

Managing Data Quality

By Sandeep Savla & Mathew Ninan

Statistical techniques are recommended to continuously monitor the quality of data and establish control limits. Multiple checkpoints for data quality verifications must be established in the entire lifecycle.

Information in any database is only as good as the data that resides within it. Data quality is crucial to the success of Data Management (DM) as well as Business Intelligence (BI) initiatives. Quality of data can be said to be directly proportional to the information value. Unless data in systems is accurate, reliable and credible, more effort will be spent on manual activities, rework and stabilization of system than on business. For successful data quality management, therefore, the solution must include tools, techniques, processes and frameworks.

Data quality management lifecycle must be clearly defined using continuous as well as iterative frameworks. Better data quality enables accurate decision support systems and helps improve business response. Automated cleansing and monitoring of data issues ensures rapid improvement in data quality and also helps prevent human errors. Deploying

statistical techniques ensures continuous monitoring of data quality and provides additional controls to track data cleansing efforts.

DATA QUALITY NEEDS

A Data Warehouse (DW) or a data migration initiative cannot be expected to deliver a satisfactory Return on Investment (ROI) unless the data within the system is accurate, reliable, and credible. To deliver the level of data accuracy and reliability, the data quality control process must include – and excel in – data analysis, data cleansing, data standardization, and data validation processes.

Improving the quality of data is imperative for accurate business reporting, which is a regulatory requirement. Any loss of data in migration process is a loss of valuable information and affects the quality of data. Enterprises can neither afford to ignore cost

savings due to reduction in redundancy of data and elimination of duplicate information nor the cost of wrong decisions based on incorrect data.

What enterprises require is a strong enforcement of business rules on data and user-friendly application environment, which makes accurate lookups for data entry operators.

COMMON ERRORS

Efficient data management is imperative to ensure data quality. Following are a few of the common errors that IT organizations commit:

- Different order entry operators creating multiple customer master records in the

transactions to the general ledger.

- Any data quality initiative imposes additional challenges due to other process related errors such as absence of data quality tracking and monitoring routines, and multiple owners and unmonitored data input.

CONTINUOUS AND ITERATIVE FRAMEWORK

A continuous and iterative framework can help IT organizations control the data quality (Figure 1). The framework continuously monitors the quality of data and provides necessary inputs to users or systems to take timely actions to control

A continuous and iterative framework can help IT organizations control the data quality as it continuously monitors and provides necessary inputs to take timely corrective actions

system resulting in inability to report overall customer revenue without establishing the links between duplicate customers.

- Lack of common standards across various products renders tracking profitability of new products and their relevant impact on product mix.
- Missing data elements such as shipping date, and order date will result in inaccurate cycle time tracking.
- Lack of links between transaction elements such as account receivable, and general ledger may result in inaccurate financial reporting and inability to match

the quality. The framework is also iterative as quality improves iteratively over a period of time and learnings are translated to further detect and improve quality. The framework consists of four steps:

1. Detection: Detecting the data quality issues is the first step towards ensuring quality. Intelligent algorithms and tools are deployed to continuously monitor and report the quality of the data and flag any anomalies in the data elements as erroneous.

2. Correction: This step involves taking necessary actions to correct the data

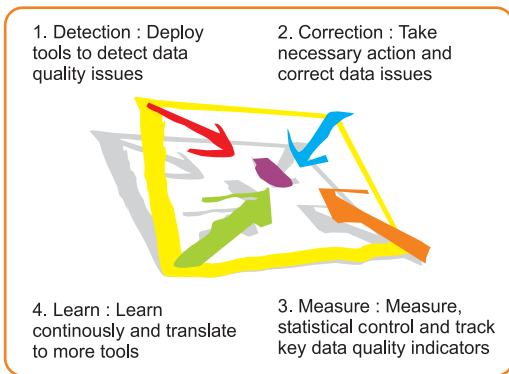


Figure 1: Effective results are achieved using continuous iterative framework

Source: Infosys Research

anomalies. While analytical algorithms are used to auto-correct some of the entries, user intervention is called for when there is more than one choice of action that can be taken.

3. Measurement: The measurement step involves measuring the data quality and tracking errors. It ensures that all reported anomalies are corrected and monitored. Key performance indicators such as average cycle rate for error fix, and total error data are captured, trended, and reported to management periodically.

Data errors need to be classified to have specific strategy based on classification. Type I error occur when relationship is assumed but in fact does not exist. This leads to additional validation and false alarms. Type II errors occur when no relationship is assumed when in fact it exists. Detecting Type II errors is difficult as they require in-depth data analysis. Defining confidence limits of the data errors helps in prioritizing the effort that needs to be spent to bring the data quality within control limits.

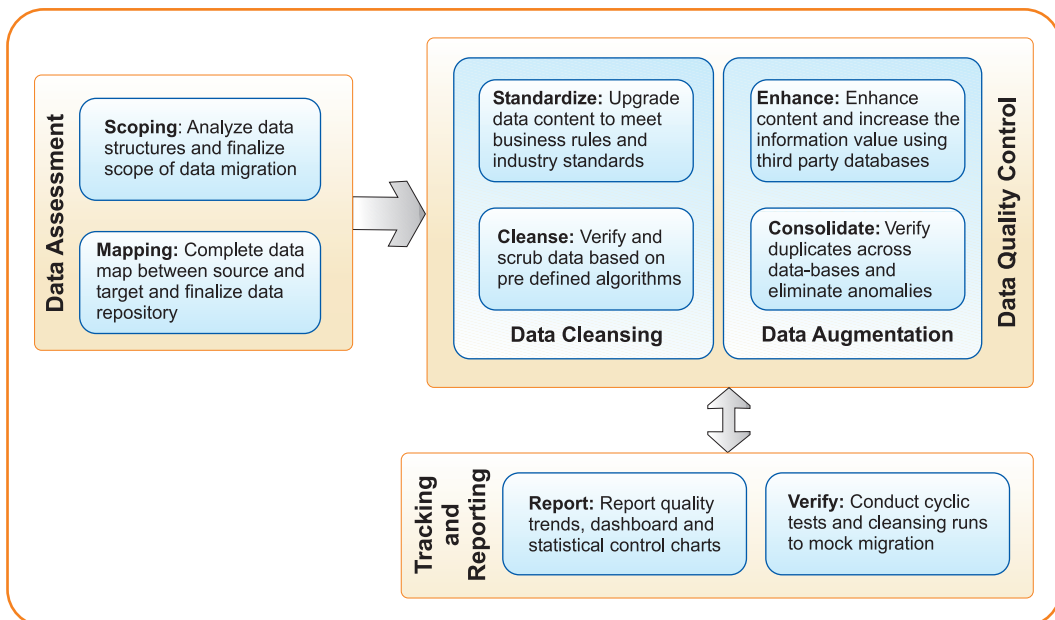


Figure 2: Data quality process involves assessment, quality control and reporting.

Source: Infosys Research

4. Learn: The step involves reviewing the process periodically and enhancing the ability to detect and correct anomalies by deploying additional tools and by reforming business processes. This step ensures that the effort required to maintain the data quality reduces over time.

The framework is scalable and can be extended for other business requirements for migration, business intelligence warehouses and so on. It also helps IT organizations reduce maintenance costs.

Mapping provides more visibility to the cleansing requirements. Cleansing reports can be created and prioritized on the entities, which are critical for migration (for example, customer entity must be thoroughly cleaned to prevent it from impacting dependent entities such as contracts, install base, and field service.

Data Quality Control: Data quality control phase focuses on correcting and standardizing the data and involves data cleansing and augmentation. Data is inspected for errors,

Detailed analysis of errors encountered during migration, statistical summarization of erred records, and tracking of success percentage of each conversion run help reduce chances of errors during subsequent runs

DATA QUALITY CONTROL PROCESS

Any data quality initiative needs well-defined processes to be followed for maximizing the control on quality. The data quality control process involves assessment of data quality issues, cleansing and augmentation of data, and tracking and reporting of data anomalies (Figure 2).

Data Assessment: Data assessment phase consists of analyzing the data structures and finalizing the scope of data quality audits. This is an important phase for data management as priorities of correction efforts is determined in this phase.

In data mapping or the profiling phase, end to end mapping between source system and destination system are carried out.

anomalies, duplications, and inadequacies. The phase involves detection of errors and correction of data to control the data integrity over time.

Standardize: The standardizing exercise is important to upgrade data content to meet business rules and industry standards. Data must be made consistent across systems, which in turn reduces redundancy. Data standardization can be done at the following two levels:

- Coding standardization involves product codes, financial codes, inventory numbers, model numbers, program types and so on. For example, all B/W printers from HP are coded starting with BWHP-PRN which makes product recognition easy.

- Address standardization Involves consistency in the order of various data fields. For example, “9575 N FARM ROAD 173” needs to be changed to “9575 N FARM RD 173”.

Cleanse: The data cleansing and preparation phase fulfills the following objectives:

- Ensure that the data meets the requirements of future state system
- Minimize migration related errors, thus reducing manual data entry effort.

Cleansing is carried out to ensure integrity of data and to prepare the data for specific migration needs. Tools and intelligent

commercial databases. Some of the constraints that can be enforced are:

- Referential constraints to verify if the key is present in referential master table before transaction record can be built
- Unique constraints to ensure that duplicates within an entity are avoided
- Default constraints to plug in default values in the absence of user entry in commercial databases

Business rules to allow custom scripting of business rules in commercial databases.

Enhance: The enhance phase involves data augmentation by enhancing the information

Frameworks and processes are imperative for controlling data quality and for managing data, the most important corporate asset

algorithms customized to business scenarios are implemented to detect the violations of the rules. Additional correction procedures are also deployed for rapid error fix. A few instances where cleansing is required are as follows:

- Inactive master records with no transactions and inspecting them for potential purges
- Transaction records with no masters
- Invoices with missing AR and vice versa.

Enforce: Commercial databases can also be leveraged for enforcing referential integrity and constraints. Additional business rules can be enforced using pre-defined triggers on

value using internal and external data sources. Industry standard databases such as Dun & Bradstreet and United States Postal Service (USPS) database are used to add information about customers and provide enhanced analytical capabilities for the sales and marketing functions. Data can also be enhanced by sourcing information from systems within the organization by matching the required contents across the applications. For example, various customers in database are linked together with information in external D&B database, which resulted in more accurate reporting of profitability for this customer.

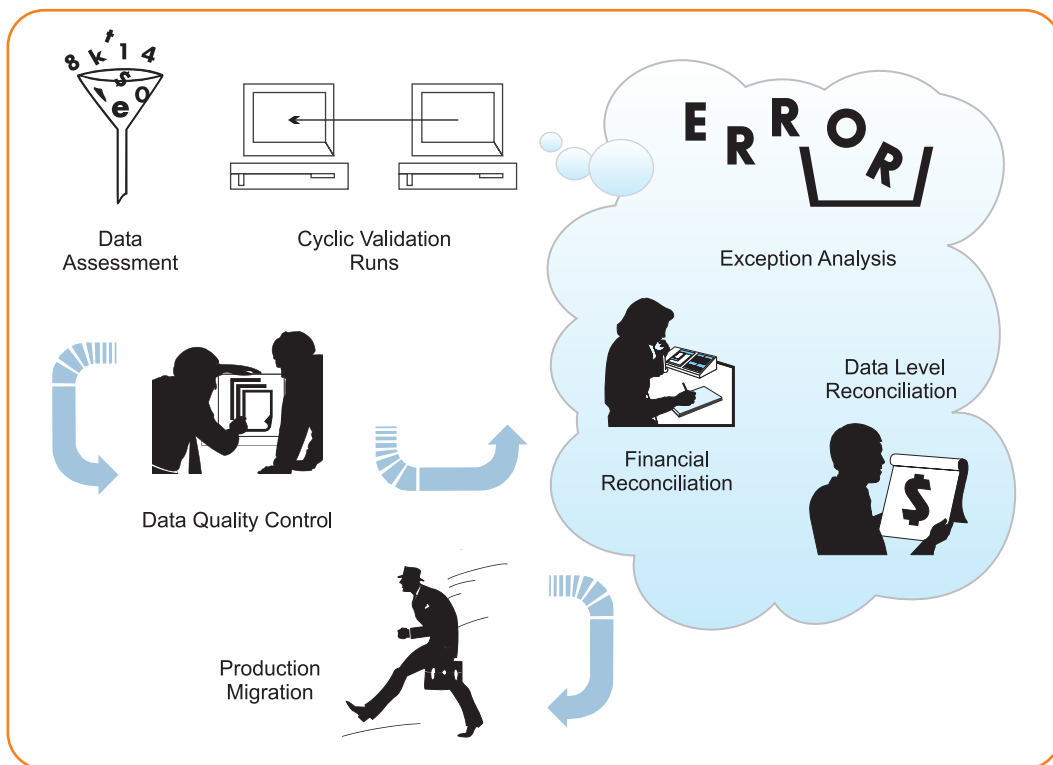


Figure 3: Migration involves data validation procedures which includes exception analysis and reconciliation **Source:** Infosys Research

Consolidate: The consolidate phase involves reducing duplicate or unnecessary information in data store. This ensures that transactions are assigned to correct dimensions making rolling up easier.

Tracking and Reporting: Monitoring and tracking the data cleansing exercise is important to understand readiness of data and also the progress. It enables users to know if the correction has been successful or not. It provides the ability to prioritize based on severity of data quality issues and trend line. The performance of data quality control teams can be assessed on the action taken on reported quality issues. Tracking also helps identify the most common causes of errors and help explore

the possibilities of automating correction procedures.

Data Quality Verification: Data quality verification focuses on additional efforts to verify the quality of data by performing live tests on the data. The data is tested in cyclic validation runs to test databases and analysis is carried out on the migration to pre-verify the quality of the data. This provides necessary information to the management about readiness of the data in terms of quality.

Exception Analysis: Data validation or error analysis plays an important role in data migration. It shows the business community the end result of data cleansing being carried out

and also gives visibility into the errors, their causes, and the corrective action that needs to be taken.

Detailed analysis of errors encountered during migration, statistical summarization of erred records, and tracking of success percentage of each conversion run help reduce chances of errors during subsequent runs and also proactively identify new areas of cleansing for achieving higher conversion rates.

Reconciliation: Reconciliation is a process through which data from both source and target systems are compared and analyzed. Scripts must be run to bring both source and target data into a common framework for comparison. Financial reconciliation, for example, specifically looks at matching revenue and expense related data from source and destination systems to ensure that revenue flow remains constant (Figure 4).

DATA MIGRATION SUCCESS STORY

The client is one of the world's leading providers of business communication products and services. The client, in late 90s, had taken the acquisitions route to growth. All the acquired companies were working on different IT systems and the challenge before the client was to bring them all into a common platform to address data integration issues. The company partnered with Infosys to implement Operational Data Store (ODS), creating necessary cleansing routines and preparing the data for migration to Oracle e-Business Suite.

Infosys' solution based on a strong data quality control framework included a data-cleansing suite with more than 200 reports to detect data errors, and a data control engine to track and record statistics and provide management dashboards. An address

standardization algorithm was deployed for standardizing address, and resolving county/taxing issues and more than 50 automated exception analysis routines were set up to analyze the errors in cyclic validation runs. Data assessment services, which included mapping between legacy and destination source system, were leveraged to maintain the quality of data.

Following are some of the best practices that were used to implement the migration exercise:

- a) Implementation of continuous and iterative framework
- b) Usage of USPS Postal database and D&B database for data augmentation.
- c) Automation of standard manual fixes
- d) Active participation in data quality control group meetings
- e) Single stop shop web service for all data management needs

The client was able to successfully migrate large-scale data for two North American regions. The solution helped attain six sigma limits for data quality: less than 3.4 defects in a million records in the database; more than 80 person months of effort saved by deploying automated error correction routines, and more than 55 percent improvement in the data migration success rate.

Successful data management at client location would not have been possible without presence of a strong foundation for data quality control and support from all the management groups involved in the initiative.

CONCLUSION

Strong frameworks and processes are imperative for controlling data quality and for managing data, the most important corporate asset. Intelligent algorithms help error detection

and correction and minimize the risks. Additional validation procedures such as exception analysis and data level reconciliation ensure high success rates in migration-related initiatives.

The challenges associated with data quality control initiatives can be effectively handled by implementing the recommended framework and processes to control data quality.

REFERENCES

1. Data Quality control in operational data store: Sandeep Savla, Infosys Data warehousing seminar series, February 2003.
2. Data Management: An Executive Briefing: A River Runs Through IT: George Marinos, DM Review, Jan 2005.
3. The Essential Ingredient: How Business Intelligence depends on data quality: Mat Hanrahan, May 2004.
4. Implementing Data Quality as a corporate service: Colin White, BI research, 2004.
5. Clean up your data: Eric Donohue, Tony Chang and Jon Bostwic, DM Direct newsletter, December 2004 