opentext™

WHITE PAPER

# Modern data collection

New imperatives and critical requirements

# Contents

## Executive summary

Electronic discovery (eDiscovery)—and consequently, data collection—deal with different data types, formats, and sources. More and more data reside in the cloud and endpoints that may not even belong to the organization, raising new concerns about data privacy. Additionally, most organizations now have flexible remote work and BYOD policies allowing employees to work from any location and use their own devices.

As a result, data is scattered across devices and geographies and needs to be collected on and off VPN. Organizations still need effective and legally defensible data collection methods, but they also need to narrow the universe of ever-expanding data into a manageable volume that they can afford to review to gain rapid insights into litigation and investigations. Reducing data at an earlier stage in the electronic discovery reference model (EDRM) process can reduce costs significantly.

Modern data collection demands discreet search, collection, and preservation of electronically stored information (ESI) in a court-admissible format with visibility across all data sources. Organizations must ensure their collection solution is up to the task. This paper shares the seven key requirements for modern data collection to help organizations choose the right eDiscovery solution.

## Introduction

Data has changed since the advent of eDiscovery. It no longer looks the same or resides exclusively in discrete, well-defined local repositories. Organizations still need data collection methods that are effective and legally defensible. However, with skyrocketing data volumes and varieties, they also need an approach that will immediately winnow down data sets.

The more data that progresses to the review stage, the more it costs for software licensing, services, lawyer fees, and overall effort to reduce the data to the most accurate responsive data set. With rising litigation and time-pressured regulatory compliance concerns, organizations need to gain rapid insights into many sources of data, whatever their origin.

New types of data, mobile devices, remote work, and data privacy mandates have changed what organizations need from data collection. This paper explores how data collection is no longer just about eDiscovery for active or anticipated litigation. It is becoming increasingly important for ongoing compliance and regulatory requests and investigations, both internal and external. Modern data collection must be:

- **Comprehensive**, accounting for all data sources.
- **Fast and efficient**, automatically deduplicating and avoiding re-collection.
- **Easy to use**, collecting all data in one process.
- **Discrete**, to not disrupt the custodian's work.
- **Fail-safe**, correcting for network disruptions and other interruptions.
- **Defensible**, meeting judicial and regulatory standards.
- **Integrated**, creating outputs that can be easily transferred to standard review platforms.

## Dynamics of data collection

Data collection used to be simpler. When eDiscovery was young, most discoverable information could be found in digital formats that emulated paper, such as emails, word-processing documents, or spreadsheets. That information was located entirely within the organization's walls, on fixed-location computers and in-house network servers, and work was done almost exclusively in the office. Even when organizations provided laptops, employees rarely took them out of their offices. While organizations had an obligation to protect data, there were fewer threats to data and fewer requirements for its protection.

Today, advances in technology and a more mobile society have driven a shift in data collection toward multicloud and off-network devices, systems, and applications.

### Ephemeral data and new data types

Much of the data that needs to be collected today has no direct corollary in the world of paper. It's not a traditional "document" that might exist in a word

processing program or an email. Laptops, desktops, servers, Internet of Things (IoT) devices, and content repositories represent the backbone of a modern organization. Today's discoverable data might include messages and integrated notifications from collaboration applications like Slack™, videos, outputs from IOT devices, and countless other new types of data.

Some of that data is ephemeral or short-lived. It is rapidly deleted or replaced by new incoming data regularly, requiring any preservation or collection effort to be undertaken promptly. Employees are creating, editing, transmitting and

deleting information every day—underscoring the importance of such data to the eDiscovery process. Data collection must now encompass every potentially relevant data type from wherever it may originate.

## Mobile data and remote work

Corporations and government agencies alike have been moving away from stationary computing resources such as desktop computers, in-house servers, and intranets. Now, most employees do their work on a combination of mobile devices, such as laptops, tablets, and smartphones. Cloud adoption has accelerated. More than 94 percent of organizations with more than 1,000 employees have a significant portion of their workloads in the cloud, according to a survey of 800 organizations.[1]

The prevalence of remote work creates security challenges. Since everybody can work from anywhere and on premises is vanishing, tech teams confront new challenges to protect data and individual information. As organizations migrate to cloud networks, a shift in focus on security and compliance is needed.

1 CloudZero, *90+ Cloud Computing Statistics: A 2025 Market Snapshot*


This leads to two data collection challenges: first, collection efforts cannot be restricted to one physical location but must instead span both on- and off-network locations. Second, the widespread use of personal devices for business purposes makes it even more difficult for organizations to collect data that belongs to the business without trampling the privacy rights of individual employees.

### Data privacy regulations and laws

The last few years have seen the proliferation of a patchwork of data privacy regulations, from the EU's General Data Protection Regulation (GDPR) to the California Consumer Privacy Act (CCPA), along with myriad other state and local laws designed to give individuals specific rights regarding their personal information. According to the IAPP, 79.3 percent of the world's population is covered by some form of national data privacy law.[3] These regulations have increased the burden on organizations to ensure their data collection gives due respect to individuals' data privacy rights.

For example, many companies allow their employees to access corporate email accounts, Slack channels, and company documents from their personal devices. While any corporate data on those devices is discoverable and must be defensibly preserved and collected, the collection methods used cannot infringe on employees' privacy. Collection practices must therefore take a holistic approach, balancing the legal and business needs of the organization with the data privacy rights of the individual device owner.

In short, data has become simultaneously more complex and more widespread, implicating new privacy considerations. At the same time, organizations are seeking to collect data for litigation, compliance and regulatory requirements and internal investigations, adding another layer of pressure.

## Use cases for modern data collection

Litigation preparedness has always been a key driver for eDiscovery needs of corporations. As per an annual litigation trend survey by Norton Rose Fulbright, corporate counsel faced an increased number of regulatory disputes in 2022 compared to the previous year,[4] as agencies stepped up enforcement on issues ranging from cybersecurity and white-collar crime to worker classification and compliance with evolving healthcare rules.

However, litigation-based eDiscovery isn't the only use case for data collection. Just as litigants seek to get a handle on relevant information to assess the strength of their arguments, companies also want to understand what their data indicates about internal behavior, regulatory compliance and potential acquisitions or C-suite hires. In fact, data collection to support investigations has become a tremendous growth area. Investigations pose new challenges, such as truncated timelines requiring quick insights to inform rapid decisions.

Today, the need for data collection spans across multiple use cases, including:

3 IAPP, *Identifying global privacy laws, relevant DPAs*, 2024

4 Norton Rose Fulbright, *2023 Annual Litigation Trends Survey: Perspectives From Corporate Counsel*, 2023.

- Government and regulatory agency inquiries.
- Internal investigations into reported or suspected misconduct, including HR disputes involving discrimination or harassment.
- Due diligence investigations during mergers and acquisitions.
- Compliance investigations.
- Vetting of new hires or potential C-suite promotions.
- Pre-assessment of litigation claims.

Organizations need to have the ability to rapidly survey their data and collect a wide variety of information from a range of sources, which may be relevant to decisions regarding litigation, regulatory compliance and internal investigations.

## New requirements for modern data collection

Here is what organizations now need from their data collection approach:

### Comprehensiveness

Modern data collection methods must collect data from all relevant sources. This includes physical sources like laptops and desktop computers; cloud sources like Box.com, Dropbox™, Google Drive™, Slack™, and Microsoft 365®; servers such as Microsoft® SharePoint®; and all manner of email formats from POP3 and IMAP to Microsoft® Exchange and Microsoft® PST, whether live or archived. Modern forensic software, such as OpenText™ Information Assurance (EnCase), includes robust connectors to all of these data sources as well as remote agents for collecting data from desktops and laptops to enable comprehensive data collection.

### Speed and efficiency

A data collection method must work quickly and efficiently. It should use advanced search filtering to aggressively cull data at the point of collection; use global deduplication to avoid re-collection of data that has already been obtained, perhaps from another custodian; and automatically deNIST data sets to remove machine files and zero-byte files. This minimizes costly, disproportional over-collection in eDiscovery, where the cost of review is directly proportional to the volume of data.

Speedy, efficient data collection also helps organizations identify conclusive information for investigations where time is of the essence.

### Ease of use and insight

**Almost 50% of corporate legal professionals think meeting deadlines with increasingly more complex data becomes more difficult.**

**Of those surveyed, 32% called "leveraging technology to facilitate targeted collections" a primary step to limiting and/or controlling the overcollection of ESI.**

**Internal investigations triggered by an external factor is the most common practice area for applying AI-driven eDiscovery technology.**

2023 State of Corporate Legal Industry Report

Learn more ›

Modern data collection methods should allow a single unified process to collect data from all sources, crawl multiple target sources in parallel to expedite collection time and enable frequently used criteria to be templated and automated.Additionally, collections should be configurable to target data flexibly in line with requirements.

For example, OpenText Information Assurance supports the ability to tailor collections to specific folders anywhere within multi-tiered folder structures. If the objective is to collect all sales contracts across the organization, the parent contracts folder can be targeted. If interest is focused on only the sales contracts of a single division within the organization, over-collection can be avoided by targeting just that sub-folder.

A sophisticated data collection approach should also make it easy to gain insight into what the data indicates. With pre-collection analytics, users can rapidly understand the scope of the data. Advanced search functions should allow targeting of specific datasets within an endpoint, network or cloud source based on keywords, hash values or metadata. It should be possible to conduct collections in parallel, including splitting jobs by folder so data can be analyzed sooner as individual jobs are completed, instead of waiting for entire processes to finish.

### Discreet operation

Data collection should run quietly in the background without monopolizing system resources or bogging down routine tasks.

### Fail-safe measures

With dispersed data, sporadic connectivity is a fact of life. Data collection solutions should maintain logs of successful processes and automatically reattempt any collection that fails, such as when a device drops off a network. OpenText Information Assurance communicates with remote agents to monitor connection attempts and automatically execute retries until devices re-appear on networks and collections are completed.

### Defensible process and results

Data collection is worthless if it results in altered metadata or a broken chain of custody. Both the process and the results must be defensible and forensically sound, with rigorous adherence to the chain of custody. Data collection methods should generate legally sanctioned output formats, such as OpenText Information Assurance's LEF (Logical Evidence File), which maintains the integrity of collected data without altering the metadata.

### Easily transferable outputs

One of the core objectives of data collection is to enable subsequent data review, so data must be collected in an industry-standard format that can be readily ported to widely used review platforms, such as OpenText eDiscovery. The ability to easily transfer data collections to review platforms is an important element of containing eDiscovery costs.

## Conclusion

Given the rising complexity of data types and sources, the surge in mobile devices, remote work and new mandates for data privacy, organizations need to take a new approach to data collection for eDiscovery and investigations.

Data collection solutions should be easy to use, comprehensive, efficient, discrete, stable, and forensically sound while outputting in widely accepted formats. Fulfilling these seven critical requirements is a best practices prescription to accommodate the new mandates of modern data collection.

## About OpenText Information Assurance

OpenText Information Assurance, a forensic data collection and preservation platform, exceeds all these requirements. It enables rapid data collection from an extensive array of sources and endpoints, automatically reduces data volumes, and thereby review costs, and allows users to gain rapid insights into the scope of a matter. Users can combine keywords, hash values or metadata properties to search across all content systems without pre-indexing.

OpenText Information Assurance operates discreetly without disruption to users and corrects for network connection issues. The solution is trusted by courts and has been cited or mentioned in many US judicial opinions and publications. Exports can be readily ingested by widely used review platforms in a variety of formats, including Concordance, EDRM XML, and E01. It generates output load files that are Relativity-ready and supports streamlined uploads to OpenText eDiscovery.



opentext™