opentext™

Supporting Evolving Discovery Use Cases Through Information Assurance

This position paper will discuss the challenges that organizations face today regarding the proliferation, protection and use of data, and how an effective Information Assurance program can quickly sift through organizational data to identify, preserve and collect information from various data sources to support key discovery objectives efficiently and defensibly and give your organization an information advantage over your competition.

opentext™

Contents

Executive Summary 3

Governance Challenges Within Organizations Today 4

Increasing Sources of Data 4

Redundant, Obsolete and Trivial (ROT) Data 7

Discovery Challenges Within Organizations Today 8

Compliance Challenges 8

Identification and Security of PII Data 9

Risk Challenges 9

- Key Cybersecurity and Risk Trends 9

Evolving and Expanding Discovery Use Cases 10

Understanding an Information Assurance Approach to Discovery 11

- Differentiating “Data” from “Information” 11
- Understanding and Utilizing Endpoints 11
- The “Left Side” of the EDRM 12
- Information Assurance Defined 12

Advantages of an Effective Information Assurance Approach: 12

- Discovery Challenges Within Organizations Today 12
- Direct Collection from Various Endpoints 12
- Early Insight 13
- Proactive Data Remediation 14
- Scalability 14
- Defensibility and Repeatability 14

Conclusion 15

opentext™

Executive summary

Organizations are practically drowning in data today. We all have so much data in so many different places and a large percentage of that data isn't even useful to helping us accomplish our business goals. Data for our organizations used to be located primarily on in-house workstations and servers; today, much of it is in the cloud within enterprise systems for everything from office productivity and email solutions like M365 and G-Suite, to customer relationship management (CRM) and human resources (HR) systems, to collaboration platforms, such as Slack and Microsoft Teams – and much more!

Business data is also regularly stored on mobile devices, which practically everyone uses today. And the sources of data are continuing to evolve, with the emergence of data from Internet of Things (IoT) devices and the specter of "the Metaverse" looming in our future.

Not only do we have so much data that using it effectively is a challenge, but the risks associated with protecting that data have never been higher, due to strengthened data privacy laws worldwide and increasing cyber-attacks on businesses everywhere. Yet, we're also expected to use that data in more use cases for discovery than ever, especially to support activities ranging from litigation and investigations to support of Second Requests in merger & acquisition activities.

In short, organizations today have too much data, but not enough timely information to support ever-evolving governance and discovery needs.

Governance Challenges Within Organizations Today

With business communications evolving to include collaboration apps, governing data has never been more challenging, given the volume of data within organizations, the increasing variety of sources of organizational data and how much of that data provides little if any benefit to those organizations.

Growth of Data

The term “Big Data” has become popular to illustrate the volumes of structured and unstructured data (predominantly unstructured) that overwhelm organizations on an increasing basis. And the growth of data within organizations has been stratospheric, doubling every 1 to 2 years. Back in 2010, the amount of data in the world was about 2 zettabytes: today it’s about 97 zettabytes! That’s over 48 times more data in 12 years! And it’s expected to almost double again by 2025 – to 181 zettabytes (which is 181 trillion gigabytes)!

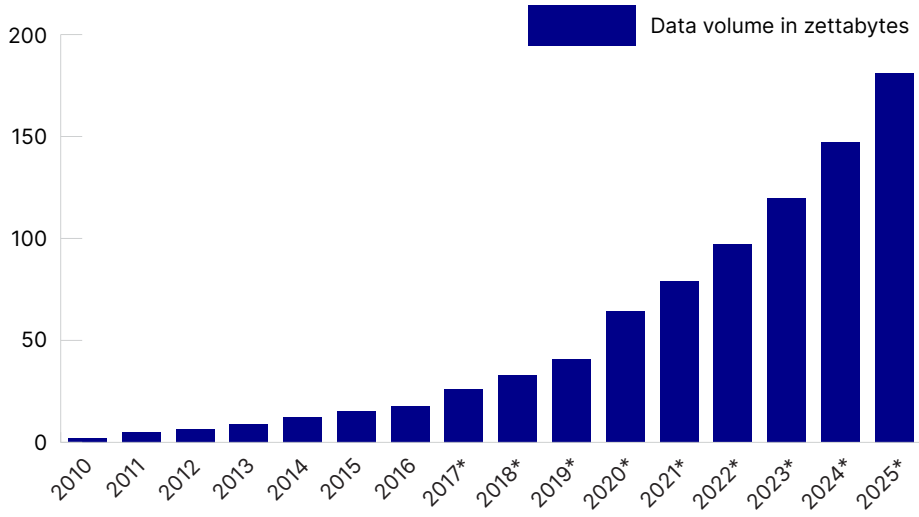

Figure 1: Volume of Data in the World from 2010 to 2025 (Source: Statista)

That level of data proliferation has completely overwhelmed organizations everywhere and has been impossible for many of them to manage and govern effectively.

Increasing Sources of Data

One of the biggest reasons for Big Data growth is the increased variety of sources of that data today, adding complexity to governance and discovery initiatives. Many organizations operate predominantly in the cloud, generating work product and communicating and collaborating within the cloud more than ever. If they’re not in the cloud, they’re on mobile devices continuing to communicate and generate additional work product. And data sources continue to evolve as our use of technology changes with data from IoT devices and the Metaverse likely future data sources we will need to address.

opentext™

Cloud Sources

Global cloud adoption is expanding and will continue to expand rapidly as organizations move to the cloud for their enterprise solutions. According to Gartner, end-user spending on public cloud services will grow from $314 billion in 2020 to $482 billion in 2022, a growth rate of over 53% in just two years. As a result, Gartner predicts public cloud spending will exceed 45% of all enterprise IT spending by 2026, up from less than 17% in 2021.

	2020	2021	2022
Cloud Business Process Services (BPaaS)	46,066	51,027	55,538
Cloud Application Infrastructure Services (PaaS)	58,917	80,002	100,636
Cloud Application Services (SaaS)	120,686	145,509	171,915
Cloud Management and Security Services	22,664	25,987	29,736
Cloud System Infrastructure Services (IaaS)	64,286	91,543	121,620
Desktop as a Service (DaaS)	1,235	2,079	2,710
Total Market	**313,853**	**396,147**	**482,155**

Figure 2: Worldwide Public Cloud Services End-User Spending Forecast in Millions$ (Source: Gartner)

And that's only **public** cloud services, it doesn't account for the large number of **private** cloud environments that are out there.

Collaboration and Remote Work

One of the biggest reasons for the growth in cloud usage is the growing popularity of collaboration apps for communication and web conferencing. Already growing in popularity before the COVID-19 pandemic, their use skyrocketed when the pandemic forced so many organizations into remote work during the pandemic. For example, the number of daily users of Zoom catapulted from 10 million in December 2019 to 300 million by April 2020!

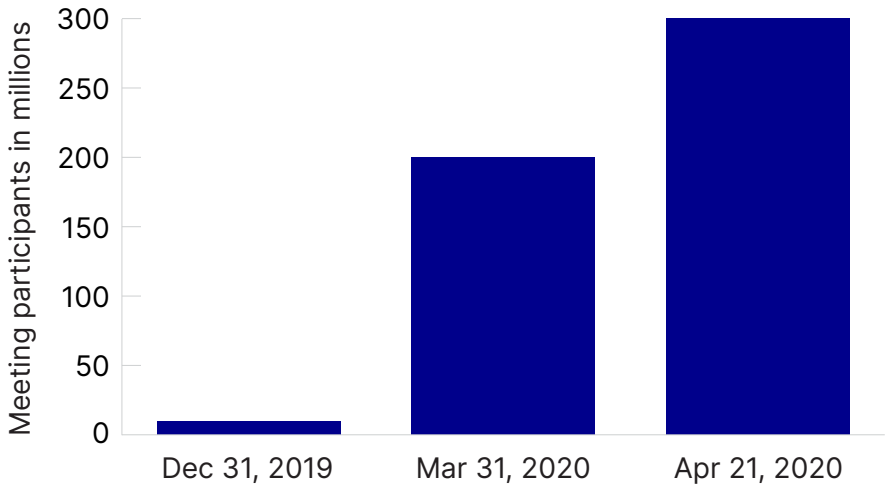

Figure 3: Daily Users in Zoom Meetings in Millions (Source: Statista)

And the number of Microsoft Teams daily users has grown from 2 million back in 2017 to 145 million by mid-2021!

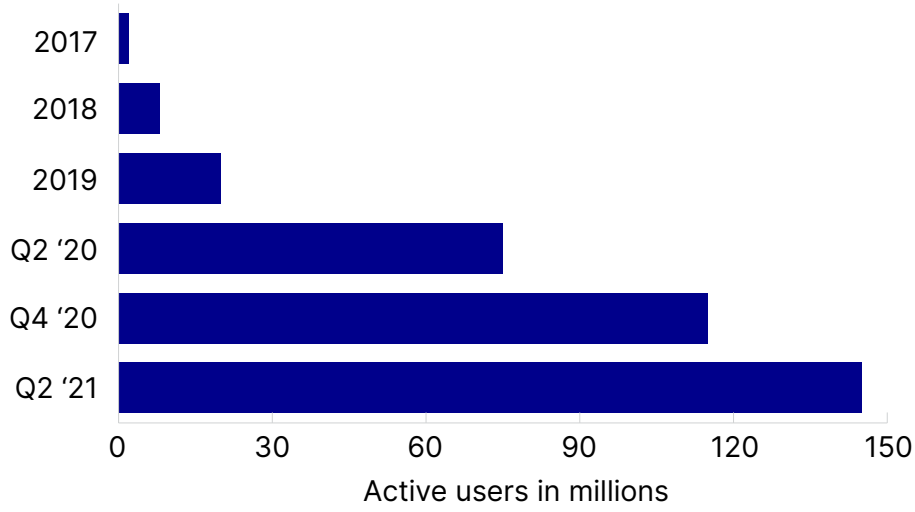

Figure 4: Daily Active Users of Microsoft Teams in Millions (Source: Statista)

Collaboration apps like Zoom, Teams and Slack have become ubiquitous in organizations today and they are starting to be routinely discoverable in litigation, as this case illustrates.

Ephemeral Data

Not all chat and communication apps keep data forever and most can be set to automatically delete data within a certain time frame. Use of ephemeral apps like Snapchat, Signal or Telegram can reduce the volume of communications ESI to govern, but that capability may be counter to enforcing your records retention policies, and the use of them once there is a duty to preserve ESI can be problematic as well. It's important to have a plan and formal policies for the use of ephemeral apps before and during litigation as reducing potentially responsive data during litigation can lead to sanctions.

Mobile Devices

Who **doesn't** use mobile devices today? Practically all of us do. As of 2021, there were 6.4 billion smartphone users worldwide. With global population of 7.9 billion, that means more than 4 out of 5 (80 percent) of humans use smartphones. In the US, we use smartphones an average of 2 hours and 55 minutes daily and all mobile devices 3 hours and 54 minutes daily.

And that mobile device use is more frequently for business purposes, including text messaging and other communications and work product for which the data may not be available anywhere else.

New Data Types

And new data types continue to be identified. That includes data from IoT devices (which is regularly important in criminal cases and starting to become more important in civil cases as well) and will eventually include virtual reality data from the Metaverse. Change is a constant when it comes to data within organizations.

Redundant, Obsolete and Trivial (ROT) Data

Unfortunately, much or most of that data has little to no value to the organization holding on to it. It's either Redundant, Obsolete or Trivial (ROT):

- **Redundant data** is data that has duplicates stored across multiple locations. A common example of redundant data are emails that include multiple employees within the organization, as each one of them may have a copy in their email mail store.
- **Trivial data** is data that isn't necessary to store or keep and provides no value to the organization. A common example of trivial data is personal email sent and received by employees every day.
- **Obsolete data** is data that is no longer accurate or no longer in use. This might include old reports that no longer provide any value to the organization.

ROT data can be found in many different locations, including desktops, mobile devices, on-premise and cloud servers. ROT data has a huge impact on productivity and occupies a large percentage of a company's data storage. A recent survey from AIIM found up to 80 percent of electronically stored information is ROT. That's a potentially huge amount of data within organizations that provides no value.

opentext™

Discovery Challenges Within Organizations Today

In addition to the Governance challenges discussed above, there are several challenges that are adding to the discovery workload for organizations, even if they don't have any meaningful litigation. Those include compliance challenges associated with data privacy compliance, risk challenges associated with threats to your data and evolving and expanding discovery use cases.

Compliance Challenges

With the volume of personal data online, countries worldwide, including Europe and the US, have passed new data privacy laws that place a greater emphasis on data privacy protection. Part of the challenges is that each data privacy law is unique, with different requirements for protecting personal data, requiring organizations to keep "moving the target" to stay compliant with the evolving data privacy landscape. Here are recent enacted data privacy laws in Europe and the US, as well as their impact on the identification and protection of personal data.

Europe

The General Data Protection Regulation (GDPR) became effective in May 2018 and it strengthens data privacy compliance requirements more than its predecessor, the 1995 European Union Directive 95/46/EC. The United Kingdom now has its own GDPR after leaving the EU via Brexit. GDPR applies to the European Economic Area (EEA), which is the EU plus Iceland, Norway, and Lichtenstein.

GDPR applies to any organization offering goods or services to European "data subjects" or organizations controlling, processing, or holding personal data of European nationals, regardless of whether the organization location is in the EEA or not, and organizations must be able to show in clear and plain language that they obtained consent for the handling of personal data.

Fines under GDPR can be huge – up to **4 percent of annual revenue** or **20 million Euro,** whichever is greater. Since GDPR was enacted in May 2018, we have seen some significant fines related to GDPR violations, with the largest to date having been assessed against Amazon in July of 2021 of $887 million!

US

The US has no comprehensive national data privacy law. There are currently only four states that have passed data privacy laws:

- California: The California Consumer Privacy Act (CCPA) was passed in 2018 and went into effect in January 2020. Californians voted to replace it in 2020 with the California Privacy Rights Act (CPRA), which significantly expands the data privacy rights of consumers over what the CCPA covers and will replace it in January 2023.
- Virginia: In 2021, Virginia passed the Consumer Data Protection Act (CDPA) (aka VCDPA), which is set to go into effect in January 2023.
- Colorado: Also in 2021, Colorado passed the Colorado Privacy Act (CPA), which is set to go into effect in July 2023.

- Utah: In 2022, Utah passed the Utah Consumer Privacy Act (UCPA), which is set to go into effect at the end of 2023.
- Connecticut: In 2022, Connecticut passed the Connecticut Data Privacy Act (CTDPA), which is set to go into effect in July 2023.

Each of these data privacy laws has differences that force organizations to continue to adjust to the requirements of each jurisdiction. For example, while all the states currently provide right of access, only Colorado, Connecticut and Virginia provide right of rectification (California will add it when CPRA goes into effect). Every state but California currently offers the right to opt-out of processing for advertising purposes. And only California offers any private right of action. These differences create a "moving target" for data privacy compliance. And there are many states to go!

Identification and Security of PII Data

Data growth within organizations makes it more challenging to identify personally identifiable information (PII) of data subjects, especially considering the different types of PII that an organization may possess about individuals. Examples of PII relating to an individual include:

- Name
- Identification number such as Social Security Number
- Home address and other location data
- Online identifiers such as e-mail addresses and IP addresses
- Personal Health Information (PHI), such as health history and medical records
- Financial data, such as credit card and back account numbers
- Racial, ethnic, sexual orientation or religious beliefs
- Genetic data and biometric data

Some PII (such as social security numbers, phone numbers and credit card numbers) can be located and secured via pattern matching searches through regular expressions, but many other types of PII don't fit those patterns and require targeted technology for the organization's proprietary information to facilitate identification.

Risk Challenges

While the stakes are high for protecting sensitive data, the risks have never been higher due to the increased threat of cyberattacks and data breaches.

Key Cybersecurity and Risk Trends

Here are eight key statistics that illustrate the increased risk and challenges associated with protecting sensitive data within your organization:

- The number of individuals impacted by data compromises was up **564%** in the first quarter of 2021, compared to the fourth quarter of 2020 (Source: Identity Theft Resource Center®).
- From 2018 to 2019 there was a **640%** increase in phishing attempts (Source: B2C).

- Global ransomware reports were **715%** higher for the first six months of 2020, compared to the first six months of 2019 (Source: Bitdefender).
- Cyber insurance premium prices jumped as much as **40%** in 2021 — in large part due to the rise in ransomware claims (Source: CyberScoop).
- In a recent survey, **83%** of respondents said they continued accessing accounts from their previous employer after leaving the company and **56%** of respondents said they had used their continued digital access to harm their former employer (Source: Beyond Identity).
- Financial services firms are 300 times as likely as other companies to be targeted by a cyberattack (Source: Boston Consulting Group)
- **56%** of organizations do not have a cyber incident response plan. And only **32%** of the remaining 44% actually think that their plan is effective (Source: McAfee).
- In 2021, organizations experienced an average cost of a data breach of **$4.24 million**, rising from $3.86 million the previous year (Source: IBM).
- The average time taken for businesses to detect and contain breaches is **287 days** – 212 to detect and 75 to contain (Source: IBM).

These are just a few of the indicators of increased risk that organizations face today in protecting sensitive data. With so much data to protect and the stakes higher than ever, disastrous data breaches are more common than ever. Because of their involvement in Governance, Risk management and Compliance (GRC), legal teams have been placed in a leadership role to ensure a program that manages and protects sensitive data and avoids these costly data breaches.

Evolving and Expanding Discovery Use Cases

Another discovery challenge is the increased use cases for which discovery workflows can be applied, which include:

- **Litigation:** Organizations are busier with litigation than ever before, with data privacy lawsuits related to CCPA on the rise. The importance of having an information advantage over opposing parties has never been more important.
- **Investigations and Audits:** With more than $4.7 trillion lost annually to occupational fraud worldwide, the need to apply discovery workflows to internal investigations has never been higher.
- **Data Subject Access Requests (DSARs):** Because of strengthening data privacy laws, individuals now have the right to submit a DSAR to a business asking to know what personal data of theirs has been collected and stored as well as how it is being used, and the need to be able to support this is a discovery workflow to identify and capture this data.
- **HSR Second Requests:** In the case of mergers and acquisitions, Second Requests to reply to the Federal Trade Commission (FTC) and Department of Justice (DOJ) regarding antitrust concerns under the HSR Act are often voluminous and with a shortened timeframe (compared to litigation) but require a discovery workflow to provide documents in response to the Second Request.

With the number of use cases for discovery technology increasing, the ability to maximize the usefulness of organizational data – while identifying and protecting sensitive data within the organization – has grown in importance considerably. Legacy methods of conducting discovery are no longer viable to adequately address the dramatic increase in discovery burden.

Understanding an Information Assurance Approach to Discovery

To address these challenges today, organizations need to adopt a different approach – one that focuses on efficiency, identification of sensitive data quickly, identification and remediation of ROT data and minimizing the volume of data through early insight and targeted collection. Let's discuss an Information Assurance approach to legal GRC and discovery.

Differentiating "Data" from "Information"

As we noted in the Introduction, organizations today have too much **data**, but not enough timely **information**. People often use the terms "data" and "information" interchangeably, but they're not the same.

Data is a collection of raw and unorganized facts. On its own, data – without analysis and interpretation – is meaningless. According to Forrester, between **60%** and **73%** of all data within an enterprise goes unused for analytics – it remains what we call "raw data" and most of it within organizations is unstructured. Much of that data is unused for a reason – it's not useful to the organization's information needs.

Information is knowledge gained through study, communication, research, or instruction. Information is the result of analyzing, interpreting, communicating, discussing or discovering pieces of data and the value within them.

Historically, discovery workflows push unstructured **data** (and decisions about that data) downstream within the discovery lifecycle. With more data than ever to manage, these legacy workflows are inefficient and costly. The focus of an Information Assurance approach is to identify potential **information** sooner within the discovery life cycle and performing a targeted, forensically-sound collection of that information (without altering metadata), saving considerable time and costs in discovery.

Understanding and Utilizing Endpoints

Accomplishing early insight into the data to uncover the information you need involves going to the endpoints within your organization. Organization data is distributed more than ever before on a series of **endpoints** throughout the organization – workstations, servers, mobile devices, etc. These endpoints are where the data lives and they are where the information needs to be discovered to collected in a targeted, forensically-sound manner. An effective Information Assurance approach utilizes the endpoints to glean the information from the vast collection of data siloes that exist within your organization.

The “Left Side” of the EDRM

Historically, the entire custodian corpus of data (including their email, file shares, etc.) has been collected and then loaded into an eDiscovery solution where the data is filtered and culled downstream. In today’s Big Data world, that approach is no longer viable. An effective Information Assurance approach focuses on the left side of the EDRM phases – Identification, Preservation and Collection – to limit the data moving downstream to the more expensive phases of eDiscovery – Processing and Review (which includes hosting the data to be reviewed).

However, not all “left side” EDRM approaches are the same. Some are limited in their ability to collect information in a defensible, forensically-sound manner. Others are limited in the data sources from which they can identify, preserve and collect data. An effective Information Assurance approach applies a “left side” EDRM centric approach to forensically-sound, targeted collection to ALL the endpoints throughout the organization – workstations, servers, mobile devices, etc. That’s a **truly** comprehensive “left side” EDRM centric approach to discovery!

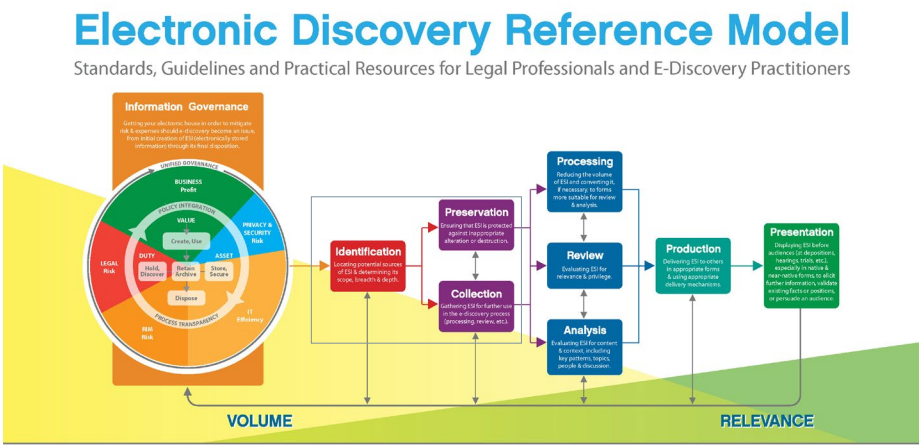

Figure 5: Electronic Discovery Reference Model (EDRM) with the “Left Side” Phases Highlighted

Information Assurance Defined

With that in mind, Information Assurance can be defined as:

Efficiently and defensibly identifying, preserving and collecting **information** from various organizational endpoint **data** sources to support key discovery business objectives.

Advantages of an Effective Information Assurance Approach

An effective Information Assurance approach has several key advantages that streamline the discovery process, including:

Direct Collection from Various Endpoints

Key to the success of an Information Assurance approach to discovery is the deployment of agents to these endpoints to support information requests remotely.

Instead of having to perform searches and collections of each of the endpoints locally and separately, the searches can be performed remotely, across many endpoints simultaneously for important, sensitive or even unimportant ROT data for remediation. For endpoints that are temporary offline, those searches are queued up until they are back online. This Information Assurance approach to discovery minimizes the **data** collected for downstream use, enabling the discovery team to focus on **information** that is much more likely to be relevant to the discovery use case.

To support those discovery use cases, it's also important to have data connectors included within the solution that can collect from email servers, document repositories and other data sources. Those included data connectors should be specific to each type of email or document store, providing secure and direct connections to allow organizations to execute seamless and efficient eDiscovery collections from across multiple sources.

Early Insight

An Information Assurance approach to discovery enables you to gather early insight on the data **where it lives** – at various endpoints within your organization.

Early insight where the data lives enables you to:

- Test search criteria before any data is collected;
- Proactively identify potential custodians and data sources;
- Search all record types including email, email attachments, calendar events, folder labels and user lists;
- Ensure that only necessary data is collected;
- Split a large folder share into smaller data "chunks" that can be collected in parallel for further processing;
- Gain insight within minutes instead of waiting weeks for solutions that first require indexing, collection and processing;

- Generate reports to clearly understand the proportionality of collections before producing them; and
- Cull at the point of collection.

In today's Big Data world, the ability to gather early insight on your data and cull at the point of collection is critical to meeting deadlines and effectively managing rising costs and is a huge advantage of an Information Assurance approach.

Proactive Data Remediation

Gathering insight about data at the endpoints serves a secondary purpose – the identification of ROT data that has no value to the organization. The ability to identify ROT within the organization facilitates location of information responsive to various discovery use cases.

But identification of ROT data isn't enough. Proactive data remediation is the identification and **removal** of data identified as ROT. Records retention and destruction **enforcement** is a necessary proactive step in controlling discovery costs. An effective Information Assurance approach applies proactive data remediation directly at the endpoints, improving the overall data health of an organization and reducing discovery costs before you even have a use case. It's a literal "life raft" to keep from drowning in data.

Scalability

In the new post-pandemic era of remote work, there are many more endpoints than ever for organizations. Centralizing this information is costly and makes it difficult to scale as repositories become larger and larger. The ability to conduct searching and culling at the endpoints is considerably more scalable as the workload is being divided across each of the endpoint nodes instead of within a centralized repository. You can literally support millions of endpoint nodes and up to a petabyte of data within an Information Assurance approach as the searching and culling process is completely distributed.

Defensibility and Repeatability

An Information Assurance approach ensures defensibility at every step in the process, providing reporting, auditing and logging to track chain of custody, while preserving collected data in a forensically sound format without any metadata being altered, using tools that are recognized in hundreds of judicial opinions from criminal and civil courts around the world. It enables you to templatize collection criteria and settings for easy re-use in future cases. Automated workflows ensure repeatability, even as the volume of data rises. **Sanity** is doing the same correct thing over and over again and getting the **same** result!

opentext™

Conclusion

There's too much **data** and not enough timely **information** within organizations today. The stakes are high, and the risks are even higher and there are more discovery use cases than ever that your organization needs to support. Organizations can't afford to continue to move ever-growing volumes of data to downstream processing, hosting and review to support those use cases. This legacy approach to discovery is no longer sustainable.

An effective Information Assurance approach provides several advantages, including direct collection from various endpoints, early insight, data remediation, scalability, and defensibility and repeatability. It addresses the Big Data Governance challenges and the Compliance, Risk and Use Case discovery challenges to support a variety of evolving discovery use cases efficiently and defensibly, giving your organization an information advantage over your competition.

Video:

- Introducing OpenText EnCase Information Assurance
- Large scale collections made simple

Product Briefs/Data Sheets:

- OpenText EnCase Information Assurance product overview
- EnCase Information Assurance data connectors
- End-to-End eDiscovery solution overview

Position papers:

- Modern Data Collection: New Imperatives and Critical Requirements
- Information Assurance and Digital Forensics: A Deepening Relationship

Customer Success Stories:

- Banner Health success story
- Novelis success story
- Liberty Mutual Insurance success story

opentext.com/contact