

A Guide to Data Quality Management: Leverage your Data Assets for Accurate Decision-Making

Access Cleaner Data for Better Reporting



Summary

Data is the lifeblood of an organization and forms the basis for many critical business decisions. However, organizations should have an actionable data management process in place to ensure the viability of data, as accurate data can be the difference between positive and negative outcomes. To capitalize on the explosive data growth, businesses need to employ a data quality framework to ensure actionable insights extracted from the source data are accurate and reliable.

This eBook will give you extensive insights into data quality management and its benefits for your enterprise data processes. Further, it delves into the causes of 'bad' data and how high-quality, accurate data can improve decision-making. Lastly, the eBook helps businesses users select an enterprise-grade integration tool that can simplify data quality management tasks for business users.

Table of Contents

INTRODUCTION TO DATA QUALITY MANAGEMENT..... 04

What is Data Quality Management? 05

UNDERSTANDING DATA: AN ORGANIZATION'S MOST VALUABLE ASSET..... 06

Why do Data Quality Issues Occur? 07

Benefits of Using High-Quality Data for Businesses 08

DATA PROFILING: THE BASICS OF UNCOVERING ACTIONABLE INSIGHTS..... 10

What is Data Profiling? 11

Where is Data Profiling Used? 11

Why Do You Need Data Profiling? 12

Challenges Associated with Data Profiling 12

HOW TO HARNESS DATA: ASSESSMENT METRICS..... 13

Measuring Data Quality 14

WHAT FEATURES TO LOOK FOR IN A DATA QUALITY MANAGEMENT TOOL?..... 15

Selecting the Right Data Quality Management Tool 16

Creating a Centralized DQM Strategy 19

An Efficient Data Quality Management Solution: Astera Centerprise 21

CONCLUSION 22

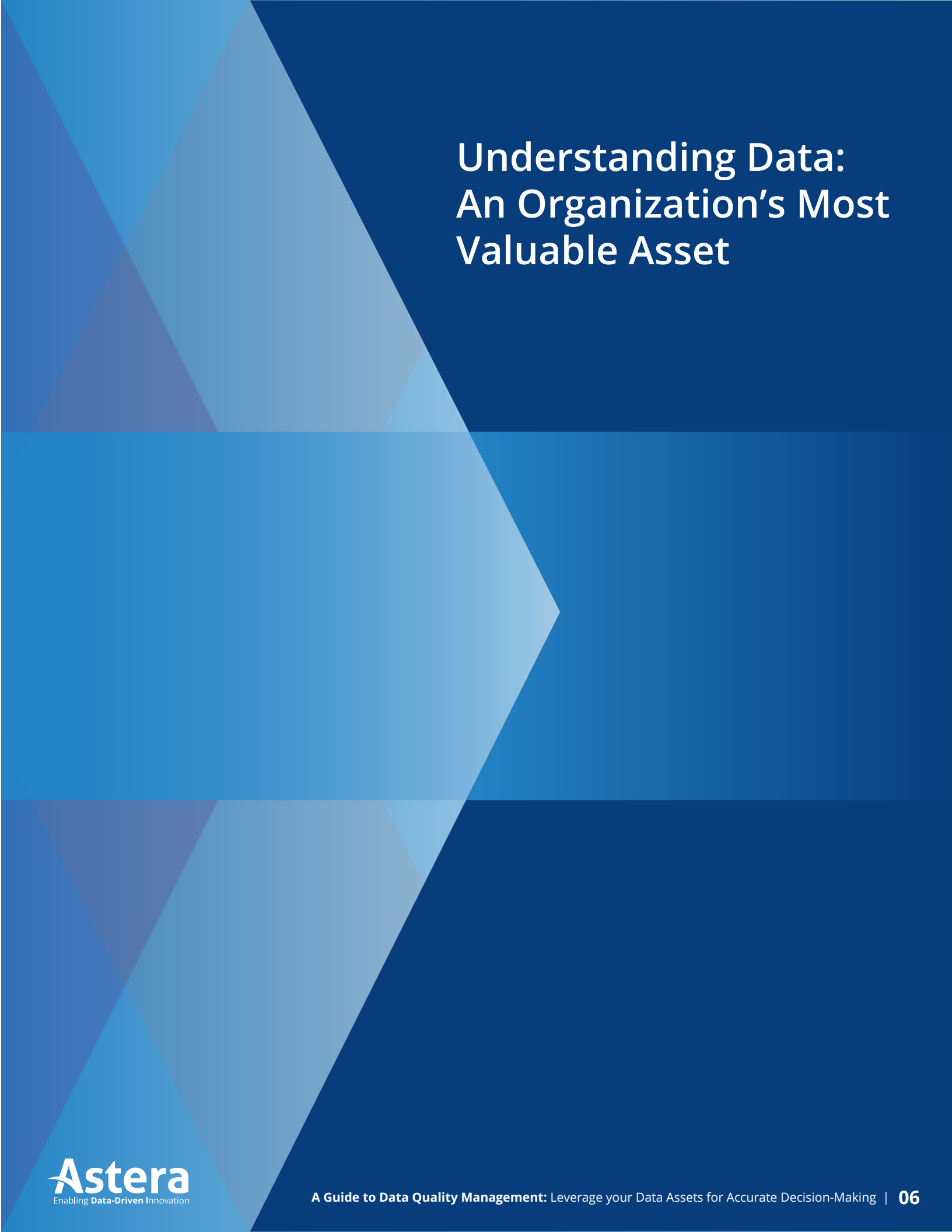
Introduction to Data Quality Management

What is Data Quality Management?

Data quality management (DQM) refers to the set of business practices that involves employing the right people, processes, and technologies to derive actionable insights from the available information. A well-established DQM framework ensures that data quality is maintained throughout the data lifecycle.

As part of a DQM plan, users specify certain data quality checks throughout the data journey to eliminate any inconsistencies or errors and ensure reliable data for analytics and business intelligence processes.





Understanding Data: An Organization's Most Valuable Asset

Why do Data Quality Issues Occur?

Research shows that 40 percent of business initiatives fail to achieve their targets due to poor data quality. Hence, it is critical for data stewards to identify the root causes of bad data quality, and build a robust data profiling and validation plan to improve the accuracy of information used for decision-making.

According to the 451 Research, the top three reasons for poor data quality include:

❖ Manual Data Entry

Many organizations rely on their employees to manually feed data into business systems, which may lead to errors due to lack of expertise, human error, or monotonous nature of work. Other common consequences of manual data keying include duplicate records and missing information.

❖ Data Migration and Conversion Projects

Data migration projects involve the transfer of data between different types of file formats, databases, and storage systems, which can often lead to duplication or missing records. Moreover, migrating from a legacy information system to a new one often involves converting data into a compatible format which, if not done correctly, can result in poor data quality.

❖ Entries by Multiple Users

In many departments, multiple employees are involved in the process of handling and modifying data. This can cause discrepancies, such as different names for the same supplier. For instance, some employees might enter the supplier's name as 'Dell', while others might use 'Dell Inc.' for the same vendor.

The Real Cost of Bad Data



Benefits of Using High-Quality Data for Businesses

High-quality data has the potential to improve business operations, and make them more efficient and profitable. A few benefits of ensuring data quality at every step of the business process are:

❖ Helps Identify New Opportunities and Improve Business Outcomes

Business decisions based on quality data are more likely to result in positive outcomes, as managers have an accurate, up-to-date, and complete picture of critical data assets. Moreover, high-quality data helps managers identify and leverage new opportunities, enabling the business to grow and stay competitive.

For example, incorrect financial information, such as overstated profits may result in misleading financial ratios, which are often used for evaluating a company's past performance. This analysis should be based on accurate and trusted data, as it lays the foundation for many important decisions, such as potential target markets and price changes. Similarly, updated financials can help the company decide which market segments are more profitable so that managers can explore new growth opportunities in those areas.

❖ Aids Successful Data Migrations

Poor data quality is one of the reasons why migration projects fail, as validating high volume data during transfer is often a challenge. To ensure a high success rate, data quality rules should be used to identify and correct any errors before the migration can take place. This helps in carrying out data migration projects faster and with greater accuracy.

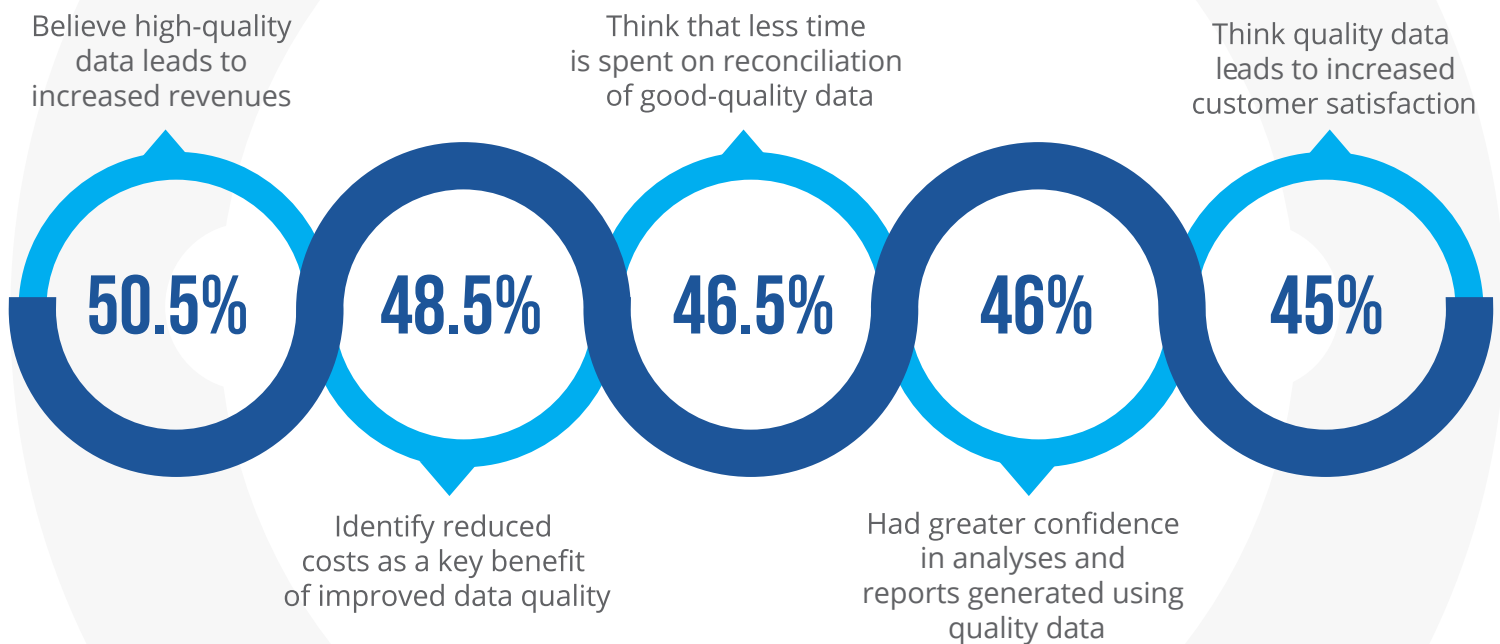
For instance, to create a unified repository for customer data, a company plans to move from a decentralized information storage system to a centralized one, such as a data warehouse. Previously, data was manually entered by employees and had errors including duplicate records and missing information. An effective data quality tool can help the business identify those errors and correct them before migrating data into a data warehouse.

❖ Ensures Data Quality Reduces Data Processing Time and Costs

According to Gartner, poor data quality can have an average financial impact of \$9.7 million per year. In addition, bad data means incorrect information is being processed which might involve rework. However, if companies make data governance a part of their overall business process, time and cost spent on rework can be minimized.

The Business Value of Good Data

IT decision makers and influencers identify the following as the top 5 benefits of data quality initiatives.





Data Profiling: Assessing Data Quality by Statistically Breaking Down Data

What is Data Profiling?

Data profiling offers critical insights into information that an organization can leverage to its benefit for decision-making and analysis.

It helps evaluate the data integrity by presenting a complete breakdown of its statistical characteristics, such as error count, warning count, duplicate percentage, and minimum and maximum value, enabling detailed data inspection.

This information assists users in identifying quality issues, risks, and overall trends. Data profiling tools use analytical algorithms to help scrutinize the data to determine its validity. These tools play a vital role in helping businesses streamline their data strategy with the company's principles and objectives.

Where is Data Profiling Used?

Generally, data profiling is used in the following processes:

❖ Data Migration

Data migration involves moving a high volume of information across heterogeneous systems, such as files, databases, etc. However, before initiating the transfer, it is essential to profile the data to identify discrepancies and resolve them to maintain consistency between the old and new systems.

Data profiling at an initial stage of migration can reduce the risk of errors, duplications, and incorrect information.

❖ Data Integration

Data integration creates a holistic view of enterprise data by merging it from disparate sources. Profiling data in the initial phase of integration ensures that there are no errors when source data is integrated and loaded into a data warehouse, data hub, or data mart.

❖ Data Cleansing

Data cleansing, a primary step in the data preparation process, helps with error rectification and deduplication to authenticate the validity and relevance of data. However, data cleansing is only beneficial for data sets you know are corrupt.

Often, poor quality data loiters in the system unnoticed and unaddressed until it is identified via data profiling.

Thus, data profiling methodically examines huge amounts of data to identify incorrect fields, null values, and other statistical irregularities that might affect data processes.

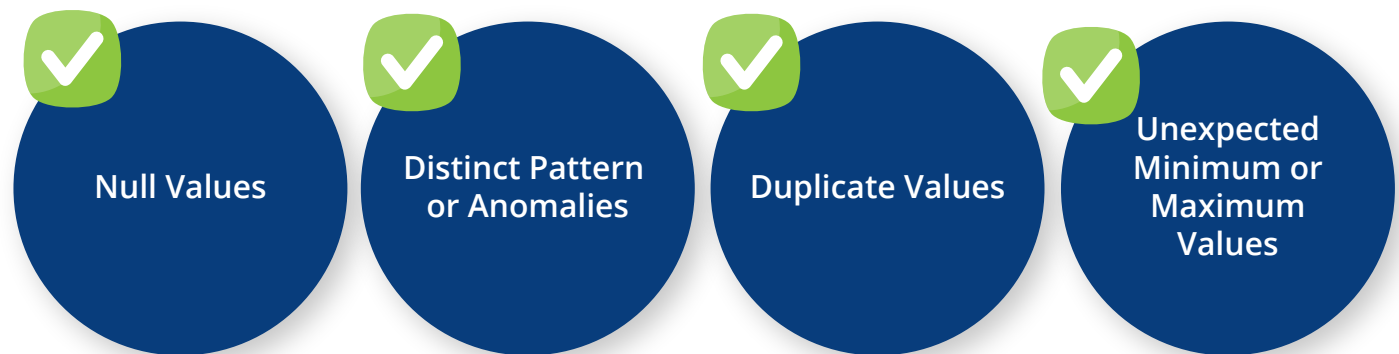
Why Do You Need Data Profiling?

Data profiling is critical to the validity of data processes as it helps you answer the following questions regarding your data:

- Does the data contain any null or blank values?
- Are there any anomalies in the data? Do they have a distinct pattern?
- Does it contain any duplicate values? What is the ratio of unique values?
- What is the range of values in the source data? Are the minimum and maximum values within your expected range?

Getting the answer to these questions can help you maintain the quality of your enterprise data and eradicate errors that can negatively influence the business processes.

You Need A Data Profiling Tool If Your Data Contains:



Challenges Associated with Data Profiling

Data profiling becomes challenging when you are dealing with large data volumes. To tackle this challenge, it is recommended to divide the data into segments and profile smaller datasets at a time.

Opting for manual data profiling presents a different set of challenges and won't be possible without the help of a professional, as it involves performing frequent queries to obtain essential insights about your data.

This is a more resource-intensive method. Moreover, chances are that you will be able to check just a subsection of your overall data as it might be time-consuming to manually profile the complete data set.

How to Harness Data: Assessment Metrics

Measuring Data Quality

Having a well-defined set of assessment metrics in place is vital for evaluating the performance of an enterprise's data quality management initiatives. It helps determine whether the data quality management strategy is bearing fruit to meet organizational goals.



Dimensions of Data Quality

Some key dimensions of data quality include:

- **Completeness** indicates whether the data gathered is enough to draw conclusions. This can be assessed by ensuring that there is no missing information in any data set.
- **Consistency** ensures that data across all the systems in an organization is synchronized and reflects the same information. An example of consistent data includes recording the shipment date in the same date format as in a customer's information spreadsheet.
- **Accuracy** implies whether the data that has been collected accurately represents what it should. This can be measured against source data and validated against user-defined business rules.
- **Timeliness** means that the data is available as and when expected to facilitate data-driven decision making. Many businesses are leveraging tools that support real-time data integration to gain up-to-date business knowledge.
- **Uniqueness** involves making sure that there are no duplicates present in the data. For example, the lack of unique data can result in multiple emails being sent to a single customer due to duplicate records.
- **Validity** measures whether the data meets the standards or criteria set by the business user. For instance, a business can place a data quality check on the order quantity field, i.e., 'Order Quantity >= 0' as negative order quantity implies invalid information.

Essential Features to Look for in a Data Quality Management Tool

Selecting the Right Data Quality Management Tool

Data drives decision-making, and therefore, managing data quality has become a top priority for businesses. However, managing data quality manually can be error-prone and time-consuming due to increased data volumes and disparity. This is where DQM tools come into play.

Here are some important factors that businesses should consider when selecting the right data quality tool:

❖ Data Profiling

An effective data quality tool should include data profiling features, which can automate the identification of metadata and provide clear visibility into the source data to identify any discrepancies.

Field Level Data Profile - Destination

Click On Field To See Profile

Record (Document)

Profile

Address (String)

CompanyName (String)

ContactName (String)

ContactTitle (String)

CustomerID (String)

Email_Address (String)

Phone (String)

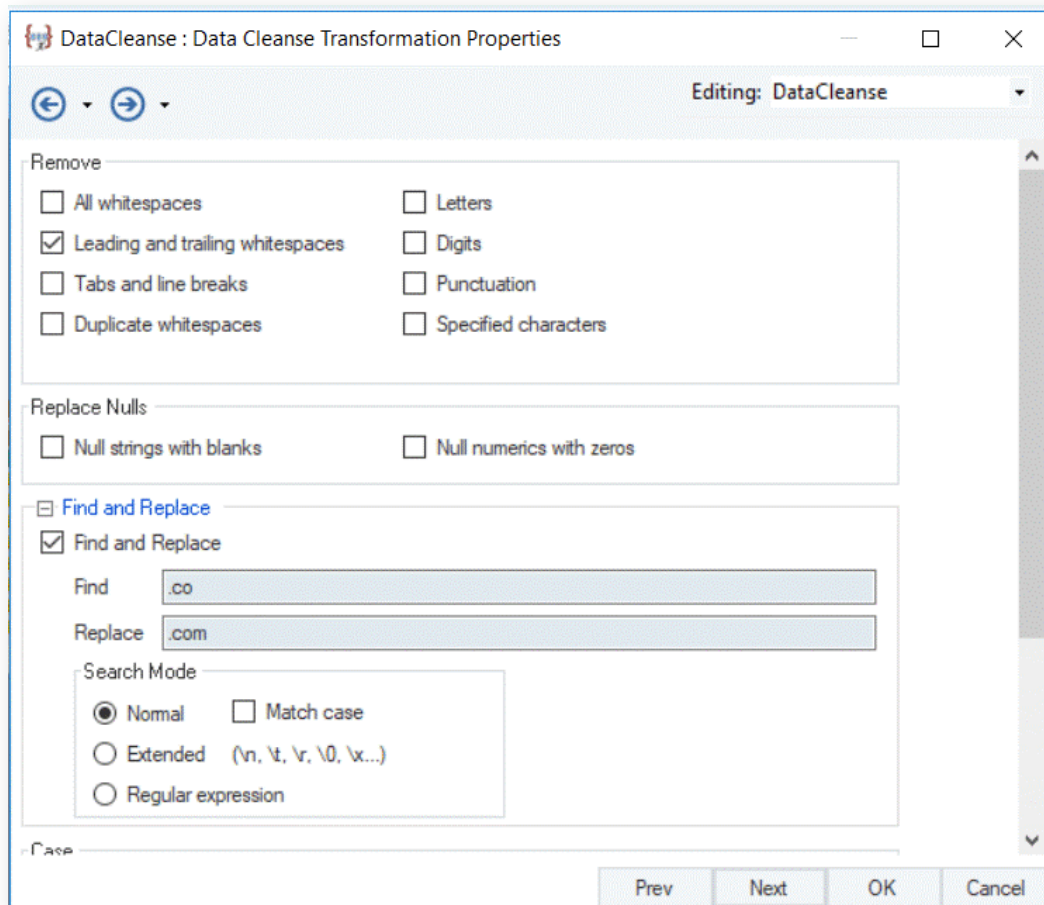
Field : Document.Profile.Address

Description	Value
Field	Address
Data Type	String
Null Count	6
Null %	6.593%
Error Count	0
Error %	0.000%
Warning Count	0
Warning %	0.000%
Minimum Value	1 rue Alsace-Lorraine
Maximum Value	Walsenweg 21
Blank Count	0
Blank %	0.000%
Minimum String Length	11
Maximum String Length	46

❖ Profiling source data

❖ Data Cleansing

Data cleansing remove errors, redundancies, and duplicates from raw data to retain accuracy before it is loaded onto a destination.



❖ Applying conditions to cleanse data

❖ Data Quality Checks

Data quality checks are objects or rules that can be integrated into the information flow for monitoring and reporting any errors that may occur while processing data. They ensure that the data being processed is validated based on defined business rules to ensure data integrity.

Expression	Description	Active	Warning	Error Message
IsMatchRegEx("[\\w\\d]+\\@[\\w\\d]+...	Contains	☒	☐	

Description : ☒ Active ☒ Is Error
 Attach rule to the field:
 Show Message :
 Expression:

Prev Next OK Cancel

Adding condition to flag incorrect records

❖ Data Lineage Management

Data lineage management helps control and analyze the flow of information by describing the data origin and its journey, such as the steps at which the data was transformed or written to the destination.

❖ Connectivity to Multiple Data Sources

With the increasing variety and number of data sources, it has become crucial to assess and validate internal and external data sets. Businesses should select DQM tools that offer support for data in any format and complexity, whether structured or unstructured, flat or hierarchical, legacy or modern.



Creating a Centralized DQM Strategy

Ensuring data quality is an on-going process, which evolves with the changing needs of the organization. This means organizations must have a centralized DQM strategy with a robust framework to address the data quality challenges and reap the benefits of high-quality data.

The steps for creating a centralized data quality management strategy include:

❖ Define the key success objectives for the data quality program

This involves defining the data quality metrics, such as data-to-errors ratio and percentage of blank records. This provides users a clear understanding of the data that is being analyzed and the dimensions, including completeness, uniqueness, accuracy, etc. that will be used to assess data integrity.

❖ Communicate the DQM plan organization-wide

Ensuring data quality is the responsibility of all information stakeholders, including data architects, business analysts, and IT. Hence, employees should know the expected data quality levels, business benefits of the set data quality standards, and assessment metrics for smooth implementation of the DQM strategy.

❖ Assess incoming business data against the set data quality metrics

Ensuring data quality is easier with an advanced data quality tool as it enables users to define data quality rules and assess incoming data based on the predefined criteria.

❖ Analyze data quality results and identify the root causes of bad data

Once the data has been processed, users can evaluate the quality of data and identify the reasons for flagged records. For instance, the screenshot below shows that one of the records was erroneous because of the incorrect email address.

Object Path	Serial_No	Order_ID	Customer_Name	Email_Address	Phone_Number	Address	City	Country	Company_Name	Industry	Product	Invoice_Number
DataQualityRules	1	10643	Maria Anders	mariaandersalfreds Futterki	030-0074321	Oberer Str. 57	Berlin	Germany	Alfreds Futterkiste	Education	Chartreuse verte	106438251997
DataQualityRules	2	10308	Ana Trujillo	anatruijillo@anatruijilloemp	(5) 555-4729	Avda. de la Const	México D.F.	Mexico	Ana Trujillo Empa	Education	Gudbrandsdalsos	103089181996
DataQualityRules	3	10558	Thomas Hardy	thomashardy@aroundtheh	(171) 555-7788	120 Hanover Sq.	London	UK	Around the Horn	Education	Perth Pasties	10558641997
DataQualityRules	4	10566	Frédérique Citeau	frédériqueciteaux@blonde	88.60.15.31	24, place Kléber	Strasbourg	France	Blondesdél père	Trading	Lakkalikööri	105666121997
DataQualityRules	5	10663	Laurence Leblan	laurenceleblan@bonapp'	91.24.45.40	12, rue des Bouc	Marseille	France	Bon app'	Trading	Singaporean Hok	106639101997
DataQualityRules	6	10742	Elizabeth Lincoln	elizabethlincoln@bottom-d	(604) 555-4729	23 Tsawassen Bl	Tsawassen	Canada	Bottom-Dollar Mar	Trading	Camembert Pierr	107421114199
DataQualityRules	7	11023	Victoria Ashworth	victoriaashworth'sbevera	(171) 555-1212	Fauntleroy Circus	London		B's Beverages	Trading	Uncle Bob's Orga	110234141998
DataQualityRules	8	10254	Yang Wang	yangwang@chop-suey.chi	0452-076545	Hauptstr. 29	Bern	Switzerland	Chop-suey Chine	Service	Guaraná Fantásti	102547111996
Validation rule [Email address does not contain @] failed.			Jdro Afonso	pedroafonso@comerciomi	(11) 555-7647	Av. dos Lusíadas,	Sao Paulo	Brazil	Comércio Mineiro	Service	Siroop d'érable	110424221998
DataQualityRules	10	11042	Pedro Afonso	pedroafonso@comerciomi	(11) 555-7647	Av. dos Lusíadas,	Sao Paulo	Brazil	Comércio Mineiro	Service	Siroop d'érable	110424221998

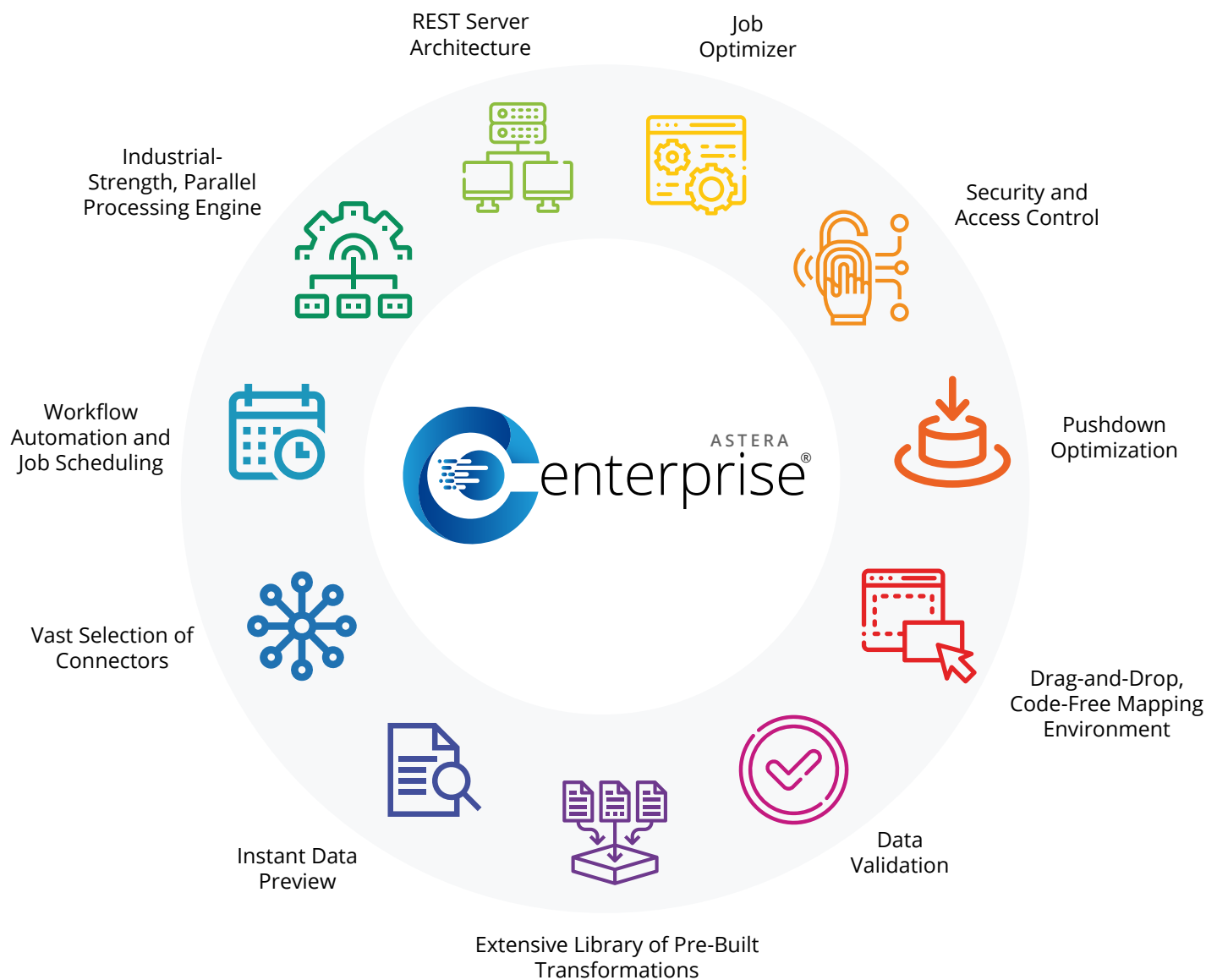
❖ Flagged Records

❖ Monitor and adjust the data quality workflows based on changing data needs

Users must verify the data quality workflows at periodic intervals to ensure that the data quality rules are in sync with the overall business goals. This also includes taking necessary actions to improve data quality standards based on prior results.

An Efficient Data Quality Management Solution: Astera Centreprise

Astera Centreprise is an end-to-end data integration solution that enables businesses to accomplish complex data integration jobs while ensuring data quality. The advanced data profiling and data quality capabilities allow users to measure the integrity of critical business data, speeding up data integration projects in an agile, code-free environment.



Conclusion

The volume of business data worldwide doubles every 1.2 years. This exponential growth in data volume makes it imperative for organizations to establish data quality checks at every step to reap significant benefits. Data drives decision-making, and therefore, managing data quality has become a top priority for businesses.

C-level executives are realizing the inherent value of data and leveraging its benefits to stay competitive. However, if not identified and validated at an early stage, bad data can adversely impact the health of an organization in the form of operational inefficiencies, financial losses, and missed opportunities.

Smart solutions, such as Astera Centerprise, are integral for ensuring reliability and accuracy of data. It combines data quality, profiling, cleansing, and integration in a single, easy-to-use environment to facilitate the transformation, movement, and consolidation of data.