7-2022

Detecting and Resolving 'Dirty' Data: Ten Steps to Better Business Insights

Dana R. Hermanson
Kennesaw State University

James G. Lawson III
Bucknell University

Daniel Street
Bucknell University, das051@bucknell.edu

Detecting and Resolving 'Dirty' Data

Ten Steps to Better Business Insights

By Dana R. Hermanson, James G. Lawson, and Daniel A. Street

In Brief

Although business's increasing utilization of technology to drive operational decisions has meant a greater reliance on data, it is exceedingly rare that an entity possesses a perfectly clean set of data. Some of it is inevitably invalid, incomplete, or inaccurate—in other words, "dirty." In order to address the issues caused by dirty data, the authors provide a 10-step program CPAs in any setting can use to help businesses make better decisions.

CPAs face increasing calls to provide business insights through data analytics, but the quality of insights depends critically upon the validity, completeness, and accuracy of the underlying data. "Dirty" data can cause analytics to be inaccurate, leading to suboptimal decisions. To assist CPAs in maximizing the benefits of their data analytics, this article presents a 10-step process to detect and resolve dirty data. The authors provide several examples to highlight dirty data issues and to show how to work toward an environment of cleaner data and more useful analytics.

The 10-step process applies to a wide variety of accounting settings, but the process may be easiest to fully implement in larger organizations where CPAs have sufficient time available to address data issues. Nevertheless, the authors also believe that it is important for CPAs in smaller firms or smaller companies to consider these steps. Although ignoring dirty data issues may save time in the short run, this will likely result in larger costs in the future as dirty data problems persist. In short, the authors believe that investing time to identify and resolve dirty data issues now may be more efficient in the long run than dealing with the same dirty data issues month after month after month. In cases where there simply is not enough time to pursue these 10 steps, perhaps the most critical step is identifying root causes when data problems appear, as this may allow for correcting the root cause, or at least anticipating future data problems.

The fundamental message that the authors want to emphasize is not to jump straight to analysis when new data arrives. Rather, take time to understand the business process and data sources; develop and test expectations about the data; and, as dirty data challenges are revealed, take steps to clean up the data and address the root causes to prevent the same problems from happening in the future.

Data Analytics and Dirty Data

Numerous articles have been written about how accountants can utilize data analytics, focusing on, for example, auditors (Susan B. Anders, "Audit Data Analytics Resources," *The CPA Journal,* June 2017), nonprofits (Amy West, "Data-Driven Decision Making for Not-for-Profit Organizations," *The CPA Journal,* April 2019), management accountants (Raef Lawson, "New Competencies for Management Accountants," *The CPA Journal,* September 2019), and tax accountants (Ernie Guerriero, Richard L. Engebretson, and Cary W. Parker, "Leveraging Data Analytics," *The CPA Journal,* December 2019), among

others. Although the use of data analytics provides unique opportunities to develop business insights, it also presents unique challenges. Specifically, successful using data analytics requires clean source data. Although others have acknowledged the need to assess the underlying data before analyzing it (e.g., Virginia Collins and Joel Lanz, "Managing Data as an Asset," *The CPA Journal,* June 2019; Ellie Hume and Amy West, "Becoming a Data-Driven Decision Making Organization," *The CPA Journal,* April 2020), many CPAs may lack a systematic process for identifying and resolving dirty data.

"Dirty data" refers to invalid, incomplete, or inaccurate data. Imagine the following hypothetical examples in the information system of a small retail store: 1) vendor invoices marked "paid," but with no corresponding check number or payment details; 2) payroll checks with employee numbers that do not match any of the employee numbers in the employee master data file; 3) a $50,001 monthly electric bill; or 4) a monthly sales summary that does not include all departments in the store. These dirty data problems could be due to the following: 1) the bookkeeper issued a check and forgot to associate the check with the vendor invoice; 2) employees were issued new ID numbers when the payroll processor changed, but the system was never updated; 3) a $500.01 electric bill was paid properly, but mis-keyed into the information system; and 4) an employee was late

Exhibit 1
10-Step Process to Detect and Resolve Dirty Data

1. Understand the business process represented by the data.
2. Analyze the source and processing of the data.
3. Determine which elements the data set should contain.
4. Scan a sample of recent data.
5. Summarize the data of each table.
6. Document expectations for the data within each field.
7. Document expectations for the relationships among fields.
8. Test each expectation by creating exception reports.
9. Evaluate exceptions and identify root causes.
10. Act to clean the data and address the root causes.

entering data for one department in the store. Without correction, the store's financial and managerial reports, and therefore decision-making process, could be dramatically affected, and the same types of dirty data issues may continue to arise in the future.

Dirty data can be caused by a variety of factors, including human error, poor information system design, ineffective internal controls over data, time pressure, merging of different systems, failure to appreciate sloppy data entry, lack of accountability for data quality, and improper incentives. Furthermore, data quality is not perfectly controllable, as it relies on people. Even after extensive training and with effective internal controls, dirty data likely will still exist to some degree—albeit hopefully less than before.

Many executives are aware of both the pervasiveness and challenges of dirty data. For example, Brad Fisher of KPMG states:

> At present, many CEOs understand in principle the value and potential that data offers, yet seem reticent to fully trust (and act on) what the data tells them. KPMG International's recent Guardians of Trust study highlighted the extent of this conflict—just 35% of executives surveyed said they had a high level of trust in their own organization's use of data analytics, while 25% admitted they either have limited levels of trust in, or actively distrust, the data they receive. It's probably not a surprise, then, that in our CEO Survey, 67% of CEOs say they have ignored the insights provided by data analysis or computer-driven models in the last three years, because they have contradicted their own intuition or experience. (https://web.archive.org/web/20210622011530/https://home.kpmg/xx/en/home/insights/2019/04/dont-doubt-the-data.html)

Given the importance of data analytics, and the costs of dirty data, the authors provide a 10-step process that can help CPAs to both detect and resolve dirty data.

A 10-Step Process to Detect and Resolve Dirty Data

The authors offer the following 10-step process to detect and resolve dirty data (see *Exhibit 1*). One author worked in data analytics and encountered many data issues, and two of the authors presented a portion of this process in an educational case that they and others use in class (James G. Lawson and Daniel A. Street, "Detecting Dirty Data Using SQL: Rigorous House Insurance Case," *Journal of Accounting Education,* 2021). The third author has used elements of this approach in his research, especially when dealing with survey-based data sets.

The authors expect that readers may already be performing some of these steps, but following this 10-step process should give CPAs a holistic approach, show them how to proceed, and guide their responses to this challenging problem. This process can also provide a shared communication framework to enable team members to respond to dirty data issues more effectively.

Before discussing the 10-step process, it is critical to make one thing clear: the first step in detecting and resolving dirty data is not to jump into the data and begin running complex models. Rather, there are some key foundational steps to take first.

Step 1: ***Understand the business process represented by the data.*** The first step in the process is to understand the business setting that the data represents. The 10-step process relies on CPAs' business knowledge and intuition, so it is critical to comprehend the business process represented by the data. If one is working with internal data (data from a CPA's own organization), then it is likely that this step will require less effort. For external data (data from another organization), CPAs should obtain an understanding of the business process, know how the information flows through the system and is used by the individuals in operations, and perhaps perform a process walkthrough. A solid understanding of the business setting is critical, because this understanding is the basis for knowing what data to expect and identifying data issues.

Step 2: ***Analyze the source and processing of the data.*** Once the CPA understands the business process that the data represents, the next step is to identify the source and analyze the processing of the data. Considerations

include the following:

- How is the data input and processed by the information system?
- Is data input manually, or is it input by electronic data interchange, barcode scanning, APIs (application programming interfaces), or other such electronic methods?
- Which internal controls affect the data input process?
- Is the data tested for accuracy and validity before it is accepted into the information system, or is the accuracy of the data assumed?
- Once the data is input into the information system, is it later processed or updated?
- Who can edit data after it is originally input?
- Are any changes made during processing or editing the data also subjected to internal controls?

Readers should take a cue from Murphy's Law: Almost anything that can go wrong in a database will go wrong. Thinking along these lines will allow CPAs to understand the existence and origins of dirty data in the information system.

Step 3: ***Determine which elements the data set should contain.*** Based on the understanding developed from Steps 1 and 2, which elements (e.g., tables, fields, IDs) should the data set contain? If available, CPAs should ask for a copy of the database schema, system flow chart, or data dictionary. In many cases, such documentation will not exist, especially in smaller companies. A lack of documentation may be a signal that dirty data issues are likely to be present.

Step 4: ***Scan a sample of recent data.*** The next step is to open the data and get a sense of what it actually looks like. CPAs should briefly scan a sample of recently updated rows in each table to become familiar with the data. Compare the actual layout of the data to the expectations developed in Step 3. If data does not make sense or does not seem right—which occurs frequently—make notes and consider those issues in Steps 6–8. With practice and experience, the amount of time needed to gain familiarity with actual data will likely decrease.

Step 5: ***Summarize the data of each table.*** With the data set open, summarize the data of each table as follows:

- Determine how much data is missing in each field and assess whether the missing data seems to be distributed randomly or systematically.
- For numeric and date fields, identify the minimum, average, and maximum amount/date.
- For categorical variables, count the frequency with which each category occurs.

A lack of documentation may be a signal that dirty data issues are likely to be present.

CPAs should make notes of any data that does not make sense or does not seem right (e.g., missing data elements, minimum or maximum values that are outside a reasonable or expected range, or patterns in the categorical variables that are not expected) and address those issues in Steps 6–8. This is a form of overall analytical review that the authors have found to be extremely helpful in detecting dirty data.

Step 6: ***Document expectations for the data within each field.*** Working one field (variable) at a time, CPAs should document their expectations for the data within each field. This step is valuable for all field types (e.g., IDs, statuses, dates, amounts). The following are examples of some expectations drawn from a purchase order data table for an organization that typically uses POs for purchases:

- Purchase order ID should not be missing nor be repeated.
- Purchase order status should contain only the following possibilities: requested, approved, denied, ordered, awaiting payment, or paid.
- Requested amount should not be negative.
- Approval date should be no later than today's date.

Step 7: ***Document expectations for the relationships among fields.*** Again, CPAs should work through each field (variable) at a time. This time, instead of thinking about expectations within each field, think about how each field should relate to the other fields in the data set (perhaps even among tables). The following are examples of some expectations drawn from an invoicing context:

- Vendor IDs from the invoice detail rows should match the IDs of the vendor table.
- "Paid" status invoices should not be missing an "amount paid" or "date paid."
- "Sale date" should be less than or equal to the "invoice date," which should be less than or equal to the "date paid."
- "Remaining balance" should be less than or equal to the "invoiced amount."

Step 8: ***Test each expectation by creating exception reports.*** Exception reports should be developed using

tools such as SQL, Idea, or ACL to test the expectations developed in Steps 6 and 7. Some CPAs may be capable of creating their own exception reports, while others will need to collaborate with an analyst or an IT professional to perform this step. Regardless of the software used, the logic underlying each exception report is the same: Display all rows in the data where the expectation is *not* met. For example, identify all invoices that have a "paid" status, but that are missing data in either the "amount paid" or "date paid" fields.

Some of the exception reports may contain no records. Assuming that the exception report was created correctly, an exception report with no records likely means that every row in the queried data satisfies the expectation that was tested. On the other hand, any exception report that *does* contain records reveals that the expectations were not fully met.

Step 9: ***Evaluate exceptions and identify root causes.*** CPAs should review the exceptions identified by the reports from Step 8. In each case, it is important to ask, "Could this be right? If this were true, what would the business process look like?" Ask other users of the data or individuals in operations or IT, "Could you help me understand why this record looks like this?" It is possible that the CPA's initial expectations were not met because the business process functions differently than one initially understood. Perhaps there are unusual circumstances that occur from time to time.

When the identified exceptions do reflect dirty data in the database, there are two important tasks:

- Clean the data by determining the proper values for each affected field/record. This addresses the short-term issue of needing clean data for analysis.
- Determine why the dirty data exists and how it can be prevented going forward. This addresses the root cause and the longer-term issue of preventing the same problems from occurring in the future.

It is important to acknowledge that there is rarely, if ever, a perfectly clean data set. CPAs should adopt a perspective of curiosity, humility, and teamwork when evaluating dirty data and trying to identify its root causes.

The root causes of dirty data can be assessed by considering patterns, people, and process.

Patterns: Do the identified exceptions represent a widespread, systematic problem with the data, or are these exceptions few and far between? Did these exceptions occur recently, or is it only old data that exhibits these problems? Is there a particular "locale" (i.e., database user, employee, department, customer, office) that generates dirty data?

People: Do the individuals who input the data need more training? Do the individuals who input and edit the data understand the downstream implications of dirty data? Do these users have a sense of how the data will be used? Are there any incentives that might encourage users who input or edit the data to bias or even falsify the data?

Process: Is the database designed correctly? Were there any internal controls that should have prevented or detected the dirty data? Are any new internal controls over data input, editing, or processing needed to ensure data cleanliness going forward?

It is important to acknowledge that there is rarely, if ever, a perfectly clean data set. CPAs should adopt a perspective of curiosity, humility, and teamwork when evaluating dirty data and trying to identify its root causes. When collaborating with others to address dirty data, focus on communicating the bigger picture objective: "Let's work together to improve the quality of this data so that the organization can leverage this data to conduct more efficient analyses and to make better decisions going forward." Data, by its nature, is not of perfect quality all of the time. The goal is to improve the quality of the data supporting analyses—to get it as clean as possible.

Step 10: ***Act to clean the data and address root causes.*** It is very important not to simply clean the current data while leaving the fundamental root causes unresolved. This approach might solve current problems, but likely would result in the same dirty data problems emerging in future analyses, requiring yet more data cleaning. The actions necessary to resolve dirty data problems will vary depending upon the CPA's role, the entity's culture, and the types of relationships in place with the personnel involved. In a managerial role, a CPA might establish accountability for data cleanliness, change incentives, retrain or reallocate staff, implement or enhance internal controls, or even change the information system design. In an audit or attest role, a CPA might increase their assessment of control risk, reevaluate management's competence or integrity, or reevaluate the reliability of other data and reports.

In an internal audit or consulting role, a CPA might plan a new engagement to clean existing data and propose additional information system internal controls. Regardless of the CPA's role, it is important to take appropriate action to ensure data cleanliness going forward.

Be Skeptical but Curious

Because perfectly clean data sets are quite rare, it is important to have a process for cleaning data and preventing the same problems from arising in the future. With this in mind, CPAs should take away four main messages about working with dirty data:

- Be professionally skeptical about the cleanliness of source data underlying reports or analyses. Remember: Almost anything that can go wrong in a database will go wrong. This includes errors, omissions, duplications, and biased or hurried data entry. Develop an understanding of the business process and the source and processing of the data to provide a basis for knowing what to expect, asking probing questions, and identifying issues. Perform a variety of analytical review steps to identify unexpected or missing data and other potential problems with the data set.
- Have an attitude of curiosity, humility, and teamwork while searching for the root cause of dirty data issues. Use this opportunity to hone one's knowledge of the organization—its processes, its people, and its controls.
- Once dirty data is identified and its root causes are understood, boldly seek process or information system improvements necessary to prevent dirty data from recurring. Do not just fix today's problem, but think about how to improve the process so that the next time others analyze this data, they will not have to devote so much time to cleaning it.
- As progress is made in cleaning data and preventing similar problems in the future, highlight the success of these efforts and communicate how they contributed to more effective and efficient business decisions. Work toward building an organizational culture that values clean data and that understands the financial and time investment needed to improve data quality.

The authors have found the 10 steps discussed in this article to be very helpful in their own data analysis efforts. CPAs in a variety of settings can use this process to identify and resolve dirty data in the short term, as well as to promote longer-term data integrity. ■

Dana R. Hermanson, PhD, *is a professor at Kennesaw State University, Kennesaw, Ga.* ***James G. Lawson, PhD,*** *is an assistant professor at Bucknell University, Lewisburg, Pa.* ***Daniel A. Street, PhD,*** *is an assistant professor, also at Bucknell University.*