MAYER·BROWN

ELECTRONIC DISCOVERY & INFORMATION GOVERNANCE

Tip of the Month

Big Data 3: Structured Data in Discovery and Class Certification

Scenario

A large e-commerce company is facing a putative class action lawsuit in which the plaintiffs allege that the company posted misleading and deceptive price comparisons on its website and that, as a result, the plaintiffs overpaid for the products that they purchased from the company. The plaintiffs have just moved for class certification, and they have asked for all of the company's pricing and sales transactions data, which are stored on the company's proprietary multi-terabyte relational databases. The company has asked counsel for advice on how to identify, preserve, collect, process and produce this data in responding to the plaintiffs' written discovery and how to offensively access and use this data in the company's own efforts to defeat class certification. To complicate matters, the data contains confidential financial and proprietary business information as well as private customer information such as customer addresses, telephone numbers and credit card numbers.

Structured Data in Litigation

Structured data resides in a defined field within a database record or file, most commonly in a relational database or spreadsheet. Examples of structured data include databases maintained in programs such as Microsoft Excel and Access, Oracle's PeopleSoft, and various relational database products from SAP, which often contain records of sales, products, employees, prices, accounting data or financial statements. By contrast, examples of "unstructured" data include memoranda and presentations, e-mails and scanned correspondence.

Structured data can be relevant to litigation in myriad ways. For instance, it may contain employee-specific information relevant in the labor and employment context or it may be a source of sales data for use in calculating damages. Structured data often plays a prominent role in fighting class certification, and companies increasingly rely on structured data to oppose motions for class certification by demonstrating that plaintiffs lack commonality and typicality. Examples of such data include: information related to compensation and salary data for all employees of a company in a putative employment discrimination class action; pricing and sales data in a putative antitrust class action; and transactions involving a company's stock in a putative securities class action.

Challenges

Identifying, preserving, collecting, processing, analyzing and producing structured data can present challenges that do not usually arise for unstructured data. For instance, structured data sources can be enormous: it's not called "Big Data" for nothing. Hundreds of terabytes of

transactional information with relatively short retention periods may be difficult to preserve and collect. Further, unlike unstructured data, once structured data has been collected, it may be necessary or desirable to validate and authenticate the data set to determine whether the data collection was accurate and complete. Companies also need to be mindful of privacy concerns, because structured data often includes private and protected information, including employee records, customer records, financial information, health records, social security numbers or credit card numbers. Analysis of structured data can be especially challenging because relational databases are capable of storing high volumes of data, which means that it can take days to run a single search query.

There are several options for producing structured data in litigation. These options include:

- Providing images or snapshots of the entire database or portions of it;
- Transferring certain fields from the database into a new database that the plaintiffs can access;
- Allowing access to the company's existing database systems;
- Running reports from the database; and
- Providing data in comma-separated value ("CSV") text files.

However, none of these options is perfect. For instance, producing images of the data in the form of PDFs is generally not helpful for the opposing party. In addition, providing snapshots and exports of the database can be meaningless when they lack context, such as the relationships contained within the system and the way in which the system evolves over time. Importing data into a new database that the plaintiffs can access is often technically difficult, especially if the data is stored in proprietary relational databases. Generally, creating a new database environment to host existing data is unwieldy and burdensome. Allowing access to the company's existing database systems is similarly problematic because plaintiffs will have access to the company's proprietary and confidential business information, as well as confidential customer and employee information.

Strategies and Best Practices

Given the challenges associated with structured data, below are several strategies and best practices for producing structured data.

Reports From the Database

Some databases are set up to generate regular summary reports that are produced within the ordinary course of business. Such reports may summarize, for example, the company's financial and sales data during the previous fiscal quarter. When the reports are created during the ordinary course of business, it may be sufficient to produce reports generated by the databases. Courts may also require defendant companies to generate custom reports from proprietary databases.

Use of Statistical Sampling and Consulting Experts

When data resides in large relational databases that contain up to hundreds of terabytes of data, it can take days to run a search. Under these circumstances, statistical sampling techniques may be the most efficient way to give the parties a random, smaller sample of the larger dataset. Statistical sampling allows parties to draw conclusions for the entire population after conducting a study on a sample taken from the same population. A consulting expert can define algorithms to select random observations, or the parties may be able to agree on the algorithms to select random observations.

Providing Data in CSV Text Files

After selecting a smaller sample of data to produce, there is still the question of how to provide the data to the opposing party. Data reports are often provided in CSV text files. CSV files are a widely accepted format for moving data between databases and store data in plain text such that each line of the file is a data record containing fields that are separated by commas. The advantage of producing data in CSV files is that they are relatively easy to generate and relatively easy to read and analyze. The recipient can import information from a CSV file into almost any commercially available database or statistical analysis program.

In conclusion, though collection and analysis of structured data can be daunting, it is likely to continue to play a prominent role in certain litigation. Indeed, despite the difficulty and cost of collecting and analyzing such data, companies are increasingly finding their own relational databases to be valuable sources of evidence that they can use to defend themselves in litigation.

For inquiries related to this Tip of the Month, please contact Ethan Hastert at ehastert@mayerbrown.com, Linda Shi at lshi@mayerbrown.com, or Kim Leffert at kleffert@mayerbrown.com.

To learn more about Mayer Brown's Electronic Discovery & Information Governance practice, contact Michael E. Lackey at mlackey@mayerbrown.com, Eric Evans at eevans@mayerbrown.com, Ethan Hastert at ehastert@mayerbrown.com, or Edmund Sautter at esautter@mayerbrown.com.

Please visit us at www.mayerbrown.com.