Information Builders

Implementing Real-Time Data Quality Management

The Best Solution for Fixing Bad Data

A White Paper

by Vincent Lam

WebFOCUS

iWay Software

Vincent Lam Vincent Lam is the product marketing director responsible for marketing iWay Software's entire product line.

Mr. Lam has a diverse background in the technology sector and has helped position iWay Software's products as best-of-breed in the competitive marketplace. Earlier in his career at the company, Mr. Lam was a strategic product manager. He introduced new strategic and innovative products to the marketplace such as WebFOCUS Magnify, the world's first real-time transactional enterprise search solution. Mr. Lam's background includes innovation at Wall St. firms, technology firms, and entrepreneurship.

Table of Contents

Executive Summary

Poor data quality is expensive. In fact, The Data Warehouse Institute (TDWI) estimates that bad information costs businesses close to $600 million each year.[1] As a result, companies are embarking on broad-scale data quality management (DQM) initiatives at a rapid pace.

Dirty data's negative impact knows no bounds. It results in poor customer service, missed sales opportunities, and lost revenue, and prevents the formulation of sound business strategies. PricewaterhouseCoopers has quantified the potential risks, claiming that without a data quality strategy:

- 18 percent of companies lack the desired efficiency in their CRM and marketing activities
- 29 percent of companies experience income loss as a result of erroneous billing
- 24 percent of companies waste time and effort on financial information processing
- 53 percent of companies face serious problems when implementing new systems and applications[2]

It is no secret that the most successful businesses are those that effectively and strategically generate, use, and reuse information. But for that data to be properly reused, it must be valid, accurate, and consistent; in a format that makes it easy to access and use; and a true reflection of the overall state of the subject area it describes, or the business as a whole.

That's where data quality management comes in. But not all techniques, or technologies, are created equal. Too few organizations have a formal, well-defined data quality plan in place. In fact, according to a PricewaterhouseCoopers study, only 12 percent of companies polled cited the existence of documented data quality policies.[3]

Not only do more organizations need to take data quality seriously; the companies that have embarked on data quality management initiatives may not be going about it the right way. While many organizations use batch methods or near real-time approaches, only true real-time data quality management can fully optimize information integrity across the enterprise.

Businesses need an approach that facilitates data quality management in a continuous, real-time fashion to catch corrupt or invalid information before it wreaks havoc on corporate systems. Enterprises that fail to catch problems until it is too late are forced to fight their data quality fires with outdated, complex, or inadequate tools.

In this white paper, we will compare the real-time approach to data quality management with higher-latency approaches. We will demonstrate how real-time data quality efforts are far more effective than other techniques when it comes to minimizing wasted time and money, as well as other damages, that can be inflicted as a result of inaccurate or incomplete information. We will also highlight how one company – Sabre Holdings – realized substantial advantages through the implementation of real-time data quality management.

[1] Eckerson, Wayne W. "Data Quality and the Bottom Line," The Data Warehousing Institute (TDWI), February 2002.

[2] "Data Quality Management," PriceWaterhouseCoopers, April 2011.

[3] "Data Quality Management," PriceWaterhouseCoopers, April 2011.

Real-Time vs. High-Latency Data Quality

Real-time data quality is the most effective way to minimize the financial and operational risk associated with bad information. However, it is not the method most companies use to identify and correct corporate data problems.

First There Was Batch

The traditional – and most commonly used – method of data quality management is the batch approach. One of the key reasons it is so popular is because of the legacy systems that still exist at many companies. In most scenarios, data quality tools are incorporated directly into extract, transform, and load (ETL) processes. Data is evaluated and cleansed as it is moved from databases and other source systems into data marts or data warehouses.

This technique, while widely relied upon, is hardly ideal, because it creates unacceptable levels of risk exposure. Let's assume that most companies move batches of data from source systems to repositories on a nightly basis. That means end users could potentially be using bad data to support key operations or critical decisions for up to 24 hours.

Sub-par information tends to permeate throughout an organization, causing problems as it moves from system to system and from person to person. The severity of data quality issues is likely to multiply at an astonishing rate until invalid records are detected and corrected.

Near Real-Time Doesn't Solve the Problem

Some companies have attempted to combat this problem with near real-time methods of managing data quality. Data is pushed to data quality tools for assessment and cleansing at specified intervals. Because those intervals are much shorter than those used in batch processes (the time between feeds could be as short as an hour), this is a significant improvement.

But there is still much exposure. In today's dynamic business world, a lot can happen in just one hour. Imagine a very busy call center, one that fields hundreds of calls each hour. During 60 minutes, a few small pieces of bad customer data could result in poor responses to questions and issues, and sub-par service delivery to dozens of callers. The accounting department, for example, may gather that same inaccurate information as it generates invoices, resulting in incorrect billing. Or the marketing team might use it to launch a major promotion, and ultimately waste precious budget funds and hinder campaign results.

Real-Time Data Quality Management: The Only Answer

Since major problems can arise the moment a single bad record hits a database, any approach to data quality management must be totally proactive and instantaneous. Corrupt or invalid data must be caught and rectified as it is being entered, so that inaccurate information never permeates the environment in the first place.

Any data quality strategy, as well as the tools that support it, must be able to monitor and evaluate each and every channel through which data enters the organization. This includes manual data entry, as well as data collected via automated processes and transactions conducted with third parties such as partners and customers. Predefined business rules that clearly demonstrate what constitutes bad data must also be applied, to catch any problems or issues as information is captured.

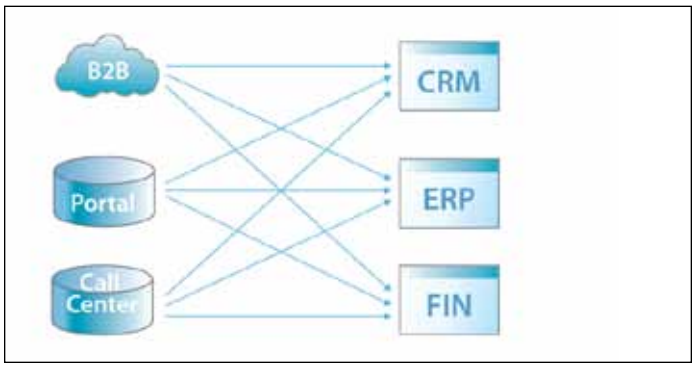
This real-time approach, in which the management of corporate information integrity is continuous and immediate, is the only way to truly minimize exposure and maximize advantage.

Applying Real-Time Quality Management Across the Data Life Cycle

Enterprise data flows into, across, and outside an organization during the course of day-to-day business activities. As it moves through this cycle, its integrity must be managed in real time.

Upstream

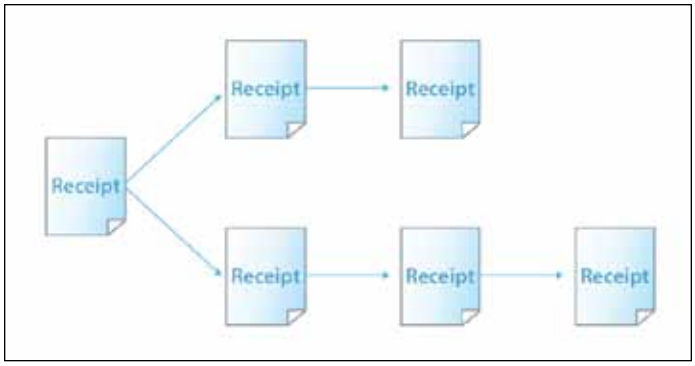
As data flows upstream and is introduced into the environment – whether typed in manually by an internal user, collected via a business-to-business (B2B) exchange with a partner, or entered by a customer via a self-service portal – it must be accurately assessed. A tremendous amount of bad data is created at the time of inception. And the emergence of new information channels and an increasing number of data touch points have exacerbated the problem.



The damage that can be caused is virtually limitless if bad data is allowed to make its way into corporate systems. That's why automated business rules and data quality standards must be applied at the point of origin.

Instream

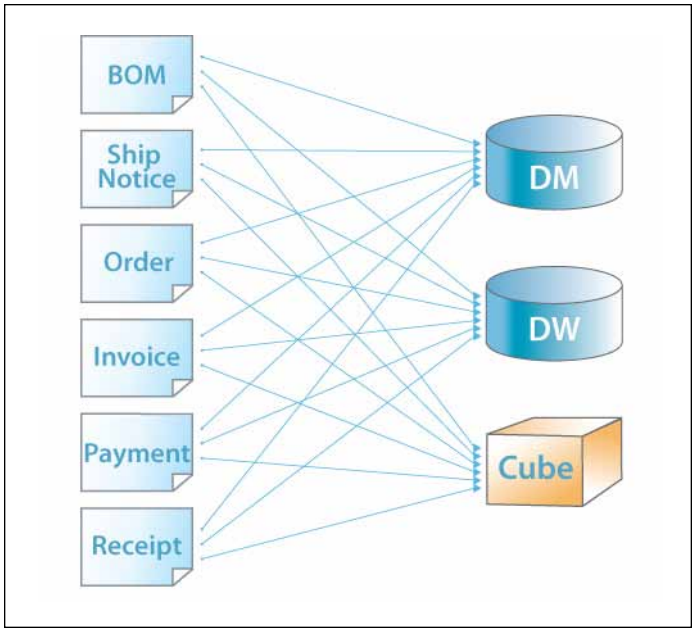
During the course of business transactions or for reporting and analysis, information moves instream, from user to user across an organization. As this data is consumed, it is often modified, extended, or combined with other records. If data is corrupted by one user, it will negatively impact many other users and processes that depend on it. For example, a single incorrect sales order will also result in an inaccurate bill of materials, shipping notice, packing slip, and invoice.



Faulty instream data must be prevented at all costs. Active, real-time checks and balances must be in place to stop information from becoming duplicated, mismatched, or misplaced. Once an error finds its way into the infrastructure, it will be difficult to stop its momentum.

Downstream

Reporting and analysis are critical tasks for organizations of all types and sizes. Information from a wide array of sources – data warehouses, data marts, multidimensional cubes, and back-end applications – is retrieved countless times each day to support strategic and tactical planning, as well as the execution of mission-critical operations.



If this information is redundant, inconsistent, hard to access, or just wrong, it can hinder operational efficiency, corporate performance, and profitability. Further problems can be created if that information is shared with or distributed to outside constituents, such as regulatory agencies, customers, or partners.

Every stage in the data life cycle creates an opportunity for bad data to propagate. Given the rate in which corrupt data can multiply, letting in just one invalid record may necessitate the location and correction of dozens more if it moves instream or downstream.

If that inaccurate information is stopped as it flows back upstream or instream, there is only one piece of bad data to rectify. Effort and cost will be minimized, and the integrity of the information environment will be fully preserved. As a result, the critical business processes that are driven by that information will be far more reliable and efficient.

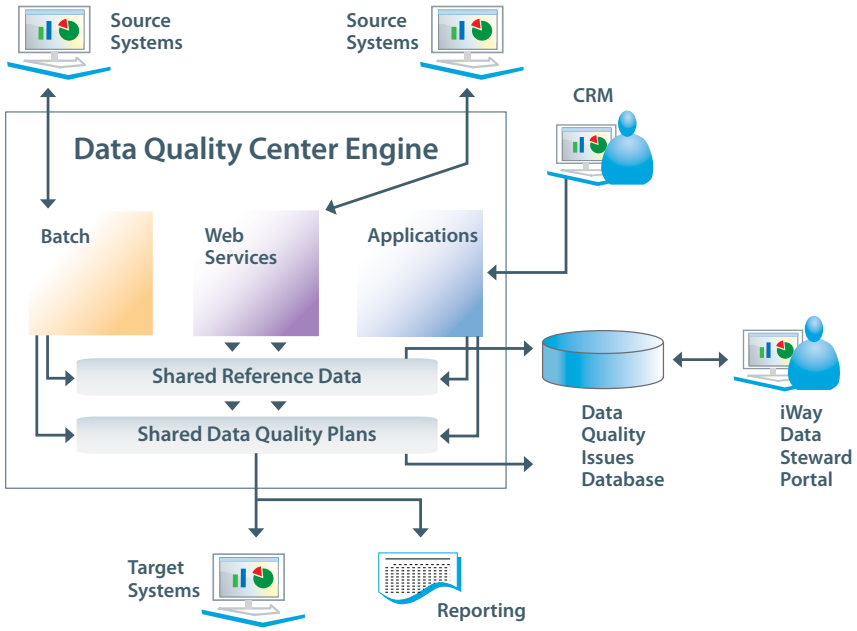
Implementing Real-Time Data Quality Management

We've discussed how important real-time data quality management is. But what is required to implement and enforce it?

For real-time data quality management to be truly successful, it must be broad-reaching and must touch the entire information infrastructure. It cannot be optional, nor can it be applied to only a handful of systems. It must be an integral part of the environment, involving every component in the architecture, as well as every user and automated process that creates and consumes data.

Data quality processes also must be part of all back-end databases. Correcting records only as they are moved into a data warehouse, data mart, or other repository will address just a fraction of the problems. Dirty data will still be left in its underlying source, creating risk for any user who accesses and uses it.

Data integrity issues are never confined to one system. One corrupt record will undoubtedly infect other systems as it moves upstream, downstream, and instream. The longer it takes to fix, the greater the problem will grow. That means more time, effort, and money will be required to clean the data once it is uncovered.



The key is to design and deploy data quality services that are reusable, so they can be exposed to all applications and systems in the environment. The same data quality policies and supporting tools – all governed under a single plan – should be used across the organization, regardless of the level of latency. For example, if real-time data quality is being applied to information as it is being captured, and batch data quality is leveraged as that data is moved to a reporting repository, the same techniques and procedures should be used to cleanse bad information when it is found.

Real-Time Data Quality Management at Sabre Holdings

Sabre Holdings Corporation is a Texas-based company that provides products, distribution, and technology solutions to the travel industry. One of its offerings, the Airport Data Intelligence (ADI) system, helps Sabre's clients to study airline traffic patterns and other trends over a seven-year period. That insight is used to predict flight capacities, plan connections, and more.

Hundreds of customers depend on Sabre ADI as an accurate, online source of demand and schedule information, so data quality is of the utmost importance. Each month, ADI is populated with seven to eight million rows of historic and future passenger booking information derived from various booking and ticketing data sources, government entities, and the International Air Transport Association (IATA) Billing and Settlement Plan. The value of this information is inestimable, since it will impact the routes, schedules, and even the fares that travelers depend on when they book a trip. Yet, in spite of manual quality assurance processes conducted by Sabre employees, the database was plagued with discrepancies and other issues.

To eliminate these problems, Sabre implemented Information Builders' iWay Data Quality Center (DQC), helping the company take a proactive, real-time approach to data quality management. iWay DQC serves as a centralized management hub for business rules, data quality, and data flows, dynamically and instantly identifying and correcting common errors as information is collected from the various third-party sources, and before it is loaded into the ADI database. As a result, all data is scrubbed, audited, and certified before customers ever see it.

Sabre is now much more confident about the integrity of the information contained in ADI. Substantial increases in data accuracy and reliability have resulted in improved customer satisfaction. The company has also improved staff efficiency by eliminating manual data quality activities.

Conclusion

Implementing any type of data quality management program is a step in the right direction, but without the right kind of data quality policies and procedures companies will leave themselves open to significant risk, and fail to realize the true value of their efforts.

Most popular methods of data quality management, such as batch and near real time, will fail to completely eliminate data integrity issues – still leaving organizations in jeopardy. While these approaches are effective at catching bad data, they often do so way too late in the game. Damage has already been caused in other systems, and the cleanup effort will likely be tremendous.

Only real-time data quality management can keep corporate systems fully protected from inaccurate, incomplete, or invalid information. By proactively catching bad data elements before they reach their source, companies can truly ensure the consistency and accuracy of all data across their enterprise.

Worldwide Offices

Corporate Headquarters

Two Penn Plaza
New York, NY 10121-2898
(212) 736-4433
(800) 969-4636

United States

Atlanta, GA* (770) 395-9913
Baltimore, MD (703) 247-5565
Boston, MA* (781) 224-7660
Channels (770) 677-9923
Chicago, IL* (630) 971-6700
Cincinnati, OH* (513) 891-2338
Dallas, TX* (972) 398-4100
Denver, CO* (303) 770-4440
Detroit, MI* (248) 641-8820
Federal Systems, DC* (703) 276-9006
Florham Park, NJ (973) 593-0022
Gulf Area (972) 490-1300
Hartford, CT (781) 272-8600
Houston, TX* (713) 952-4800
Kansas City, MO (816) 471-3320
Los Angeles, CA* (310) 615-0735
Milwaukee, WI (414) 827-4685
Minneapolis, MN* (651) 602-9100
New York, NY* (212) 736-4433
Orlando, FL (407) 804-8000
Philadelphia, PA* (610) 940-0790
Phoenix, AZ (480) 346-1095
Pittsburgh, PA (412) 494-9699
Sacramento, CA (916) 973-9511
San Jose, CA* (408) 453-7600
Seattle, WA (206) 624-9055
St. Louis, MO* (636) 519-1411, ext. 321
Washington DC* (703) 276-9006

International

Australia*
Melbourne 61-3-9631-7900
Sydney 61-2-8223-0600

Austria Raffeisen Informatik Consulting GmbH
Wien 43-1-211-36-3344

Bangladesh
Dhaka 415-505-1329

Belgium*
Brussels 32(0)2-743-02-40

Brazil InfoBuild Brazil Ltda.
São Paulo 55-11-3285-1050

Canada
Calgary (403) 437-3479
Montreal* (514) 421-1555
Ottawa (613) 233-7647
Toronto* (416) 364-2760
Vancouver (604) 688-2499

China
Beijing 010-51289680, ext. 8010

Croatia InfoBuild CEE
Strmec Samoborski 385-1-23-62-400

Czech Republic InfoBuild CEE
Praha 420-221-986-460

Estonia InfoBuild Baltics
Tallinn 372-5265815

Finland InfoBuild Oy
Espoo 358-207-580-840

France*
Sèvres +33 (0)1-45-07-66-00

Germany
Eschborn* 49-6196-775-76-0

Greece Applied Science Ltd.
Athens 30-210-699-8225

Guatemala IDS de Centroamerica
Guatemala City (502) 2412-4212

Hungary InfoBuild CEE
Budapest 36-1-430-3500

India* InfoBuild India
Chennai 91-44-42177082

Israel Malam Team – SRL Products
Petah-Tikva 972-3-7662040

Italy
Milan 39-02-92-349-724

Japan KK Ashisuto
Tokyo 81-3-5276-5863

Kuwait InfoBuild Middle East
Safat 965-2-232-2926

Latvia InfoBuild Baltics
Riga 371-67039637

Lebanon InfoBuild Middle East
Beirut 961-4-533162

Lithuania InfoBuild Baltics
Vilnius 370-5-268-3327

Mexico
Mexico City 52-55-5062-0660

Netherlands*
Amstelveen 31 (0)20-4563333

Nigeria InfoBuild Nigeria
Garki-Abuja 234-803-318-4750

Norway InfoBuild Norge AS
Oslo 47-4820-4030

Poland InfoBuild CEE
Warszawa 48-22-657-0014

Portugal
Lisboa 351-217-217-400

Qatar InfoBuild Middle East
Doha 974-4-466-6244

Russian Federation InfoBuild CIS
Moscow 7-495-797-20-46
■ Armenia ■ Azerbaijan ■ Belarus ■ Kazakhstan
■ Kyrgyzstan ■ Moldova ■ Tajikistan
■ Turkmenistan ■ Ukraine ■ Uzbekistan

Saudi Arabia InfoBuild Middle East
Riyadh 966-1-479-7623

Singapore Automatic Identification Technology Ltd.
Singapore 65-6286-2922

Slovakia InfoBuild CEE
Bratislava 421-232-332-513
■ Bulgaria ■ Romania ■ Serbia ■ Slovenia

South Africa Fujitsu (Pty) Ltd.
Cape Town 27-21-937-6100
Johannesburg 27-11-233-5432

South Korea Uvansys
Seoul 82-2-832-0705

Spain
Barcelona 34-93-452-63-85
Bilbao 34-94-452-50-15
Madrid* 34-91-710-22-75

Sweden InfoBuild AB
Solna 46-8-578-772-01

Switzerland
Dietlikon 41-44-839-49-49

Taiwan Galaxy Software Services, Inc.
Taipei (866) 2-2586-7890

Thailand Datapro Computer Systems Co. Ltd.
Bangkok 66(2) 301 2800

Turkey InfoBuild Turkey
Ankara 90-312-266-3300
Istanbul 90-212-351-2730

United Arab Emirates InfoBuild Middle East
Abu Dhabi 971-2-627-5911
■ Bahrain ■ Egypt ■ Jordan ■ Oman
Dubai 971-4-391-4394

United Kingdom*
Uxbridge Middlesex 0845-658-8484

Venezuela InfoServices Consulting
Caracas 58212-763-1653

* Training facilities are located at these offices.

Information Builders

Corporate Headquarters Two Penn Plaza, New York, NY 10121-2898 (212) 736-4433 Fax (212) 967-6406 DN3601490.1111
Connect With Us informationbuilders.com askinfo@informationbuilders.com