




Eric Dull

Data Lakes: An Intelligent Solution to Information Storage

October 31, 2023



Challenges to Cyber Analysis

Nearly 30 years ago, I went online for the first time and used a 2400 baud modem to dial into a Bulletin Board System to set up an email account. Since then, there has thankfully been exponential growth in bandwidth. For instance, last weekend, I made a large download on my home broadband, seeing sustained speeds of 192 MB/s, which is 160,000 times faster than my first venture online. This is just one example of the significant growth in Internet access and usage.

This increased digital activity creates “digital footprints” reflecting that activity in the form of structured metadata and semi-structured data. Information stored in its native format such as images, audio, signals, or other myriad formats allows for rich analysis. Organizations wishing to take advantage of this new abundance of information and turn it into intelligence may find themselves having to address challenges of scalable and secure storage, consumption, and usage.

Traditional data storage options typically use preset structures with consistent formats. These traditional databases are highly searchable, but their implemented schemas may preclude the rapid and efficient import of semi-structured data. Data fields may be dropped during extraction, transformation, and load, which can reduce the data’s analytic value and result in undiscovered insights. An improved system capable of handling the volume and variety of semi-structured data will greatly benefit organizations.

How Are Data Lakes Used for Cybersecurity?

Adding to the technical challenges are policy directives. In recognition of the changing cyber landscape, the Biden Administration released the *Executive Order on Improving the Nation’s Cybersecurity* ([EO 14028](#)) in May 2021. It calls on federal agencies to adopt practices to improve and accelerate cybersecurity modernization, including a requirement ([M-21-31](#)) to centralize and streamline data access for cyber risk analysis. As discussed in [Clearing Roadblocks to AI-Enabled Cybersecurity](#), agencies must contend with the changing data landscape and evolution of digital threats. To comply with EO 14028 and M-21-31, organizations will have to reconsider how they store and access their information.

Gathering cybersecurity insights requires the collection, observation, and analysis of broad amounts of information gathered from diverse sources such as firewalls, routers, endpoint security software, server logs, dedicated network sensors, and enrichment sources. Cyber threat analysts use these details to detect, analyze, and respond to potential security risks. The steep increase in structured and unstructured data volumes has created an overwhelming workload for cyber professionals who must sift through it all to identify anomalies and detect patterns. This information burden can lead to false positive threat alerts and other errors. Sustainable threat detection and alert evaluation requires overcoming these data challenges and harnessing the large amounts of available information to separate false alerts from threats posing true security risks.



Enter: [data lakes](#)

Data lakes are large, centralized data repositories that can store both structured and unstructured data in their native formats regardless of origin or schema. Scalable processing and storage available with data lakes enables retrospective and real-time analytics regardless of data size or type. Organizations can store, manage, and access information from different sources and layer on artificial intelligence (AI) and machine learning to analyze the data points in their original forms and correlate multiple insights. This enables more [dynamic and in-depth intelligence](#) to be extracted from stored data than previously limited methods.

Organizations with deployed data lakes can label information and apply relevant security controls such as encryption, tokenization, and identity management practices in-flight or at-rest. These capabilities can reduce the workload and help to more efficiently comply with government requirements and authorizations. They can also capture data access and movement to identify consumption patterns and—more importantly, anomalies—within the data lake.

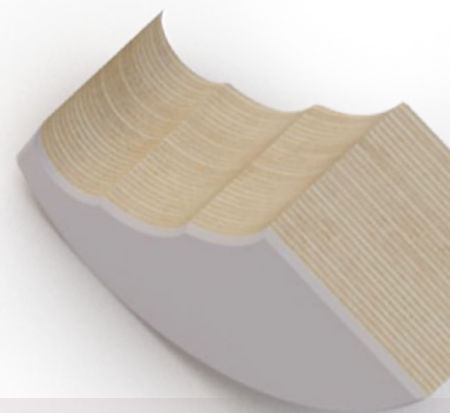
Cyber analysts working in data lake-enabled platforms can incorporate analytics and [AI](#) to create analytic automations to improve both efficiency and efficacy. Data lakes do not come with pre-existing schemas and are designed for both unsupervised or supervised algorithm development and execution. Teams often use data lakes to identify behaviors that should be investigated, otherwise known as “bumps in the night”. Teams also can use data lakes to automate routine analytic tasks and workflows using supervised machine learning, driving higher analytic quality and reducing the time required to analyze security alerts as well as behaviors identified by other workflows running on the data lake.

By using data lakes, agencies can not only better handle the challenges of increased information volume, but also utilize the rich digital footprints about online activities occurring on their networks to gain deeper insights into network activities, proactively detect more sophisticated cyber threats, automate cyber analytic workflows, and enable their cyber analysts to get time back and ask higher order, more sophisticated questions. They get to become analysts again.

Where can I learn more?

Learn more about data lake solutions by visiting our GPS Google Cloud Alliance site [here](#) or by reviewing our new cyber data lakes placemat [here](#).

This posting contains general information only, does not constitute professional advice or services, and should not be used as a basis for any decision or action that may affect your business. Deloitte shall not be responsible for any loss sustained by any person who relies on this posting.



Eric Dull

Managing Director

Deloitte & Touche LLP

edull@deloitte.com

This communication contains general information only, and none of Deloitte Touche Tohmatsu Limited (“DTTL”), its global network of member firms or their related entities (collectively, the “Deloitte organization”) is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser. No representations, warranties or undertakings (express or implied) are given as to the accuracy or completeness of the information in this communication, and none of DTTL, its member firms, related entities, employees or agents shall be liable or responsible for any loss or damage whatsoever arising directly or indirectly in connection with any person relying on this communication. DTTL and each of its member firms, and their related entities, are legally separate and independent entities. Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited (“DTTL”), its global network of member firms, and their related entities (collectively, the “Deloitte organization”). DTTL (also referred to as “Deloitte Global”) and each of its member firms and related entities are legally separate and independent entities, which cannot obligate or bind each other in respect of third parties. DTTL and each DTTL member firm and related entity is liable only for its own acts and omissions, and not those of each other. DTTL does not provide services to clients. Please see www.deloitte.com/about learn more.