

## Towards Quantifying Data Quality Costs

**Won Kim**, Cyber Database Solutions, Austin, Texas

**Byoungju Choi**, Department of Computer Science and Engineering, Ewha Women's University, Seoul, Korea

### Abstract

Today most organizations run their daily operations using data at their disposal. However, a vast majority of the organizations do not have adequate process and tools to maintain high quality operational data at all times. One of the key reasons for this is the lack of appreciation of the damages that low quality data can bring to an organization, and the cost of ensuring high quality of data. This article provides a basis for quantifying in monetary terms the costs of both low quality data and ensuring high quality data. A comparison of the costs of low quality data and ensuring high quality data can be a simple and compelling basis for an organization to determine the extent of the efforts it must expend to ensure high quality of its operational data.

## 1 TYPES OF OPERATIONAL DATA

To provide a basis for understanding the costs of both low quality data and ensuring high quality data, we need to understand the types of data that organizations use in running their daily operations. There are at least five types:

- **“Front-office” data**

Front-office data is data that is used in running primary daily operations of an organization; for example, customer account data in a bank to support daily debit-credit transactions.

- **“Back-office” data**

Back-office data is data that is used for analyses in support of strategic and tactical decision-making by an organization; for example, monthly or quarterly customer transactions data to support profit-loss, inventory, and merchandizing analyses by a nationwide retail chain. Often, back-office data consists of data extracted and transformed from front-office data. As such, low quality front-office data propagate to back-office data; and when low quality data is detected in front-office data, the part of the back-office data that was derived from the front-office data must be repaired. Back-office data also includes web log data that captures website visitors' navigation patterns.

- **Archival data**  
Archival data is data that has been transferred from hard disk to tertiary storage devices such as CDs, tapes, floppy disks, etc. for backup purposes and to support mobile users. Again, low quality data propagate from secondary storage to tertiary storage; and when low quality data is detected in secondary storage, the part of the archival data that originated from secondary storage must be repaired.
- **Metadata**  
Metadata is data about data, and includes such data as the system catalogs in relational databases. It also includes data dictionary, which contains such information as versions of software that work together, etc.
- **Control data**  
Control data is data that data-processing software uses for proper operation and control; for example, the access control list of ordinary users and their privileges, system access audit trail, etc. Control data is usually managed by privileged users such as system administrators and database administrators.

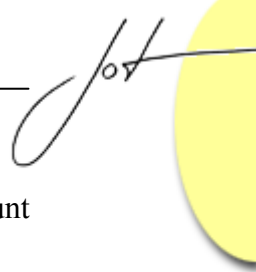
Different types of operational data have different uses for an organization, and cause different types of damages. Some data, such as the web log data and system access audit trail, are automatically generated and maintained, and therefore are normally not susceptible to errors, while operational data entered by people are subject to input errors of various types. Further, errors in data, such as missing data, wrong data, etc., can propagate, thereby complicating the tasks of detecting and repairing errors.

## 2 COST OF LOW QUALITY DATA

There have been limited efforts to systematically understand the effects of low quality data. The efforts have been directed to investigating the effects of data errors on computer-based models such as neural networks, linear regression models, rule-based systems, etc. [Ballou et al 1987] [Kauffman et al 1993]. In practice, low quality data can bring monetary damages to an organization in a variety of ways. As observed earlier in [Kim 2002], the types of damage it can cause depend on the nature of data, nature of the uses of the data, the types of responses (by the customers or citizens) to the damages, etc. For businesses they all boil down to financial losses, and for the government they boil down to loss of trust and waste of taxpayers' money. The types of damages include at least the following.

### 1. Loss of revenue

Organizations may actually lose money if they under-charge customers or citizens based on incorrect account information, such as the amount of charges patients incurred in hospital stays, the license fees and taxes citizens owed the government, etc. The shortfall on revenues can be determined manually on a periodic basis, and the historical amount of shortfall should be used as a guide to determine the level of efforts to prevent errors in data. In the case of a particular



major hospital in the US, as much as 30 percent of the customer account information contained errors.

## **2. Waste of money**

Organizations that mail marketing materials to incorrect postal addresses clearly waste money. This problem arises when organizations use misspelled or incomplete addresses or outdated addresses to send materials via postal mail. The problem does not arise if organizations use such incorrect addresses for other purposes, such as customer segmentation based on cities within a state or sections within a city, etc. The amount of wasted money can be computed if undeliverable mails are returned. If the mails are not delivered but also not returned, the rate of undelivered mails should be estimated by other means.

Organizations that make mistakes often have to take corrective actions. People resources and expenses incurred to correct the mistakes represent waste of money. For example, when organizations over-charge customers based on incorrect account information, they usually have to process refund requests at a later date. When organizations distribute reports, marketing materials, or product user manuals, etc. with important errors in them, they often have to re-print and re-distribute such materials.

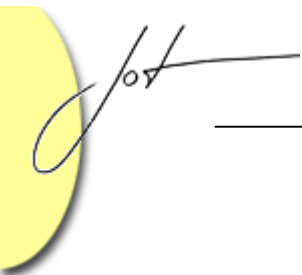
## **3. Lost opportunity**

Organizations that make inaccurate strategic and tactical decisions based on inaccurate analyses of customer segmentation, customers' purchase behaviors, etc. suffer lost opportunity. For example, suppose that a manufacturer incorrectly decides, based on customer data that contains a disproportionate amount of low quality data, to target customers "in the 30s and 40s age groups and \$60K to 70K income groups" to market a new product, rather than "40s and 50s age groups and \$70K to 80K income groups, and with no children". Then the manufacturer's marketing campaign will be partially misdirected, and if the effective time window for the marketing campaign closed before the mistake is discovered, the customer group that the manufacturer did not pursue represents a lost opportunity, and a loss of potential additional revenue. The wrong customer group the manufacturer pursued represents waste of money (at least a less-than-optimal use of money).

Lost opportunity often comes in conjunction with waste of money. When organizations take corrective actions upon discovering mistakes, they end up doing the same work twice or more. The people resource and the time they spend to correct the mistakes could have been used for productive work, and thus represent a lost opportunity. For example, customers of software products, when given user manuals that contain errors, will call help desk (customer technical support line). Customer support engineers then must spend time answering such calls, rather than helping customers who call with other types of request for help.

## **4. Tarnished image (or loss of goodwill)**

Repeated mistakes due to persistence of low quality data, like repeated mistakes from other reasons, can lead to desertion of customers, since after a while



disappointed and frustrated customers will switch to competitors for goods and services. This is somewhat difficult to quantitatively assess, since customers switch vendors and brands for a variety of reasons. However, the customer churn rate and the consequent decline in revenue and increase in the cost of acquiring new customers can be fairly easily determined, and a certain fraction of the figures may be assigned to persistent uses of low quality data.

#### **5. Invasion of privacy and civil liberties**

Invasion of privacy and civil liberties occurs largely due to unauthorized dissemination of private data about individual customers and citizens. However, they can also occur as a result of a proper use and dissemination of low quality data maintained by organizations. Incorrect, as well as correct, information about the diagnosis and treatment of patients in doctors' offices and hospitals, memberships in certain organizations, subscriptions for certain publications, purchase records of certain types of product, etc. can all potentially place individuals in bad situations. Such information can lead to discrimination in employment, a police investigation, detainment at airports and customs when traveling, etc.

Incorrect or outdated control data can also lead to invasion of privacy. If system administrators or database administrators do not properly manage the access control list, for example, by not updating it after certain employees have been terminated, unauthorized employees may access the computer systems or the databases, and damage the computer systems and databases, and/or disseminate certain private information about certain individuals found in the databases.

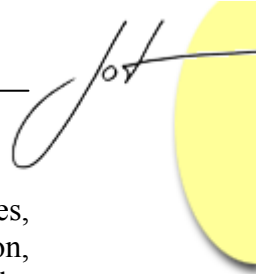
It is difficult to quantify monetary damages caused by invasion or privacy and violation of civil liberties, unless they result in lawsuits. But the costs of certain corrective actions such as a bank's invalidating and re-issuing credit cards to millions of customers after a security breach in its computer systems can certainly be estimated or even measured.

#### **6. Personal injury and death of people**

It is rare but still possible for the use of incorrect data or unauthorized dissemination of data to lead to a destruction of properties, personal injuries and deaths of people. For example, invasion of privacy may lead to mental distress and subsequent medical treatment or even suicides. Wrong instructions, due to wrong or outdated data, for operating certain types of machines (e.g., construction equipment, and other hazardous equipment) or controlling certain environments (e.g., chemical plants, nuclear power plants, and oil refineries) can cause accidents and even disasters. The damages done to properties and people can translate to replacement costs, hospital charges, legal costs, and lost workdays for people involved (waste of money and lost opportunity).

#### **7. Lawsuits**

Damages attributable to low quality data can, under certain circumstances, lead to lawsuits by customers against businesses, and by citizens against the government. Invasion of privacy, personal injuries and deaths, and significant revenue losses,



etc. are likely causes for lawsuits. The cost of a lawsuit includes legal expenses, human resources (other than the lawyers) and time spent on the legal action, judgments (i.e., damages awarded to the plaintiffs, fines levied on the organization and officers of the organization), and tarnished image due to bad press.

Although it is often difficult to estimate the cost of low quality data in monetary terms, this must be done. As they say, “money talks”, and only when expressed in explicit monetary terms, decision makers of organizations can relate to the consequences of low quality data on their organizations. The cost of low quality data should be estimated for all fields in all tables in a database, and for all conceived uses of data in all fields. However, as not all damages are guaranteed to occur (e.g., a lawsuit, death of people), probability should be assigned to each type of damage when estimating the total damages.

### 3 COST OF ENSURING HIGH QUALITY DATA

The cost of ensuring high quality data is the cost of preventing, detecting, and repairing low quality data. This does not include the cost of taking corrective actions for damages done due to low quality data. [Kim et al 2003] provides a rather comprehensive taxonomy of low quality data (also called dirty data) for the purpose of understanding the types of low quality data and the techniques appropriate for addressing them. Let us examine the three primary elements in maintaining high quality data.

#### 1. Preventing low quality data

There are two methods to prevent low quality data from entering databases and files: automatic and manual. The automatic method is embedded in software, such as database management systems, and end-user applications (e.g., hospital patient records systems, order entry systems). Software systems usually reject certain wrong data types (e.g., numeric data for the person\_name field, character data for the person\_age field). Transaction management systems prevent bad updates through concurrency control and crash recovery mechanisms.

The automatic method, however, is rather limited in its coverage; for example, it cannot prevent misspelled words or many types of wrong data from entering the databases or files. The manual method is usually necessary to augment the automatic method, possibly assisted by software tools. Before storing data in the database or a file, people should check correctness of data. For example, people should check the spelling of character data, aided by a spell checker; check the correctness of both numerical and character data by examining the original sources of data (e.g., faxed purchase orders, employees’ resumes, filled-in surveys) and/or by cross-referencing other related and redundant sources of data. The manual method is typically costly, since it requires people resources, and possibly licensing of software tools.

## 2. Detecting low quality data

In general, it is not possible to prevent most types of low quality data. As such, it is necessary to detect low quality data after it has entered the databases or files. Detection may be done periodically (e.g., once a week, a month, a quarter) or before certain events (e.g., the release of financial data of a business to shareholders, a government tax audit). There are also two methods to detect low quality data: automatic and manual.

The automatic method involves the use of software applications that examine the database in many different ways. For example, the names of employees in a department of a corporation (in one table) may be compared against the names of employees reporting to each department manager (in another table); or a value range test may be done against the ages of all employees to make sure nobody's age lies outside the range [18..65], if such a test was not done when the employees' age data was entered; or if the organization's rule is to ensure that "the customer technical support staff be at least 20% of the engineering development and quality assurance staff", the counts of the employees in the two groups should be checked.

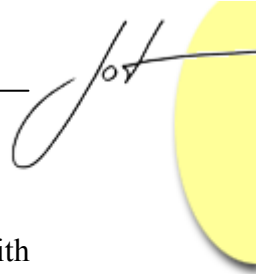
The manual method for detecting low quality data is similar to that for preventing low quality data. For detection, however, people need to examine the data stored in the databases and/or files, rather than the data being entered or updated. The manual method is also usually required to identify all data propagated from the low quality data. Low quality data may propagate in a few different ways. Low quality data in one field (e.g., the Salary field) of a table may be copied to another field (e.g., the Salary+Commission field) in the same table or the same database. It may be copied to a separate remote database as backup against disk crashes or to improve performance or for autonomous administration of databases in a distributed environment. It may also have propagated into summary tables or materialized views created in the same database or in remote databases.

## 3. Repairing low quality data

Once low quality data has been detected, it has to be repaired. Again, there are two methods to repair low quality data: automatic and manual. The automatic method typically involves the use of a series of update statements against a database to replace wrong or outdated data with correct data. The manual method augments the automatic method, and involves people who manually examine data and repair them. Repairing low quality data usually requires a combination of automatic and manual methods.

## 4 BALANCING THE COSTS

A data quality strategy for an organization includes several elements:



- Understanding the semantics of all the operational data.
- Determining all the uses of the data and consequences of low quality data with respect to such uses.
- Determining the costs of low quality data.
- Determining the cost-effective methods of ensuring quality of data appropriate for the conceived uses of the data.
- Establishing a process to ensure quality of data as determined, including the installation of necessary software tools and deployment of an appropriate level of human resources.

To determine the cost-effective methods of ensuring appropriate quality of data, organizations need to compare the cost of ensuring high quality data against the total estimated cost of low quality data, and select an appropriate combination of the methods for ensuring high quality data that collectively would be significantly lower than the cost of low quality data. This tradeoff decision is akin to that proposed between the accuracy and timeliness dimensions of data quality [Ballou and Pazer 1995].

## REFERENCES

- [Ballou et al 1987] Ballou, DP, Pazer, HL, Belardo, S., and Klein, B. "Implications of Data Quality for Spreadsheet Analysis," *Data Base*, vol. 18, no. 3, pp. 13-19.
- [Ballou and Pazer 1995] Ballou, DP, Pazer, HL, "Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff," *Information Systems Research*, vol. 6, no. 1, pp. 51-72.
- [Kauffman et al 1993] Bansal, A, Kauffman RJ, and Weitz, RR, "Comparing the Modeling Performance of Regression and Neural Networks as Data Quality Varies: A Business Value Approach," *Journal of Management Information Systems*, vol. 10, no.1, pp. 11-32.
- [Kim 2002] Kim, W. "On Three Major Holes in Data Warehousing Today," *Journal of Object Technology*, vol. 1, no. 4, Sept/Oct., ETH.
- [Kim et al 2003], Kim, W., Choi, BJ, Hong, EK, Lee, DH, Kim, SK. "A Taxonomy of Dirty Data," *The Data Mining and Knowledge Discovery Journal*, vol. 7, no. 1, pp. 81-99.

## About the authors



**Won Kim** is President and CEO of Cyber Database Solutions ([www.cyberdb.com](http://www.cyberdb.com)) and MaxScan ([www.maxscan.com](http://www.maxscan.com)) in Austin, Texas, USA. He is also Dean of Ewha Institute of Science and Technology, Ewha Women's University, Seoul, Korea. He is Editor-in-Chief of ACM Transactions on Internet Technology ([www.acm.org/toit](http://www.acm.org/toit)), and Chair of ACM Special Interest Group on Knowledge Discovery and Data Mining ([www.acm.org/sigkdd](http://www.acm.org/sigkdd)). He is the recipient of the ACM 2001 Distinguished Service Award.



**Byoungju Choi**, associate professor of computer science and engineering, (component-based software engineering, software testing, data and software quality), Ph.D. in computer science, Purdue University, [bjchoi@ewha.ac.kr](mailto:bjchoi@ewha.ac.kr)