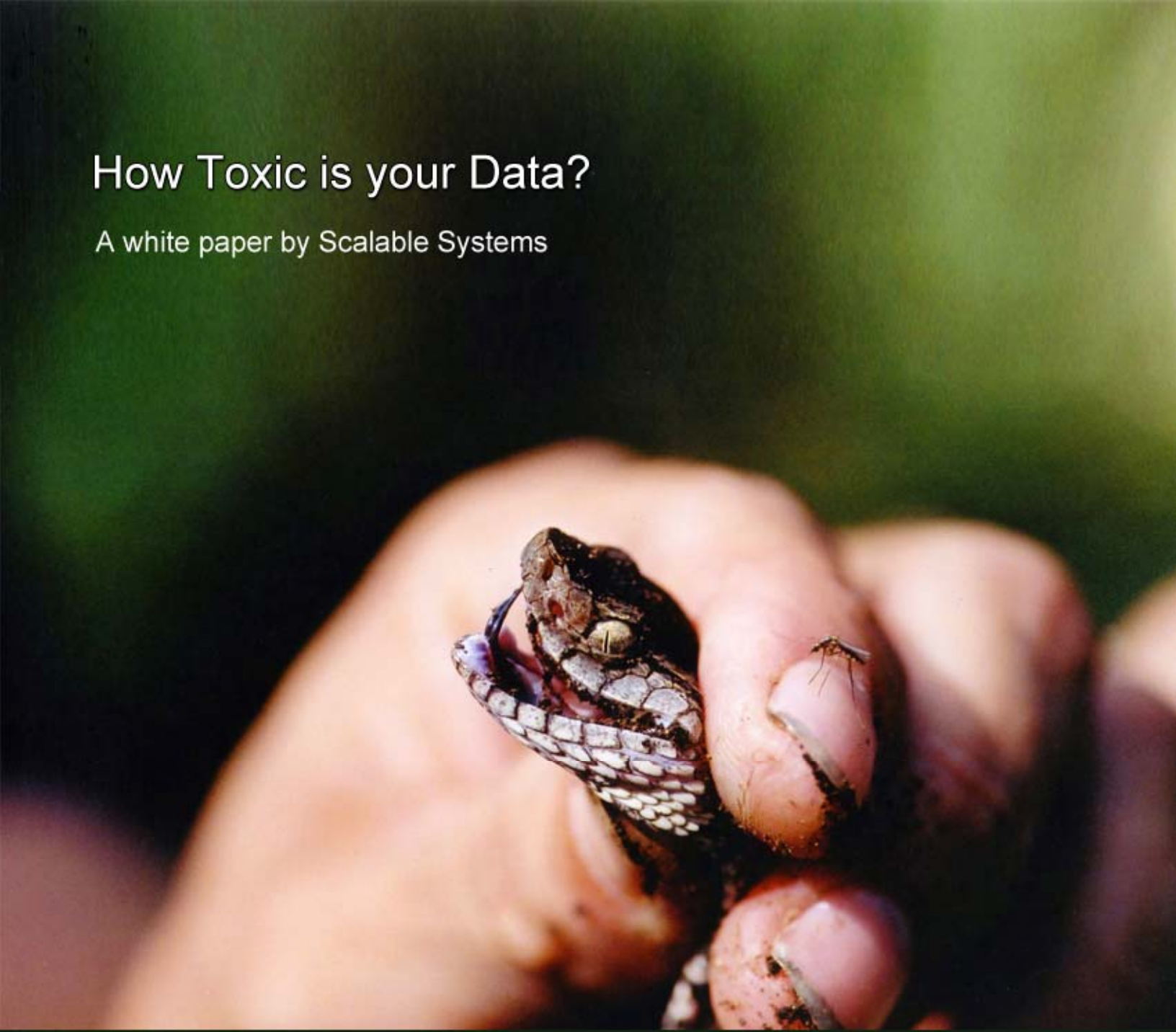


# How Toxic is your Data?

A white paper by Scalable Systems



Transforming Data into Intelligence

## Executive Summary

Business of all sizes is experiencing a massive explosion in the volume of the data. Though Data is being recognized as one of the most important assets of the company many times Data is overlooked. As a result the quality of data goes from bad to worse making it toxic. Toxic data could be damaging to any organization. Many times data toxicity may not be recognized but works as a silent killer damaging to business in a variety of ways. Ongoing harmful intrusion of toxic data over a time period can make organization sick. If toxic data is not cleaned and prevented it may lead to various organizational hazards like declining sales, poor customer service, inadequate response to customer demands and inability to innovate and sustain in difficult business environment. During earlier days almost any business could survive if they focus on quality customer services, good product or service line and reasonable marketing effort. Today even best businesses with exceptional customer services, innovate products or services, highly experienced management and dedicated team of employees cannot be guaranteed success due to global competitiveness, aggressive price reduction, easy information availability and bad market condition. It seemed now there is numerous exceptions to every business rule defined in past decades.

Data is the lifeblood of any organization. To maintain a healthy and happy organization there should be a healthy data flow throughout organization. Toxic data at any department of the organization over a time period may lead to data plaques which hampers the smooth flow of data. When smooth flow of data hampers and data becomes toxic here is a typical scenario in an infected organization. The operational data management team works harder to meet the data demand. Data pressure increases from marketing team since customers are leaving. Finance department feels dizzy with regular short of quarterly targets. IT department experiences fatigue while doing everything they can but not getting enough results. Organization Management experiences increased sweating and cancels golf meetings. Suddenly organizational vision gets blurred and the negative spiral continues.

Had the organization taken the preventive steps to keep the healthy data flow by Scalable Data Quality framework, regular data governance exercises, frequent data quality check, identifying and preventing toxic data entry points, data cleansing it could have positioned itself as a healthy, strong and growing organization.

Bad Data can be very costly, particularly for small and medium sized businesses where the difference between survival and closure can rest on the ability to recover from a disaster. At the very least, critical data loss will have a financial impact on companies of all sizes. The financial impact on a company is a combination of loss of business, low productivity, legal action, and cost of re-creating data. At its very worst, critical data loss can lead to business collapse.



## Overview

According to Gartner Inc. organizations may lose more money in operational inefficiency due to data quality issues that they spend on Data warehousing and CRM activities. Quality of Data is important to get the desired business result for every organization. Though most of the organization recognize the importance of data quality many time data quality initiative gets lost in the while coping with changes, planning for growth, resolving day to day operational challenges. Most of the organization agrees 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 use of data profiling or data cleansing tools to cleanse the toxic dirty data with Extract-Transform-Load (ETL) tools for Data Warehouse and other important applications. All these technology oriented data quality efforts are 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. Most of what is stated following is obvious, and may also relate to the incidents they must have faced in real-life.

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. Organization data should be valued and treated as assets as other tangible assets like building, employees, customers are treated.

# How Toxic Data Happens

Data Quality becomes a problem when organization do not treat information as an asset which can bring measurable value for growth and profit. When data is not cleaned, scrubbed, checked, validated, measured or simply not cared it gets contaminated. With the same data having many duplicate versions especially from legacy systems leads to multiplicity syndrome creating great confusion and inaccuracies. Data Quality will not be problem, if there were a few data entry points and sporadic data usage. But “the environment doesn’t sit still” and when external applications such as ERP, CRM are running in an organization, it is very difficult of enforcing data quality standards. When data is not complete, correct and consistent, businesses often put the blame on IT. It is natural for organizations to think of data as being IT responsibility. However, IT departments cannot manage data quality by themselves. According to a study published by the Data Warehousing Institute entitled “Taking Data Quality to the Enterprise through Data Governance” data quality is mostly related to business issues.

Data can become toxic over a time period due following reasons:

## **Poor data entry habits without adequate validation and quality check**

Many legacy systems developed many years back does not have enough validation and checks to prevent the data entry errors and anomalies. Also many times if there are some validation data entry operator have found easier ways to override it. For example if a telephone number entry have a validation that the telephone number should have xxx-xx-xxxx format an operator can easily override the validation by entering 111-11-1111 which is of no value.

## **Lack of clarity in business rules definition**

If business requirements were not articulated in a precise manner that leads to speculation and speculation spirals down in a wrong path of incorrect data modeling and great applications that brings, processes and reports incorrect data.

## **Improper interpretation of Business Rules**

Many times business users who are involved in intermittent data entry activities might be clear about some business rules. In that case they might enter the data which they think are correct. This leads to data inconsistency and if not resolved can become toxic data.

## **Poor Data Capture**

During system requirements definition we rarely bother to gather the data requirements from down-stream information of consumers such as from 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 elements for users of the marketing department are 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 often is the reason why so many data elements in operational systems do 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 added it should withstand the pressure of changes. Poor architecture will lead to duplicate entries, redundant data all across the systems, improper correlation between tables. Eventually as the 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 system. If there is improper data mapping which remains undetected then toxic data starts circulating throughout the veins of organization data centre. Many time 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 harmful effects of like customer dissatisfaction whose name always spelled wrong when they get a letter 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 of customer information. This opens doors for software programs to accidentally execute tasks that affect thousands or millions of records at a time incorrectly.

## **Data errors during application migration**

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

## **Data errors during Database upgrade and Updates**

When the database is upgraded there is a likelihood that the new database might not be supporting some old functions for calculating data values. For example there are certain data functions which might work in Oracle 8, but not work in Oracle 10g. In that the application might not handle the data the way it should after the migration. If undetected in production environment the application will start bringing bad data which can become toxic over a time period.

## **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 been struggling to implement applications in a time frame that is acceptable to their business community. Because a day has only 24 hours, something has to give, and 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 organizations rely to make key decisions. Decisions based on such data can cause direct financial loss, spoil customer relations, and damage an organization's credibility in the market place. As more organizations 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 those business processes which depend on data**

Almost every business process is dependent on data in some way or other. From customer order entry, invoicing, reporting, business analytics data plays an important role. Even all business processes are near perfect poor data with 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 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 wrong account.

## **2. Poor data quality gives rise to poor decisions**

A decision can be no better than the information upon which it's based, and critical decisions based on poor-quality data can have very serious consequences. This is another reason why you should make sure that your data actually represents reality.

Congressional investigators said recently that two-thirds of the U.S. health-insurance industry used a faulty database that undercompensated patients when they saw doctors outside their insurance network, costing Americans billions of dollars in inflated medical bills.

## **3. Poor data quality creates mistrust**

Poor data quality can reflect adversely on your organization, lowering customer confidence. If the data's 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 U.S. businesses admit to losing 7.3 percent of revenue due to poorly managed customer data and 77 percent of companies confessing shortcomings in data quality that are having a detrimental effect on their bottom line.

Most of what is stated following is obvious, but could be used for people to create business case for data quality program. It may also relate to the incidents they must have faced in real-life.

Processes	Impact
Customer Retention impact	<ul style="list-style-type: none"> <li>○ 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</li> <li>○ Incorrect customer names, addresses decreases customer trust</li> <li>○ Customer complaints leading to customer attrition</li> </ul>
Operational Inefficiency	<ul style="list-style-type: none"> <li>○ When wrong data is detected, corrective actions takes away organization focus</li> <li>○ Internal impacts like investigation, root cause analysis, fixing process/IT issues and conducting constant monitoring</li> <li>○ External impacts like addressing all the implications of wrong data like stop-payment for faulty cheques issued, Recall of credit cards issued on wrong addresses</li> </ul>
Customer Acquisition Impact	<ul style="list-style-type: none"> <li>○ Undelivered mail leads to failed mailer campaigns</li> <li>○ Mailed products getting returned due to errors in names</li> <li>○ Dissatisfied sales and distribution channels, due to erroneous compensation</li> </ul>
Operational effectiveness	<ul style="list-style-type: none"> <li>○ Not able to track the status of the delivery</li> <li>○ Errors in the delivered product</li> <li>○ Bad analysis of Business</li> <li>○ Bad campaign</li> </ul>
Business impact	<ul style="list-style-type: none"> <li>○ 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, customer payments applied to wrong accounts. The results are unhappy customers and vendors, frustrated and inefficient employees, as well as the controls those are compromised</li> </ul>
Reputation Impact	<ul style="list-style-type: none"> <li>○ A data issue impacting a wide set of stakeholders could lead to media coverage.</li> <li>○ Major product recall</li> </ul>
Shareholder Impact	<ul style="list-style-type: none"> <li>○ Faulty financial statements and less than appropriate audit ratings could lead to loss of confidence on the shareholders and investing public</li> </ul>
Regulatory impact	<ul style="list-style-type: none"> <li>○ Faulty regulatory submissions, leading to considerable legal exposure</li> <li>○ Bad Data quality leading to customer OR shareholder impact, could result in lawsuits</li> </ul>
Decision impact	<ul style="list-style-type: none"> <li>○ Quality of decision depends on quality of data. Bad data quality leads to misinformed or under-informed decisions</li> </ul>
Business Management impact	<ul style="list-style-type: none"> <li>○ Lack of data on the key performance indicators, hampers objective performance management</li> </ul>

## 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 the data quality might be poor. But all data might not be cleaned immediately. Focus on the immediate problems which might be caused by poor data. Once the problem area is identified then following methods can used to analyze, identify, clean the data.

- **Data Profiling**

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

- **Data Cleansing and Enhancement**

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

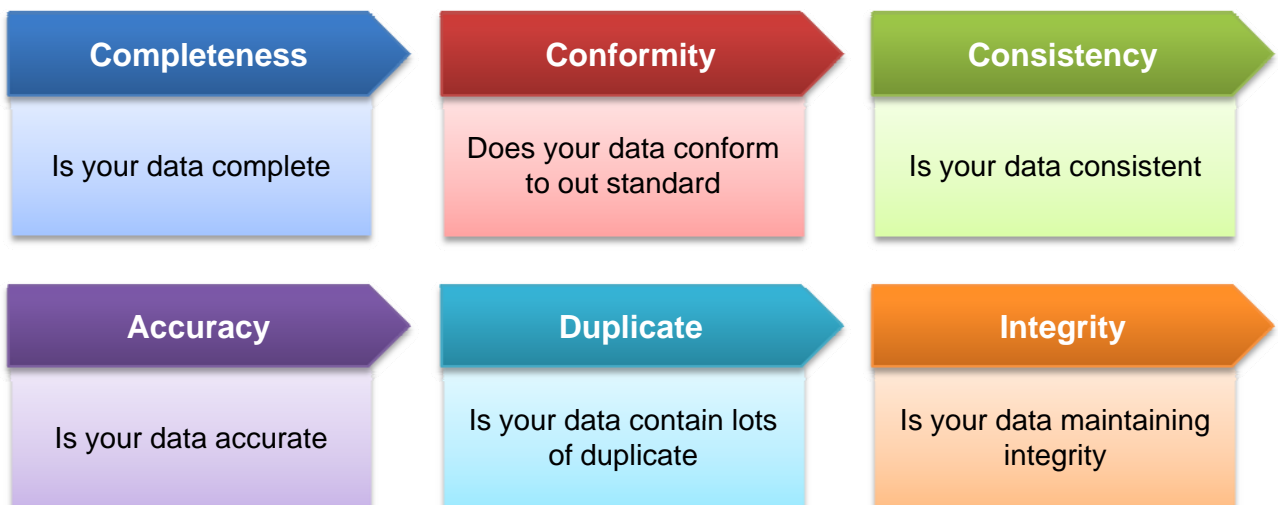
- **Data Matching and Consolidation**

Match similar records and perform de-duplication and consolidation based on set criteria. Matching is done on various business rules like name, address, SSN, DUNS. Once the duplication is determined, merge is performed on duplicate records if they are same identity. This directly benefits into removal of duplicate from your database, Improve the accuracy of Customer Information in Customer database

## How to prevent Toxic Data and manage Data Quality

90% of the toxic data enters at various data entry point in an organization. If bad data entry can be prevented at source checkpoint organization can save a significant amount later spent in data correction and detoxification. Entering your data correctly for the first time is the best way to ensure the integrity of your data. Be prepared to spend money at this data collection stage. It saves more money in long term. Use auto complete or other data validation application at the data entry data stage. Using controls and input masks during data entry can 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. Repeat the data quality process in an iterative fashion 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. Here are the steps involved in building a data quality scorecard to summarize the overall cost associated with low data quality and help identify the best opportunities for improvement:

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 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 purpose”. Data as an asset must be usable by organization 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 solution for the most important asset of your organization – Data.

## 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](mailto:info@scalable-systems.com)

Web: [www.scalable-systems.com](http://www.scalable-systems.com)

Copyright © 2008 Scalable Systems. All Rights Reserved.

While every attempt has been made to ensure that the information in this document is accurate and complete, some typographical errors or technical inaccuracies may exist. Scalable Systems does not accept responsibility for any kind of loss resulting from the use of information contained in this document. The information contained in this document is subject to change without notice. Scalable Systems logos, and trademarks are registered trademarks of Scalable Systems or its subsidiaries in the United States and other countries. Other names and brands may be claimed as the property of others. Information regarding third party products is provided solely for educational purposes. Scalable Systems is not responsible for the performance or support of third party products and does not make any representations or warranties whatsoever regarding quality, reliability, functionality, or compatibility of these devices or products.

This edition published August 2008