Cribl

>WHITE PAPER_

Navigating the future of IT and Security Data management:

embracing schema-on-need for flexibility and control.

>WHITE PAPER_

Navigating the future of IT and Security Data management:

embracing schema-on-need for flexibility and control.

As the volume and complexity of IT and security data continue to grow exponentially, traditional data management approaches are struggling to keep pace. The one-size-fits-all mentality of legacy databases and tools has led to fragmentation, high costs, and vendor lock-in. Addressing these challenges and unlocking the full potential of their data means organizations must adopt a more nuanced and adaptable approach that recognizes the unique characteristics of IT and security data. Schema-on-need is an emerging concept that combines specialized, intelligent technology for data in motion with automated, scalable architectures for data at rest to address the full lifecycle of IT and security data.

Key findings.

- IT and security data is fundamentally different from traditional business data, requiring a tailored approach to how it is managed and analyzed throughout its lifecycle.
- Legacy databases and tools, with their rigid schemas and high costs, are ill-suited to handle the scale and complexity of modern IT and security data.
- Tiered data management strategies, while effective in containing costs, can lead to fragmentation and skill set challenges.

Recommendations.

- Embrace schema-on-need: Adopt a flexible data management approach that allows for selective acceleration, indexing, and storage based on the value and usage of the data.
- Invest in your data lake: Centralize your IT and security data in a cost-effective, open data lake that facilitates large-scale queries, fast search, and smart data acceleration.
- Prioritize integration and unification: Ensure seamless integration between your data lake, existing tools, and edge devices, while providing a unified data experience that empowers users to explore and analyze data across formats and locations.
- Avoid vendor lock-in: Choose solutions that store data in open formats and allow for flexibility in data management, preventing lock-in to proprietary tools and enabling adaptability to future needs.

> As the volume and complexity of IT and security data continue to grow exponentially, traditional data management approaches are struggling to keep pace.

The evolution of data management.

Schema-on-write: the limitations of traditional databases.

Traditional databases, with their schema-on-write approach, have struggled to keep pace with the scale and complexity of IT and security data. The need to define data structures in advance hinders the ability to quickly adapt to changing data formats and limits the incentive for developers to capture all the necessary information.

With the structures defined, the work of transforming multiple, high volume data sources into the overall format begins. This is a process known as extract, transform, and load (ETL). ETL is cumbersome for any data, but this pain is especially acute for IT and security data. ETL development involves working with large volumes of complex data from different sources, data integration, transformation, and validation. It also demands performance optimizations, scalability considerations, and ongoing maintenance to handle data source changes and evolving business requirements. Moreover, ETL development often uses specific tools and frameworks for data integration and workflow orchestration, with the corresponding specialized skills.

Schema-on-read: indexing and time series databases.

The advent of purpose-built indexing and time series databases has addressed many of the ergonomics and scale issues associated with traditional databases. These solutions offer improved query performance, easier data ingestion, and better handling of complex schemas. However, they come with their own challenges, including high costs, vendor lock-in, and the need for specialized skill sets.

While schema-on-read approaches don't require the ETL overhead of schema-on-write, you often find manual processes in place to make sense of less structured data, or to integrate two sources of data into some unified result.

As organizations grapple with the challenges of managing IT and security data, many have turned to tiered data management strategies to balance cost and complexity.

Tiered data management: balancing cost and complexity.

As organizations grapple with the challenges of managing IT and security data, many have turned to tiered data management strategies to balance cost and complexity. This approach involves using a combination of legacy tools, cloud data warehouses, lakehouses, and data lakes to store and analyze data based on its value, usage patterns, and retention requirements.

Legacy tools, such as purpose-built indexing and time series databases, are often maintained to leverage existing investments and user training. However, to contain the growth and cost of these tools, organizations are increasingly exploring alternatives for managing their ever-expanding data volumes.

Cloud data warehouses and lakehouses have emerged as attractive options due to their separation of storage and compute, which allows for cost savings through the use of cheap object storage and on-demand compute resources. These solutions also offer the potential for more flexible data management and analysis. However, adopting these technologies often requires significant investments in data engineering resources and skill sets, which can be a barrier for many organizations.

Data lakes, largely built on object storage, have gained popularity as a cost-effective way to store vast amounts of raw data in open formats. While data lakes offer the potential for cheaper storage and greater flexibility, organizations often struggle to derive value from the data once it lands in the lake. Most analytics solutions today require data to be moved into their own systems for querying, limiting the ability to analyze data directly in the lake.

Despite the challenges, tiered data management strategies have proven effective in helping organizations contain costs and begin exploring more flexible and scalable approaches to managing their IT and security data. However, these strategies also lead to fragmented management across the data lifecycle, with data spread across multiple systems and formats, each requiring different skill sets and tools for analysis. This fragmentation can hinder an organization's ability to gain a comprehensive view of their data and quickly derive insights.

The challenges of increasing costs and decreasing flexibility.

Purpose-built IT and security data management solutions, while addressing the ergonomics and scale issues of traditional databases, come with their own set of challenges. These solutions can be expensive, particularly when lifted and shifted from the datacenter to the cloud. The tight coupling of storage and compute, along with application-level replication and full-text indexing, can result in costs that are nearly 100 times higher than storing data in open formats on object storage.

Moreover, these solutions often store data in proprietary formats, leading to vendor lock-in. Once data is ingested into a particular tool, it can only be read back using that same tool, effectively locking organizations into that solution forever. This lack of flexibility and control over data can hinder an organization's ability to adapt to changing needs and take advantage of new technologies.

The future of IT and Security Data management.

Schema-on-need: flexible data management for the modern era.

In the future of IT and security data management, the concept of schema-on-need will play a central role in enabling organizations to achieve the right balance between structure and flexibility. Schema-on-need is a data management approach that allows for the selective acceleration, indexing, and storage of data based on its value and usage patterns. Ultimately, it combines the performance attributes of schema-on-write with the flexibility of schema-on-read.

Schema-on-need is a data management approach that allows for the selective acceleration, indexing, and storage of data based on its value and usage patterns.

Under this approach, data is initially stored in its raw, unstructured form in a centralized data lake. As the data is accessed and analyzed, the system dynamically applies the necessary structure and optimizations to facilitate fast querying and analysis. This selective application of schema ensures that resources are only consumed for data that is actively being used, reducing overall costs and complexity.

Schema-on-need also enables organizations to adapt quickly to changing data formats and requirements. As new data sources and use cases emerge, the system can easily incorporate them without the need for extensive upfront modeling or schema design. This flexibility is particularly crucial in the rapidly evolving world of IT and security, where new threats, technologies, and data sources are constantly emerging.

Moreover, schema-on-need allows for the coexistence of structured and unstructured data within the same system. This means that organizations can leverage the benefits of structured data for fast, targeted queries and analytics, while still retaining the ability to explore and analyze unstructured data for new insights and patterns.

By embracing schema-on-need, organizations can break free from the rigidity and limitations of traditional schema-on-write and schema-on-read approaches. They can achieve a more agile, cost-effective, and future-proof data management strategy that empowers them to extract maximum value from their IT and security data.

The Data Lake: centralized, open, and cost-effective.

The data lake is the cornerstone of the future IT and security data management strategy. By centralizing all relevant data in a single, open repository, organizations can break down silos, reduce complexity, and enable a more holistic approach to data analysis and insight generation.

The data lake leverages cheap object storage to provide a cost-effective foundation for storing vast amounts of raw, unstructured data. This approach eliminates the need for expensive, proprietary storage solutions and enables organizations to scale their data storage seamlessly as their needs grow.

To maximize the value of the data lake, it is crucial to store data in open, standards-based formats such as Apache Parquet, Apache Avro, or JSON. These formats ensure that data can be easily accessed and analyzed by a wide range of tools and platforms, avoiding vendor lock-in and promoting interoperability.

The data lake also serves as the foundation for smart data acceleration and fast search capabilities. By selectively accelerating and indexing data based on usage patterns and value, organizations can achieve the performance and responsiveness needed for real-time analysis and decision-making, without incurring the high costs associated with traditional indexing and acceleration approaches.

Moreover, the data lake enables organizations to retain data for longer periods, allowing for historical analysis, trend identification, and machine learning applications. By keeping data in its raw form, organizations can always go back to the original source data, ensuring data lineage and enabling them to ask new questions and apply new analytical techniques as their needs evolve.

The data lake leverages cheap object storage to provide a cost-effective foundation for storing vast amounts of raw, unstructured data.

The intelligent Edge: efficient and centrally managed.

In addition to the centralized data lake, the future of IT and security data management also encompasses an intelligent edge. The intelligent edge leverages the growing computing power and storage capacity of edge devices, such as servers, laptops, and IoT devices, to enable efficient data processing and analysis at the source.

By processing data at the edge, organizations can reduce the amount of data that needs to be transferred to the central data lake, minimizing network bandwidth requirements and improving overall system performance. This is particularly important for use cases such as real-time monitoring, anomaly detection, and security incident response, where fast, local decision-making is critical.

The intelligent edge is centrally managed, ensuring that edge devices are properly configured, monitored, and secured. This central management allows IT and security teams to maintain control over the edge infrastructure, apply consistent policies and updates, and ensure compliance with organizational standards and regulations.

To enable efficient data processing at the edge, lightweight data processing engines and analytics tools are deployed on edge devices. These tools allow for local data transformation, aggregation, and analysis, while still being able to seamlessly integrate with the central data lake for more advanced analytics and long-term storage.
The intelligent edge also plays a key role in enabling centralized querying and analysis across the entire IT and security data landscape. By exposing a unified query interface, the intelligent edge allows users to query data across edge devices and the central data lake, without needing to worry about the underlying data location or format.

This unified querying capability empowers organizations to gain a comprehensive, real-time view of their IT and security posture, enabling faster, more informed decision-making.

Seamless integration with existing tools.

In the future of IT and security data management, seamless integration with existing tools will be a critical success factor. Organizations have invested heavily in a wide range of tools for monitoring, analytics, security, and more, and these tools will continue to play a vital role in their data management strategies for years to come.

To maximize the value of existing investments and minimize disruption, the future data management platform must prioritize seamless integration with these tools. This involves developing robust APIs, connectors, and adapters that enable bi-directional data flow between the centralized data lake, intelligent edge, and existing tools.

Integration should cover both data ingestion and data consumption. On the ingestion side, the data management platform should be able to efficiently collect data from a wide range of sources, including legacy tools, and transform it into a common format for storage in the data lake. This process should be automated and scalable, allowing organizations to easily onboard new data sources as their needs evolve.

By enabling seamless integration with existing tools, the future data management platform allows organizations to continue leveraging their existing investments while gradually modernizing the data infrastructure.

On the consumption side, the data management platform should provide flexible interfaces that allow existing tools to query and analyze data stored in the data lake and at the intelligent edge. This can be achieved through standard APIs, such as SQL and REST, as well as through custom connectors that leverage the native query languages and protocols of specific tools.

Seamless integration also involves addressing the challenge of data format and schema translation. With data coming from a wide range of sources and tools, it is essential to be able to map and translate between different formats and schemas in real-time. The data management platform should include powerful data transformation capabilities that can handle this complexity automatically, ensuring that data is always available in the format required by each consuming tool.

By enabling seamless integration with existing tools, the future data management platform empowers organizations to continue leveraging their existing investments while gradually modernizing their data infrastructure. This approach reduces risk, minimizes disruption, and allows organizations to realize the benefits of a more flexible and scalable data management strategy incrementally over time.

A unified data experience.

In addition to seamless integration with existing tools, the future of IT and security data management also demands a unified data experience. With data spread across multiple systems, formats, and locations, it is essential to provide users with a single, consistent interface for accessing, querying, and analyzing data.

A unified data experience starts with a common data model and schema that can accommodate the diverse types of data generated by IT and security systems. This data model should be flexible enough to handle structured, semi-structured, and unstructured data, and should be able to evolve over time as new data sources and requirements emerge.

On top of this common data model, the data management platform should provide a unified query language that allows users to interact with data consistently across the entire data landscape.

This query language should be intuitive and easy to learn, enabling both technical and non-technical users to access and analyze data without needing to master multiple tools and syntaxes.

The unified data experience should also include a rich set of tools and interfaces for data visualization, exploration, and collaboration. These tools should be tightly integrated with the query language and data model, allowing users to seamlessly move between querying, visualizing, and sharing data insights.

To support a wide range of user personas and skill sets, the unified data experience should offer both code-based and no-code interfaces. Code-based interfaces, such as notebooks and IDEs, enable power users and data scientists to leverage the full flexibility and power of the data management platform. No-code interfaces, such as drag-and-drop dashboards and visual query builders, empower business users and analysts to access and analyze data without needing to write complex queries or code.

By embracing a data-centric, open, and unified approach to data management, organizations can achieve the agility, efficiency, and insights they need to thrive in an increasingly complex and dynamic IT and security landscape.

Finally, the unified data experience should be highly performant and scalable, enabling users to interact with large volumes of data in real-time. This requires a highly optimized data processing engine that can efficiently query and analyze data across the centralized data lake and intelligent edge, as well as smart caching and acceleration techniques that can deliver sub-second response times even for complex queries.

By providing a unified data experience, the future data management platform democratizes access to IT and security data, enabling organizations to harness the collective intelligence of their entire workforce. This empowers teams to make faster, more informed decisions, identify emerging threats and opportunities, and continuously optimize their operations based on real-time data insights.

Conclusion.

As IT and security data continues to grow in volume and complexity, organizations must embrace a more flexible and adaptable approach to data management. By adopting a schema-on-need strategy, investing in a centralized data lake, leveraging the intelligent edge, prioritizing seamless integration with existing tools, and providing a unified data experience, organizations can unlock the full potential of their data while maintaining control and cost-effectiveness.

The future of IT and security data management lies in empowering organizations to make informed decisions based on their unique needs and resources, rather than being constrained by rigid, one-size-fits-all solutions or proprietary tools that limit flexibility and control over data.

ABOUT CRIBL

Cribl, the Data Engine for IT and Security, empowers organizations to transform their data strategy. Customers use Cribl's vendor-agnostic solutions to analyze, collect, process, and route all IT and security data from any source or in any destination, delivering the choice, control, and flexibility required to adapt to their ever-changing needs. Cribl's product suite, which is used by Fortune 1000 companies globally, is purpose-built for IT and Security, including Cribl Stream, the industry's leading observability pipeline, Cribl Edge, an intelligent vendor-neutral agent, Cribl Search, the industry's first search-in-place solution, and Cribl Lake, a turnkey data lake. Founded in 2018, Cribl is a remote-first workforce with an office in San Francisco, CA.

Learn more: www.cribl.io | Try now: Cribl sandboxes | Join us: Slack community | Follow us: LinkedIn and Twitter

WP-0018-EN-2-0624