View Point

Data Quality

MAVERIC
Transforming Assurance

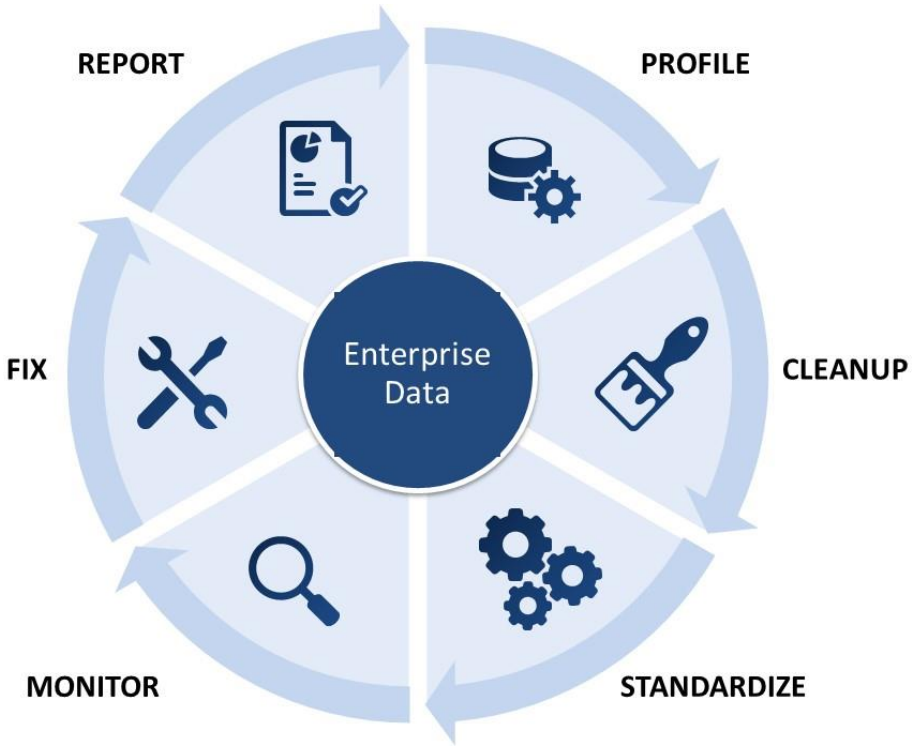
Overview

Quality of data for business operations is considered to be a critical component of enterprise success. With the exponential rise in ways and means by which data is generated and consumed, organizations are more and more focusing on ensuring data quality. Studies indicate that fewer than 50% of IT decision makers have confidence in their organization's data quality initiatives, although more than 90% acknowledge the growing importance and volumes of data that they have to grapple with in future.

This whitepaper aims to throw light on the importance of data quality and means to assure quality of data, focusing on key facets and tools. Ensuring data quality in today's environment of Big data is not only necessary but also critical to an organization's survival.

Data Quality Framework

A robust framework that can address data quality needs to look at the following 6 focus areas.



Profile: Analyze the data to ensure conformance to data quality dimensions outlined below

- **Completeness**: Field level (e.g., missing data) or data level (e.g., missing line item for an order) check to ensure all requisite details are available. This will ensure that the key fields are populated with relevant data
- **Conformity**: Validation to ensure whether the data conforms to the required formats i.e. whether values are correlating with the attributes and structure of the target field
- **Consistency**: Check to ensure values in one data set are consistent with the same values in another data set. This can be within table (e.g., inconsistency between Title and Gender columns) or across tables (e.g., aggregates and parent/child relationship)
- **Accuracy**: Aim to ensure the data represents its intended purpose/requirement. Data inaccuracy may happen due to typo, using acronyms or untimely arrival of data
- **Duplication**: Check carried out at field level or record level to ensure same data set is not stored more than once within a table

Cleanup: Revamp existing data to improve data quality by using the following methods:

- **Lookup**: Identify the clean data set and use it as lookup to cleanup the data. This may be within a table (e.g., Gender column having values like M, Male, Mr, etc.) or by looking up value in another table (master table or a manually created lookup table). We may also use this method for cleaning up missing values (e.g., populate Gender column based on Title or vice versa)
- **Auto/Manual:** Create scripts based on requirements to fix the data issues. However, certain data issues cannot be fixed through any of the methods e.g., mandatory columns are missing, but that data is stored in a free text column, say 'Comments' column

Standardize: Regulate the data to improve accuracy through the following:

- **Geo-coding**: Adopt various measures to ensure conformance to geographical addresses like standardizing address columns to meet postal rules (e.g., USPS for US addresses) and getting and storing the Latitude & Longitude of the physical location using the standardized address
- **Matching/Linking**: At times, we may be able to get all details about a field from multiple data sources or notice duplicate data that may not be exactly matching but closely relating to an existing value. To address these issues and create accurate data, we need to create business logic that enables matching of such data and removing anomalies

Monitor: Establish systems or business processes to track quality of incoming data. Some of these tracking mechanisms include the following.

- **Business rules**: Create quality checks based on business rules (e.g., daily transaction limit cannot exceed a specific amount per account) and execute the checks on a periodic basis in line with requirements and raise alert when a check fails
- **Trending/Threshold:** Analyze data and identify the trend of data (e.g., customer creation on a daily basis should be between 'x' and 'y', sum of all loan amounts disbursed on a day should be between 'x' and 'y'). Quality checks can then be created using the analysis and run periodically based on requirement. The boundary values for analyzing the trends also need to be changed periodically based on the patterns observed

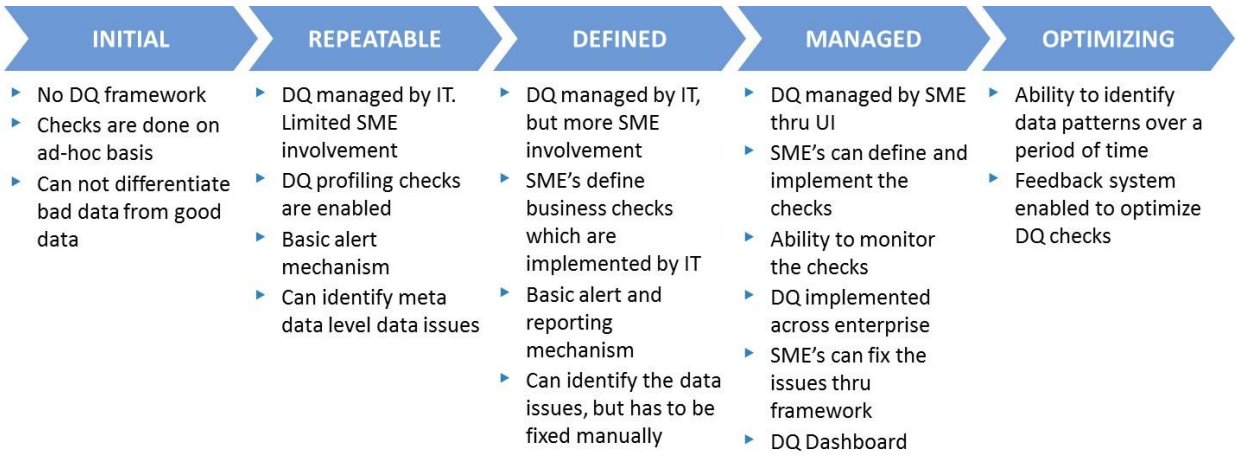
Fix: Clean up existing data to improve quality. This needs to be carried out in a manual way by analyzing for each data quality fail alert, the data set and identifying the root cause (it can be missing data, data duplication, data corruption, business logic not implemented correctly, etc.). Further, it also needs to be checked upfront if the alert that has been given is fake or requires attention.

Report: Build a reporting system to have a better understanding of the quality of data using the following means:

- **Dashboard**: Provides details of the quality checks that are being run, execution outcome details, and issues identified and resolved, etc. in addition to the current status, they also provide information on trends over a period of time
- **Email**: Email based alert system should be established to provide the results of each data quality check execution. The system should have the ability to be configured at various levels

Data Quality Maturity Model & Tools

Using the data quality framework, organizations can move towards implementing and executing initiatives that have a profound effect on the quality of data. However, organizations need to be cognizant of the state of data quality maturity they are in currently and take steps to move towards higher maturity levels. The maturity model outlined below provides a snap shot of the evolution of various data quality maturity initiatives.



An organization needs to move towards the final stage of achieving the managed and optimizing level in terms of quality of data. A plethora of tools, manual approaches and customized solutions are available to help firms move from one maturity stage to another. These include profiling and checking tools like SQL scripting, UI and dashboard formats like HTML, CSS and PHP, and various third party end to end solutions like Informatica data quality, IBM quality stage, Talend data quality and Open source data quality.

Conclusion

Ascertaining and ensuring the quality of data is paramount for business success. With personalized customer outreach mechanisms increasing in prominence, dependence on data to understand customer preferences and buying behavior has become more important than ever. Add to it, exponential increase in the amount of data generated across multiple devices and access mechanisms. The importance of data quality cannot be more emphasized. The framework and maturity model provided in this article will only serve as a mode to move towards being a more data quality intensive organization. Understanding the importance of data quality, analyzing the level of data quality maturity currently in and taking steps to move towards the highest level are all dependent on the individual organization. Businesses who understand this and move on the path to execution are the ones to succeed in today's highly competitive landscape.