What Every IT Auditor Should Know About Transforming Data for CAATs

For several decades now, advances in computers and information technology have led to the ubiquitous employment of computers in the business community, whether the entity is small or large. That growth led to a concomitant increase in the amount of data entities have and keep, due to the significant decline in the cost of storage. Today, the outcome is referred to as “big data.” Considered a hot topic in IT, big data is also the result of collecting nonfinancial data along with financial data. Everything from network logs to industry and economic data is being collected. These facts drive the need for experts in using computer assisted audit tools/techniques (CAATs).

Another consideration is the number of paperless transactions that occur. These days, many transactions do not necessarily have any paper at the point of sale (e.g., e-commerce sales). How can the primary source document (i.e., data) for such a transaction be evaluated without a CAAT?

None of this is new to the average business person, but there is something many may not realize. Big data is going to get bigger—much bigger—and faster. According to the Berkeley School of Management, more data have been created in the three years 2009-2011 than in the whole of human history.¹,² According to one expert, in 2005, there were 130 exabytes of data in the digital universe. By 2010, it was 1,227 exabytes—a 10-fold increase. By 2015, it is predicted that the digital universe will contain 7,910 exabytes.³ In addition, from 2011 to 2020, data being managed by IT professionals are expected to increase 50 times, with the number of IT professionals only increasing 1.5 times.⁴

Combining these facts, there is an inescapable conclusion about auditing or analyzing data, whether financial or operational objectives: Expertise in CAAT-like techniques and tools is needed to examine big data. More than ever, entities are in need of IT auditors who can properly extract and analyze mounds of data and turn them into useful information. That requires an effective methodology and an effectual set of tools.

The data warehouse segment of IT has developed a sound methodology to do just that by using extract-transform-load (ETL) techniques and tools to get various source data sets into a single warehouse, where business analytics tools are used to analyze the huge amounts of data amassed in the warehouse. This methodology should be equally beneficial to IT auditors using CAATs and big data.

USING ETL METHODOLOGY FOR CAATS AND DATA MINING

The ETL approach can be applied to CAATs and data. Most experts agree that the most difficult part of using CAATs is the data extraction phase. In ISACA Journal, volume 6, 2010, this column addressed the issues of extract. In summary, the primary issue was getting data from operational computers, generally online transaction processing systems, into a form and format that are compatible with the CAAT. Often, data are exported into a format that is not immediately suitable for use in the CAAT. The ideal format was presented in the previous article: a flat file with column headings. Various potential extraction formats were discussed in that article, as were preferences based on suitability of the resulting format. The need to verify the data integrity of the extracted data, compared to the original data, was also discussed.

There are three major types of issues that cause extracted data to need to be corrected or cleaned after they are extracted:

1. Formatting issues related to the way data are formatted in a particular extraction approach. For example, the export could be in a report format (e.g., a digital PDF report) and have a lot of extraneous lines of data (e.g., headings, subtotals). In addition, some reports take a single transaction and list the information on two lines.

2. The various idiosyncrasies in the way that specific data values are presented and/or formatted. For example, a negative number could have a negative sign in front of or behind the figure, or negative numbers could be put in
Some CAATs provide specific commands related to the problems described in figures 1 and 2. In fact, a tool that is good at transforming data could be cost-effective even if a different CAAT is being used to perform the procedures after the data are transformed and loaded into the CAAT. Microsoft Excel can do most of these transforming procedures, and using macros, it is likely that all of them can be done in Excel. However, CAATs exist for most, if not all, of these transform needs, and are much easier and more reliable than using macros in Excel.

CONCLUSION
Efficiency and effectiveness in data mining and data analytics are highly dependent on reliable, clean data provided to the CAAT. A useful approach is the data warehouse ETL process. An earlier article addressed the process of extracting data (volume 6, 2010). This article addresses the transform process. In a data warehouse, the transform process is usually considered the most time-consuming; the same is true in using CAATs. The good news is once the extract and transform steps are properly performed, the load part, to get the data into the CAAT, is usually a simple process.

The first major point regarding transform is that, for it to be effective, the IT auditor needs to understand the procedures and commands of the CAAT in order to determine what needs to be addressed or changed in the data for those commands to execute properly. The second point is to understand the need for a sound methodology or approach to the transform process. Finally, the IT auditor needs an effectual tool that can perform most, if not all, of the three types of transform issues discussed.

ENDNOTES