

View Point



Enterprise Data Quality

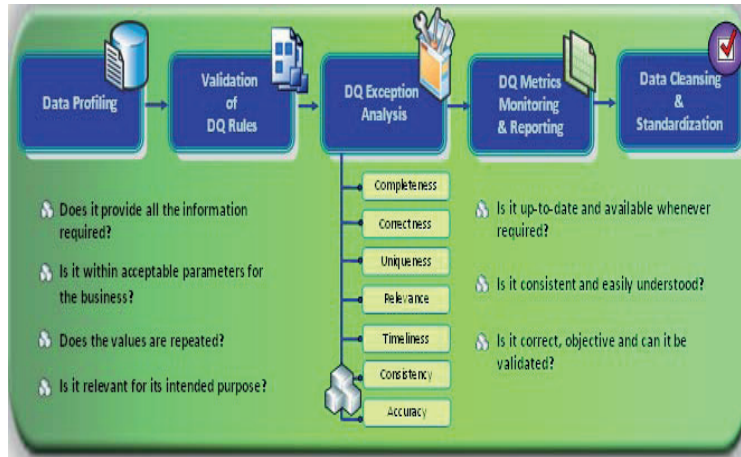
An Approach to Improve the Trust Factor of Operational Data

Sivaprakasam S.R.

Given the poor quality of data, Communication Service Providers (CSPs) face challenges of order fallout, billing disputes, revenue leakage and customer dissatisfaction. Closely monitoring and filtering both technical and business exceptional data from the voluminous data maintained in heterogeneous applications/a database is a huge task. It becomes imperative to formulate a strategy for continuous monitoring of enterprise data and develop a dynamic, generic Data Quality (DQ) framework to replace manual analysis. To enable such quality analysis, data must be consolidated, cleansed and integrated in a single instance. Our white paper offers a step-by-step approach to improve data quality by focusing on building trust into data.

Executive Summary

In today's competitive market, enterprises are spending most of their time reconciling and validating business analysis since the underlying data originates from disparate source systems. The trust factor in such metrics is very low due to inconsistency in data. To improve the trust factor of business analysis, CSPs are compelled to continuously monitor the quality of their data, and improve its completeness, correctness, timeliness, accuracy, uniqueness, integrity and relevance.



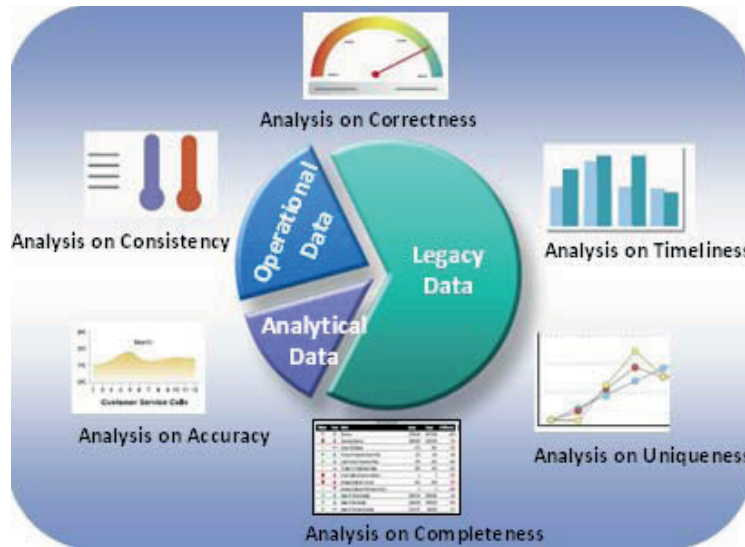
The business goals of enterprise data quality are:

- Increase efficiency and effectiveness of business processes for better decision making
- Reduce uncertainty and risk in missing SLAs
- Improve corporate performance and operational efficiency
- Improve competitive advantage
- Reduce costs of iterations and rework
- Maximize ROI from enterprise applications
- Provide a measurable and accurate view of data for better customer service

Need for Data Quality Strategy

The communications industry is evolving horizontally in fixed-line, mobile, ISP, broadband and satellite carriers; and vertically in customer base, range of products, services and price plans. Enterprises invest hugely in their operational and analytical applications to meet growing business needs. The performance and productivity of applications is dependent on the quality of data processed by it. Hence, enterprises should have detailed insights into data such as business exceptions, consistencies, etc. They must approach these issues strategically. CSPs must formulate a data quality strategy that involves data stewards, subject matter experts, the data governance team and decision makers. There must be a step-by-step approach for profiling enterprise data, framing the quality rule, validating data, analyzing exceptions, cleansing data and standardization, ongoing monitoring and data quality reporting, as shown in the diagram below.

A data quality strategy is required to implement organization-wide initiatives such as increasing revenue by customer cross-selling or up-selling, increasing SLA adherence, etc. Strategic goals need a single view of enterprise data that is enabled by integration, cleansing, de-duplication and standardization. Data quality rules, defined as part of the data quality strategy, will distinguish between valid and invalid data; cleanse and standardize invalid data; manage and measure the trust factor and report quality metrics.



Challenges in Data Quality Assurance

An established telecom player that manages terabytes of data volume with several legacy operational applications may be challenged by:

- Lack of unique data quality validations across the enterprise
- Unavailability of source data owner and formal data quality strategy
- No active participation from business users
- Lack of cross-application expertise and data governance team
- Frequent mergers and acquisitions without strategic goals
- Lack of source data documentation and resulting incorrect table relationships and dependencies
- Lack of source code expertise for legacy applications
- Lack of advanced and automated source data profiling
- Data duplication and silos
- Lack of validation during data entry

The inconsistent and incomplete customer details being maintained by the telecommunication service provider leads to a loss in its monthly revenue.

Data Quality Process – Not an Event

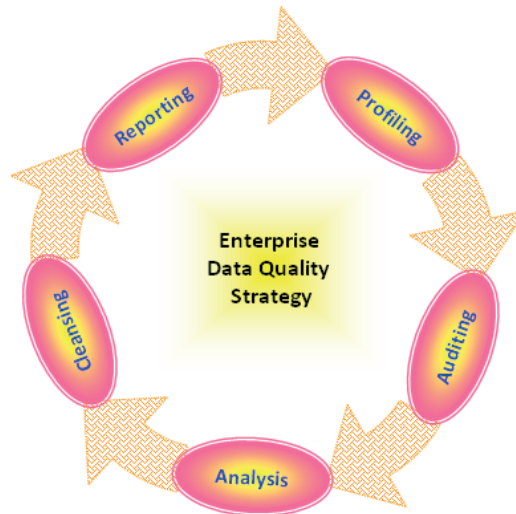
To overcome the issues of poor quality of data, data duplication, etc., CSPs are centralizing their legacy applications and data silos. As a result of this strategic move, Common Data Model (CDM) and data integration processes have emerged.

While integrating data from multiple and heterogeneous source systems, data quality validations are enforced to filter exceptional data and maximize the trust factor of the CDM. Data integration being a continuous process, it is essential to impose quality monitoring, analysis, correction and reporting whenever the CDM is refreshed. This will improve data quality in a holistic, cross-functional way.

In addition, the following enterprise-wide activities should be incorporated and improved over time:

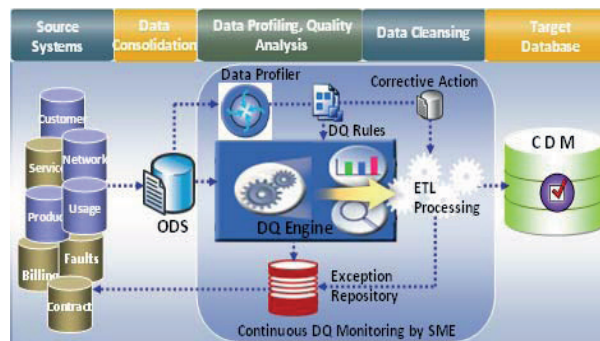
- A strong data governance team to guide source data owners and data stewards for better data quality
- Domain and subject matter experts to lead the data quality assurance team
- A data quality enforcement policy
- Data quality and standardization rules
- Data quality metrics to measure the trust factor

The data quality process flow is illustrated in the next section.



Data Quality Architectural Framework

Due to telecom convergence and the increasingly competitive market, CSPs have to focus on scientific and accurate analysis of data. Quality analyses can be a part of any data intensive project such as data integration, data migration, master data management, data warehousing initiatives, etc., because Data Quality analysis is a costly exercise. The architecture diagram below describes a Data Quality analysis as part of Enterprise Data Integration (EDI).



Basically, EDI consists of architectural components such as data sources, Operational Data Store (ODS), data profiler, Data Quality Analysis (DQA), ETL and target database. The target database should be compliant with the industry standard SID framework. A SID-based CDM is a catalyst in the integration architecture since it enables CSPs to integrate disparate data sources and data from mergers and acquisitions with minimal rework.

Data Profiling: It is a methodical approach of discovering metadata, semantic and syntactic relationships within and across different data sources. It is an on-going activity in data sources that assists the data governance team to evolve a Data Quality rules repository for assessing and providing Go/No-go decisions. Data profiling also helps discover data problems such as inconsistency, missing data, duplicate data, data that does not meet business rules and orphaned data. The table below depicts classic examples of data anomalies:

Name	Age	DOB	Gender	Height	Anomalies
John	21	13-01-86	M	5.8	Nil
Mary	45	27-8-62	5.3		Lexical Error
Diana	36	20-1-71	F	5-9-2	Domain Format Error
Paul	27		T	7.1	Irregularities
Jim	30	30-9-67	M	5.7	Contradiction
Smith	0	26-9-78	M	6.1	Integrity Constraint Violation
Jim	40				Duplicates
#@5%	^^	@	T	0	Invalid Tuple
	40	30-9-67	F	5.7	Missing Values
					Missing Tuple

Systematic analysis of incoming data is mandatory to avoid exceptional data landing in the CDM or any target system. Data profiler helps data analysts to find metadata information such as data type, decimal places, data length, cardinality, primary key, unique count, unique percent, pattern count, minimum value, maximum value, minimum length, maximum length, null counts, blank counts, etc. Such analysis helps determine whether the data adheres to metadata definitions and the business expectations. Pattern matching is an advanced analysis on any given attribute whose distinct value is very low. This analysis helps to find out whether the data adheres to the expected format, is consistent across the data source, is completely numeric, and consistent in length. The table below shows the result of a typical pattern matching analysis on the phone number attribute. It helps a subject matter expert to frame Data Quality rule on phone numbers not adhering to the business standard.

Pattern	No. of Rows	Percentage
9999999999	180000	85.50
999-999-Aaaa	20000	9.50
{999} 999-9999	10000	4.75
999-999-9999-9999	500	0.24
999-Aaaa	15	0.01

Other advanced profiling techniques such as frequency distribution, domain value analysis, pattern outliers, and percentile analysis help subject matter experts in discovering the data and data quality. This helps them frame data de-duplication strategies and business rules. Data profiling also helps identify the relationship between attributes originating from different source systems, which in turn helps find records adhering to referential integrity. The benefits of leveraging data profiling include:

- Data meeting business expectations
- Discovering the level of quality of each attribute
- Deciding the type of cleansing, standardization or de-duplication required to meet business expectations

Data Auditing: Data quality rules and business rules discovered with the help of data profiling is configured in the data quality engine. The Data Quality engine triggered during data refreshment from the source to target systems screens incoming data against the rules and pipes data that does not adhere to the expectation of the exceptions repository. This process can be executed on an ad-hoc basis on any set of data to confirm validity of the quality rules. The screening process performs validation that is similar to profiling in column screening, structure screening and business rule screening to filter exceptions and categorize them by quality dimensions, type, priority and severity.

1. Column screening involves validations such as mandatory value checks, range checks, pattern checks, length checks, spell checks, etc.
2. Structure screening includes validations like relationship checks across attributes, relativity check between the primary key of one table and the foreign key of another table, etc.
3. Business rule screening takes care of validations that are very complex and checks for records that do not meet business mandatory expectations. Generally, business rule tests are very complex as they involve attributes from multiple tables and operations such as data aggregation, derivation, etc.

Typically, the Data Quality engine is configured with Data Quality rules and corresponding corrective measures in the metadata repository. The corrective measures repository is continuously evolved by subject matter experts by brainstorming with data stewards, exceptions that cannot be fixed at the source systems because of system behavior, legacy applications, non-availability of system documentation, etc. Exceptions that do not have any associated corrective measures are piped to an exceptions repository. The exceptions repository is scanned for fresh exceptions and data issues are communicated to the data governance team, data stewards and source system owners. Source system records that are fixed are processed during the next ETL to target CDM.

Data Cleansing (Part of ETL): Data that passed quality testing and can be fixed by corrective measures are passed for ETL processing. Consolidating data in a single physical common data model has proven to be a valuable approach to provide integrated access to trustworthy and relevant information. It involves a lot of data manipulation such as matching, merging, schema-related data transformation, reconciliation, cross matching, de-duping, data formatting, data standardization and cleansing. Data cleansing is the method of locating and repairing data anomalies by comparing the domain of values within the database. Data is scrubbed or cleansed as per the corrective measures suggested for exception like misspellings, missing information, invalid data, varying value representations, etc. Anomalies are fixed automatically by ETL processing. The data cleansing and standardization functions should be driven by properly designed metadata tables, without time and labor consuming pre-formatting or pre-processing. The DQ and ETL framework must understand, interpret and re-organize data automatically within any context, from any source, to the standard target form. Once the standard transformation and cleansing has been completed, enterprise data is standardized to work effectively. This improves operational efficiency as well.

Exception Analysis: During Data Auditing, if the Data Quality engine finds that any data item “stands-out” (holds statistically significant variance from a mean population), then the engine flags it as an exception and stores it in the exception schema. Such exceptions are thoroughly analyzed by the data governance team or subject matter experts. Depending on the category, exceptions are communicated to:

1. Data stewards to fix data anomalies at the source database
2. ETL experts to modify the algorithm for transformation, cleansing, standardization, etc.
3. Quality experts and business users to frame new quality rules/corrective measures

Quality Dimension	Accuracy					Uniqueness
	Integrity		Consistency		Density	
	Completeness	Validity	Schema Conformance	Uniformity		
Lexical Error						
Domain Format Error						
Irregularities						
Constraint Violation						
Missing Value						
Missing Tuple						
Duplicates						
Invalid Tuple						

■ Indicates direct downgrading of the quality dimension
 ■ Indicates that the occurrence of this anomaly hampers the detection of other anomalies downgrading the quality dimension

Based on feedback of the above communication and brainstorming sessions between subject matter experts, data stewards, ETL experts and quality experts, the Data Quality repository can be refreshed with exception priority, data attributes that benefit from new data quality rules and corrective measures. The matrix below contains a list of standard data anomalies against the quality dimensions. Correlating the configured Data Quality rules, based on fundamental data quality dimensions, allows subject matter experts to represent different measurable aspects of data quality. It can also be used in characterizing against the exceptions and finding how it impacts the business. A correlation matrix, as shown below, is used by the data governance program for quickly understanding the nature and severity of exceptions.

Exception Reporting: Exceptions captured during data auditing and ETL processing are analyzed to generate scorecards and dashboard reports. Trend reports based on quality dimensions, exception category, priority and severity help enterprises measure data quality performance at different levels of the operational business hierarchy, enabling monitoring of both line-of-business and enterprise data governance. The data quality framework automatically measures, logs, collects, communicates and presents the results to those entrusted with data stewardship and the source data owners. Such closed-loop monitoring ensures enterprise data quality.

Conclusion

Systematic roll out of Data Quality strategy and process helps enterprises deliver trustworthy strategic and tactical information to their business users. Successful implementation of data quality principles in the enterprise needs in-depth knowledge, awareness, expertise and hard work. The quality of information that flows across an enterprise's operational/ analytical applications, databases, networks, business processes and business rules, and linking to customers, channel partners, switching partners and helpdesk representatives are extremely critical in setting strategic directions for the enterprise. Enterprises must assess the level of quality of information at the technology, process and people layers to provide a quality framework.

About the Author

Sivaprakasam S.R. is a Principal Architect and mentors the Database and Business Intelligence track in the Communication, Media and Entertainment practice at Infosys. His areas of interests include Enterprise Data Modeling, Enterprise Data Integration, Enterprise Data Warehousing, Enterprise Data Quality Management and Semantic Data Integration. He can be reached at <<sivaprakasam_s@infosys.com>>



For more information, contact askus@infosys.com

About Infosys

Many of the world's most successful organizations rely on Infosys to deliver measurable business value. Infosys provides business consulting, technology, engineering and outsourcing services to help clients in over 30 countries build tomorrow's enterprise.

For more information about Infosys (NASDAQ:INFY), visit www.infosys.com.