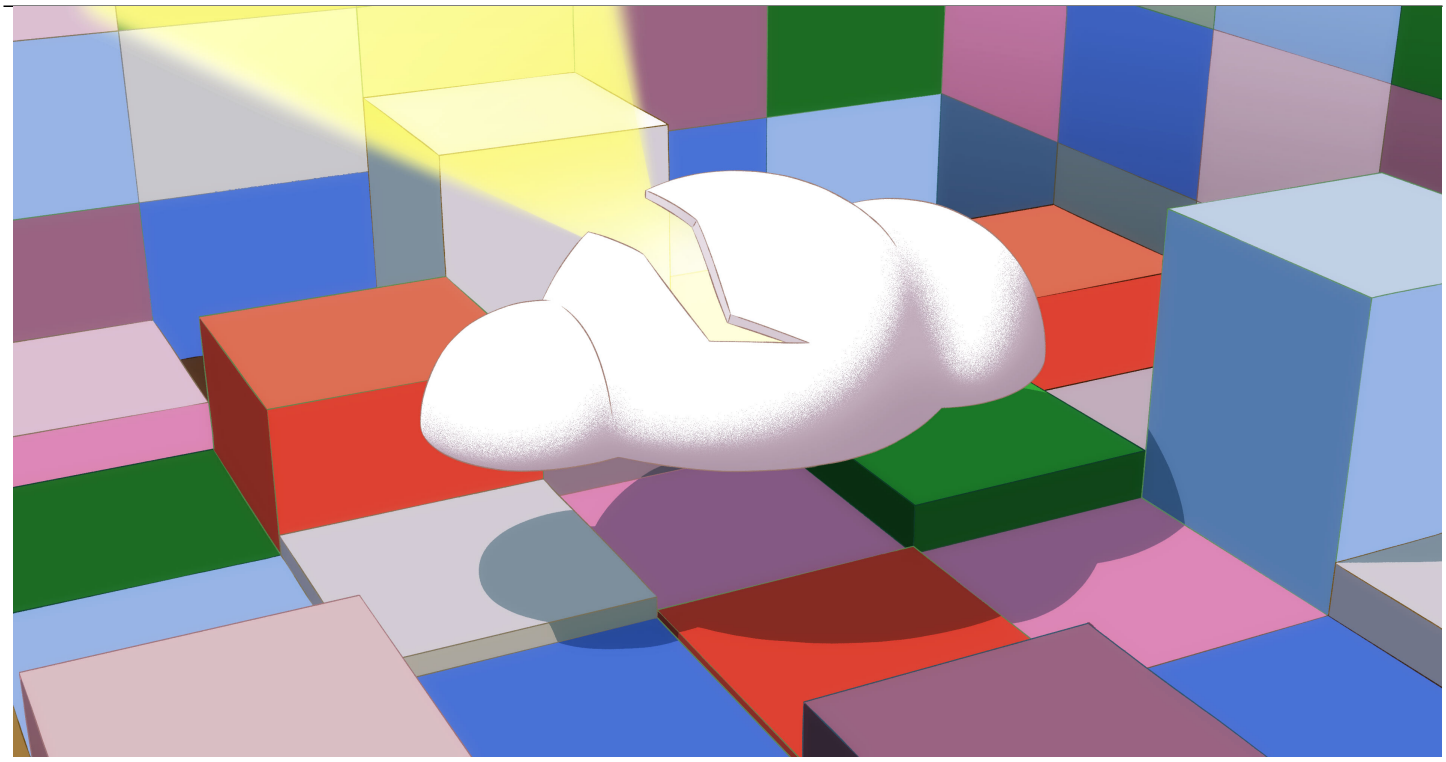
Investigations Across the Cloud

Litigation and Dispute Resolution

Technology, Privacy, and eCommerce



Original banner artwork by Lili des Bellons

Cheat Sheet

- **Challenges in data forensics**. Over 40 percent of general counsel face new challenges with emerging cloud data sources and messaging apps.
- **Data deluge**. Data volume and variety is growing exponentially.
- **Case study insights**. Examples involving the US Department of Justice highlight the complexities of collecting and reviewing data.
- **Legal and technical nuances**. Legal teams must evolve by first understanding the risks associated with the new technology and then utilizing innovative processes to address them.

Cloud data sources and the prevalence of messaging and collaboration applications used within the workplace have spurred a new cycle of disruption in digital forensics, investigations, and information governance. More than 40 percent of general counsel have reported that their legal departments are experiencing new challenges relating to emerging data sources and in a study of global CIOs, 77 percent reported that increased cloud and collaboration app use are causing their IT environments to change every minute or less.

Overall user numbers for cloud-based platforms, and the variety of applications tapped for business purposes, will only continue to grow, especially as tech giants innovate with artificial intelligence and release new collaborative interfaces. The upshot is that counsel, courts, regulators, and investigators

are grappling with how to uphold accepted principles for forensic defensibility, electronic evidence collection, data preservation, and information governance across rapidly changing populations of data created by and stored in new technology.

One DOJ investigation required a shipping provider to collect, review, and produce documents from cloud data sources. Is your organization prepared to do the same?

As cloud-based messaging, productivity, collaboration, and document management platforms become more prevalent within corporations, the process of responding to requests for company documents and data is becoming increasingly complex.

For example, in one investigation into alleged price fixing within an international shipping provider, the company was required by the US Department of Justice and numerous international competition bureaus to collect, review, and produce documents from sources including Google Workspace (its primary internal productivity suite) and AppSheet (a Google tool for workflow automation).

A large volume of documents in scope in the matter were dynamic, shared documents with multiple versions that were created, accessed, revised, and shared by numerous employees with varying levels of permissions, and were shared between employees as dynamic links rather than static attachments — all which had to be forensically collected and matched according to their “parent” message and creator.

Moreover, critical data and artifacts generated by AppSheet were not compatible with traditional forensic collection and eDiscovery processing and review platforms, further compounding the numerous technical challenges of defensibly collecting and reviewing the pertinent data. This example is not an outlier, but instead will be the norm as organizations move most of their business information into cloud-based enterprise software like Microsoft 365 or Google Workspace.

This article will cover the fundamentals of what emerging data sources are, the specific legal and technical challenges around them, solutions for handling them in a forensically defensible way, and recent court decisions that will help guide legal teams as they adjust to this new paradigm.

Emerging cloud-based technology and data sources

So, what are emerging cloud-based technology and data sources, specifically? And what are the primary challenges surfacing around them? An emerging data source is any cloud-based platform, collaboration tool, or messaging application used for business purposes and communications. The most known are Microsoft 365, Google Workspace, Box, Dropbox, Slack, Zoom, and WhatsApp. There are thousands of emerging data sources currently in use around the world, with more arriving all the time.

While most people are familiar with these tools and use them with ease in day-to-day work and personal communications, they are highly complex under the hood. These technical nuances have created persistent challenges in the context of disputes, investigations, regulatory compliance, and information governance.

Emerging data sources are popular among users but have underlying technical complications. David Urbinati / giphy

> An emerging data source is any cloud-based platform, collaboration tool, or messaging application used for business purposes and communications.

These include:

Volume

Data and document volumes are growing all the time, as is the volume of usage across cloud-based platforms. The global data universe is expected to reach 175 zettabytes of data by 2025 (IDC). Analysts predict the use of persistent messaging at work to reach 247.6 million daily active users 2025. And spending on social and collaboration software is estimated to grow by more than 70 percent between 2023 and 2027 (Gartner). All this points to ongoing exponential growth in data volumes.

Variety

The variety of emerging data sources is equally impressive — new sources are emerging all the time. Variety is growing via new platforms, and in the numerous formats that exist within each platform and the types of metadata they include.

The variety of chat tools and message formats is especially challenging. Many formats are not compatible out of the box with traditional forensic and eDiscovery tools. Additionally, short-form messages are difficult to organize in a way that eliminates arbitrary chatter while capturing relevant information in context. The use of shorthand, slang, images, and emojis further complicates the task of synthesizing chat conversations to infer meaning and relevance. And the coming incorporation of artificial intelligence into these enterprise cloud-based data sources will create massive amounts of a new type of data.

Velocity

Continuous improvement, evolution, and cycling of change across cloud platforms is happening at a

rapid and accelerating pace. New capabilities and tools are added, features are phased out, and systems are adjusted to optimize user experience. Yet for digital forensics purposes, as soon as a solution, process, or workaround is developed for one technical challenge, it changes, or another emerges. This is happening all the time, across every platform, and when compounded alongside the scale of variety in data sources, it makes it incredibly difficult for traditional digital forensic tools to maintain compatibility. Similarly, the underlying APIs that are available from cloud providers are constantly changing.

Constant changes and improvements to new cloud platforms create difficulty for traditional digital forensic tools to maintain compatibility. Abdul-Jalil / *Shutterstock.com*

Access and preservation

Emerging data sources are prevalent across company systems, employee computers, and personal devices being used for work. Gaining access to and preserving traditional company documents and email is a routine process, however doing so for encrypted WhatsApp data on an employee's personal phone is not. Aside from the many legal and practical challenges of accessing certain applications and devices, traditional forensic tools and workflows do not have plug-and-play capability to execute a complete, forensically-sound collection for many cloud sources.

Duplicates, near-duplicates, and changeable files

With dynamic documents (i.e., linked content, or sometimes referred to as modern attachments), there are a vast number of systems creating and saving numerous versions of a single document. As legal teams are tasked with collecting and reviewing documents linked in messages, they must determine and record which version of that document is relevant, whether there's a need to review and/or produce all versions, and if changes from version-to-version are relevant or need to be analyzed against a broader dataset. With the ability of multiple users to have access to the same document, often at the same time, and the creation of hundreds of "versions" of a single document with a few keystrokes, the very concept of versions may no longer be sustainable.

Authentication

With the ability of multiple users to have access to the same document, often at the same time, and the creation of hundreds of "versions" of a single document with a few keystrokes, the very concept of versions may no longer be sustainable.

How can teams defensibly explain and demonstrate the flow of data coming in and out of a record stored in a collaboration application? Justin Mezzel / giphy

Traditional processes for authenticating information come into question when cloud-based sources are involved. For example, when data is connected from numerous, separate repositories into a single chat message, how is the original source determined and verified? How can teams defensibly explain and demonstrate the flow of data coming in and out of a record stored in a collaboration application? Authenticating to cloud endpoints is another challenge. Each cloud endpoint implements a variety of authentication protocols, and while standards have emerged for authentication and authorization, cloud providers implement these options in varying degrees.

Custodian identification

The process of appropriately or efficiently targeting individuals/custodians and their activities must now also be gauged through the lens of their access, permissions, and participation in certain documents or chat channels. Can a single custodian be defined when a single document contains content from multiple contributors in a shared workspace? Does membership alone, without active participation in a communication channel constitute custodianship?

This was a central issue in *People v. N.T.B.*, 457 P.3d 126, 131 (2019), which notes, "[C]loud -based files lack many of the readily identifiable characteristics that often make authentication…possible.

Specifically, files uploaded to remote servers are not necessarily shared with other users, which forecloses the opportunity for a recipient to authenticate them. And cloud storage providers may not require detailed profiles of their users, which eliminates another avenue to corroborate ownership of the account's contents."

Generally, even the notion of a custodian is problematic when dealing with cloud sources. Usually, cloud sources have entities, like users or shared drives. Attempting to identify a custodian obscures the source and often doesn't work for modern cloud platforms, which runs contrary to traditional forensic collection philosophies. Indeed, for many applications, even a location designated and "owner" by an individual as a personal drive does not mean that others do not have access to edit and modify a document in that individual's personal drive.

Data retention and defensible disposal

For many of the same reasons that forensic investigation workflows are difficult across emerging data sources, information governance is also challenging. When data can't be easily collected and preserved, it's also difficult to enable automated, sustained and routine retention and disposal processes in a legally defensible way. Yet without those, volumes and data risk continue to grow, creating a reinforcing cycle of data bloat and complexity.

Achieving forensic defensibility

Fundamentally, traditional concepts and approaches are no longer applicable to the new digital landscape. Still, it is possible to remain faithful to digital forensic principles of admissibility, authenticity, repeatability, and chain of custody, while leaning into new approaches and tools that are better aligned to the technical nuances of emerging data. For example, with the right approach, legal teams can still authenticate a point in time based upon metadata within a linked document, even if a static, local attachment isn't available.

APIs (application programming interfaces) are an important element in the quest to modernize approaches to defensibility in forensic investigations and eDiscovery. Millions of developers and billions of users globally rely on APIs provided by credible technology companies to provide secure, consistent, and trustworthy methods for extracting and exchanging data and developing solutions. They are foundational infrastructure in the modern data ecosystem. Understanding how they function and documenting the protocols in place and the details about the API when using them, can provide a strong foundation for defensibility.

APIs play a critical role in advancing defensibility in forensic investigations and eDiscovery. JulsIst / *Shutterstock.com*

> Millions of developers and billions of users globally rely on APIs provided by credible technology companies to provide secure, consistent, and trustworthy methods for extracting and exchanging data and developing solutions.

Particularly as many emerging data sources do not have export features or have significant limitations in the format and throughput of data exports, APIs are integral to collecting data in an investigation and converting it into a format that can be ingested and reviewed. In a forensic collection, APIs may be used, for example, to retrieve a native file and other content, capture specific metadata and/or retrieve versioning information.

Note this issue per *Hoyt v. Career Sys. Dev. Corp.*, 2010 U.S. Dist. LEXIS 43584, at *2–3 (S.D. Cal. May 3, 2010), "If authentication of electronically stored information becomes an issue at trial, metadata may be essential in ensuring the admissibility of documents." Also, in *Summit Fire & Sec. LLC v. Kolias*, 2022 Del. Ch. LEXIS 197, at*6, the court advised, "Metadata can be useful, and in many circumstances I would require it be produced…"

Many platforms that offer APIs also offer documentation and test environments, which enable quality control and validation of outputs. Moreover, APIs can provide a foundation for further development and versioning, providing teams a way to establish repeatability for workflows and thus reinforce defensibility. Test environments are an effective way to validate an approach for collection, but actually using modern applications to perform actual production collections using the provider's APIs is often the most effective approach. This means that the applications developed to perform a collection must be built for rapid updates and include robust instrumentation to rapidly track down issues that might only appear during a production collection effort.

Our teams have addressed countless matters wherein APIs provided the only viable and proportionate option for obtaining access to the data needed and collecting it in a sufficient format.

For example, in an ongoing, high-profile criminal investigation, data of interest included five years' worth of historic calendar items from a cloud-based calendaring application, which was typically installed and used on personal devices. The application was not compatible with digital forensics and investigations tooling, and the information could not be accessed from custodian devices, leaving APIs as the only defensible way to extract the required information. In a highly technical exercise, the team used the API to obtain access to the data and operations of the calendar items. Through this approach and additional bespoke work, the team collected and authenticated roughly 6,000 calendar items. This work was supported by a detailed error log of items that could not be recovered, with documentation of the reasoning, to further illustrate the defensibility and completeness of the collection.

Legal teams must be involved in these processes, understanding the legal implications made both when designing information governance for a data source (retention and preservation practices, etc.), and when conducting forensic collections for a legal matter. This includes in-house and outside counsel in a matter, as the risk of misrepresentations to courts and regulators grow as legal teams are called upon to explain the technical nuances of these new data sources.

Maintaining data and artifact integrity in investigations

The heart of forensic defensibility in investigations lies in maintaining data or artifact integrity. During any investigation or legal matter, the original data collected must remain unaltered to maintain its potential evidentiary value. This integrity is particularly important when dealing with cloud-based platforms, as they often contain unique or supplemental artifacts that might not be present in traditional storage media.

Artifacts unique to cloud platforms can range from user access logs, network traffic data, application metadata to remnants of shared access environments, versions of documents, and chat messaging data. These artifacts are crucial components of investigations, as they can provide insights into user behavior, data modification history, and other key elements. In many cloud platforms, it is common for activity logs or version history to exist for a trailing time period. Without swift action to preserve these logs in their entirety, data critical to an investigation may be destroyed permanently.

Ensure original data is not permanently destroyed by maintaining data or artifact integrity. Visumrun / giphy

The preservation of data integrity calls for robust protocols and mechanisms designed to protect the original state of data. Unauthorized modifications, editing, or data corruptions can potentially alter or even destroy the valuable information that could be obtained from these artifacts.

> By adopting forensically sound data collection practices and focusing on data integrity, legal teams and digital forensic specialists can ensure that the collected information retains its full evidentiary value, making it reliable and admissible in legal proceedings.

By adopting forensically sound data collection practices and focusing on data integrity, legal teams and digital forensic specialists can ensure that the collected information retains its full evidentiary value, making it reliable and admissible in legal proceedings. This level of diligence also offers an

added layer of confidence across the broad scope of digital forensics assignments, as it reassures stakeholders and legal teams that the evidence collected is both authentic and untampered.

Shortfalls in traditional methods

Traditional methods of data collection, such as physical hard drive imaging are falling short in the context of cloud-based forensics due to a combination of technological, legal, and practical challenges. Imaging of physical hard drives has for years been considered the “standard” and typically involves creating an exact bit-for-bit copy of a target hard drive. This image is then analyzed for evidence, allowing investigators to conduct their work without altering the original data.

Utilizing standard traditional methods such as physical hard drives to capture data has become outdated and presents many challenges for forensic investigations. Jovanovic Dejan /

Shutterstock.com

However, this method is becoming increasingly inadequate when dealing with emerging data sources. Some of the key limitations relating to this form of data capture include verifying authenticity, accessibility to the data, limitations in the legality of capturing certain types of data and sources intermingled with personal information. Therefore, new methods and tools designed to interface directly with cloud services and account for the unique characteristics of cloud data are increasingly needed to conduct effective, defensible investigations.

Upholding the integrity and efficacy of investigations

The ever-changing nature of today's digital landscape, marked by the exponential growth of cloud data sources, collaboration tools and messaging applications, has fundamentally reshaped the field of digital forensics and other adjacent workstreams.

The proliferation of cloud-based platforms and the sheer volume, variety, and velocity of data being created have created new technical, legal, and practical challenges. These challenges range from the difficulties in accessing and preserving data across diverse platforms, handling duplicates and near-duplicates, authenticating information, and ensuring appropriate data retention and disposal. The traditional methods like imaging of physical hard drives are falling short in the context of cloud-based forensics due to these ever changing variables.

However, these challenges are not insurmountable. With an in-depth understanding of emerging data sources and a willingness to adapt and innovate, legal teams, investigators, regulators, and courts can uphold the principles of forensic defensibility regardless of the source of data. Leveraging technologies such as APIs and adopting forensically sound data collection practices can provide a strong foundation for defensibility, ensuring that the evidence collected retains its full evidentiary value.

One of the critical takeaways that should be gleaned from this paper is the need for a continuous evolution of methods, tools, and frameworks that align with the technical nuances of emerging data. There should be an emphasis on the need to maintain data and artifact integrity, which reflects the heart of forensic defensibility, especially in the face of cloud-based platforms that present unique or supplemental artifacts. It's also critical to apply modern software engineering practices to adapt quickly to a constantly changing technical landscape.

Enhancing digital forensics and across cloud data sources is both an imperative and a complex task. Meeting the challenges posed by emerging data requires a symbiotic relationship between understanding the nature of the data and the integration of new technologies and approaches that respect the principles of forensic defensibility. It is imperative that we pave the way for future innovations and adjustments to continue to uphold the integrity and efficacy of investigations in an ever-changing digital world.

Zander Brandt

Discovery Litigation Manager

Netflix

Zander Brandt is the eDiscovery Litigation Manager at Netflix, where he oversees and runs Netflix's eDiscovery practice. He has more than 15 years of experience in the eDiscovery space, ranging from handling forensic collections at a large vendor to managing projects as a Litigation PM at an AM 100 law firm. Prior to joining Netflix, Brandt ran and built the Legal Technology team at Lyft, and also managed and built the eDiscovery team at LendingClub. A frequent speaker, Brandt has led discussions at events such as CLOC and ABA conferences.

Anthony Diana

Partner

Reed Smith

Anthony Diana is a partner in the Emerging Technologies Group and focuses his practice on commercial litigation, internal and regulatory investigations, electronic discovery and information governance, and data privacy and security.

Therese Craparo

Partner

Reed Smith

Therese Craparo is a partner in Reed Smith's Emerging Technologies and Records & E-Discovery Groups. She focuses her practice on enterprise data risk management and information governance, data privacy and security, cyber security, records and eDiscovery.

Michael Khoury

Senior Managing Director

FTI Technology

Michael Khoury is a Senior Managing Director within FTI Technology, based in Australia. He specializes in forensic technology, cybersecurity, electronic discovery and dispute advisory, with more than 19 years of industry experience. He is frequently retained as an expert witness in matters requiring the application of computer forensic skills to provide answers to legal and investigative cases.

Tim Anderson

Senior Managing Director

FTI Technology

Tim Anderson is a Senior Managing Director within FTI Technology, based in San Francisco. He has more than twenty years' experience in legal technology as an e-discovery consultant, paralegal, programmer, and application development manager. Anderson currently leads the Emerging Data Sources Team and is responsible for directing data identification, integration, preservation, collection, and analysis strategy and workflows to support clients in an ever-changing digital and regulatory landscape.