Big Data – What It Is And Why You Should Care

In a series of articles originally published by InsideCounsel Magazine, David J. Walton discusses the concept of big data and explores its impact on the way we conduct business.

David J. Walton
Co-Chair, Privacy and Data Security Group
610.832.7455
dwalton@cozen.com
# Table of Contents

Technology: Why Big Data is a Big Deal for Lawyers (February 14, 2014) ................................................................. 1

Technology: All Databases Are Not Created Equal, and Counsel Should Know the Difference (March 7, 2014) ............................................................ 2

Technology: How Exactly Are Businesses Using Big Data? (March 14, 2014) ........................................................................................................ 4

Big Data Raises Big Legal Issues (March 28, 2014) ........................................................................................................ 5

How Lawyers and Law Firms Operate in a Big Data World (April 11, 2014) ........................................................................................................ 7

You Thought ESI Was Complicated – Now Add Big Data (April 25, 2014).................................................................................... 9
Technology: Why Big Data is a Big Deal for Lawyers (February 14, 2014)

Big data is really about data analytics — sophisticated algorithms that are being applied to incomprehensibly large volumes of data

Yahoo CEO Marissa Mayer said that “big data” will have a bigger impact than the Internet. Consider how the Internet completely changed our lives. It’s hard to imagine anything, let alone the vague concept of “big data,” having that type of impact.

Yet, if you have read any article the past year on a legal technology issue, you have undoubtedly heard about big data. There’s still a lot of confusion about big data, its power, its potential and what it means for lawyers. This article is the first in a series that will explore these issues and illustrate why big data really is (and will continue to be) a big deal for the legal profession.

The first step to understanding big data is to define it. Many people think big data just means a lot of data. That’s only partially true. It is generally accepted that big data “refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.” Yet, at its core, big data is really about data analytics — sophisticated algorithms that are being applied to incomprehensibly large volumes of data. We create a staggering amount of data each day. For several years, computer scientists have been developing more and more powerful ways to harness the incredible volume of data for all sorts of purposes, such as marketing, medical research and business intelligence. This is not a recent phenomenon. The big data revolution is a quiet one. It has been going on for years, right under our noses.

Every time we visit a website or send an email, it is likely that some computer somewhere is tracking our movements and adding to a database that contains our online profile. Researchers use these databases, through highly complex mathematical algorithms, to find patterns in data so they can predict future buying preferences and decisions based on our online activities. This information is then used to sell highly focused and effective advertising. This type of data analytics has been going on for years, but many of us have been completely oblivious to it.

Big data has become today’s next phenomenon because the science behind data analytics has continued to grow and is now being used in numerous areas of our lives — more than just advertising. At the same time, our ability to analyze data has improved, the amount of data we create is increasing dramatically, and our ability to store, process and transfer that data has improved tremendously. We have so much data about so many different aspects of the world, and we now have the capacity to store and collect it. This is a dream for big data researchers. They are figuring how to combine and review these immense data sets together. The result is that they are finding patterns in human conduct and nature that would have never been found without the ability to analyze these large data sets.

In the past, in order to discover or research something new, researchers would postulate a theory, gather data to test it, use statistical sampling to extrapolate from that data and then reach a conclusion. But this process has a major limitation: The researcher must pose the questions before the sample data is collected.

Big data is fundamentally changing this process. Rather than creating a theory and gathering sample data to test it, which may in itself skew the results, researchers are gathering massive amounts of data and then looking for patterns and correlations. In doing so, they are letting the data speak for itself. By looking at massive amounts of data objectively (rather than sample data), researchers are now making discoveries that are not limited by human instinct and intuition. Now, this does not mean that big data replaces human instinct or intuition. But sometimes, human instinct and intuition are skewed by the natural desire to figure out why something happened; for example, why a disease starts. Instead of looking for why, however, big data focuses on what — i.e., what is likely to happen next.
In their excellent book, *Big Data: A Revolution That Will Transform How We Live and Work*, authors Kenneth Cukier and Viktor Mayer-Schoenberger discuss this dichotomy and give a great example of how big data is being used to predict the what in medicine instead of the why. Researchers in Canada used big data to spot infections in premature babies before any overt symptoms appeared. They took 16 vital signs, like heartbeat, body temperature, respiration and blood-oxygen levels, and turned that into a stream of information with more than 1,000 data points per second. Using this data set, they were able to find correlations and connections in the data that helped predict the existence of an infection before it surfaced. Big data doesn't explain why the infection starts, but it can help predict what is likely to happen next when certain factors are present at the same time.

In this sense, big data is giving researchers a view of the world never seen before. We are moving from a world where data was used to explain or support a discovery to a world where data — its connections and correlations — is the discovery. In this sense, big data is a collision between math and sociology that promises to change the way we see and analyze our world.

Today, businesses of all kinds are using big data to improve customer service, analyze their competition, manage supply-chains, monitor customer markets, follow societal trends, maintain employee relations, target advertising, find emerging markets, and expedite product innovation. Because technology developments are making it easier to derive value from analyzing data, sophisticated businesses are focusing more than ever on data analytics. These analytics are being used to generate revenue. This monetization of data is fueling the big data explosion. And, as we will explore in upcoming articles, this is the main the reason why lawyers need to understand big data.

**Technology: All Databases Are Not Created Equal, and Counsel Should Know the Difference (March 7, 2014)**

To explore current data, it is helpful to have a little background in the history of data management and analysis

The amount of data we are creating as society has exploded over the last decade. Consider this fact: Each day, we create more than 70 times the amount of information in the Library of Congress. Or this one: Approximately 2.5 billion Internet users generate 2.5 quintillion (2,500,000,000,000,000,000) bytes of data every day. Why are we producing so much data? Because we can.

Bandwidth, computer memory and computer-processing capabilities have improved exponentially over the last decade. By 2016, it is estimated that the gigabyte equivalent of all movies ever made will cross global IP networks every 3 minutes. The average smartphone now has more computing power than NASA did when Neil Armstrong landed on the moon. At the same time, each one of us is a walking content generator. Our use of the Internet, social media and mobile devices is creating a tsunami of electronic data. And, as mobile devices get smaller, faster and more powerful, they will enable us to generate even more bytes of “likes” every year.

While the mere existence of so much data is interesting on a phenomenological level, it is not, in economic terms, worth much. The key development of big data analytics is our growing ability to turn this data into valuable information. In order to understand these analytics, it is helpful to have a little background in the history of data management and analysis.

When companies moved from paper records to electronic records, they needed a system to store, manage and analyze data — and thus, **structured data** was born. Structured data is digital information that has been organized into a common and intentionally designed structure or scheme. Examples of this include stock-trading data, customer relations management systems, customer-sales orders, supply-chain documentation and inventory data.
All digital data that could not be put in a form that was easily manipulated or analyzed became known as *unstructured data*. Both humans and machines generate unstructured data. Many of today’s software applications and electronic devices create machine data that users do not even know about. This machine-generated data typically contains information regarding applications’ or devices’ status and activity. For example, smart meters automatically send data regarding electronic usage by a household to a server located at the electric company. Other examples of machine-generated data include search data, network data and health monitor or medical device data. This machine-to-machine communication is becoming known as “the Internet of Things.”

Compared to machine-generated data, human-generated data is infinitely more difficult to manage and organize. It varies widely in its structure, format, nomenclature and style. It is also more context dependent than any other data source. Often, it is necessary to understand something about the data’s context in order to understand the data itself. Examples of unstructured human-generated data include emails, text messages, video files and social media feeds.

As long as there has been digital data, businesses have been trying to analyze it. In the 1960s and 1970s, IBM and Oracle developed *relational database* software. A relational database is simply a table with rows and columns that allows the user to categorize, compare and analyze data. Relational databases can also be linked together to form several layers of related data. These databases are still the primary means for businesses to analyze data. An Excel spreadsheet, for example, is a basic relational database.

One of the key characteristics of a relational database is that it does not analyze data in its native or original form. In other words, relational databases require that data be entered into the database after it has already been analyzed and processed by the user.

The problem is that processing data in order to log it in a database can take a long time and be very expensive. If a company wants to analyze its customer relations data using a relationship database, it has to make an investment in IT infrastructure and staffing to build the database, link it to other data systems within the company, and build the interface. Because this process is so resource intensive, companies got in the habit of only saving data that was immediately valuable, cost effective to analyze, or had to be kept for risk-management purposes. It simply wasn’t worth saving data where the potential cost of preserving, collecting and structuring the data outweighed its immediate analytical value.

Another limitation of having to process data before it goes into a relational databases is that decisions about the design of the database have to be made before any data has been analyzed. In essence, you have to ask the questions first and review the data later. Thus, the value of a relational database depends on whether you’ve asked or anticipated the right questions from the beginning.

What has changed over the last several years is our ability to analyze unstructured data. Based on advances in computer technology and computer science, we can now analyze massive amounts of data in real time using analytics that are not limited to relational interfaces. In other words, big data analytics lets us analyze data in its native state without having to boot-strap it into the columns and rows of a relational database. In a macro sense, this greatly enhances our ability to analyze more kinds of data much faster. But even more important, because we can review the data in its native state, we can see patterns and relationships that are not limited by prior suppositions, biases or assumptions. Instead of defining what data is relevant before seeing the data itself, as occurs with a relational database, we can now let the data speak for itself.
Technology: How Exactly Are Businesses Using Big Data? (March 14, 2014)

The use of big data will continue to increase exponentially — as will its applications

Today, almost every large company collects data about its customers — reams and reams of raw, unstructured data. And they aren’t storing it for posterity. They are using it to do what businesses always try to do: Sell more widgets. More specifically, companies are using big data to identify new customers, advertise more effectively, and develop new products and services.

An apocryphal-sounding story about how companies are using big data to find customers recently made its way around cyberspace. Target tracks customers’ purchases to determine what products to advertise. According to The New York Times, an enterprising researcher at Target thought to analyze purchase data for women and developed a list of 25 products that, when considered together, generated a pregnancy-prediction score. The researcher analyzed every woman in Target’s database, and Target sent women with high scores baby-related coupons and advertisements.

A teenage girl, using her parents’ Target Value Card, bought four items: scent-free soap, extra-large bag of cotton balls, hand-sanitizer and washcloths. Target concluded that she was pregnant and sent baby-related coupons to her home address. The girl’s father called the local Target store manager to complain. The manager apologized. But when the manager followed up a few days later, the father said he had learned that Target was right. His daughter was pregnant.

To the extent that big data is being used to identify the right customers — it is also being used to identify the wrong ones. Kreditech, a European company, uses more than 8,000 sources including social media, to create a unique credit score for consumers, which it then sells to banks and other lenders. And they have discovered some surprising correlations between social media behaviors and financial stability. For example, if your Facebook friends use all capital letters, your score is docked. Title insurance underwriters are using the same tools to evaluate specific risks.

In the same way, big data analytics are being used to reinvent advertising. The name of the game used to be broadcasting, which worked on a carpet-bomb theory: The bigger the audience, the more likely you are to hit a target. Big data has turned that theory on its head. Now it is all about narrowcasting: Knowing exactly who your target consumer is and how to reach them as efficiently as possible.

Google and Facebook have built multibillion-dollar business on their ability to provide narrowcasting services. Google, of course, is free. But through its search engine and email services, Google collects a vast amount of data that it uses to create a data profile for each visitor. These profiles allow companies to strategically place their ads in front of individuals who are likely to be interested in their products. Facebook operates the same way. It encourages individuals to use its site by offering a free service. Then it tracks all user activity and creates an incredibly valuable profile for marketing purposes.

One example of narrowcasting is President Obama’s 2012 reelection campaign. The campaign hired more than 50 data analysts and behavioral science experts and housed them in a bunker-like room called “the Cave” located in the Chicago campaign headquarters. These experts worked 16-hour days over 16 months analyzing traditional campaign data from call-centers with publicly available and social media data about voters in swing states. They identified roughly 15 million undecided voters who might be persuaded to vote for the president. Using data from cable boxes, the analysts determined what those 15 million voters watched on TV. (Turns out they watched “Judge Joe Brown” in the afternoon and late-night repeats of “Inside Edition.”) Obama targeted his TV advertising to those viewers. Because Romney was not using this system, Obama’s ads went uncontested. And, because he eschewed the standard prime-time slots, the Obama campaign was able to purchase 40,000 more advertising spots than Romney and spend $90 million less.
Big data is not only driving sales, it is also driving innovation. In their book, *Big Data: A Revolution That Will Transform How We Live and Work*, Kenneth Cukier and Viktor Mayer-Schoenberger describe how this process of “datification,” taking something that has never been treated as data before and transforming it into a numerically quantified format, is drastically changing how we conduct R&D.

For example, one researcher figured out a way to assign a numerical value to 360 pressure points on a person’s back, hips and buttocks. Using software that can read these values, a car company could teach a car to recognize who is sitting in the seat. This could prevent auto theft, by requiring a code for someone whose “seat” is not recognized. In another case, IBM patented a special type of floor that datafies foot traffic. By digitizing a footprint, the floor can determine when a person walks into the room. This information could be used to automate lights or initiate security protocols. Retailers, convention centers and sports arenas could also use this technology to measure, analyze and manage foot traffic.

And big data is not just about big profits. In their book, Cukier and Mayer-Schoenberger discuss ways in which our ability to manage full data sets is improving government services and public safety. For example, the Center for Disease Control (CDC) has historically had to rely on doctor reports of influenza outbreaks in order to track epidemics. This process had obvious drawbacks. Patients usually waited a few days before they go to the doctor with symptoms, and doctors usually waited a few days before they report to the CDC. The CDC was typically weeks behind a flu outbreak in providing vaccines.

A couple years ago, Google proposed a different approach. Using historical data from the CDC, Google compared search term queries against geographical areas that were known to have had flu outbreaks. Google found spikes in certain search terms where flu outbreaks occurred and identified 45 terms that were strongly correlated with the outbreak of flu. Google then started tracking the use of those terms and is now able to accurately predict when a flu outbreak is occurring in real time. Using this data, the CDC can act immediately.

Big data is a big deal for businesses. The use of big data will continue to increase exponentially — as will its applications. The promise of big data extends to other areas including health care and education. These new uses will raise a host of legal issues that will impact the way that lawyers litigate and advise clients — and how lawyers run their own law firms. These topics will be discussed in the next article in this series.

**Big Data Raises Big Legal Issues (March 28, 2014)**

*As the laws and regulations within the United States evolve, companies must be extremely attentive*

As companies realize the benefits of big data on their research & development, marketing, sales, branding and revenue growth, they will increasingly have to reckon with its risks. Utilizing and monetizing big data raises enormous legal questions and potential liabilities. The most salient of these legal issues, at least in the near term, revolve around privacy, regulatory compliance and duty to intervene.

When companies analyze extremely large pools of data, they often attempt to protect the privacy of individuals through “anonymization,” the process of removing or replacing individual identifying information from a communication or record. Communications and records can be made completely anonymous by removing all identifiers or made pseudonymous by assigning each individual replacement identifiers, like a 10-digit code.
Of course, stories of incomplete or ineffective anonymization are rife. In one of the most infamous incidents, the Massachusetts Group Insurance Commission released anonymized data on state employees’ hospital visits in the mid-1990’s as part of a study. In order to prove the existing limitations of anonymization, then-graduate student, Latanya Sweeny, publicly identified Governor William Weld without difficulty. Continuing her work on this topic, Sweeney showed in 2000 that 87 percent of all Americans could be identified using only three data points: birthdate, gender and zip code.

In August 2006, AOL released three months of search queries by 650,000 of its users to the public, with the hope that this data would be useful for academic research. Despite AOL’s efforts to anonymize the data, many of the users could be identified based solely on the pattern and substance of their searches. This anonymization failure was widely reported by the media and sparked significant public backlash. In 2011, users of AOL’s website brought a class action suit against the Internet giant for disclosing search queries to the public. The action was settled on May 24, 2013, to the tune of nearly $6 million, along with a stipulation that AOL maintain policies and procedures to prevent future privacy leaks.

Similarly, in October 2006, Netflix released an anonymized database of 100 million movie ratings and offered $1 million to the first team who could use that data to “significantly improve” Netflix’s recommendation algorithm. Using publicly available user ratings in the Internet Movie Database (IMDb) for 50 Netflix members, researchers were easily able to identify to a statistical near-certainty two users in the Netflix database.

Although class actions based on data breaches and ineffective anonymization are exorbitantly expensive to pursue, litigation of this type will continue to mount. Companies should exercise the utmost caution when utilizing seemingly anonymized data or they might find themselves facing significant legal troubles.

One of the biggest challenges today for companies working with big data is that the regulatory regime is in a state of tremendous flux. Lawmakers and agency officials are trying to regulate technologies that are themselves changing on a daily basis, and they are trying to satisfy competing demands for privacy protection and commercial freedom. As the laws and regulations within the United States evolve, companies must be extremely attentive.

And the domestic legal landscape is just one of many relevant jurisdictions. Every country has its own patchwork of laws and regulations that concern data and privacy. Keeping track of all of these laws in real time is nearly impossible. Merely keeping track of where the data resides is a job in and of itself. As data warehouses manage their load balance, they can, without users’ knowledge, shift data from one data center to another. Those data centers may be located in completely different parts of the world and each governed by a different regulatory scheme.

The difficulty of tracking data, managing data and protecting privacy in an international economy will only intensify over the coming years. As the Internet grows and more people have access to mobile devices and broadband frequencies, data proliferation will increase. Workers and data will be utterly globalized. Governments will try to keep pace, and laws and regulations will abound. These issues will not be the province of privacy lawyers alone. Litigators in general will need to understand how to advise clients about privacy and data protection, how to access data that resides on foreign soil, and what rights they have to use “foreign” data in U.S.-based litigation.

Another very interesting, although nascent, legal question for corporate users of big data is whether the predictive capabilities of big data analytics impose greater duties to identify risks and intervene before incidents occur. In other words, if companies use big data analytics to look at historical data to predict where problems, accidents or financial irregularities are likely to arise, do they have a greater duty to act to prevent problems before they cause injury? If so, how will big-data applications be used to prove that notice existed and that the company should have acted sooner? And if big data applications do not analyze the data correctly, are these providers liable for failing to identify the potential for injuries or unfortunate events?
Many of the questions raised here cannot be answered right now. And big data will undoubtedly give rise to other as-yet-unforeseen legal challenges over the next decade. It is important, however, for lawyers to be thinking about these issues and preparing clients for the legal realities of commerce in a big-data driven world.

**How Lawyers and Law Firms Operate in a Big Data World (April 11, 2014)**

Most observers agree that big data is likely to have more an effect on the practice rather than the business of law.

Big data has sparked a revolution in how corporate America conducts research, identifies customers, advertises itself and pursues profits. But just as big data has enormous implications for corporate clients, it has equally important consequences for the business and practice of law — *business of law*, meaning how firms are structured, managed and financed, and *practice of law*, meaning how attorneys perform legal work on behalf of clients.

Compared to other modern professional services organizations, law firms generate very little actual data. In addition to providing great fodder for lawyer jokes, this fact means that law firms are unlikely to use internally produced big data to drive client outreach or management strategy.

But it doesn’t mean that big data won’t affect how law firms do business. Already, some of the most sophisticated clients are using big data to manage their legal spending. Some clients are vetting law firms based on meta-analyses of bills; others are holding reverse auctions and using requests for proposal heavily based on law firm use of technology. Some clients, especially insurers, are using extensive relational databases to supervise litigation and manage costs. On the backend, clients are using big data to analyze their bills. They can determine whether there were any violations of client guidelines, and reject fees and expenses that violate set parameters. They can also do comparative analyses to make sure that attorney tasks are being billed in line with market trends and historical precedent.

Most observers agree, though, that big data is likely to have more an effect on the practice rather than the business of law. In an important sign of the times, in August 2012, the American Bar Association (ABA) House of Delegates adopted recommendations by the ABA’s Ethics Commission regarding the definition of “competence.” It has been amended as follows: “To maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology, engage in continuing study and education, and comply with all continuing legal education requirements to which the lawyer is subject.” In other words, being a dinosaur when it comes to technology is no longer idiosyncratic; it’s grounds for malpractice.

Recognition of that fact comes not a moment too soon. Big data is already changing how various types of litigation are conducted, including employment, insurance, intellectual property, antitrust and general commercial litigation. One area where its influence is keenly felt is in discovery. Put simply, formal document requests are becoming irrelevant as they are increasingly replaced by Rule 26 conferences. At these conferences, attorneys agree on the names of ESI custodians, the sources of ESI to be collected by these custodians, and the methods for searching each source. ESI vendors will likely develop applications that will automatically identify potentially relevant data. Advancements in this area could profoundly alter calculations about when to pursue litigation. Being able to avoid common discovery disputes would streamline the process of preparing for trial, reduce costs and potentially make it possible for certain cases to go to trial, rather than be forced into settlement due to discovery cost alone.
To engage in trial work lawyers must understand the nature of electronic data — where it exists and how to obtain it, authenticate it and get it admitted. More specifically, though, lawyers today must also understand the potential value of social media evidence, how to access unstructured data, and when to marshal big data analytics to build a case. Many types of data automatically generated by social media applications and mobile devices constitute a potential treasure trove of evidence that most attorneys currently ignore. For example, in a claim for emotional distress, an attorney can gather a great deal of evidence about the plaintiff’s everyday activities based on his or her use of social media and mobile devices. Once identified, practitioners need to understand how to get this evidence admitted or excluded.

Big data might at some point even be an important tool for setting litigation strategy. Because it uses historical analyses to predict future outcomes, big data could be channeled to predict the outcome of individual lawsuits. At least one company is already giving it a try in the intellectual property realm. Lex Machina is a Menlo Park, Calif.-based data-sourcing company that has compiled data from 147,000 cases, 120,000 filings, 16 million docket entries, 185,000 documents, 1,400 judges and 28,000 decisions. Using data analytics, Lex Machina churns through this material to try to predict the likely outcome of current intellectual property cases. Can similar types of analytical models really be that far off, especially in the areas of employment and personal injury?

Another area of potential impact is on the process of jury selection. There are already applications that help lawyers track information about prospective jurors. Big data adds the possibility of having the ability to pull information about prospective jurors from their publicly available data in real time. Instead of relying on self-reported information, trial lawyers in the near future might know a lot more about those 12 strangers sitting in the box. And if political candidates can use such data to target their messages, perhaps attorneys will use it to mold their arguments as well.

Big data is already influencing the kinds of arguments made in class actions and other lawsuits that typically invoke statistical sampling, which has historically been used to extrapolate about both cause and effect. To the extent that big data, by making it possible to review the entire data set and not just a sample, trumps probability-based calculations, will statistical sampling still be allowed? The answer likely depends on the specific factual and legal issues in dispute. But there can be no doubt that the use of big-data analytics will create many opportunities for lawyers to tell a much richer and more real story than they have in the past.

Finally, information governance is likely to become a central practice area in most law firms, as its importance to C-suite executives continues to increase. It is not yet clear whether particular lawyers will specialize in information governance or whether every lawyer will need to participate in data security and data management. (Remember that when the Internet was first developed, many people predicted that there would be Internet practice groups at every major law firm.) At the very least, every lawyer will need to understand the basics of big data so that he or she can advise clients about new technology and comply with ethical obligations.
You Thought ESI Was Complicated - Now Add Big Data (April 25, 2014)

There are seven core issues concerning the interplay of ESI and big data that attorneys today must study and comprehend

And you thought the issues swirling around electronically saved information (ESI) were difficult to understand and even harder to manage. Now throw big data into the mix, and it becomes clear that attorneys have their technological work cut out for them. To simplify as much as possible, there are seven core issues concerning the interplay of ESI and big data that attorneys today must study and comprehend: storage structure; balance between big data analytics and document destruction; importance of centralized data sources; use of big data tools with integrated litigation hold and ESI functionality; applicability of technology-assisted review and predictive coding; management of ESI created in real-time; and proportionality.

STORAGE STRUCTURE

Finding the right system of record is particularly important in the context of big data, given the sheer quantity of information at issue. Attorneys should ask clients about how data is stored and retained. Creating a map of client data will enable more effective action if and when litigation begins. As part of the initial review, attorneys should identify the “systems of record.” A system of record is the central data source; data might be stored in other places as well, such as mobile phones, archive storage or tablets, but the system of record is the most central data source. By locating the right system of record, you can streamline your preservation and collection efforts and ignore truly redundant data sources.

BALANCE BETWEEN BIG DATA ANALYTICS AND DOCUMENT DESTRUCTION

The big data revolution has created an incentive for clients to keep and store much more data. But by saving (or over-saving) fodder for big data analytics, companies are creating a new e-discovery challenge: keeping discovery costs manageable. The more data that exists, the more time-consuming and expensive it is to preserve, collect and review when litigation arises. With the help of counsel, companies need to reach a balance between retention and destruction. Savvy lawyers have started to advise their clients to delete as much unusable data as possible in a careful and systematic way that ensures no potentially relevant data is deleted once litigation has become reasonably foreseeable.

IMPORTANCE OF CENTRALIZED DATA SOURCES

A key aspect of effective data preservation and collection is the use of centralized data sources. Having all data of a certain type contained in one place makes it easier to find, identify, preserve and collect. Especially in the era of big data and vast accumulation of ESI, lawyers should encourage clients to establish a centralized data source from the start. This will make it much easier to manage data in the event of litigation.

USE OF BIG DATA TOOLS WITH INTEGRATED LITIGATION HOLD AND ESI FUNCTIONALITY

Many newer relational database tools now have “baked-in” litigation hold and ESI capabilities, which make it much easier to preserve and manage data during a lawsuit. Some creators of big-data applications are likely to do the same — but not all. People who develop big data applications are not concerned with having to produce the same data in a lawsuit. Their focus is on getting value from analyzing the data and doing it as quickly and inexpensively as possible. Thus, advise clients to use big data tools that have litigation hold and ESI capabilities “baked” into the application.
APPLICABILITY OF TECHNOLOGY-ASSISTED REVIEW AND PREDICTIVE CODING

Big data analytics already are being employed in technology-assisted review (TAR), otherwise known as predictive coding. When using TAR, an attorney or team codes a “seed set” of documents, and the TAR program uses the seed set to predict how a reviewer would classify the document. In the past year, prominent courts have approved the use of TAR to conduct discovery. Litigators must understand the advantages and disadvantages of using TAR in a given case. And TAR is not just for data collection and production anymore. One major TAR vendor is launching a new product that can be used for document retention and management, particularly for companies using big data. A seed set is created for the type of documents that an enterprise wants to retain, and the system independently searches internal networks for relevant documents.

MANAGEMENT OF ESI CREATED IN REAL-TIME

One of the most difficult aspects of managing ESI during active litigation is preserving, collecting and reviewing new data that is created after the litigation hold has been issued. This difficulty is compounded by big data applications that introduce loads of new data into the system in real time. Lawyers need to understand what data is being fed into the system, the sources, its format, what changes occur after it enters the database, and whether a big-data database is the system of record. In some circumstances, a report from a database can be used instead of preserving and producing a huge amount of raw data.

IMPACT OF BIG DATA ON PROPORTIONALITY

The Federal Rules address proportionality and basically state that the burden of a particular discovery request has to bear some relation to the size, scope or severity of the issues at stake in the litigation. Proportionality is a very hot issue in the ESI world, and new amendments to Rule 26 may expand the principle to the preservation of potentially relevant ESI. It is not difficult to imagine litigants trying to leverage the cost of preserving and producing big data to force early, unjustified settlements. Defense lawyers must be ready to challenge discovery of big data by arguing that the cost of producing this information is disproportionate to the nature of the case, the amount in controversy, and the issues at stake in the litigation. When faced with an asymmetrical big data-ESI burden, create a plan for using the principle of proportionality to fairly limit your client’s discovery obligations.

It's an open secret that many lawyers have tried to avoid learning about the technical aspects of electronically stored information (ESI). But how can a lawyer help clients with document preservation if he doesn’t understand the structure of that client’s data? How can a lawyer negotiate format issues with adversaries if she doesn’t understand the specifics of the data being sought or produced? And how can a lawyer prevent an adversary from taking advantage of an asymmetrical burden, if she cannot apply Rule 34 of the Federal Rules of Civil Procedure to big data? In a world where data is king and the vast majority of data is electronic, ignorance is no longer an option.