

Bringing the Future Forward



Real-world ways data
can solve some of today's
biggest challenges



Data. It's Everything.

As technologists and software developers, we face challenges and roadblocks every day. One of the most potentially damaging challenges is out of our control: hype. Startups hoping to land a round of funding or get acquired wait for that sweet spot between the Peak of Inflated Expectations and the Trough of Disillusionment, to borrow Gartner's terms. I've seen promising technologies fail because the timing wasn't right. And of course, the opposite is just as true.

I firmly believe that fundamentals will always beat hype in the long run. And when it comes to technology, software and business, nothing is more fundamental than data. If you want to judge a technology trend, strip away the hype and see if the technology in question is built on a sound foundation of data.

Blockchain provides a perfect example. Not so long ago, and thanks in large part to the hype surrounding "Bitcoin millionaires," investors started throwing money at blockchain startups which, unsurprisingly, haven't turned into unicorns yet. But that isn't the fault of blockchain. The technology itself still has the potential to transform online security and revolutionize the financial services industry. Not because of hype. Because of data.

In this book, we explore the most promising use cases for data-driven technology and Splunk, the Data-to-Everything™ Platform. In the chapter on blockchain, for example, we ignore the hype and talk about the practical ways blockchain can break down barriers of trust in multiparty interactions and enable the next digital revolution.

We discuss how real-time monitoring of medical device data can transform healthcare organizations by making them more responsive, more secure and more effective. Imagine how different 2020 would have been if hospitals, epidemiologists and public health officials had access to a predictive, data-driven monitoring system.

In 2020 we often found ourselves questioning what was and wasn't real — especially online. Social media manipulation in general and bots in particular can be used to alter public opinion and affect the outcome of elections. We talk about bot detection and the ways in which data can be used to mitigate attempts to divide, discredit and disorient.

And you can't talk about 2020 and elections without acknowledging the unprecedented confusion that followed Election Day in the United States. In a data-driven world, that confusion wouldn't happen. The time series data generated by voting machines can be a valuable measure of reliability. We have a chapter devoted to this use case as well.

In this book, we also look at new ways data is being applied to solve some of the oldest challenges of the digital revolution, including advances in the prevention of financial crime and more comprehensive, predictive ways of using data to improve information security.

Hype is phony. Data is real. Like science, data provides clarity and certainty to forward-thinking, intelligent leaders who seek it out and use it to make their most important decisions. If you want to become a successful data-driven leader, this book is a great place to start.



Tim Tully
Former CTO, Splunk

Foreword

The Future Is Already Here

“Immer weiter” is one of German pop star Nena’s favorite phrases. Loosely translated, it means “on and on,” or “ever further.” This phrase aptly describes the evolution of using data to solve big problems, one that goes on and on as the analytics of information leads to more discoveries each day. That’s the purpose of this book — to examine how data in today’s world continues to unlock the use cases of tomorrow. The use cases we discuss, however, are not merely the stuff of futuristic fiction. They are actually available today with existing products and technology, just waiting to be manifested.

This book explores lesser-known use cases that have begun to be actualized by Splunk, the Data-to-Everything Platform. Each chapter, written by an expert in their respective data-centric field, starts with a brief explanation of a data-driven use case, and then provides enough technical detail so readers can envision implementing it themselves.

For instance, we have a chapter devoted to blockchain, explaining how the ledger-based technology breaks down barriers of trust in multiparty interactions. Another chapter explains the real-time monitoring of time series data emitted by medical devices, an important topic for a world that needs reliable medical infrastructure to combat disease. Yet another chapter discusses the detection of bots in social media, a critical measure that would guard against the spread of misinformation for industries, consumers, elections and more.

Mitigating financial crime has also been top-of-mind, and we have addressed this issue with a chapter on a new way to use graph algorithms for fraud detection. On the general topic of security, information security has historically operated reactively, responding after the fact to security alerts. But we want to become more prescriptive with handling security incidents and make informed decisions driven by data. The chapter on security metrics addresses this topic. For instance, what if we speed up incident resolution using data from Splunk to power a virtual-reality-lead whiteboard, providing better context and automation and ultimately enhancing the SOC experience?

For veteran information technologists, we close out the chapters with a new way to ingest and interpret syslog, for easier-to-create use cases from its raw data. This can be applied to multiple industries and will upgrade the overall experience with syslog use cases.

We hope this diverse set of use cases compels you to try them out in one form or another and inspires you to come up with your own, as data continues to revolutionize every aspect of our businesses. The use cases of tomorrow are only figuratively named as such, because they are being implemented somewhere, today, as we speak. Enjoy and “immer weiter.”

Nimish Doshi
Principal Systems Engineer, Splunk

Nimish Doshi is a principal solutions engineer at Splunk, where he has worked for over a decade. He works with some of Splunk’s largest customers to ensure their customer success. A prolific contributor to Splunk’s app site, Splunkbase, Nimish is also the main author for Splunk’s Essentials for Financial Services Industries app and the SplunkStart app.

Table of Contents

Certifying Election Results With Greater Confidence	5
Accelerating Enterprise SOC Incident Triage and Collaboration in VR	12
Detecting and Preventing Financial Crimes With Graph Algorithms	19
Real-Time Social Media Bot Moderation Solutions That Could Save Democracy.....	26
How Data Keeps Hospitals Healthy.....	33
How Data Can Help Score Your Cloud and Organization's Security.....	38
Straight Outta Syslog: A New Look at an Age-Old IT Data Collection Problem.....	49
A Future We Can Trust.....	58





“Those who cast the votes decide nothing. Those who count the votes decide everything.”

– Anonymous (often incorrectly attributed to Josef Stalin)



Certifying Election Results With Greater Confidence

Each U.S. election seems more volatile than the last. And it's not just on the debate stage — think of all the partisan ads bombarding every platform and device during election season. In light of the growing noise, there have been accusations and evidence that the accuracy and the validity of our elections have been under attack from both domestic and foreign actors. It's not surprising that many people approach voting and elections with some degree of skepticism. But the truth is that the election process has been, and continues to be, extraordinarily accurate and **almost impossible to hack** on any meaningful scale.

Bad actors can try to affect elections but only in limited areas. Investigation and facts, responsible media channels and, ultimately, an educated electorate all thwart disinformation campaigns aimed at social engineering. IT security tools, good security hygiene and constant vigilance can prevent traditional hacking techniques from infiltrating local election boards' IT systems. Trying to alter the actual voting or tabulation systems either electronically or physically on a scale large enough to affect an election has been unrealistic. In part, that's because most of the machines involved in elections aren't connected to a network. Furthermore, 80% of the United States still uses paper ballots or paper audit trails.

But all that doesn't mean that devices used in the election process aren't susceptible to malfunction or human error that can have costly results. Fortunately, being able to see all the data and take action accordingly can be done with Splunk.

How do U.S. elections work?

Elections encompass curious collections of state and local agencies, various types of voting equipment and myriad civic-minded citizens who staff the elections as poll workers. Each election, poll workers are trained to handle ballots correctly throughout the voting process, much like law enforcement handles evidence through a chain of custody. While some states have paper ballots and pens, other states use ballot-marking devices (BMDs). BMDs are essentially electronic tablets that present the ballot electronically, allow voters to select their candidates, and then produce a human-readable version of the ballot that is submitted into the tabulation machine. That process happens millions of times during every election, including on early voting days and on Election Day until the polls close.

Once the polls close, the votes need to be counted, and election officials need to certify those results. The higher the degree of confidence that the certification can be given, the better everyone can accept the outcome. Timing is key in this process. Election Day is the last day someone may cast a vote, but not the last day election workers can count it. County and state election boards have a limited and [varied time span](#) after voting ends to certify the results of an election, which varies by state, can range from a few days to a month, and offers significant pressure for the board to certify elections as soon as possible.

Because of these compressed time frames, and the importance of providing an accurate and trustworthy result, election boards need tools to evaluate the voting process more quickly and broadly than the traditional manual methods allow. Manual methods are prone to human error and often restrict the breadth and depth of verification that can occur within the certification window — which can lead to a lower level of confidence in reported results.

Most states use voting machines from Election Systems & Software (ES&S), Dominion Voting Systems or Hart InterCivic to tabulate results. These three companies own close to 90% of the domestic market. Each of their voting machines creates audit logs of all the activities that occur during an election, including user interactions and system events.

Enter Splunk. The platform can easily ingest and interpret logs to look for any anomalies that may affect the confidence and certification of an election — all with greater speed and accuracy than manual analysis. Here are a couple of areas in which Splunk delivers value to electoral boards looking to fine-tune Election Day processes.

Splunk value	How is value measured?	Impact for the electoral boards
Faster and full-fidelity examination of all voting machines	All events on thousands of voting machines can now be examined for anomalies in minutes, instead of taking days or weeks to manually sift through a representative subset of randomly chosen machines	Highest confidence in the election certification
Discovery of unexpected information	Ensuring things go as expected leads to higher confidence in certification	Better insight into improving elections practices
Improved resource allocation and appropriation	Examining voting metrics can identify which precincts need more or fewer machines	More effective asset deployment, and data can substantiate budget requests for additional equipment
Poll worker training opportunities	Ballot machine events can uncover any number of inefficient practices or missteps, either intentional or unintentional	Reducing inefficient practices or missteps provides smoother and more accurate elections, and data can substantiate budget requests for additional training
Enhanced ADA reporting	The American Disabilities Act provides modifications for voters with disabilities, which elections must follow and report	Event data can be used to satisfy ADA reporting requirements and aid in appropriate equipment request justifications

But let's dive deeper into just what the big rocks are and how Splunk uses data to help users move them in the right direction.

The challenges of election certification

Without Splunk, election officials manually collect compact flash cards from each voting machine and transport them to election headquarters once polls close. These flash cards contain the audit files (logs) for the machine with a record of every event that occurred on that machine. Because there are often hundreds, if not thousands, of voting machines in a county, and each machine can have hundreds of events transacted during the elections, there is too much data to manually sift through before the certification deadline.

This leads to different strategies to determine if any tampering happened. One method is examining a random sampling of machines across multiple precincts. If officials uncover anything unusual, they can take a closer look at that precinct. But this process depends to some degree on chance, and officials can realistically only check a small sample of information. Yet certain machine tasks need to be validated. For instance, officials must check when machines are opened and closed for voting to ensure that voting started and ended at the correct times. Just the process of checking the appropriate time stamps for every machine can take hours to validate. And because the election boards typically focus their money and resources on the voting process, this critical validation is a manual process with little, if any, automation.

Meeting the need with data

Search and report in moments

Here is a sample of event data from a ballot device:

1672,11/4/20,2057,Terminal Opened,PS251233,426,17:01:10,V5135875

There are a few fields that are easily understandable, like the date, time and the message of what is being reported. But there is other critical information contained here, such as the event *Code* (1672), the machine number (V5135875) and the *Precinct Code* where this machine was located (426).

The first advantage is Splunk's ability to rapidly ingest data and index all of the fields, making quick searching and reporting possible. Election workers can transfer the audit files from the flash cards to a designated directory on the Splunk server to make the data almost instantly ready to use. Whereas in the past it would take hours to determine that all of the machines that were opened were also closed, Splunk can provide an up-to-the-moment validation as that data is loaded into it.

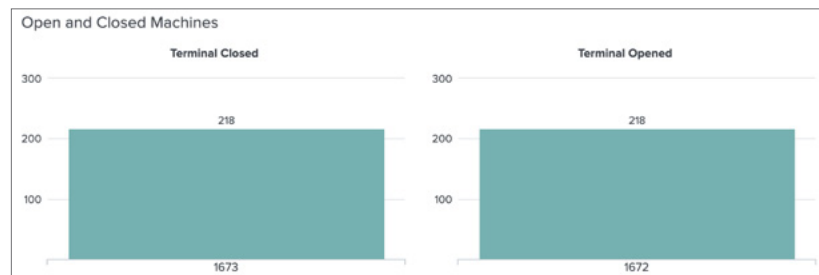


Figure 1

The query for producing this result is straightforward:

```
index=polling Code IN (1672,1673) | stats count by Code, Event
```

Deconstructed, the “index=polling” refers to the searched events being kept in a Splunk repository (index) named *Polling*. We are looking for all the events where the event *Code* is either “1672” or “1673.” Then they can be totaled by event *Code*. There is a little more that is not seen in this figure — the selection of a time frame (in our case, Election Day) and a visualization option to display our results in a column graph.

Users can then examine or group together with other machine events to discover anomalies or quickly validate that expected actions occurred, as noted in the following figure.

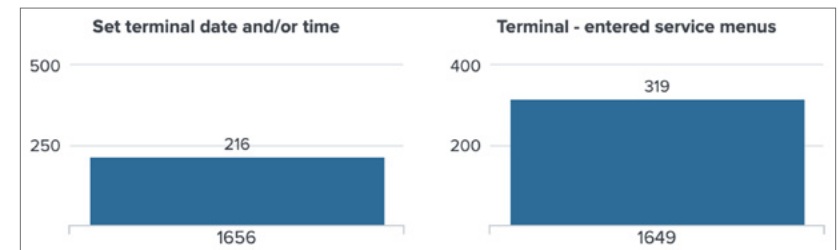


Figure 2

The query for the above chart is:

```
index=polling Code IN (1656,1649) | stats count by Code, Event
```


Tap into health of critical devices

There is also the opportunity to look closer at situations that may be occurring at the polls that may not normally be uncovered. The ballot-marking devices (BMDs) may receive a low battery warning (Code 1619, in figure below) at the end of an election day, but a critically low battery (Code 1622) might be an issue.

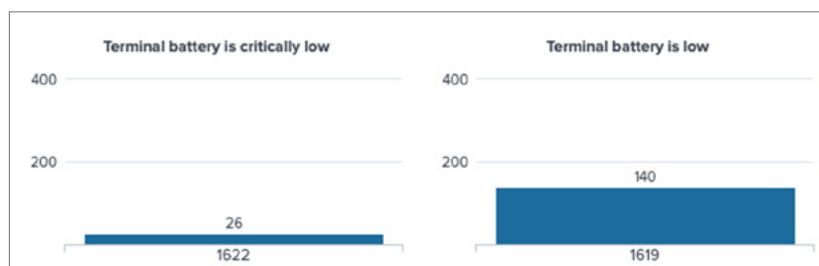


Figure 3

In this case, by clicking the chart, Splunk allows the user to drill down to the specific events that show which devices reported this error and where they are located. Notice the far left-hand column in the figure below shows contextual field names like *Election ID*, *Precinct Code* and *Event*.

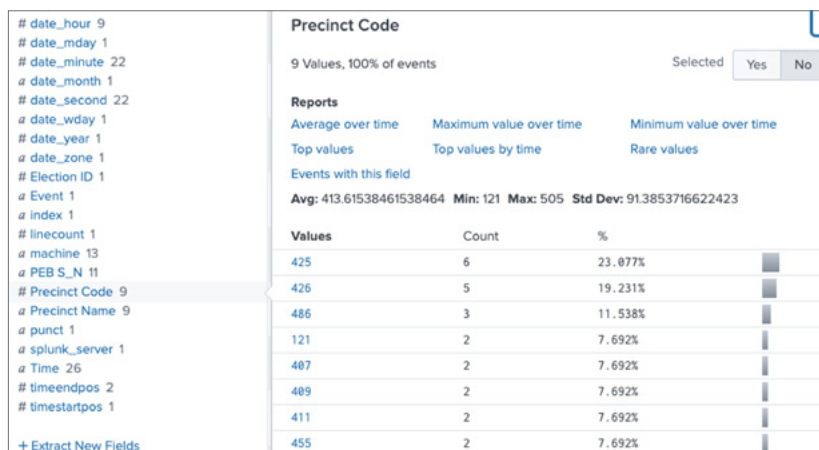


Figure 4

Here we can see that a quarter of the warnings come from one precinct. A more detailed examination (which would require local knowledge of the precinct and the machine's numbers) found that the machines receiving these warnings are the ones that are used for curbside voting. Curbside voting is an accommodation for voters who can not easily physically enter the polling place, so a BMD is brought to the voter to cast their vote. When poll workers return the BMD to the polling place, the standard procedure is to return it to a charging station. These machines were not recharged, hence the warning. Before the next election process, officials can change the poll worker training to explicitly call out returning curbside voting machines to their charging stations to avoid potential situations that might prevent voters from casting their ballots because no charged machines are available.

Another view of this data can provide the number of votes that are cast on curbside devices by precinct. This can later be used for ADA reporting or appropriate allocation of curbside BMDs into precincts with higher populations of voters who need that accommodation.

Finally, by using lookup tables that map onto event data, users can get an even more comprehensive picture. In this case, precinct names — which are not contained in the record — can easily be returned along with the events to offer more meaningful information than what is in the event by itself; this way, anyone can understand what gets reported.

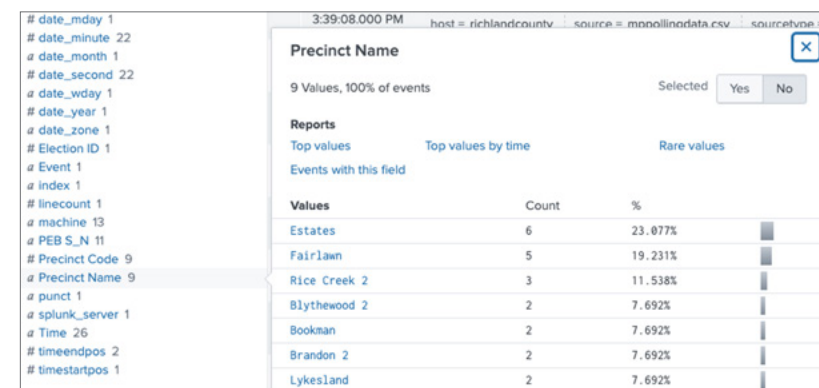


Figure 5

See what's taking place when and where

Aside from examining discrete events, using Splunk allows us to examine transactions. In the case of an election, that means the actual vote itself. To be clear, the audit logs do not contain the ballot information (which candidates were chosen or whose vote was cast). Rather, there are two events shown: that a vote was started and a vote was completed. Using the timestamps in the events, we can then calculate how long a vote took, what the average vote time was on a machine or in a precinct, the longest and shortest vote times, and how many votes may have begun but were not submitted normally. Splunk can show all of these almost instantly once the software has ingested the data.

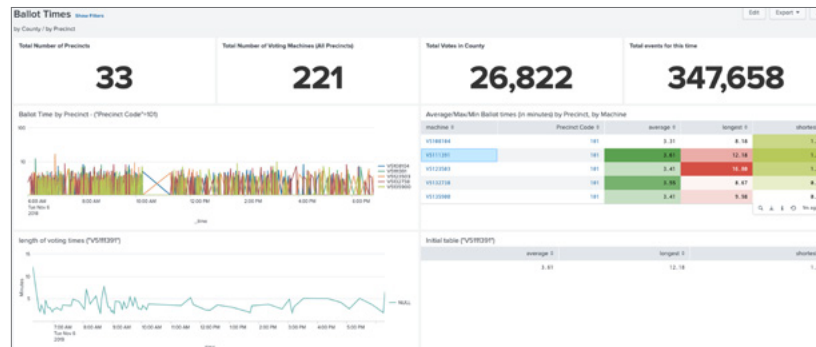


Figure 6

In addition to some countywide statistics, the following figure is a visualization of ballot times for a single precinct in aggregate and in a per-machine view. In this case, we are looking at the length of time each vote took, but we could also look at votes per minute in a precinct to determine busy or light times, which could then inform us if more or less voting machines are needed in this precinct for the next election.

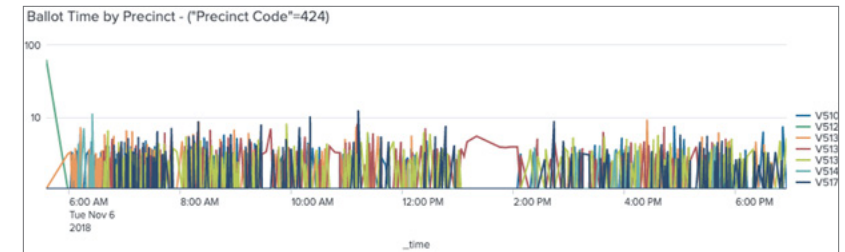


Figure 7

The query to investigate this is:

```
index=polling ("Precinct Code"=424) Code=2401 OR Code=1502
| transaction machine startswith=(Code=2401) endswith=(Code=1502)
| eval dur=(duration/60)
| chart values(dur) over _time by machine
```

Here we look for all of the events in a single precinct ("Precinct Code" = 424) where the event Code is either 2401 or 1502, the events that begin or end an individual voting session. Splunk builds time-based transactions with these events; one artifact of using these transactions is the creation of a Duration field, which we divide by 60 to change our units from seconds to minutes. Our Chart command creates the visualization and plots the calculated values across the day, broken out by the individual ballot machines.

We can discover anomalies by looking at the data in this manner. In the "Ballot Time by Precinct" example above, there is a very obvious space from 10 a.m. until 11 a.m. This is a large concern because it looks like the precinct was closed for an hour in the middle of the election. When other precincts were checked, that same empty block appeared but not necessarily at the same time.

Additionally, polls opened at 7 a.m. and closed at 7 p.m. But the events in the above visualization show differently. In fact, it notes that voting events started at 6 a.m. This could be a critical problem for an elections board certifying an election if the polls were opened too early. And time is the issue here. Here is the relevant event from one of the polling machines:

1656,11/4/20,2057,Set terminal date and/or time,PS250352,424,14:06:28,V5107761

We can see that in the middle of the election, sometime during the day in each precinct, the date/time was reset. Looking closer, it was determined that the voting machines had never been reset from Daylight Savings Time to Standard Time. That gave the appearance in the data that poll workers had opened the terminals for voting too early. In reality, they were open only for the requisite 12 hours.

There were two outcomes here. First, the election officials were able to explain what the discrepancy in the data was so there was no confusion. Secondly, there was another opportunity to refine the poll worker training processes to include checking and, if needed, properly resetting the machine time for accurate reporting.

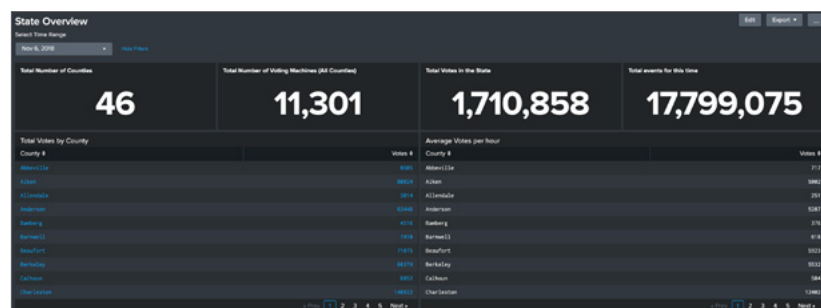


Figure 8

More data leads to better election processes

Using a wider lens, this data can be collected by a precinct and aggregated up to the county level, and then again up to the state level. State auditors can then drill down into specific counties and precincts to validate certifications, or to evaluate resource requests from those counties. If the state were to host a Splunk deployment for all of the counties, it could provide visibility to the counties, have the aggregate views of the entire state, and offer custom views to the election board liaisons whose role it is to interface between the state and a set of assigned counties.

With this in place, the counties and the state can certify their elections more rapidly, more accurately and with a much higher level of confidence than with existing manual methods. All levels of electoral board agencies charged with the administration of the elections process — not to mention campaign finance and lobbying disclosure and compliance — can benefit from the instant insights, as well as from uncovering the various unknown trends and situations that might have never been exposed using traditional solutions. These deeper insights, coupled with faster and more confident election certifications, offer greater peace of mind to the agency and to the voters. Isn't that the outcome we all are aiming for?

Matt Portnoy

Matt Portnoy has over 30 years of technology experience across a variety of disciplines. He is a Splunk staff sales engineer specializing in data management and virtualization and supports public sector and higher education accounts in North and South Carolina.

“The future is already here — it’s just not very evenly distributed.”

– William Gibson

Accelerating Enterprise SOC Incident Triage and Collaboration in VR

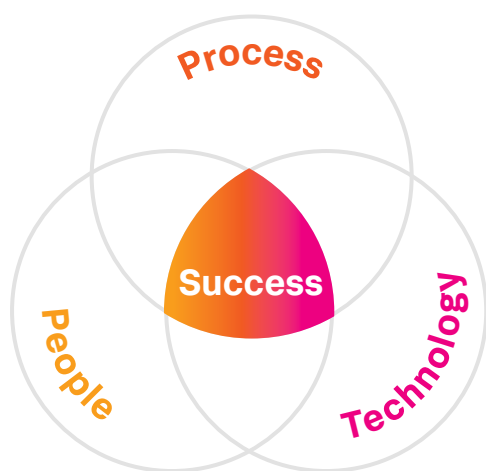
The modern security operations center (aka SOC) is the nerve center of a high-functioning secure enterprise. But for years now, SOC analysts and the analysts who power them have been increasingly strained for time and resources.

With the sudden shift to remote work into 2020, SOC analysts are now facing even greater challenges. Effective communication has become a critical priority for enterprise cyber defense and incident triage. Today and in the future, distributed SOC analysts powered by remote-working analysts must adapt.



Unfortunately, it's not enough merely to return to previous levels of productivity and effectiveness. This change in the way we work has done nothing to slow or stem the tide of attackers who have themselves become increasingly organized, and who continue to exploit the weakest links in enterprise defense at an ever-accelerating pace.

The infosec industry at large has long represented the key to successful modern cyber defense as a Venn diagram of equal parts people, process and technology.



While there is undoubtedly truth in this, many organizations struggle, especially when it comes to making meaningful improvements, to focus on the people and the processes needed in cyber defense to make effective use of technology.

In recent years, attackers have seized upon the people element, taking advantage of the fact that people often lack the necessary knowledge and planning. The rise of weaponized social engineering and deep fakes in place of technical hack attempts like malvertising and trojans shows that attackers have fully embraced and are actively seeking to exploit this global weakness and lack of investment or attention to people in particular. Modern attackers have figured out that they don't need to circumvent sophisticated technology which has been iterated on and refined for more than a decade. Instead, they can strike at the weak links which are today most commonly people and lack of process.

So what does this have to do with virtual reality (VR)? And if we're writing a book about the bleeding edge and near future, shouldn't we be talking about augmented reality instead? The answer is that in order to enhance collaboration and accelerate communication by a significant and meaningful margin, we need to bust a few of these common misconceptions.

Augmented reality today is excellent for enhancing or developing wholly new interactions with things. A lineman's ability to tag power poles with virtual specs and call up metadata about past maintenance, for example, is a fantastic use case for augmented reality. Virtual reality, on the other hand, which fully takes over the user's visual field of view, would be completely impractical and downright dangerous to use in a situation like that. However, while augmented reality excels at attaching data to real-world objects, virtual reality excels at enhancing the interactions of people working against a large, complex and already digital dataset.

Which brings us to the next common misconception: that if virtual reality is involved, it must look like something Morpheus would have created in Excel, with 3D pie charts and nth-dimensional data tables. Based on several years of research into VR usability in the enterprise, the future looks more like the gesture-controlled computers of "Minority Report" than it does "Lawnmower Man" frantically searching for the exit.

Since the invention of the computer screen, people have been trying to make them larger. We even stack and tile multiple monitors today as a standard configuration for most knowledge workers, all in an effort to visualize the work in as large a virtual space as possible. If we can see it all, it's automatically more intuitive to manage and simpler to multitask. VR takes this concept and flips it on its head (pun intended) by placing you the person in the center of your already digital data, thus allowing you an infinitely flexible workspace of any size, depth, orientation or intensity to interact not only with that data, but with other people too.

Because of the convergence of VR and the internet, cyberspace is finally coming into its namesake. Collaboration in VR can take any form imaginable. From a skeuomorphic representation of a conference room complete with virtual whiteboards, virtual terminals and virtual participants, to a highly abstract empty vessel on a mountaintop where data is freed from the 2D confines of a rectangular plane — like an artist with a brush and a blank canvas, the only limits are your imagination.

The business problem

After studying real enterprise users interacting with VR for security incident triage over the past several years, one of our most surprising discoveries has been the realization that 2D data is actually really great. It makes sense to us, and it's practical and efficient. If you're looking at an IP address, nothing is added to that experience by making the font 3D and thinking along those lines could lead us to overlook the true untapped potential of eliminating the screen. Superfluous augmentation like adding a depth dimension to fonts or charts might be good for the movies, but doesn't make a productive impact when it comes to work.

Instead, think about what it would mean if that same IP address could be freed from the page altogether. With a gesture or a flick of your finger, you could literally lift an IP off the page and then, using security orchestration, automation and response (SOAR) technologies like Splunk Phantom in the background, instantly run a number of reputational checks, geolocation, user and entity behavioral analytics (UEBA) attribution, and more, so that by the time you set it on the virtual whiteboard, all of the enrichment and metadata is right there along with it.

Now, imagine doing this with an audience of observers and collaborators and you'll start to see a glimmer of the potential of this new paradigm to accelerate the way we work in very practical and measurable terms. Playbook and runbook creation would no longer need to be relegated to hierarchical, Vizio-style logical diagrams. A system powered by machine learning could build playbooks automatically based on analysts performing those parts of their daily work in VR that are well defined and ripe for total automation. Playbooks don't need to be static snapshots in a binder. Instead, they could take on a life of their own, constantly fitting and adapting parameters and actions according to what they learn from their human analyst trainers.

Realizing such a vision requires many unique advances in domains spanning the full spectrum of people, process and technology, some of which have already been fully realized, and others that soon will be. For example, people need VR headsets that are comfortable and portable, with reasonable battery life and high-enough resolution to make reading text comfortable. You need that headset to support software that's compatible out of the box with a large variety of existing technologies. You need processes which are so well-defined as to be nearly to the point of automation.

And you need a platform capable of the proper levels of abstraction and open-ended data exploration these use cases will depend on, like Splunk, the Data-to-Everything Platform. In the rest of this chapter we will explore the details of a solution and method of implementation that satisfies all of these requirements. Not only do all of these capabilities already exist, now they're also affordable and mainstream enough that enterprise adoption is possible and even practical. As the prescient analysts at Forrester predicted in their [report](#): "The future SOC analyst will use VR and gesture control to analyze an event with 3D links in virtual reality while commanding an intelligent assistant to capture forensic data from a host — simultaneously. For the first time, S&R pros will be able to analyze and take action in near time. With faster analysis and decision making, combined with automation and orchestration, CISOs should set expectations that security operations must move faster."

For our SOC-specific example, we'll use Splunk Enterprise Security (ES) and Splunk User Behavior Analytics (UBA) to triage a security incident with remote collaborators playing the roles of SOC and forensic analysts. This combination will allow us to take full advantage of the strengths of human-driven investigation and analytics as well as machine learning to automate threat detection and response. Splunk's notable event framework will help us identify the security incident of concern and reduce the human effort involved in examining vast amounts of event data by automatically filtering, tagging and sorting it for us so that our people can focus their valuable time and attention on a security incident of real concern.

We'll also need the appropriate VR hardware, of course, as well as software designed specifically for VR collaboration, telepresence and virtual desktop functionality.

How it works in practice

Hardware

One of the most common criticisms of VR in enterprise work is that the headsets can become uncomfortable when in use for an extended period of time. Though they have come a very long way in recent years and will continue to improve, when used strategically for key communication events like incident triage and follow-the-sun shift changes, it is not necessary for users to wear them for extended periods at all.

For our use case demonstration, we have opted to use the recently launched Oculus Quest 2, currently the highest resolution, affordable VR headset available on the mass market. The headset will be interfaced with a modest-spec PC running Windows 10 on an Intel Core i5 processor with 16GB RAM and a mid-range ATI Radeon RX570 video card. The total retail cost of all hardware used is less than \$1,000. VR is more affordable than ever, and it doesn't need to involve a high-end gaming PC, dozens of feet of wires, external sensors or any of the other expensive, specialized hardware requirements just a year or two ago. Here is the Oculus Quest 2 and its included controllers:



VR software

There are a handful of software offerings available today that provide collaboration, telepresence and virtual desktop functionality. These options offer a quick way to get started and test the VR collaboration waters without having to roll out your own or change anything about the way you access tools on a traditional desktop. In fact, many of these software offerings offer hybrid clients for those who may not be in VR at all. One such offering is the software we'll be using to drive our virtual meeting space in these examples, [vSpatial](#).

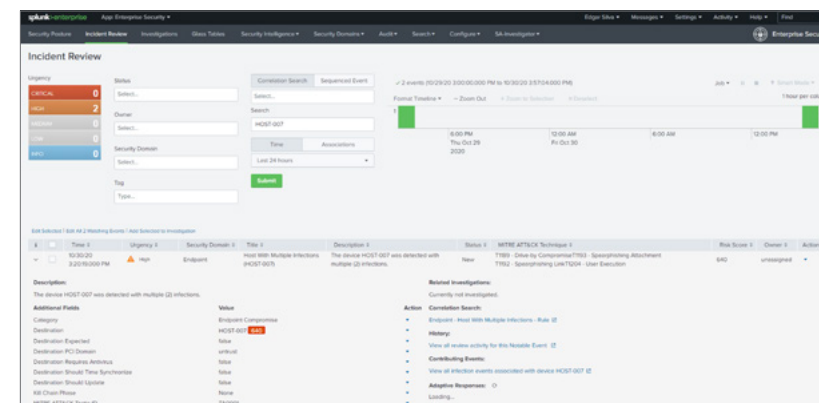
There is a free version as well as paid versions that support larger meeting spaces and more simultaneous attendees. Similar platforms include The Wild, Spatial, Facebook Horizon (formerly Facebook Spaces), hubs and MeetinVR, with many more in development and on the way. We chose vSpatial because of the diverse platform support available (Oculus Rift/ Rift S, Oculus Quest/Quest 2, Oculus Go, SteamVR, Valve Index, HTC Vive, Windows MR, 2D on Mac, as well as 2D on PC). It also supports hybrid 2D as well as VR users in the same virtual meeting space, giving us maximum flexibility, depending on what each attendee requires and how they prefer to interact.

Now that we have the foundation of a virtual meeting and collaboration environment in place, let's take a look at what we're going to do with it and how it all comes together.

As we walk through the example, try to think about how you perform these tasks today, and how much time it takes. How many different people do you need to communicate with and bring up to speed? If you have a particularly large 24/7 operation, how do you communicate about the issues that are still open at the time of a shift change and hand-off? How many different systems of record and screens and tools and Slack messages does a typical investigation take? And does everyone working understand all of the processes and technology well enough to be equally effective on their own?

The incident

As you can see from the screen capture below, we have experienced a possible breach. Thanks to Splunk UBA's ability to automatically map security incidents to specific users, we also know right away who is impacted. Fortunately, the individual incident appears to be isolated to this one user, but we can't be sure yet how this user was compromised, or what that might mean for the risk to the rest of the company.



In order to understand those things and fully resolve this incident we need to communicate. This is the point where technology ROI is greatly accelerated by having made similar investments in people and processes. Today, we communicate via Slack or Zoom, and before COVID-19, it might even have occurred in a conference room or the SOC itself. Regardless, communication is crucial as we introduce the problem and timeline of events as currently understood to people who are hearing about this incident for the first time. They could be fellow analysts, escalation engineers, colleagues in IT, forensic analysts, business stakeholders, SOC management, risk management, etc. Each person will have a different set of questions, level of expertise and level of expected detail. No wonder incidents like this in a large organization often take so much time to fully triage!

Start your VR engines

Everyone who has ever misread tone or been misunderstood via email or text (which is everyone) knows the value of body language and face-to-face communication. The current generation of virtual collaboration options, including vSpatial, solves for this by showing all participants as avatars whose virtual head and hand movements are synchronized with the actual user's movements. What's more, when you speak in a virtual meeting, your avatar's mouth moves in sync with yours, and sometimes with facial expressions as well. Being able to gesture, nod and see while communicating with a group of people is extremely valuable in terms of reading the nuances and subtext of the conversation.

Analysts have gathered in vSpatial to discuss this security incident and identify, via a forensic investigation of the endpoint, how the incident occurred. They are not all in VR! Only one user (Jason) has a VR headset (our Oculus Quest 2); the other two users are on traditional screens with mice and keyboards. Still, they are able to view and interact with Jason with almost as much detail as he can with them.

Jason is the analyst who originally identified the incident and has worked it solo so far. Now he's bringing it to his forensics counterparts who will assist with the investigation of the endpoints. Jason is using Splunk ES and UBA, so he is sharing those views in a collaborative virtual space with the other two users, Monty and Jasmine, who are not in VR. Monty and Jasmine have their own tools and processes for performing forensic investigations. As the investigation proceeds, all three are able to collaborate in this virtual war room in real time — sharing data, gesturing, nodding and brainstorming potential explanations and solutions in a way that is far more engaging than Zoom or Slack could ever be.

An investigation among remote colleagues that would have taken countless back-and-forths in email or multiple phone calls is now happening as if everyone were in the same room. Even better, because this room is digital, the entire thing is a whiteboard with a suite of augmentations specifically tailored to support the team and mission.

As Monty and Jasmine zero in on the forensic evidence, they are able to explain to Jason how the end user was fooled into opening a malicious attachment in email which led to the malware infection.

Now that this incident has been classified as a successful phish, additional processes kick in. For starters, the entire organization's mailboxes need to be scanned for similar emails. If found as still unread in any other user's mailbox, they need to be deleted. If found as opened, those endpoints need to be scrutinized.

Fortunately, Jason has access to Splunk Phantom, which has been configured to execute these playbooks in just such an incident. In fact, Phantom has already run those playbooks, and the results are at Jason's fingertips. Within moments, Jason is able to tell Monty and Jasmine that the message was found in five other user's mailboxes unopened and so it was automatically deleted — the rest of the organization is unaffected and safe. All three can immediately see one another's virtual screens and now, having reviewed all the data at hand and resolved the issue, they can close the meeting and return to their daily work, whether inside or outside of VR.

The next experience

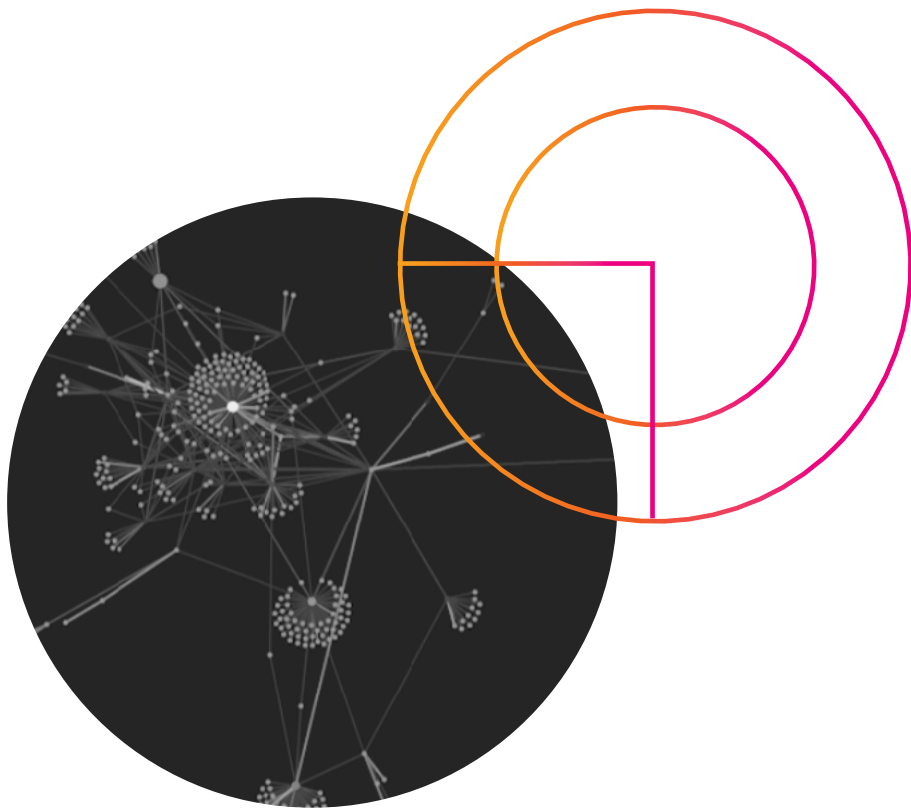
Given the density and complexity of information VR can provide and the speed at which it can be delivered en masse, the benefits of VR in enterprise security work already far outweigh any drawbacks, even in the current generation of the technology. As integration between enterprise software tools and virtual meeting spaces continues to evolve and mature, we'll find ourselves spending more and more time in both virtual as well as mixed or augmented reality. Like punch card readers before them, the screens that many of us have been stuck behind for so long will finally give way to an entirely new computing paradigm which will come to define the next great leap of progress and productivity.

Not only does this evolution have massive ramifications on human-to-computer interaction, we are poised to see the same revolution play out in human-to-human interaction as well. Especially with so many people now physically isolated from each other at work, we are forced to explore new ways to bridge gaps in communication and understanding that none of us could have ever expected.

For information security in particular, it isn't enough to get back to where we were pre-COVID-19. We need to adopt this accelerated and more information-dense ability to interface not only with our technology, but also with our people and processes, so that we can protect our organizations from cyber threats. Hopefully, future generations can say about us what [William Gibson said](#) when he tried VR for the first time: "They did it!"

Jason Landers

Jason Landers is a solutions engineer at Splunk. He has spent a number of years working in technologies that combine virtual reality and the security operations center. He has presented a VR collaboration solution for security at multiple conferences and events.



Detecting and Preventing Financial Crimes With Graph Algorithms

In June of 2020, the German payments company Wirecard collapsed as a result of sophisticated global fraud. Wirecard managers had just continued blindly with fraudulent activities that resulted in massive collateral damage for shareholders, business partners, individuals, employees and their families.

“Wirecard collapsed on Thursday owing creditors almost \$4 billion after disclosing a gaping hole in its books that its auditor EY said was the result of a sophisticated global fraud.”

– Reuters

Though financial services seems to be one of the most regulated industries globally, the Wirecard scandal may only be the tip of the iceberg. Financial crime and fraud is an increasingly pressing topic in the financial services industries. According to the [ACFE Report to the Nations](#), the largest global study on occupational fraud, the damage and global cost of fraud reached \$3.6 billion in 2020. Given the scope and severity of the problem, it's clear there is no silver bullet, despite the variety of approaches and solutions available.

This chapter does not promise to be a silver bullet, but it will highlight a novel approach that can help improve existing fraud detection approaches with Splunk. Not only is it effective, it's also a smart option for Splunk customers who want to complement and enrich existing fraud solutions with new methods and techniques, or start building from scratch. Using concrete examples, this chapter will help you replicate these detection patterns and extend your existing Splunk Enterprise use cases while leveraging existing and publicly-available Splunk components and apps.

The cost of not preventing financial crimes

It's no secret that Splunk Enterprise enables companies to ingest a huge variety of unstructured and structured data that can be searched and analyzed to tackle various business problems. With such robust data collection, the question becomes: how best to analyze all of that data in a meaningful way that can help solve business problems?

When it comes to fraud detection, that data is gold. And graph algorithms that go beyond basic statistical measures to reveal hidden structures in interconnected datasets can reveal traces of malicious and fraudulent activities in a multitude of systems and services. Traditional Splunk searches can easily be transformed into rules that help detect aspects of fraud, and those results can be further aggregated into risk-score-based metrics to help prioritize investigations. This is a valid and useful approach. However, the wealth of heterogeneous data ingested in Splunk provides a unique foundation for a much more sophisticated approach to fraud detection.

There are algorithms and techniques that allow us to add more intelligence and make existing data sources even more valuable. The "secret sauce" here is to focus explicitly on the relationships between entities captured by data. From a network of interconnected entities, the structure and topology that approach reveals allows you to build a different type of dataset derived from the raw data.

If this sounds a bit too abstract, just think of a social network in which people are connected to each other by relationships and areas of interest. From the social network structure, you can derive insight into who the most popular influencers are and which communities and areas of interest overlap. You can even predict how likely you are to connect with another person in the network because of mutual friends and shared interests. That's why applying graph algorithms to such datasets can make up the essential ingredient for a use case of tomorrow: graph-powered models that contribute additional intelligence to existing Splunk-based fraud solutions.

Unlocking Splunk with algorithms to fight financial crimes

This solution uses Splunk Enterprise or Splunk Cloud and a set of apps that can be downloaded for free on [Splunkbase](#). These apps will allow you to use graph algorithms for the more advanced fraud use cases we will discuss later. Here is the full list of apps that work together seamlessly for the listed compatible version numbers:

- [Splunk Enterprise](#) or [Splunk Cloud](#)
- [Splunk Machine Learning Toolkit 5.2](#)
- [Python for Scientific Computing 2.0](#)
- [3D Graph Network Topology App for Splunk 1.2](#)
- [Deep Learning Toolkit 3.3](#) (optional, for Splunk Enterprise only)

Once installed, you can find some graph algorithm examples in the 3D Graph Network Topology App for Splunk. The app also contains a Graph Analysis Framework which allows you to easily apply different graph algorithms directly to your data in Splunk, giving you access to the following graph algorithms you can use right away:

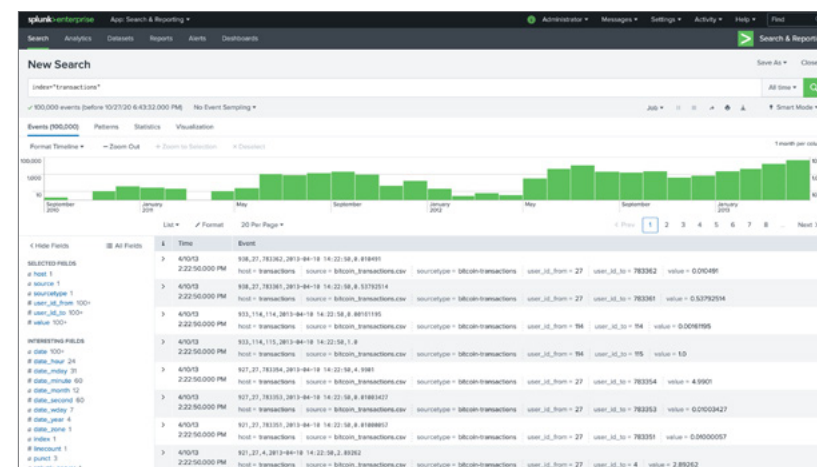
- Graph Centrality Measures
 - Degree Centrality
 - Betweenness Centrality
 - Eigenvector Centrality
- Clustering Coefficient
- Connected Components
- Label Propagation
- Minimum Spanning Tree
- Louvain Modularity (optional, in Deep Learning Toolkit only)

Implementation details

In this section, we'll focus on two specific examples that provide all the necessary elements you'll need to add these new tools to your fraud investigation approach. But first, let's have a look at an important preprocessing step that you need to take before you can analyze any graphs — using SPL to retrieve the relationships from your raw data.

Data preprocessing for graph analysis tasks

Let's assume you already have data in Splunk that you want to analyze with [graph algorithms](#). The first thing you need to do is to define which data sources you want to connect and the fields you want to use to do it. Typically you define and extract the fields of interest from raw log data, or have them automatically extracted if it's a known source type. Let's assume you have a datasource that contains transaction records of an amount (value) transferred between two entities (user_id_from, user_id_to) at a given time (_time):



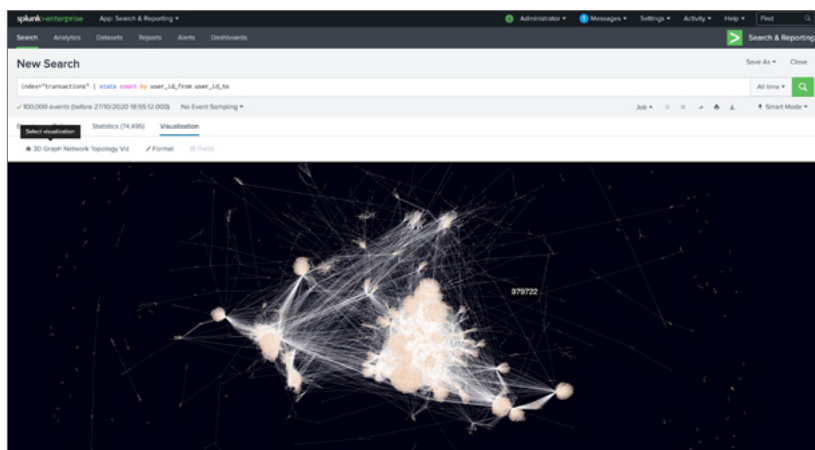
From this point, there are many ways you can use **SPL** to further analyze your data. In our case, we want to prepare the data so that it can be used by graph algorithms. One way to construct a graph from our data is to retrieve a so-called edge list which contains all the connections of interest. In SPL, there is one simple search pattern that you can leverage for this task which aggregates, for example, the count of transactions between entities in **the selected timeframe**:

... | stats count by user_id from user_id to

When we apply this to our search, we can inspect the calculated results in the statistics tab:

user_id_from	user_id_to	count
453676	11	1284
3958	74	478
25	25	389
1714687	37261	216
1533889	74	286
26	25	181
453676	13	178

Now we can switch from the Statistics tab to the Visualization tab and display the dataset as a graph with the help of the 3D Graph Network Topology visualization:

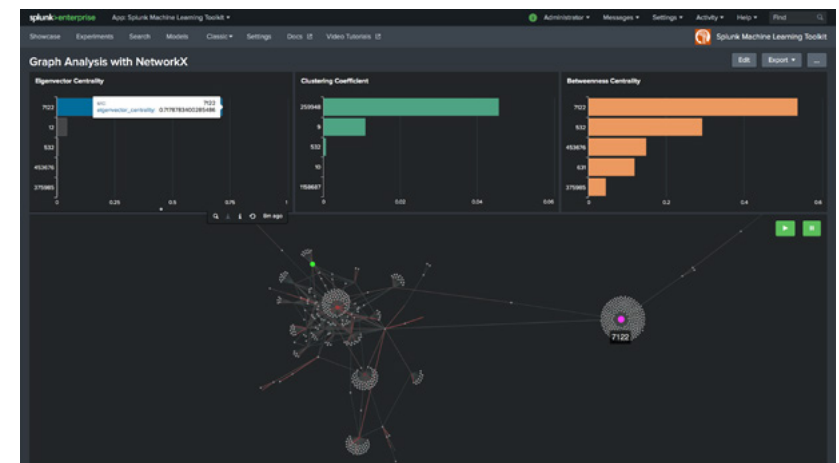


As you can see, this is a fairly complex graph structure, hard to read and interpret visually, which is why we want to use graph algorithms to retrieve only those entities or structures of interest.

Example 1: Identify suspicious actors with graph centrality measures

Fraudulent actors can take on many behavioral types and follow various characteristics in order to hide suspicious behavior. One such behavioral pattern can be described in terms of how important an actor is within the network. A graph can reveal this information with the help of centrality measures. The PageRank algorithm became popular to solve a similar problem for internet search to show the best matching and, in many cases, the most popular sites. That same underlying concept can be used for our purposes as well.

We can compute the eigenvector centrality of every entity in the graph and analyze which actors are most influential. This helps us find and prioritize the most important entities in this network of transactions quickly. Another interesting measure is the betweenness centrality, which takes into account how central an entity is in relation to all other transactions flowing through it as an intermediary node. Think of this as “middle man” behavior that can be specifically measured and identified. Here are the results combined in a Splunk dashboard:



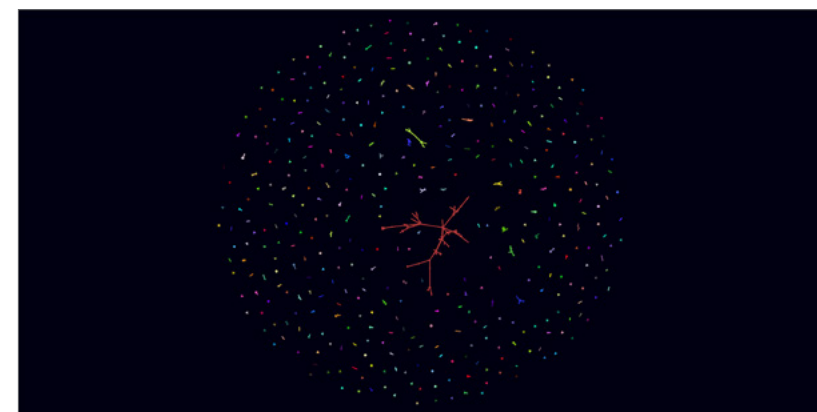
This shows a subset of bitcoin transactions and highlights its top five nodes, which have either a high eigenvector centrality, betweenness centrality or clustering coefficient. Clearly, the pink highlighted node 7122 stands out. It shows a high eigenvector centrality, connects the graph structure on the left with another structure off-screen on the right, and also leads to the highest betweenness centrality. For fraud analysts, these can be interesting finds that reveal important patterns in a large dataset and lead to insights into areas for further investigation.

The computed measures can also contribute to existing risk-based profiling approaches. You can read the centrality measure as a score that contributes to the overall risk score of an actor. Not only that, but these centrality measures can be additional features in subsequent machine learning models that incorporate information retrieved from topological aspects of the graph.

Example 2: Identify suspicious groups of actors

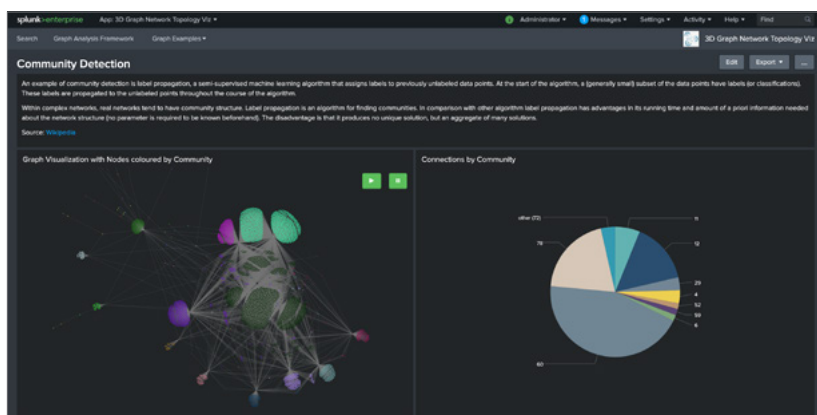
Graphs can reveal other fraudulent behaviors and actors through their connections with each other. In financial transactions, this approach reasonably assumes that it's unlikely that individuals act as both middlemen and as brokers between many other actors. The betweenness centrality described earlier can be helpful in measuring this. If we were to identify an individual actor, the next step would be to find out who is connected to them in a given time frame and how connected they all are to each other. Such a pattern is often described as a "fraud ring," a structure within a network of transactions that connects individuals who collaborate for a fraudulent purpose. For example, a fraud ring might move money through multiple actors to an agreed-upon destination, or exchange other values among each other.

In a sparse graph, such a structure can easily be revealed with the connected component algorithm. The algorithm assigns each group of connected entities a number which can be used as a label to separate the groups. The following graph visualizes the findings of the results of a connected component algorithm being applied on a real-world dataset of financial transactions (displayed anonymously here).



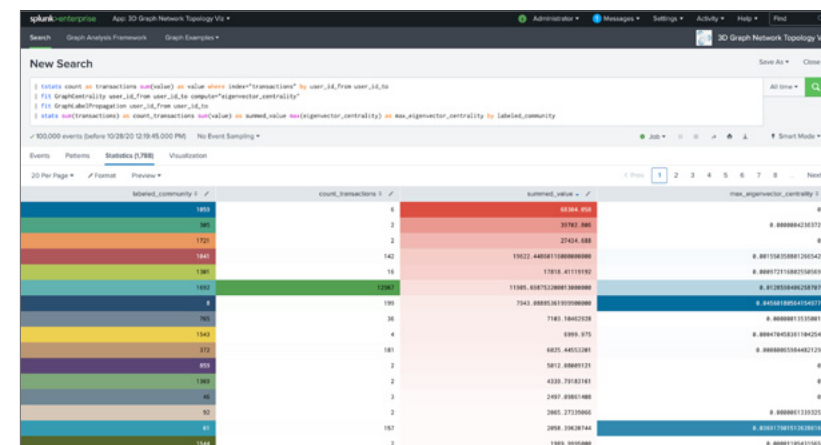
Each color represents a connected component of monetary transactions between individuals. The bigger red component in the center could indicate potential fraud rings and serve as an entry point for further investigations. The graph can easily be combined with drill down functions on a Splunk dashboard to show other relevant data for the selected entities. And the group labels of the connected components can be used for further statistical aggregations, such as the sum of money transferred within the group and the number of transactions. This can help us integrate further specific business constraints and retrieve other meaningful statistics to add to existing risk measures or models. We can also combine multiple graph algorithm results with each other to glean further insights. For example, we could incorporate centrality measures from the first example above and derive statistics on the connected components by using the SPL stats command for further aggregations.

While the connected components algorithm is helpful for sparse graphs, in cases where all entities are connected with each other, there is only one connected component identified, rendering the whole approach useless. In such cases, [Label Propagation or the Louvain modularity method](#) are far more useful. The label propagation algorithm is a semi-supervised machine learning algorithm that assigns labels to previously unlabeled data points. It propagates initial labels on a subset of the data through the graph. Let's have a look at how this algorithm works on a subset of the bitcoin transactions dataset:



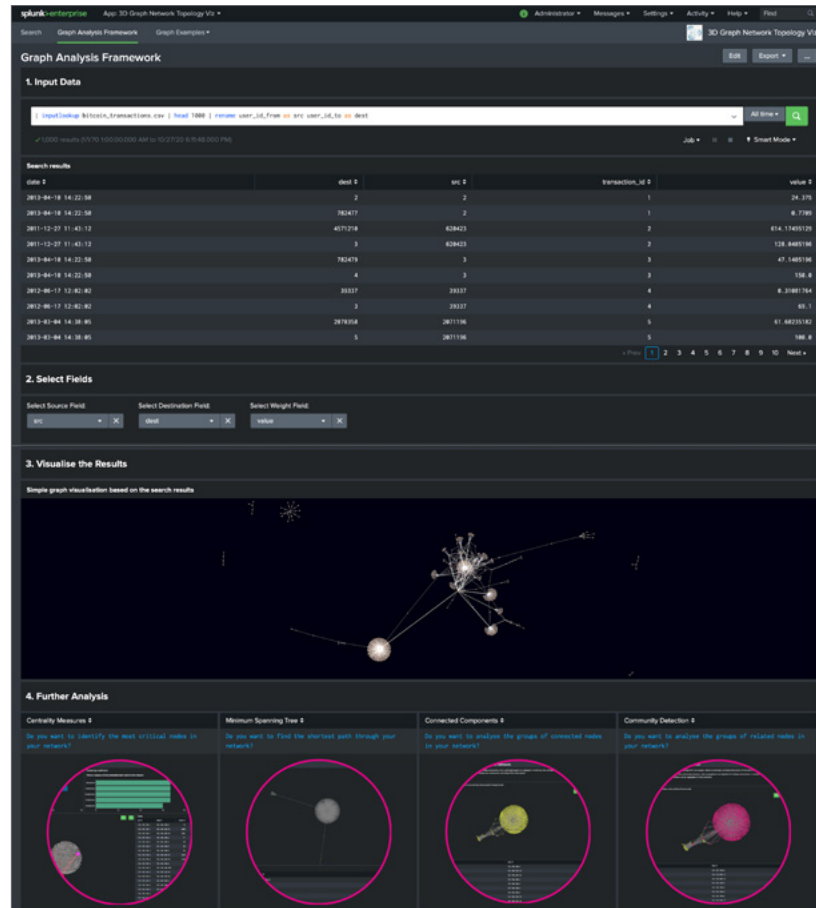
Here we see how parts of the graph are labeled as communities with the same color, indicating groups which are close with each other based on their connectedness. In contrast to the connected components algorithm, this algorithm provides us with a different perspective that is revealed in the underlying graph structure.

Again, we can use this approach to identify structures of interest in a graph that should be investigated further. For example, finding groups of individuals who interact closely within specific business constraints may indicate potentially suspicious behavior worthy of deeper investigation. Let's conclude with an analysis example that combines both methods as presented above in a simple, elegant SPL statement:



The results of this analysis allows us to easily pivot across the dimensions of the total number of transactions, the summed transferred value and the maximum eigenvector centrality by community. In the example above we sorted the summed value in descending order and we can read that the first five communities have quite high sums transferred with only a few transactions. This might be valid high volume transactions, but the sixth and seventh row with community labels 1692 and 8 or 61 at the bottom show additional high eigenvector centrality and/or transaction counts. These can be quickly identified in the table and provide direct entry points for further investigations.

We have discussed two methods for using graph algorithms to retrieve new measurements from interconnected data points. This methodology not only works for very specific datasets but it can also be extended and applied to other connected datasets.



By applying the concepts and algorithms described above, you can easily identify fraud-related data from any data retrieved from a search in Splunk. The following screen shows the example dataset of bitcoin transactions within the Graph Analysis framework that you can use to quickly derive meaningful graph-based features and new and useful investigative approaches.

Philipp Drieger

Philipp Drieger works as a principal machine learning architect at Splunk. Over the past six years, he has accompanied Splunk customers and partners across various industries in their digital journeys, helping to achieve advanced analytics use cases in cybersecurity, IT operations, IoT and business analytics.



Real-Time Social Media Bot Moderation Solutions That Could Save Democracy

“If it is on the internet, it must be true and you can’t question it.”

– George Washington

Due to the rapid growth of social media use in the last decade, the world is now connected in a way unlike any other in history. We can share information and ideas across geographic and social borders, and a single post can raise widespread awareness for a particular cause or even, in the case of the #MeToo and #BlackLivesMatter movements, inspire real social change. Though social media has certainly had positive impacts, the unfortunate flip side is that malicious misinformation can travel just as quickly as the truth — if not faster.



While the term “fake news” has in recent years become part of our everyday lexicon in a variety of different contexts, and often reduced to a punchline, it’s a very serious matter. **Studies show** that on Twitter, false news stories are 70% more likely to be retweeted, and they reach people six times faster than true ones. These statistics are particularly alarming given that social media is becoming **the most common way people consume news online**.

The spread of misinformation can cause serious health risks, such as the spread of falsified data supporting **unproven protection strategies against COVID-19**, or the promotion of **scientifically-debunked cancer remedies**.

Another grave danger posed by the spread of misinformation on social media, one with perhaps an even more devastating impact, is the threat it poses to democracy around the globe. Authoritarian governments are using social media for propaganda and societal control, and others are using it to manipulate public opinion, sow division and influence **elections**. As our social media feeds become the frontlines of the 21st-century political battleground, real-time online content moderation could be crucial to retaining the sanctity of truth, and the health of democracy.

It’s a use case of tomorrow, but we need it today.

The cost of doing business with bots

To understand the solution to preventing malicious posting of misinformation on social media, we must first understand how it’s spread. Computational propaganda, as it’s known, is **defined** by those who coined the term at the **Oxford Internet Institute** as “the use of algorithms, automation and human curation to purposefully distribute misleading information over social media networks.” While there are undoubtedly malicious actors manually disseminating false information, the vast majority of content is spread through bots, automated accounts that attempt to emulate human behavior in order to achieve their goals of widespread manipulation and social tampering.

Bots are now the main tool for **online smear campaigns** and as many as 48 million — 15% of all Twitter accounts — **are believed to be bots**. Though not all automated accounts are malicious, identifying those that are and understanding how they spread misinformation is vital.

Bots are built using machine learning (ML), a technology that has grown massively in recent years and will only continue to do so in the future because of its applicability to big data. As the volume and accessibility of data produced globally skyrockets, algorithms have larger, more detailed datasets to generate models. In the case of computational propaganda bots, more social media posts from a higher number of users means the algorithms training the bots have a higher number of user posts to study and learn from before attempting to replicate with their own posts.

The social media giants are aware of the threat posed by bots, but have struggled to solve it. Despite their efforts at manual moderation using huge teams of fact-checkers, there’s such a large virtual world for bots to hide in. Facebook alone has 2.6 billion users, which makes getting ahead of the problem like trying to find a needle in a field full of haystacks. In the real-time universe of social media, even if moderators can identify a post as originating from a misinformation bot, it’s a race against the clock.

Silencing the bots

In a world where machine learning algorithms have enabled the creators of these bots to mimic human behaviors, infiltrate the online discourse and distribute automated propaganda in real time and at a huge scale, surely the most scalable and effective solution is to leverage these technologies and utilize big data pipelines in conjunction with machine learning algorithms and automated action workflows to detect, classify and remove social media posts and accounts which promote computational propaganda.

Splunk's position as an AI-infused Data-to-Everything Platform means it's uniquely positioned to act as the real-time, end-to-end monitoring center, data science modeling environment and automated response orchestration center for any business faced with moderating a social platform.

Using Twitter as a use case, our solution leverages Splunk's capabilities to ingest social media data in real time through Representational State Transfer (REST) read input of Twitter's filter API. It also uses a separately formatted static dataset of bot tweets/accounts that Twitter discovered and **published** as part of their ongoing election integrity campaign against bot-farms, such as the Russian Internet Research Agency.

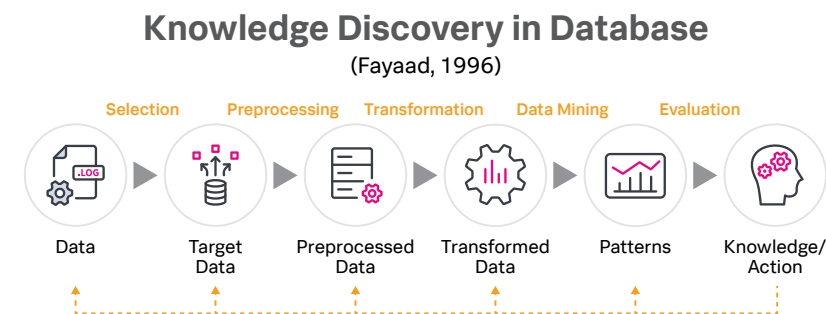
Splunk provides us with the framework for the entire iterative data science pipeline: data can be cleaned, munged and extended with contextual lookups to enable the two datasets to be compared. With **Splunk Machine Learning Toolkit** (MLTK) we can select features and build, test and improve models that can detect bots using an open-source ecosystem of classification algorithms.

We can then operationalize these models and apply them to real-time data traveling through the ingestion pipeline from the social media platform. When a bot is detected, it triggers an automated response either to a ticketing system for manual moderation or, via Splunk's two-way REST API, to the platform itself, to quarantine the account and remove its posts.

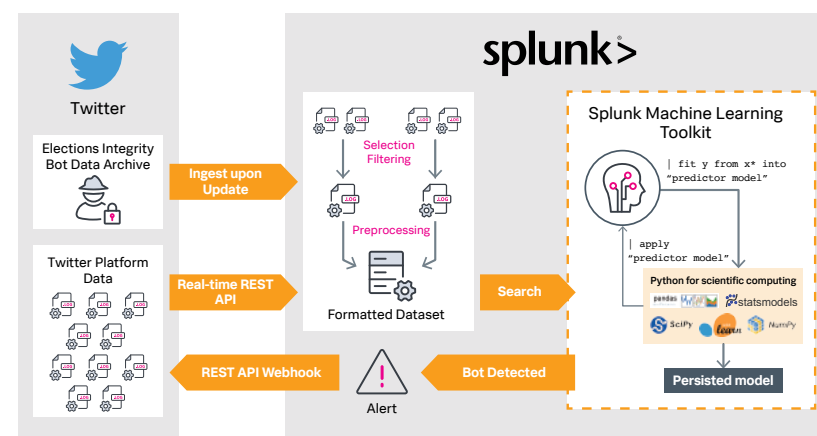
With this solution online platforms can protect themselves against the growing threat of automated misinformation and locate, identify and remove malicious bots before they're able to influence the social network.

How Splunk can detect bots

Before diving into the technical intricacies of implementing the social media bot detection system with Splunk, we'll outline the broader methodology this method follows, Fayaad's Knowledge Discovery in Databases. Broken down into five sections, this well-known iterative framework is a blueprint for solving data-mining problems, from knowledge discovery and data procurement to analysis and interpretation.



Deconstructing the solution in such a way both enables logical step-by-step implementation and showcases how Splunk uniquely serves as an end-to-end platform for extracting and applying knowledge during the data-mining process.



Selection

The first stage is selecting, and therefore ingesting, a relevant target dataset with which to generate the bot detection models. As we're using Twitter as a use case, the solution would rely on real-time data ingestion from Twitter itself.

Unlike other major social media sites, the vast majority of users on Twitter have no security blocking their profile content, so their posts or tweets can be viewed and retrieved by Twitter's real-time search feature, which developers can embed in their applications through a number of APIs.

Depending on how we want to apply the bot detection system, it may be more prudent to [utilize the sample API](#) (which provides a sample of all live tweets), or [the filter API](#) (which provides a sample of all tweets filtered on provided keyword(s), of which it allows up to 400 tweets and up to 5,000 user ids). Either way, the process to access and ingest the data is essentially the same.

First, you request a Twitter developer account through their portal at developer.twitter.com. This requires that you link a developer account with a general Twitter account, which may take several days to be granted. Next, create an application through the developer portal that includes a description of your reason for use and any necessary URLs you'll be using. This provides the necessary access keys and tokens to make the requests within Splunk.

The easiest way to ingest data via REST APIs is through the [Splunk Add-On Builder](#), an application that helps you construct add-ons that take inputs via REST API, Python script or shell command. You can download the app and install it through the UI using the "Manage Apps" and "Browse More Apps" screens, or by downloading it from Splunkbase and installing from file.

The app gives you the option to create a new add-on before configuring a new REST API data input by entering its parameters. On the "Inputs & Parameters screen," you can set the source type and desired display and input names before determining the frequency to call the API input. This effectively enables real-time data flow into Splunk. And this is where the REST call itself is defined with the URL of the desired API — whether it be the sample or filter entered along with the chosen parameters and headers. The example below uses the sample API, so all that's required is the authorization header with a value of "Bearer," followed by the bearer token found on the keys and tokens screen of your created application on the Twitter developer portal.

Create Data Input

Input Method | **Inputs & Parameters** | Define & Test

Data Input Properties | **Data Input Parameters** | Add-on Setup Parameters

Define the properties of your data input. [Learn More](#)

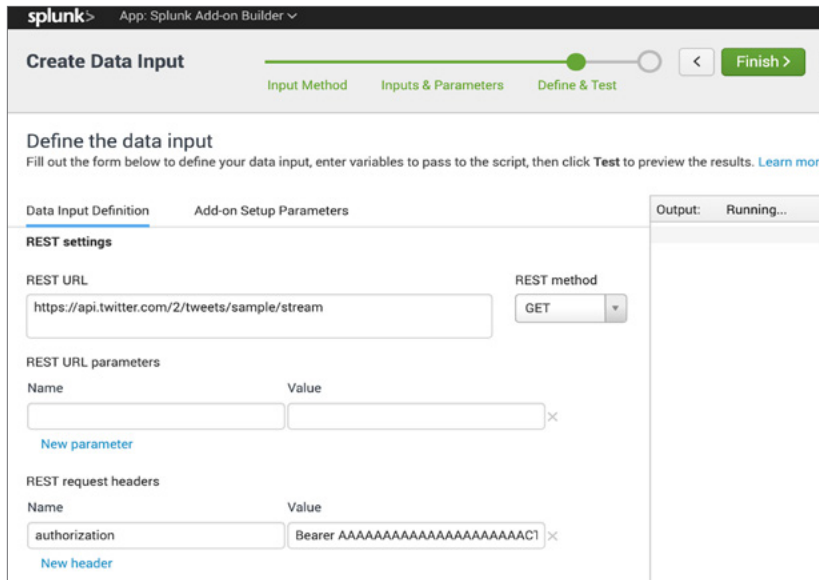
*Source type name:

*Input display name:

*Input name:

Description:

*Collection interval: seconds



Once this is completed, the final step is to enter the newly created add-on from the home screen of your Splunk environment and configure an input, determining which index you want the data to sit in, and how often the input will poll the collected data. Once entered, JSON format social media data containing both post content and contextual information from the posting account can now travel into Splunk in real time, enabling us to gain insight and take action.

In order to then assess whether a post has been generated by a malicious bot, we need examples of data containing confirmed bot posts. Twitter publishes such data as part of its [Election Integrity initiative](#) in CSV format, so ingesting this data is as easy as a one-time pass through the Add Data > Upload workflow and selecting the CSV sourcetype.

Pre-processing/transforming

If you've ever conducted a data science investigation, you know that 60% of the work is cleaning and organizing the data after collection. Not surprisingly, data scientists [report](#) this is the task they enjoy least. Fortunately, Splunk simplifies and expedites this stage of the investigation by providing greater control over data aggregation and manipulation than traditional modeling environments.

Start by searching across the datasets and identifying commonalities in the content contained in the fields. While the names of the fields may be different, they both contain the body texts of the tweets as well as metadata on the posting accounts. This efficiently eliminates information not covered by or engineerable from one of the datasets and allows you to downsize the dataset into lookups with the outputlookup command.

You can also engineer new fields using calculated fields — fields added to events at search time that perform calculations with the values of two or more fields already present in those events. This allows you to analyze components drilled from a single field, such as the number of hashtags within the body text and parameters drilled from the relationship between multiple fields (e.g., the ratio of account followers to accounts followed).

Name	Field aliases	Owner	App	Sharing	Status	Actions
csv(tweet) : FIELDALIAS-followers	follower_count AS followers	admin	search	Global / Permissions	Enabled	Clone Move Delete
csv(tweet) : FIELDALIAS-following	following_count AS following	admin	search	Global / Permissions	Enabled	Clone Move Delete
tweets : FIELDALIAS-bio	"user.description" AS bio	admin	search	Global / Permissions	Enabled	Clone Move Delete
tweets : FIELDALIAS-followers	"user.followers_count" AS followers	admin	search	Global / Permissions	Enabled	Clone Move Delete
tweets : FIELDALIAS-following	"user.friends_count" AS following	admin	search	Global / Permissions	Enabled	Clone Move Delete
tweets : FIELDALIAS-number_of_usermentions	"entities.user_mentions[id] AS number_of_usermentions	admin	search	Global / Permissions	Enabled	Clone Move Delete
tweets : FIELDALIAS-retweet_time	"retweeted_status.created_at" AS retweet_time	admin	search	Global / Permissions	Enabled	Clone Move Delete
tweets : FIELDALIAS-sweet_language	lang AS sweet_language	admin	search	Global / Permissions	Enabled	Clone Move Delete

```

csv(tweet) EVAL- hashtag_number case(hashtags="[]", 0, (mcount(split(hashtags, " "))) / (mcount(split(hashtags, " "))) * 100
csv(tweet) EVAL-h_reply h_reply case(is_in_reply_to_userid="0", "true", isnull(is_in_reply_to_userid), "false")
  
```

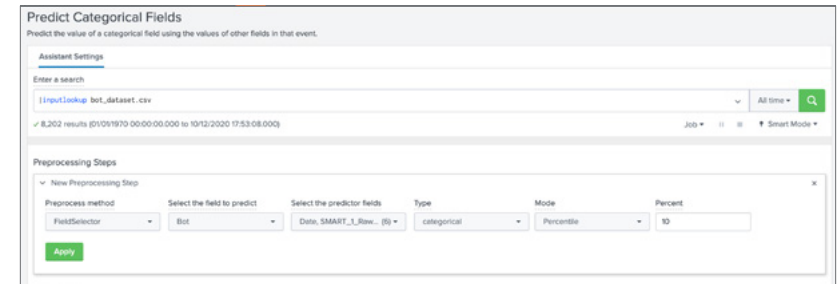
Having selected the desired fields for a given iteration of data modeling, your final task is to transform the two datasets into one cohesive, universally formatted training set. You can do this by using field aliases/ tags to generate permanently referable standardized names for the fields, and/or by renaming fields with the rename SPL command and outputting the results to an updated lookup covering both datasets.

Mining for data

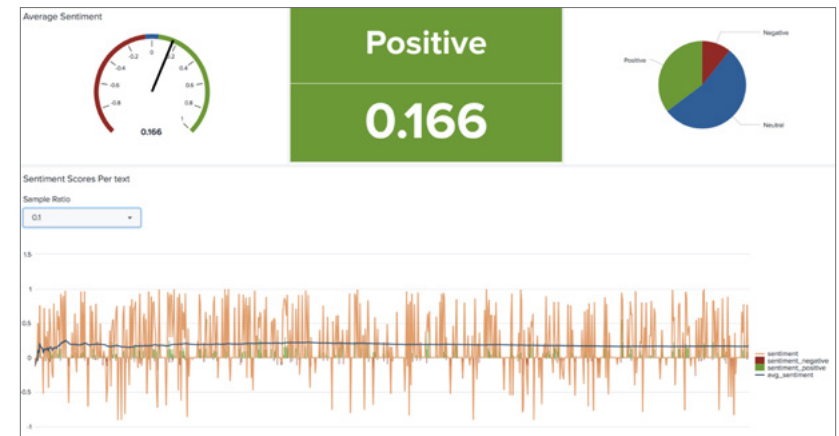
Machine learning and data mining are linked by process and desired outcomes, but while machine learning automates significant portions of the iterative learning approach that could be adopted by a more statistical-based data mining project, data mining is the process of analyzing the contents of a dataset to detect hidden patterns. The first step in this process is to identify the appropriate data mining (or machine learning) methods you need. Here, we want to determine whether a post belongs to the bot class or human class, which requires binary classification. Since these classifications will be based on examples from the existing dataset, this is also a case of supervised learning.

The Splunk Machine Learning Toolkit provides the algorithms we need for this process. It comes pre-bundled with sci-kit learn algorithms and allows you to import custom-built algorithms that make use of the Python for Scientific Computing library in addition to ML-specific commands that extend the SPL instruction base, and an Experiment Management Framework, which provides an interface for model versioning and lineage.

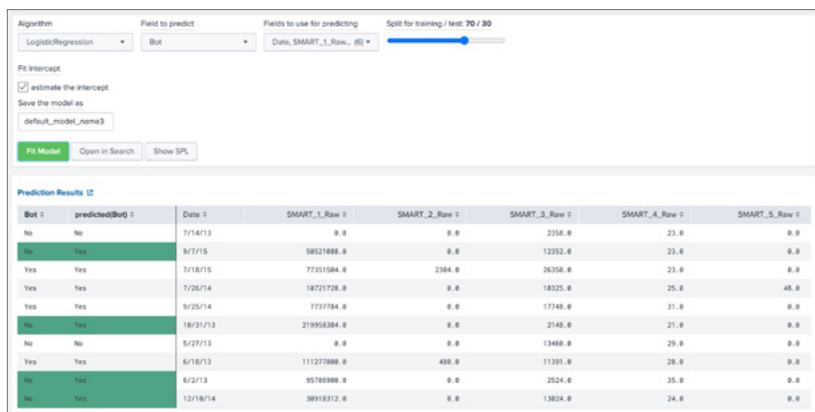
Using this toolkit, we can then conduct exploratory analysis by further filtering to create an optimal dataset for modeling. The Predict Categorical Fields workflow provides a UI to perform preprocessing tasks, such as improving the quality of the data through scaling numerical values, or reducing the number of fields to a set number of uncorrelated dimensions via Principal Component Analysis. However, at this stage, the most useful feature is the FieldSelector algorithm, which uses the scikit-learn GenericUnivariateSelect to select the best predictor fields and reduce the less useful ones that might otherwise lead to overfitting and compromise the quality of the model.



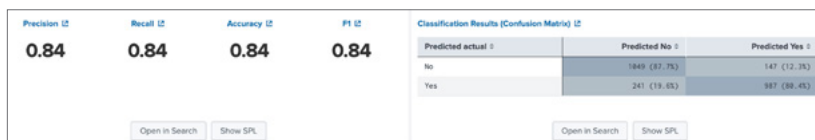
Another option at this stage would be to leverage [Splunk NLP Text Analytics](#), a Splunkbase application which extends the MLTK to provide natural language processing capabilities. With this, you can mine patterns within the dataset, such as the frequency of certain language terms and sentiments displayed, that could help differentiate bot from human and train the eventual detection model.



The final stage is to apply your chosen machine learning algorithms to the pre-processed data in order to discover patterns. For this, the MLTK offers a number of classifier algorithms out of the box, such as those for logistic regression, random forest and support vector machines. Here, you'll select an algorithm along with the prediction field; for example, the prediction could be whether or not a post belongs to the bot class, and the fields used to determine that. A slider also allows tinkering with the split ratio of training and test data for varied validation strategies, depending on your desired variance.



Fitting a model with the desired settings produces the individual prediction results, with incorrect predictions highlighted as well as broader recognized measures of classifier quality including accuracy, recall and precision. In the image below, the value of each of these measures is 0.84 — meaning that 84% of classification predictions are correct. Though you can try to optimize by increasing this value to reach as close to 100% as possible, it's worth noting that achieving this can sometimes be the result of overfitting, where the model generated is overtrained on the specific training dataset and won't perform well on new unseen data. Thus, a seemingly near-perfect prediction model could have an accuracy value of 0.84 or even less.



Using these results, along with the model saving and versioning we covered earlier, you can then compare different algorithms based on the quality of their generated models and make improvements by altering the fields and configuring hyperparameters in the UI.

Evaluation

Optimizing these models should be an ongoing process as more real-time data enters Splunk through training schedules within the MLTK.

For this, we can use the MLTK commands which extend the base Splunk Processing Language (SPL), specifically `| apply`, to apply the current model to the new data. Then, we can create saved searches to have this defined system run on a given schedule and learn quickly when a suspected bot has tweeted.

At that point, we can use alert actions to launch a ticket through Splunk On-Call or any other ticketing application to have a human moderator investigate. Depending on privileges, we could also connect directly with the social media platform in question via REST API Webhook to automatically remove the tweet.

The massive threat of automated misinformation bots on social media can seem overwhelming. Hopefully reading this chapter inspires you to implement this solution and contribute to our ongoing fight against malicious misinformation by turning the very technology that created these bots to infect our online discourse against them.

Rupert Truman

Rupert Truman works as a solutions engineer at Splunk, covering the commercial business across EMEA. He first joined Splunk as his region's first technical intern in 2017 and subsequently returned in 2019 after completing his degree in computer science at Newcastle University. He enjoys experimenting within the domains of machine learning and IoT, and working with customers to realize the value of their data across IT, application and security monitoring use cases.



“A combination of end user behaviors and technical issues were responsible for poor adoption of the technology, threatening a \$7 million investment.”

– Anonymous

How Data Keeps Hospitals Healthy

Companies spend millions, if not billions, of dollars trying to accelerate innovation, whether it's by changing the way they do business, adding new tools and technology or even acquiring other businesses. Too often, these initiatives fail — largely due to a lack of adoption. This, in turn, translates into little to no tangible return on investment. At the root of this problem is a lack of data. Only when organizations know when something is happening (or not) in a timely manner can adoption truly be successful. Otherwise, everything can be done right — the right tech, the right processes, the right long-term plan — but without proper insights on adoption and use it can all mean nothing. It's a common issue and one we've seen happen in spaces like healthcare, where change can be difficult to enact. We've witnessed it ourselves when new, smart medical devices were really taking center stage.

Health systems go smart

In 2014, for example, the Center for Medicare and Medicaid Services (CMS) created a program to increase the use of electronic health records (EHR) within its system to accomplish a variety of goals. It was part of the 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act to encourage healthcare providers to use systems which were better at sharing information and streamlining record-keeping processes. To encourage digital adoption, CMS was willing to reimburse organizations that were able to effectively deploy and leverage technology to meet that end. So, one of the largest hospital ecosystems in the country — with over 160 hospitals and other facilities — decided to roll out mobile vital signs monitoring systems to improve data collection and accuracy, provide a more seamless integration with the EHR system and improve patient outcomes.

This was no trivial project. Over 3,500 devices were distributed to about 160 sites for roughly 8,000 nurses to use on a daily basis with the goal of getting over 90% of vitals taken electronically and sent directly to the EHR. That is a lot of people in a lot of places, which opens the door for potential problems. User buy-in for such projects can be a challenge, especially in medical settings — which can lead to failure. How can institutions achieve consistent participation?

The short answer: data.

The ups and downs of implementation, adoption and use

People don't like change, especially when something has seemingly been working just fine. This is true for medical practitioners, too. When this hospital system began to integrate new practices with the devices, staff asked questions like, "Why now?" and "Are they going to replace us with robots and computers?"

Then there were technical issues. Batteries weren't lasting long enough. Devices weren't sending data properly. Manual input was required despite the promise of automation. But leaders in charge of the transition lacked the visibility into how 3,500 new medical devices across more than 100 sites were performing and being used. Back at headquarters, admins were running tests, using sample models, seeing vitals and everything seemed fine. There was clearly a disconnect, and visibility was integral to addressing it.

The role of visibility in device health

The organization had a few options. They could scrap the entire project and write off \$7 million — but they would also forfeit any chance of CMS reimbursing millions of dollars. They were also deciding whether to purchase extended-life batteries from the device manufacturer, which might resolve many of the complaints but would cost an additional \$3 million. Before they were going to spend another penny, leadership asked for more visibility into what was going on in the field. Not everyone was having the same problems with battery life, wireless connectivity, etc.

So, they started collecting and analyzing data from devices, servers and databases with Splunk. They also got verbal feedback from the medical professionals using the devices and watched what was happening on the ground more closely. Device data for battery life, wireless status, signal strength and usage metrics were now being collected, analyzed and visualized. Leadership could see visualizations and metrics reflecting data day over day, week over week.

One such chart was related to battery life. Most of the time it showed a typical slow drain from fully charged down until the device would begin re-charging and race back up to full capacity. On occasion, some devices would not behave according to this pattern, which the graph of percentage battery life over time clearly showed. Instead of draining and then zooming to the top, it just went to zero and stayed there.

“I can’t complete my rounds with this device. The battery doesn’t last long enough,” some of the nurses were saying. Others would report, “This device keeps dropping off the wireless network and can’t send the data from my patients’ rooms.” All of this was true, but not for everyone. More data was needed to better understand what was going on, and it needed to get into the right hands.

Splunk dashboards were built so everyone from the CEO to the nursing unit manager could see what was happening. Who was using the devices? Who wasn’t? Did they need more training or just more encouragement from their managers?

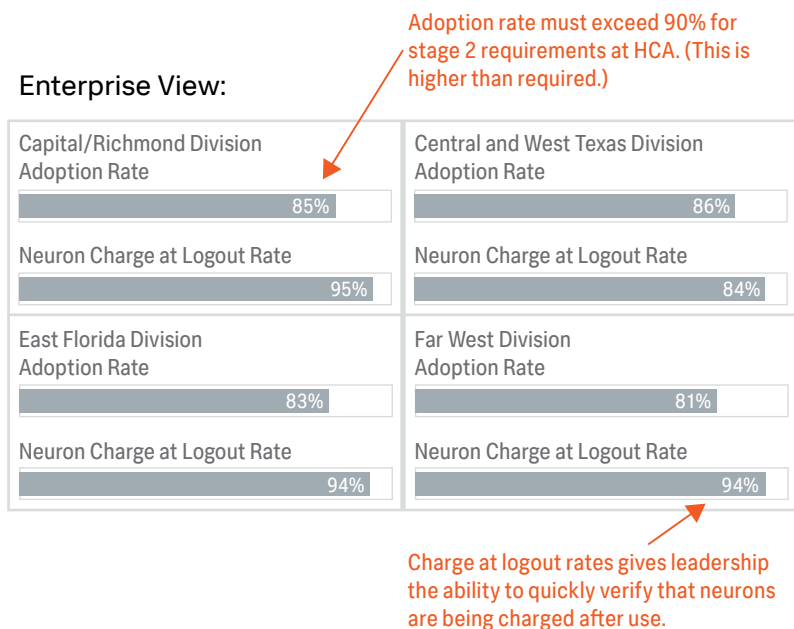


Figure 1: Examining adoption rates at the enterprise level, by division.

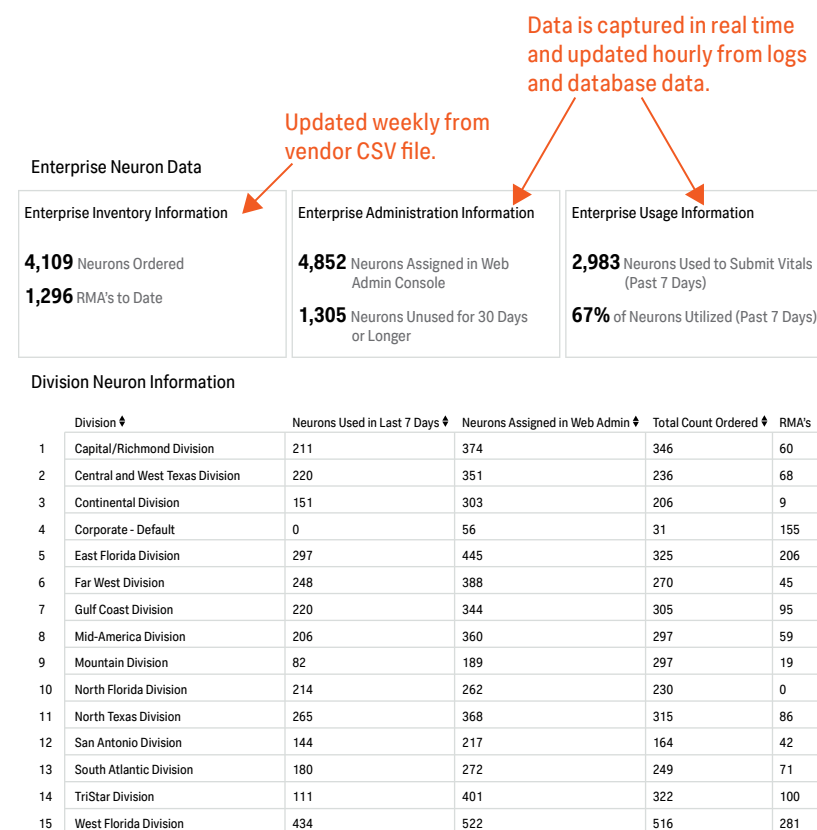


Figure 2: Aggregating information from CSV, databases and logs into a view to show usage.

With clear access to this data, it became clear why the devices were acting up and why users were having problems. When a nurse comes to work, there is typically a hand-off between the previous nurse who is getting ready to leave and the incoming nurse. After that, the on-duty nurse completes round one, where they visit each patient, take vitals, check status and create notes for any follow-up needed. After visiting each patient, they return to the nurse’s station and get necessary supplies so they can tend to their patients. This happens again at the middle of their shift and finally another hand-off happens.

Each time after rounding, nurses were supposed to plug in the devices to recharge the batteries — that just didn't always happen, and the data confirmed it. Sometimes it was related to urgent interruptions; other times, they simply forgot. Because patient and nurse information was available on each reading, admins could even identify which nurses were not plugging them in and address the issue accordingly.

There was also a problem with how devices were handling low power situations. It turned out that when battery percentage fell below 10%, the devices immediately dropped off the network. After conversations with the manufacturers, the team discovered that that device behavior was designed as a feature intended to conserve battery life. The feature was causing problems mostly due to lack of knowledge about how the devices really worked in certain situations. A little knowledge went a long way, and with some retraining, that issue was resolved.

Solving the battery problem saved the organization \$3 million.

The importance in reliability when it comes to adoption

But the problem of user adoption remained — users were still having trouble with the network. Fortunately, the previous battery issue had gotten the data wheels turning. Looking at wireless signal strength in conjunction with recorded patient vitals (which included room numbers) proved very helpful. It was easy to find out where signal strength was so weak that the device's data transmission failed.

Because the data so clearly showed what was going on with the network strength, the team made network improvements to ensure better coverage throughout the facilities. This resulted in better readings and increased patient satisfaction. Nobody likes being in a dead zone, especially in the hospital!

Collecting and relaying the right information

Frequently, technical faults get in the way of broader usage and adoption. But just as often, the issue can be political. Contextualizing data and relaying that information to the right people can solve this problem, especially when the right data is available to optimize feedback loops. If you know who your most active users are, they are the right people to be asking. It also reveals who the laggards are on the adoption curve, the ones who might need a nudge in the right direction.

More than solving people problems, reliable data also helped with security and compliance. Eventually, this hospital system also used several more dashboards to show readings tracking data on its journey from the medical device through the web and finally into the EHR.

What happens once data is streamlined?

With all of this data now available through the Splunk platform, it became easier to adopt and use the new devices effectively. For example, sometimes there were reports of missing devices. The nurse had it, was rounding, got interrupted and now could not easily locate the device. With the data readily available to them via Splunk, it was easy to see which room the device was last used in based on the dashboard readings. Data about the Wi-Fi signal strength could also help triangulate a device's approximate location.

Data as a detective tool struck again when another few groups began reporting that they did not have enough devices for all of the on-duty personnel to use. Back at headquarters, administrators asked for a list of devices that were assigned to those locations; an asset list was provided which included MAC addresses. These addresses could be easily linked to the network information already being collected. After quickly comparing

the list and all the MAC addresses on the network for the past 30 days, IT was able to come up with a list of devices that had never connected to the network. The suggestion? To check the closet or loading docks. Sure enough, facilities were searched, and the “missing” devices were found. Now everyone had enough devices to use to accomplish the goal and drive adoption.

More than devices

And data touch points come in handy even when not dealing with physical devices. Anyone using single sign on (SSO)? It’s a useful technology that makes software easier for everyone to access and requires fewer passwords to remember. Users log in one place, one time and get access to dozens of systems. Which begs the question: If a user has access to a system from SSO and never opens it, is some other company still receiving a license fee for that user? Probably. Which leads us to another way this hospital system used data to save more money.

There was a clinical application doctors were using throughout the system to enter diagnosis information related to patients. This system had two interfaces: a full application installed on the physician’s laptop and a web-based interface accessed through an SSO portal on a browser. After collecting some data from the systems using Splunk (e.g., the web logs, the application logs, the device information) the data showed something interesting. Out of thousands of doctors in the system, only one was using the full client application rather than accessing it through the web. That application was costing the business over \$1 million per year in maintenance and other costs. Here we had the opportunity to retrain one doctor and save a million dollars. That was easy. How many other opportunities do organizations worldwide have to save money on unused license costs associated with inactive users?

Granted, sometimes adoption stalls because many applications are overly complicated, or not user-friendly. Regardless, leaders need visibility at all the right levels to see who is using a system, who seems to be having trouble, who is spending too much time on it, and who is using it in unexpected ways. They need data to make informed decisions about what they can do to drive adoption, determine who really needs access to the systems and even make better purchasing decisions. And what easier way to start than with SSO data for user information and Splunk.


By evaluating the process as a series of steps, you can even understand where users are getting frustrated or need help getting to the next step. This information can be useful when considering what strategies to implement to drive higher adoption rates among user groups.

Data and adoption

Change, change, change. More often than not, in business and in life, projects and ideas fail. There are many reasons for this which we weren’t able to cover in this short chapter. But leaders can more readily drive adoption and usage of new technology and devices if it can be validated and verified throughout the rollout process. As we saw in our hospital example, these improved data practices increased visibility and attention to detail to make all the difference as the team adopted new processes. With data, leadership was able to quickly identify if their message was heard, understood and translated into action; they also could combat the political and technical headwinds that might have delayed or even destroyed efforts to make meaningful changes. The next time your organization decides to make a change, an improvement, or take a complete 180, think about how data can play a pivotal role in driving usage and adoption from the very beginning.

Eric Motz

Eric Motz is a senior manager leading the security and IT specialists covering New England for Splunk, where he has worked for more than 10 years. Eric and his team help provide customers with the knowledge they need to make better decisions.



“When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.”

– Lord Kelvin

How Data Can Help Score Your Cloud and Organization’s Security

This quote, by the inventor of the international system of absolute temperature that bears his name, often comes to mind when working with security metrics, key performance indicators (KPIs) and scorecards.

Though he wasn’t thinking about cybersecurity in 1891, the insight is as true today as it was 129 years ago: We need meaningful measurement in order to understand and improve things.



Measuring the security performance of today's modern companies with their complex organizational structures and distributed cloud-based infrastructures can seem daunting. But wherever data can be collected and measured in a meaningful way, we should be able to generate valuable metrics and indicators to understand and improve security performance.

This chapter will share a tried and trusted framework for identifying, developing and presenting security metrics, KPIs and scores. By using these calculations in a thoughtful and structured way, the leadership of an organization can better understand the importance of security in terms of business drivers — and help make the case for additional resources where needed.

The cost of not securing your business

The need for security will never go away, and will become only more important as business systems and services become increasingly interconnected.

There is no shortage of statistics showing the continuous rise of cybercrime, and the cost and frequency of breaches, as well as of regulatory penalties and fines. Yet security professionals have frequently struggled to get board-level buy-in for security investments. But times are changing, and now the board is paying closer attention.

As Gartner recently [reported](#), board members now realize how critical security and risk management is, and they're asking more complex and nuanced questions. As they become more informed, they're also more prepared to challenge the effectiveness of their companies' programs, asking questions like: Are we appropriately allocating resources? Are we spending enough? Why are we spending so much?

The board wants reassurance that their security and risk management leaders are not standing still, and they'll want to know about metrics and ROI. Security professionals and leadership need to find better ways of communicating about security performance and the value of security, and the board will need security professionals to help answer their questions and analyze metrics in a meaningful way. As a [NASDAQ report](#) highlighted, though business leaders are increasingly involved in cybersecurity, 91% of board members are unable to interpret a cybersecurity report.

Tackling these problems requires that security teams provide more transparency into the effectiveness and efficiency of security programs, and in a way that is clear, and clearly relevant, to organizational leadership. This transparency will also bring focus to better allocating resources, increasing accountability across the organization and demonstrating compliance.

For example, let's say a security team knows they have a weakness in configuration management in one particular cloud provider, and would like to implement an additional service to plug this gap. The board or budget-holder may find it hard to understand the technical benefits, and hence justify the spend, unless they're given clear, easily understandable statistics that demonstrate both the need and the value. This is just one example of the many challenges nearly every security team faces across the wide range of detective, preventative, investigative and responsive technologies they employ.

Securing the grade

Splunk Enterprise Security (ES) provides clear measurements over time and aligns with any security framework. Splunk ES can generate security metrics that can be normalized and aggregated into KPIs and scores. These indicators can then be used to create a dashboard that provides a high-level view of an organization's security, one that will make sense to security teams and board members alike.

Here are the key terms to understanding and implementing this solution. For our purposes, **measurement** is the primary calculation/aggregation performed to obtain the relevant data, such as the number of uses of default user names over a given time period. A **metric** is the measurement when analyzed over time. Multiple metrics combine to create **KPIs**. In security, KPIs will often relate to a particular domain, such as access control or vulnerability management, though KPIs can be anything that makes sense for the organization. Here, a **score** refers to a simplified, normalized value derived from the metric value. Measurements may include **dimensions** to provide more granular reporting. When analyzing security within an organization, this is usually related to a business unit or a cloud or technology vendor.

Planning

Generating security scores, KPIs and metrics does require some planning upfront.

Relevance

First, you'll need to figure out the security metrics, KPIs and scores that are most relevant and useful to you, as this varies widely from organization to organization. Some security metrics may be qualitative rather than quantitative, if there isn't easily measurable data available; for example, a risk analyst's gut feeling or the output of a security assessment. These qualitative metrics still have value, and can still be used. You can put them into Splunk (commonly via CSV format) and use them to widen the scope of the security scoring, or as a mechanism to validate or be validated by the real-world data.

Presentation

Consider from the start what the dashboards should look like, what questions might be asked of the data and what drill downs might be useful. The final presentation of the data is important. It should be clear, concise and show both successes and underlying issues still to be addressed.

Documentation

Be sure to document everything, and not just from the usual sustainability and supportability point of view. You'll need to be ready to answer the question of why your numbers are the way they are. Consider creating a table for each metric that defines what it means, why it exists and exactly how it's calculated. Here is an example:

Field	Description
Metric ID	Privileged Access Metric 1
Goal	The misuse of administrative privileges is a primary method for attackers to spread inside a target enterprise
Reference	CIS 4: Privileged Access
Metric Type	Implementation_X_Effectiveness_Y_Efficiency_X_Impact_X_
Metric (Search name)	sm_access.privileged
KPI	access
Calculation	Monitor number of uses of default account names relative to total authentications
Calculation	from datamodel:"Authentication.Authentication" stats count as total count(eval(is_Privileged_Authentication=1)) as value by dest_bunit rename dest_bunit as bunit
Target	N/A
Data Sources	ES datamodel: "Authentication.Authentication"
Data Owners	SOC (data is currently in Enterprise Security)
Frequency	Metric: Weekly Report: Monthly
Reporting Format	Current security metrics dashboard

You'll also want to document the KPIs that will be generated by aggregating and weighting the individual metrics into more understandable, business-centric insights. For example, there might be a KPI for access that combines metrics for privileged access, default access, failed logins and cleartext logins into a single, simple to understand number. Here's an example:

KPI	Metric	Weight	Invert
access	sm_access.privileged	0.2	1
access	sm_access.cleartext	0.4	1
access	sm_access.default	0.3	1
access	sm_access.lockout	0.1	1
vulnerability	sm_vulnerability.sev4_5	0.75	1
vulnerability	sm_vulnerability.missing_device	0.25	1

Selecting metrics

All of the methods we cover here are useful for both compliance and risk reporting, but it's important to be consistent. In compliance reporting, we tend to be looking for the good, like how many machines have antivirus installed, whereas in risk reporting, the opposite is more often the case (how many don't have antivirus installed). See the implementation details section on the next page for a method to invert a metric where needed.

When considering which metrics to use, the built-in KPIs in Splunk ES are a great place to start. You can view these in Configure > Content Management > Type: Key Indicator.

Start small, with a single metric and a single KPI, and don't try to implement metrics for the entire NIST Cyber Security Framework or CIS Controls from the outset. Consider starting with either a specific area of concern, or an area where visibility for the organization needs to be improved. In the current security climate, most organizations get the most immediate value from access controls, endpoint and vulnerability metrics.

Implementation details

Splunk ES automatically generates security KPIs over a wide range of domains (access, incidents, vulnerability, etc.). However, KPIs are not persistent in the system, because traditionally there is no need, as they can be generated on the fly. These KPIs are commonly unbounded numbers, so it can be tricky to compare like-for-like across business areas where volumes of data may be highly varied, and scoring can be difficult. What follows is a framework for generating and sustaining these metrics over time so they can be monitored and analyzed at a high level as well as at a technical level. It assumes a basic level of knowledge of Splunk, for example, that you know how to create an index, report and dashboard. If you don't commonly perform these tasks, you can easily learn how in [Splunk Docs](#).

The framework could be useful for other security reporting purposes beyond security scorecards, for example for risk or compliance reporting.

The data

Any data within Splunk can be mined to create metrics, KPIs and scores, but with security in mind, the best place to start is with the Splunk ES data models. Here the data is already classified, normalized and enriched, making it easier for you to find and query the data you need. However, if the data you want to report on is not in a data model, you can still perform these steps, as long as the appropriate methods are implemented.

First, create a store for your security measurements to ensure they can be quickly and easily queried and stored for as long as necessary.

Create a small new event index, called security_measures, and it doesn't need to be too large because the data volumes stored here will be relatively small. In this example, we are using 1GB.

New Index

General Settings

Index Name: security_measures
Set index name (e.g., INDEX_NAME). Search using index=INDEX_NAME.

Index Data Type: Events (selected) | Metrics

Home Path: optional
Hot/warm db path. Leave blank for default (\$SPLUNK_DB/INDEX_NAME/db).

Cold Path: optional
Cold db path. Leave blank for default (\$SPLUNK_DB/INDEX_NAME/colddb).

Thawed Path: optional
Thawed/resurrected db path. Leave blank for default (\$SPLUNK_DB/INDEX_NAME/thaweddb).

Data Integrity Check: Enable (selected) | Disable

Max Size of Entire Index: 1 GB

Max Size of Hot/Warm/Cold Bucket: auto GB

Frozen Path: optional

App: Enterprise Security

Storage Optimization

Tsidx Retention Policy: Enable Reduction (selected) | Disable Reduction
Warning: Do not enable reduction without understanding the full implications. It is extremely difficult to rebuild reduced buckets. [Learn More](#)

Reduce tsidx files older than: Days

Age is determined by the latest event in a bucket.

Save Cancel

Measurement searches

The second step is probably the most time-consuming part of developing scorecards: creating the searches that will populate your measurements index. You already know the metrics and KPIs you want to develop. The next task is to build the search to create the measurement.

1. It's usually best to store a basic measurement, as calculations can be performed at reporting/visualization time, giving you the flexibility to change them over time as required. It also means you can easily drill down from the abstracted metrics and score to display the actual data points if needed. Using the example of default authentications, we would want to store the number of default authentications that occur over the search time period, so the value might be 16,249.

New Search

| from datamodel:"Authentication.Authentication" | stats count(eval(is_Default_Authentication=1)) as value

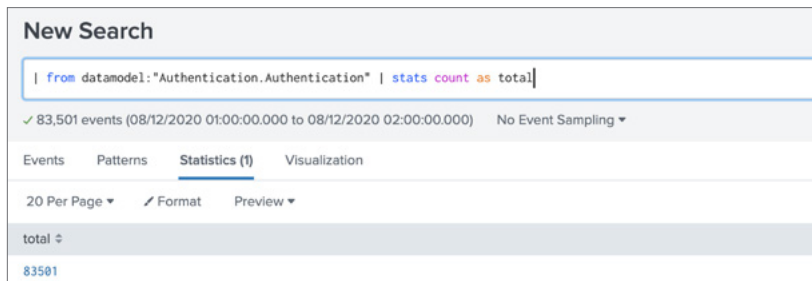
✓ 83,477 events (08/12/2020 01:00:00.000 to 08/12/2020 02:00:00.000) No Event Sampling

Events Patterns **Statistics (1)** Visualization

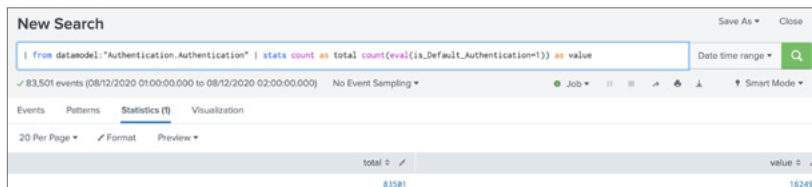
20 Per Page Format Preview

value
16249

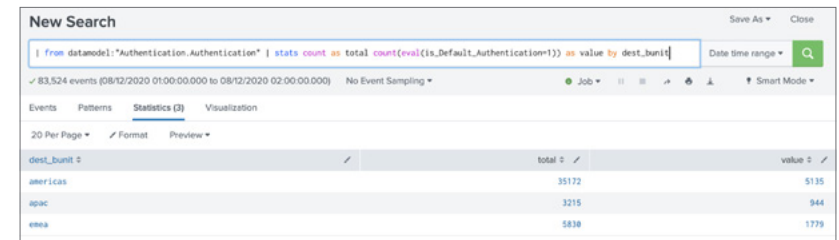
- To ensure the data is as usable as possible, you want it to be a comparative measure so you can compare one environment to another without having to worry about the relative sizes of different environments. Using the above example, the total number of authentications might be 83,501. If the measure you are looking at is already a percentage, use the search to set the total to 100 with the eval command.



- When analyzing and reporting on the data, you'll likely need to break down the metrics and score by various dimensions. This is where the pre-planning pays off. Most commonly this is performed either as a business unit like sales or finance, or technology vendors like AWS, Azure or Symantec. If this is your first time implementing scorecards, it might be worth skipping this step initially to focus more closely on how to generate the most meaningful organization-wide KPIs and metrics. Normalizing all of the measurement data into one structure will allow you to extract whatever data is required through a single, simple search. Once you've done all this, the output from any search should look something like the examples below.



Output for a business unit dimension:



Output for technology vendor dimension:

cloud_vendor	value	total
google	320	8,142
amazon	101	1,398

I won't provide an example for a search here, because cloud enhancements to the data models mean the data available to you is version-dependent. However, look at the data available within your data model, create a normalized field name across all searches, and populate it with the relevant field in your data (e.g. product_vendor, sourcetype) that contains the pertinent information for your technology vendors.

Always run the search manually to ensure it returns the expected results before scheduling the search. It's a good idea to run the search over a long period of time to get a feel for how these numbers will change over time.

Search reports

Now that the basis of the search is complete, it's time to schedule the search as a report. The simplest way to perform this from a search window is to Save As > Report. Of the five values needed, the fourth is the **name**. Give the report a name that reflects the metric you're gathering (for example: sm_access.default). Follow a naming convention for clarity later on. For example:

sm_access.default

Where:

sm_ denotes it is a security metric
access denotes the KPI the metric rolls up to
default denotes the metric name

Edit the schedule for the report, and select to schedule every week or month. This will appear in the output data as **search_name**.

To simplify reporting, it's preferable to have all reports scheduled over the same time frame.

Edit Schedule

⚠ Scheduling this report results in removal of the time picker from the report display.

Report: **sm_access.default**

Schedule Report: [Learn More](#)

Schedule: Run every day

At: 1:00

Time Range: Yesterday

Schedule Priority: Default

Schedule Window: No window

Trigger Actions: + Add Actions

Cancel Save

Edit the summary indexing for the report, and enable and select the summary index created previously (security_measures). This will write the data into the index every time the report runs.

Edit Summary Index

Report: **sm_access.default**

Enable Summary Indexing: [Learn More](#)

Select the summary index: security_measures

Only indexes you can write to are listed.

Add Fields: [Add another field](#)

Cancel Save

The last required value is the **timestamp**, which you don't need to add manually. Splunk will automatically add multiple timestamps to the data including the search time range (start and end) and the time the search was run. The **default_time** field is the end of the search time range, which is usually the best timestamp to use.

Once the scheduled reports have run, a normal search should show you your measures within the **security_measures** index.

index=security_measures | table _time search_name <dimension> value total

_time	search_name	dest_bunit	value	total
2020-10-28 08:00	sm_test1	emea	5824	5824
2020-10-28 08:00	sm_test1	apac	3195	3195
2020-10-28 08:00	sm_test1	americas	34944	34944

Create a security metric matrix

Part of the planning process is to map metrics to KPIs. This forms the basis of the metric matrix, but you'll still need to add some additional information.

This is commonly done by creating a spreadsheet as a CSV file and uploading it as a lookup. There should be one line for each search report. The metric field should be the name of the report, and the KPI field should be the name of the KPI to which it contributes. Though you can extract this from the metric name if this naming scheme is followed, it's clearer to have it defined separately in a spreadsheet.

At minimum, a weight field is required to provide the relative weighting of each metric in the calculation of the KPI. For each KPI, these should add up to 1.0. It's helpful to create a scorecard health dashboard where you can check this, along with a timechart that monitors measurement generation. Splunk ES can be configured to generate an alert in these circumstances.

Including an invert field can be very useful in simplifying and standardizing the SPL code across the metrics and KPIs. Many security measurements are based on finding the bad rather than the good, and an invert field can quickly and easily flip a metric to ensure that all of your metrics are consistent and aligned. Typically, the output of this process looks something like the chart below. Here, "0" means don't invert, "1" means to invert.

metric	kpi	weight	invert
sm_access.privileged	access	0.50	1
sm_access.remote	access	0.25	1
sm_access.default	access	0.25	1
sm_malware.detection	malware	0.50	1
sm_malware.coverage	malware	0.50	0
sm_vulnerability.sev45	vulnerability	0.50	1

...

Once you've created your metric matrix, you'll create a new managed lookup in Splunk ES and drop your csv file in. In this example, the following searches assume the definition name of metric_weights, and the csv field names as in this table.

The screenshot shows the 'Manage New Lookup' dialog box. It features a title bar with the text 'Manage New Lookup' and a close button (X). Below the title bar are two tabs: 'Create New' (which is selected) and 'Select Existing'. A dashed box contains the text 'Drop your file here or browse...'. Below this are several form fields: 'App' (dropdown menu set to 'Enterprise Security'), 'Destination File Name' (text input 'metric_weights.csv'), 'Definition Name' (text input 'metric_weights'), 'Lookup Type' (dropdown menu set to 'Manually edited'), 'Label' (empty text input), 'Allow Lookup Editing' (checkbox checked), and 'Description' (empty text input). At the bottom right are 'Cancel' and 'Save' buttons.

Reporting

Now that the basic data structures are in place, here comes the good part, when you actually get to use them. If you have followed the previous steps, the following search will calculate each metric as a percentage, and as a score between 0 and 5, and invert data, if required. It might be worth saving the lookup and evals statements as a macro, as it will likely be the base of most of the searches you use to visualize the data.

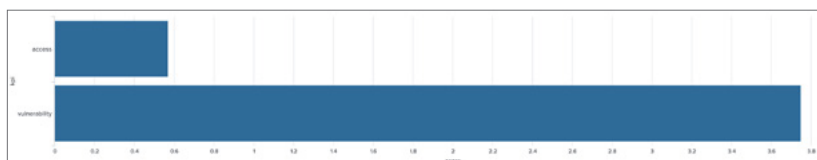
```
index=security_measures | lookup metric_weights metric as search_name | eval perc=case(invert==0, (value/total*100), invert==1, 100-(value/total*100)), score=round(perc*weight/20, 2)
```

Dashboards

Now you are ready to deploy all of your visualization and dashboarding skills to present the data in the most compelling way. The planning phase should have provided a view to the information that would be useful to present on the final dashboard(s). Here are some examples of the calculations and visualizations you can present:

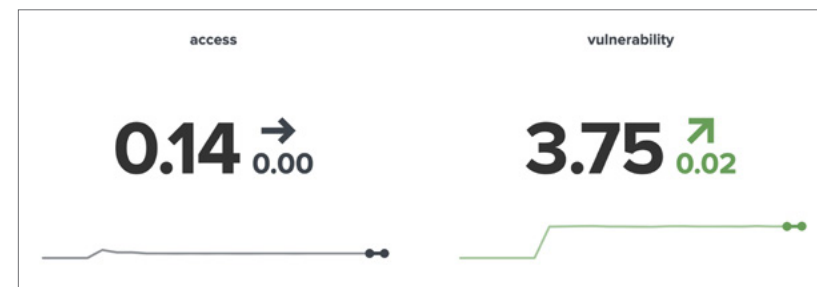
Business unit breakdown by KPI

```
index=security_measures bunit=americas | lookup metric_weights metric as search_name | eval perc=case(invert==0, (value/total*100), invert==1, 100-(value/total*100)), score=round(perc*weight/20, 2) | chart sum(score) as score by kpi
```



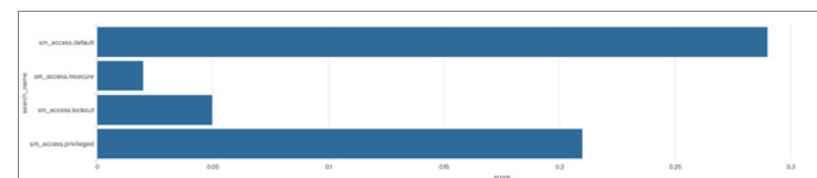
Business unit trend by KPI

```
index=security_measures bunit=americas | lookup metric_weights metric as search_name | eval perc=case(invert==0, (value/total*100), invert==1, 100-(value/total*100)), score=round((perc*weight/20, 2) | timechart span=1w avg(score) as score by kpi
```



KPI score breakdown by metric

```
index=security_measures bunit=americas | lookup metric_weights metric as search_name | search kpi=access | eval perc=case(invert==0, (value/total*100), invert==1, 100-(value/total*100)), score=round(perc*weight/20, 2) | chart avg(score) as score by search_name
```



The beta dashboard app includes useful new features for improving the look and feel of the resulting dashboards. The following is an example of a dashboard created to present a security scorecard designed to aggregate and score the organization's on-premises data centers alongside the cloud providers in use.

Here additional non-metric-based, real-time threat and event data is displayed alongside the scorecard to provide a holistic view of the organization's security posture.



Providing more transparency into the effectiveness and efficiency of security programs is a challenge many security teams face. Hopefully this chapter has highlighted some methods and techniques that can be used to help achieve this aim.

These methods can be applied across the whole range of security domains to provide information on where security controls are working effectively and efficiently, and hence what areas are in need of investment of either tools or time.

Additionally, these same methods, metrics and KPIs can be used as the basis for any risk and/or compliance reporting requirements incumbent on the security organization.

By conveying this information in a clear and pertinent manner that aligns with the asks of the board or security management, the security organization can both better focus the currently available resources as well as be able to better justify the need for additional resources.

Graeme Sinden

Graeme Sinden is an account sales engineer at Splunk specializing in network security. Through his more than 20 years of experience, first in post-sales and now pre-sales security roles, he has gained a thorough understanding of security issues, helping customers and sales teams alike.

“Everything old is new again.”

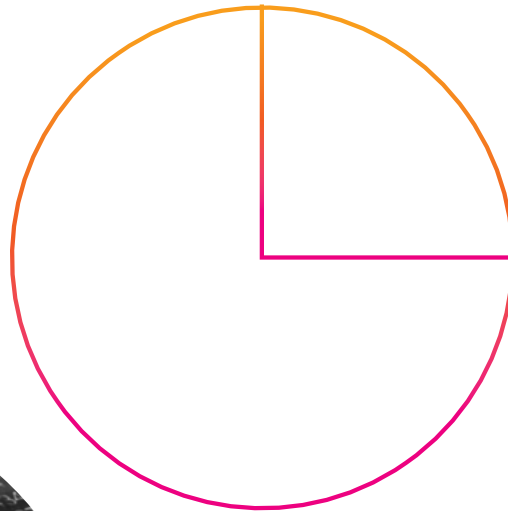
– Peter Allen



Straight Outta Syslog: A New Look at an Age-Old IT Data Collection Problem

Wait, a book about the future and you're talking about syslog?

Yes. One of the earliest itches the founders of Splunk wanted to scratch some 17 years ago was to analyze Cisco PIX logs. Prior to Splunk, just about all that could be done was to save these logs to disk and look at them by hand, or with basic search tools such as `grep`. A primary reason for this was the lack of a suitable database, or more specifically a database structure (or schema) for this type of free-format data.



A hallmark of Splunk since its founding is its ability to handle unstructured data. A large part of why this was (and continues to be) so revelatory is because log data, and particularly network device data, has essentially no format or structure requirements on the event payload. Therefore, event formats each vendor uses for their device families are vastly different, making it incredibly difficult (until Splunk came along) to analyze different network device logs as a group — which is key to unearthing the real story that the data is telling.

How does Splunk handle unstructured data, and why is it so effective? The key is a technique called “schema-on-the-fly” or “schema at read.” Data storage and retrieval in Splunk differs from that of a standard relational database in the fact that the schema is not required when the data is stored, but rather gets applied when the data is read. This fundamental difference allows you to collect, store and index (catalog “tokens,” or individual portions of a log event) into a time-based data store for quick retrieval — regardless of the format of the data and questions you may wish to ask of it. A schema is applied only when the data is queried (read) for analysis. Splunk simply delays the schema application until query time, making collection of unstructured data (and maintenance of the schema) far simpler and analysis much richer via the search processing language (SPL). But remember a critical fact: A schema is still required; furthermore, a separate one is required for each kind of data. We will explore how this is fundamental in the sections ahead.

Fast-forward more than 17 years to today. Splunk has revolutionized the analysis of network and security device data but traditional data collection and ingestion methods have severely hampered the effectiveness of schema-on-the-fly for critical parts of IT infrastructure. This had huge implications on the ability to collect network and security device data at scale — until we introduced a revolutionary change in how we treat this part of the process. Let’s explore the details of how we unlocked the power of schema-on-the-fly by rethinking our approach to IT infrastructure data collection to provide greater value in the Splunk ecosystem.

The syslog challenge

In a majority of Splunk installations, network and security devices comprise up to 50% of all data ingest. Handling this type of data efficiently and at scale is critical to the success of almost every Splunk deployment, so it is imperative to arm ourselves with the best tools and methods when architecting this portion of a Splunk installation. But here is the challenge: Network devices send logs using a transport protocol that was designed close to 40 years ago. This is close to the practical age of the internet itself! Think of the application-layer protocols that have come and gone since then, like internet relay chat (IRC), gopher, finger and even file transfer protocol (FTP). These are rarely used anymore and have been replaced with more functional and secure versions or completely new protocols. Just a few of the older ones remain in use — and one is even undergoing expanded adoption by many vendors.

That protocol is syslog. And the fact that the syslog protocol is so old and remains an integral (and expanding) part of the data center after so much time warrants a new look at an age-old IT problem.

Let’s dive into the specifics of syslog, which involve both technical and behavioral challenges:

- The syslog protocol is so old that it dictates very basic network architectures. Attempts to centralize data collection with load balancers for scale and redundancy are simply not possible with this protocol, which was designed for efficiency above all else. As the volume of data has skyrocketed, unique scale challenges have arisen with traditional Splunk architectures.
- The protocol enforces little payload structure, essentially allowing vendors a completely unstructured approach to their event formats. One vendor has close to 50 different payload formats. This one characteristic and its effect on Splunk drive much of the requirement for extensive coding just to get the data into Splunk (see more below).

- The traditional tools used to collect such data (and attempts to categorize it properly for effective Splunk use) are themselves more than 20 years old and, while technically solid, are essentially programming toolkits. This dictates a substantial learning curve and forces administrators to write code simply to collect and categorize this kind of data and effectively send it to Splunk. These coding efforts are almost always one-offs that are applicable only to the enterprise for which it was written and even to the unique desires (and capability) of the administrator who wrote it. Rarely are they documented, which often sees administrators inheriting old designs and starting over in frustration. Rinse and repeat.
- This requirement to effectively write code to properly categorize and prepare syslog data for ingestion into Splunk has resulted in many organizations simply punting and not doing it at all — which has huge implications for Splunk’s operational efficiency. Though Splunk offers the rudimentary capability of listening on a syslog port (with the allure of Splunk accepting any data), that route omits the very important categorization step needed for the schema at read process to effectively operate.

These challenges have plagued administrators since Splunk’s inception. It was clear that a new approach to effectively collect and prepare syslog data for analysis in Splunk was needed. Not taking the necessary step has huge business implications: Customers either undertake a significant development and architecture step on their own with little guidance, or (even worse) they do nothing at all, severely compromising the value of their Splunk investment. Fortunately, Splunk Connect for Syslog has taken a 40-year-old protocol and turned it into a best practice of tomorrow.

Splunk Connect for Syslog: best practices of tomorrow

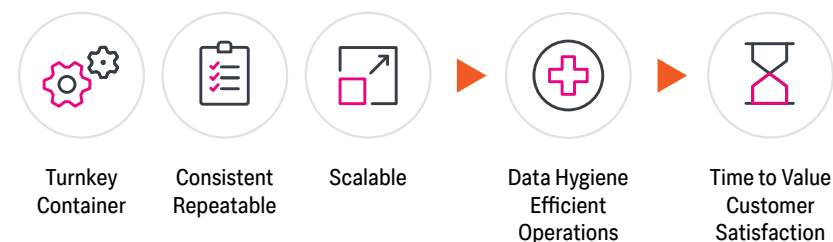
In our quest for a thoughtful response to the challenges above, we sought to distill them into a single question:

As an administrator, how do I easily ingest syslog data at scale while removing the requirement of upfront design work and “syslog-fu?”

To effectively address this question, we created Splunk Connect for Syslog (SC4S):

Splunk Connect for Syslog

A solution for Splunk’s oldest data source



Working left to right in the diagram, the goal was to ensure that the data arriving into Splunk is properly curated, which results in far more efficient Splunk operation. The resulting short time to value ensures a satisfied customer base for what may account for 50% of their entire Splunk data complement. In short, SC4S must be:

- *Turnkey*: Little to no syslog configuration (i.e., coding) knowledge is assumed or needed.
- *Consistent and repeatable*: The tool must operate and administer the same for all enterprises and be thoroughly documented so that one-off solutions are avoided.
- *Scalable*: It must account for the limitations of the syslog protocol, and must scale — both up and down.

In the year since SC4S was released we have seen a huge reduction in the number of hours spent — by customers, professional services, partners and sales engineers — on re-architecting syslog collection architectures. Challenges remain, however, on educating customers about the fundamentals of sound network architecture, as they run counter to much of today's network engineering philosophy for more modern protocols. Again, this stems from the fact that we're still working with a protocol that is entirely unchanged from 40 years ago. If this chapter can shed light on this one topic alone, it will be a success.

Let's dig into these challenges one by one and explore how they have been met with the SC4S design and a sound architectural approach, and flip this 40-year-old data collection challenge into the 21st century!

Meeting the challenges with SC4C

Network architecture

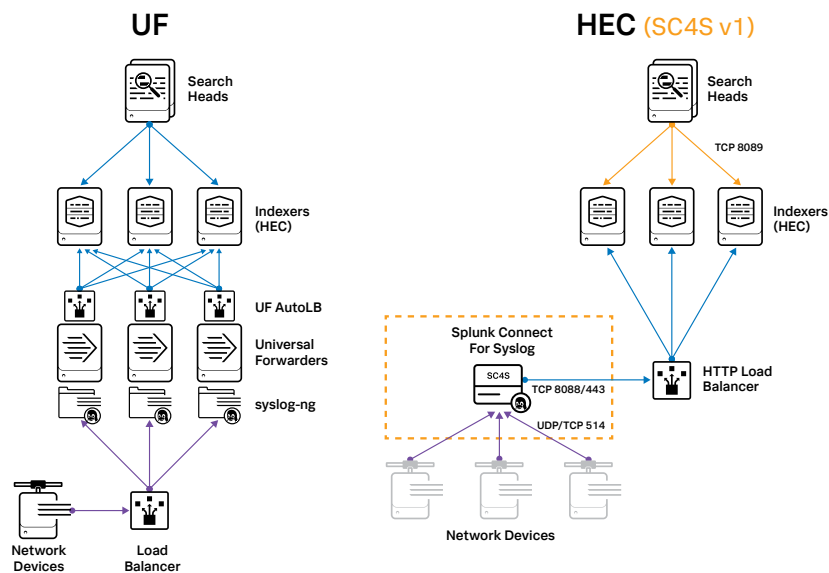
Designing a proper network architecture for syslog demands that you step way back in time. It's important to understand that a key tenet of the syslog protocol is that it's totally stateless and unacknowledged at the application layer. Again, the protocol was designed in the 1980s to accommodate the very limited processing power of network devices. Their job was to process packets and get them on their way, not create logs. Only the table scraps of CPU cycles were available for log production, which meant that the logging protocol had to stress efficiency over resiliency. This philosophy has carried over to today's devices — they can process far more packets than those from four decades ago, but CPU availability for logging still remains a premium. This is a fundamental reason why syslog has endured (and even expanded) to this day.

For years, the recommended best practice for syslog data collection was this: Data is sent to a syslog server (typically syslog-ng or rsyslog) which listens on one or more UDP or TCP ports (most often the agreed-upon well-known port of UDP 514). Events are then written to disk, at which point a standard universal forwarder (UF) monitors the log files and sends them to Splunk.

As the years went on, though, a couple of key changes took place. Firstly, Splunk developed a new collection mechanism for data collection in the form of HTTP event collector (HEC). This allows direct communication from clients without having to write events to disk first. Secondly, this happened at about the same time as the volume of syslog data in many Splunk deployments started to explode, causing scale issues that exacerbated challenges with the UF when used for aggregated data collection.

This issue led us to explore alternatives to this tried-and-true design philosophy about three years ago. The issue of scale exposes a key problem of using the UF in an aggregated manner, for which it was not designed. Rather, the UF was designed to be deployed on thousands of clients, all contributing roughly equal to the whole. By being distributed over so many clients, random data ingest is relatively assured. Each indexer will be connected to a random and relatively equal number of forwarders at any given time. This ensures each indexer is contributing equally to data ingest as well as search.

This notion is flipped on its head when just a few UFs collect aggregated data. In that scenario, correspondingly few indexers are participating in the ingest process as well. This leaves some indexers drinking from a firehose while others are starving for data. Ultimately, this leads to degradation of search performance too, since for any given time slice a relatively few indexers have the data needed — limiting parallelization. This has huge implications for short-window searches, such as those used for data model acceleration.



Therefore, a new architecture was developed which forms the foundation of Splunk Connect for Syslog. Using the HTTP event collector (as shown above on the right), data can be “sprayed” to the indexers from an aggregated source far more effectively than the Universal Forwarders. This is partly due to the protocol itself (HTTP), which enterprise load balancers handle very well.

You will note one other critical difference between the diagram on the left and the one on the right. On the recommended architecture on the right, you won’t see a load balancer between the devices and the collector. This is a fundamental characteristic of the protocol itself, and is independent of the collector itself (e.g., SC4S, rsyslog, etc.). In the quest for efficiency, syslog was designed as a stateless/connectionless protocol.

Traditional load balancers used for high availability or scale depend on acknowledgement and retry on the client side to avoid data loss. Being used with syslog, a “send and forget” protocol, ultimately results in more data loss over time (not less). There are [several scenarios](#) which can result in this loss, which renders a protocol that, at best, can only be made “mostly available.” If there is one takeaway from this chapter, it’s to resist all urges to insert a load balancer between the devices and the collectors. Use OS-level clustering with a shared IP, or even VMware vMotion, to provide for a “mostly available” (and simple) solution.

Best practice of tomorrow: SC4S utilizes HEC for data transport to Splunk and scales down as well as up so that small collectors can be placed at the edge, as close to the devices as possible. Syslog demands true edge data collection. No amount of desire for centralized collection will magically make a 40-year-old protocol bend to today’s modern networking approaches.

Payload structure

In the introduction, we outlined the key difference between a traditional relational database and Splunk, and that was when the schema was applied — at ingest time and at read, respectively. But in both cases, a schema is required for effective data analysis. We also noted that in Splunk, a separate schema is needed for each kind (format) of data. This requirement poses a challenge with syslog, because the protocol places essentially no boundaries or structure on the data itself. Newer but far less ubiquitous versions of the syslog protocol help somewhat, but the easier and far lazier approach is to use the older version. Therefore, each vendor requires its own schema, and often a separate schema for each device from a particular vendor. Read on to discover how this has traditionally been handled.

Best practice of tomorrow: Each event passing through SC4S is assigned the proper Splunk metadata (e.g., time, index, host, source and sourcetype). In many cases, simple elements like time and host are not in the syslog header where you would expect to find them but instead located deep in the event payload. SC4S accounts for these nuances for 40 of the most common device types in the industry. The goal is to give a well-curated event from which to start TA development.

Existing tools and processes

Splunk users are not the only ones that want to make sense of syslog data and, of course, this kind of data existed far before Splunk did. For this reason, the Unix community developed what has been distilled into two main tools (or “syslog servers”) to handle this kind of data: `sysLog-ng` and `rsysLog`. Over the years, both have developed into very solid open source software packages, with decades of development resulting in extremely capable operation. What has not developed is any semblance of out-of-the box operation. Both have their own domain-specific languages that are used primarily to parse and categorize data, which is exactly what we need for Splunk. But their effective use requires software development, something most Splunk administrators and users have not signed up for. Nothing prior to the release of SC4S provided anything approaching turnkey operation for the majority of device types.

Best practice of tomorrow: Here is where a considerable amount of development effort was made in the last two years. How do you abstract the process of building filters or log paths (the development exercise when using traditional syslog servers) while affording their flexibility? SC4S retains the benefits of existing, well-entrenched open source software (`syslog-ng`) while providing an abstraction that eliminates most of the development tasks. For most installations, simple environment-variable substitutions replace the traditional development work and lookup files to assign Splunk metadata. For those needing to handle custom sources, [even the development tasks are template-driven](#), while none of the domain-specific language is restricted.

The perils of punting

Given the requirement to effectively develop code just to categorize data, there is a strong urge to simply throw up your arms and just say, “Forget it. Splunk can ingest anything, so we’ll deal with it when the data lands in Splunk.” At that point, many just send their syslog data directly to Splunk infrastructure like UFs, HWFs or even indexers. Let’s look at an analogy to see why that is exactly what you do not want to do.

Imagine a small town sheriff’s department that wants to collect and analyze traffic citations in their jurisdiction. They have set up a database to collect the data, including the route and street where the citation was issued, the make and model of the car, and ending with a free-form “detail” field where the officer can enter anything else that may be relevant. Here is an example of that database, where the columns are mapped to their Splunk equivalent fields:

Ingres (source)	Make (sourcetype)	Style (field)	Details (event)
Hwy 125	BMW	Sedan	Blue, BMW, Sedan, 2018, 4 passengers
Elm St.	BMW	Hybrid	Grey, BMW, Hybrid, 2019, 2 passengers
Hwy 35	Ford	Sedan	Red, Ford, Sedan, 2018, 1 person
Hwy 125	Ford	Truck	Black, Ford, Truck, 2018, 1 person, crew-cab
Hwy 35	Mercedes	Sedan	White, Mercedes, Sedan, 2018, 1 pax
Main St.	Mercedes	SUV	Grey, Mercedes, SUV, 2015, 3 pax
Interstate	Interstate		Blue, Ford, Sedan, 2018, 2 persons
Interstate	Interstate		Red, Tesla, Sedan, 2019, 1 passenger, electric
Interstate	Interstate		Grey, BMW, Sedan, 2016, 1 passenger
Interstate	Interstate		Blue, Mercedes, SUV, 2018, 2 pax
Interstate	Interstate		White, Toyota, Truck, 2018, 4WD, 1 passenger
Interstate	Interstate		Blue, Tesla, Sedan, 2019, Dual Motor, 2 passengers

You will note a quirk in the data above, and one that presents a continual challenge for the sheriff. All of the citations on the interstate are categorized as “Interstate” for the make of the car, and have no data whatsoever for the style because the highway patrol has a database that is not fully compatible with the sheriff department’s. Fortunately the sheriff does receive a full detail record from the highway patrol that can construct the missing fields. That has to be done manually and is very time-consuming and expensive, but does yield what is needed to make the database whole.

Ingres (source)	Make (sourcetype)	Style (field)	Details (event)
Hwy 125	BMW	Sedan	Blue, BMW, Sedan, 2018, 4 passengers
Elm St.	BMW	Hybrid	Grey, BMW, Hybrid, 2019, 2 passengers
Hwy 35	Ford	Sedan	Red, Ford, Sedan, 2018, 1 person
Hwy 125	Ford	Truck	Black, Ford, Truck, 2018, 1 person, crew-cab
Hwy 35	Mercedes	Sedan	White, Mercedes, Sedan, 2018, 1 pax
Main St.	Mercedes	SUV	Grey, Mercedes, SUV, 2015, 3 pax
Interstate	Ford	Sedan	Blue, Ford, Sedan, 2018, 2 persons
Interstate	Tesla	Sedan	Red, Tesla, Sedan, 2019, 1 passenger, electric
Interstate	BMW	Sedan	Grey, BMW, Sedan, 2016, 1 passenger
Interstate	Mercedes	SUV	Blue, Mercedes, SUV, 2018, 2 pax
Interstate	Toyota	Truck	White, Toyota, Truck, 2018, 4WD, 1 passenger
Interstate	Tesla	Sedan	Blue, Tesla, Sedan, 2019, Dual Motor, 2 passengers

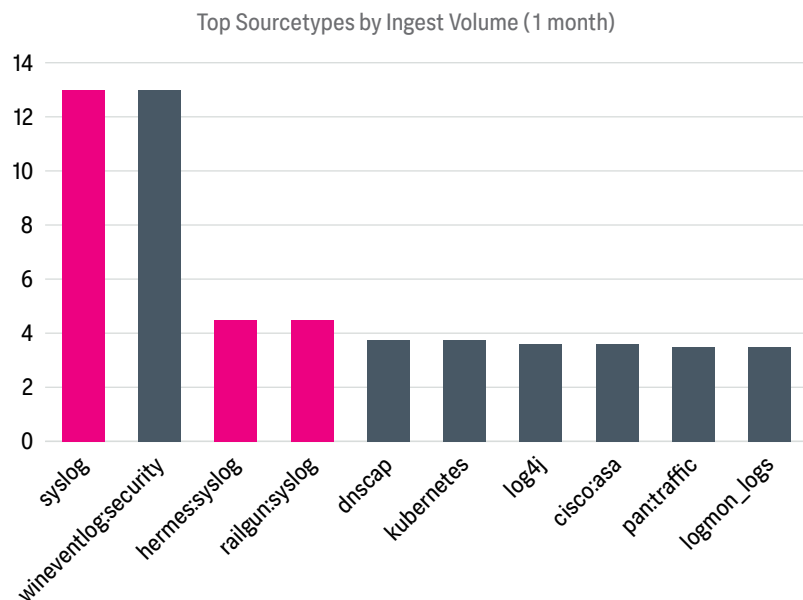
After this expensive and manual work, the fields are populated and a final sort yields:

Ingres (source)	Make (sourcetype)	Style (field)	Details (event)
Hwy 125	BMW	Sedan	Blue, BMW, Sedan, 2018, 4 passengers
Elm St.	BMW	Hybrid	Grey, BMW, Hybrid, 2019, 2 passengers
Interstate	BMW	Sedan	Grey, BMW, Sedan, 2016, 1 passenger
Hwy 35	Ford	Sedan	Red, Ford, Sedan, 2018, 1 person
Hwy 125	Ford	Truck	Black, Ford, Truck, 2018, 1 person, crew-cab
Interstate	Ford	Sedan	Blue, Ford, Sedan, 2018, 2 persons
Hwy 35	Mercedes	Sedan	White, Mercedes, Sedan, 2018, 1 pax
Main St.	Mercedes	SUV	Grey, Mercedes, SUV, 2015, 3 pax
Interstate	Mercedes	SUV	Blue, Mercedes, SUV, 2018, 2 pax
Interstate	Tesla	Sedan	Red, Tesla, Sedan, 2019, 1 passenger, electric
Interstate	Tesla	Sedan	Blue, Tesla, Sedan, 2019, Dual Motor, 2 passengers
Interstate	Toyota	Truck	White, Toyota, Truck, 2018, 4WD, 1 passenger

Only now can a proper and fast analysis take place. For instance, we can instantly check to see how many Fords were cited during the window of this search.

So what does this have to do with Splunk and network (syslog) devices in particular? An examination of Splunk’s customer metrics shows a very interesting situation.

Syslog Is Splunk’s “Interstate” ... and is Splunk’s dominant sourcetype



You will see that “syslog” is the number one source type and represents more than 40% of the incoming data. And here is the problem: syslog is a protocol, not a source type! It is the interstate upon which events from network devices flow. The network device events are the cars on that interstate, each having a different make, style, number of occupants and other identifying features.

The ramification (and cost) of this situation becomes apparent when the data is analyzed. Remember when we talked about the schema-on-the-fly operating on each kind of data? If that kind is simply “Interstate,” that will not work when analyzing the data together (e.g., Cisco IOS, Palo Alto PanOS, etc. in an Enterprise Security dashboard). The reason is that each device (source type) needs its own schema, as a single schema cannot account for all of the format variations from each vendor. Historically, the Technology Add-on (TA) author has been the one left holding the bag here.

Let’s look at a typical TA and examine where this expensive manual work takes place in the Splunk realm. Here is a section of the `props.conf` file from a very popular Cisco IOS TA:

```
[syslog]

TRANSFORMS-force_sourcetype_for_cisco_ios = force_sourcetype_for_cisco_ios, force_sourcetype_for_cisco_ios-xr, force_sourcetype_for_cisco_ios-xe
```

The `[syslog]` stanza (“Interstate”) directs each event (“car”) to be further processed by the listed stanzas in the `transforms.conf` file to look for an event match (just like the manual searching through the event detail in the example above). If the event matches one or more of these “force” stanzas:

```
[force_sourcetype_for_cisco_ios]

DEST_KEY = MetaData:SourceType

# This also gets process_name for IOS XE

REGEX=(?:(?:\S+)\s)?(?:?:(?:\d+)?\:\s(?:\.\S+\s)?(?:[\.\*])?(?:\.\s)?\:\s+(?:%|#)(?:?!POLICY_ENGINE|UCSM|FWSM|ASA|FTD|PIX|ACE)[A-Z0-9_]+)-(?:?:[A-Z012_]*(?:-[A-Z_][^-]*)-)?(?:?:[0-7])-(?:[A-Z0-9_]+):(?:?:[A-Za-z0-9_]+):)?\s(?:.+)
```

```
FORMAT = sourcetype::cisco:ios
```


The respective source type is assigned and the TA can now operate on that just-assigned source type. Now, imagine every TA looking through every syslog event using a similar approach to the Cisco example above to check whether or not it should even process the event at all. At hundreds of thousands of events per second, you can see how this gets very expensive, resource-wise, for Splunk infrastructure. Instead, imagine if this first step did not need to take place, and the source type of `cisco:ios` was simply handed to Splunk.

Best practice of tomorrow: Given how easy it is to configure SC4S, as well as the extensive device support, Splunk administrators don't have to choose between the difficulties of administering traditional syslog server software versus simply sending to Splunk. Effectively, with SC4S the administrator can send directly to Splunk and the final database will be sent to Splunk rather than the incomplete one that needs manual overhaul.

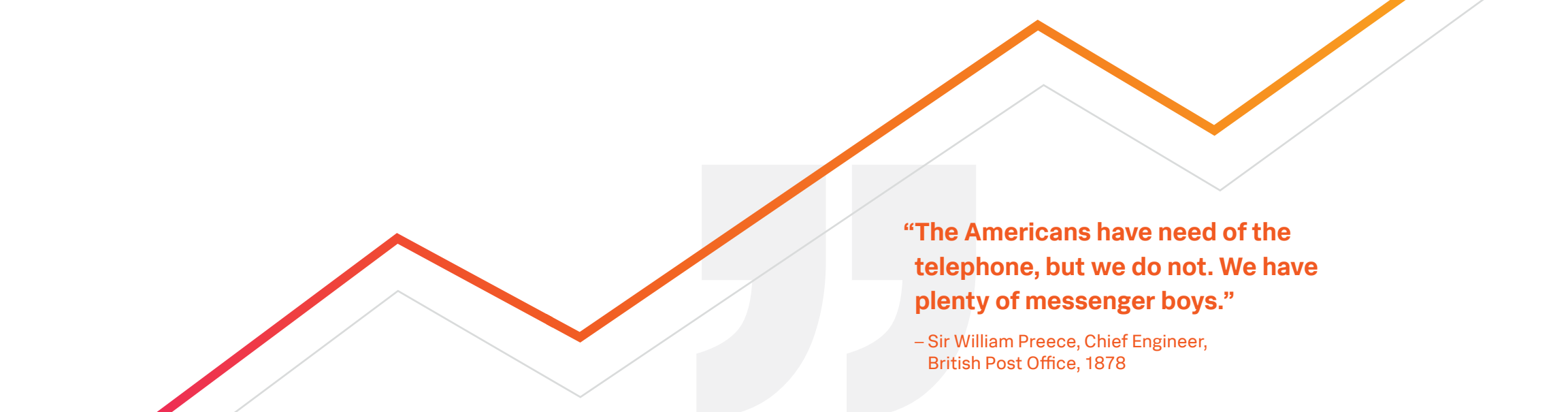
Old and simple

We can see that with a new best practice of tomorrow applied to syslog network design, coupled with the ease of use and scale of SC4S, we can address these decades-old issues with syslog data collection.

Splunk Connect for Syslog is the culmination of four years of effort to completely rethink how to collect network and security device data at scale. Hobbled by a 40-year old protocol, the latest devices process huge amounts of data, yet were mired in decades-old approaches to data collection. SC4S, along with a significant educational effort through many blogs, presentations and direct customer interaction, has ushered in a true best practice of tomorrow for syslog data collection that offers performance, scale, ease of use and maintainability — all in one simple container package.

Mark Bonsack

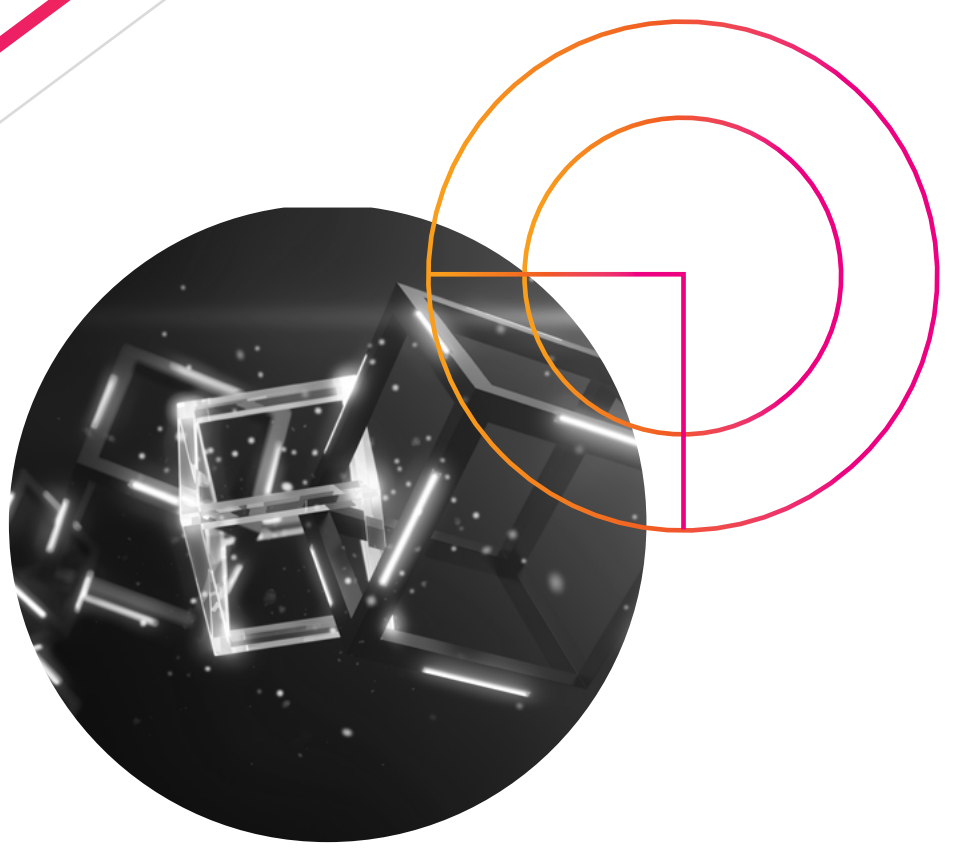
Mark Bonsack is a principal sales engineer at Splunk. During his nine-year Splunk career, he has developed a particular interest in data acquisition (aka "GDI") and has guided Splunk's largest customers in this area.



“The Americans have need of the telephone, but we do not. We have plenty of messenger boys.”

– Sir William Preece, Chief Engineer,
British Post Office, 1878

A Future We Can Trust



Like technologies of the past, many have said there is no need for blockchain and that it won't scale. There is always resistance to change, especially when incumbent technologies have been working “good enough.” For instance, horses were preferred over automobiles because they didn't get stuck in the mud. With email, many people didn't understand why anyone would want to type on a keyboard and dial up to the internet instead of making a phone call. Technology seems to be more broadly adopted only as the infrastructure behind it improves (e.g., roads and high-speed internet), or if it becomes clear that tools and their underlying legacy infrastructure can no longer keep up.

We're seeing this today as infrastructure for distributed ledger technology (DLT) — such as blockchain — becomes more widely used where legacy infrastructure can't keep up, even before the former reaches maturity. In this chapter, you will learn what the current use cases for DLT infrastructure are today and the biggest pains organizations are facing when adopting it.

But first, how is DLT making waves across industries? Simply put, blockchain and other distributed ledger technologies enable data-sharing among multiple parties in a way that can be automatically trusted and verified without human interference. Every use case below is based on this simple concept; multiple parties that don't necessarily trust each other can share data in an instantly verifiable manner without an intermediary. It may sound simple, but the ramifications are significant.

From increasing transparency across supply chains to detecting counterfeit medicine, blockchain does more than transform the way we share data. It transforms business deal making, counterfeit drug detecting, healthcare process making — and will continue to touch even more facets of our businesses and very lives for years to come. Splunk is embedded as a critical component in order to gain the necessary level of monitoring and observability to take action on data.

Managing supply chains

At first glance, supply chains seem boring — almost imperceptible. The general consumer orders something, it shows up in two days and no questions are asked. Consumers have no way to know the details of what it took to get a package to their door. Details like whether the product was manufactured in a sweatshop or whether it's authentic just can't be seen. This lack of visibility even permeates the steps involved in the procurement and distribution of goods. While it may seem good not to know the likely tedious intricacies of supply chains, their opacity creates massive inefficiencies that come with great costs.

It is clear that traceability and transparency of supply chains are major problems in logistics, and solving them can be a boon to any organization involved in the process. **Seventy-nine percent** of organizations with superior supply chain capabilities achieve significantly above-average revenue growth. And that's where DLT comes in, by enabling efficient and accurate logistics.

This benefit has never been more important. COVID-19 has already pushed supply chains past their breaking point as services and organizations across the board suffer under the strain of the pandemic. In 2020, some consumers could not even buy meat, toilet paper or other goods.

Imagine the complexities that COVID-19 vaccine distribution entails. It is the biggest supply chain challenge in history. Multiple organizations are manufacturing billions of vaccines that have to be transported across air and land while kept in ideal conditions to keep them viable. Not to mention that all parties involved want clear visibility into the what, when, where and how.

Without a distributed ledger, each party will store their version of the truth in their own system of record, even paper. An error or dispute could mean lost or destroyed vaccines, as the process of finding root causes and other backward tracing could take weeks or months.

Fortunately, with a shared distributed ledger, all parties can see the status, amounts, temperature ranges, customs documents and other data along the entire journey.

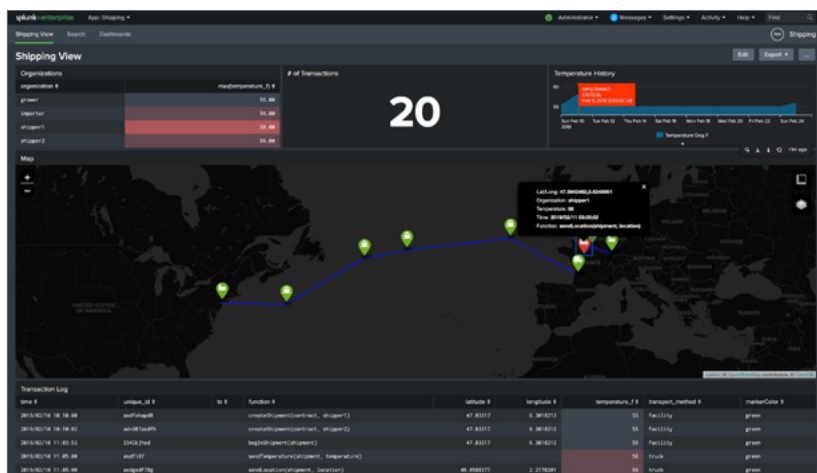


Figure 1: Supply chain visibility

So are there companies out there using distributed ledgers for their supply chains already? CISCO and DHL [lists their top 8 benefits](#) of their blockchain-based supply chain solution, and Microsoft and its partners are using a [Consensys Quorum](#) blockchain platform to improve the resiliency, traceability, and predictability of [their cloud hardware supply chain](#). Upon launch, the platform uncovered tens of millions of dollars in hidden costs that resulted from new levels of transparency into the responsible sourcing of critical or raw materials. It is a little-known fact that hardware manufacturers spend hundreds of millions of dollars and millions of person hours a year to trace their supply chains for conflict minerals — and even after all this, [90% of those companies](#) still couldn't confirm their products were conflict-free. In addition to reducing costs and increasing speed and efficiency, lives are being improved simply by enabling data to be shared in a verifiable way between parties that don't trust each other.

Impact: Consumers and businesses benefit from reduced waste, increased efficiencies, decreased costs, faster deliveries and more sustainably resourced goods. Not to mention that increased transparency can ensure better working conditions.

Let's make a deal

Like in any business, deals are made in supply chains. Let's look at the hypothetical situation of Acme Co. It was willing to pay \$1 each for 3 million face masks if they were delivered in three months. Four months later, only 2.5 million masks arrived and 300,000 of them didn't pass quality inspection. This caused all parties to go to their systems of record to start the long, tedious and expensive reconciliation process to answer questions such as: Where did the missing masks go? Which masks didn't pass inspection and why? What caused the delivery delay? Who is responsible for paying, and how much is due?

The problem with figuring out what went wrong was that the parties used different systems of record including databases, SAP, Microsoft Dynamics and (sadly) paper documents. Side agreements and updates were made along the way that were not updated across every system of record, and paperwork was lost. The only way to solve the issue was for every party to baseline what their system of record showed and then verify that it matched the other parties' records.

This situation illustrates the use case for the baseline protocol.

"The baseline protocol is an open source initiative that combines advances in cryptography, messaging, and blockchain to execute secure and private business processes at low cost via the public Ethereum Mainnet. The protocol enables confidential and complex collaboration between enterprises without leaving any sensitive data on-chain."

This means any organization can make deals and stay in sync without changing their backend systems such as their external resource provider (ERP), enabling multiple enterprises to share data and logic.

Impact: Organizations that can confidently synchronize their data in real time and empower themselves through their data insights will leap ahead of competition. The baseline protocol enables automation of business processes by leveraging data in new or traditional systems while maintaining integrity and confidentiality. This unlocks data in ways previously impossible, quite literally turning data into doing.

Fraud

Ten percent of all medicine globally is counterfeit, and 30% of drugs sold in Asia, Africa and Latin America are. For instance, 64% of antimalarial drugs in Nigeria were found to be counterfeit, [according to a recent study](#). The seriousness of this problem cannot be overstated. Counterfeit drugs cost the pharmaceutical industry \$100 billion a year and have killed hundreds of thousands of people. Current systems lack effective detection, tracing and removal of counterfeit drugs. The lack of a global shared database allows criminals to introduce counterfeits. To solve the problem, LedgerDomain has created a shared ledger that provides accurate inventory tracking, detects counterfeit assets and satisfies compliance requirements. It is a cloud-agnostic solution that uses Splunk on the ledger data with the application and infrastructure data, enabling debugging, as well as auditing the path of individual drugs and associated API calls.

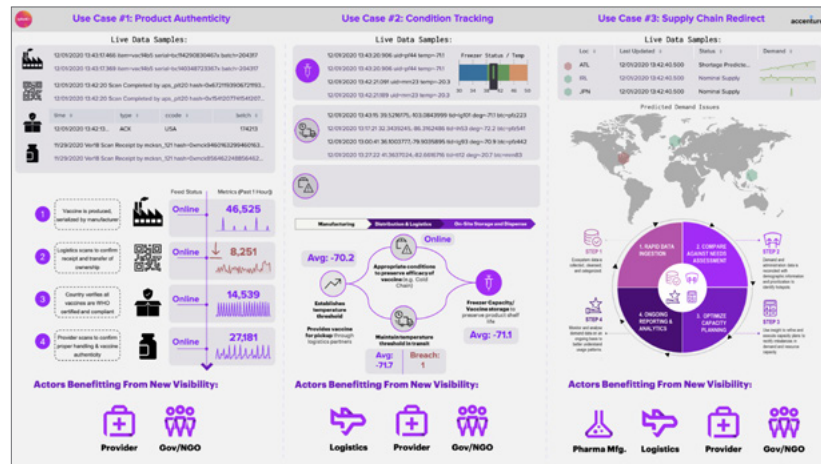


Figure 2: Example of Splunk use cases

Many vaccines require a booster shot from the same lot/batch and manufacturer. Vaccine administrators can verify that the vaccine is authentic and from the correct batch and manufacturer.

Impact: A distributed ledger where all parties can verify the data will save lives and prevent billions of dollars of fraud.

Fraud and risk in finance

Fraudsters are often early adopters of technology, and the media often portrays digital currencies as the new frontier for fraud and money laundering. It turns out that laundering with fiat has quite a high success rate — 99.9% of money-laundering enforcement fails. As longtime financial crime expert [Raymond Baker notes](#), “Total failure is just a decimal point away.”

Against common perception, digital currencies can actually enable vastly improved fraud detection because most distributed ledgers contain the full history of transactions. Organizations can index this data in Splunk and enrich it with external data to identify fraud, track down criminals and even proactively determine the risk of transacting with entities.

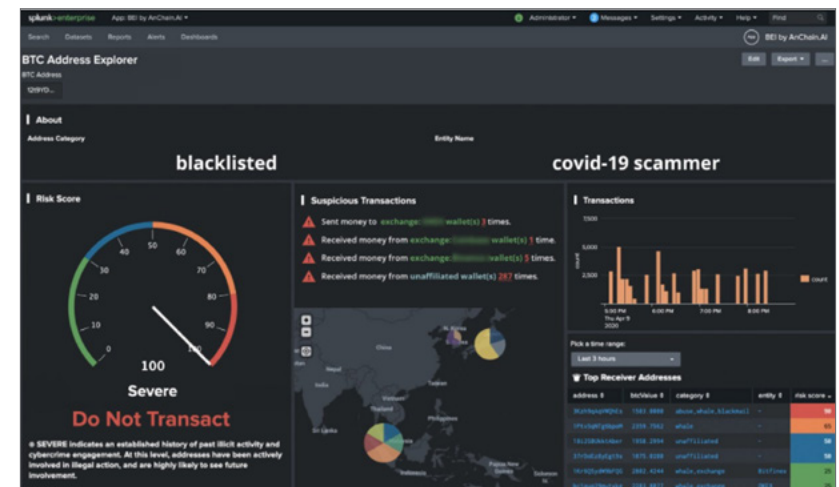


Figure 3: Proactively determine risk

In addition to decreasing fraud, digital payments can dramatically decrease the amount of time it takes for payments to process and the cost to do so.

“You need merely lodge your transactions in cyberspace. This will, of course, be illegal in many jurisdictions. But old laws seldom can resist new technology. In the 1980s, it was illegal in the United States to send a fax message. The U.S. Postal Service considered faxes to be first-class mail, over which the U.S. Post Office claimed an ancient monopoly. An edict to that effect was issued reiterating the requirement that all fax transmissions be routed to the nearest post office for delivery with regular mail. Billions of fax messages later, it was unclear whether anyone complied... The advantages of operating in the emerging cybereconomy are even more compelling than sidestepping the post office in sending a fax. Widespread adoption of public-key/private-key encryption technologies will soon allow many economic activities to be completed anywhere you please.”

The excerpt from above was published more than 20 years before the famous Bitcoin paper. It details how economic activities will become truly digital and that the speed and efficiency means organizations will be forced to adopt it or be left behind by those who do. Many businesses in the late 80s that only used the postal service were left in the dust compared to their competition that sent faxes. “Messenger boys” were good enough until they weren’t. Those that abided by the existing regulations of the time were, in a way, punished because they couldn’t compete. Similarly, the regulations from the Telecommunications Act of 1934 hindered the internet in the 1990s. Regulations are not bad, but as they age the regulations fit the past not the present. It would be hard to imagine a large, successful business today that uses mail or fax as their primary way to send information or payments. Without encryption, online payments would not be safe. Yet in the 1990s, it was still [illegal to export browsers with SSL encryption](#).

In 2021, we are in a similar scenario where the existing regulations aren’t quite ready to cope with digital assets. Some believe digital currencies (a type of digital asset) are inevitable. Dismissing digital currencies will come to be seen as short-sighted as dismissing automobiles for horses, faxes for snail mail and the internet for the telephone.

As [Dan Schulman, the president and CEO of PayPal said](#), “The shift to digital forms of currencies is inevitable, bringing with it clear advantages in terms of financial inclusion and access; efficiency, speed and resilience of the payments system; and the ability for governments to disburse funds to citizens quickly.”

Even central banks are rapidly developing or integrating digital currencies. Although a few countries have already launched their central bank digital currency (CBDC), this is still a use case of the future because of the need for proper regulation while engineering continues to design. Some countries are directly creating and releasing CBDCs while others believe the market can create the best digital currency. The controller of the United States points out that this is analogous to telecommunications, where government specific regulations, such as what spectrums are allowed, and the industry builds phones around the regulations rather than the government building a phone. The “crypto” industry will create new forms of digital currency such as U.S. dollar-backed [USD Coin](#) as well as algorithmic stablecoins such as [Dai](#). The government can specify “know your customer/client” (KYC) and anti-money laundering (AML) policies and other rules, and many of these rules can be embedded into the currencies themselves. While creating a basket of these stablecoins can increase design robustness, more consideration and testing is needed before this use case becomes widespread.

Cross-border payment regulations and engineering designs are better understood and are already benefiting from distributed ledger technology. These payments are a \$125 trillion global money transfer market of which 80% of cross-border payments are B2B, accounting for \$100 billion in revenues. Often these transactions take several days to complete as they are routed through intermediary banks. Distributed ledgers enable payment providers to streamline global business payments in a secure and predictable manner where transactions complete in moments rather than days at much lower costs. If international payments are gaining the ability to transact faster with fewer fees why not have systems such as the automated clearing house (ACH)?

The Federal Reserve Banks [are upgrading to a new service](#) to support faster payments in the United States. Today, the Fedwire system rings similar to the days before the DTCC (Depository Trust Clearing Corporation). Processing hours are limited from 8:30 a.m. to 6 p.m. ET and exclude weekends and holidays. The effect of these limitations is analogous to when stock certificates had to be delivered by human messengers. The volume of transactions increased and the stock certificates piled up on tables. Many were delivered to wrong addresses, or not at all. Overwork night became mandatory, followed by stock exchanges closing every Wednesday to catch up; even the trading hours were shortened. In 1973, a temporary measure was developed where shareholders would cede their ownership and enable clearing through centralized intermediaries. This temporary measure enabled decades of success in markets; the DTCC processes over \$2 quadrillion in securities per year. DTCC believes [distributed ledgers represent a generational opportunity](#) for post-trade infrastructure. Adopting the technology has a number of benefits, such as 24/7, year-round operations instead of the limited hours and days we have been accustomed to for decades.

Impact: Low cost, low latency, greater transparency transactions with less fraud and the addition of micro-transactions will benefit every nation, corporation and individual by letting value flow as freely as information on the internet does today.

Healthcare, insurance and more

The U.S. healthcare system wastes a staggering [\\$1 trillion annually](#). A more efficient system will both reduce costs and improve customer experience. The major crux in healthcare has been sharing data while maintaining trust. While overall digitization has increased, outcomes lag in part due to the lack of interoperability and technological alignment with people and processes.

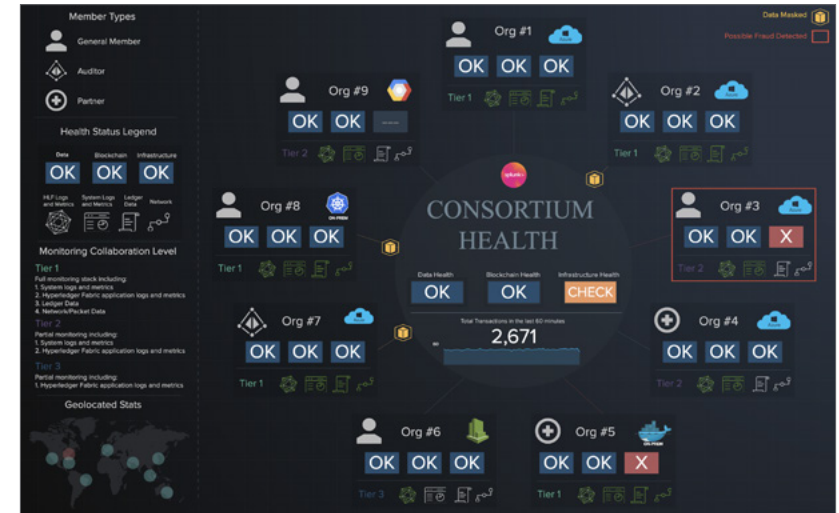


Figure 4: Splunk helps manage operational health

There are many pioneering efforts to improve upon these challenges especially revolving around data access and sharing. One such example is the health insurance company Anthem, which is using blockchain technology to keep patient information private and secure, while allowing it to be shared on demand to a specified audience. Data access can then be revoked after its purpose has been met. The goal is to provide trust and usability of data among patients, insurance companies and providers. Anthem is using Hyperledger technologies such as Hyperledger Fabric and Hyperledger Indy to serve [all 40 million of its members](#).

While blockchain technology can play a role in solving some of these complex problems, the infrastructure, systems and applications powering the solutions require insights into their performance and activity. Interoperability between different technologies and organizations further reduces visibility across the solutions. Splunk is helping customers gain full observability into their blockchain solutions by providing insights into logs, metrics, traces and the distributed ledger data all in one place.

Impact: The future of healthcare is about helping with decision-making to improve outcomes and wellness using analytics and big data. Splunk and blockchain can help remove the barriers between data and action so that healthcare can thrive.

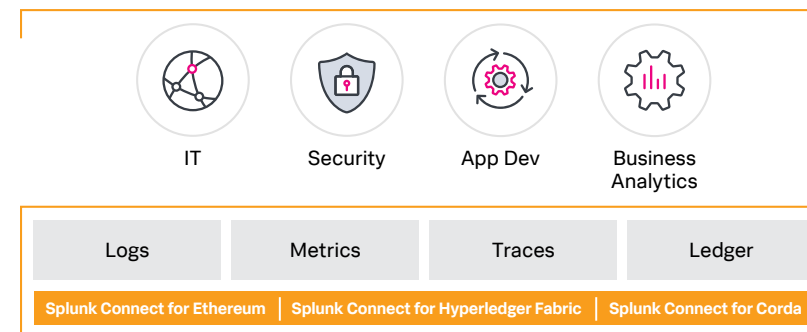
No gain without some pain

Multiparty systems break down the trust gap by enabling organizations to share data and auto-execute decisions nearly instantly in a verifiable manner. From banking the unbanked to drastically increasing supply chain efficiency and decreasing fraud, distributed ledger technology holds immense promise. But as with most emerging technology, enterprises have found it painful to adapt, due to the following factors:

- Difficult to gain end-to-end visibility among all components
- Many options to choose for infrastructure including on-premises, cloud, hybrid, unmanaged, managed and multi-organization
- Diverse set of data sources with differing formats and velocities
- Different blockchain technologies and platforms
- Interoperability and collaboration between consortiums
- Disparate tools for logging, metrics, tracing, transaction analytics and security
- Different tools (if any) for load testing and development performance versus production monitoring and investigation

The organizations that successfully use data combine logs, metrics, traces and ledger data into one platform to gain superior observability. They are able to break down silos among data sources, infrastructure providers, organizations and operators.

Blockchain Observability



The IT team uses the data to prevent downtime, such as downtime due to failed transactions. Security can thwart attacks, detect and prevent fraud. Application developers can measure the performance of their latest releases and see how smart contracts are performing. Business analytics teams can identify the least efficient and most expensive parts of the supply chain. It's the same data, just different lenses.

The logs, metrics and traces can all be implemented using the [Splunk Observability suite](#). Depending on the ledger, there are specific open source Splunk connectors that get the ledger data (e.g., transactions and metadata, etc.) such as [Splunk Connect for Ethereum](#) and [Splunk Connect for Hyperledger Fabric](#).

Blockchain will seem to disappear

As Mark Weiser predicted of the 21st century computer back in 1991, “The **most profound technologies** are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.”

Blockchain is not an application, it is a set of protocols with a new model of achieving consensus that further enables technologies such as AI, IoT and cloud. As blockchain technology matures it will become more of a given. Today it is being applied to many use cases beyond those included in this chapter, and in time no one will say they have a blockchain-enabled application just as no one says they have an internet-enabled application — they say “check out this app”.

Just as Sir William Preece said they don’t need telephones because they have plenty of messengers, many organizations will say the same about multiparty communications enabled through distributed ledger technology. Organizations that do not have a way to share data in a trusted and verifiable manner will be outperformed by those that do.

Nate McKervey

As head of blockchain and distributed ledger technology at Splunk, Nate McKervey leads the product strategy and development of distributed ledger technologies. Previously he ran technical marketing to help drive the value Splunk creates by creating compelling product demos and narratives, influencing industry analysts and media, presenting on stage at events and creating technical thought leadership content.

Stephen Luedtke

Stephen Luedtke started his career as a network and systems engineer for one of the world’s largest and most sophisticated communication networks, the FAA’s telecommunications infrastructure. After seven years and catching the data analytics bug, he joined Splunk as a professional services consultant, working with clients to design and implement data analytics solutions before moving to technical product marketing. He recently joined the blockchain team as a data engineer.



Launch into the future

Visit splunk.com to learn more.

Contact Us

