

IT'S ALL ABOUT THE DATA

Martin D. Solomon

Data problems come about when little regard is paid to the consequences of data replication. As IT infrastructure costs decline, investing in the accuracy and currency of the corporate data asset will become paramount.

OVER 100 YEARS AGO, MARK TWAIN coined the phrase “everyone talks about the weather, but no one does anything about it.” Applying that phrase to 21st century technology, it frequently seems that organizations talk about their data quality problems, issues, and concerns but not many do anything about it. This is unfortunate because, unlike the weather, we do have the methodology, models, software, and infrastructure available to us to actually do something about it. In addition, as machines get faster and storage gets cheaper, and the information technology (IT) infrastructure becomes more of a commodity instead of an advantage, the need to be proactive in keeping data clean and healthy becomes more imperative than ever. So it is inevitable that the principal differentiating factor between competing corporations could become considerably data focused. Many shortcomings in the enterprise IT architecture can be overcome or at least mitigated at an expense, with more hardware to make applications run faster and more software to make them behave more efficiently. In regard to data content and data quality, the same axiom is not that easily applied. Strong data stewardship and process changes must be coupled with all expenditures to improve data quality through software and hardware infrastructure additions. This expense, however, is commonly anticipated as being much smaller than what ultimately reaches the bottom line. The false expectation can be attributed to the background reverberations of poor data governance and

lack of data discipline throughout the organization.

EVENTS AND DATA

A significant portion of data stored today is the representation of an event, whether scientific, financial, or otherwise. The event may be a transaction describing a bank withdrawal, a claim payment, a customer service phone call, the sale of a product, or the temperature at a given location, date, and time. Even data that is a “non-event,” such as the elements representing a “customer name,” were initially entered as part of an event. Whether manually typed in at company headquarters, by a new customer over the Internet, or through a batch load, it came into existence as part of an experience with supporting event data points. To make decisions based on these events, whether tactical or strategic, it is paramount that *all* of the data representing these events is consistent and accurate and in a limited number of locations. We know through experience that sometimes data must be copied or duplicated to allow for easier access due to a physical limitation. When data replication is required, it is imperative to maintain the integrity of the original event by maintaining the original content and context of the data describing it. A guideline to live by is that an environment with too much data replication is often costly and unwieldy, but one that has data that is replicated and different is often devastating.

MARTIN D. SOLOMON, Architecture Director for CIGNA, Inc., is responsible for the overall architecture and design of Information Management and Customer Acquisition initiatives. Solomon previously was a director in a large international consultancy. He can be reached at marty.solomon@cigna.com.

This flow of events is akin to the old elementary school game of “telephone,” where through each handoff the context and contents of the first statement...get lost to a point where a different message is delivered.

Businesses get into untenable and sometimes nearly intractable data problems when data is replicated and the content is changed such that the original event it describes is not wholly maintained (Redman, 2001). While it is justified to extract, transform, and load (ETL) data from normalized source systems into denormalized data structures for analytic purposes, discipline and governance must be in place to ensure that data elements do not get removed or modified in the process. This discipline must be maintained when other users in the organization move enterprise analytic data into their own localized file server databases (e.g., Access) or desktop spreadsheet environment. Unfortunately, when this occurs and the data gets further removed from the “book of record” or original source, the laws of entropy are usually overwhelming and the data becomes in whole or in part modified and separated from the initial event it described.

A primary justification given for replicating and migrating data several generations removed from its source and in an ungoverned fashion (and to assume the risks that go with that process) is that the IT enterprise cannot accommodate these end-user needs. The common business case behind this is that the resources required for building and designing the appropriate replicated data structures such as additional or extended data marts or fact and dimension tables are not cost justified. Unfortunately, these cost justifications are simple and usually only point to a comparison of the extra storage costs (disk) versus the design and programming effort to code and implement a more robust solution. Not brought to the forefront are the extensive hidden costs of poor data quality and expanded infrastructure costs precipitated by the organization’s continuous decision-making process that supports data redundancy. It is therefore critical that a comprehensive cost-benefit analysis (CBA) be performed for each data replication request based on the full set of requirements and the hardware, software, programming time, and the long-term risk and cost of *not implementing* structured analytic data stores in a governed environment.

Figure 1 displays the process through which a set of actions that individually appear innocent enough causes the data describing the original event to be misrepresented or misinterpreted. In this example, a customer contacts customer service to request that four bounced check fees be removed from his personal account. The bank representative sees

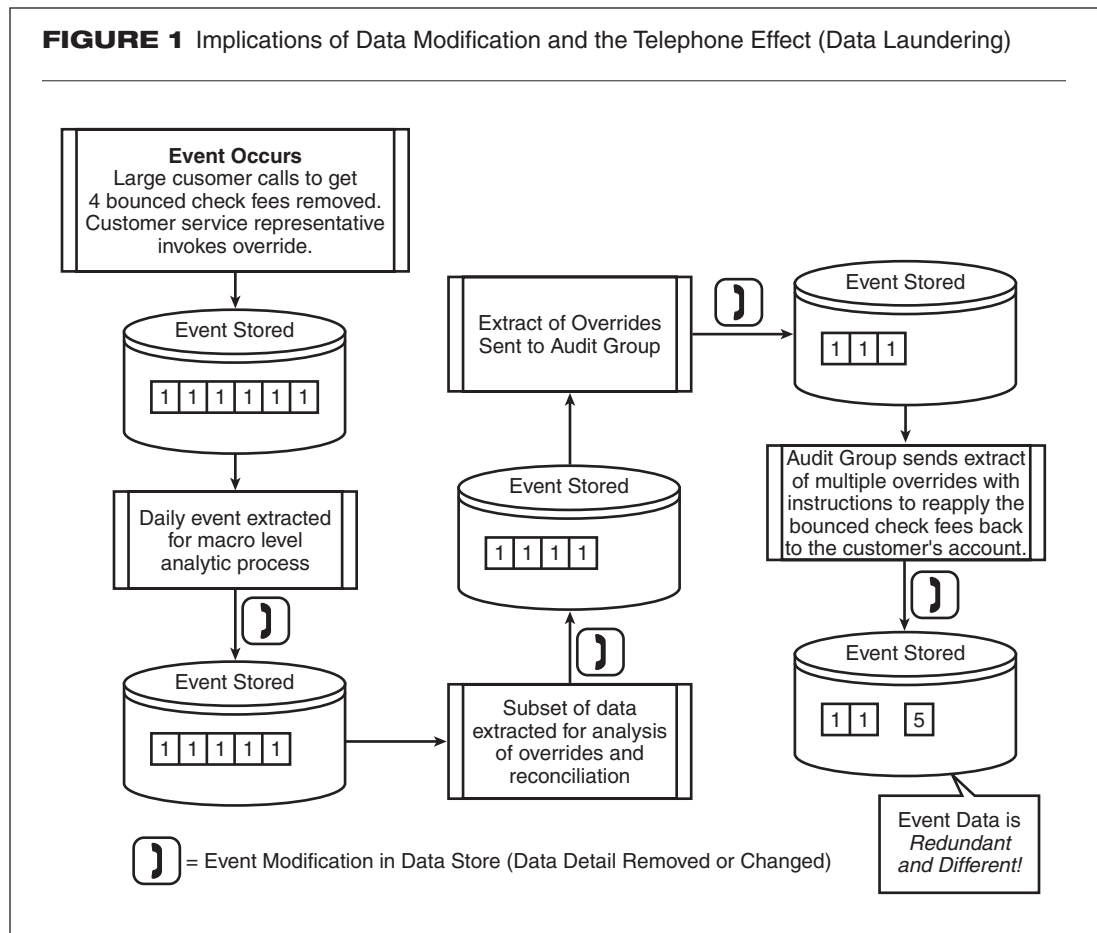
that these fees are probably legitimate and that the customer’s request is a bit of a stretch and never usually performed, but the customer also points out that he is also a very large commercial customer that the bank is privileged to have. The customer service representative reverses the fees and the activity is recorded. Downstream from this event, a portion of this activity data is extracted nightly to a system used for analytic purposes. From the analytic system another portion of the daily event data is then extracted for auditing purposes. After this, yet another partial data extract containing overrides is taken and analyzed. This analysis shows four legitimate bounced check fees being overridden. Unfortunately, it does not contain the appropriate data context under which the original event occurred. The data is then sent back to the system responsible for applying them and instructions are sent to re-apply them. The context and intent of the original event and message are lost.

This flow of events is akin to the old elementary school game of “telephone,” where through each handoff the context and contents of the first statement or sentence (e.g., Jane is running with Sally this Thursday) get lost to a point where a different message is delivered (e.g., Jane and Sully ran last Thursday.) Organizations and businesses frequently face this in one form or another. This could be in the form of wasted marketing attempts (disinterest was already expressed by the targeted consumer), or improperly handled claims (approval was already received that should have prevented the denial) or a host of other examples. These are all brought about by improperly interpreted or compromised data. The outcome can also be more than just an annoyance, depending on customer impact, as alluded to in the example in Figure 1.

The following three sections further elaborate the hidden costs and implications of data redundancy.

INFRASTRUCTURE IMPLICATIONS

For quite some time, the perceived infrastructure cost of replicating data was based on the additional disk storage price quote that is utilized as input at the time of the CBA for the given project or initiative. Examining Figure 2, however, shows that this assumption is incorrect. Whether adding a gigabyte or terabyte of disk, there will be far-reaching consequences from an infrastructure perspective. Each byte will require an incrementally larger network,

FIGURE 1 Implications of Data Modification and the Telephone Effect (Data Laundering)

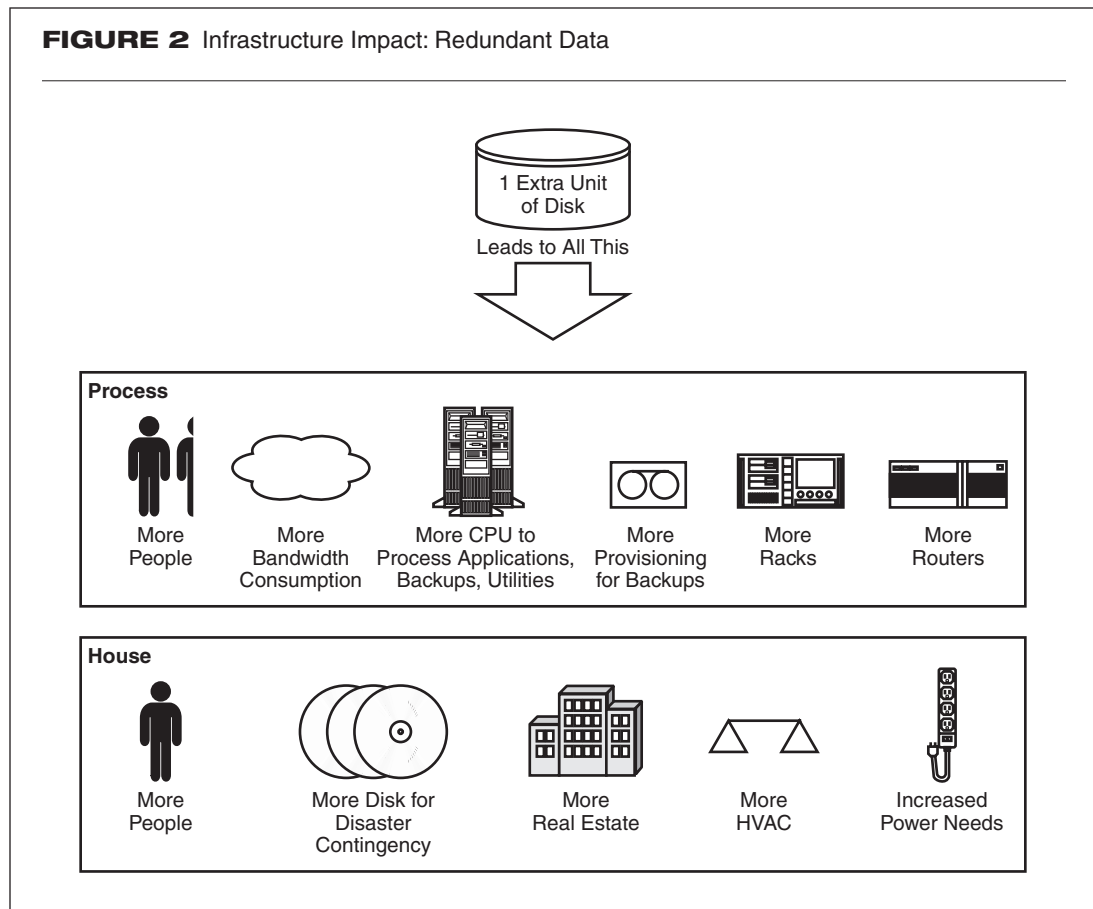
more hardware added to the disaster contingency plan (DCP) contract, extra computer processing power, more real estate, more HVAC, and the additional incremental labor to manage and support the larger disk footprint.

In addition to the labor to support this data from a pure infrastructure perspective is the human resources that will be consumed in writing programs, macros, and queries to analyze yet another set of data. These customized queries and data stores will manifest themselves in the form of spreadsheets, local desktop databases (e.g., Access), and eventually, more robust stores in the form of various moderately sized relational database management systems (e.g., small SQL Server, Oracle, UDB, etc. databases in "one off" or stand-alone installations). Every time one of these additional data stores is instantiated, another increment of processing power will be required to maintain and run them, whether on the desktop or a larger platform. Additional LAN (local area network) and WAN (wide area network) bandwidth will be consumed when spreadsheets and desktop database instances get moved from location to location. Proliferation of desktop databases and spreadsheets will mean increased PC hardware

requirements and more complex workstation configurations.

In addition to the local database manifestations stated above is the quantity of extract, transform, and load (ETL) logic and associated processing that is required to put the data into these redundant data stores. In some cases, this code may be relatively simple, such as when a subset of the data is placed into a spreadsheet or desktop database with no modification. In others, it can be complex, with subsets of data being transformed into different formats and blended with data from other sources. The complexity of the ETL will dictate to what degree additional physical resources will be required in the form of computer processing power and network bandwidth.

In summary, it is observed that data redundancy and its seemingly innocent requirement of more disk space for the enterprise will instead insidiously tax the desktop size and configuration, network capacity, and enterprise processing power to levels far exceeding initial expectations. These additional capacity requirements are not intuitively recognized. They need to be researched and articulated as input

FIGURE 2 Infrastructure Impact: Redundant Data

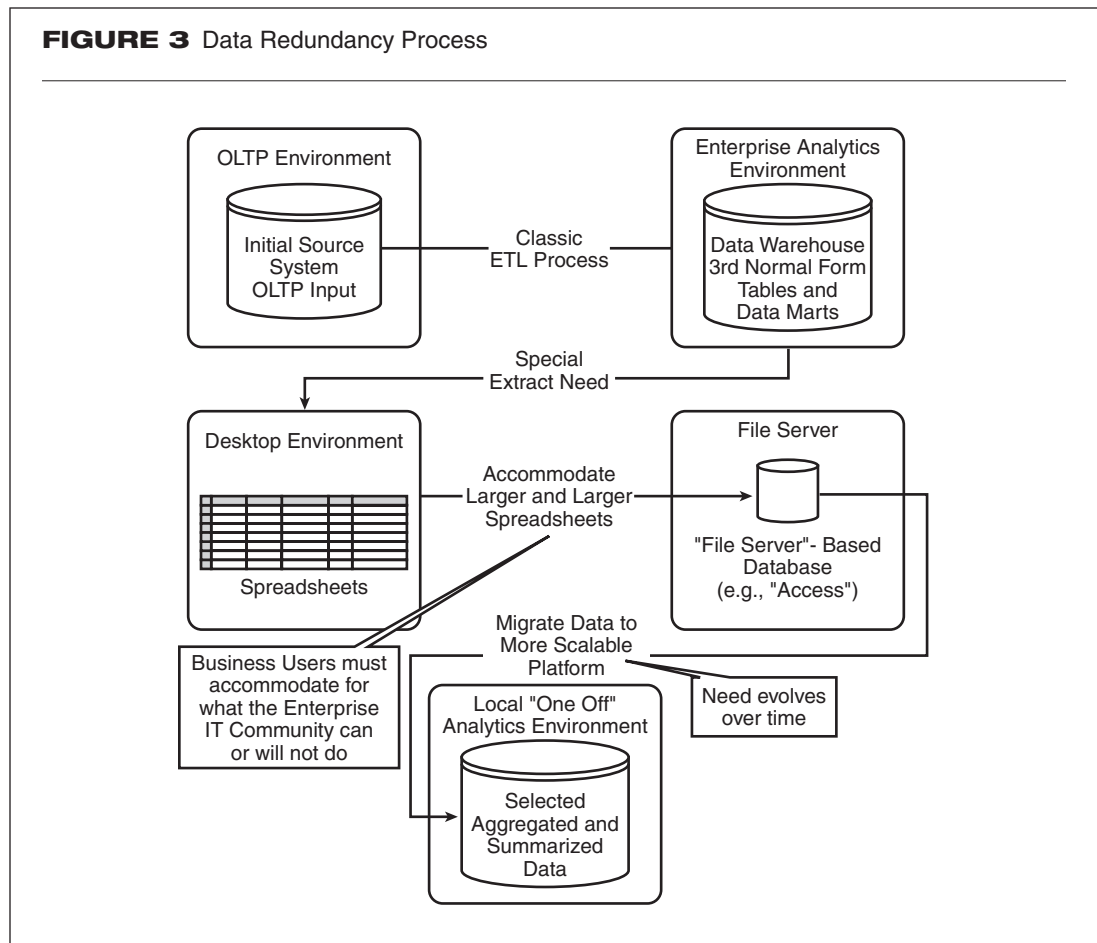
to any CBA analysis that seeks to replicate data in the enterprise.

SUPPORT IMPLICATIONS

From an organizational standpoint the new data stores and databases created by replication are spread throughout the organization, which encourages the development of secondary and tertiary IT support areas. These are quite frequently referred to as "shadow" IT groups. Referring to [Figure 3](#) it is shown how these can develop. As depicted, we can see how as the data gets removed from its original source and enterprise analytic environment, new data stores of varying sizes and technology are spawned over time (Donohue, 2005). In many cases, the data goes almost full circle, required to go back into an enterprise environment that is now required to support the growing appetite of the redundant and frequently morphed data. Compounding the hardware impact of this path is the fact that these shadow support groups are more business skilled and less IT savvy. As a result, the processes and queries executed against and in support of

the data are not as tuned and efficient as they could be. This breeds even more infrastructure resource consumption of processing power and network usage. Interviews with IT professionals at any large corporation about the size and proliferation of Access databases and Excel spreadsheets are sure to bring about tales of server meltdowns and multi-megabyte items being shipped across their LAN and WAN thousands of times a day. The root cause of these issues can usually be found in data replication evolving over time that was far beyond the original intent.

The nature of the support implications will also extend to that required for maintaining the ETL code as described previously. The code must be maintained to accommodate changes to the redundant data store(s) and the business logic behind the ETL programs and procedures housed within the shadow IT group. When this occurs out of the enterprise IT mainstream, efficiencies and disciplines are generally lost in regard to application design, maintenance, and support software and processes.

FIGURE 3 Data Redundancy Process

QUALITY AND INTEGRITY IMPLICATIONS

Among those performing analytic functions, the results of which are used to make important corporate business decisions, there is virtually no disagreement with the statement that “no data is better than incorrect data.” With that in mind, a conclusion can safely be drawn that no business or organization would intentionally allow one of its prized possessions, data, to become inaccurate. Extending this one step further and referring back to [Figure 1](#), we see that it is only through many small, innocent, tactical iterations that the data becomes corrupt and threatens to compromise business decision-making processes. On one hand are the physical costs that affect the bottom line in terms of maintaining additional hardware, software, and people resources to support duplicitous and inconsistent data. On the other is the potentially harmful consequence of relying on that data to make strategic choices or for delivering excellent customer service.

Beyond any physical products they may be producing, organizations are all about the data they collect, analyze, and distribute. A compromise to each data element a corporation owns is a compromise to its integrity as a business (*NY Times*, 2004). The customer perception of a bank that pays a tenth of a percent less in interest payments than its competitor on a certificate of deposit is quite different than if the same bank is calculating interest with a one tenth of a percent error.

Every occurrence of duplication and iteration of a given piece of data comes associated with an equal amount of effort and time to “undo” these actions. With it also comes the fact that the duplicative process and content has now increased the number of ways an organization is doing the same thing. Like the decision of whether to let a snowball loose at the top of a hill, organizations must consider each of these activities with an eye toward the sum total of their data actions. Each one brings the increased risk of poor data quality, a weakened

foundation for decision making, and the potential for diminishing customer satisfaction.

CONCLUSIONS

The cost savings from having good data quality are often demonstrated through the simple assumption of reduced disk and maintenance costs as a result of eliminating physical data replication. However, the larger cost is in the time and expense devoted to write hundreds of thousands of lines of code to design, transport, profile, clean, and reconcile myriad subject area data from the different parts of an organization. These extra reams of programming code also require lots of care and feeding. This comes in the form of application maintenance specialists and their business counterparts necessary to enhance and modify the sophisticated program logic to handle future requirements. Compounding the situation is the fact that the extra code and extra disk requirements then ripple through the infrastructure in the form of more supporting hardware and software to accommodate increased processing, network, and storage requirements.

Investment in the surrounding infrastructure to provide a safe home for the corporate data asset is a large component of most business operations budgets. Consequently, a similarly large investment should be found in the form of the governance and discipline applied to this data. It can be argued that as the pure infrastructure costs of the environment decline (e.g., storage) in the years to come, those components will become less relevant to creating a competitive edge and the accuracy and currency of the data will become paramount. Therefore, the increased complexity and additional programming logic components required specifically for data quality management in OLTP systems and ETL processes in the analytic environment must be demonstrated. This is manifested due to the need to govern, cleanse, scrub, and reconcile data not once, but multiple times. Also, the "ripple" effect on many infrastructure components due to redundant and poor-quality data must continue to be articulated beyond the conventional thinking in regard to added storage expenses.

Data problems come about when little regard is paid to the consequences of data replication and integrity beyond their original intent and sources. To fix this problem, a new and higher level of thinking must be applied than what got us here. A good place to start is by carefully analyzing and articulating each of the consequences: the ripple effect in the infrastructure, shadow IT organizations, and a loss of data meaning, content, and integrity. In concert with these efforts, organizations also need to understand and distinguish bad data from *bad processing* of data, whether through bad queries or lack of user training.

Examples are often given of how a gradual lack of monetary discipline or reduced accounting integrity, whether personal or corporate, will eventually lead to financial instability and the ensuing painful consequences. The same is true with data. ▲

Author Note

The content of this article does not reflect any view of the author's employer, CIGNA.

References

- Donohue, E., Change T., and Bostwick, J., Clean Up Your Data, *DM Review*, February 2005
<http://www.datawarehouse.com/article/?articleId=5071>.
- [New York Times], Reuters, Morningstar Data on Fund May Lead to Suit by S.E.C., *The New York Times*, September 25, 2004.
- Olson, J., *Data Quality: The Accuracy Dimension*, Morgan Kaufmann, 2003.
- Redman, T., *Data Quality: The Field Guide*, Digital Press, 2001.
- Suggested Readings**
- Buretta, M., *Data Replication: Tools and Techniques for Managing Distributed Information*, New York: John Wiley & Sons, 1997.
- English, L., *Improving Data Warehouse and Business Information Quality*, New York: John Wiley & Sons, 1999.
- Halsall, F., *Data Communications, Computer Networks & Open Systems*, Addison-Wesley, 1992.

Data problems come about when little regard is paid to the consequences of data replication and integrity beyond their original intent and sources.