WHITE PAPER

A framework to increase ROI through quality data

- Kuriakose K. K.
Senior Project Manager

Infosys®

The perception of data across organizations is changing. Data is no longer just one of the components of business. It has turned out to be 'the business.'

Today, information is viewed as a lifeline by senior management and by many other departments for decision making, customer interactions, and optimum operations. An organization's success heavily depends on how it is able to understand and leverage its data. Unfortunately, there continues to be a high amount of inaccurate data within enterprises today, despite developing multiple solutions/ systems to counter the same. Nowadays, a major portion of data for decision making is collected from external sources through a variety of channels. This often results in poor data quality, which has a negative effect on an organization's ability to take decisions, run business, budget, market, and gain customer satisfaction. Organizations which fail to control the quality of data, is unable to sustain in today's data-centric world. Any data-driven effort needs to have strong focus on data quality, which implies that organizations looking for success are mandated to prioritize data accuracy and accessibility. It is essential for them to interact with consumers, vendors, suppliers, and third parties in countless ways, by exploring diverse new methods of communication. Information is the key for areas like inventory management, shipment, and marketing. The objective of this paper is to analyze principle challenges with data across few key business functions and discuss a framework which can bring down the erroneous data that is getting pumped in and out of an enterprise system.

Marketing:

If we have accurate information on who our customers are and what are their needs are, we have hit gold in Marketing terms. This is more easily said than done as today we neither have an accurate nor enough information about customers. We can surely gather information about customers from various sources like- website, physical store, mobile application, call center, face to face, catalogues etc. But, one can never be sure if they are same or different set of people consuming your services. There's no surety of the information being accurate as most of these channels accept data directly with limited to no validations.

Now let's assume that we have done all possible validations and found out the target group of customers but there's still no defined method of reaching them-- should it be through emails, telephonic conversations, social media, physical address etc. Let's drill this down into one of the mediums as--physical address. The catch- the customer has many addresses like for credit card, savings bank account, driving license, and for office purposes.

Shipping

The current status of shipments is constantly added to enterprise systems through shipping vendors like DHL Express, DHL Parcel, United States Postal Service (USPS), United Parcel Service (UPS), FedEx, Canada Post, LaserShip, OnTrac, and Hermes. Most of these vendors do not even share shipment history, hence organizations are forced to store and link this continuous flow of information. Many times incorrect data gets fed into system through these external sources. We see data like order shipment data being before order book data etc. This results in:

- Order lost in transit
- Incorrect shipping address
- Order sent to wrong address
- Shipping wrong items
- Late shipments

Inventory

Inventory management can help a manufacturer / supplier in improving accuracy, cost savings, and speed. This in turn will help organizations have better control on operations and reduce cost of goods. Today, most of the manufacturers are facing challenges in inventory management systems. Few challenges listed below:

- Limited standardization in management systems, business users, inventory integration, and movement checkpoints
- Limited inventory reconciliation on regular intervals
- Data discrepancies between demand planning and inventory planning systems
- Improper logging of inventory information
- Inaccurate data fed into forecasting systems

Insurance

When it comes to insurance industry, data not only helps run operations, but also helps them ensure that claims have the required and correct information. Generally, the following issues are found in the claims data:

- Invalid diagnosis codes
- Incorrect pin codes of hospitals
- Incomplete forms with missing crucial data like gender, patients' stay in hospital, etc.
- Inaccurate age data

Banking

Financial organizations are required to meet regulatory compliance requirements according to the law of the land to avoid instances such as housing crisis. At the same time, data quality issues lead to transparency and accountability problems. Hence, quality of data for banking needs to be measured along the dimensions of completeness, accuracy, consistency, duplication, and integrity. There is also a need to ensure information that is being shared complies with information privacy and protection laws and regulations.

Pharma

Pharmaceutical industry gets warnings on regular intervals for falsifying, altering, or failing to protect essential quality data on their drug manufacturing process and its validation, resulting in huge business risks. According to US Food and Drug Administration (USFDA) regulations, pharma companies are mandated to maintain manufacturing and drug testing data. Many times, issues occur due to human data entry errors and machine errors like data recording failures. These regulations have even resulted in shutdown of plants causing huge losses.

Today's state

Many organizations are constantly investing in data quality to improve their efficiency and customer interactions through data insights. Majority of the companies are suffering from common data errors like incomplete or missing data, outdated information, and inaccurate data. This level of inaccurate data jeopardizes business that relies on business intelligence for taking key decisions. An organization's current level of maturity can be assessed from data quality maturity model given below:

Data quality maturity model				
Think & Act Local	Think Global & Act Local	Think Global & Act collectively	Think & Act Global	Matured Data Governance model
• Limited awareness of data quality • Certain defined rules for data quality and integration for a specific module based on production issues encountered • Duplication of data across systems with no established system for data quality validation	• Data inconsistency across systems • Organizations / programs accepting the impact of inconsistent, inaccurate, or unreliable data • Steps initiated to identify corrupt data • Gains are defined more at project level	• Data inconsistency across systems • Organizations / programs accepting the impact of inconsistent, inaccurate, or unreliable data • Steps initiated to identify corrupt data • Gains are defined more at project level	• Establishment of a well defined and unified data governance model • Regular checks and reporting of established business rules and data quality on a defined frequency • Organization shift toward management of data as business critical asset	• System information and data is trusted • Key metrics for data quality is tracked against the defined variance percentage • Action items are tracked to closure for any variances beyond the agreed limit • ROI for data quality projects is tracked

Automated data quality framework:

This calls for a need of a strong quality framework, which can validate standard business rules against the processed data coming from external sources into enterprise systems. This framework should be able to report incorrect data and related information.

A framework which has easily configurable rules and threshold values can be set by business using simple text through a user interface directly into framework. The framework can connect to almost all kinds of data sources — mainframes, file systems, relational database management system (RDBMS) systems, analytical databases such as columnar, massively parallel processing (MPP), in-memory data base, NoSQL databases, Hadoop, web services, packaged enterprise applications, OLAP applications, software as a service, and cloud-based applications.

The details of common business rules are also collected by our subject matter experts (SMEs) in retail, consumer packaged goods (CPG), logistics, manufacturing, banking, healthcare, insurance, life sciences, capital markets, financial services, cards and payments, energy, utilities, communications, and services. This helped in the creation of a backbone for our standard quality framework where one can add / remove rules according to the specific business need.

The user can pick a set of business rules and schedule it according to their need. An automated report gets generated which is emailed to the concerned parties. It is recommended to go with open source solution to bring down the cost of development and maintenance of the tool. We have used a combination of tools--Talend and Python scripts for the development. This framework can be based out of other open source solutions like KETL, Pentaho Data Integrator - Kettle, Talend Open Source Data Integrator, Scriptella, Jaspersoft ETL, GeoKettle, CloverETL, HPCC Systems, Jedox, Python, Apatar. The framework can also be enhanced further to carry out data profiling and data cleansing on an "as-needed" basis.

Infosys Automated data quality framework consists of a configurable data quality (DQ) framework with built-in, reusable rules across domains with the following:

- Standard business rules which can validate the processed data and information from third parties
- Framework reports incorrect data and related information crossing the thresholds
- Capability to easily configure rules and threshold values independently
- Daily automated report generation, post job completion, enables independent operations for business

Realizing the return on investment (ROI) for data quality

Today, businesses need relevant data to make informed decisions. Decisions and communications based out of bad data carries substantial risks to business performance. For any data-driven organization, it is important to ensure that utmost standards of data quality are met and the organization has scheduled processes to validate quality along with the data that is being pumped in and out of the organization. We also need to ensure that a structured methodology is being followed in data quality metric definition and its validation on regular intervals. Few possible outcomes of successful implementation of a strong data quality framework are:

Marketing:

Accurate data helps drive more effective campaigns for the intended target audience

Shipping:

Cost savings and operational efficiencies achieved with basic address validation and order-related data quality checks

Inventory:

Faster turnover of stock

Insurance:

Complete information about client's risk exposure enabling more accurate decisions on policy costs

Banking:

Ability to detect fraud patterns and improved customer service

Pharma:

Gain more compliance as per FDA regulations

For more information, contact askus@infosys.com

Stay Connected