

## White Paper

# VAST Cyber Lakehouse

This document explores the key features, benefits, scalability, and data storage capabilities of the VAST Cyber Lakehouse, emphasizing its role in achieving real-time continuous vigilance in the realm of cybersecurity.

# Enabling Real-Time Continuous Vigilance

In today's rapidly evolving cybersecurity landscape, the need for real-time continuous vigilance is paramount to ensure the proactive mitigation of threats. The VAST Cyber Lakehouse seamlessly integrates with Apache Spark, a powerful data processing engine, to provide an all-encompassing solution. This comprehensive solution combines Apache Spark's data processing prowess with VAST's capabilities for real-time data ingestion, curation, and normalization. Furthermore, the integration extends to include real-time querying and visualization from Security Information and Event Management (SIEM) systems. This document explores the key features, benefits, scalability, and data storage capabilities of the VAST Cyber Lakehouse, emphasizing its role in achieving real-time continuous vigilance in the realm of cybersecurity.

Memorandum 21-31, issued by the Office of Management and Budget (OMB) on August 27, 2021, in accordance with Executive Order 14028, Improving the Nation's Cybersecurity, established federal agency requirements to increase the government's visibility "before, during, and after a cybersecurity incident." M-21-31 outlines the necessity for agencies to capture specific logs and maintain them for a mandatory 18-month retention period. The VAST Cyber Lakehouse's capability to efficiently store data for 18 months perfectly aligns with Memorandum 21-31's requirements and reinforces its position as the ultimate choice for comprehensive cybersecurity data management.

## Capabilities and Benefits

**Data Collection:** The VAST Data Lakehouse employs Logstash to gather data from various sources. Logstash is responsible for orchestrating data collection within the enterprise. It collects a wide range of data types and efficiently routes them to Apache Kafka. Kafka acts as a message broker and data streaming platform, enabling real-time data ingestion. This streamlined approach allows organizations to collect diverse datasets from different sources in real-time, which is crucial for comprehensive cybersecurity data ingestion.

**Data Processing with Apache Spark:** Once data is collected by Logstash and published to Kafka, Apache Spark comes into play. Apache Spark is used for data transformation and enrichment. It consumes the data from Kafka and enhances it with contextual information. This enrichment process significantly increases the value of the data for subsequent analysis and threat detection. Apache Spark's capabilities in data processing are crucial for making the data ready for further analysis.

Implementing the VAST Cyber Lakehouse with Apache Spark integration enables organizations to meet stringent cybersecurity mandates while significantly reducing the total cost of ownership for data management. This approach offers scalable, high-performance analytics capabilities, ensuring both compliance and cost-effectiveness in a single, unified platform.



*Data processing pipeline*

**Multi-Source Ingestion and Scalability:** Apache Spark's capabilities go beyond data transformation and enrichment. It also offers multi-source ingestion and scalability features. This means that the system can handle data from multiple sources simultaneously, making it suitable for large-scale cybersecurity data analysis. Scalability ensures that the system can handle increasing volumes of data as needed.

**Data Normalization:** Spark's data transformation capabilities play a pivotal role in normalizing data into a data frame within the Vast Cyber Lakehouse. This normalization process ensures that data is prepared effectively for subsequent analysis and threat detection.

## Scalability

**Unparalleled Scalability:** VAST Cyber Lakehouse is designed to offer exceptional scalability, eliminating the need for data partitioning. This means it can handle growing data volumes without requiring complex data splitting or sharding strategies.

**Handling Trillions of Rows:** VAST Cyber Lakehouse can effortlessly store trillions of rows per table. This capability is crucial for organizations dealing with large-scale data, especially in the context of historical threat data, enabling easy access for deep forensic investigations and addressing advanced threats.

**Efficient Data Management:** The VAST Database efficiently manages massive datasets, allowing organizations to confidently respond to evolving cyber threats. This efficiency ensures that data can be stored, retrieved, and analyzed without compromising performance, even as the volume of data increases.

**Native C++ Plugin for Apache Spark:** VAST Cyber Lakehouse offers a native C++ plugin for Apache Spark's distributed architecture. This integration ensures seamless scaling as data volumes grow. Apache Spark is a popular distributed computing framework, and the plugin's support means that organizations can harness Spark's capabilities without worrying about performance bottlenecks.

**Deep Write Buffers with Persistent Memory:** VAST utilizes deep write buffers and low-cost persistent memory to instantly store every ACID transaction. This ensures that data is stored immediately, guaranteeing data integrity and consistency.

**Migrating to Hyperscale Flash Storage:** As data tables become populated, VAST seamlessly migrates them to low-cost hyperscale flash storage while adopting an efficient columnar format. This transition optimizes data distribution and organization, leading to faster query execution and improved overall query performance.

**Query Execution:** VAST allows queries to cut across both the long-term data store and the write buffer. Even though data accessed in the buffer is row-based, the underlying persistent memory structures enable rapid row reads, ensuring swift data retrieval.

**Columnar Data Format:** VAST uses a columnar data format optimized for flash storage, distinguishing it from systems that use Parquet. The chunk size in VAST’s Database is significantly smaller (32KB) compared to a typical Parquet row group, resulting in efficient query filtration and reduced query engine processing.

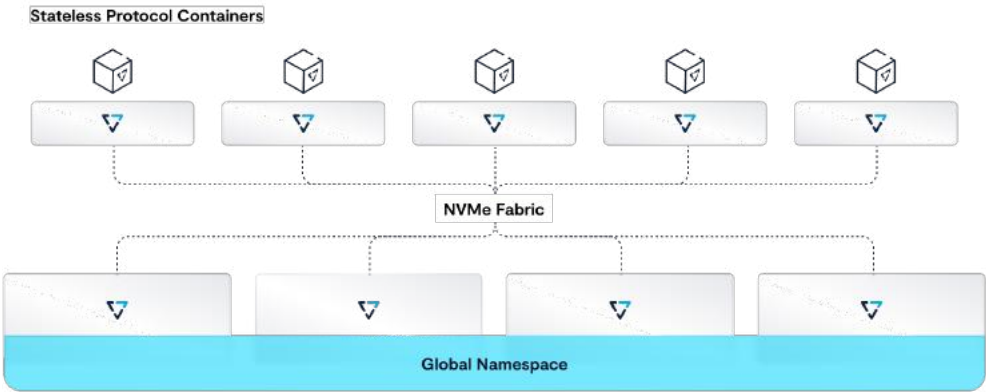


VAST Table Element

**Easy Updates:** VAST’s 32KB Database columnar payload is straightforward to update. This feature allows customers to instantly update tables for various purposes, such as GDPR requests or retention policy enforcement, without the need for complex database vacuuming processes.

**Prioritizing Speed and Efficiency:** VAST prioritizes speed and efficiency in data management and analysis, aiming to provide a seamless experience for users.

**Dynamic Resource Allocation:** Spark’s dynamic resource allocation capabilities intelligently distribute computing resources in response to workload demands. This dynamic allocation ensures that the Lakehouse efficiently assigns resources for real-time analysis and anomaly detection, adapting to varying data processing requirements.



VAST Disaggregated Shared-Everything Architecture

# Achieving Real-Time Proactive Threat Mitigation

The combination of Apache Spark and the VAST Cyber Lakehouse results in a powerful arsenal for achieving real-time proactive threat mitigation:

**Rapid Anomaly Detection:** By leveraging Spark's real-time processing capabilities, the VAST Cyber Lakehouse excels in swiftly identifying anomalies and deviations from normal network behavior. Spark's ability to process data in real-time enables the Lakehouse to trigger alerts instantaneously, providing security teams with the agility to respond promptly to potential threats as they arise.

**Threat Hunting:** Within the Lakehouse, Spark's robust machine learning algorithms and advanced data querying capabilities empower security analysts to engage in effective threat hunting activities. This includes the ability to conduct complex queries and investigations in real-time, enabling the discovery of concealed threats and vulnerabilities with remarkable efficiency. Spark's support for various machine learning models enhances the precision of threat detection.

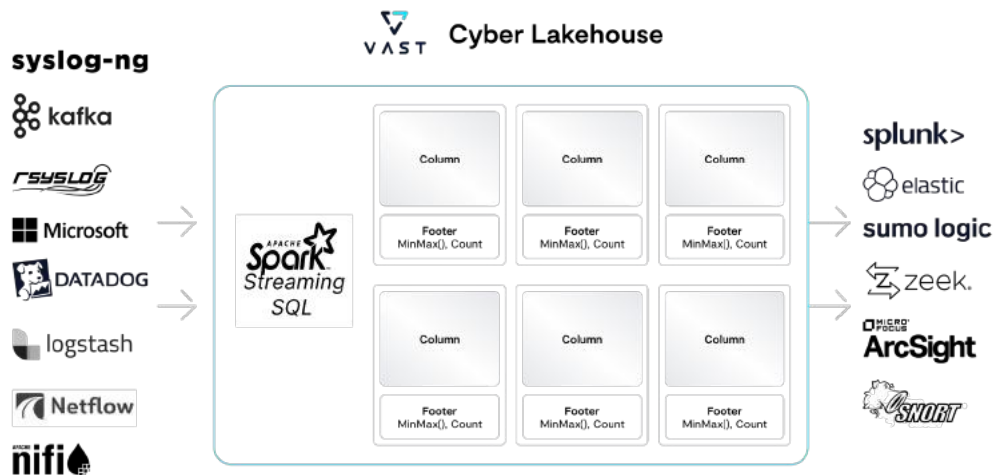
**Predictive Analysis:** The Lakehouse effectively harnesses Spark's rich library of machine learning models to perform predictive analysis, thus proactively forecasting potential security threats based on historical data patterns. This proactive approach empowers organizations to stay ahead of emerging threats by identifying unusual patterns or trends in cybersecurity data. By leveraging Spark's predictive capabilities, the VAST Cyber Lakehouse enhances overall cybersecurity preparedness and response, ensuring a proactive stance against evolving threats.

**Alerting:** When Apache Spark detects suspicious activity or security threats, it can generate alerts and seamlessly publish them to designated Kafka topics. Kafka acts as a central message broker, ensuring reliable and real-time distribution of these alerts to your SIEM.

## Integration with SIEMs for Real-Time Monitoring & Visualization

The VAST Data Lakehouse offers seamless integration with SIEM(s), placing a particular emphasis on enhancing the user experience within Splunk. Through its integration with Splunk via the DB Connect Application, the VAST Data Lakehouse enables Splunk users to tap into the potential of curated and enriched data for real-time insights and enhanced threat detection.

Furthermore, the flexibility of the VAST Data Lakehouse extends beyond Splunk, encompassing various SIEM platforms such as Sumo Logic, Elastic, ArcSight, DataDog, and more. This versatility ensures that organizations have the freedom to leverage their preferred security monitoring and analytics tools while still reaping the benefits of the advanced data processing and storage capabilities offered by the VAST Data Lakehouse.



## Conclusion

The VAST Cyber Lakehouse's capability to efficiently store data perfectly aligns with Memorandum 21-31's requirements and reinforces its position as the ultimate choice for comprehensive cybersecurity data management. The Vast Cyber Lakehouse is more than just data storage; it excels in data curation and transforms information into actionable insights. By harnessing Spark's exceptional data processing capabilities, the Lakehouse enables organizations to swiftly detect and mitigate security threats while also significantly reducing data storage costs. Moreover, its seamless integration with various SIEM platforms ensures interoperability, eliminating vendor lock-in and promoting flexibility. Furthermore, the Lakehouse extends its reach into advanced technologies by integrating with AI/ML frameworks, providing a proactive approach to threat mitigation. In essence, the VAST Cyber Lakehouse is a unified and cost-effective solution that transcends technology boundaries, enabling organizations to navigate the ever-evolving cybersecurity landscape with agility and confidence while meeting regulatory compliance needs and historical data retention requirements.



For more information on the VAST Data Platform and how it can help you solve your application problems, reach out to us at [hello@vastdata.com](mailto:hello@vastdata.com).