is needed for the decision-making – for instance keeping the current means in place, undertaking further analysis, etc. Let's focus on the nine dimensions of Data Quality and evaluate the impact of poor quality on each of them for Arthur's objectives probability/severity matrix:



The values for the levels of probability and severity should be defined by the main stakeholders, who know the data in question best.

> For example, if we focus on the completeness of the emails, poor quality will curtail the number of communications by 6%. The evaluation of this percentage can only be achieved by combining data profiling with theexperience of Arthur and his employer. Bearing in mind that marketing campaigns turn 0,5% of emails into leads, we can deduce the financial impact of Data Quality on completeness. We can therefore define the risk of poor quality on completeness as very high/very likely.

We could run the same exercise on each quality dimension and the other identified risks.

4. **Risk processing:** this processing phase aims to set out the available options to reduce risk and roll them out. This processing also involves the ability to assess the usefulness of the actions taken, determining whether the residual risk is acceptable or not - and in this last case - consider further processing.

Improving the quality of the data is clearly not a goal in itself

- Its cost must be evaluated based on company objectives;
- The treatments to be implemented must be evaluated through each dimension of quality.

> In order to reduce the risks of poor quality on the completeness dimension, it could make sense to complete all the emails by collecting data from other CRMs and asking business people to fill in missing emails manually. ROI seems important given the impact on the level of severity and the relatively low cost of processing.
>
> We could think of a different approach to processing for the accuracy dimension. We saw that, in Arthur's case, the problems were linked to invalid email addresses. To correct the data, we could get help from an external email verification service, with an invoice per transaction. In this case, the impact on the objectives isn't as strong as it is with completeness. Indeed, many letters will still reach the intended recipients despite errors in the addresses. Here it would make sense to define criteria for the extent of data correction needed. We could use the verification service to reduce the percentage of invalid addresses from 10% to 5% (or we could also opt to not do anything for that dimension)

The features of a Data Quality Management tool.

One way to better understand the challenges of Data Quality is to look at the existing Data Quality solutions on the market.

From an operational point of view, how do we identify and correct Data Quality issues? What features do Data Quality Management tools offer to improve Data Quality?

Without going into too much detail, let's illustrate the pros of a Data Quality Management tool through the main evaluation criteria of Gartner's **Magic Quadrant for Data Quality Solutions**

Connectivity

A Data Quality Management tool has to be able to gather and apply quality rules on all enterprise data (internal, external, on-prem, cloud, relational, non-relational, etc.). The tool must be able to plug into all relevant data in order to apply quality rules.

> **Taking the example of Arthur, it would make sense to connect directly to the data warehouse to evaluate the quality of the data, and to apply, for example, the rules of coherence between genders and titles in order to avoid having "Male" associated with "Mrs".**

Data profiling, data measuring, and data visualization

You cannot correct Data Quality issues if you cannot detect them first. Data profiling enables IT and business users to assess the quality of the data in orderto identify and understand the Data Quality issues.

Monitoring

The tool must be able to monitor the evolution of the quality of the data and warn management at a certain point.

> **Staying with Arthur's example, it could make sense to receive an alert as soon as 5% of data lacks an email address. It would entail control and surveillance solely on the completeness dimension for the email attribute.**

Data standardization and data cleaning

Then comes the data cleaning phase. The aim here is to provide data cleaning functionalities in order to **enact norms or business rules** to alter the data (format, values, page layout).

Data matching and merging

The aim is to **identify and delete duplicates** that can be present within or between data sets.

> Arthur had, for instance, identified duplicates for the Lisa Smith entry within his marketing data. He, therefore, needs an algorithm or rules to identify and merge these duplicates.

Address validation

The aim is to **standardize addresses** that could be incomplete or incorrect.

> Arthur noticed in the accuracy dimension that an address wasn't formatted as it should be. A validation and email correction service could solve the problem.

Data curation and enrichment

The capabilities of a Data Quality Management tool are what enable the **integration of data from external sources** and improve the completeness, thereby adding value to the data.

> Arthur noticed that certain email fields were empty. A data integration tool would help get additional emails in another CRM system to solve the issue. If the data isn't in another system, he could ask people to insert the missing data manually.

The development and putting in place of business rules

The capabilities of a Data Quality Management tool are what **enable the creation, deployment, and management of business rules**, which can then be used to validate the data.

> The Client Type is not always correct (Csutomer instead of Customer for example). Arthur could establish a dictionary of possible values (Customer or Prospect) which can then be exploited to quickly identify quality issues.

Problem resolution

The quality management tool helps both IT and business users to **assign, escalate, solve, and monitor Data Quality problems.**

Metadata management

The tool should also be capable of **capturing and reconciling all the metadata related to the Data Quality process.**

User-friendliness

Lastly, a solution should be able to **adapt to the different roles within the company**, and specifically to non-technical business users.

Data catalogs and Data Quality Management

As indicated in the introduction of this ebook, I am a Product Manager at **Actian Zeenea** - a data catalog solution pure player. It is important at this stage to highlight the link between the two disciplines of Data Quality Management and data catalog use at the enterprise level.

A data catalog is not a DQM tool

An essential element is that **a data catalog should not be considered as a Data Quality Management tool per se**, as we describe **in this article.**

First of all, one of the core principles at the heart of Data Quality is that **controls should ideally take place in the source system**. Running these controls solely in the data catalog - rather than at the source and the data transformation flow - increases the global cost of the undertaking.

Furthermore, a data catalog must be both comprehensive and less intrusive to facilitate its rapid deployment within the company. This is simply incompatible with the complex nature of data transformation and the multitude of tools used to carry out ./these transformations.

Lastly, a data catalog must remain a **simple tool** to understand and use, as described in article 3 of our **Data Democracy**.

A data catalog's contribution to DQM

While the data catalog isn't a Data Quality tool, its contribution to the upkeeping of Data Quality is nonetheless substantial. Here is how:

- **A data catalog enables data consumers to easily understand metadata and avoid hazardous interpretations** around the data. It echoes the clarity dimension of quality;
- **A data catalog gives a centralized view of all the available enterprise data**. Data Quality information is therefore metadata like any other that carries value and should be made available to all. They are easy to interpret and extract, an echo to the dimensions of accuracy, validity, consistency, uniqueness, completeness, and timeliness.

> It's key that a data catalog has the capacity of retrieving and updating all this information via APIs. This is all the more important for organizations that use different Data Quality Management tools for different siloes and/or data dimensions.

- **A dvA data catalog usually allows direct access to the data sources**, echoing the availability dimension of quality.

The implementation strategy of the DQM

The following table details how Data Quality is taken into account depending on the different solutions on the market:

Data Quality Dimensions	Source System	Data Catalog **without** Integrated Quality Rules	Data Catalog **with** Integrated Quality Rules	Specialized Data Quality Tool	Data Catalog Integrated to a Specialized Data Quality Tools
Completeness	✔ (Verification upon input)	✘	Partial (Data Quality not integrated to the transformation flow and in the source system)	✔	✔
Accuracy	✔ (Verification upon input)	✘	Partial (Data Quality not integrated to the transformation flow and in the source system)	✔	✔
Validity	✔ (Verification upon input)	✘	Partial (Data Quality not integrated to the transformation flow and in the source system)	✔	✔
Consistency	✔ (Verification upon input)	✘	Partial (Data Quality not integrated to the transformation flow and in the source system)	✔	✔
Timeliness	✔ (Informed stored and filtered in the source system)	✘	Partial (Data Quality not integrated to the transformation flow and in the source system)	✔	✔
Uniqueness	✔ (Verification upon input)	✘	Partial (Data Quality not integrated to the transformation flow and in the source system)	✔	✔
Traceability	✘	✔ (Lineage capabilities)	✔ (Lineage capabilities)	✘	✔ (Lineage capabilities)
Clarity	✘	✔	✔	✘	✔
Availability	✘	✔ (Sources access capabilities)	✔ (Sources access capabilities)	✘	✔ (Sources access capabilities)

As stated above, quality testing should by default take place directly in the source system. Quality test integration in a data catalog can improve user experience, but it isn't a must in light of its limitations - as Data Quality isn't integrated into the transformation flow.

That said, when the systems stacks become too complex and we need, for example, to consolidate data from different systems with different functional rules, a Data Quality tool becomes unavoidable.

The implementation strategy will depend on use cases and company objectives. It is nonetheless apropos to put Data Quality in place incrementally to:

1. Ensure the source systems have put in place the relevant quality rules;
2. Implement a data catalog to improve quality on the dimensions of clarity, traceability, and/or availability;
3. Integrate Data Quality in the transformation flows with a specialized tool, while importing this information automatically in the data catalog via APIs.

Take away

Data Quality refers to the ability of a company to maintain the sustainability of its data over time. At Actian Zeenea, we define it through the prism of nine of the sixty dimensions described by DAMA International: completeness, accuracy, validity, uniqueness, consistency, timeliness, traceability, clarity, and availability. Data Quality Management has become a real challenge for all data-driven organizations, which are implementing strategies for continuous improvement of quality and its dimensions to meet compliance and risk reduction requirements. To achieve this objective, they can equip themselves with Data Quality Management tools, the main features of which we have detailed in this document. As a data catalog provider, we reject the idea that a data catalog is a full-fledged quality management tool. In fact, it is only one of several ways to contribute to the improvement of Data Quality, notably through the dimensions of clarity, availability, and traceability.

About Actian

Actian makes data easy. We deliver cloud, hybrid, and on-premises data solutions that simplify how people connect, manage, and analyze data. We transform business by enabling customers to make confident, data-driven decisions that accelerate their organization's growth. Our data platform integrates seamlessly, performs reliably, and delivers at industry-leading speeds. Learn more about Actian, a division of HCLSoftware: **www.actian.com**

Tel +1 512 231 6000 Fax +1.512.231.6010 710 Hesters Crossing Road, Suite 250, Round Rock, TX 78681 **Actian.com**

ACTIAN™ a division of HCLSoftware | zeenea