Top 10 Root Causes of Data Quality Problems

White Paper

Table of Contents

We all know data quality problems when we see them. They can undermine your organization's ability to work efficiently, comply with government regulations and make revenue. The specific technical problems include missing data, misfielded attributes, duplicate records and broken data models to name just a few.

But rather than merely patching up bad data, most experts agree that the best strategy for fighting data quality issues is to understand the root causes and put new processes in place to prevent them. This white paper discusses the top ten root causes of data quality problems and suggests steps the business can implement to prevent them.

Typographical Errors and Non-Conforming Data

Despite a lot of automation in our data architecture these days, data is still typed into Web forms and other user interfaces by people. A common source of data inaccuracy is that the person manually entering the data just makes a mistake. People mistype. They choose the wrong entry from a list. They enter the right data value into the wrong box.

Given complete freedom on a data field, those who enter data have to go from memory. Is the vendor named Grainger, WW Granger, or W. W. Grainger? Ideally, there should be a corporate-wide set of reference data (metadata) so that forms help users find the right vendor, customer name, city, part number, and so on.

Business Strategy

- Training – Make sure that those people who enter data know the impact they have on downstream applications.

- Metadata Definitions – By locking down exactly what people can enter into a field using a definitive list, many problems can be alleviated. This metadata (for vendor names, part numbers, and so on can) become part of data quality in data integration, business applications and other solutions.
- Monitoring – Make public the results of poorly entered data and praise those who enter data correctly. You can keep track of this with data monitoring software such as the Talend Data Quality Portal.
- Real-time Validation – In addition to forms, validation data quality tools can be implemented to validate addresses, e-mail addresses and other important information as it is entered. Ensure that your data quality solution provides the ability to deploy data quality in application server environments, in the cloud or in an enterprise service bus (ESB).

Information Obfuscation

Data entry errors might not be completely by mistake. How often do people give incomplete or incorrect information to safeguard their privacy? If there is nothing at stake for those who enter data, there will be a tendency to fudge.

Even if the people entering data want to do the right thing, sometimes they cannot. If a field is not available, an alternate field is often used. This can lead to such data quality issues as having Tax ID numbers in the name field or contact information in the comments field.

Business Strategy

- Reward – Offer an incentive for those who enter personal data correctly. This should be focused on those who enter data from the outside, like those using Web forms. Employees should not need a reward to do their job. The type of reward will depend upon how important it is to have the correct information.
- Accessibility – As a technologist in charge of data stewardship, be open and accessible about criticism from users. Give them a voice when processes change requiring technology change. If you're not accessible, users will look for quiet ways around your forms validation.
- Real-time Validation – In addition to forms, validation data quality tools can be implemented to validate addresses, e-mail addresses and other important information as it is entered.

Renegade IT and Spreadmarts

A renegade is a person who deserts and betrays an organizational set of principles. That's exactly what some impatient business owners unknowingly do by moving data in and out of business solutions, databases and the like. Rather than wait for some professional help from IT, eager business units may decide to create their own set of local applications without the knowledge of IT. While the application may meet the immediate departmental need, it is unlikely to adhere to standards of data, data model or interfaces. The database might start by making a copy of a sanctioned database to a local application on team desktops. So-called "spreadmarts," which are important pieces of data stored in Excel spreadsheets, are easily replicated to team desktops. In this scenario, you lose control of versions as well as standards. There are no backups, versioning or business rules.

Business Strategy

- Corporate Culture – There should be a consequence for renegade data, making it more difficult for the renegades to create local data applications.
- Communication – Educate and train your employees on the negative impact of renegade data.
- Small Data Management – Having tools around that can help business users and IT professionals alike manage the data is crucial. Solutions like Talend Master Data Management (MDM) help bridge this gap between expensive and clunky IT applications and effective business management of data.
- Locking Down the Data – A culture where creating unsanctioned spreadmarts is shunned is the goal. Some organizations have found success in locking down the data to make it more difficult to export.

After the Merger

Corporate mergers increase the likelihood for data quality errors because they usually happen fast and are unforeseen by IT departments. Almost immediately, there is pressure to consolidate and take shortcuts on proper planning. The consolidation will likely include the need to share data among a varied set of disjointed applications. Many shortcuts are taken to "make it happen," often involving known or unknown risks to the data quality.

On top of the quick schedule, merging IT departments may encounter culture clash and a different definition of truth. Additionally, mergers

can result in a loss of expertise when key people leave midway through the project to seek new ventures.

Business Strategy

- Corporate Awareness – Whenever possible civil division of labor should be mandated by management to avoid culture clashes and data grabs by the power hungry.
- Document – Your IT initiative should survive even if the entire team leaves, disbands or gets hit by a bus when crossing the street. You can do this with proper documentation of the infrastructure.
- Third-party Consultants – Management should be aware that there is extra work to do and that conflicts can arise after a merger. Consultants can provide the continuity needed to get through the transition.
- Agile Data Management – Talend's open source data management solutions will keep your organization agile, giving you the ability to divide and conquer the workload without expensive licensing of commercial applications.

Change is Good… Except for Data Quality

Organizations undergo business process change to improve. Good, right? Prime examples include:

- Company expansion into new markets
- New partnership deals

- New regulatory reporting laws
- Financial reporting to a parent company
- Downsizing

If data quality is defined as "fitness for purpose," what happens when the purpose changes? It's these new data uses that bring about changes in perceived level of data quality even though underlying data is the same. It's natural for data to change. As it does, the data quality rules, business rules and data integration layers must also change.

Business Strategy

- Data Governance – By setting up a cross-functional data governance team, you will always have a team who will be looking at the changes your company is undergoing and considering its impact on information. In fact, this should be in the charter of a data governance team.
- Communication – Regular communication and a well-documented metadata model will make the process of change much easier.
- Tool Flexibility – One of the challenges of buying data quality tools embedded within enterprise applications is that they may not work in all enterprise applications. When you choose tools, make sure they are flexible enough to work with data from any application and that the company is committed to flexibility and openness.

Hidden Code

Databases rarely begin their life blank. The starting point is typically a data conversion from some previously existing data source. The problem is that while the data may work perfectly well in the source application, it may fail in the target. It's difficult to see all the custom code and special processes that happen beneath the data unless you profile.

Business Strategy

- Profile Early and Often – Don't assume your data is fit for purpose because it works in the source application. Profiling will give you an exact evaluation of the shape and syntax of the data in the source. It also will let you know how much work you need to do to make it work in the target.
- Apply Data Quality Tools When Possible – Rather than custom code in the application, a better strategy is to let data quality tools apply standards. Data quality tools will apply corporate standards in a uniform way, leading to more accurate sharing of data.

Transaction Transition

More and more data is exchanged between systems through real-time (or near real-time) interfaces. As soon as the data enters one database, it triggers procedures necessary to send transactions to other downstream databases. The advantage is immediate propagation of data to all relevant databases.

However, what happens when transactions go awry? A malfunctioning system could cause problems with downstream business applications. In fact, even a small data model change could cause issues.

Business Strategy

- Schema Checks – Employ schema checks in your job streams to make sure your real-time applications are producing consistent data. Schema checks will do basic testing to make sure your data is complete and formatted correctly before loading.
- Real-time Data Monitoring – One level beyond schema checks is to proactively monitor data with profiling and data monitoring tools. Tools like the Talend Data Quality Portal will ensure the data contains the right kind of information. For example, if your part numbers are always a certain shape and length, and contain a finite set of values, any variation on that attribute can be monitored. When variations occur, the monitoring software can notify you.

Metadata Metamorphosis

Metadata repository should be able to be shared by multiple projects, with audit trail maintained on usage and access. For example, your company might have part numbers and descriptions that are universal to CRM, billing, ERP systems, and so on. When a part number becomes obsolete in the ERP system, the CRM system should know. Metadata changes and needs to be shared.

In theory, documenting the complete picture of what is going on in the database and how various processes are interrelated would allow you to completely mitigate the problem. Sharing the descriptions and part numbers among all applicable applications needs to happen. To get

started, you could then analyze the data quality implications of any changes in code, processes, data structure, or data collection procedures and thus eliminate unexpected data errors. In practice, this is a huge task.

Business Strategy

- Predefined Data Models – Many industries now have basic definitions of what should be in any given set of data. For example, the automotive industry follows certain ISO 8000 standards. The energy industry follows Petroleum Industry Data Exchange standards or PIDX. Look for a data model in your industry to help.
- Agile Data Management – Data governance is achieved by starting small and building out a process that first fixes the most important problems from a business perspective. You can leverage agile solutions from Talend to share metadata and set up optional processes across the enterprise.

Defining Data Quality

More and more companies recognize the need for data quality, but there are different ways to clean data and improve data quality. You can:

- Write some code and cleanse manually
- Handle data quality within the source application
- Buy tools to cleanse data

However, consider what happens when you have two or more of these types of data quality processes adjusting and massaging the data. Sales has one definition of customer, while Billing has another. Due to differing processes, they don't agree on whether two records are a duplicate.

Business Strategy

- Standardize Tools – Choose tools that aren't tied to a particular solution. Having data quality only in SAP, for example, won't help your Oracle, Salesforce and MySQL data sets. When picking a solution, select one like Talend Data Quality that is capable of accessing any data, anywhere, at any time. While Talend's solution is versatile, it won't cost you a bundle to leverage a common solution like this across multiple platforms and solutions.
- Data Governance – By setting up a cross-functional data governance team, you will have the people in place to define a common data model.

Loss of Expertise

On almost every data intensive project, there is one person whose legacy data expertise is outstanding. These are the folks who understand why some employee date of hire information is stored in the date of birth field and why some of the name attributes also contain tax ID numbers.

Data might be a kind of historical record for an organization. It might have come from legacy systems. In some cases, the same value in the

same field will mean a totally different thing in different records. Knowledge of these anomalies allows experts to use the data properly.

If you encounter this situation, there are some business processes you can follow.

Business Strategy

- Profile and Monitor – Profiling the data will help you identify most of these types of issues. For example, if you have a tax ID number embedded in the name field, analysis will let you quickly spot it. Monitoring will prevent a recurrence.
- Document – Although they may be reluctant to do so for fear of losing job security, make sure experts document all of the anomalies and transformations that need to happen every time the data is moved.
- Use Consultants – Expert employees may be so valuable and busy that there is no time to document the legacy anomalies. Outside consulting firms are usually very good at documenting issues and providing continuity between legacy and new employees.

Conclusion

Today, most businesses realize that their success is increasingly tied to the quality of their information. Companies rely on data to make significant decisions that can affect customer service, regulatory compliance, supply chain and many other areas. As you collect more and more information about customers, products, suppliers, transactions and billing, you must attack the root causes of data quality. Data quality tools are just that, tools to attack problems at their root and resolve issues if they slip through. The tools work in

combination with people and process to drive high-quality, high-value information.

Want to learn more about open source Talend Data Quality? Watch online webinars or download the latest open source data management software at Talend.com.

About Talend

Talend is the recognized market leader in open source data management. A single development studio provides consistency across integration, quality and master data projects so resources can be shared and reutilized, all the while remaining open, intuitive and economical. Tackling the challenges of data management doesn't need to be overly expensive or locked to one application. Talend shatters the traditional proprietary model by providing open source technology that is proven on the basis of performance, ease of use, extensibility and robustness.

Talend's data quality products provide organizations with detailed insight and monitoring of any business data and include a comprehensive set of features to improve the quality and effectiveness of critical data assets.

Talend offers two data quality options: Talend Open Profiler, an open source data profiling product, which is available on the Talend Web site for free download, and Talend Data Quality, which includes additional enterprise-grade features including cleansing and standardization, matching and de-duplication, an integrated data integration tool for quick and easy data transformations, and Web-based data quality dashboards.