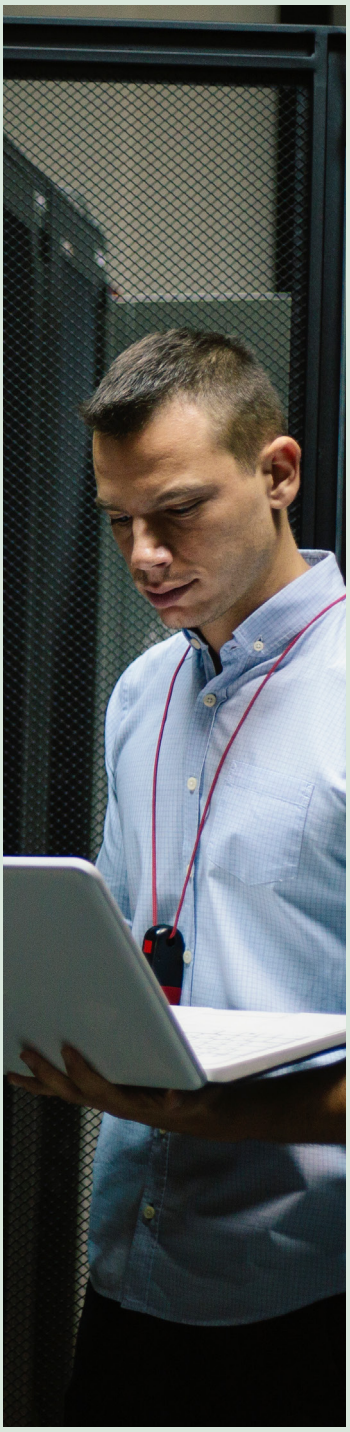
Infosys cobalt

WHITE PAPER



A DATA QUALITY APPROACH FOR ENTERPRISE DIGITAL TRANSFORMATION PROGRAMS: A PRACTITIONER'S PERSPECTIVE

Abstract

Data quality is a prime concern for organizations owing to its impact on operations as well as transformation initiatives. Yet many enterprises lack a firm understanding of what comprises quality data. This knowledge is crucial to formulate an effective data quality strategy, which, when executed right, helps organizations unlock maximum potential from enterprise data. This paper looks at the dimensions that make up data quality, the steps within enterprise data quality journeys, and how organizations can improve their data quality for successful digital transformation programs.

Infosys®
Navigate your next

Introduction

Organizations across the globe are investing in and running digital transformation programs to keep pace with changing customer demand, business models, and innovative technologies. However, not all of these programs succeed as expected. An analysis of the figures around failed digital transformation programs shows that 'poor data quality' is often the root cause. According to IBM, businesses lose US $3.1 trillion every year due to poor quality data.

Ensuring good quality of data is an integral part of digital transformation success. As a discipline, data quality has matured over the years and the market has several tools to handle data quality needs. But first, organizations should understand how data travels across an enterprise along with the metrics of measuring data quality, so they can devise the right data quality approach for their transformation programs.

Defining Data Quality

The quality of data influences how enterprise data is handled during any IT initiative within organizations. For instance, it affects how data is converted, integrated, analyzed, secured, and reported.

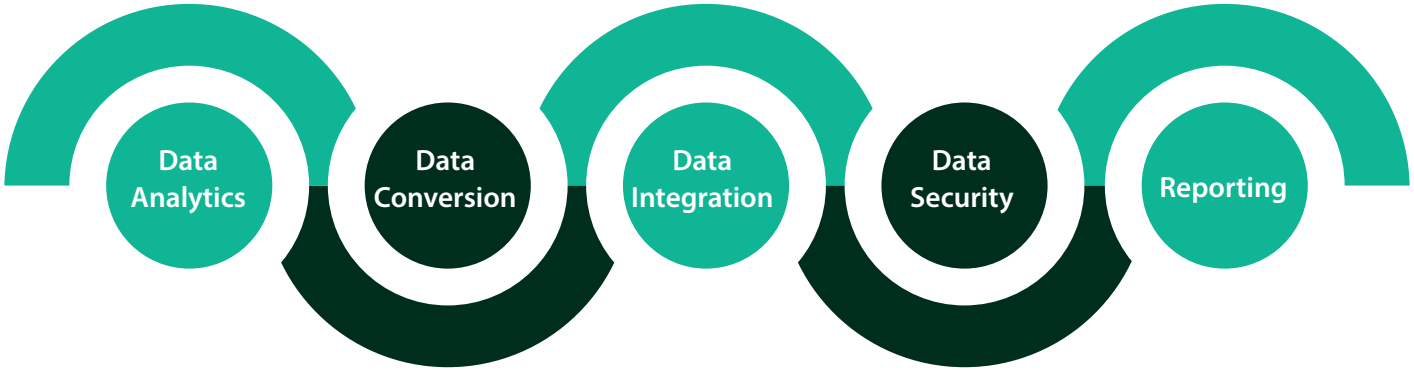


Figure 1 – Impact of data quality on enterprise data

Some of the defining and measurable dimensions of data quality are its completeness, accuracy, consistency, conformity, integrity, and duplication. Figure 2 examines each of the dimensions in greater detail.

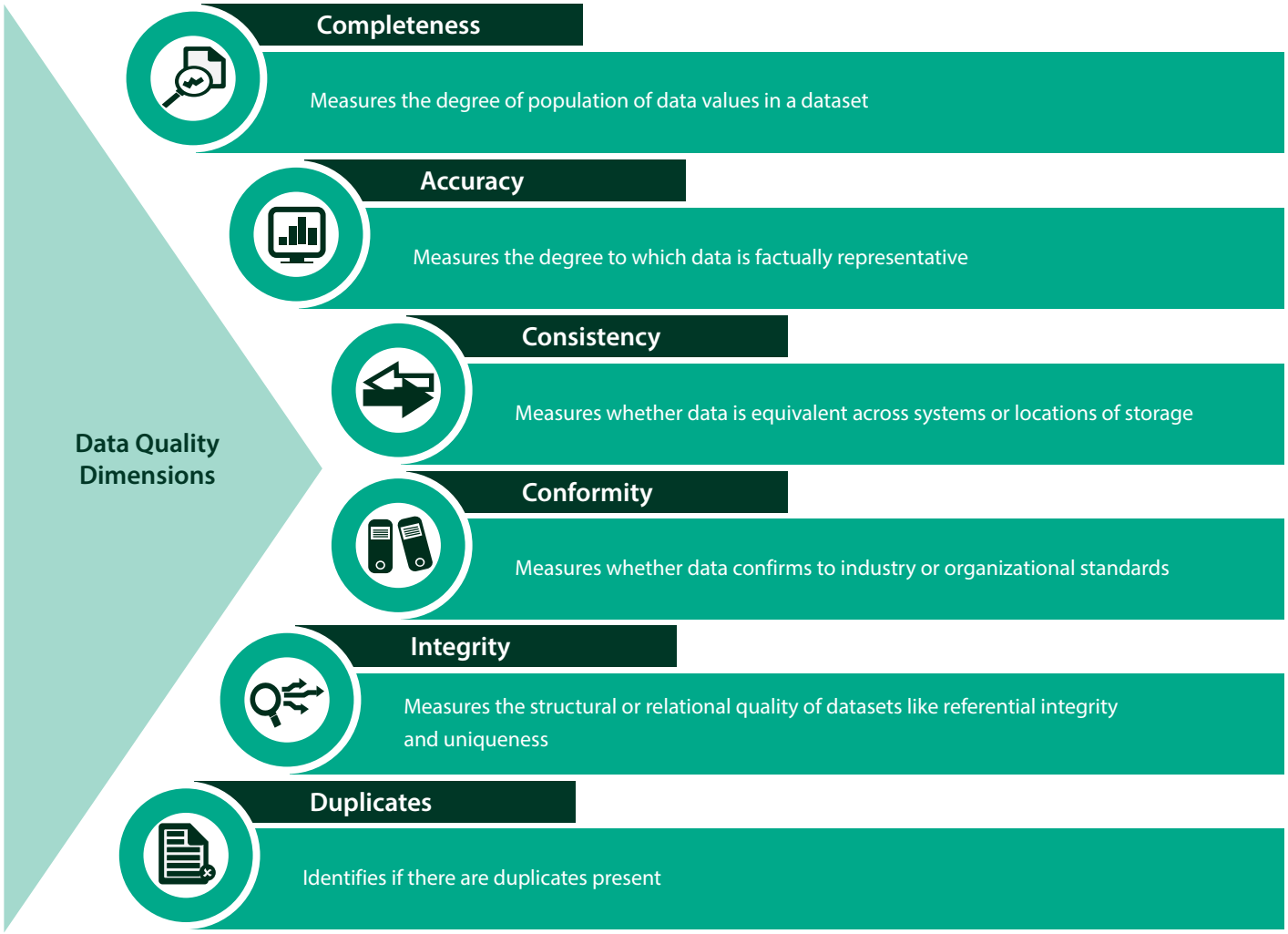


Figure 2 – Dimensions of data quality

The Data Quality Journey for Large Enterprises

The data quality journey in an enterprise must be a continuous one with the objective of getting clean data and keeping data clean. Unless data quality initiatives are envisaged and executed as a continuous journey, data is bound to get tainted or 'dirty' over time.

The data quality journey in an enterprise is a step-by-step process that includes the following stages and sub-stages:

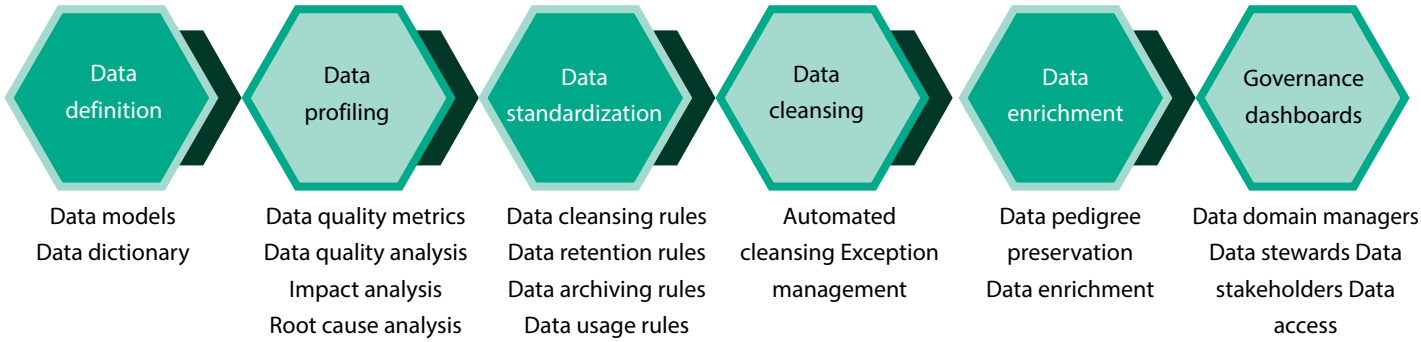


Figure 3 – The data quality journey for large enterprises

1. Data definition

Data definition is the foundational phase for any large-scale data cleansing program in an enterprise. It includes two steps:

- **Data modeling:** The data model is critical to model the data entity including its structure, hierarchy, and attributes in the tool. Most tools come with pre-built best practice-based data models that can be leveraged to get a head start in data model definition. Both logical and physical data models must be defined during this phase.
- **Data dictionary:** This is one of the most critical documents, which serves as the baseline to maintain metadata information for all data elements that are included as part of the data entity. The data dictionary contains critical information regarding description, ownership, uniqueness, and technical details about data attributes. Data dictionaries must be kept updated by data stewards so that these reflect the latest attribution status for critical data entities.

2. Data profiling

A comprehensive data profiling exercise is necessary for all data cleansing activities. Data profiling gives an overall view of the quality of data across various applications and helps users validate any overall hypothesis around the quality of data. It also catalogs the true scope of existing data quality issues. Data profiling reveals certain unknowns about the current state of data and identifies the areas of focus on the data quality journey.

3. Data standardization

Data standardization ensures that everybody in the organization speaks the same language regarding data. In order to have a common baseline for creating data, it is important to have a predefined set of data standards that all users can follow. Data standards must include naming conventions, formatting rules, and adherence to international naming standards, among others. Data standards must also be kept up-to-date and maintained on a regular basis with clearly defined ownership.

4. Data cleansing

Data cleansing is the actual stage of cleaning the data, either manually or through data quality tools. There are two modes in which data can be cleansed:

- **Data cleansing at source/real-time cleansing:** Data cleansing at source is carried out when there is no impact on open transactional data during the data cleanup.
- **Offline data cleaning:** Offline data cleansing is carried out by extracting data from the application and then cleaning it. This method is used primarily when the extracted/cleansed data needs to be migrated to a new application.

5. Data enrichment

Data enrichment is the process of enriching the data with additional attributes either though a business-driven enrichment process or through external sources like D&B Optimizer for customer master data. Data enrichment can be automated or rule-based through a data quality tool, provided the rules are established upfront in the application.

6. Data governance dashboards

Data governance dashboards are needed to monitor ongoing improvements in the quality of data. It also helps provide visibility to leadership on the status of the data cleansing initiatives. It is important to set up data governance dashboards with the right set of key performance indicators (KPIs) so as to better govern and manage the key master data entities in the enterprise.

Measuring and Monitoring Data Quality KPIs

Organizations looking to improve the success of digital transformation programs need a systematic and KPI-driven approach to ensuring data quality. This means setting data quality objectives and defining measurable KPIs. Some of the guidelines for a successful data quality initiative are:

- Identify the right KPIs for data quality
- Ensure the KPIs are measurable and align with data quality dimensions
- Establish clear business ownership around the KPIs
- Continuously monitor KPI performance
- Implement continuous upkeep of data quality

For example, an organization seeking to improve the quality of customer addresses in their application adopted the following workflow for data quality:

- Build a profile of the address data and arrive at current state for data quality
- Define data quality standards
- Set up a data dictionary
- Clean the address data (batch cleansing)
- Use ongoing address cleansing (real-time address cleansing)
- Adhere to data standards (manually or through tools)
- Report data quality improvements in customer address data over time

How to Improve Data Quality

1. Choose the right tool

Tools play a significant role in automating and scaling data quality programs. Organizations must be careful in choosing the right data quality tool that will serve their needs. Some criteria to consider when evaluating data quality tools are:

- Is the tool easy to use (i.e., minimal coding) by data stewards?
- Is it scalable to handle large data volumes?
- Does it have strong data profiling capabilities to support data pulls from different sources?
- Does it offer strong dashboard capabilities?
- Is it a cloud-first solution with extensible architecture?

Infosys Data Workbench (iDW) is a data quality tool that assists enterprises in their data quality journey by providing end-to-end capabilities from data profiling to data cleansing. It also leverages AI/ML prediction capabilities to auto-identify and auto-cleanse data.

Apart from this, there are other tools in the market like Cloud Data Quality from Informatica, Talend, Stibo, Ataccama, Oracle Enterprise Data Quality, Syniti, SAP, etc.

2. Find appropriate business stewards for data quality

Business stewards play a significant role is ensuring high-quality data across the enterprise. Organizations ought to identify the right data governance structures to maintain and manage data.

Data stewards can help in this regard by monitoring the data quality reports, carrying out manual data merges, maintaining the data dictionary, and periodically auditing the data. They are responsible for the overall maintenance of data quality.

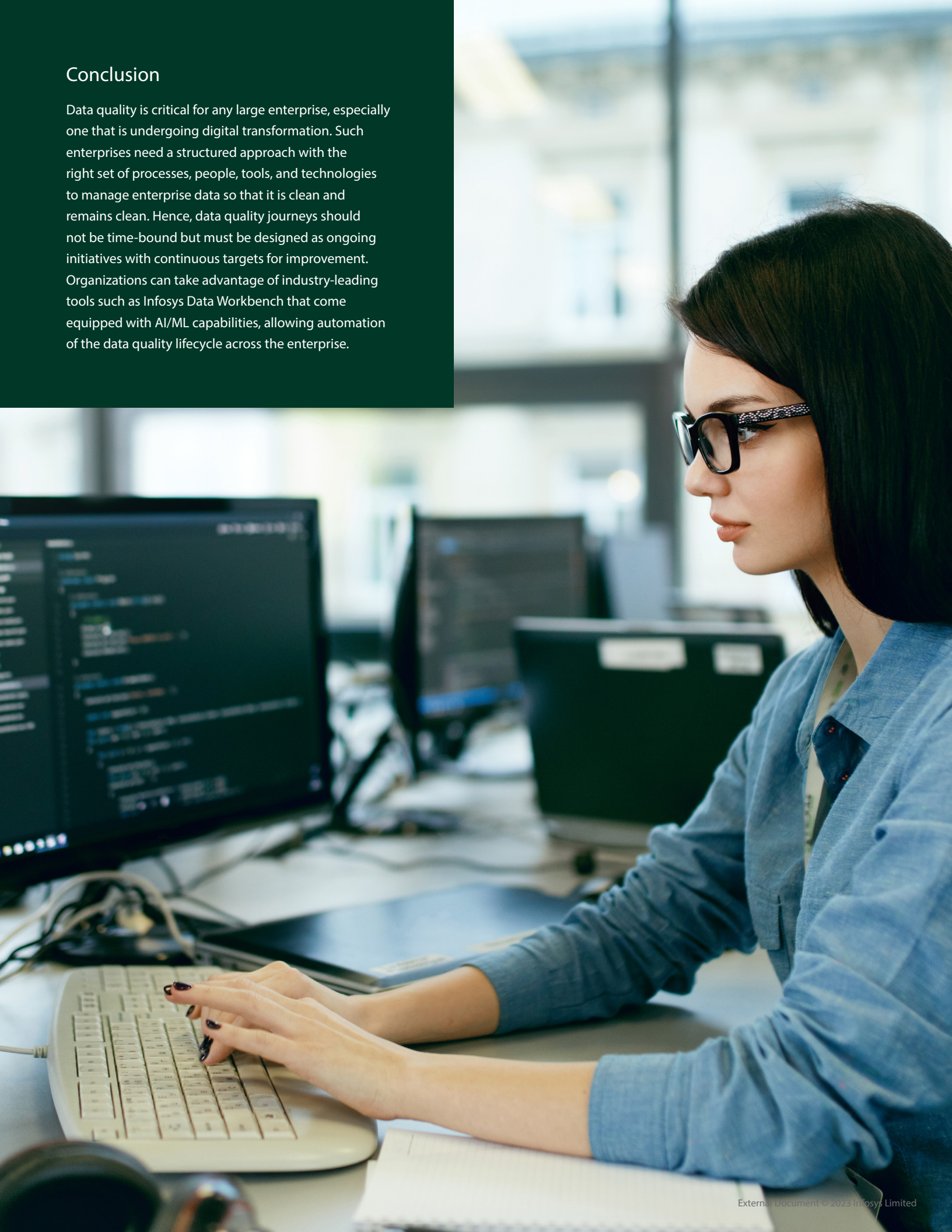
3. Leverage AI/ML technologies

AI/ML capabilities can help organizations enhance data quality in the following ways:

- **Uncovering data quality problems:** Advancements in AI/ML have changed the way organizations approach data quality problems. AI/ML capabilities help discover patterns in data quality issues and can predict where the data can go wrong. ML can also help with auto-matching and merging of data. These proactive steps ensure that data stays clean in the long run.
- **Auto correction:** Here, AI/ML provides suggestions to correct data and can even make the corrections automatically, without human intervention. It can autofill values in the data based on history. This helps improve productivity and reduces the total cost of quality around data maintenance.
- **Auto classification:** Auto classification of data refers to classifying items into primary or secondary categories. It also automates the customer classification into industry classification codes and segments customers based on attributes.
- **Identifying sensitive data:** AI/ML can help identify and mask sensitive data in large datasets. It ensures that the data complies with various privacy requirements.

Conclusion

Data quality is critical for any large enterprise, especially one that is undergoing digital transformation. Such enterprises need a structured approach with the right set of processes, people, tools, and technologies to manage enterprise data so that it is clean and remains clean. Hence, data quality journeys should not be time-bound but must be designed as ongoing initiatives with continuous targets for improvement. Organizations can take advantage of industry-leading tools such as Infosys Data Workbench that come equipped with AI/ML capabilities, allowing automation of the data quality lifecycle across the enterprise.



About the Author

Somnath Majumdar
Senior Industry Principal with Infosys

He has over 22 years of experience in master data management (MDM) and supply chain transformation programs. Somnath has driven MDM strategy and deployments for global clients in the manufacturing, hi-tech, retail, and financial services industries. He is a thought leader in the data management and supply chain management space.

References

1. https://www.gartner.com/smarterwithgartner/how-to-improve-your-data-quality
2. https://firsteigen.com/blog/how-to-scale-your-data-quality-operations-with-ai-and-ml/
3. https://www.talend.com/resources/definitive-guide-data-quality/
4. Part 2: The cost of Poor Data Quality | LinkedIn

For more information, contact askus@infosys.com

Infosys®
Navigate your next