LAW360

Portfolio Media. Inc. | 111 West 19th Street, 5th Floor | New York, NY 10011 | www.law360.com
Phone: +1 646 783 7100 | Fax: +1 646 783 7161 | customerservice@law360.com

Harnessing The Power Of Big Data In Litigation

By **Jorge Gallardo-García, Benjamin Scher and David Barth**
(August 30, 2022, 2:57 PM EDT)

Jorge Gallardo-García

Benjamin Scher

David Barth

With the digital transformation, the amount of data available in litigation matters has grown dramatically in scope and volume.

A data set containing a million records was considered significant just a decade ago. Now it is commonplace to have data sets with billions of records.

This increase is not surprising. Companies are collecting and analyzing all the data they can get their hands on. Vendors that track and sell data are seemingly everywhere. Government entities are making large public data sets available. This explosion of available data has affected litigation, production and the parties involved in disputes.

Data — when analyzed correctly and explained effectively — have always provided a valuable way to find objective insights in litigation matters, answer key liability and damages questions, and support critical discovery efforts.

The growth in the volume, scope and utility of available data is transforming the way data are analyzed, and it requires new technological tools. Those tools can be used to harness big data, pull insights from it and ultimately help inform case strategy based on information and insights the data provide.

Simple tools like spreadsheets are not equipped to handle the volume of data now available in litigation matters. The need to derive insights from litigation data in a timely way has rendered even "large data tools" from years past suboptimal and often unworkable.

New tools, including Hadoop, Spark, Databricks and high-performance computing, are now available to manage and analyze today's big data in litigation matters.[1] How do those tools change the approach to analyzing data for litigation?

Big Data Tools for Different Types of Data Processing

At their core, these new technologies take advantage of a concept known as parallel processing. Parallel processing did not start with the big data tools discussed here, but these new tools have taken it to

unprecedented levels. The basic idea is that, rather than running a long process from start to finish linearly, the process is broken into multiple components that are then run simultaneously.

Imagine you have 1 billion transactional records of the sales of a product, and you want to see how each transaction price compares to a benchmark price. Using standard tools, the first record would be compared to the benchmark, then the second record would be compared to the benchmark, then the third, the fourth — linearly to the billionth. Not surprisingly, even on advanced servers, these types of processes can often take quite some time.

With big data tools that utilize parallel processing, the process can instead work as follows. First, a billion records are broken up, for example, into 100 groups of 10 million each. Then the comparison to the benchmark is run on each of the 100 groups simultaneously. Since the 100 groups are analyzed in parallel, the task can be accomplished roughly 100 times faster.

Instead of taking a couple of weeks, the analysis can be done in a couple of hours. In the context of tight litigation timelines, this time savings can be critical.

There are three key benefits of using big data tools in litigation matters.

More Analyses in the Same Amount of Time

Litigation is inherently deadline oriented. Big data technologies can accomplish analyses in a fraction of the time it would have taken without them.

Whether attempting to perform complex computations on large data sets prior to a report submission or evaluating large amounts of information prior to the close of discovery, big data tools may allow for the previously impossible to be accomplished within tight time frames.

Better-Informed Decisions

The time saved in processing data enables better insights and conclusions than are possible with more limited computing resources. If an analysis is done earlier, there is more time to think through the results, to run different scenarios, to identify different possibilities and adjust strategy.

Cost Efficiencies

By speeding up the analytical process, using big data tools can save money. The time gained in running the data can be used for other aspects of the case or just saved altogether by obtaining similar answers but in less time. That is, more can be done in less time or the same analysis can be done for less.

In other words, using big data tools can improve the quality of the work and the reliability of the analysis and can result in cost efficiencies.

But big data has also opened the door to other important considerations.

Points to Consider With Big Data

Deciding Whether to Use a new Data Tool

Big data tools are not self-aware. They cannot tell us if they should be used or not. For each situation, it's important to determine the best approach, given the problem and the available data. The matter may, in fact, dictate that analyzing a statistically representative sample of data is a better approach than using a new data tool.

For instance, in litigation over price-fixing with billions of transactional records of the sales of a product available, there may be two approaches for comparing them to a benchmark price. One option could be to use big data tools and calculate the difference for each of the billion records.

A different option is to draw a random sample from the billion records to perform the analysis. Big data tools make the choice easier in some cases: If the data are electronic and of uniform quality, and one has the means to use all the data, then using all the data is both feasible and efficient.

The resulting analysis will be more precise, more reliable and even less expensive. Also, if there isn't a need to design, implement or interpret a sample, that can free up budget and allow more time to analyze the results in greater detail.

Sampling may be an appropriate solution if, for instance, the information is only available in handwritten notes in hard copies stored at various locations. In such instances, collecting all this information and converting it into a database could take a long time and be cost prohibitive.

For such a circumstance, a representative sample could be drawn, collected, analyzed and extrapolated to estimate the results for the full population. The key, of course, is to have a sample design that will be representative of the population and useful to answer the question at hand.

Sampling can be appropriate even when the data are entirely electronic. Although big data tools have increased the amount of data that can be analyzed, they are not limitless, and some data sets are too big even for these modern tools. In this case, one can use a sample of millions of observations instead of billions or trillions.

In some cases, data may be widely distributed and even when possessed by a small number of firms, enormously large, even by the standards of big data. In litigation matters involving credit card transactions or social media postings, there can be billions of records.

For example, in the U.S. in 2020 alone, there were 124 billion transactions on certain credit cards and debit cards[2] spread across 11 million merchants.[3] There were 1.93 billion daily average users of Facebook in December 2021.[4] If that number held throughout the year, and if each daily active user created on average five posts, comments or likes per day, there would have been 3.5 trillion such interactions in 2021.

It may not be practical to collect information on all card transactions or all Facebook interactions for analysis, but even when possible, the benefits of collecting, storing and analyzing all the data may be outweighed by the costs. The best approach in such situations could be to review data in a properly chosen sample.

Considering Nuances of the Data

No machine or technology can answer certain questions about data, such as how they were compiled and whether they are useful for a particular analytical question.

In many instances, a company may collect data as part of the normal course of business with one purpose in mind. Although the data can be very useful for that purpose, those data may need to be audited and supplemented to become useful to answer the question posed in litigation.

Or the data may not be relevant enough for the question at hand. Knowing what is in the data — and conversely knowing what is not in the data — is critical because not accounting for those factors can result in erroneous conclusions no matter how much data there are or what data tool is used.

For example, in False Claims Act litigation with Medicare and other health data, there may be concerns about publicly disclosing personally identifiable or protected health information. As a result, many health care databases mask or exclude infrequent events, such as medical diagnoses of a rare disease.

Then, for instance, not taking a rare medical condition into account for the analysis can cause the prevalence of certain conditions computed from the available data to be incorrect and misleading. The province of the data can determine how it affects the analysis.

Avoiding Distorted Results

Outliers and erroneous data values can bias conclusions and significantly skew results. With a small data set, potential outliers and apparent errors may be visually perused and individually evaluated in a relatively straightforward manner. However, with big data, outliers and apparent errors can easily number in the thousands or even millions.

For example, in a data set of 1 billion records, 1 million records would represent just 0.1% of the data. Visually perusing or individually evaluating 1 million records is likely not feasible, so more sophisticated and complex procedures are necessary to properly evaluate such records.

With relatively small data sets — such as monthly pricing data spanning a few years — key points can be easily illustrated. A basic line chart showing the month-to-month change in prices may be sufficient to show trends or patterns.

In contrast, determining the key visual to identify and subsequently depict the pattern of interest is much more challenging when the data include millions of daily transactions across geographies, products and customer types. In such situations, all the relevant permutations of the available information need to be taken into account.

Conclusion

Modern technology can be used to harness the power of big data in litigation matters. What would have been unworkable just a few years ago today is readily achievable now.

Attorneys need to make sure they understand what data should be collected and what the right tools are to analyze them for accurate and timely results.

Jorge Gallardo-García, Benjamin Scher and David Barth are partners at Bates White LLC.

clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.

[1] Hadoop is an open-source framework used to efficiently store and process large data sets ranging in size from gigabytes to petabytes of data; instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive data sets in parallel more quickly. Spark is another open-source unified analytics engine for large-scale data processing that provides an interface for programming clusters with implicit data parallelism and fault tolerance. Databricks is a Big Data processing platform that provides a just-in-time cloud-based platform for big data processing.

[2] Board of Governors of the Federal Reserve System, "Federal Research Payments Study," updated Jan. 14, 2022, https://www.federalreserve.gov/paymentsystems/december-2021-findings-from-the-federal-reserve-payments-study.htm.

[3] CNBC, "99% of Merchants in the U.S. Who Accept Credit Cards Now Take American Express," Oct. 19, 2021, https://www.cnbc.com/select/american-express-merchant-acceptance/.

[4] Meta, 2021 10-K, https://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/14039b47-2e2f-4054-9dc5-71bcc7cf01ce.pdf, p. 56.