Open Access Library Journal
2024, Volume 11, e11496
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Independent Study of Splunk

Xueying Pan

Department of Computer Science & Engineering, Oakland University, Rochester, USA
Email: ivypan89@gmail.com

How to cite this paper: Pan, X.Y. (2024) Independent Study of Splunk. *Open Access Library Journal*, **11**: e11496.
http://doi.org/10.4236/oalib.1111496

Received: March 28, 2024
Accepted: April 25, 2024
Published: April 28, 2024

Abstract

The paper's objective is to easily search data and integrate all data sources or tools into one place for people to identify issues in visualizable ways based on correlating multiple data sources. On the other hand, we would use Splunk to build customized dashboards depending on critical success factors (CSF) and critical-to-quality (CTQ) from a single "pane of glass" that gives us a powerful search engine in root cause analysis, data analytics, and integration of multiple logs. In the paper, we introduce various methods to integrate all data sources or tools into one place for authorized users to access and view all of them from a single screen. Typical dashboards are designed based on monitoring log files, viewing the trend of hung threads of a server, or tracking recent changes and critical incidents. Furthermore, we offer customizable dashboard functionality for different technical departments to smoothly work on their complex daily tasks. To analyze huge data results from the Splunk searching tool, we could annotate the data stream with metadata keys including host, server, source, source type, and index. However, some limitations and disadvantages are in the Splunk tool. Therefore, we provide different scenarios that could make Splunk run slowly. Then, we not only discuss what root causes exist in Splunk itself or inside of companies themselves but also describe what aspects of Splunk still need to be improved. Finally, we could take advantage of Splunk to build various functional dashboards to get a quick view of overall system health, application performance, and end-user ramifications for fulfilling business purposes. Additionally, we summarize beneficiations using Splunk and discuss current related works on Splunk tool.

Subject Areas

Applications of Communication Systems, Information Management, Technology

Keywords

Splunk, Customizable Dashboard, Data Source, Database, System Health, Performance

1. Introduction

As database management is growing day by day, most people want to find a tool to be a powerful search engine to has features including root cause analysis capability, data analytics, and integration of multiple logs for server data. At the same time, they want to find a way of an effective tool to help identify various technical issues in a quicker way, which could be a way of correlating multiple data sources. Their main purpose is to easily search data by a simple query and integrate all data sources or tools into one place for authorized users who can access all of them from a single screen. On the other hand, using this tool is not only to build a high-end dashboard through which people can monitor critical success factors and critical quality from a single screen but also to track recent changes and critical incidents. As a result, the Splunk engine tool is a good tool for most people to design and create customizable dashboards for monitoring kind of servers in different environments. Additionally, they could proactively investigate notable events for alert creations when a metric falls below the SLA/KPI for specific services.

Former researchers may face challenges in detecting and interpreting security threats amidst a large volume of log information because they need to stay updated on evolving attack techniques and security trends. It is difficult to correlate log information from various sources to identify the root cause of technical issues since advanced analytics techniques are needed to predict and prevent future system failures. For analyzing business-related logs such as website traffic, sales transactions, and customer interactions, former researchers may deal with data quality issues in missing and inconsistent data that would affect the accurate analysis. The challenge in integrating data from multiple sources is to gain a comprehensive understanding of business operations. There may be a privacy and security issue for previous researchers who want to analyze healthcare data such as electronic health records (EHRs), medical device logs, and patient monitoring data. To identify disease outbreaks, optimize healthcare operations, and improve patient care, they have to develop advanced analytics models to extract complex and heterogeneous healthcare data. Development of scalable and efficient data processing pipelines is difficult in monitoring environmental conditions, and improving efficient operations when previous researchers handle large volumes of streaming data generated by Internet of Things (IoT) devices such as smart appliances, sensors, and actuators in real-time.

2. Design a Dashboard Based on Log Files to Help Identify Processes are Updated or Not

To help identify whether system processes are updated or not, we design a dashboard to monitor local files such as syslog files, and config files. It is a systematic way to bring new data sources into Splunk and then break them into events based on timestamps.

For dashboard design, firstly we need to understand specific requirements including monitoring processes, updating constitution to process, and information of log files as indicators of update. Data collection and ingestion is the second step to collect and index the log files including the relevant data about the processes an end-user wants to monitor. Then, the user needs to make sure the structure facilitates easy extraction and analysis of log data. Building search and query is to identify changes or updates to defined processes based on the criteria in the defined requirements and to extract the relevant information from the log data when writing Splunk queries. Data visualization is to design charts, graphs, and tables for effectively communicating the data extracted from the log files. To identify trends and patterns, we would use line charts for trends over time or pie charts for distributing updates across processes. Visualization needs to be organized on the dashboard in a logical and intuitive manner. To design the layout of the dashboard, we need to consider specific needs and target audiences and then make sure the dashboard is easily navigated and understood. Dashboard capabilities would be used for creating actual dashboard based on the designed layout. Configured visualization would reflect the output of the search queries and update in real-time as new ingested log information. Dashboard testing is to accurately reflect process status and identify corrected updates. We would make sure designed dashboard that meets users' needs from the validation process. Depending on targeted users or stakeholders, we would deploy the dashboard and monitor it regularly so that the function is kept correctly. Throughout the design process, we would use users' feedback and changing requirements to iterate and refine the dashboard to make sure it remains effective for identifying updated processes. Splunk is not only a purchased product for big enterprises but also a powerful platform that collects and indexes server data from virtually any source in real-time. There are kinds of data which could be syslog, tools logs, any type of custom logs, server access logs, etc. We could use Splunk to give us a view of all this server data in one timeline to identify whether system processes are updated or not. Finally, we could report on, analyze, save, and export this data, and share it with other related IT departments.

For searching specific server data, we could start typing like any search engine, add a clicked term to the search bar, and exclude a clicked term from the search by pressing Alt + Click combination buttons, etc. as basic search methods in Splunk. Moreover, relative time, real-time, and custom time are three-time pickers in Splunk for different searching purposes. (See Figure 1)

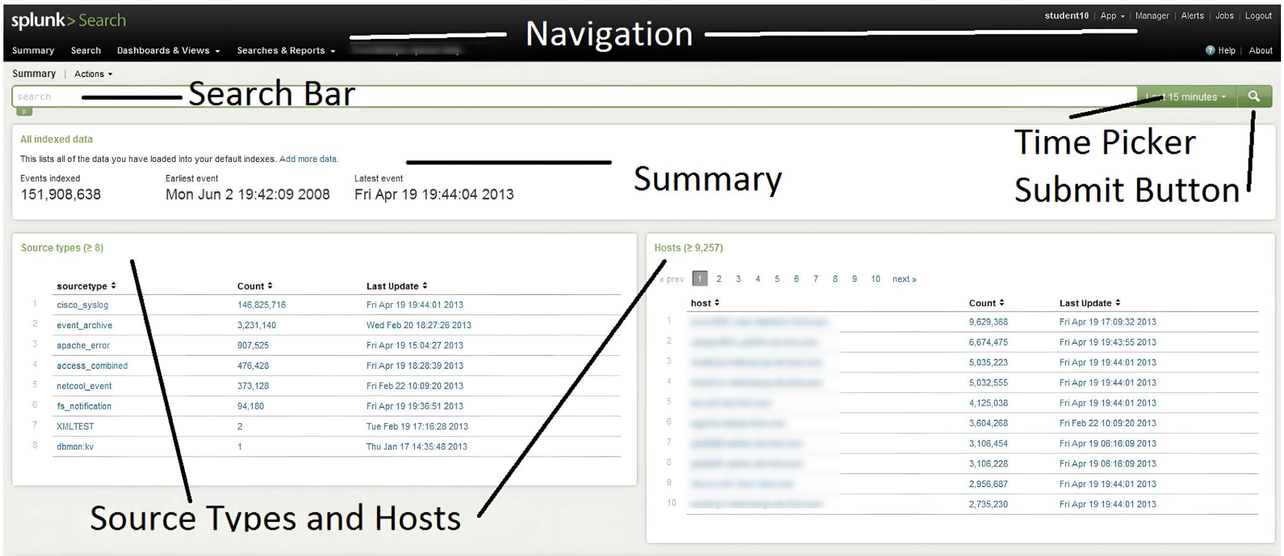
3. Design a Dashboard Based on Viewing the Trend of Hung Threads of a Server

To detect hung threads for Java virtual machine, we try to design a dashboard in

Splunk based on analysis of the Java thread dump and the WebSphere SystemOut.log files. (See Figures 2-6) Since a hung thread is a thread that could be blocked by a blocking call. We could use Splunk to receive output messages from SystemOut.log files which simply indicate that a thread may be hung or a previously reported hung thread completed its work. [1]

4. Design a Dashboard Based on Tracking the Recent Changes and Critical Incidents

To track recent changes and critical incidents for production servers, we specifically designed a dashboard in Splunk that must follow a common information



All indexed data

This lists all of the data you have loaded into your default indexes. Add more data.

Events indexed	Earliest event	Latest event
151,908,638	Mon Jun 2 19:42:09 2008	Fri Apr 19 19:44:04 2013

Source types (≥ 8)

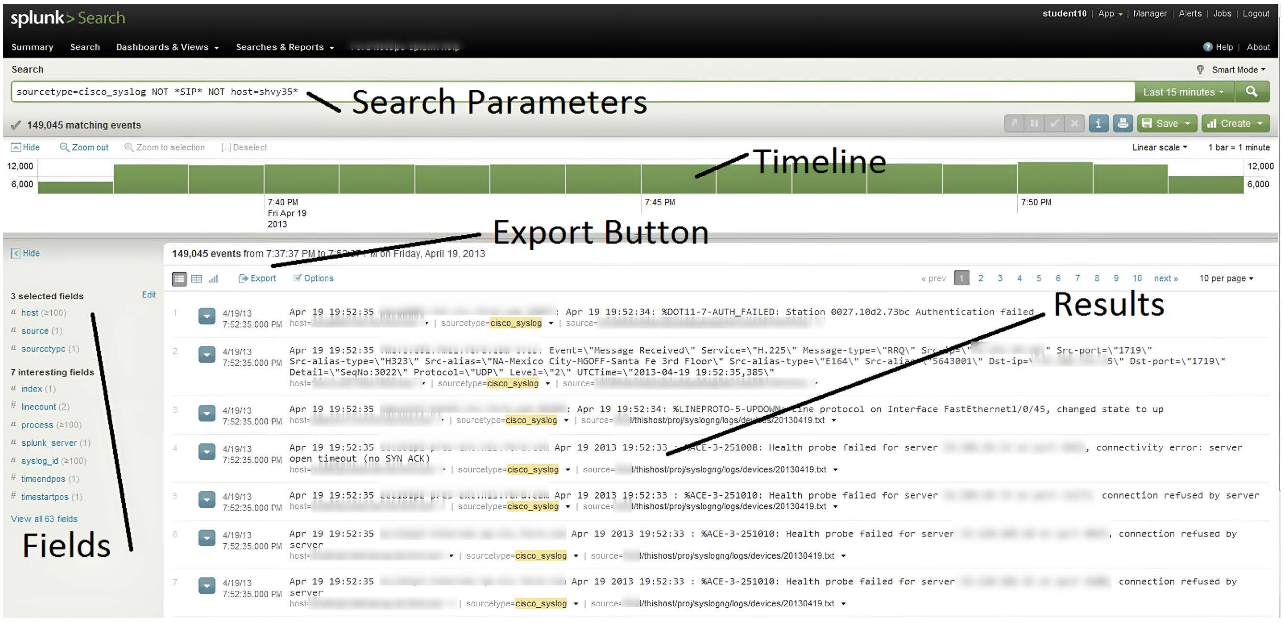
	sourcetype	Count	Last Update
1	cisco_syslog	146,825,716	Fri Apr 19 19:44:01 2013
2	event_archive	3,231,140	Wed Feb 20 18:27:26 2013
3	apache_error	907,525	Fri Apr 19 15:04:27 2013
4	access_combined	476,428	Fri Apr 19 18:28:39 2013
5	netcool_event	373,128	Fri Feb 22 10:09:20 2013
6	fs_notification	94,180	Fri Apr 19 19:36:51 2013
7	XMLTEST	2	Tue Feb 19 17:16:28 2013
8	dbmon:kv	1	Thu Jan 17 14:35:48 2013

Hosts (≥ 9,257)

« prev 1 2 3 4 5 6 7 8 9 10 next »

	host	Count	Last Update
1	[redacted]	9,629,368	Fri Apr 19 17:09:32 2013
2	[redacted]	6,674,475	Fri Apr 19 19:43:55 2013
3	[redacted]	5,035,223	Fri Apr 19 19:44:01 2013
4	[redacted]	5,032,555	Fri Apr 19 19:44:01 2013
5	[redacted]	4,125,038	Fri Apr 19 19:44:01 2013
6	[redacted]	3,604,268	Fri Feb 22 10:09:20 2013
7	[redacted]	3,106,454	Fri Apr 19 06:16:09 2013
8	[redacted]	3,106,228	Fri Apr 19 06:16:09 2013
9	[redacted]	2,956,687	Fri Apr 19 19:44:01 2013
10	[redacted]	2,735,230	Fri Apr 19 19:44:01 2013

Figure 1. Search page in the Splunk.


splunk> Search
student10 | App ▾ | Manager | Alerts | Jobs | Logout
Summary Search Dashboards & Views ▾ Searches & Reports ▾
Help | About
Search
Smart Mode ▾
Last 15 minutes ▾
Save ▾
Create ▾
Hide Zoom out Zoom to selection Deselect
Linear scale ▾ 1 bar = 1 minute
12,000
6,000
7:40 PM
Fri Apr 19
2013
7:45 PM
7:50 PM
Hide
Export Options
« prev 1 2 3 4 5 6 7 8 9 10 next » 10 per page ▾
Search Parameters
Timeline
Export Button
Results
Fields

```
sourcetype=cisco_syslog NOT *SIP* NOT host=shvy35*
```

✓ 149,045 matching events

149,045 events from 7:37:37 PM to 7:52:37 PM on Friday, April 19, 2013

3 selected fields Edit

- host (≥100)
- source (1)
- sourcetype (1)

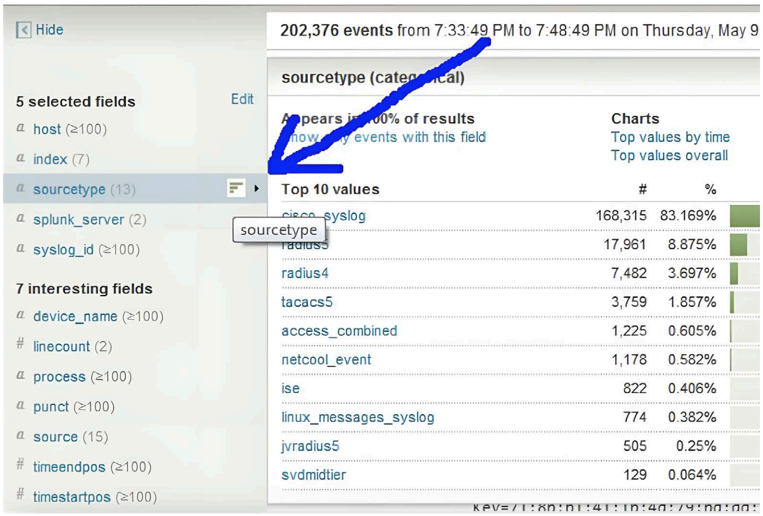
7 interesting fields

- index (1)
- linecount (2)
- process (≥100)
- splunk_server (1)
- syslog_id (≥100)
- timeendpos (1)
- timestartpos (1)

View all 63 fields

```
1  4/19/13           Apr 19 19:52:35 [redacted]: Apr 19 19:52:34: %DOT11-7-AUTH_FAILED: Station 0027.10d2.73bc Authentication failed
   7:52:35.000 PM    host=[redacted] | sourcetype=cisco_syslog | source=[redacted]

2  4/19/13           Apr 19 19:52:35 [redacted] Event=\"Message Received\" Service=\"H.225\" Message-type=\"RRQ\" Src-ip=\"[redacted]\" Src-port=\"1719\"
   7:52:35.000 PM    Src-alias-type=\"H323\" Src-alias=\"NA-Mexico City-MGOFF-Santa Fe 3rd Floor\" Src-alias-type=\"E164\" Src-alias=\"5643001\" Dst-ip=\"[redacted]5\" Dst-port=\"1719\"
                     Detail=\"SeqNo:3022\" Protocol=\"UDP\" Level=\"2\" UTCTime=\"2013-04-19 19:52:35,385\"
                     host=[redacted] | sourcetype=cisco_syslog | source=[redacted]

3  4/19/13           Apr 19 19:52:35 [redacted]: Apr 19 19:52:34: %LINEPROTO-5-UPDOWN: Line protocol on Interface FastEthernet1/0/45, changed state to up
   7:52:35.000 PM    host=[redacted] | sourcetype=cisco_syslog | source=[redacted]/thishost/proj/syslogng/logs/devices/20130419.txt

4  4/19/13           Apr 19 19:52:35 [redacted] Apr 19 2013 19:52:33 : %ACE-3-251008: Health probe failed for server [redacted], connectivity error: server
   7:52:35.000 PM    open timeout (no SYN ACK)
                     host=[redacted] | sourcetype=cisco_syslog | source=[redacted]/thishost/proj/syslogng/logs/devices/20130419.txt

5  4/19/13           Apr 19 19:52:35 [redacted] Apr 19 2013 19:52:33 : %ACE-3-251010: Health probe failed for server [redacted] connection refused by server
   7:52:35.000 PM    host=[redacted] | sourcetype=cisco_syslog | source=[redacted]/thishost/proj/syslogng/logs/devices/20130419.txt

6  4/19/13           Apr 19 19:52:35 [redacted] Apr 19 2013 19:52:33 : %ACE-3-251010: Health probe failed for server [redacted], connection refused by
   7:52:35.000 PM    server
                     host=[redacted] | sourcetype=cisco_syslog | source=[redacted]/thishost/proj/syslogng/logs/devices/20130419.txt

7  4/19/13           Apr 19 19:52:35 [redacted] Apr 19 2013 19:52:33 : %ACE-3-251010: Health probe failed for server [redacted] connection refused by
   7:52:35.000 PM    server
                     host=[redacted] | sourcetype=cisco_syslog | source=[redacted]/thishost/proj/syslogng/logs/devices/20130419.txt
```

Figure 2. Search page in the Splunk.



Hide

202,376 events from 7:33:49 PM to 7:48:49 PM on Thursday, May 9

5 selected fields Edit
- host (≥100)
- index (7)
- sourcetype (13)
- splunk_server (2)
- syslog_id (≥100)

7 interesting fields
- device_name (≥100)
- linecount (2)
- process (≥100)
- punct (≥100)
- source (15)
- timeendpos (≥100)
- timestartpos (≥100)

sourcetype (categorical)

Appears in 100% of results
Show only events with this field

Charts
Top values by time
Top values overall

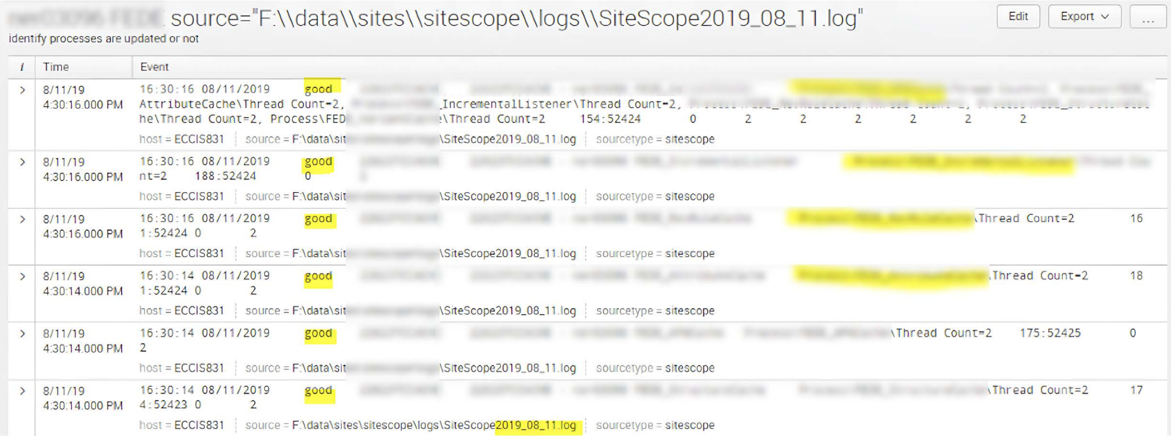
sourcetype

Top 10 values

Top 10 values	#	%
cisco_syslog	168,315	83.169%
radius3	17,961	8.875%
radius4	7,482	3.697%
tacacs5	3,759	1.857%
access_combined	1,225	0.605%
netcool_event	1,178	0.582%
ise	822	0.406%
linux_messages_syslog	774	0.382%
jvradius5	505	0.25%
svdmidtier	129	0.064%

key=71:8b:b1:41:1b:4d:79:bd:dd:

Figure 3. Search page in the Splunk.



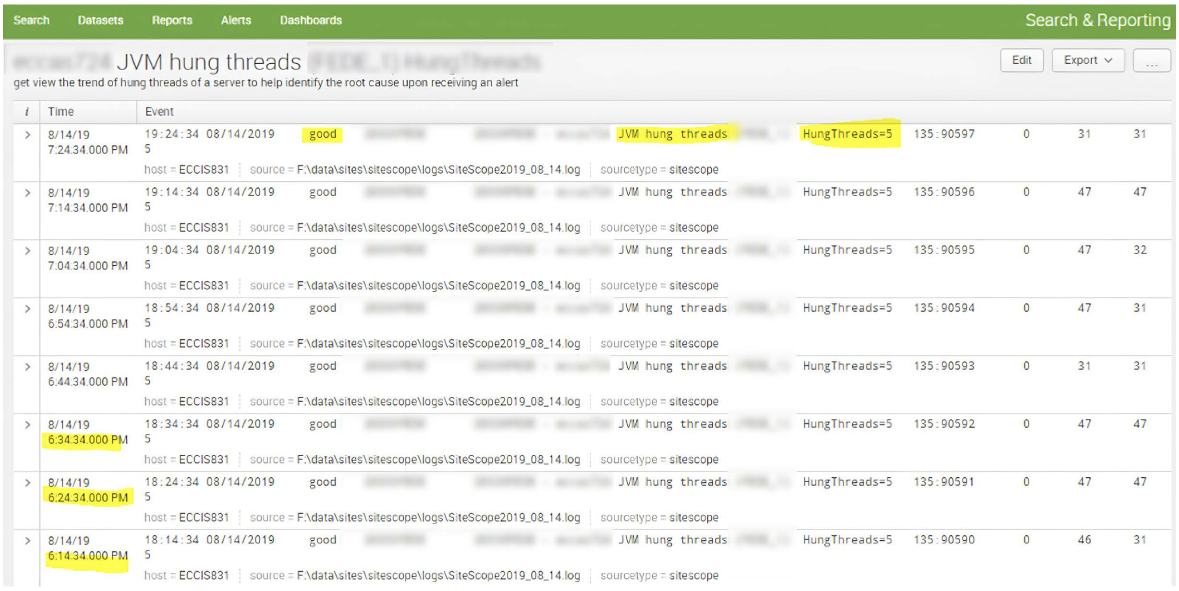
[redacted] source="F:\\data\\sites\\sitescope\\logs\\SiteScope2019_08_11.log"

identify processes are updated or not

Edit Export ...

```
i  Time                 Event
>  8/11/19              16:30:16 08/11/2019   good   [redacted]
   4:30:16.000 PM       AttributeCache\Thread Count=2, [redacted]_IncrementalListener\Thread Count=2, [redacted]
                        he\Thread Count=2, Process\FED[redacted]\Thread Count=2    154:52424    0    2    2    2    2    2    2
                        host = ECCIS831 | source = F:\data\sit[redacted]\SiteScope2019_08_11.log | sourcetype = sitescope
>  8/11/19              16:30:16 08/11/2019   good   [redacted]
   4:30:16.000 PM       nt=2     188:52424    0
                        host = ECCIS831 | source = F:\data\sit[redacted]\SiteScope2019_08_11.log | sourcetype = sitescope
>  8/11/19              16:30:16 08/11/2019   good   [redacted]\Thread Count=2     16
   4:30:16.000 PM       1:52424 0      2
                        host = ECCIS831 | source = F:\data\sit[redacted]\SiteScope2019_08_11.log | sourcetype = sitescope
>  8/11/19              16:30:14 08/11/2019   good   [redacted]\Thread Count=2     18
   4:30:14.000 PM       1:52424 0      2
                        host = ECCIS831 | source = F:\data\sit[redacted]\SiteScope2019_08_11.log | sourcetype = sitescope
>  8/11/19              16:30:14 08/11/2019   good   [redacted]\Thread Count=2     175:52425     0
   4:30:14.000 PM       2
                        host = ECCIS831 | source = F:\data\sit[redacted]\SiteScope2019_08_11.log | sourcetype = sitescope
>  8/11/19              16:30:14 08/11/2019   good   [redacted]\Thread Count=2     17
   4:30:14.000 PM       4:52423 0      2
                        host = ECCIS831 | source = F:\data\sites\sitescope\logs\SiteScope2019_08_11.log | sourcetype = sitescope
```

Figure 4. Customizable dashboard based on log files.



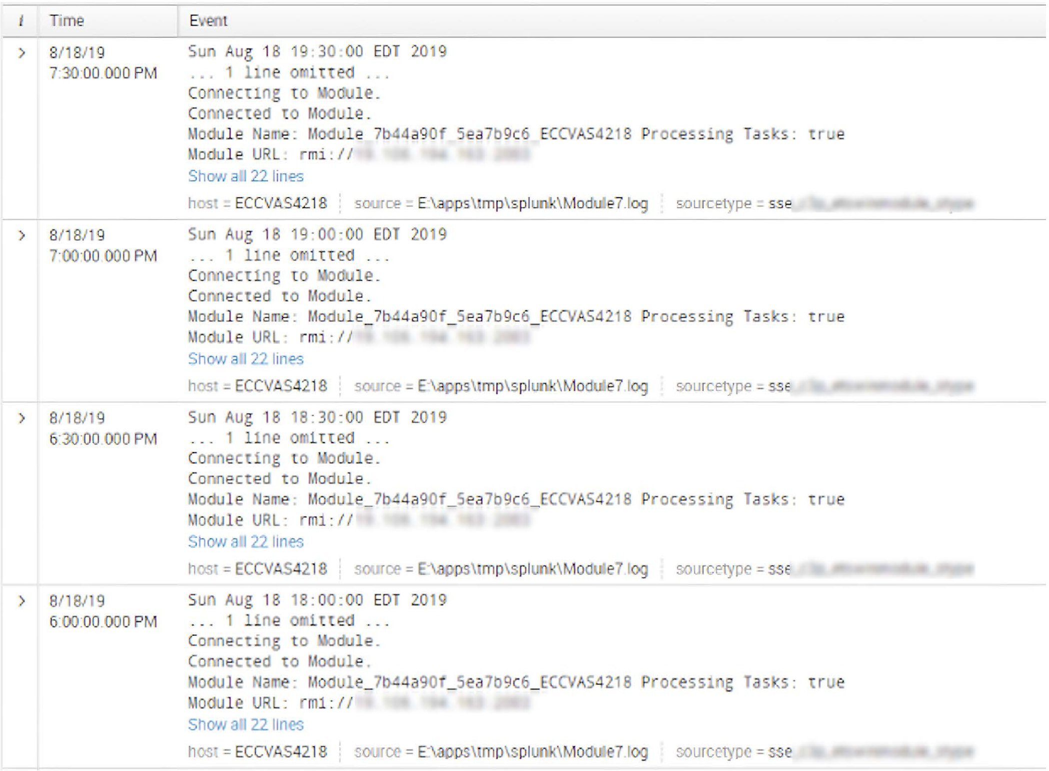
[redacted] JVM hung threads [redacted]

get view the trend of hung threads of a server to help identify the root cause upon receiving an alert

Edit Export ...

```
i  Time                 Event
>  8/14/19              19:24:34 08/14/2019   good   [redacted]   JVM hung threads [redacted]   HungThreads=5   135:90597   0   31   31
   7:24:34.000 PM       5
                        host = ECCIS831 | source = F:\data\sites\sitescope\logs\SiteScope2019_08_14.log | sourcetype = sitescope
>  8/14/19              19:14:34 08/14/2019   good   [redacted]   JVM hung threads [redacted]   HungThreads=5   135:90596   0   47   47
   7:14:34.000 PM       5
                        host = ECCIS831 | source = F:\data\sites\sitescope\logs\SiteScope2019_08_14.log | sourcetype = sitescope
>  8/14/19              19:04:34 08/14/2019   good   [redacted]   JVM hung threads [redacted]   HungThreads=5   135:90595   0   47   32
   7:04:34.000 PM       5
                        host = ECCIS831 | source = F:\data\sites\sitescope\logs\SiteScope2019_08_14.log | sourcetype = sitescope
>  8/14/19              18:54:34 08/14/2019   good   [redacted]   JVM hung threads [redacted]   HungThreads=5   135:90594   0   47   31
   6:54:34.000 PM       5
                        host = ECCIS831 | source = F:\data\sites\sitescope\logs\SiteScope2019_08_14.log | sourcetype = sitescope
>  8/14/19              18:44:34 08/14/2019   good   [redacted]   JVM hung threads [redacted]   HungThreads=5   135:90593   0   31   31
   6:44:34.000 PM       5
                        host = ECCIS831 | source = F:\data\sites\sitescope\logs\SiteScope2019_08_14.log | sourcetype = sitescope
>  8/14/19              18:34:34 08/14/2019   good   [redacted]   JVM hung threads [redacted]   HungThreads=5   135:90592   0   47   47
   6:34:34.000 PM       5
                        host = ECCIS831 | source = F:\data\sites\sitescope\logs\SiteScope2019_08_14.log | sourcetype = sitescope
>  8/14/19              18:24:34 08/14/2019   good   [redacted]   JVM hung threads [redacted]   HungThreads=5   135:90591   0   47   47
   6:24:34.000 PM       5
                        host = ECCIS831 | source = F:\data\sites\sitescope\logs\SiteScope2019_08_14.log | sourcetype = sitescope
>  8/14/19              18:14:34 08/14/2019   good   [redacted]   JVM hung threads [redacted]   HungThreads=5   135:90590   0   46   31
   6:14:34.000 PM       5
                        host = ECCIS831 | source = F:\data\sites\sitescope\logs\SiteScope2019_08_14.log | sourcetype = sitescope
```

Figure 5. Hung threads dashboards for JVM.

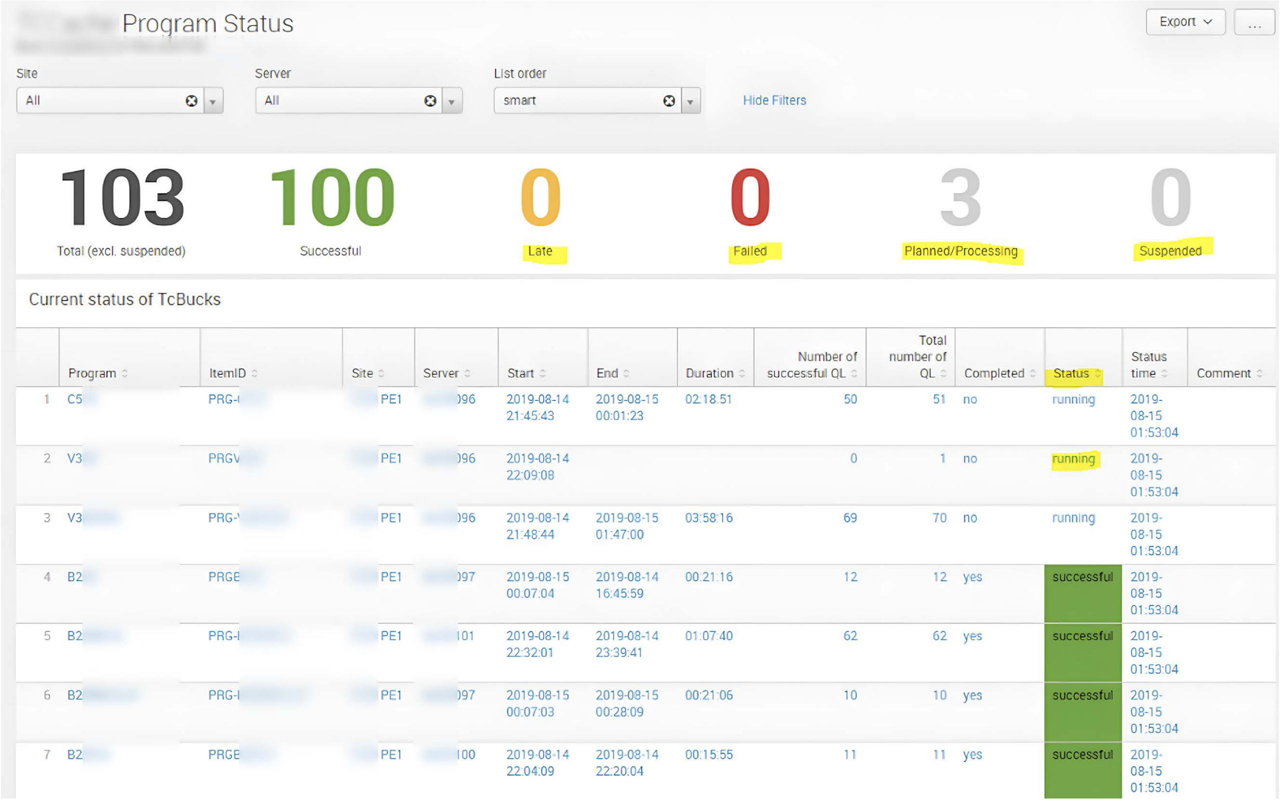


```
i  Time                  Event
>  8/18/19               Sun Aug 18 19:30:00 EDT 2019
   7:30:00.000 PM        ... 1 line omitted ...
                         Connecting to Module.
                         Connected to Module.
                         Module Name: Module_7b44a90f_5ea7b9c6_ECCVAS4218 Processing Tasks: true
                         Module URL: rmi://[redacted]
                         Show all 22 lines
                         host = ECCVAS4218   source = E:\apps\tmp\splunk\Module7.log   sourcetype = sse[redacted]
>  8/18/19               Sun Aug 18 19:00:00 EDT 2019
   7:00:00.000 PM        ... 1 line omitted ...
                         Connecting to Module.
                         Connected to Module.
                         Module Name: Module_7b44a90f_5ea7b9c6_ECCVAS4218 Processing Tasks: true
                         Module URL: rmi://[redacted]
                         Show all 22 lines
                         host = ECCVAS4218   source = E:\apps\tmp\splunk\Module7.log   sourcetype = sse[redacted]
>  8/18/19               Sun Aug 18 18:30:00 EDT 2019
   6:30:00.000 PM        ... 1 line omitted ...
                         Connecting to Module.
                         Connected to Module.
                         Module Name: Module_7b44a90f_5ea7b9c6_ECCVAS4218 Processing Tasks: true
                         Module URL: rmi://[redacted]
                         Show all 22 lines
                         host = ECCVAS4218   source = E:\apps\tmp\splunk\Module7.log   sourcetype = sse[redacted]
>  8/18/19               Sun Aug 18 18:00:00 EDT 2019
   6:00:00.000 PM        ... 1 line omitted ...
                         Connecting to Module.
                         Connected to Module.
                         Module Name: Module_7b44a90f_5ea7b9c6_ECCVAS4218 Processing Tasks: true
                         Module URL: rmi://[redacted]
                         Show all 22 lines
                         host = ECCVAS4218   source = E:\apps\tmp\splunk\Module7.log   sourcetype = sse[redacted]
```

Figure 6. Designed dashboard based on system module log.

model to log information about IT systems because this is an industry standard to help enable log aggregation tools like Splunk to utilize data in standardized dashboards, alerts, and other knowledge artifacts quickly and easily. To efficiently operate a production system, it is critical to choose the correct log levels for a message Splunk. For example, there are kinds of log levels including Error, WARN, INFO, DEBUG, and TRACE which stand for different meanings in Splunk. Error message always indicates unhandled or fatal exceptions, including failed business logic. This mostly triggers when service is interrupted. For WARN messages in Splunk, they are handled as non-fatal exceptions, execution of the process was able to continue but in a compromised state such as failure to meet NFR (non-functional requirement). The main reason is we always use NFR to specify criteria for making judgments of operations of a system, rather than specifying behaviors or functions. In a production environment, every transaction should have logs as the outputs which could be informational messages. Especially, process entry and exit, key sub-transactions, status of any external calls, and transaction summarization are treated as key information. For DEBUG message, it is similar to info but with further depth in that may often be turned on in production to allow additional debugging. We typically use the TRACE message for the most detailed level of logging in the preproduction environment.

Additionally, if there is not enough data available at the INFO level, live monitoring and forensic research info incidents and poor performance will be severely compromised. Therefore, choosing the correct log level leads the support team to run services in DEBUG instead of finding key messages they need to operate the system, incurring additional costs. Conversely, if too much information is emitted at INFO, the performance and storage costs of logging may be higher than is necessary. (See Figures 7-9)



[redacted] Program Status

Site: All

Server: All

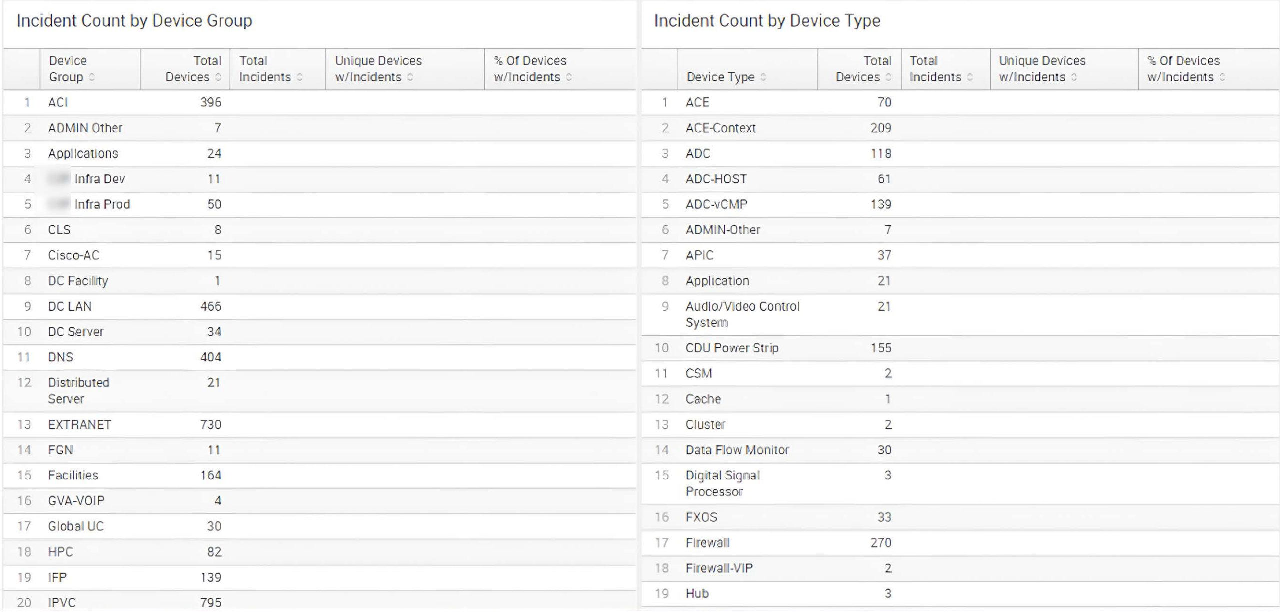
List order: smart

103	100	0	0	3	0
Total (excl. suspended)	Successful	Late	Failed	Planned/Processing	Suspended

Current status of TcBucks

	Program	ItemID	Site	Server	Start	End	Duration	Number of successful QL	Total number of QL	Completed	Status	Status time	Comment
1	C5[redacted]	PRG-[redacted]	[redacted]PE1	[redacted]096	2019-08-14 21:45:43	2019-08-15 00:01:23	02:18:51	50	51	no	running	2019-08-15 01:53:04	
2	V3[redacted]	PRGV[redacted]	[redacted]PE1	[redacted]096	2019-08-14 22:09:08			0	1	no	running	2019-08-15 01:53:04	
3	V3[redacted]	PRG-[redacted]	[redacted]PE1	[redacted]096	2019-08-14 21:48:44	2019-08-15 01:47:00	03:58:16	69	70	no	running	2019-08-15 01:53:04	
4	B2[redacted]	PRGE[redacted]	[redacted]PE1	[redacted]097	2019-08-15 00:07:04	2019-08-14 16:45:59	00:21:16	12	12	yes	successful	2019-08-15 01:53:04	
5	B2[redacted]	PRG-[redacted]	[redacted]PE1	[redacted]101	2019-08-14 22:32:01	2019-08-14 23:39:41	01:07:40	62	62	yes	successful	2019-08-15 01:53:04	
6	B2[redacted]	PRG-[redacted]	[redacted]PE1	[redacted]097	2019-08-15 00:07:03	2019-08-15 00:28:09	00:21:06	10	10	yes	successful	2019-08-15 01:53:04	
7	B2[redacted]	PRGE[redacted]	[redacted]PE1	[redacted]100	2019-08-14 22:04:09	2019-08-14 22:20:04	00:15:55	11	11	yes	successful	2019-08-15 01:53:04	

Figure 7. Customizable dashboard for recent changes and status.



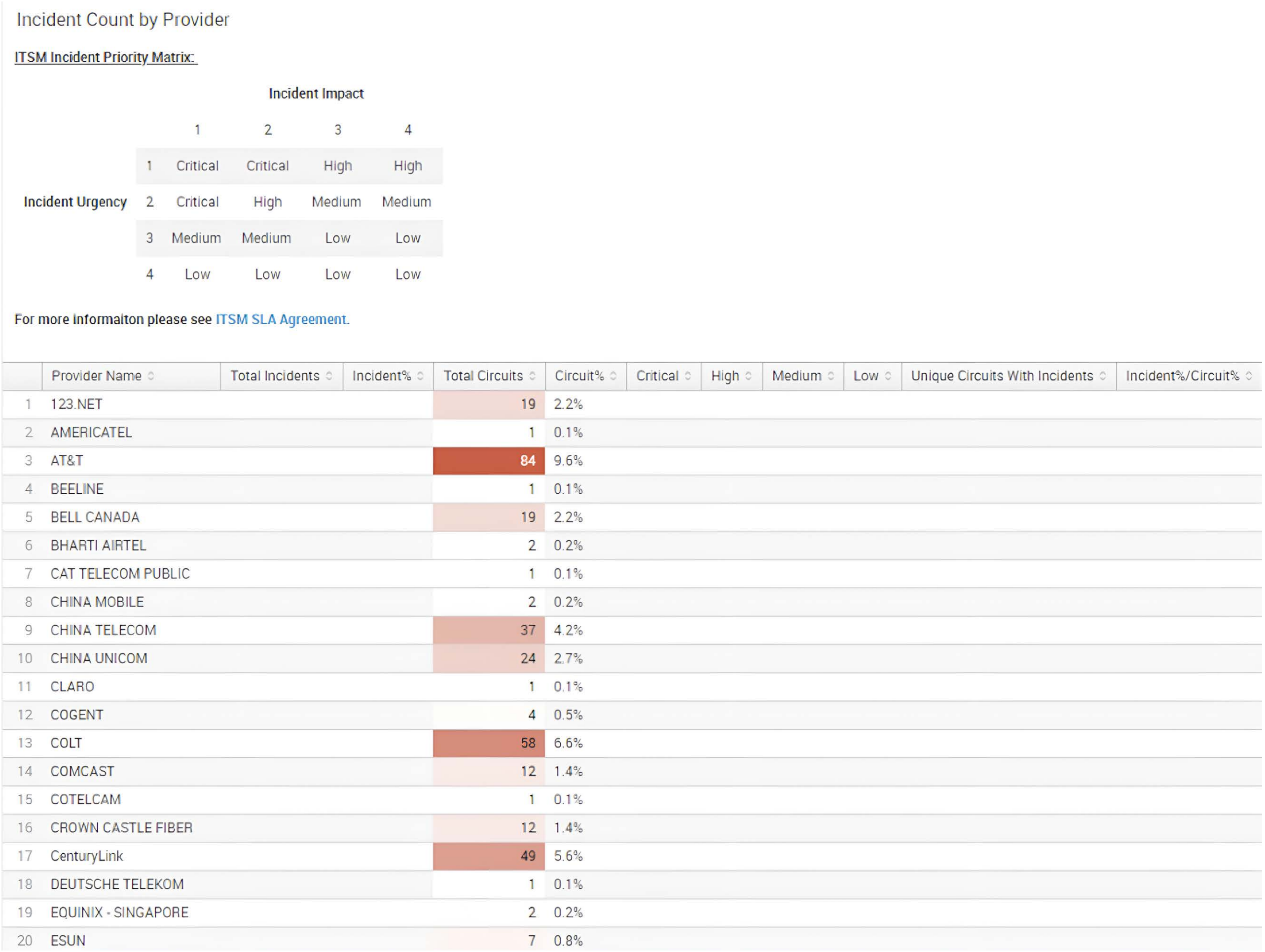
Incident Count by Device Group

	Device Group	Total Devices	Total Incidents	Unique Devices w/Incidents	% Of Devices w/Incidents
1	ACI	396			
2	ADMIN Other	7			
3	Applications	24			
4	[redacted] Infra Dev	11			
5	[redacted] Infra Prod	50			
6	CLS	8			
7	Cisco-AC	15			
8	DC Facility	1			
9	DC LAN	466			
10	DC Server	34			
11	DNS	404			
12	Distributed Server	21			
13	EXTRANET	730			
14	FGN	11			
15	Facilities	164			
16	GVA-VOIP	4			
17	Global UC	30			
18	HPC	82			
19	IFP	139			
20	IPVC	795			

Incident Count by Device Type

	Device Type	Total Devices	Total Incidents	Unique Devices w/Incidents	% Of Devices w/Incidents
1	ACE	70			
2	ACE-Context	209			
3	ADC	118			
4	ADC-HOST	61			
5	ADC-vCMP	139			
6	ADMIN-Other	7			
7	APIC	37			
8	Application	21			
9	Audio/Video Control System	21			
10	CDU Power Strip	155			
11	CSM	2			
12	Cache	1			
13	Cluster	2			
14	Data Flow Monitor	30			
15	Digital Signal Processor	3			
16	FXOS	33			
17	Firewall	270			
18	Firewall-VIP	2			
19	Hub	3			

Figure 8. Splunk dashboard based on incident count by device group and type.



Incident Count by Provider

ITSM Incident Priority Matrix:

		Incident Impact			
		1	2	3	4
	1	Critical	Critical	High	High
Incident Urgency	2	Critical	High	Medium	Medium
	3	Medium	Medium	Low	Low
	4	Low	Low	Low	Low

For more informaiton please see ITSM SLA Agreement.

	Provider Name	Total Incidents	Incident%	Total Circuits	Circuit%	Critical	High	Medium	Low	Unique Circuits With Incidents	Incident%/Circuit%
1	123.NET			19	2.2%						
2	AMERICATEL			1	0.1%						
3	AT&T			84	9.6%						
4	BEELINE			1	0.1%						
5	BELL CANADA			19	2.2%						
6	BHARTI AIRTEL			2	0.2%						
7	CAT TELECOM PUBLIC			1	0.1%						
8	CHINA MOBILE			2	0.2%						
9	CHINA TELECOM			37	4.2%						
10	CHINA UNICOM			24	2.7%						
11	CLARO			1	0.1%						
12	COGENT			4	0.5%						
13	COLT			58	6.6%						
14	COMCAST			12	1.4%						
15	COTELCAM			1	0.1%						
16	CROWN CASTLE FIBER			12	1.4%						
17	CenturyLink			49	5.6%						
18	DEUTSCHE TELEKOM			1	0.1%						
19	EQUINIX - SINGAPORE			2	0.2%						
20	ESUN			7	0.8%						

Figure 9. Splunk view dashabord based on incident priority matrix.

5. Annotate the Data Stream with Metadata Keys

In Splunk, there are 2 important default fields including indexes and sourcetypes. For instance, sourcetype = radius5, and index = netcool. (See **Figure 10**)

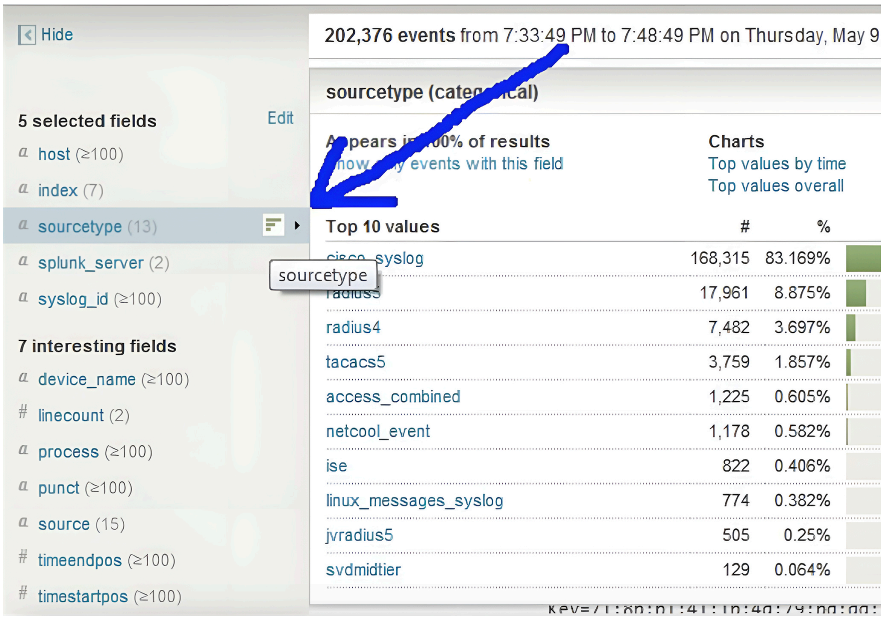
The date/time stamp field is what Splunk uses as the time field, otherwise, it will use the time the data was indexed. So, if an IT administrator queries data from yesterday, he does not have a time field, it will index it with today's date. Because incoming "machine data" is given a distinct source type based on where the data comes from. This data is then started in an index. What's more, roles are granted access at the index level. Only administrators could have access to the indices that contain the data that they are authorized to see. Otherwise, the indices that unauthorized users don't have access to will be invisible, along with the data stored inside since those who may be restricted to a certain app in Splunk. Such as created date, modified date, last updated date, etc. are timestamps in Splunk.

On the other hand, we need to pay more attention to the names of fields when trying to use them to search in Splunk in that is case sensitive so that they must match how it is defined in Splunk.

Every authorized user may have various fields depending on what data he has access to.

Generally, when we use particular fields to search data in Splunk, it should be a faster search because it is doing a metadata search rather than digging into events. For example, host = fcvas4217, and syslog id = "LINK-3-UPDOWN" are treated as particular fields to be searched in Splunk. (See Figure 11 and Figure 12)

The more descriptive the name, the easier to find and share later. If you name it with one of several reserved terms, it will automatically be saved under the "Searches and Reports "folder in Splunk such as Netcool, RADIUS, Wireless, etc. Please do not use your name or ID in the name field in case you ever want to share it. The timeframe of your search is saved along with the parameters, so best to include in the name the last 24 hours, etc.



202,376 events from 7:33:49 PM to 7:48:49 PM on Thursday, May 9

5 selected fields Edit
- host (≥100)
- index (7)
- sourcetype (13)
- splunk_server (2)
- syslog_id (≥100)

7 interesting fields
- device_name (≥100)
- linecount (2)
- process (≥100)
- punct (≥100)
- source (15)
- timeendpos (≥100)
- timestartpos (≥100)

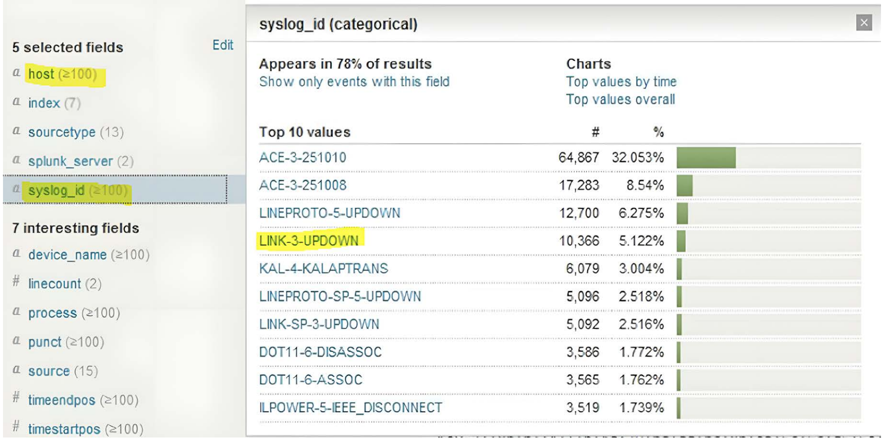
sourcetype (categorical)

Appears in 100% of results
Show only events with this field

Charts
Top values by time
Top values overall

Top 10 values	#	%
cisco_syslog	168,315	83.169%
radius5	17,961	8.875%
radius4	7,482	3.697%
tacacs5	3,759	1.857%
access_combined	1,225	0.605%
netcool_event	1,178	0.582%
ise	822	0.406%
linux_messages_syslog	774	0.382%
jvradius5	505	0.25%
svdmidtier	129	0.064%

Figure 10. Default field: sourcetype.



5 selected fields Edit
- host (≥100)
- index (7)
- sourcetype (13)
- splunk_server (2)
- syslog_id (≥100)

7 interesting fields
- device_name (≥100)
- linecount (2)
- process (≥100)
- punct (≥100)
- source (15)
- timeendpos (≥100)
- timestartpos (≥100)

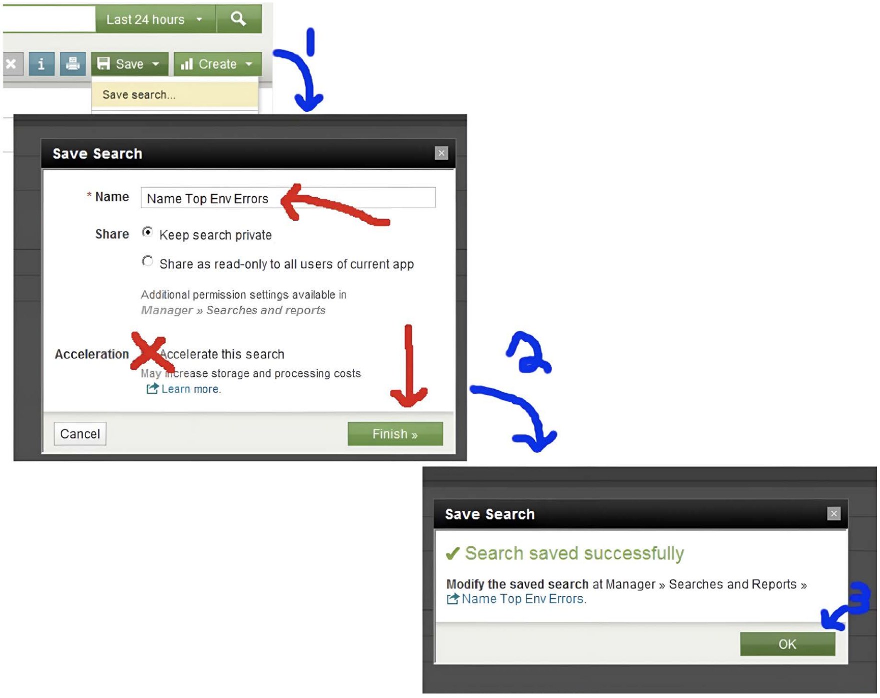
syslog_id (categorical)

Appears in 78% of results
Show only events with this field

Charts
Top values by time
Top values overall

Top 10 values	#	%
ACE-3-251010	64,867	32.053%
ACE-3-251008	17,283	8.54%
LINEPROTO-5-UPDOWN	12,700	6.275%
LINK-3-UPDOWN	10,366	5.122%
KAL-4-KALAPTRANS	6,079	3.004%
LINEPROTO-SP-5-UPDOWN	5,096	2.518%
LINK-SP-3-UPDOWN	5,092	2.516%
DOT11-6-DISASSOC	3,586	1.772%
DOT11-6-ASSOC	3,565	1.762%
ILPOWER-5-IEEE_DISCONNECT	3,519	1.739%

Figure 11. Syslog id field.



Save Search

* Name: Name Top Env Errors

Share: ● Keep search private
○ Share as read-only to all users of current app

Additional permission settings available in *Manager » Searches and reports*

Acceleration: Accelerate this search
May increase storage and processing costs
Learn more.

Save Search

✔ Search saved successfully

Modify the saved search at Manager » Searches and Reports » Name Top Env Errors.

Figure 12. Saving searches.

6. Analyze Root Causes Based on Operational Metrics

To analyze what root causes, make Splunk run slowly, we will specifically explain different scenarios based on operational metrics.

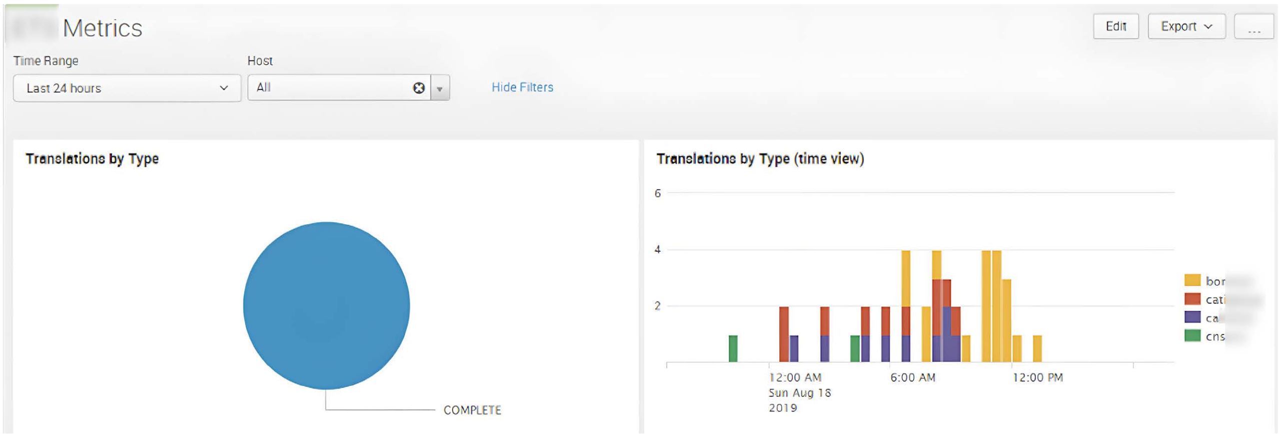
Sometimes, if you find some executive commands of Splunk commands take more than ten or so seconds to complete, they indicate an issue on the shared storage so please try using basic commands in Splunk to search. In the Splunk searching tool, any search taking over 30 seconds to return is standing for a slow search. If that case happened, we could validate whether the mount point is healthy by running some commands outside of Splunk when slow problems are only searches. There may be some results from looking in metrics for two to five minutes before and after the period of the slowly running search. The first one is this period is high which has SoS installed, and then we need to make sure a system load is not seen by us from CPU graphs on SoS in the same period. When this issue was caused by search load, we could have seen high CPU usage during that period of slow search. [2]

When specific departmental administrators want to use business criteria to judge the operation of a system, they must follow non-functional requirements which are not behavior or functional specifications to their application but are operational metrics. These measurable values demonstrate how effectively their application or system is achieving key business objectives. Therefore, key performance indicators and critical success factors are both as important operational metrics.

In Figure 13, "Translations by Type" is the key performance indicator that has 4 different types of translations based on the last 24 hours. We would see how long each of types of translations finished process and know which typed of translation had huge volumes in the translation services. In Figure 14, we would see average performance for bomlines per hour based on last 30 days of 2019 within 5 source sites. We would know how many accumulated numbers of bomlines had been built based on the five source sites. From these multisite metrics, we would see which source site has the highest volumes of bomlines built within 30 days and which one had the lowest volumes of accumulated bomlines.

7. Analyze the Limitations and Disadvantages of Splunk

For big business enterprises, they would like to use Splunk as a data analyzer, a way of correlating multiple data sources, or a metrics generator. However, Splunk has also kind of limitations and disadvantages. Small or middle-size companies, have to maintain a lot of resources which are required by Splunk since Splunk is not only very expensive to them but also requires a well-skilled technical team of employees to work on the Splunk platform. To use Splunk effectively, people need to have onboarding and special training to learn SPL (Splunk Processing Language). [3]

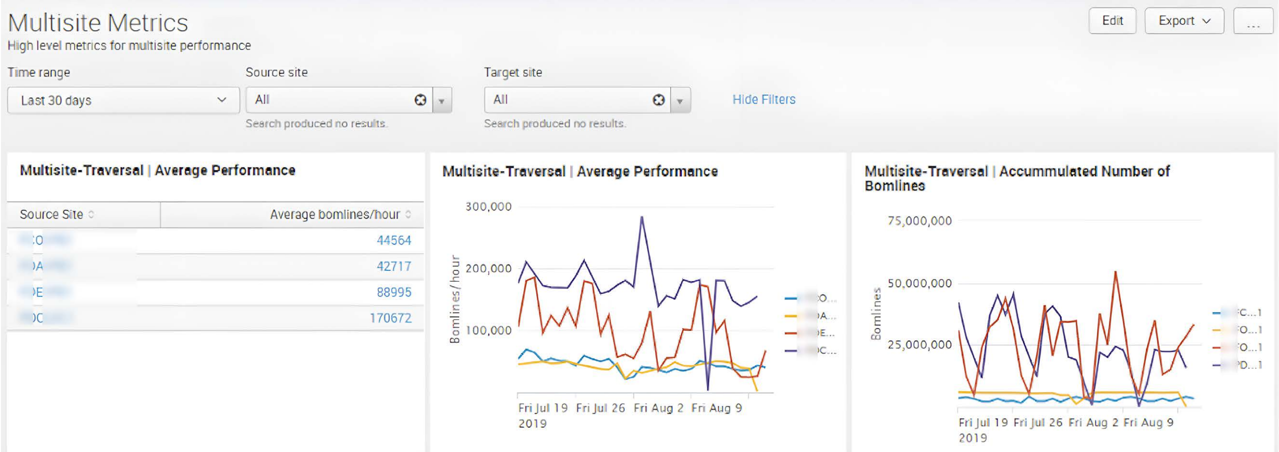


[redacted] Metrics

Time Range: Last 24 hours

Host: All

Figure 13. Engineering translations service metrics.



Multisite Metrics

High level metrics for multisite performance

Time range: Last 30 days

Source site: All

Search produced no results.

Target site: All

Search produced no results.

Multisite-Traversal | Average Performance

Source Site	Average bomlines/hour
[redacted]	44564
[redacted]	42717
[redacted]	88995
[redacted]	170672

Figure 14. Multiple metrics for different site performance.

In the IT infrastructure field, it is hard to implement Splunk in that is difficult to look at data integrity and data quality. Splunk has not developed any automated alerting with a new data integrity feature to automatically notify administrators when some users have changed something but only has deployed a script on their indexers to check the integrity of all indexes with the Splunk CLI command in the inputs.conf. [4]

If searching data over a long-time range or a large amount of data, Splunk would have a very slow query speed. Many developers always complain about searching speed from Splunk so they have to limit the amount of data retrieved from the disk range by setting a narrow time window. [5] In other words, Splunk is inability to handle large amounts of data but must use the most specific terms in users' search that they can.

The setup process is complex for Splunk enterprise inside docker containers because it requires the user's current system to meet kind of requirements before trying to deploy it to their existing system such as supported operating systems, hardware requirements, disk volumes, supported file systems, and storage requirements, etc. [6]

The model of licensing could become expensive due to the increasing volume of data which is one of the significant drawbacks of Splunk. Resource intensive would be one of the features based on hardware requirements and operational overhead. For large-scale deployments, it may require significant hardware resources to deploy and maintain a Splunk instance. Expertise and ongoing effort would be required to manage and optimize Splunk configurations and searches. There is a steep learning curve for new users to the Splunk platform because they are not familiar with search query languages such as Splunk Processing Language (SPL). Time-consuming would be in both of training and onboarding processes for new users. Performance issues may have existed in large volumes of data or complex search queries even if Splunk has robust indexing capabilities. The more volumes of data ingested; the more indexing performance degraded. As a result, longer search times and potential delays are to obtain insights from log information. Data silos may be from Splunk which is a separate platform in analyzing logs and requires extra effort and customization to integrate with other systems or tools. It is difficult to get a unified view of data across the organization based on fragmented data management practices. For highly regulated industries, security and compliance features of Splunk may not meet their requirements of operations within organizations. Extra investments and configurations may be requirements for compliance with relevant regulations and implementing robust security measures. Vendor lock-in is from adopting Splunk when business organizations heavily dependent on Splunk to analyze and monitor logs. There is a challenge to switch to an alternative solution since the organization has significant investment in Splunk infrastructure and training. Specific requirements and use cases would make the availability of support be the difference in a large and active user community of Splunk. Timely support and expertise would be issues in troubleshooting problems and optimizing Splunk deployments.

8. What Aspects of Splunk Need To Be Improved

There are some kind of aspects in Splunk I have to say since Splunk is hard to compare two or more times series data in a single graph. Moreover, we cannot search and make suggestions on Splunk commands as we are typing on the search windows. Splunk requires some learning to use all of its features but for most businesspeople, maybe it is hard to easily understand its SPL during short-term periods. Therefore, Splunk should make the dashboard and reports less technical for kind of non-technical users to get more value from this tool. On the other hand, Splunk processing language goes very deep so if we want to do some advanced formatting or statistical analysis, there is a bit of a learning curve. During that time, we have to pay for Splunk training to learn this special language to manipulate our data. [7]

Splunk needs to integrate AI to understand its system logs. Splunk alerting should be based on auto-learning. Due to module-driven in Splunk, we have to constantly add modules and costs to get newer functionality. In addition, Splunk lacks offline and email features. There is not only an easy way to back out integrating a log that suddenly balloons a user over their license limits but also does not have an easier way to help themselves parse log types which means an administrator must be able to tell Splunk how random log is formatted. Otherwise, they only have limited ability to search on random logs.

Even if Splunk as a searching tool exists in the IT field, we still have to read through a lot of documentation to find the right answers we are looking for and sometimes we do not find them. As a result, the helping function in Splunk still needs to be improved to be more intuitive. There are not many tools or features for customization in Splunk reporting such as adding font options and colors for graphs, graphics, and text so that stops us from easily making important reports or panels at the top of the Splunk dashboard. I think Splunk should improve functionality to give users more opportunities such as adding URL links to other related dashboards or offering kind of smart ways for users to emphasize their important data in their customizable dashboard.

When it comes to the visualization capabilities of the Splunk dashboard, we should use kind of development tools such as XML, JavaScript, and CSS to improve its visualization functioning.

9. Current Related Works on Splunk Tool

For the Splunk tool, I want to introduce some related works about it. First of all, Fred Speece discussed using Splunk as their SIEM for a Windows majority network when their InfoSec team worked with Active Directory in an organization with an immature security posture preparing for the first Penetration Test since Splunk could alert on abnormal behavior. Secondly, Cisco enterprise talks about how their Cisco HyperFlex systems are working with Splunk to gain an easy, fast, and secure way to analyze massive streams of data generated by information technology systems, security devices, and technical infrastructure. The main reason is Splunk tool could monitor and analyze data from any source such as computing, storage, and networking activities; service health; firewall access; and customer clickstreams and call records. Finally, Splunk could turn machine-generated data into their specific business insight. [8]

To explain how we could improve our security posture, Splunk enterprise positively uses Splunk software as their SIEM for supporting the full range of information security operations including posture assessment, monitoring, alert and incident handling, CSIRT, breach analysis and response, and event correlation. In addition to detecting known and unknown threats, they introduce how Splunk software could investigate threats, determine compliance, and use advanced security analytics for detailed insight. [9]

To correct and index any machine data, Splunk enterprise offers a leading platform for real-time operational intelligence to deliver new levels of visibility, insight, and intelligence for information technology and the business since Splunk software would be treated as software as a service (SaaS) offering which could be able to read data from virtually any source such as network traffic or wire data, web servers, custom applications, application servers, hypervisors, GPS systems, stock market feeds, social media, sensors and preexisting structured databases. Then, we could learn to know what is happening in real time and would be given by deep analysis of what's happened across our IT system and technology infrastructure so that we could make informed decisions. [10]

For example, Oracle Enterprise collects data from Oracle ZFS Storage Appliance based on a Splunk infrastructure which is allowed to distribute the load of collecting the data and configures Splunk Enterprise for indexing the data in that is an operational intelligence platform for many firms to gather machine data from kind of resources. [10] [11] As a result, Splunk is not only used to monitor the end-to-end infrastructure to help avoid service degradation and to troubleshoot different problems like performance bottlenecks but also offers predictive analysis tools for helping determine resources may be overburdened.

As users' applications and infrastructure change, Splunk company provides us with a proactive monitoring tool, which is called Splunk software to support those changes since Splunk software can collect, index, and analyze the users' machine data, and then programmatically search for anomalous events and derive aggregate measures based on multiple events or queries. Finally, these could be aggregated and summarized on performance dashboards tailored to different constituencies.

10. Summary of Beneficiations of Using Splunk

Using Splunk is to collect and index any syslog, tools logs, any type of custom logs, server and device access records, monitoring events, change control records, and incident record data from virtually any source in real time because Splunk could give us a view of all these data in one timeline, or a single “pane of glass”. Moreover, we could not only use it to analyze, save, and export data and share them with others, but also provide customizable dashboard functionality.

We could take advantage of Splunk in building different functional dashboards to get a quick view of overall system health, issue areas, and end-user ramifications for fulfilling our business purposes. For web analytics dashboard, is used for analyzing usage trends and performance. The incident notification dashboard is an alert when services are broken. The cyber security anomalies dashboard is to read huge network logs. [12] For example, with Splunk, one team member could be able to have a possible root cause within one business day for a wireless network outage. Because Splunk not only allows us to allow those closest to the problem to quickly search all data, in one single app, but also helps us save on hours spent troubleshooting and quickly provide the most accurate data to vendor support cases.

Author Contributions

Xueying Pan as the only author confirms her contribution to the paper as follows: Design of paper; data collection; analysis and interpretation of results; draft manuscript preparation. Xueying Pan reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials

The author confirms the data supporting the findings of this paper are available within the article.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Fred Speece (2016) Detecting Penetration Testers on a Windows Network with Splunk. SANS Institute Information Security Reading Room Site.

[2] Splunk Inc (2024) Splunk Enterprise Overview 7.3.0. https://docs.splunk.com/Documentation/Splunk/7.3.0/Overview/AboutSplunkEnterprise

[3] Splunk Inc (2021) Search Tutorial. https://docs.splunk.com/Documentation/SplunkCloud/latest/SearchTutorial/Usethesearchlanguage

[4] Splunk Inc (2023) Monitor Files and Directories with Inputs.Conf. https://docs.splunk.com/Documentation/Splunk/9.2.0/Data/Monitorfilesanddirectorieswithinputs.conf

[5] Splunk Inc (2023) Quick Tips for Optimization. https://docs.splunk.com/Documentation/Splunk/7.3.1/Search/Quicktipsforoptimization

[6] Splunk Inc (2018) Splunk Enterprise Distributed Search.

https://docs.splunk.com/Documentation/Splunk/7.3.0/DistSearch/Whatisdistributedsearch

[7] Splunk Inc (2022) Splunk Enterprise Search Manual 7.3.1 about Search Normalization. https://docs.splunk.com/Documentation/Splunk/7.3.1/Search/Searchnormalization

[8] Cisco (2018) Cisco HyperFlex Systems for Splunk Enterprise. https://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-systems/hyperflex-hx-series/solution-overview-c22-739511.pdf

[9] Kidd, C. (2023) SIEM: Security Information and Event Management Explained. https://www.splunk.com/en_us/blog/learn/siem-security-information-event-management.html

[10] Siddiqui, L. (2024) The SaaS Security Guide: Best Practices for Securing SaaS. https://www.splunk.com/en_us/blog/learn/saas-security.html

[11] Hartley, J. (2015) Splunk and the Oracle ZFS Storage Appliance. Oracle Technical White Paper September 2015 Version 2.1.

[12] Splunk White Paper (2017) Splunk Security Use Case Detecting Unknow Malware and Ransomware. https://cyberoregon.com/wp-content/uploads/2017/11/White-Paper-Security-Use-Case-Detecting-Unknown-Malware.pdf