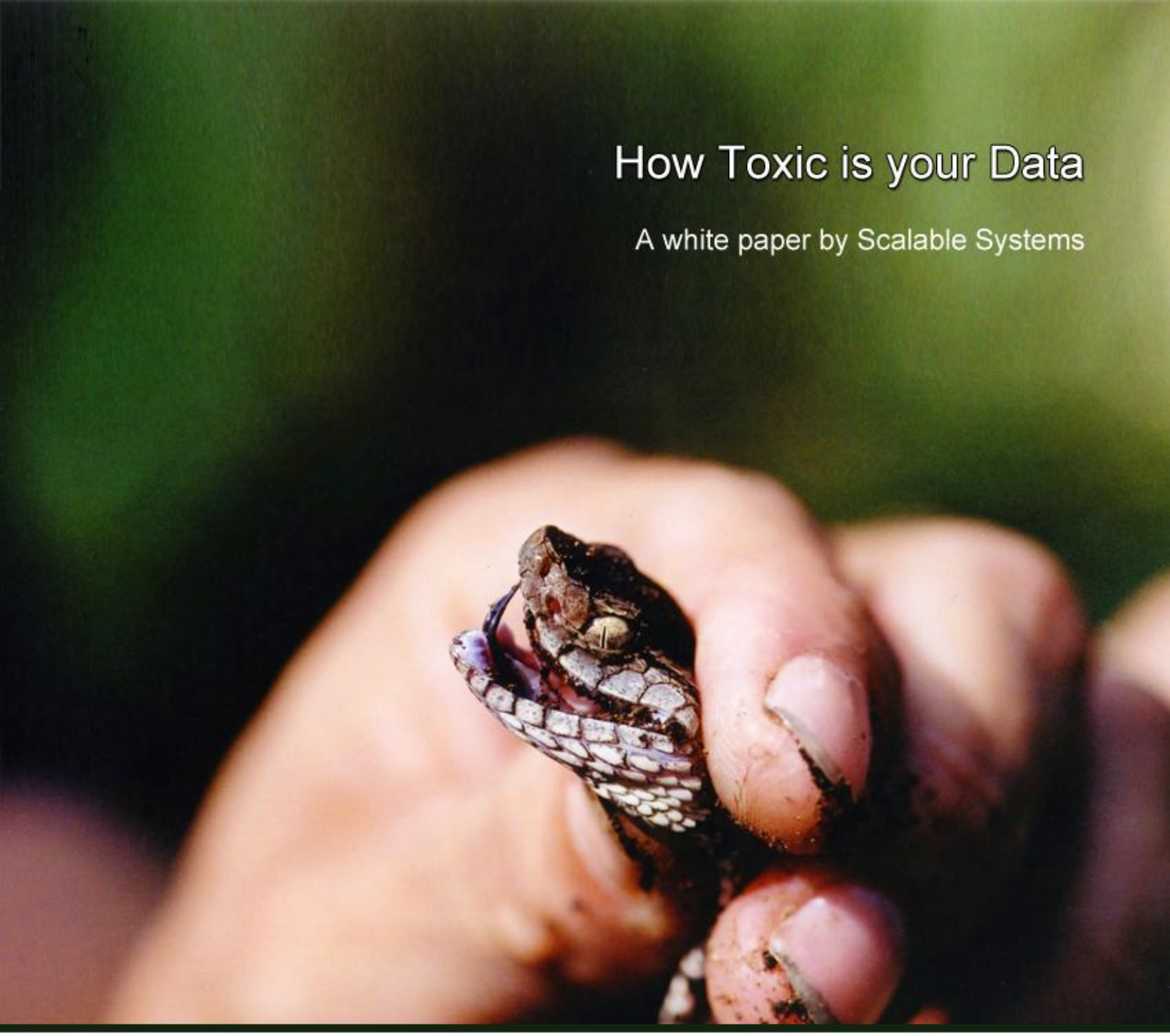


How Toxic is your Data

A white paper by Scalable Systems

SCALABLE SYSTEMS | Transforming Data into Intelligence

Executive Summary

Businesses of all sizes are experiencing a massive explosion in the volume of data. Although data is recognized as one of the most important assets of a company, many times the quality of data is overlooked.

Toxic data works as a silent killer that oftentimes results in catastrophic downstream consequences. Declining sales, poor customer service, inadequate response to customer demands, and inability to innovate and sustain clients are only some of the potential negative outcomes.

Global competition, aggressive price reductions, and challenging market conditions are today's harsh reality. The company with the competitive edge is the company that will survive market downturns. Data, then, is the lifeblood of any company.

A typical scenario in a company infected with toxic data.

Toxic Data Spiral	Effects
The operational data management team works harder to meet the data demand	Frustration rises
Data pressure increases from the marketing team because customers are leaving	Customers leave
The finance department feels dizzy falling short of quarterly targets	Revenues disappear
The IT department does everything it can, but not getting enough results	Fatigue appears
Organization Management experiences increased sweating and cancels golf meetings	Morale is down
Organizational vision becomes blurred.	Teamwork evaporates
The downward spiral continues	Customers and employees abandon ship

All of the above could have been avoided had the company taken the preventive steps to keep its data quality healthy.

Scalable Data Quality framework provides:

- ✓ Regular data governance exercises
- ✓ Frequent data quality checks
- ✓ Identifies and prevents toxic data entry points
- ✓ Data cleansing before the database is infected
- ✓ Avoid the Multiplicity Syndrome

Studies by National Archives and Record Administration showed that 80% of companies without well-conceived data management and recovery strategies go out of business within two years of a major disaster.

Overview

According to Gartner Inc., companies may lose more money in operational inefficiency due to data quality issues than they spend on data warehousing and CRM activities. Though most companies recognize the importance of data quality, many times data quality initiatives are lost while coping with long-range plans and resolving day-to-day operational issues. In hindsight, most managers agree that they did not provide sufficient attention to data while developing operational systems.

Many Organizations have a system to improve data quality by using various methods such as data profiling or data cleansing tools to cleanse the toxic data with Extract-Transform-Load (ETL) tools for Data Warehouse practices and other important applications. All these technology oriented data quality efforts are gradual steps in the right direction. The fact, however, remains that technology solutions alone cannot eradicate the root causes of poor quality data because poor quality data is not as much an IT problem as it is a business problem.

Toxic data is a serious concern in today's business environment. Data quality issues must be addressed systematically and organizationally. Enterprise-wide data quality discipline must be established and constantly nurtured. Company data should be valued and treated as a business asset, much like other company assets such as valuable employees and long-term customers.

When data is not complete, correct or consistent, businesses often fail. However, IT departments cannot manage data quality by themselves. According to a study published by the Data Warehousing Institute, "Taking Data Quality to the Enterprise through Data Governance," data quality is mostly related to business issues.

The Data Warehousing Institute, which provides research and business intelligence for the data warehousing industry, estimated that data quality problems cost U.S. businesses over US $ 600 billion a year.

How Toxic Data Happens

Data Quality becomes a problem when companies do not treat information as an asset that can bring measurable value for growth and profit. When data is not cleaned, checked, validated, measured, or simply not cared for, it becomes contaminated.

“Multiplicity Syndrome” – when duplicate versions, especially from legacy systems, leads to system-wide data confusion and inaccuracy. In companies where there are few data entry points and sporadic data usage, the multiplicity syndrome may not apply. However, even in those cases, data does not remain static. When external applications, such as ERP or CRM are running, it is very difficult to enforce data quality standards.

Reasons for Toxic Data

Poor data entry habits without adequate validation and quality checks

Many legacy systems developed years ago do not have enough validation and checks in place to prevent data entry errors and anomalies. Also, if some validation issues exist, data entry operators have found easier ways to override the problems. For example, if a telephone number entry has a validation that the telephone number should have an xxx-xx-xxxx format, a data entry operator can easily override the validation by entering 111-11-1111, which is of no value.

Lack of clarity in business rules definition

Speculation happens when business requirements are not articulated precisely. When speculation spirals – usually downward - incorrect data modeling, faulty processes, and inaccurate reports follow.

Improper interpretation of Business Rules

Many times, business users who are involved in intermittent data entry activities might be clear about some, but not all, business rules. In this scenario, they might enter the data they think is correct. This leads to data inconsistency and can contribute to toxic data if not resolved.

Poor data capture

During system requirements definition we rarely bother to gather the data requirements from down-stream information, such as the marketing department. For example, if we build a system for the lending department of a financial institution, the users of that department will most likely list Initial Loan Amount, Monthly Payment Amount, and Loan Interest Rate as some of the most critical data elements. However, the most important data element for marketing department users is probably Gender Code, Customer Age, or Zip Code of the borrower. Thus, in a system built for the lending department, data elements such as Gender Code, Customer Age, and Zip Code might not be captured at all, or only haphazardly. This scenario is often the reason why so many data elements in operational systems have missing values or default values.

Poor data modeling and data architecture

Data Modeling and Architecture needs to be designed in a scalable way so that when the data size grows and new applications are added, it should withstand the pressure of change. Poor architecture will lead to duplicate entries, redundant data across the systems, and improper correlation between tables. Eventually as the data load increases, the systems may collapse without any prior warning.

Improper data mapping from other ERP and CRM systems

In a complex business environment there is a continuous flow of data from one system to another. If there is improper data mapping that remains undetected, then toxic data starts circulating throughout the veins of the company's data center. Oftentimes, these type of errors are hard to detect because most of the time data mapping between heterogeneous systems focuses on field type matching and standard validation. For example, if first name in one ERP system is mapped to last name in CRM systems, then it might pass through since both are string values. Now, this toxic data creates negative downstream consequences if a customer's name is always misspelled when the customer receives correspondence from the company.

Technical errors that occur during the transmission of data

Along with the growth of e-commerce, there came an increasing dependence on software programs to automate tasks involving databases comprised of customer information. This opens doors for software programs to accidentally and incorrectly execute tasks that affect thousands or millions of records at a time.

Data errors during application migration

When applications are upgraded to other platforms for better performance and better user interfaces, there is a possibility during the migration process that the application code that was used to handle data in a specific way might not handle the same data element in the same manner after migration. Therefore, it is important to give special attention to data sensitive applications.

Data errors during database upgrade and updates

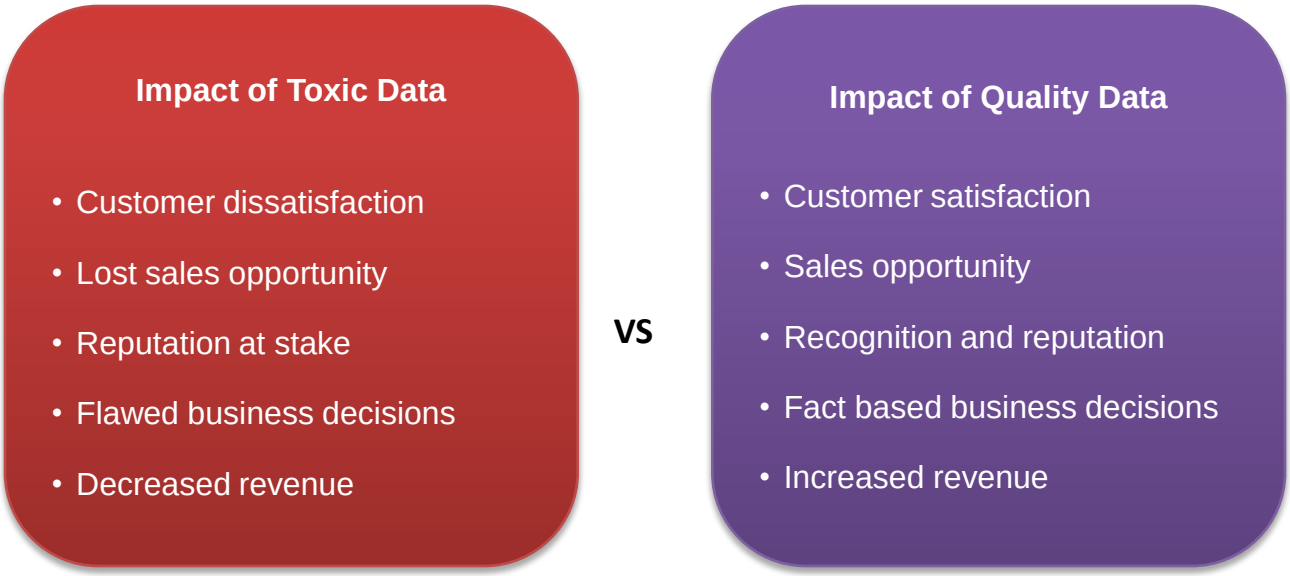
When the database is upgraded, there is a likelihood that the new database might not support some older functions for calculating data values. For example, there are certain data functions which might work in Oracle 8 but not work in Oracle 10G (the application might not handle the data the way it should after the migration). If undetected in the production environment, the application will start utilizing bad data that can become toxic over time.

Data quality as a non-priority issue

Many companies realize that they did not pay sufficient attention to data while developing systems during the last few decades. While delivery schedules have been shrinking, project scopes have been increasing, and companies have struggled to implement applications in a time frame that is acceptable to their business community. What usually gives is quality, especially data quality.

Effect of Toxic data on business

The consequences of toxic data quality are real. At the most basic level, toxic data can affect revenue, costs, and customer loyalty. Data quality is a critical component of business success. Poor quality data jeopardizes the performance and efficiency of operational systems. It also undermines the value of business intelligence systems on which companies rely to make key decisions. Decisions based on such faulty data can cause direct financial loss, spoil customer relations, and damage a company's credibility in the marketplace. As more companies recognize data as a strategic asset, business leaders are increasingly being held accountable for ensuring the accuracy, quality, and reliability of information. Poor data quality adversely affects your organization in three key ways:



1. **Poor data quality causes inefficiencies in business processes which depend on data**

   Almost every business process is dependent on data in some way or another. From customer order entry, invoicing, reporting, and business analytics data plays an important role. Even if every business process is performed perfectly, poor data will change everything. These inefficiencies result in very expensive rework efforts to "fix" the data in order to meet the requirements of various processes.

   As an example, the following losses may occur due to poor data quality and lack of data validation in financial transaction. The bank amount in the books may not agree with the amount at hand in the bank. Duplicate invoices might be paid resulting financial losses. Payment to a vendor may be made when there is a large outstanding receivable from that company. The discount amount may be calculated incorrectly. Payments made may be posted in the wrong account.

Poor data quality gives rise to poor decisions

A decision can be no better than the information upon which it is based. Critical decisions based on poor quality data can have very serious repercussions, yet another reason why companies should make sure that their data actually represents reality.

Congressional investigators said recently that two-thirds of U.S. health insurance industries use a faulty database that undercompensated patients when patients saw doctors outside their insurance network. This faulty calculation costs Americans billions of dollars in inflated medical bills.

Poor data quality creates mistrust

Poor data quality can reflect adversely on a company, lowering customer confidence. If the data is wrong, time, money, and reputations can be lost. The cost of losing one customer is, according to some studies, four times higher than obtaining that same customer due to advertisement costs and marketing staff expenses. A recent report by Experian marketing services division says that U.S. businesses admit to losing 7.3 percent of revenue due to poorly managed customer data. The report goes on to say that 77 percent of companies that confess shortcomings in data quality acknowledge that the shortcoming has a detrimental effect on their bottom line.

Most of what is stated below is obvious, but it could be used for people to create business cases for data quality programs. It may also relate to the incidents organizations must face in real-life.

Processes	Impact
Customer Retention impact	○ A better CRM system does not guarantee customer data quality and is unable to generate return on investment on its own. It is the quality of the data that is fed into the system that makes all the difference ○ Incorrect customer names and addresses decrease customer trust ○ Customer complaints lead to customer attrition
Operational Inefficiency	○ When wrong data is detected, corrective actions take away an organizations focus ○ Internal impacts like investigation, root cause analysis, fixing process/IT issues and constant monitoring ○ External impacts like addressing all the implications of wrong data like stop-payment for faulty checks issued, and recall of credit cards issued to wrong addresses
Customer Acquisition Impact	○ Undelivered mail leads to failed mailer campaigns ○ Mailed products getting returned due to errors in names ○ Dissatisfied sales and distribution channels, due to erroneous compensation
Operational Effectiveness	○ Not able to track the status of the delivery ○ Errors in the delivered product ○ Bad analysis of business ○ Bad campaign
Business Impact	○ Poor data hurts business operations in many ways. Errors may be made in fulfilling customer orders, invoices sent to incorrect locations, duplicate payments made to vendors, and customer payments applied to wrong accounts. The results are unhappy customers and vendors, as well as frustrated and inefficient employees.

Reputation Impact	○ A data issue impacting a wide set of stakeholders could lead to media coverage. ○ Major product recall
Shareholder Impact	○ Faulty financial statements and less than appropriate audit ratings could lead to loss of confidence from shareholders and the investing public
Regulatory Impact	○ Faulty regulatory submissions, leading to considerable legal exposure ○ Bad data quality leading to customer or shareholder impact, could result in lawsuits
Decision Impact	○ Quality of decision depends on quality of data. Bad data quality leads to misinformed or under-informed decisions
Business Management Impact	○ Lack of data on the key performance indicators, hampers objective performance management

How to Detoxify Toxic Data

For detoxification the existing data needs to analyzed in a careful and scalable manner. There might be lots of places where data quality might be poor. But all data may not be cleaned immediately. Focus on the immediate problems which might be caused by poor data. Once the problem area is identified then the following methods can be used to analyze, identify, and clean data.

- **Data Profiling**
  Data profiling is a process where data content, structure, and relationship are evaluated and measured. During data profiling various data anomalies like empty columns, unused data values, overused data values, duplicated data columns, violation of structure rules, violation of business data rules, and representation of missing values are discovered. Data profiling can be done by many popular data profiling tools or it can be conducted in house by SQL queries.

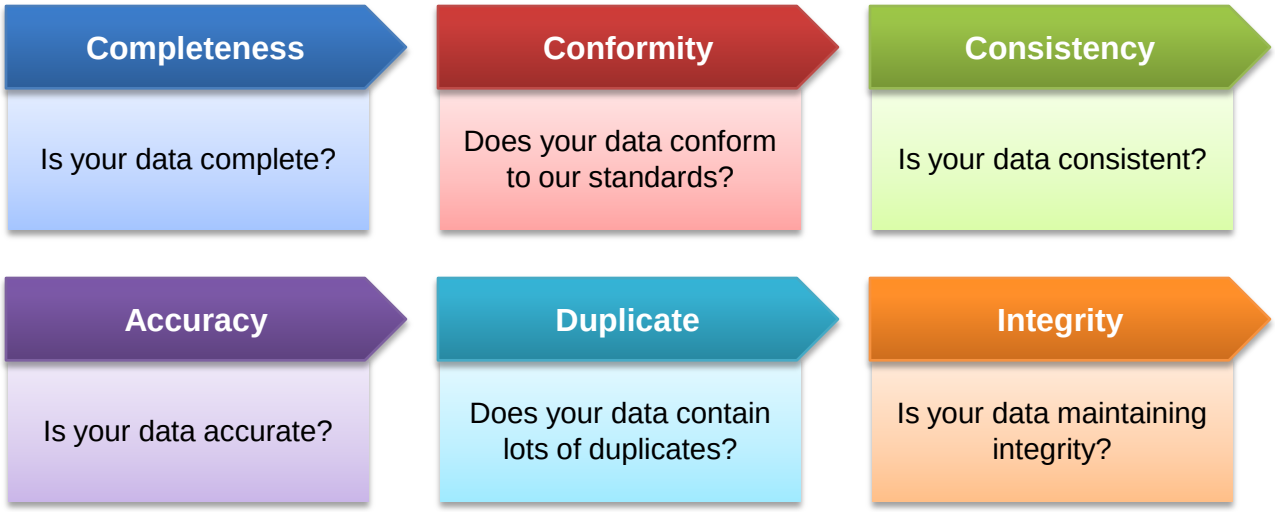
- **Data Cleansing and Enhancement**
  Data cleansing is the process of correcting or removing toxic data. Data cleansing is a repeating process until all data issues are cleared. Data cleansing requires a thorough understanding of the business objective and involves looking for and handling data errors, outliers, and missing values.

- **Data Matching and Consolidation**
  Match similar records and perform de-duplication and consolidation based on set criteria. Matching is done on various business rules like name, address, and SSN. Once the duplication is determined, merge is performed on duplicate records if they are the same identity. This directly benefits the removal of duplicates from your database and improves the accuracy of customer information in the customer database.

How to Prevent Toxic Data and Manage Data Quality

About 90% of toxic data enters at various data entry points in an organization. If bad data entry can be prevented at the source checkpoint, organizations can save a significant amount of capital spent on data correction and detoxification. Entering your data correctly for the first time is the best way to ensure the integrity of data. Be prepared to spend money on this data collection stage. It saves more money in the long term. Using auto complete or other data validation applications at the data entry data stage is highly suggested. Use of controls and input masks during data entry can also help to correct entry in formatted fields.

For ongoing data quality improvement a data quality framework should be established and followed. The data quality framework is intended to provide a common objective approach to assessing data quality. There are six data quality dimensions.



Data quality is an iterative process of assessment, planning, and implementation. Organizations should repeat the data quality process in a constant fashion in order to measure ongoing effort and effectiveness. Assessment results can be used to build an economic model that evaluates the costs associated with instituting improvements. This model can be viewed as a scorecard that documents data quality levels associated with a set of data quality dimensions measured at specific locations in the information chain.

The data quality scorecard is a framework for calculating the return on investment for improved project implementation. The scorecard can be used as a management tool, in which any suggested improvement is connected with the cost of designing and implementing the improvement along with a time frame for implementation. Ultimately, this scorecard can be used as the basis for an ongoing data quality improvement project that will subsequently enhance all of the company's intelligence efforts.

Conclusion

Data quality is an ongoing process and cannot be achieved overnight. As per the Japanese Kaizen principle, small daily improvements eventually result in huge advantages. Data quality is a broad umbrella term for the accuracy, completeness, consistency, conformity, and timeliness of a particular piece or set of data and for how data stores and flows through the enterprise. Different organizations will have different definitions and requirements for data quality, but it ultimately boils down to data that is “fit for a purpose.” Data, as an asset, must be usable by companies for constant growth.

We at Scalable Systems view our solutions approach to customer data as an art form - because it is both a creative and constantly evolving process. Rather than merely cleansing and organizing your database, our preference is to continually nurture, organize, cherish, and maintain your data to ensure it does not become toxic at any point now or in the future. With our expertise in data model architecture, database administration, data migration, sound database development, data quality framework, and master data management, we provide holistic and long-term solutions for the most important asset of your organization – Data.

- Bibliography

- 1. Data Strategy: by Sid Adelman, Larissa T. Moss, Majid Abai
- 2. Enterprise Knowledge Management: The Data Quality Approach (The Morgan Kaufmann Series in Data Management Systems)
- 3. Data Quality Assessment by Arkady Maydanchik
- 4. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (TM) by
- Danette McGilvray

About Scalable Systems:.

Scalable Systems is a global software consulting, development and IT outsourcing company providing both onshore and offshore software solutions and integration services to business enterprises around the globe. Scalable Systems has proven expertise in encompassing low cost, but high quality and reliable software solutions and services in areas like Data Management, Business Intelligence, Content Management and Application Development.

Scalable Systems
Email: info@scalable-systems.com
Web: www.scalable-systems.com

This edition published August 2008