White Paper
Information Security
September 2019

IT@Intel: Transforming Intel's Security Posture with Innovations in Data Intelligence

Intel's new Cyber Intelligence Platform provides a context-rich environment that provides value across our entire Information Security organization. It has transformed how Information Security works with a data advantage.

"With the data in the right place and reskilling of our people, we created a force multiplier."
— **Brent Conran**
Chief Information Security Officer, Intel

Intel IT Authors

Jac Noel
Security Solution Architect

Todd Glasgow
Product Owner

Victor S. Colvard
InfoSec Engineer

Dennis Kwong
InfoSec Engineer

Ted Mahar
Global Cyber Response

Eric M. Monroe
Data Architect, Data Scientist

Elaine Rainbolt
Industry Engagement Manager

Aubrey Sharwarko
Data Scientist

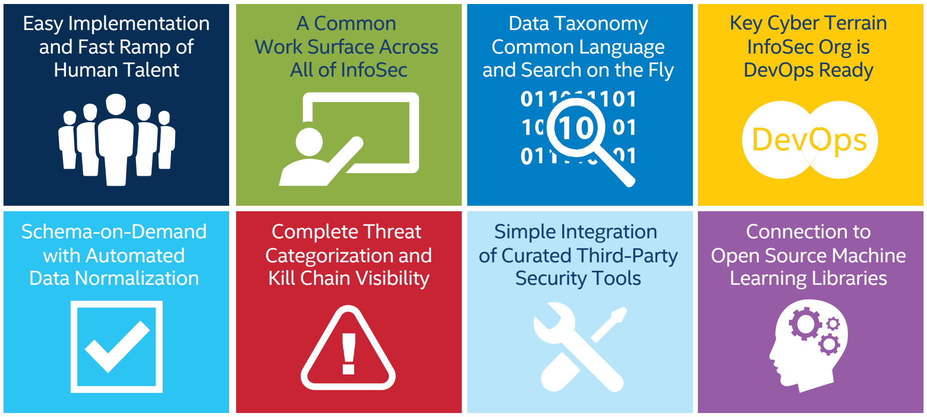
Table of Contents

Executive Summary

Intel IT is transforming our approach to Information Security (InfoSec) by deploying a new Cyber Intelligence Platform (CIP) based on leading-edge technologies, including Splunk and Apache Kafka. Our new platform ingests data from hundreds of sources and security tools, providing context-rich visibility and a common language and work surface around our data. It significantly improves productivity, efficiency, and effectiveness across our entire InfoSec organization. Access to real-time data, streams processing, machine-learning tools, consistent data models, and orchestration and automation capabilities decrease the time it takes to detect and respond to increasingly sophisticated threats and ultimately leads to faster insights for prevention.

Our team deployed this big data solution in just five weeks and immediately started realizing business value. Our CIP infrastructure is based on Intel® Xeon® Platinum processors, Intel® 3D NAND Solid State Drives, and Intel® Optane™ SSDs, providing the compute power our security experts need to gain faster and more intelligent insights while reducing time to pivot between security tools.

Some of the key benefits of our CIP are:


Easy Implementation and Fast Ramp of Human Talent

A Common Work Surface Across All of InfoSec

Data Taxonomy Common Language and Search on the Fly

Key Cyber Terrain InfoSec Org is DevOps Ready

Schema-on-Demand with Automated Data Normalization

Complete Threat Categorization and Kill Chain Visibility

Simple Integration of Curated Third-Party Security Tools

Connection to Open Source Machine Learning Libraries

Brent Conran, Intel's chief information security officer, said, "Deploying and scaling a cyber analytics platform of this size is a journey. We took out a 17-year-old security information and event management system and a nine-year-old logging infrastructure. We put in a brand-new data lake and we modernized our tools. With the data in the right place and reskilling of our people, we created a force multiplier. We are using artificial intelligence and machine learning to significantly increase the depth and speed of our cyber intelligence."

Background

Intel IT's Information Security (InfoSec) organization first implemented a legacy security information and event management (SIEM) system and log data warehouse many years ago. Although state-of-the-art when it was first deployed, the system did not evolve or scale with modern industry standards and presented several challenges:

- **Poor usability.** Developing new detection logic, queries, or advanced statistical analysis required significant expertise, which meant most of the organization had to rely on a small number of experts to create change, hunt for threats, or extract data needed for incident response. We invested substantially in training for new employees, and sometimes lost that talent once they became proficient with the tools.
- **Data dead end.** With very limited data processing capabilities, even simple data transforms, enrichment, or filtering was challenging or impossible. It was also not easy to accommodate new sources of data—which arise every day with the rapid advancement of technology, such as Internet of Things (IoT), cloud, and smart buildings.
- **Data inconsistencies.** Because data existed in silos, it was often interpreted differently by different InfoSec teams, leading to disjointed efforts to prevent, detect, and respond to security threats.
- **Significant technical debt.** The data lake became a data swamp without normalized data and was so customized it was difficult to maintain. Different InfoSec teams used multiple disparate solutions to analyze the same data, often leading to different and/or conflicting results.

These challenges prevented the legacy system from keeping up with the rapidly changing threat landscape. It hindered the efficient prevention, detection, and response to security threats and vulnerabilities, and served only a small portion of the InfoSec organization.

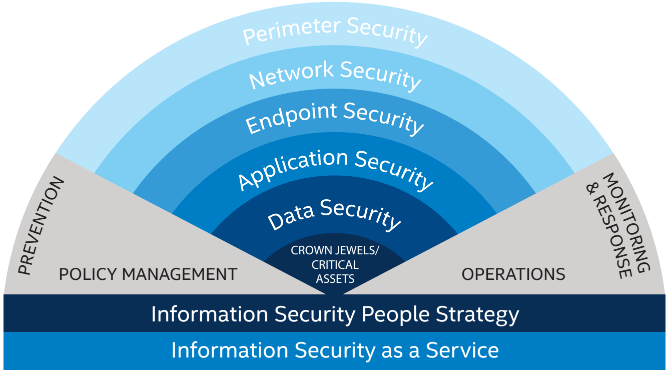
The new solution required a resilient, highly available, and scalable platform that ingests multi-terabytes of data each day from a large number of sources—in near real time. The entire InfoSec organization needed to be able to work in concert to quickly and easily use and construct search queries, detection logic, and dashboards, to better identify, triage, and mitigate threats and effectively communicate InfoSec metrics. We needed a new platform that went far beyond traditional SIEM capabilities to increase the effectiveness of everything the InfoSec organization does: vulnerability management, security compliance and enforcement, threat hunting, incident response, risk management, and more. We needed a modern set of industry-leading and open capabilities that could make data more valuable, usable, and accessible to all of our InfoSec organization.

Intel's Defense-in-Depth Strategy

Automated prevention, detection, and response handles 99 percent of threats—freeing Intel's threat hunters to pursue the 1 percent of advanced adversaries attempting to penetrate Intel's environment.

Intel IT uses a vast array of automated tools and capabilities to help protect Intel from and respond to security threats. Our strategy is to protect at many layers, including the perimeter, network, endpoints, applications, and data. We focus on prevention, detection, and response while simultaneously engaging and educating our employees about information security best practices.

Our defense-in-depth strategy successfully detects and remediates the majority of threats. But advanced cyber-security threats continue to grow in frequency and sophistication. Rooting out that 1 percent of threats—those that are so sophisticated as to escape the extensive automated security environment—requires providing threat hunters with data collected from multiple sources *and* tools. Our new Cyber Intelligence Platform (CIP) provides a means for integrating data sources and multiple tools.



Learn More: "IT@Intel: Advanced Persistent Threats: Hunting the One Percent" white paper

Creating a New CIP

Establish Key Requirements

Evaluate Various Solutions

Educate Decision Makers

Intel IT's New CIP

Our first step toward replacing our legacy systems was to establish key requirements, evaluate various solutions, and educate decision makers about the technology and business value to gain their commitment and support.

Key Requirements

In 2017, we began to evaluate commercial products and open source technologies against our key requirements:

- **Empowerment of the Security Operations Center.** We needed our incident responders and threat hunters to have the access and knowledge necessary to develop and deploy new detection logic on their timeline, without waiting for an expert correlation rule or analytics developer.
- **Ability to scale.** We have a large and diverse group of security data sources. Initially, we needed the platform to ingest about 12 TB per day, which corresponds to about 22 billion security events per day or 8 trillion events per year. But given the projected data growth, we need the platform to scale beyond 50 TB per day in the foreseeable future. We also need to retain a majority of the data for 12 months.
- **Efficient data sharing and transformation.** We needed a message bus to significantly reduce data integration work, promote efficient sharing of data between security systems, and enable data filtering, enrichment, and streaming transformations.
- **Effective user interface.** We needed an easy-to-use, search engine-like user interface with off-the-shelf capabilities for search, reporting, analysis, and visualization to foster rapid adoption by all of our InfoSec organization.
- **High performance.** InfoSec users must be able to find the data they need quickly, whether the event happened an hour ago or nine months ago.
- **Flexibility.** We desired a platform that accommodates rapid change so that we could quickly adjust data models and parsing rules in response to changes—for example, a new source of data or a logging change due to a product update or configuration change.
- **Industry-standard.** We needed a platform that was widely adopted, so that we could tap into the large ecosystem of readily available security tools and capabilities.

  A robust ecosystem includes user communities, public documentation and shared software, third-party training, third-party integrations and support, and other things that increase the value of a product.
- **Extensible.** Intel is committed to a hybrid and multi-cloud model. The solution must easily extend across the enterprise and into the public cloud and software as a service.
- **Highly available.** Intel InfoSec never sleeps. The solution must be fully operational 24/7/365 and through site and data center outages.

Solution Overview

After evaluating a number of off-the-shelf and open source technologies, we chose the following main components for our new CIP:

- **Splunk Enterprise** for the Data Lake and Common Work Surface. Splunk Enterprise ingests and indexes data from a variety of sources into a searchable repository from which users can generate metrics, reports, alerts, dashboards, and visualizations. It contains a collection of data models designed around common security data sources such as firewalls, authentication, and data loss prevention. Splunk Enterprise serves as the "central nervous system" for our CIP.
- **Splunk Enterprise Security** adds SIEM capabilities on top of Splunk Enterprise.
- **Splunk IT Service Intelligence** provides operational mentoring of our Key Cyber Terrain to verify that our key security capabilities are operating as expected at all times.
- **Splunk Phantom** for security orchestration and automated response, enabling automation of mundane tasks, which frees up InfoSec staff to work on higher-value and more sophisticated activities.[1]
- **Apache Kafka** for the enterprise message bus. Kafka is a Pub/Sub messaging system that provides a high-throughput and low-latency solution for ingesting and producing data feeds and performing in-stream data transformations (also known as streams processing). Kafka serves as the "circulatory system" for data across InfoSec systems.
- **High-performance Intel® architecture** provides the necessary hardware foundation for our CIP. This includes Intel® Xeon® Scalable processors and Intel® Solid State Drives (Intel® SSDs).

For more details, see "Solution Architecture."

[1] Intel's InfoSec organization is currently performing a production pilot of Splunk Phantom.

Besides supporting data enrichment and filtering, Kafka allows data to be acquired once and consumed many times, providing economies of scale across all of our security capabilities. Splunk is one of many consumers and is primarily used to analyze, visualize, and report on the data. We currently have hundreds of disparate data sources feeding into Kafka and Splunk. Examples of these data sources include over 200,000 client devices, 800,000 servers, hundreds of firewalls, web proxy servers, network monitoring tools, plus many other types of contextual data, such as IP address mapping, geolocation data, and human resources data.

As shown in Figure 1, our CIP supports the entire InfoSec organization. We found that modernizing our tools and moving our data to a platform with built-in analytics capabilities increases the effectiveness of our InfoSec employees. Our CIP acts as a force multiplier across the entire organization.

Solution Benefits

Our CIP provides Intel with enhanced security in an environment of increasingly sophisticated threats. We educated decision makers about these benefits:

- **Agility.** Our CIP enables us to detect and respond to incidents within hours or minutes, instead of the days or weeks that is typical of SIEM systems built on older technology with point-to-point integration that feeds data in a serial fashion. For example, when we collect data from client devices, we do not have to wait for the endpoint detection and response (EDR) system to process the data. In fact, we recently detected a threat (hacking tool) from a correlation rule before the EDR system that collected the data knew about the threat. Normally, it would have taken several hours for the EDR system to detect the threat; our CIP detected it in minutes.
- **Efficiency.** Threat intelligence is also enhanced with our CIP. We can publish indicators of compromise (such as IP addresses or file hashes) on the message bus, which gives all security systems (such as firewalls and proxy servers) the opportunity to automatically consume the threat information from the bus and keep their internal threat tables current.
- **Easy implementation and fast ramp of human talent.** Splunk is an industry leader in log management—in part due to an excellent user experience, good documentation, and supplier-provided training sessions. Our employee productivity increased, even for those who had never used Splunk before. And because of Splunk's industry popularity, we were able to hire external talent to accelerate our implementation and adoption.
- **A common work surface across the entire InfoSec organization.** The platform has drastically reduced the number of consoles that incident responders need to perform their analysis, which greatly improves incident response times. The threat landscape is constantly changing; with the improved access to data provided by our CIP, all of our incident responders can quickly create or update correlation searches and detection logic that identify events of interest. They have access to a wealth of

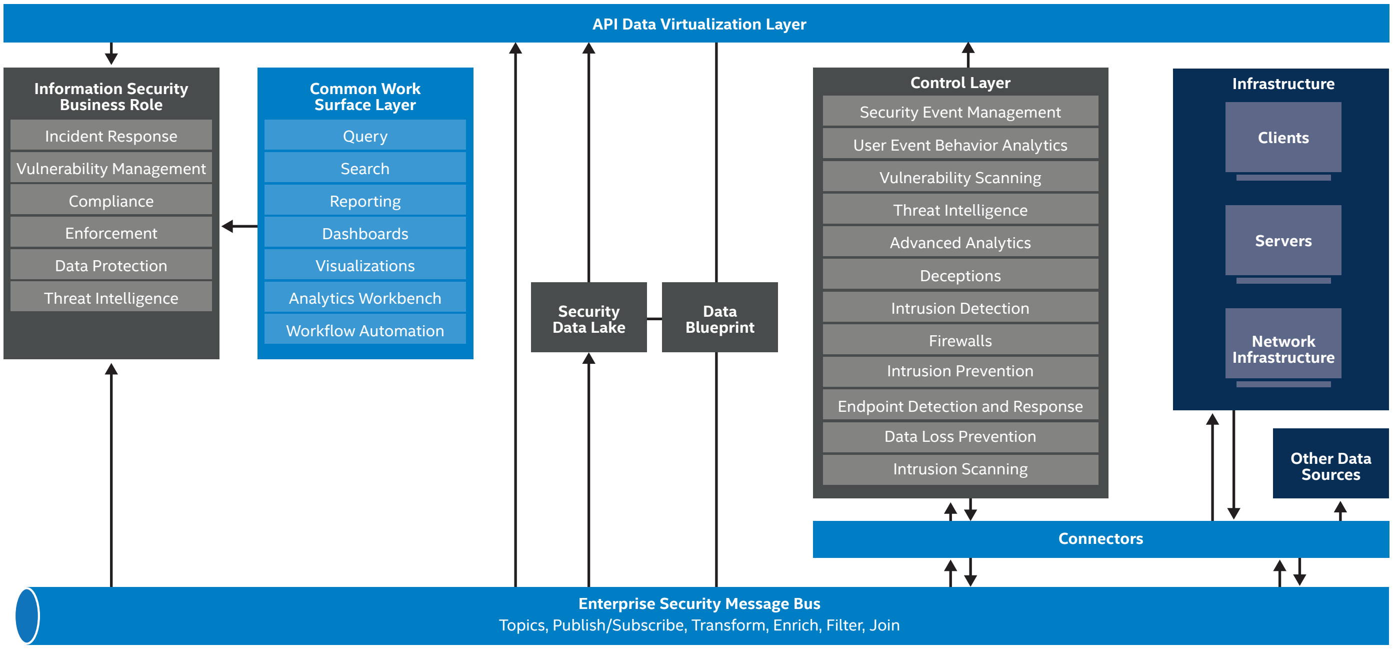


Figure 1. Our CIP supports the entire InfoSec organization and all security control functions, using a common work surface that is integrated with all of our data sources.

historical data to rapidly develop and test detection logic. Once a detection occurs, they can easily search TBs of data to discover the root cause of an event. We can also quickly generate metrics to help us understand the effectiveness of our correlation searches—a capability that our prior generation of SIEM tools lacked.

- **Data taxonomy, common search language, and schema-on-demand.** The entire InfoSec organization can speak the same language and easily retrieve data from our CIP. Everyone in InfoSec has the ability and skill to work directly with the data, increasing the impact of their efforts and the effectiveness of InfoSec as a whole.
- **DevOps-ready.** Splunk's support of incremental integration of data sources and schema-on-demand capability with automated data normalization are key enablers for us to "go fast." See "Agile, DevOps Process" for more details.
- **Reduced technical debt.** As we migrate legacy applications into our CIP, we can significantly reduce outdated, redundant, or custom applications. This allows our technical resources to focus on higher-value solutions.
- **Granular control.** Using Splunk indices, we can grant entitlements to select portions of the data repository, based on "need to know" business cases. We can also manage view permissions for groups outside of InfoSec. For example, a business unit risk manager can access information related to their business unit, without needing access to any other security data.

These combined benefits (see Figure 2) provide us with the ability to detect and respond to sophisticated threats faster than ever before. We use our CIP to respond in near real time—which is a key part of our mission to help keep Intel legal and secure.

Rapid Deployment Best Practices

We used a number of best practices to deploy our CIP into production in just *five weeks*. The following sections provide an overview of how we accomplished this.

A Broad Scope: People, Data, and Technology

While our CIP is in essence technology, it is the people and data that bring value through the technology. As shown in Figure 2, we combine technology, people, and data to transform how InfoSec works.

We used feature-rich, future-ready, industry-leading technologies that can scale well beyond our current needs, to help find and respond to new types of threats. People know how to use these industry-standard technologies, making them a force multiplier across collaboration, efficiency, and adoption. For example, Splunk's large community of users makes it easy to hire people who already have experience with the key components of our CIP. Far more people at Intel know how to use Splunk than knew how to use the legacy SIEM system and log data warehouse. The open ecosystem, featuring app stores and connectors, is a catalyst for collaboration and adoption and helps with alignment to other standards and frameworks, such as open source machine-learning libraries.

Using a common query language and a common data lake across all of InfoSec facilitated the broad adoption of our new CIP across all InfoSec groups. Plus, Splunk products feature highly integrated functionality, such as with our existing DevOps incident management and IT alerting platform. The solution's real-time streaming capabilities enable us to gain faster detection and response, along with greater insights.

The message bus enables us to connect different sources of data, which advances the entire InfoSec organization's thought process. In essence, our CIP provides a feedback loop that instills ideation that leads to better prevention, which helps to protect the Intel brand and reputation.

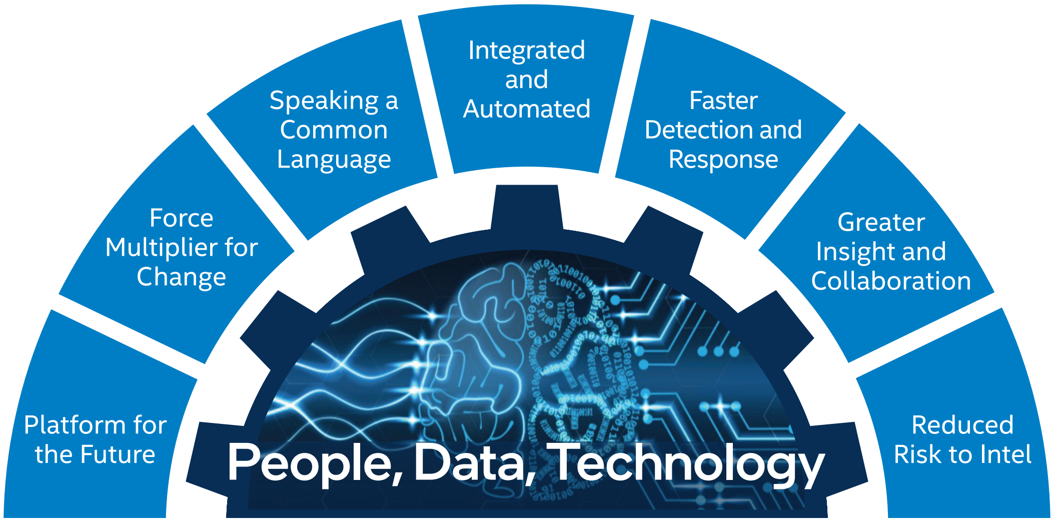

Platform for the Future
Force Multiplier for Change
Speaking a Common Language
Integrated and Automated
Faster Detection and Response
Greater Insight and Collaboration
Reduced Risk to Intel
People, Data, Technology

Figure 2. Our Cyber Intelligence Platform provides a wide range of benefits to Intel.

People and Data Onboarding

While information security is, of course, primarily the purview of InfoSec, collaboration with Intel's business units continues to be key to our success. Much of the data that pours into our CIP daily comes from these business units.

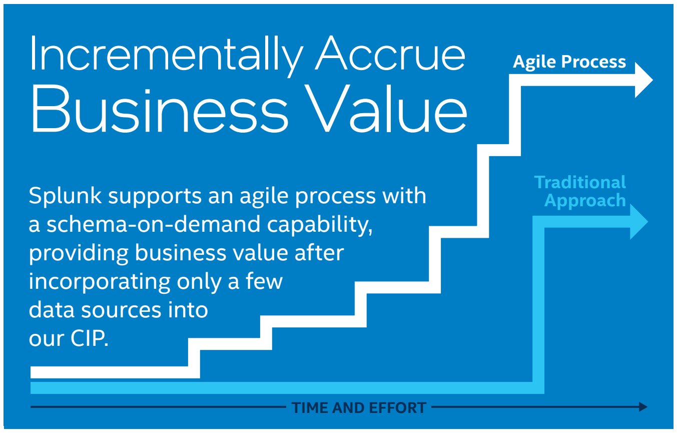
A key enabler to go fast was to prepare stakeholder data. We did not try to prepare all data sources at once. Instead, for each data source, we followed a six-step process (see Figure 3):

1. **Resource alignment.** We worked with the business stewards of the data—they needed to be committed.
2. **Training.** We trained these data stewards to understand the Splunk vernacular.
3. **Business workflow.** We identified a key partner at Intel, who explained how the data relates to a business result. That is, how is the data combined with other data to enable a decision, and can we automate that decision?
4. **Data characterization.** We relentlessly strove to understand the details of the data. Does it need to be joined, filtered, parsed, or enriched? Does a new derived variable need to be calculated?
5. **Data modeling.** How does the data logically flow into a business decision and map into the Common Information Model? At this point, the data source was ready for the Splunk onboarding team.
6. **Incorporation into our CIP.** Once the previous five steps were completed, the data was ready for the Splunk and Kafka onboarding team. The team first developed a Kafka connector or selected an off-the-shelf connector from Splunkbase. Then they mapped elements of the data stream into the Splunk data model.

Agile, DevOps Process

Our CIP engineering and deployment team was small, so we had to use our time wisely. As described in the previous section, we used an Agile process to onboard people and data—an approach that Splunk supports very well.

In the case of a traditional relational database management system (RDBMS), analysts need to plan the complete end-state and define keys; decide how many tables and fields are needed; load the data; then perform optimizations. That means a very long time before business value accrues. Splunk Enterprise is quite different. Its support for schema-on-demand (also known as "schema-on-read," "schema-on-need," and "schema-on-use") provides an opportunity to ingest data and learn what is needed along the way. Even if things changed, we could ebb and flow and pivot nimbly. We did not need the complete library of all data sources before we saw benefits. Instead, we prioritized which data sources created a minimal viable product. We onboarded five or six data sources in a couple days and showcased our CIP's functionality. Incremental wins created enthusiasm for the new platform and provided immediate business value.


Incrementally Accrue Business Value

Splunk supports an agile process with a schema-on-demand capability, providing business value after incorporating only a few data sources into our CIP.

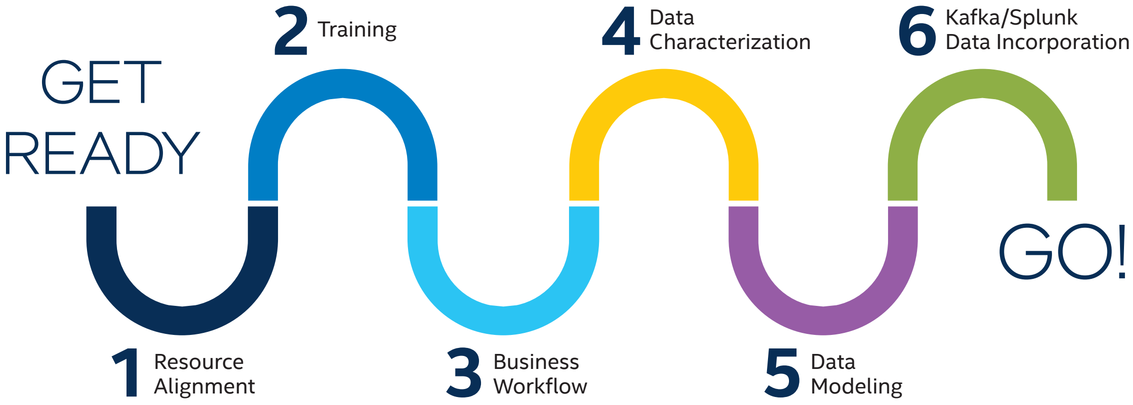


Figure 3. To enable quick deployment of our CIP, we used an Agile process to collaborate with Intel's business units to prepare both people and data.

To further speed adoption and results, we set up a 30-day "challenge" in the incident response group to complete formal training and achieve a baseline level of competency. Then other InfoSec teams conducted their own challenges. These motivational events accelerated our ability to see benefits from the platform.

BDAT Framework

The Open Group Architecture Framework (TOGAF) defines four layers: business, data, application, and technology (BDAT). We map our CIP to this framework as follows:

- **Business.** Splunk's common work surface means InfoSec employees do not have to log into multiple tools; they have one tool for all of their needs.
- **Data.** We developed an understanding of how data maps into business capabilities and we search that data on demand.
- **Application.** Kafka provides an effective way to distribute data (acquire once, consume by many different applications), and as things change, we make the change in only one location. It provides an abstraction layer for many different technologies. Splunk provides an industry-leading set of capabilities to operate on the data. Together, Kafka and Splunk allow orchestration of an automated playbook: a particular event or alert occurs, which triggers an appropriate decision, which in turn generates a specific, defined reaction.
- **Technology.** We are using best-of-breed software technology, such as Splunk, Kafka, and open source machine-learning capabilities. We encourage our suppliers to provide native support for Kafka and Splunk, to facilitate the onboarding of new data sources and system integration. When needed, data sources can be enriched with stream processing on the Kafka message bus before being consumed by Splunk and other applications.

Data Analytics

We focus on bringing in clean, rich data and then using out-of-the-box tools as much as possible to produce data-driven insights. This means we do not need to maintain a lot of custom code—helping to prevent technical debt. We also use threat intelligence feeds, open source machine-learning libraries, and continuous integration/continuous deployment automation to speed our data analytics.

Depending on the data and the goal, our data analytics techniques include classification, clustering, feature selection, or a combination of all of these. Ultimately, our data analytics efforts result in business value across operations, system health, workflow alerting, and anomaly detection.

Splunk's support for schema-on-demand—ingesting data first and then imposing structure on it later—is an important enabler of our data analytics efforts.

Customers Benefit from Splunk and Intel Collaboration

Almost every company in the high-tech industry wants to deliver better artificial intelligence (AI) and machine-learning solutions for big data challenges. However, "better" typically requires close collaboration from a variety of experts. To that end, earlier this year Splunk and Intel began working together to deliver greater business value to joint customers through a more performant, lower-cost customer experience. Several Intel and Splunk solution architects and product engineers formed a team with a broad range of expertise and identified multiple opportunities across a variety of products and technologies.

One of those opportunities is to glean insights from Intel IT's internal Information Security (InfoSec) organization, which uses Splunk products for its CIP. The collaboration team is incorporating feedback from multiple Intel IT InfoSec roles, such as security operations analyst, data scientist, data architect, security architect, and security engineer. As a result, the collaboration team learned, and continues to learn, about customer challenges, their goals, and their success criteria.

The collaboration team is poised to release its first joint reference architecture with Splunk Enterprise running on servers based on 2nd Generation Intel® Xeon® Scalable processors, Intel® NVMe Solid State Drives, and Intel® Optane™ SSDs. This reference architecture provides two important benefits to customers:

- It makes it easy to choose the right solution ingredients (compute, storage, and network) for their Splunk implementations.
- It provides clear guidance for sizing Splunk clusters.

By sharing their optimizations with Intel and Splunk's joint customers, the collaboration team can provide improved performance for customers' business processes while saving customers time, resources, and money.

With a more performant platform, customers achieve faster time to critical insights. OEMs that Intel works with benefit as well, with access to technical documentation that highlights how to design and deploy the most efficient solutions for their customers' specific needs.

splunk> intel.

Solution Architecture

When we began evaluating our SIEM and logging system replacement, we didn't want the same old thing in a new wrapper. Instead, we took a "green-field approach." We needed our new CIP to transform how InfoSec uses and learns from data. We did not want to simply migrate legacy rules, logic, and outdated security practices. We chose transformation over migration. The result is a platform that is built for change. As data sources and data volumes grow, as threats and vulnerabilities evolve, and as business needs change, our CIP will adapt—continuing to provide business value for years to come.

Figure 4 illustrates our CIP architecture. At the foundation are high-performance Intel® hardware technologies, such as high-core-count Intel Xeon Scalable processors, Intel SSDs, and a 10 GbE network. On top of the hardware are the data sources, integrated into the Kafka message bus, which ties all of our security capabilities to the data using a Pub/Sub model. Splunk Enterprise provides the bulk of the user-facing functionality. Other Splunk products, such as Splunk ES, Splunk IT Service Intelligence, Splunk Machine Learning Toolkit, Splunk Phantom, and other third-party applications provide a wide variety of additional capabilities and value on top of Splunk Enterprise.

The following sections describe our use of Splunk and Kafka in more detail, as well as discuss sizing considerations for compute and storage resources.

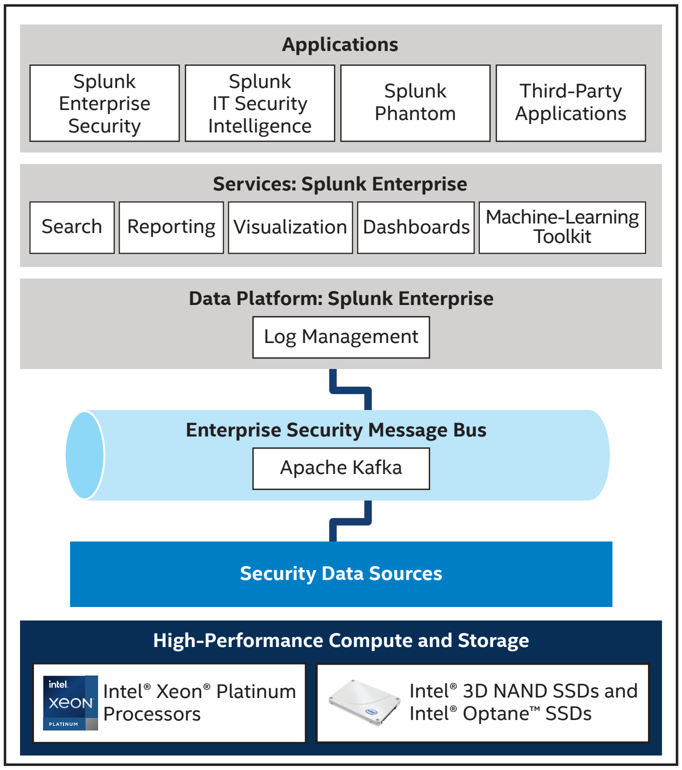

Applications
Splunk Enterprise Security
Splunk IT Security Intelligence
Splunk Phantom
Third-Party Applications
Services: Splunk Enterprise
Search
Reporting
Visualization
Dashboards
Machine-Learning Toolkit
Data Platform: Splunk Enterprise
Log Management
Enterprise Security Message Bus
Apache Kafka
Security Data Sources
High-Performance Compute and Storage
Intel® Xeon® Platinum Processors
Intel® 3D NAND SSDs and Intel® Optane™ SSDs

Figure 4. Our high-performance CIP, built on industry-leading technologies, enables us to identify and respond to sophisticated adversaries.

Data Ingestion and Processing Using Splunk

Splunk Enterprise is a time-indexed event database that ingests all the data and makes it available for search, reporting, visualization, machine-learning, and much more. Splunk helps us create easy-to-understand, data-driven narratives with context and relevance. Because it integrates with popular machine-learning libraries, it enables us to quickly and easily make predictions and test them.

For consistency, we establish operational metrics and indicators that are used across business units. This helps everyone conduct business on the same foundational knowledge and metrics. This is coordinated and enforced by our Knowledge Object Manager, and re-evaluated during planning sessions for use case onboarding. Splunk also provides each InfoSec employee with the ability to create any ad hoc metric, at any time. This flexibility is very powerful and allows InfoSec employees to build custom analytics from the data.

The key advantage of using Splunk for our CIP is that it provides a single version of the truth. The whole InfoSec organization can communicate about and share the data and the queries/searches. This creates efficiency and more consistent, higher-quality data products.

Time Index Event Database

The key advantage of using Splunk Enterprise is that it provides a single version of the truth. The whole InfoSec organization can share the data and queries/searches.

Data Transformation Using Apache Kafka

Historically, data is connected to InfoSec applications using point-to-point integration (see the left side of Figure 5). This approach has several drawbacks. The connections are complex, difficult to maintain, and guaranteed to fail over time due to tightly coupled solutions and brittle integrations—all of which increased technical debt. Adding new data sources or connecting existing data sources to new capabilities is time-consuming. Orchestration is non-existent. Therefore, if something changes in the system, custom code must be added in multiple places. And finally, monitoring and governance is difficult.

In contrast, with a Kafka message bus, the data is acquired once and any number of applications can consume data off the bus. Using a message bus such as Kafka creates a data abstraction layer for upstream producers and downstream consumers, which requires less work to maintain. Data sources and applications can be added or dropped without disrupting the entire system.

Other key benefits are being able to slice, filter, enrich, aggregate, and normalize the data in-stream and in near real time. Examples of these data transformations include the following:

- Associating an IP address with a host name (enrich)
- Combining an event that contains a user account with other data that contains worker-specific information such as from a business unit (join)
- Selecting data based on details such as date, site, or device type (slice)
- Dropping extraneous or redundant data (filter)
- Evaluating a text string to test conformity to a structural pattern, such as validating that an email address has the correct format abc@xyz.com (parse)

Kafka enables us to orchestrate multiple activities from a single platform, which makes monitoring and governance much easier than with the point-to-point approach.

Sizing CIP Infrastructure

Our CIP requires high-performance compute and storage to scale with the ever-growing amount of data streaming into our InfoSec environment. We have designed our CIP to take us well into the future with unlimited scale.

Note that sizing Splunk and Kafka implementations involves much more than the volume of data ingested or published. For example, the number and type of searches, and the number of times each Kafka data stream is transformed are critical factors. The following sections discuss what we have learned about how to size compute and storage for our CIP.

Compute Requirements

Splunk and Kafka are both performance hungry when it comes to CPU and storage. Currently, our CIP consists of 275 servers based on Intel® Xeon® Platinum processors with approximately 10,000 high-performance cores. We recently deployed additional servers to support increased demand for higher compute from new data sources, applications, and users. Our production and pre-production CIP environments store more than 10 PB of data, all hosted on Intel SSDs.

- **Splunk Compute Sizing Considerations.** Splunk requires a high-performance core for every query. This is because the Splunk search tier sends the query to the Splunk indexers—where the event data is indexed and stored—and spawns a thread to run the search. Examples of threads include data model acceleration, machine learning, correlated search, ad-hoc search, scheduled search, report, and dashboard generation for hundreds of users and numerous applications.

  The Splunk search heads and management server are two-socket servers with 24-core Intel Xeon Platinum 8168 processors. The Splunk indexers are equipped with two-socket 24-core Intel Xeon Platinum 8160 processors. These processors provide the necessary compute power to ingest over 12 TB of data per day while supporting at least 48 concurrent searches. It was important to us to achieve the highest amount of storage with the largest number of high-performance cores per server chassis. The two-socket configuration allowed us to hit that sweet spot.

  By scaling out Splunk's indexing tier horizontally, each server examines less data and the search occurs faster. For example, if 10 servers search 1 TB, each server looks through 100 GB; if 100 servers search the same amount of data, each server looks through only 10 GB.

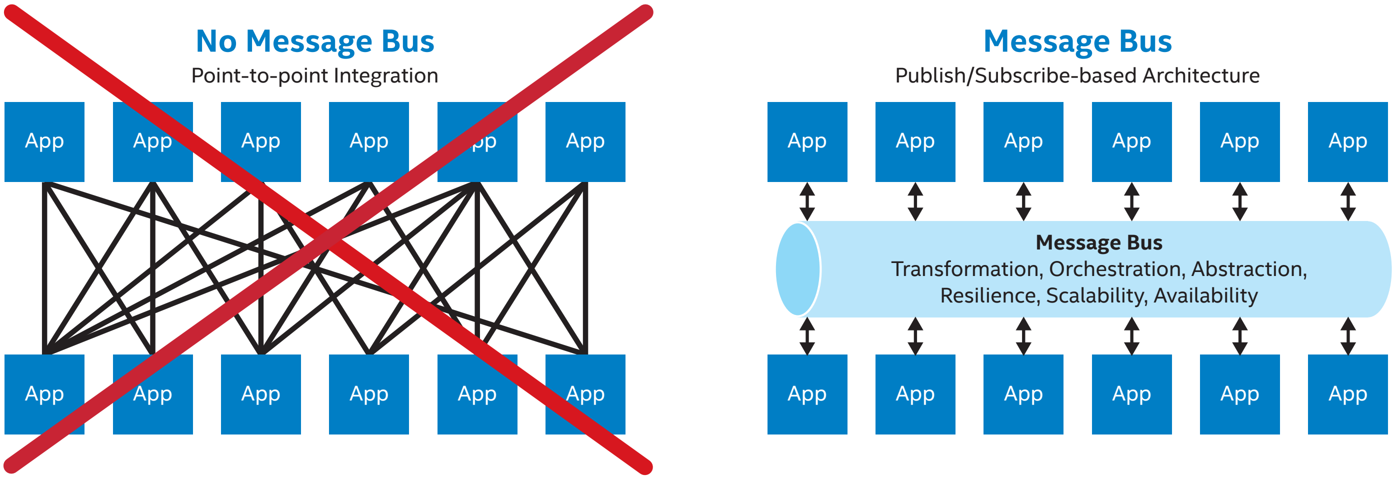

No Message Bus
Point-to-point Integration
App
Message Bus
Publish/Subscribe-based Architecture
App
Message Bus
Transformation, Orchestration, Abstraction,
Resilience, Scalability, Availability

Figure 5. Instead of hard-to-manage point-to-point integration, our CIP uses a message bus (such as Kafka) to improve our ability to transform data, orchestrate activities, and improve resiliency and scalability.

- **Kafka Compute Sizing Considerations.** The Kafka broker servers and Kafka Connect* worker servers run on two-socket 14-core Intel® Xeon® Gold 6132 processors (each server has 28 physical CPU cores). The existing Kafka Stream Processor servers also run on two-socket 14-core Intel Xeon Gold 6132 processors. But as we expand our streams processing capacity, we are equipping new Kafka Stream Processor servers with two-socket 24-core Intel Xeon Platinum 8160 processors (each existing server has 28 physical CPU cores, each new server has 48 physical CPU cores).

  Kafka is designed to take advantage of parallel processing; therefore, scaling out with multiple servers and multiple CPU cores can speed up data stream processing in real time as data volume increases. For instance, we have found that distributing the workload by using 10 broker servers to service 10,000 Kafka topic partitions provides significantly faster results, compared to using three broker servers.

Storage Requirements—Seconds Matter

Our Splunk infrastructure currently uses 24 SATA-based Intel® SSD DC S4500 Series (3.8 TB each) per indexer, providing 72 TB of total data storage per indexer. We chose SATA SSDs because a RAID controller is required to achieve the total amount of storage per Splunk indexer server and the data volume necessary to accommodate Intel's one-year data-retention requirement. We also want to ensure that the 72 TB of data stored on each indexer has the hardware redundancy that RAID provides. With RAID redundancy, a single 3.8 TB drive failure does not force a 72 TB volume to fail and require data restoration. The Splunk Enterprise application provides a high level of resiliency at the software layer, but we wanted this resiliency at the hardware layer as well to offset the risk of losing an entire indexer, which could cause a cascading negative performance effect across the platform.

Every second matters when it comes to searching for security threats (also known as hunting), both to the end user waiting for search results, and to the engineers responsible for the organization's cyber analytics platform. The engineers typically seek to:

- Decrease the total average time required for a search including both start-up search time and search run time.
- Increase the number of concurrent searches across the platform, even with increases in the number of users and ever-increasing volumes of data.

Figure 6 illustrates that Splunk averaged approximately 56,000 searches per day in mid-2018. As we continue to add users, search volume has more than doubled to approximately 115,000 searches per day.[2] During this time, our data ingestion rate increased from approximately 5.5 TB on average per day to approximately 12 TB on average per day.[3]

This magnitude of data growth and number of searches requires our CIP engineering team to constantly look for ways to optimize at all levels of the software stack, including the OS, Splunk Enterprise core, and the Splunk applications. But we also wanted to improve the platform speed and efficiency. In particular, we believed we could reduce our platform's storage I/O wait time, so we could take full advantage of the platform's servers running high-performance Intel Xeon Platinum 8168 processors.

We started our storage optimization work with the Splunk Enterprise Security search heads (ESSH). We evaluated multiple new storage technologies and decided to add Intel® Optane™ SSD DC P4800X Series (750 GB) in our production environment for active application storage.

Our next step is to deploy Intel Optane SSDs in the general use search heads (GUSH). We are also evaluating Intel® Optane™ DC persistent memory to further improve system performance. The active storage on the Splunk indexers could take advantage of the large, persistent, and affordable memory provided by this technology.

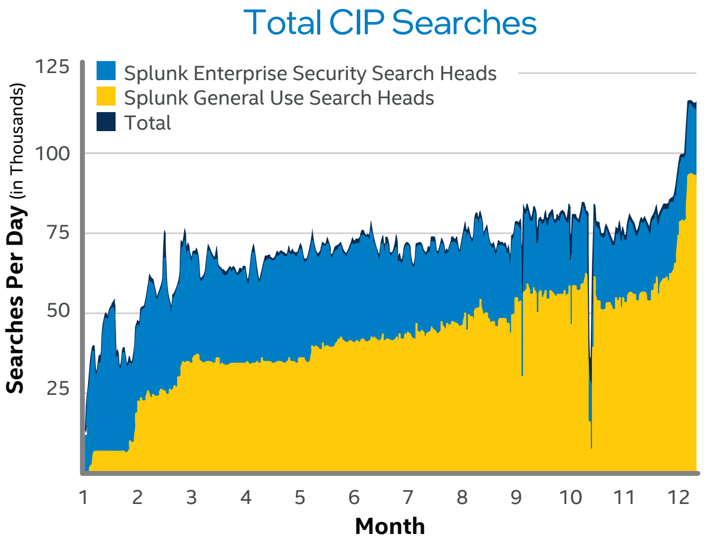


Figure 6. The number of searches performed on our CIP has more than doubled over the last year.

[2] Number of Splunk searches as of July 24, 2019.

[3] The CIP production environment started data ingestion for the Splunk data lake on July 5, 2018. As of July 24, 2019, the Splunk data lake ingests approximately 12 TB per day.

Next Steps

Our CIP provides excellent business value and has quickly gained popularity with a go-to set of capabilities across the InfoSec organization. We plan to continue to increase adoption and add new capabilities over the next few months. Here are some examples:

- **Ramping up adoption.** Our initial deployment of our CIP was focused on replacing our older-generation SIEM and log management system, primarily used by our incident response team, including the Security Operations Center. Additional teams, including vulnerability management, patching/compliance, risk management, and governance are now also using our CIP. We continue to identify opportunities to add more capability and value to the CIP including migration of legacy applications into our CIP, thereby reducing our technical debt.
- **Working with data.** We are expanding use cases for data streams processing through enrichment, transformation, parsing, time-slicing, and filtering on the Kafka message bus. This will not only improve cyber intelligence, but can also help reduce the cost of processing data by downstream systems.
- **Incorporating machine learning.** Integrating our CIP with additional machine-learning tools can increase threat intelligence and help us create workflows that deliver business value in many additional areas, such as operations, system health, orchestration of workflows and alerts, and incident response teams. We currently have machine-learning models at work in the Splunk data lake and applications. In the future we plan to explore deploying trained machine-learning models directly in Kafka.
- **Enhancing performance.** We plan to add Intel Optane SSDs for other use cases that require very fast I/O. For example, the hot or warm storage on the Splunk indexers and the temporary working datasets on Splunk Enterprise and Splunk IT Service Intelligence search heads. Replacing SATA-based SSDs with SSDs that use the Non-Volatile Memory Express (NVMe) with PCIe should help reduce the performance bottleneck created by the RAID controller. And, depending on the size of the SSDs, we may be able to reduce the number of SSDs per server.
- **Improving reliability.** Another area where Intel Optane DC SSDs may be beneficial is improving endurance. Most of the data in our CIP is written only once and read many times, so in general we are less concerned with SSD endurance than with performance for Splunk indexers. However, Splunk search heads are different. Searches are launched continuously from a number of search heads. They query the data and the results are written to temporary areas, and summary data is created. Therefore, search head SSDs do not need to be as big as indexer SSDs, but they do need high endurance. Intel Optane DC SSDs are designed for high write environments and offer ultra-high endurance.

> "Intel Information Security is much more agile than we've ever been in the past. But we need to continue to hone our skills."
>
> *–Brent Conran, CISO, Intel*

Conclusion

Our CIP helped our organization become more productive and efficient. We transformed the way InfoSec works with technology and data for the entire InfoSec organization, giving employees a common language and work surface. We made data accessible, usable, and valuable to all of InfoSec. As our chief information security officer, Brent Conran, says, "Intel Information Security is much more agile than we've ever been in the past. But we need to continue to hone our skills. With artificial intelligence and machine learning there is more data, which means more risk. But it also means more rewards…not just cost avoidance…we are saving real green dollars."

Built with a combination of Intel architecture, Splunk, and Kafka, our CIP enables us to respond to threats faster, provides insights into faster prevention, and helps reduce risk.

A crucial aspect of our CIP is the ability to combine machine learning with streams processing and rules-based logic to automate and orchestrate the mundane, and filter out false positives—which can be as many as thousands or millions every day. We are delivering contextually rich data and true threat detections to an InfoSec analyst or downstream system. Ultimately, our CIP not only lets us move faster, but also helps us continually improve the effectiveness of our entire InfoSec organization. This agility is essential in keeping Intel legal and secure as we continue to develop new products, enter new markets, and support new customers.

IT@Intel

We connect IT professionals with their IT peers inside Intel. Our IT department solves some of today's most demanding and complex technology issues, and we want to share these lessons directly with our fellow IT professionals in an open peer-to-peer forum.

Our goal is simple: improve efficiency throughout the organization and enhance the business value of IT investments.

Follow us and join the conversation:

- Twitter
- #IntelIT
- LinkedIn
- IT Peer Network

Visit us today at intel.com/IT or contact your local Intel representative if you would like to learn more.

Related Content

If you liked this paper, you may also be interested in these related stories:

- Advanced Persistent Threats: Hunting the One Percent
- Security Architecture Enables Intel's Digital Transformation
- Enterprise Architecture: Enables Intel's Digital Transformation
- Securing the Cloud for Enterprise Workloads: The Journey Continues
- Enterprise Technical Debt Strategy and Framework

For more on Intel IT best practices, visit intel.com/IT.

Contributors

Bill Brasse, Program Manager, Intel IT
Frank Ober, Enterprise Architect, NVM Solutions
Jeff Sedayao, Industry Engagement Manager, Intel IT
Merritte Stidston, Technical Specialist, Sales and Marketing
Jerome Swanson, InfoSec Engineer, Intel IT
Sandeep Togrikar, Enterprise Architect, Data Center

Acronyms

BDAT	business, data, application, and technology
CIP	Cyber Intelligence Platform
EDR	endpoint detection and response
ESSH	Enterprise Security search head
GUSH	general use search head
InfoSec	Information Security
IoT	Internet of Things
RDBMS	relational database management system
SIEM	security information and event management
SSD	solid state drive

Intel technologies may require enabled hardware, software or service activation.
No product or component can be absolutely secure.
Your costs and results may vary.
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

0621/WWES/KC/PDF 340007-002