

The Next Generation Cyber Lakehouse



Table of Contents

What is a Cyber Lakehouse?	3
What do Cyber Lakehouses need to Deliver?	3
Next Generation Cyber Lakehouse	4
Cyber Lakehouse	5
Real-Time Data Ingestion and Curation	5
Data Optimization & Governance	6
Data Analysis	6
Connectivity to 3rd Party Systems	7
Enhanced Security Posture	7
Scalability and Flexibility	7
Cost Effective Scaling	8
Analytics	8
Data Processing	8
Data Sharing	9
Conclusion	9

What is a Cyber Lakehouse?

Over the last several decades, security & cyber capabilities moved from point solutions with vendor specific tools to broad based data platforms leveraging the latest scale out technologies. The volume, variety and variation of cyber relevant data has increased dramatically, requiring cyber professionals to take advantage of the latest data platform technologies to performantly, and cost effectively, address their cyber needs. Nearly all data has become cyber relevant in today's digital world – requiring a fundamentally different approach.

Data lakehouses have emerged as the architectural pattern by which to store, manage and analyze large scale structured, unstructured and event-based (think logs) data. Their architectures feature storage that is separated from compute, which allows organizations to achieve low cost storage and analytics against dramatically increasing data volumes. The scale out computing components allow customers to meet the performance requirements while not breaking the budget with regard to fixed compute resources. While these approaches are certainly cloud-centric, the advancement of scale-out self-managed technologies like Kubernetes have allowed providers to deliver a cloud-like data lakehouse experience for cyber data in a hybrid infrastructure topology.

A next generation cyber lakehouse provides a technology platform that enables cyber security professionals to collect, store, analyze, manage, alert and action on all their cyber security data at exabyte scale in a hybrid topology.

What do Cyber Lakehouses need to Deliver?

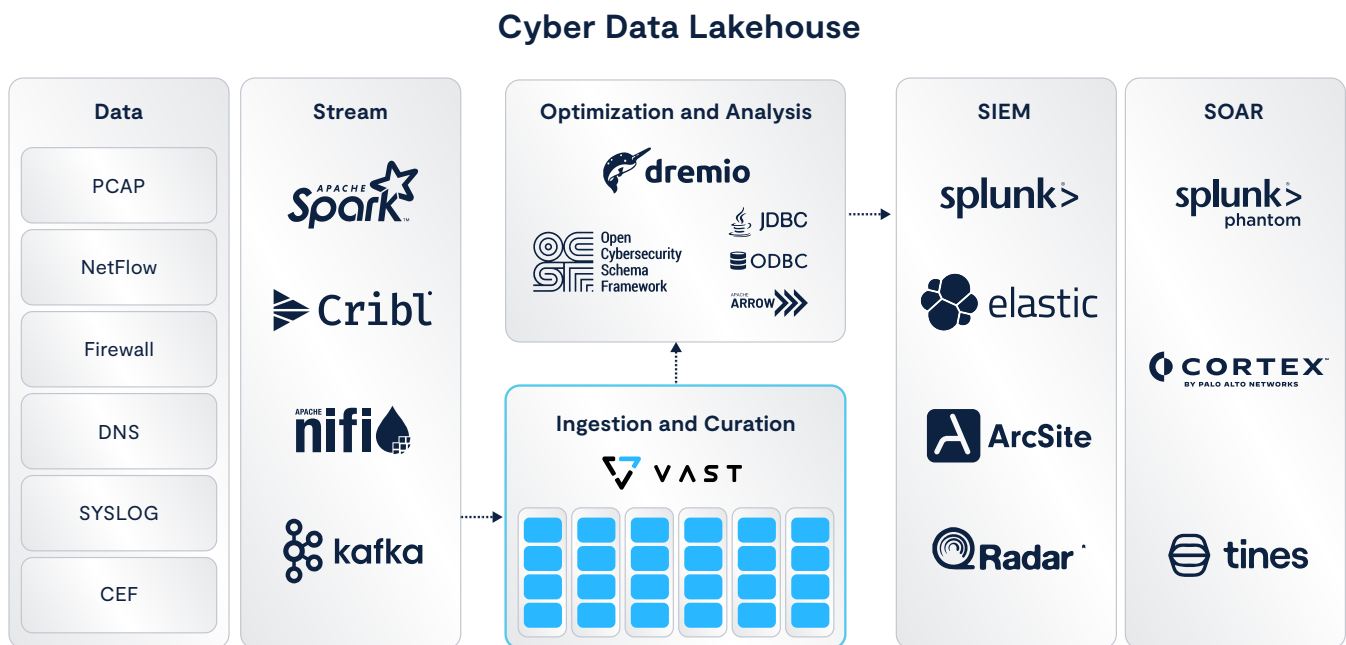
A compelling cyber lakehouse should provide an exabyte-scale architecture addressing every element of the data lifecycle in a cost-effective manner:

- **Data Ingestion:** Ability to ingest data in open data formats from agents and pcap/network/security/observability/event pipelines. This includes data on-premise as well as in the cloud. The volume and velocity of data should be easy to scale and ensure lossless collection.
- **Data Curation:** Ability to write raw data to scale out storage and subsequently transform as needed in a cost effective manner. This includes any apriori data processing for known analytics needs.
- **Data Optimization:** Dynamically optimize data layout, data summarization, and lifecycle for the myriad of continuously changing analytics and compliance needs.
- **Data Analysis (Alerting, Reporting, Root Cause Analysis):** Conduct analytics on structured and unstructured data at scale to ensure: (1) performant structured querying of structure data (2) effective summarizations and aggregations and (3) fast needle-in-the-haystack analysis.

- **Connectivity to 3rd Party Systems:** Provide a commonly used language (SQL) to connect other analytics systems to the data. This eliminates the need to copy data or conduct the same analytics multiple times. This also ensures data is available as SIEM, SOAR, XDR and other security solutions.
- **Data Sharing:** Increasingly, due to regulation and practical concerns, it is important to be able to share cyber data on-demand with other entities within and outside of the organization. Cyber data lake is particularly capable in that it allows large amounts of data to be exported or shared without straining compute resources or introducing complex concerns about granting outsiders access to internal SIEM and analytics systems.

Next Generation Cyber Lakehouse

The integration of VAST Data and Dremio enables any SQL-compliant SIEM, analytics engines, alerting system or visualization tool to create a powerful cybersecurity solution that leverages the strengths of each platform to provide real-time and continuous vigilance. Example 3rd party SIEMs include Splunk, QRadar, Grafana, Elastic, etc. The Cyber Lakehouse is designed to address the challenges of modern cybersecurity by combining real-time data ingestion, high-performance analytics, and seamless integration with Security Information and Event Management (SIEM) systems. Below, we explore the key features, benefits, scalability, and data warehouse capabilities of this combined cyber data lake.



Cyber Lakehouse

Real-Time Data Ingestion and Curation

The VAST Data Platform is a comprehensive, AI-focused data platform that integrates storage, database, and compute capabilities. This approach allows for the handling of massive volumes of data, including structured and unstructured data, at exabyte scale, while maintaining exceptional performance and scalability.

The Platform's DASE (Disaggregated Shared Everything) architecture, developed by VAST Data, leverages Storage Class Memory (SCM) as a write buffer and QLC NAND to deliver a scalable, exabyte-scale architecture optimized for high-performance data ingestion. This enables the platform to support real-time data ingestion and processing, making it suitable for applications that require immediate insights and actions.

Disaggregated Shared Everything (DASE) Architecture: DASE is designed to provide a scalable, high-performance architecture that can handle large amounts of data. It disaggregates storage services onto DPUs, allowing for independent scaling of compute and storage resources. The architecture ensures that data is accessible across all nodes in the cluster, providing a globally visible namespace.

The VAST DataBase is the first modern architecture aimed at delivering the transactional performance of a database, the analytical performance of a data warehouse, and the scale and affordability of a data lake. It takes advantage of the DASE architecture, its ability to log structure writes, columnar compress, is highly parallel, and scalable.

- **Log Structuring and Columnar Storage:** Log structuring enables efficient data migration from rows to columns, preserving a global, queryable namespace. This allows for efficient querying and analysis of data. Columnar storage enables high-performance querying and filtering of data, making it suitable for real-time analytics and data processing
- **Parallelism and Scalability:** The DASE architecture is designed to handle high-performance data ingestion and processing. It can stream data from real-time sources into an Enterprise Data Warehouse (EDW) at high speeds, making it suitable for applications that require immediate insights and actions. The architecture is non-partitioned and can scale to exabyte levels, making it ideal for large-scale data processing and analytics.

This creates a stream processing engine that can convert hundreds of terabytes of event data on a daily basis to rows in a table. This ingestion capability allows us to collect event data from Apache Kafka, Confluent, Pulsar, Apache Spark, and Cribl efficiently while reducing the workload on the SIEM.

Dremio SQL Query Engine: Dremio provides high-speed, automated summarization of data at scale. This makes it easy to apply data transformations, including filtering, sorting, aggregating, joining, casting. These transformations can be quickly built as Dremio Views – governed virtual data sets with layered transformations. This allows for easy structured data access and analysis once the data is ingested and stored in VAST.

Data Optimization & Governance

- **VAST Data Platform:** The VAST Data Platform ensures data integrity and security through advanced features such as Attribute-Based Access Control (ABAC), granular access control, and intelligent threat detection. By leveraging ABAC, the platform provides fine-grained access control policies that consider user attributes, environmental conditions, and resource attributes, ensuring that only authorized users can access sensitive data.
- **Dremio Lakehouse Management:** Dremio's modern data catalog, provides automated data optimization and data-as-code versioning. Dremio automatically compacts small files and groups similar rows together to optimize for performance. All changes to data and metadata are tracked, and the catalog provides fine-grained privileges to control column and row level access.
- **Dremio Reflections:** Reflections map and transform event data collected by the VAST Data Platform into the Open Cybersecurity Schema Framework (OCSF) format, enhancing the efficiency and performance of data queries. OCSF is a vendor-agnostic core security schema that simplifies data ingestion and normalization, making it easier for security teams to analyze and respond to threats. These transformations are saved as Dremio Views, which are governed virtual datasets with layered transformations, ensuring that the data is consistently formatted and optimized for analysis.

Data Analysis

- **VAST DataBase Columnar Storage:** Data is stored in a columnar format, similar to Apache Parquet, but with finer granularity. This allows for efficient query processing and better data compression, which is crucial for handling large-scale data analytics.
- **VAST DataBase Transactional and Analytical Processing:** The VAST DataBase is designed to support both transactional and analytical workloads. It combines the capabilities of a traditional database, data warehouse, and data lake into a single platform, providing ACID-compliant transactions and high-performance analytics.
- **VAST Data's embedded Apache Spark:** Enables custom analytics and machine learning pipelines leveraging the wide ecosystem of Spark libraries, like the machine learning algorithms from the MLlib. This accelerates the processing of cybersecurity analytics, threat detection, and advanced behavioral analysis.
- **Dremio Query Engine:** Dremio's intelligent query engine provides sub-second analytics performance, enabling users to perform live and interactive queries on cloud and on-prem data lakes. This is achieved through advanced query acceleration and optimization techniques such as the use of optimized relational cache, e.g. Reflections, C3 Columnar Cloud Cache and native push-downs.

Connectivity to 3rd Party Systems

The VAST Data Platform provides a comprehensive Python SDK, which can be leveraged to interact with the VAST Database. This SDK enables schema and table management, efficient data ingestion, querying, and modification of columnar data.

- **Dremio Unified Analytics:** Dremio provides seamless unified data access with data visualization, federation, and interactive analytics across all of an organization's data. The connector ecosystem features dozens of integrations with a large number of sources, including object storage, metastores, and databases in the cloud and on premises
- **Dremio Unified Access:** Security Information and Event Management (SIEM) systems can seamlessly connect and interact with data using a variety of protocols and drivers, including Apache Arrow Flight SQL, JDBC, and ODBC. By leveraging Apache Arrow, Dremio enables high-performance data transfer and query execution, significantly accelerating data processing compared to traditional methods. Now you can ensure efficient and scalable data access enabling users to perform live and interactive queries on cloud and on-premises data lakes.

Enhanced Security Posture

- **Proactive Threat Detection:** The combined solution offers proactive threat detection capabilities, leveraging VAST's real-time data analysis and Dremio's high-performance analytics to identify and mitigate threats before they can cause significant damage.
- **Continuous Vigilance:** By integrating with Splunk, the solution provides continuous monitoring and real-time visibility into the organization's security posture, enabling swift detection and response to potential threats.

Scalability and Flexibility

- **Scalable Architecture:** Both VAST and Dremio are designed to scale with the growing needs of modern organizations. VAST's platform supports exabyte-scale data management, while Dremio's cloud-native architecture allows for elastic scaling of compute and storage resources.
- **Flexible Data Management:** The solution supports a wide range of data types, including structured, semi-structured, and unstructured data, making it suitable for diverse use cases such as real-time fraud detection, customer analytics, and IoT analytics.

Cost Effective Scaling

By decoupling the compute layer from the storage layer, organizations can leverage Dremio as a high-performance SQL query engine integrated with VAST Data's scalable and cost-effective data storage platform. This approach offers several advantages over traditional SIEM solutions:

- **Cost Optimization:** Traditional SIEMs often bundle compute and storage together, leading to higher costs as data volumes grow. By separating these components, organizations can scale compute and storage independently based on their needs. VAST Data provides cost-effective, exabyte-scale data storage, while Dremio acts as a powerful yet cost-efficient query engine, allowing organizations to optimize costs for their specific workloads.
- **Flexibility and Scalability:** The decoupled architecture enables organizations to scale compute and storage resources independently, adapting to changing requirements without being locked into a monolithic SIEM solution. As data volumes increase, VAST Data can seamlessly expand storage capacity, while Dremio can scale out its query processing power to handle more complex analytics workloads.
- **Performance and Efficiency:** Dremio's high-performance query engine, built on Apache Arrow, is designed to handle complex queries across large datasets with speed and efficiency. By offloading the query processing to Dremio, organizations can achieve faster time-to-insight without being constrained by the limitations of traditional SIEM query engines.
- **Open Architecture:** Both Dremio and VAST Data embrace open standards and interoperability, allowing organizations to integrate with their existing data ecosystems and avoid vendor lock-in. This open approach fosters flexibility and future-proofing, enabling organizations to adapt to evolving data requirements and technologies.

Analytics

- **Dremio Semantic Layer:** Dremio provides a unified semantic layer that abstracts the underlying complexities of the data lake, offering a consistent data model for end users. This enables seamless integration with business intelligence tools and supports self-service analytics.
- **Real-Time Analytics:** The combined solution supports real-time analytics, allowing organizations to quickly analyze and respond to changing business conditions. This is particularly important for cybersecurity, where timely insights can make a significant difference in threat mitigation.

Data Processing

- **Schema Evolution, Time Travel, and Transactional Support:** The VAST DataBase plugin for Dremio ensures that data is stored and managed efficiently, with support for features like schema evolution, time travel, and transactional operations such as inserts, updates, and partitioning. These capabilities are crucial for maintaining data integrity and enabling seamless data processing workflows.

- **Apache Spark Integration:** The VAST Data Platform integrates with Apache Spark, a powerful open-source engine for large-scale data processing. This integration allows organizations to leverage Spark's machine learning library (MLlib) and other advanced analytics capabilities, enabling them to build custom behavioral analytics models and gain deeper insights from their data.
- **Predictive Analytics and Automated Threat Detection:** By combining the power of VAST Data's high-performance data platform and Dremio's query engine, organizations can perform real-time analytics and machine learning tasks, such as predictive analytics and automated threat detection. This capability is particularly valuable in cybersecurity and other domains where timely insights and proactive measures are crucial.

Data Sharing

- **Bulk Data Sharing & Export:** Conventional SIEM systems are not built to support large data export, unlike the Cyber Data Lake. Because the underlying storage is decoupled from compute, it becomes possible to share large amounts of cyber data with internal and external stakeholders without impacting internal SIEM and analytics systems with computationally-expensive data export jobs. Additionally, due to zero-cost copy semantics of Vast, sharing of the same data between internal stakeholders does not incur any additional storage costs due to data duplication.

Avoiding Complex Access Control Issues. Since the Cyber Data Lake design allows sharing of cyber data without having to provide access to SIEM and analytics systems, it becomes possible to avoid headaches and complexities of granting outside entities access to the same systems that are in use by the internal cyber and operations teams.

Conclusion

The integration of VAST Data and Dremio with Splunk provides a comprehensive cybersecurity solution that combines real-time data ingestion, high-performance analytics, and seamless integration with SIEM systems. This solution enhances the organization's security posture by enabling continuous vigilance and proactive threat detection, while also offering scalability, flexibility, and cost efficiency. By leveraging the strengths of each platform, organizations can achieve real-time continuous vigilance and effectively mitigate cybersecurity threats. The VAST DataBase, with its advanced capabilities for transactional and analytical processing, further strengthens this solution by providing a robust and scalable data management platform.



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.