Hexis CYBER SOLUTIONS
a KEYW company

7740 Milestone Parkway, Suite 400
Hanover, Maryland 21076
info@hexiscyber.com
443-733-1900

# HawkEye AP: Overcoming Today's Compliance and Security Data-Related Challenges

## Executive Summary

With increasing focus on security and compliance, companies today are challenged with collecting, storing, and analyzing large amounts of event data. Making sense of all this data enables companies to more cost effectively address ever-changing compliance regulations and more quickly thwart ever-increasing security threats.

Most compliance and security-related data consists of semi-structured or unstructured event data. However, standard relational database management systems (RDBMS) technology, which is the foundation for most data warehouse and security information and event management (SIEM) systems, performs poorly in the presence of unstructured data and even worse when datasets exceed several petabytes or more. When that happens, enterprise users are forced to make compromises such as reducing the timeframes of stored data or employing a two-tier storage architecture where older data is archived. Consequently, this can seriously reduce an enterprise's ability to detect threats or recall critical data for compliance purposes.

HawkEye AP is a high performance event data warehouse that overcomes the limitations of traditional RDBMS architectures. It is based on advanced technologies such as a columnar database, massively parallel processing, distributed querying and loading, data compression, and a non-transactional model. HawkEye AP eliminates the overhead imposed by standard RDBMS technology, and materially increases the performance and capacity to manage massively large volumes of event data.

HawkEye AP provides numerous business benefits, such as executing queries in minutes or hours, versus the hours or days required by RDBMS systems, or keeping pace with enterprise-wide event collection in multi-gigabit networks. Its clustered and modular architecture enables enterprises to add server nodes incrementally to increase capacity and throughput over time.

Numerous organizations have leveraged HawkEye AP for compliance reporting, policy enforcement purposes and for security and infrastructure monitoring. HawkEye AP has been either the main data warehousing technology or the long-term backend for SIEM solutions where their RDBMS database could store only limited amounts of data. As a high-end example, one customer collects over 6 billion records totaling 1.5 terabytes of event data every day and runs over 1,000 near-real time queries per day to support forensics, HR, and legal investigations.

## Data Challenges Associated with Compliance and Security

Today, companies encounter a number of business imperatives that involve collecting, storing, managing, and analyzing large amounts of event data. These data-related challenges can be classified into two broad areas:

1. Compliance with government regulations and corporate governance

2. Internal and external threats to the security of data, particularly from hard-to-detect advanced persistent threats (APTs)

Let's take a closer look at both of these.

1. **Compliance and Governance-Related Data Challenges**

Fulfilling compliance and corporate governance requirements creates new data challenges. Expanding regulatory compliance requirements and audit challenges make it even more essential for organizations to answer multiple questions:

- What data and computing assets do they have?
- Where are they?
- Who has access to them and under what circumstances?
- When and how are they accessed?
- How are they secured?

Event data originates from a myriad of sources including but not limited to infrastructure devices (switches and routers), banking transactions, Telco call detail records (CDRs), updates to shipping status in RFID records, GPS tracking information, and manufacturing sensor data, and many others. Here are specific examples of how data may be required for compliance or governance purposes:

- Telecommunications companies are required to store call detail records for a variable period but which is commonly seven years or more. Externally, they may be required to provide organizations such as Homeland Security and the police with requested data, such as all phone numbers called by a phone number and then all calls originating from those subsequent numbers. Internally, call detail records can be analyzed for capacity planning purposes.

- Companies may need to store all Web proxy logs and system logs to protect against being sued by a disgruntled employee or another company. Log and e-mail data are often required for forensic investigations, reconstructing past application activity or Web sites accessed by a specific user.

- ATM or other financial transaction records may be needed for fraud detection purposes or to enable forensic or operational analysis.

2. **Security-Related Data Challenges**

Information infrastructures used today by businesses and governments are more complex and dynamic than ever, making them more vulnerable to unprecedented attacks both from an ever-changing threat landscape as well as inside the organization. Dramatic increases in cybercrime and stories detailing large-scale data theft and other data disasters make headlines daily. These attacks cost organizations billions of dollars a year through interrupted operations, data loss, lawsuits, and damage to customer confidence.

The problem is intensifying as cyber-attacks have become much more sophisticated and malicious. Threats now come not only from solo hackers, but from a new breed of cybercriminals who are using innovative combinations of hacking, phishing, and botnet schemes to make money or maximize disruption of our businesses.

The threat is also increasing from inside the organization. Disgruntled or displaced employees can exploit inside access and conduct malicious attacks, either for personal gain or to damage the organization's information networks. Non-malicious insiders – average well-meaning users who disclose data unintentionally or inadvertently let attackers into the enterprise – pose an even greater insider threat.

One of the resources available to organizations in locating and responding to threats is the large amounts of event data

generated by security and IT infrastructure. Data sources can include network and security devices, physical access systems, identity management systems, workstations and servers, and database activity. Organizations can better defend themselves against increasingly sophisticated and targeted attacks by using advanced analytics and correlation across multiple, voluminous data sets. Many of these threats are specifically designed to avoid defenses such as anti-malware and intrusion detection or prevention systems that are based on signatures or reputations associated with known Web sites, vulnerabilities or exploits. Sophisticated adversaries who possess considerable resources and patience, often referred to as advanced persistent threats (APTs), use the element of time to their advantage launching attacks over months or even years to try avoiding detection.

Organizations can perform rate trending or behavioral modeling on key security metrics over long timeframes, developing a baseline of expected behavior for hosts. They can then pinpoint behavior that is abnormal or suspicious by how it deviates from the baseline.

Here is a list of typical metrics that can potentially be used for this type of analysis, especially over extended timeframes that can go 60 days, 90 days, or even spanning multiple months or years:

- **Failed logins:** An abnormal increase in the number of failed login attempts can potentially indicate that an attacker is trying to use brute force to guess passwords.

- **Successful logins:** Successful logins at times atypical of a user's behavior (3 am on Tuesday morning, for example) may indicate account compromise or identity theft.

- **Access counts for application or service access:** As with successful logins, activities typically performed by users such as printing, file sharing, accessing Web sites, or interprocess communication can be significant if they are performed in a manner atypical to a user's normal behavior.

- **Total bytes uploaded or downloaded for Web access:** The amount of Web traffic sent or received can also be useful for identifying malicious activity if it surpasses the standard levels.

## RDBMS Scalability Limitations

Relational database management systems (RDBMS) have long been the de facto standard for almost all data management problems. It is therefore not surprising that RDBMS architectures are the basis for data warehouse and security information and event management (SIEM) systems.

Initially, RDBMS technology was adequate, but as the demand to manage greater volumes of security or compliance-related event data emerged, the limitations of RDBMSs became apparent.

The crux of the issue is that RDBMSs are optimized to supported transactional data, whereas security and compliance-related data consists of events. Transactional data is structured in nature and can change over time. This structured nature means transactions can be summarized in different ways (i.e. show the total sales in North America for Q1 or the average revenue per customer). Transactional data is also continually changing; for example, a customer's order history is updated to reflect new orders or changes in information such as phone number or address.

Event data, on the other hand, can be structured (a transaction is an example of a structured event) but can also be semi-structured (a log message or an e-mail address with a standard header) or unstructured (a Twitter or Facebook status update). Event data also does not change. When an event occurs then this is a historical fact and it cannot be altered or deleted.

The following RDBMS technologies are critical for transaction data but are not required for event data and only incur additional overhead:

- RDBMSs support the commit/rollback protocol where only complete transactions are permanently stored and visible. As a result, RDBMSs have elaborate logging sub-systems that log every change when a transaction rollback occurs.

- Because of continual updates to data, RDBMSs must support locking sub-systems at the row level to prevent data from being changed by other users while a user is performing updates.

- RDBMSs are optimized for unique key queries such as customer number or invoice number. Precise information is known about the data before a query is formulated and the data has been structured to fit into a predetermined model or schema.

To this last point: RDBMSs are sub-optimized for range or pattern matching style queries, whereas most event data is semi-structured or unstructured and must be searched using a pattern search. This means that an RDBMS may provide suboptimal performance when querying event data for specific strings or patterns.

## Mitigating RDBMS Limitations for Event Data Management

Faced with the limitations of RDBMS-based solutions for event data management, SIEM vendors and their customers are forced to adopt a number of mitigation strategies. Unfortunately each strategy is insufficient, risky, or both. Among these flawed strategies are the following:

- **Limited Time Range Queries:** Most SIEM solutions can only keep 30-90 days of data online and the rest offline. This is due to technical limitations of index-based systems which experience performance degradation over time as more data is added to the index. Unfortunately, successful threat detection frequently requires analysis of data over longer time horizons.

- **Data Filtering:** To reduce storage requirements, users use filters to reduce both the number of events and the amount of data stored per event. But filtering pre-supposes that the nature of queries needed in the future is known in advance, which is not always the case. In addition, government compliance requires that all original event data be available.

- **Data Aggregation:** SIEM vendors may group events together to conserve storage consumption. For example, a SIEM may generate an event with a count of 10 rather than store the 10 distinct events. This solution, however, loses individual details about each event that may be required for investigations or other types of queries.

- **Limited Component Monitoring:** Storage requirements can be curtailed by reducing the number of

system components to be monitored. But like data filtering, this tactic pre-supposes the nature of future event analysis. Discovery of new security and system management scenarios may expose the need to access event data from non-monitored components.

- **Two-tier Storage Architecture:** Another option for reducing storage requirements is to remove older events from the RDBMS and archive them into lower-cost compressed storage. If these events are needed then they must be uncompressed and restored to the database. But removal and restoration of event data are time-consuming operations that often require database and system administration resources. A two-tier strategy is not a substitute for adequate online event data storage.

The end result is that RDBMS-based SIEM and log management systems are not able to query the full set of collected data. This can compromise the integrity and effectiveness to combat APTs. Because APTs are long-term subversive efforts, the key to discovering and mitigating them is taking a long-range, broad-based approach to collecting, managing, and analyzing event data.

> "40% of enterprises will actively analyze patterns using data sets of at least 10 terabytes in order to flag potentially dangerous activity by 2016."
> **Gartner Group**

This is supported by the Gartner Group who estimates that by 2016, 40% of enterprises – led by the banking, insurance, pharmaceutical and defense industries – will actively analyze patterns using datasets of at least 10 terabytes in order to flag potentially dangerous activity. One of the keys to defending against APTs is therefore a storage infrastructure that can support collection and querying for increasingly large sets of data.

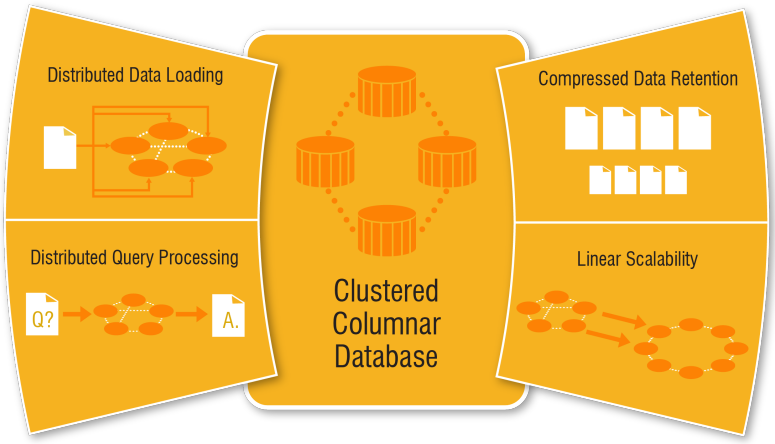
## HawkEye AP: A High Performance Event Data Warehouse

For organizations generating tens to hundreds of gigabytes of data each day, HawkEye AP enables analysts to query their data warehouses for security and infrastructure monitoring and compliance enforcement purposes. Organizations can centrally store and analyze massive amounts of event data over long periods of time while retaining the original source data.

HawkEye AP eliminates the unnecessary overhead imposed by standard RDBMS technology, and materially increases the performance and capacity to manage massively large volumes of event data.

HawkEye AP's core technology is a set of the following innovations:

- Columnar database
- Data compression
- Server clustering via a massively parallel processing architecture
- A non-transactional model
- Open access to other tools via ODBC/JDBC interfaces



## Columnar Database

HawkEye AP is based on columnar database technology. The major difference between row-based RDBMS and columnar databases is that row-based systems store data by row whereas columnar databases store data by column. For example, if a database table consists of the following data:

| EmpId | Lastname | Firstname | Salary |
|---|---|---|---|
| 10 | Smith | Joe | 40000 |
| 12 | Jones | Mary | 50000 |
| 11 | Johnson | Cathy | 44000 |
| 22 | Jones | Bob | 55000 |

Row-based databases store data serialized by row:

```
001:10,Smith,Joe,40000;002:12,Jones,Mary,500
00;003:11,Johnson,Cathy,44000;004:22,Jones,B
ob,55000;
```

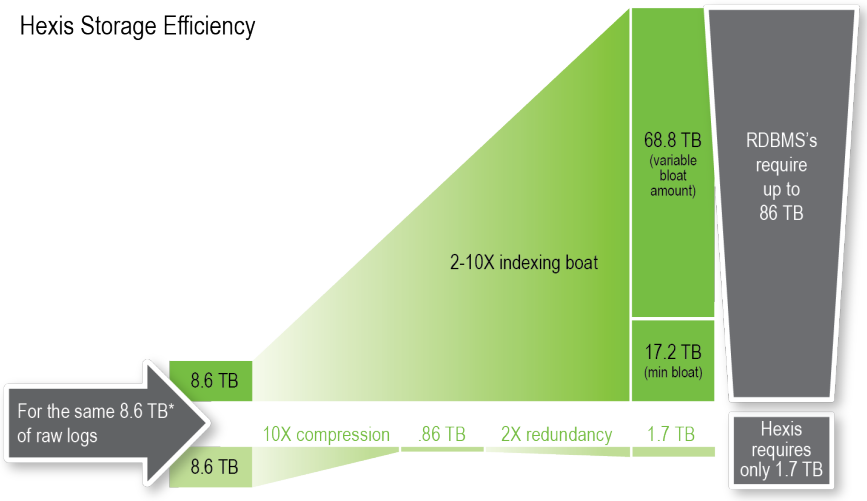
Columnar data store data serialized by column:

```
10:001,12:002,11:003,22:004;Smith:001,Jones:
002,Johnson:003,Jones:004;Joe:001,Mary:002,C
athy:003,Bob:004;40000:001,50000:002,44000:0
03,55000:004;
```

Columnar databases have the following advantages over RDBMS for data warehousing use cases:

- **Superior Query Performance:** All the data for a column is stored together so a pattern can be queried quickly across many rows of data. This difference in speed is further magnified across billions or trillions of records.
- **Higher Compression Rates:** Storing column data serially enables much higher compression rates than row-based RDBMS. Compression is made even more efficient because columnar databases do not need to store the page and row overhead data required by RDBMSs. Compression is increased still further in event data warehouses due to the repetitive nature of event data within a column. Compared to the volume of data stored in an RDBMS, HawkEye AP achieves up to a 40:1 compression ratio, depending on the complexity of the data.
- **No Indexes:** Columnar databases do not need indexes because the columns themselves serve as indexes. In event data warehouses,

Hexis Storage Efficiency




*40 million records per day, at 300 bytes per record, produces 12GB per day or 8.6 TB for a two year storage period.

indexes offer even less value because of the unstructured nature of most event data. The absence of indexes means there is no need for a database administrator to create and drop indexes to balance between query and load performance. There is also no overhead of index maintenance during loading. As a result, the event data load rate will remain constant, no matter how much data has already been loaded.

**Server Clustering**

HawkEye AP leverages a clustered server architecture to distribute workloads and achieve parallel computing on a massive scale. Key elements of this approach include:

- **Near Real-Time Loading:** Event data is created in real time and must be loaded as fast as it is created. To address this requirement, HawkEye AP offers a "trickle-feed load" feature, loading and making data available for querying in near real-time. This is done through special data structures that capture the near real-time data and make it available for querying before it is merged into the data store.

- **Distributed Queries:** Query requests are evenly distributed across servers. Each server conducts its portion of a table scan in parallel with others. The final results from each server are aggregated and returned to the user.

- **Data Redundancy:** Every event is recorded twice in the HawkEye AP server cluster. Each copy is stored on a separate server. Should a server fail, the server that holds the copy of the failed server's event data automatically takes over all query operations for the failed server.

- **Unlimited Scalability:** Many massively parallel processing databases can only scale to a single cluster. HawkEye AP has overcome this limitation by distributing data evenly across multiple clusters and returning results from a single query that spans multiple clusters. HawkEye AP users can deploy federated deployments (multiple clusters) as needed and access the data across these clusters without compromising on load and query speed.

**Non-Transactional Model**

HawkEye AP delivers unparalleled performance versus RDBMS-based SIEM and log management products, largely because of its non-transactional model that minimizes overhead and optimizes the use of computing resources. Key features of this approach include:

- **No Concurrency and Locking Overhead:** Because event data is never updated, the HawkEye AP solution has no RDBMS overhead of row and table locking. Queries never need to wait for updates.

- **No Transaction Log:** Because the commit/rollback model used in RDBMSs is not meaningful for much event data, the HawkEye AP solution avoids CPU, I/O, and storage capacity overhead required to maintain a transaction log.

**Open Access Via ODBC/JDBC**

HawkEye AP's event data warehouse supports an open access interface to event data using database connectivity (ODBC/JDBC) APIs. These APIs enable any third-party Business Intelligence (BI) tool to query data stored within HawkEye

AP. This open access to established BI tools enables faster and deeper analysis, permitting customers to extend the investment and knowledge they have in their BI tools to gain additional insight about their security environment and broader IT infrastructure.

## HawkEye AP Business Benefits

HawkEye AP provides significant business benefits to customers needing advanced event data management:

- **High Performance Queries:** Execute queries in minutes or hours, whereas RDBMS searches often taken hours or days.
- **High-Volume Loading:** Data loading keeps pace with enterprise-wide event collection for gigabit-class networks, with no degradation based on the volume of data stored.
- **High-Volume, Low-Cost Storage:** HawkEye AP uses low-cost Linux-based physical or virtual servers to store highly compressed data. Any combination of SAN, NAS, CAS, or JBOD can also be used. No expensive RDBMS licenses are required. Servers are more efficiently utilized due to the elimination of RDBMS overhead.
- **Low Cost of Ownership:** The solution requires no database administration resources. Data organization is simple and self-tuning.
- **Linear Scalability:** Additional servers can be scaled linearly to provide increased capacity and throughput to match business growth.
- **High Availability:** Built-in redundancy allows continued operation even with a server failure.
- **Data Protection:** Event data is protected against any modifications by outside sources. Data redundancy protects against loss of data in the event of component failure.

## Real World Uses of HawkEye AP

Numerous organizations have leveraged HawkEye AP for compliance enforcement purposes and security and infrastructure monitoring. HawkEye AP has been either the main data warehousing technology or the long-term backend for RDBMS-based SIEM solutions that could only store limited amounts of data.

Here are five real-world examples:

- A Fortune 100 high-tech manufacturer collects over 6 billion records, spiking to over 25 billion records at times, totaling 1.5 terabytes of event data every day from a diverse set of data sources. The manufacturer uses HawkEye AP for multiple purposes such as an event data warehouse for its RDBMS-based SIEM deployment and to help with various investigations. The company conducts over 1,000 near real-time queries a day to support forensics, HR, and legal investigations in seconds or minutes, versus hours or days with its previous RDBMS architecture.
- A major European country's national health service consolidated SIEM and log management services for clinical and nonclinical data for over 77 million patients. Key criteria included advanced threat detection,

correlation, forensic analysis across vast amounts of critical data including electronically protected health information (ePHI), and extensive role-based access and controls required by mandates to separate clinical and non-clinical data.

- A U.S government defense agency complemented its existing Cisco Security solution with support for heterogeneous event data coming from multiple sources, comprehensive security monitoring, and long-term analysis to improve its insider threat detection and analytics capability.

- A European national telecommunications company implemented a corporate-wide cyber-security and log management solution for law enforcement, internal thread detection, and internal security monitoring. This system collects and correlates log data from over 180 source types, including 3 billion call detail records (CDRs) per day.

- A household name financial services provider collects hundreds of gigabytes per day of event data from application, firewall, database, and Windows logs for security investigations and forensics. It currently keeps 4 years of event data accessible, estimated at over 150 terabytes. The organization plans to keep data accessible for 7 years, which their current storage infrastructure will enable them to do.

## Summary

Organizations today face the challenge of responding to increased compliance regulations and audits. In addition, they are at risk from new security threats that are much more frequent and sophisticated than ever before. To address compliance and audit requirements and also to protect themselves, organizations need to implement enterprise-wide solutions that enable them to better manage and analyze their data to make the best strategic decisions.

While traditional SIEM, log management, and data warehouse point solutions based on RDBMSs are appropriate for many business requirements, they are unable to scale to handle the growing amounts of data collected across all parts of the enterprise. RDBMSs are optimized for structured transaction data but most event data is semi-structured or unstructured.

HawkEye AP is a high performance event data warehouse optimized for semi-structured or unstructured event data. It enables organizations to centrally store and analyze massive volumes of events over long periods of time. This empowers enterprises to fortify broad audit compliance processes, respond to business threats, conduct thorough investigations and perform complex correlations on large data sets.

## For More Information

For more information about HawkEye AP and its event data warehousing capabilities, contact sales@hexiscyber.com.

## About Hexis Cyber Solutions

Hexis Cyber Solutions (Hexis), a subsidiary of The KEYW Corporation (NASDAQ: KEYW), provides complete cybersecurity solutions for commercial companies and government agencies. Hexis' HawkEye products offer active, multi-disciplined approaches to achieve a higher standard of cybersecurity that is based on our expertise supporting cybersecurity within the US. www.hexiscyber.com.