edureka!

SPLUNK
E-BOOK

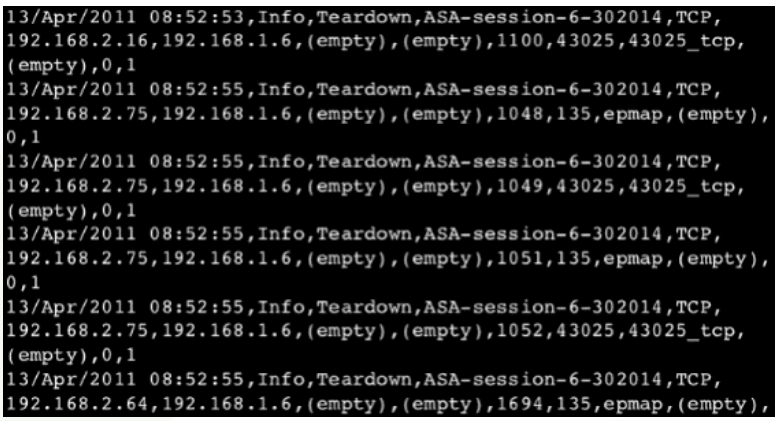
Agenda

- Need For Splunk
- What Is Splunk?
- Splunk vs ELK vs Sumo Logic
- Splunk Careers
- Splunk Use Case – Domino's Success Story
- Stages in Data Pipeline
- Splunk Forwarder
- Splunk Indexer
- Splunk Search Head
- How Splunk Works?
- Splunk Architecture
- Splunk Knowledge Objects
- Splunk Tutorial
- Splunk Interview Questions

Need For Splunk

You must be aware of the exponential growth in machine data over the last decade. It was partly because of the growing number of machines in the IT infrastructure and partly because of the increased use of IoT devices. This machine data has a lot of valuable information that can drive efficiency, productivity and visibility for the business. Splunk was founded in 2003 for one purpose: *To Make Sense Of Machine Generated Log Data*.

Look at the below image to get an idea of how machine data looks.



Now imagine if you were a SysAdmin trying to figure out what went wrong in your system's hardware and you stumble upon logs like the one's in the above image, what would you possibly do? Would you be able to locate in which step your hardware failed you? There is a remote chance that you might be able to figure it out, but even that is only after spending hours in understanding what each word means. To tell you in a nutshell, machine data is:

- Complex to understand
- In an unstructured format
- Not suitable for making analysis / visualization

This is where a tool like Splunk comes in handy. You can feed the machine data to Splunk, which will do the dirty work(data processing) for you. Once it processes and extracts the relevant data, you will be able to easily locate where and what the problems were.

Splunk started off this way, but it became more prominent with the onset of Big Data. Since Splunk can store and process large amounts of data, data analysts like myself started feeding big data to Splunk for analysis. Dashboards meant for visualization was a revelation and within no time Splunk was extensively used in the big data domain for analytics.

What Is Splunk?

Splunk is a software platform to search, analyze and visualize the machine-generated data gathered from the websites, applications, sensors, devices etc. which make up your IT infrastructure and business.

If you have a machine which is generating data continuously and you want to analyze the machine state in real time, then how will you do it? Can you do it with the help of Splunk? Yes! You can. The image below will help you relate to how Splunk collects data.

Pull data from multiple systems in real time

Real time processing is Splunk's biggest selling point because, we have seen storage devices get better and better over the years, we have seen processors become more efficient with every ageing day, but not data movement. This technique has not improved and this is the bottleneck in most of the processes within organizations.

If you already think Splunk is an awesome tool, then hear me out when I say that this is just the tip of the iceberg. You can be rest assured that the remainder of this blog post will keep you glued to your seat if you have an intention to provide your business the best solution, be it for system monitoring or for data analysis.

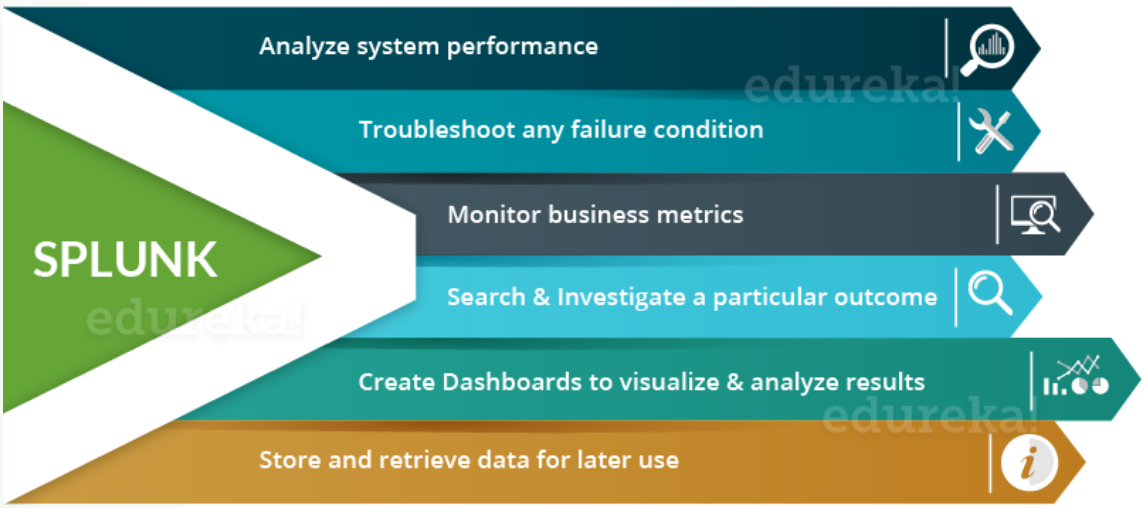
The other benefits with implementing Splunk are:

- Your input data can be in any format for e.g. .csv, or json or other formats
- You can configure Splunk to give Alerts / Events notification at the onset of a machine state
- You can accurately predict the resources needed for scaling up the infrastructure
- You can create knowledge objects for Operational Intelligence

What Is Splunk?

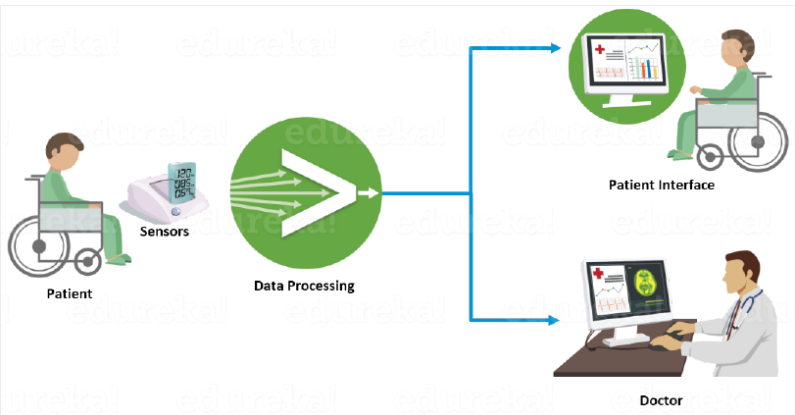
For those of you who don't know what is a knowledge object, it is a user-defined entity using which you can enrich your existing data by extracting some valuable information. These Knowledge objects can be saved searches, event types, lookups, reports, alerts or many more which helps in setting up intelligence to your systems.

The infographic below mentions some of the functionalities for which Splunk can be used.



To give you more clarity on how Splunk works, I am going to tell you how Bosch used Splunk for data analytics. They collected the healthcare data from the remotely located patients using IoT devices(sensors). Splunk would process this data and any abnormal activity would be reported to the doctor and patient via the patient interface. Splunk helped them achieve the following:

- Reporting health conditions in real time
- Delve deeper into the patient's health record and analyze patterns
- Alarms / Alerts to both the doctor and patient when the patient's health degrades



Splunk vs ELK vs Sumo Logic

There are a plethora of tools available in the market today for storing and processing machine data, but how do you choose the right tool? Do you know which among Splunk vs ELK vs Sumo Logic is the best option to handle the data generated by your machines?

This will help you know the pros and cons of each of these tools, after which you will be able to zero in on the tool most appropriate for your organization's needs. I learnt the differences between these tools when I was doing research for my project, where abnormal system state and frauds needed to be alerted in real time. That was when I learnt that there were tools dedicated to monitor the systems by processing machine data. So I have published this blog to share what I learnt:

- Differences between Splunk vs ELK vs Sumo Logic
- How to choose the right tool?

Splunk, ELK and Sumo Logic are among the most widely used in the market and they provide a good representation of the different types of tools available. Other popular tools being Loggly, Graylog and PaperTrails.

You can go through the below table to get an overview of the features supported by the three tools.

Splunk vs ELK vs Sumo Logic

Features	Splunk	Sumo Logic	ELK
Searching	✓	✓	Only possible with Integrations
Analysis	✓	✓	Only possible with Integrations
Visualization Dashboard	✓	✓	Only possible with Integrations
SaaS Setup	✓	✓	✓
On Premise Setup	✓	✗	✓
Plugins & Integration	✓	✓	✓
Input any data type	✓	Needs Plugins	Needs Plugins
Customer Support	✓	Available; but not proficient	Available; but not proficient
Documentation & Community	✓	✗	✓

Proprietary / Open-Source

Splunk is a proprietary tool which provides both an on-premise and a cloud setup. The difference between on-premise and cloud setup lies in where you are storing your data. If you are going for an on-premise setup, you can choose between Splunk Enterprise or Splunk Light. If you prefer a cloud setup, then you can opt for Splunk Cloud, which is a SaaS version of Splunk Enterprise.

Splunk vs ELK vs Sumo Logic

Sumo Logic is again a proprietary tool, but it offers only a cloud setup. This means all your data is stored in cloud.

ELK on the other hand is a combination of three open source tools(Elastic Search-Logstash-Kibana). Similar to Splunk, ELK can be installed on-premise as well as setup on the cloud. Their cloud platform is called Elastic Cloud. If you are an AWS user, then you have another option: AWS Elastic Search. In October last year, AWS released this as a hosted solution for ELK.

Bottom line: Splunk and Sumo Logic are proprietary software and you pay for a wide range of functionality. Whereas ELK is open source and cheaper. So if you work for a small or a medium sized company, proprietary software might not be the best option because you might be paying for a whole lot of features that you might not use.

Searching, Analysis & Visualization

With **Splunk** and **Sumo Logic**, you have a complete data management package at your disposal. Once you have imported the data, you can search and investigate that data. You can perform analysis to gain insights and formulate business strategies accordingly. You can even showcase your findings in a visual form by using visualization dashboards.

Since **ELK** is a combination of three tools, Searching, Analysis & Visualization will only be possible after the ELK stack is setup. Elastic Search does data storage and works as an analytics engine, Logstash is a data collection and transferring agent and Kibana is used for visualizing data. These three tools together are called the ELK stack (Elastic search – Logstash – Kibana).

Bottom line: Searching, Analysis & Visualization can be done with all three tools, but they are done in different ways in different tools.

Data Type

I did some research on the different data types that these tools accept and I learnt some interesting facts about Splunk and Sumo Logic. **Splunk** claim that their tool can accept data in any format, for e.g. .csv, or json or any other log format. Even **Sumo Logic** claim that their tool can 'collect logs from almost any system in nearly any format'.

Splunk vs ELK vs Sumo Logic

In case of **ELK**, Logstash is responsible for data on boarding. Even though Logstash does not support all data types by default, plugins can be setup for different data types. But the downside with Logstash is its long startup time and difficulty to debug errors since it uses a non-standard configuration language.

Another detail to be considered here is the difference in the way data is parsed. I noticed that in ELK and Sumo Logic, the data fields must be first identified and then configured before it is shipped. But with Splunk I can do it after the data comes to the system. This makes data onboarding easier by separating shipping and field labeling.

Integrations & Plugins

I found that **Splunk** is very good for setting up integrations with other tools. It has around 600 plugins for IT operations, security and other needs. Although **ELK** is seeing an increased number of available plugins, it does not support as many integrations as Splunk does. Logstash which is responsible for the data on boarding in ELK, has only around 160 plugins at present and work is ongoing for more integrations.

Customer Support & Documentation

Splunk has a big customer base, thus a very strong community. I found the Splunk community helpful and many of my questions got answered there. This is why I feel Splunk would offer better support than Sumo Logic and ELK.

I also found that Splunk's knowledge base has an accurate documentation for setting up clusters and plugins, but with Sumo Logic I did not find the documentation to be as good as I expected and I had a tough time navigating through the documentation.

All three of these tools have their own advantages and categories in which they are better than the other. My only intention here is to help you in your decision making. So, it is necessary that you choose the tool that can be tailored to your needs.

I found Splunk to be the most comfortable among these tools because it was very easy to use and it was a one stop solution for my needs. It let me do Searching, Analysis, Visualization all on the same platform and offered me good support when I needed it.

Splunk Careers

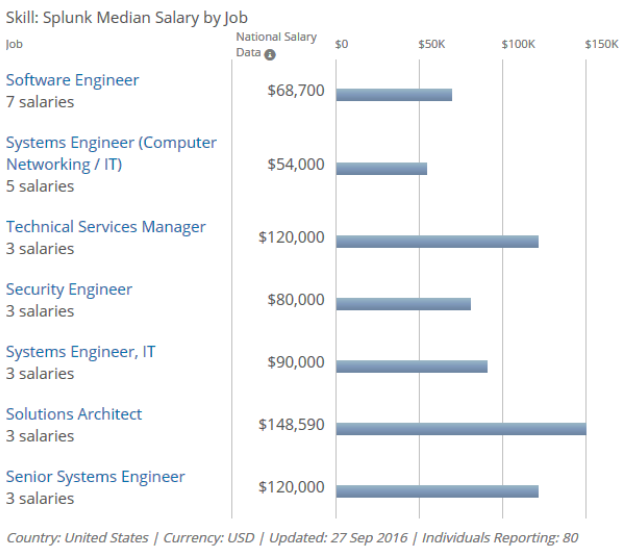
As someone who has been closely following the SMAC (Social, Mobile, Analytics & Cloud) revolution since the last couple of years or so, I used to wonder where all the 'waste' data that SMAC generated was going. Was somebody even trying to see if it contained stuff that was useful in some way? Voila! My hunch was right. Unstructured data is apparently a treasure trove of information that can be leveraged by businesses all over the world. What's more? They even provide invaluable insights about customers, their behaviour, their expected future behaviour etc. It didn't take too long for the brainy ones to create tools that helped you get these insights, and Splunk leads the pack of such tools. Consequently, Splunk careers are one of the most sought after, in the industry today.

Currently, Splunk has 1200 odd apps that help make sense of different formats of log data, providing visibility across on-premise, cloud and hybrid environments. Between these apps and the sudden proliferation of Splunk in organizations — big and small — careers around Splunk have skyrocketed in the last couple of years, and there is a fair indication that things will only get better from here on.

Splunk careers

Forbes says that big data related jobs pertaining to unstructured machine data and Internet of Things (IoT) have seen unprecedented growth percentages in excess of 704% globally over the last five years. Specific job roles that promise lucrative Splunk careers include:

- Software Engineer
- Systems Engineer
- Programming Analyst
- Solutions Architect
- Security Engineer
- Technical Services Manager



According to Indeed, Splunk related jobs dictate paychecks of up to $148,590 for a solutions architect and $120,000 for a senior systems engineer. Even starting salaries are attractive in comparison to other software development and IT jobs across the world. I've pulled out the numbers for you and they do look promising.

Splunk Careers

I did a bit of digging around to see who actually uses Splunk. While you can have a flourishing Splunk career in virtually any domain of technology, five industry sectors have been the flag-bearers of Splunk in recent times. These include Finance and Insurance, Manufacturing, Information Technology, Retail Trade, and Technical Services. More than 9,000 enterprises, government agencies, universities, and service providers in more than 100 countries currently use Splunk for in-depth business and customer understanding, mitigating cyber security risk, prevent fraud, improve service performance, and reduce overall costs. The rate at which Splunk is being adopted by organizations worldwide, the number is only slated to grow exponentially.

Take a look at the graph here (Source: Indeed). This shows how Splunk jobs have been skyrocketing since the last couple of years, with no signs of slowing down.

Splunk career tracks

If you are a Splunk enthusiast (or aspire to be one anytime soon), the future is filled with possibilities that are challenging and lucrative. This is just the perfect time to learn and master Splunk. A career in Splunk has three primary spokes – Architect, Administrator and Developer. If you decide to bring Splunk on to your resume, here are the list of tasks you'd be expected to perform, depending on the job title you choose.

If you are a Splunk Architect, your organization will expect you to leverage Splunk for searching and reporting data, creating Splunk Knowledge Objects, operating Dashboards and Visualizations, Architecting and Deploying Splunk across your organization. On the other hand, if you decide to become a Splunk Admin, you need to perform Installation, License Management, management of Splunk Apps, Configuration and Index Management, Event Parsing, Distributed Search and Performance Tuning. Finally, if you aspire to be a Splunk Developer, you are entirely responsible for end-to-end solution development as well as maintenance. Primary responsibilities of a Splunk developer include implementation, workload management, architecture, design and estimation.

Splunk Careers

Organizations are committed to Splunk

As on October 2016, the following companies had attractive openings for Splunk competencies worldwide. Take a good look at these companies, that are a mix of industries ranging from technology to Iot to manufacturing, and everything in between. With the right knowledge and training of Splunk, you can soon have a business card with one of these big names on it.

Bright future for Splunk

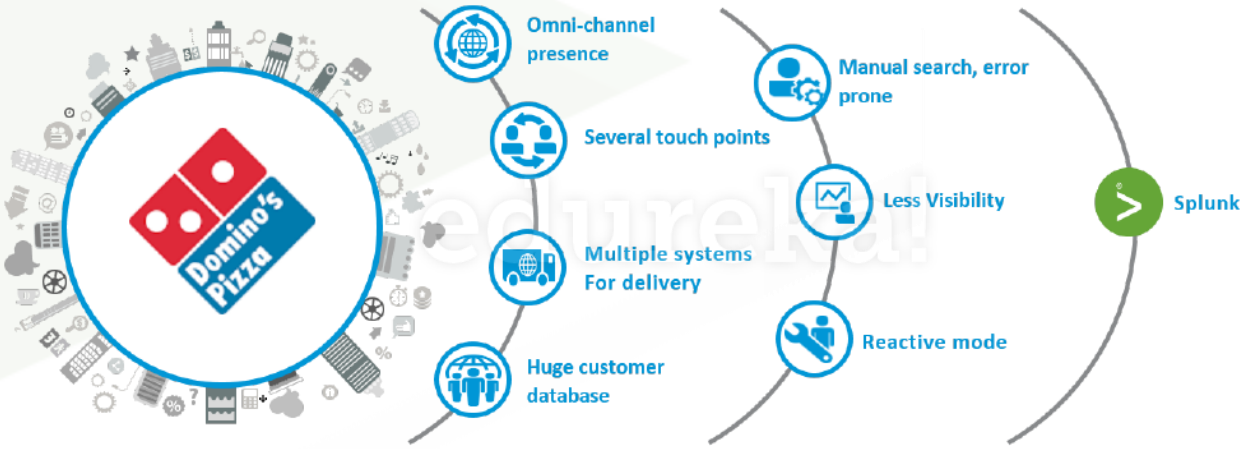
As some wise soul once said, "Inefficiency is the mother of all opportunities". Looks like the wise soul had Splunk in mind! Earlier this year, Gartner sounded a warning alarm – Unstructured data growth is rapidly outpacing structured data and is poorly controlled, stored and managed on file shares, on personal devices and in the cloud. This realization stems from the fact that 50% of IT teams in the world are completely unprepared for the humongous growth of unstructured data. What is more shocking is the fact that by 2022, 93% of all data in the digital universe will be unstructured in nature (Source: IDG). It doesn't take rocket science to understand that enterprises cannot survive with insights from just 7% of data. Data from social chatter, usage logs, customer clickstreams etc. are critical data that are invaluable to organizational strategies and success, and the evolution of offerings like Splunk are heavily capitalizing on the mismanagement of unstructured data. In simple words, this means just one thing — There is a pressing need for Splunk specialists in the world. Needless to say, you can be in the forefront of this revolution of sorts.

Splunk Use Case: Domino's

While many companies and organizations have used Splunk for operational efficiency, here I will talk about how Domino's Pizza used Splunk to analyze consumer behaviour to build data driven business strategies. This Splunk use case shows how Splunk can be used extensively in any domain.

Splunk Use Case: Domino's Pizza

You might be aware that Domino's Pizza is an e-commerce cum fast food giant, but you might be unaware of the big data challenge they were facing. They wanted to understand their customers' needs and cater to them more effectively by using Big Data. This is where Splunk came to the rescue.
Look at the image below which depicts the circumstances that were building up to cause big data problems at Domino's.



Lot of unstructured data was generated because:

- They had an omni-channel presence for driving sales
- They had a huge customer base
- They had several touch points for customer service
- They provided multiple systems for delivery: Order food in-store, order via telephone, via their website and through cross-platform mobile applications
- They upgraded their mobile apps with a new tool to support 'voice ordering' and enable tracking of their orders

The excess data generated gave rise to the following problems:

- Manual searches being tedious and error prone
- Less visibility into how customer need/preference varies
- Unpreparedness and thus working in reactive mode to fix any problem

Domino's felt that the solution to these problems would lie in a tool which can easily process data. That was when they implemented Splunk.

Splunk Use Case: Domino's

"Up until implementing Splunk, managing the company's application and platform data was a headache, with much of its log files in a giant mess"
– according to their Site Reliability & Engineering Manager, Russell Turner

Turner mentioned that using Splunk for Operational Intelligence in place of a traditional APM tool helped him to lower the cost, search the data faster, monitor performance and get better insights into how customers were interacting with Domino's. If you look at the below image, you will find the different applications that were set up by implementing Splunk.

- Interactive Maps, for showing orders in real time coming from all across US. This brought employee satisfaction and motivation
- Real time feedback, for employees to constantly see what customers are saying and understand their expectations
- Dashboard, used to keep scores and set targets, compare their performance with previous weeks/ months and against other stores
- Payment Process, for analyzing the speeds of different payment modes and identifying error free payment modes
- Promotional Support, for identifying how various promotional offers are impacting in real-time. Before implementing Splunk, the same task used to take an entire day
- Performance Monitoring, to monitor the performance of Domino's in-house developed point of sales systems

Splunk proved to be so beneficial to Domino's that teams outside the IT department started exploring the possibility to use Splunk for gaining insights from their data.

Splunk Use Case: Domino's

Splunk For Promotional Data Insights

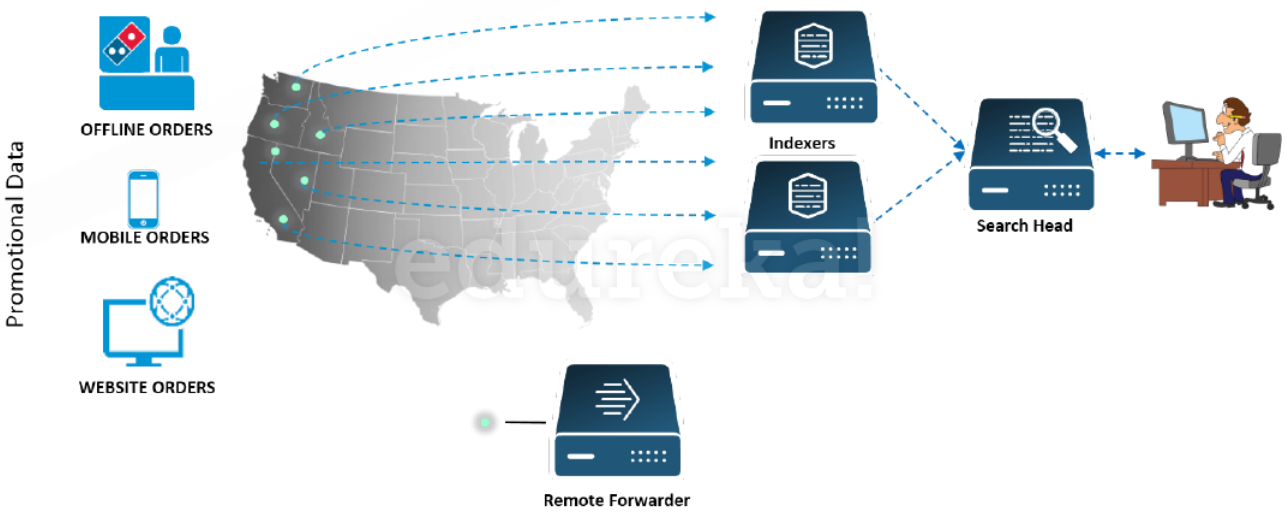
I am going to present a hypothetical Splunk use case scenario which will help you understand how Splunk works. This scenario demonstrates how Domino's Pizza used Promotional data to get better clarity as to which offer/coupon works best with respect to different regions, order revenue sizes and other variables.

*Note: The example of Promotional data used is hypothetical in nature and data present might not be accurate.

Domino's had no clear visibility into which offer works best – in terms of:

- Offer type (Whether their customers preferred a 10% discount or a flat $2 discount?)
- Cultural differences at a regional level (Do cultural differences play a role in offer choice?)
- Device used for buying products (Do devices used for ordering play a role in offer choices?)
- Time of Purchase (What is the best time for the order to be live?)
- Order revenue (Will offer response change wrt to order revenue size?)

As you can see from the below image, promotional data was collected from mobile devices, websites and various outlets of Domino's Pizza(using Splunk Forwarders) and sent to a central location(Splunk Indexers).



Splunk forwarders, would send the promotional data generated in real time. This data contained information about how customers responded when they were given offers, along with other variables like demographics, timestamp, order revenue size and device used.

Splunk Use Case: Domino's

Customers were divided into two sets for A/B Testing. Each set was given a different offer: 10% discount offer and flat $2 offer. Their response was analyzed to determine which offer was preferred by the customers.

The data also contained the time when customers responded and if they would prefer to buy in-store or do they prefer to order online. Most importantly, it contained Order revenue data – to understand if offer response changes with the order revenue size.

Once the raw data was forwarded, Splunk Indexer was configured to extract the relevant information and store it locally. Relevant information being the customers who responded to offers, time at which they responded and the device used for redeeming the coupons/offers.

Typically, the below information was stored:

- Order revenue based on customer response
- Time of purchase of products
- Device preferred by customers for placing the order
- Coupons / Offers used
- Sales numbers based on Geography

For performing various operations on the Indexed data, Search head was used. It is the component which gives a graphical interface for searching, analyzing and visualizing the data stored in the Indexers. Domino's Pizza gained the below insights by using the visualization dashboards provided by the Search head:

In USA and Europe, customers preferred a 10% discount instead of a $2 offer. Whereas in India, customers were more inclined to a flat $2 offer
10% discount coupons were used more when the order revenue size was large, whereas flat $2 coupons were used more when order revenue size was small.

Mobile apps were the preferred device for ordering during the evening and orders coming in from the website was most during the noon. Whereas ordering-in-store was highest during the morning

Domino's Pizza collated these results to customize the offers/coupons with respect to order revenue sizes for customers from a particular geography. They also determined which was the best time to give offers/coupons and targeted the customers based on the device they were using.

Stages In Data Pipeline

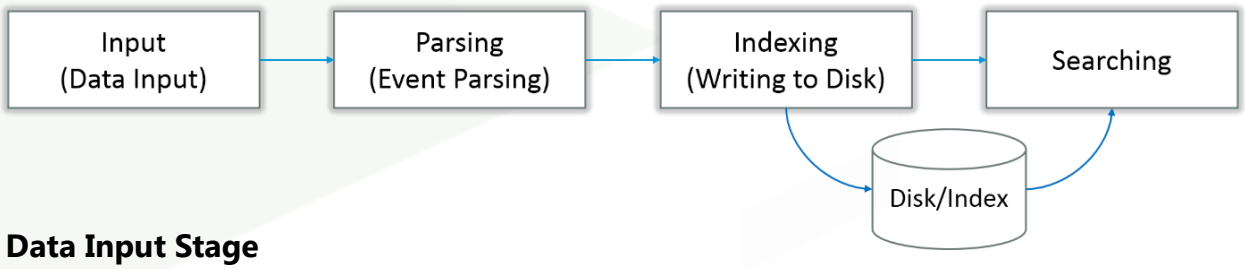
If you want to implement Splunk in your infrastructure, then it is important that you know how Splunk works internally. I have written this to help you understand the Splunk architecture and tell you how different Splunk components interact with one another.

Before I talk about how different Splunk components function, let me mention the various stages of data pipeline each component falls under.

Different Stages In Data Pipeline

There are primarily 3 different stages in Splunk:

- Data Input stage
- Data Storage stage
- Data Searching stage



Data Input Stage

In this stage, Splunk software consumes the raw data stream from its source, breaks it into 64K blocks, and annotates each block with metadata keys. The metadata keys include hostname, source, and source type of the data. The keys can also include values that are used internally, such as character encoding of the data stream and values that control the processing of data during the indexing stage, such as the index into which the events should be stored.

Data Storage Stage

Data storage consists of two phases: Parsing and Indexing.

1. In Parsing phase, Splunk software examines, analyzes, and transforms the data to extract only the relevant information. This is also known as event processing. It is during this phase that Splunk software breaks the data stream into individual events. The parsing phase has many sub-phases:

Stages In Data Pipeline

i. Breaking the stream of data into individual lines
ii. Identifying, parsing, and setting timestamps
iii. Annotating individual events with metadata copied from the source-wide keys
iv. Transforming event data and metadata according to regex transform rules

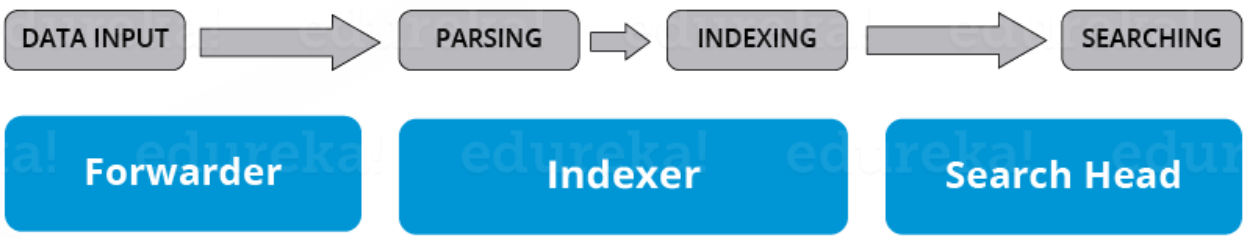
2. In Indexing phase, Splunk software writes parsed events to the index on disk. It writes both compressed raw data and the corresponding index file. The benefit of Indexing is that the data can be easily accessed during searching.

Data Searching Stage

This stage controls how the user accesses, views, and uses the indexed data. As part of the search function, Splunk software stores user-created knowledge objects, such as reports, event types, dashboards, alerts and field extractions. The search function also manages the search process.

Splunk Components

If you look at the below image, you will understand the different data pipeline stages under which various Splunk components fall under.



There are 3 main components in Splunk:

- Splunk Forwarder, used for data forwarding
- Splunk Indexer, used for Parsing and Indexing the data
- Search Head, is a GUI used for searching, analyzing and reporting

Splunk Forwarder

Splunk Forwarder is the component which you have to use for collecting the logs. Suppose, you want to collect logs from a remote machine, then you can accomplish that by using Splunk's remote forwarders which are independent of the main Splunk instance.

In fact, you can install several such forwarders in multiple machines, which will forward the log data to a Splunk Indexer for processing and storage. What if you want to do real-time analysis of the data? Splunk forwarders can be used for that purpose too. You can configure the forwarders to send data to Splunk indexers in real-time. You can install them in multiple systems and collect the data simultaneously from different machines in real time.

Compared to other traditional monitoring tools, Splunk Forwarder consumes very less cpu ~1-2%. You can scale them up to tens of thousands of remote systems easily, and collect terabytes of data with minimal impact on performance.

Now, let us understand the different types of Splunk forwarders.

Universal Forwarder

You can opt for an universal forwarder if you want to forward the raw data collected at the source. It is a simple component which performs minimal processing on the incoming data streams before forwarding them to an indexer.

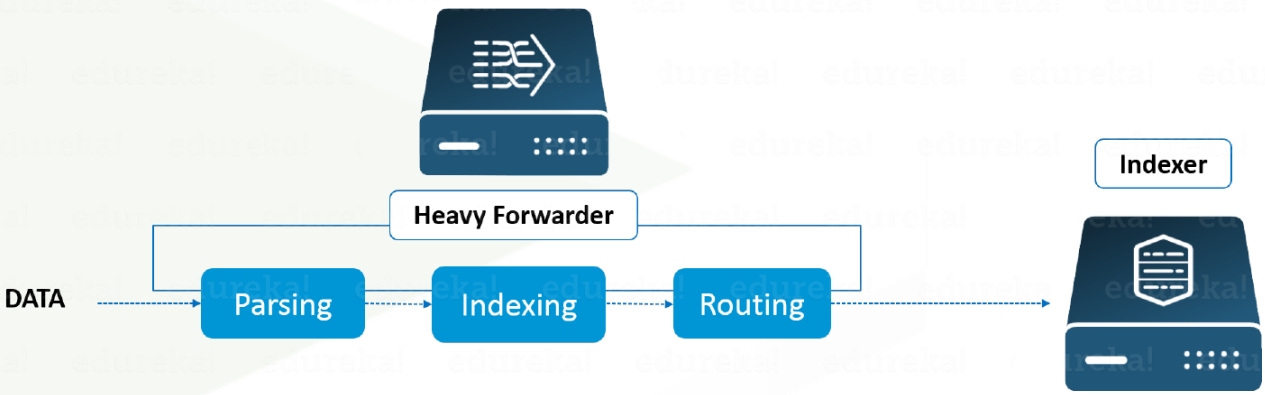
Data transfer is a major problem with almost every tool in the market. Since there is minimal processing on the data before it is forwarded, lot of unnecessary data is also forwarded to the indexer resulting in performance overheads.

Why go through the trouble of transferring all the data to the Indexers and then filter out only the relevant data? Wouldn't it be better to only send the relevant data to the Indexer and save on bandwidth, time and money? This can be solved by using Heavy forwarders which I have explained below.

Splunk Forwarder

Heavy Forwarder

You can use a Heavy forwarder and eliminate half your problems, because one level of data processing happens at the source itself before forwarding data to the indexer. Heavy Forwarder typically does parsing and indexing at the source and also intelligently routes the data to the Indexer saving on bandwidth and storage space. So when a heavy forwarder parses the data, the indexer only needs to handle the indexing segment.



Splunk Indexer

Splunk Indexer

Indexer is the Splunk component which you will have to use for indexing and storing the data coming from the forwarder. Splunk instance transforms the incoming data into events and stores it in indexes for performing search operations efficiently.

If you are receiving the data from a Universal forwarder, then the indexer will first parse the data and then index it. Parsing of data is done to eliminate the unwanted data. But, if you are receiving the data from a Heavy forwarder, the indexer will only index the data.

As the Splunk instance indexes your data, it creates a number of files. These files contain one of the below:

- Raw data in compressed form
- Indexes that point to raw data (index files, also referred to as tsidx files), plus some metadata files

These files reside in sets of directories called buckets.

Let me now tell you how Indexing works.

Splunk processes the incoming data to enable fast search and analysis. It enhances the data in various ways like:

- Separating the data stream into individual, searchable events
- Creating or identifying timestamps
- Extracting fields such as host, source, and sourcetype
- Performing user-defined actions on the incoming data, such as identifying custom fields, masking sensitive data, writing new or modified keys, applying breaking rules for multi-line events, filtering unwanted events, and routing events to specified indexes or servers

This indexing process is also known as event processing.

Another benefit with Splunk Indexer is data replication. You need not worry about loss of data because Splunk keeps multiple copies of indexed data. This process is called Index replication or Indexer clustering. This is achieved with the help of an Indexer cluster, which is a group of indexers configured to replicate each other's' data.

Splunk Search Head

Splunk Search Head

Search head is the component used for interacting with Splunk. It provides a graphical user interface to users for performing various operations. You can search and query the data stored in the Indexer by entering search words and you will get the expected result.

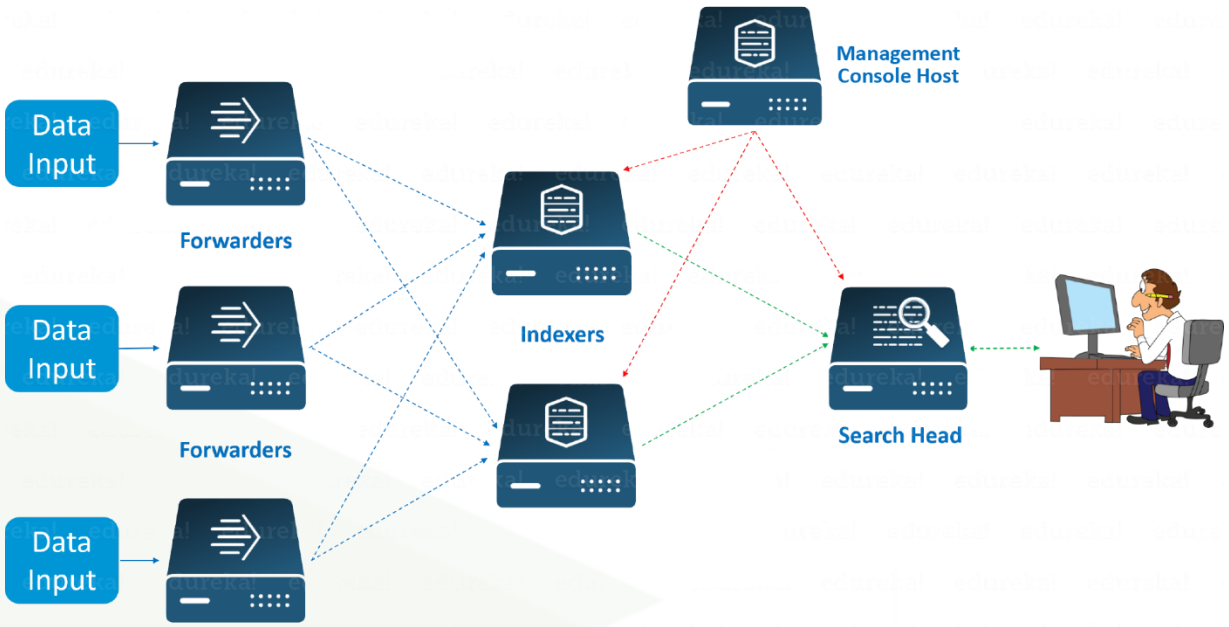
You can install the search head on separate servers or with other Splunk components on the same server. There is no separate installation file for search head, you just have to enable splunkweb service on the Splunk server to enable it.

A Splunk instance can function both as a search head and a search peer. A search head that performs only searching, and not indexing is referred to as a dedicated search head. Whereas, a search peer performs indexing and responds to search requests from other search heads.

In a Splunk instance, a search head can send search requests to a group of indexers, or search peers, which perform the actual searches on their indexes. The search head then merges the results and sends them back to the user. This is a faster technique to search data called distributed searching.

Search head clusters are groups of search heads that coordinate the search activities. The cluster coordinates the activity of the search heads, allocates jobs based on the current loads, and ensures that all the search heads have access to the same set of knowledge objects.

How Splunk Works?

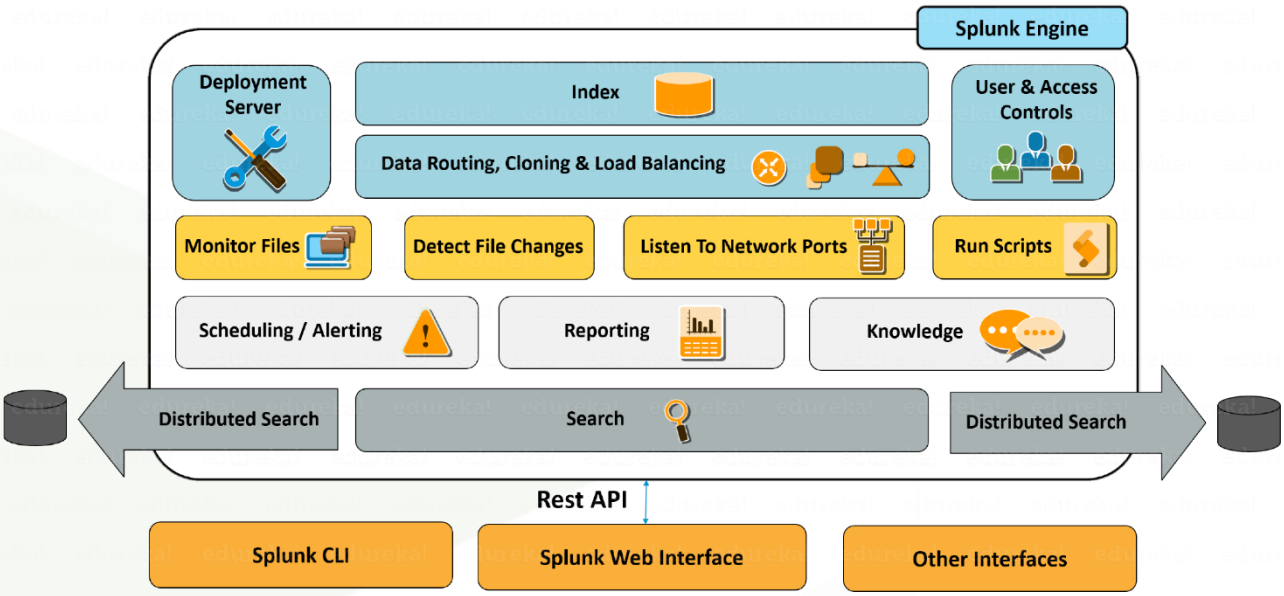


Look at the above image to understand the end to end working of Splunk. The images shows a few remote Forwarders that send the data to the Indexers. Based on the data present in the Indexer, you can use the Search Head to perform functions like searching, analyzing, visualizing and creating knowledge objects for Operational Intelligence.

The Management Console Host acts as a centralized configuration manager responsible for distributing configurations, app updates and content updates to the Deployment Clients. The Deployment Clients are Forwarders, Indexers and Search Heads.

Splunk Architecture

If you have understood the concepts explained above, you can easily relate to the Splunk architecture. Look at the image below to get a consolidated view of the various components involved in the process and their functionalities.



- Use Splunk CLI / Web Interface to communicate with the Search Head. The communication happens via a REST API
- Use the Search Head to make Distributed searches, set up Alerts and Reminders, perform Reporting and setup knowledge objects for gaining Operational Intelligence
- Run Scripts to automate data forwarding from remote Splunk forwarders to pre-defined Network Ports
- Monitor files and logs coming in at real time by specifying the path of a file to where the data needs to be stored
- Perform Routing, Cloning and Load Balancing of the forwarded data and customize your deployments by using a Deployment Server
- Create multiple users with restricted access to perform operations on the indexed data

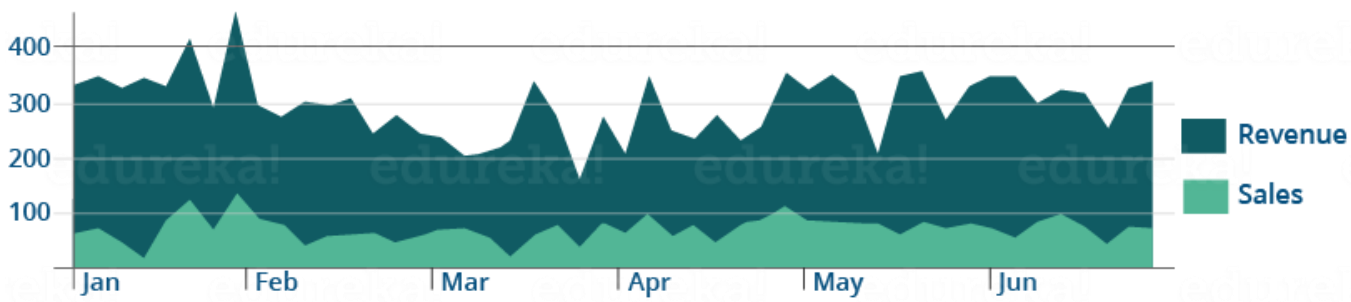
Splunk Knowledge Objects

We will understand the relevance of knowledge objects and the role it plays in bringing operational efficiency to your business. Here, I will explain the 3 main Knowledge objects: Splunk Timechart, Data models and Alert.
Look at the below image to get an idea on how knowledge objects work.

The data is first stored in an indexer and then you can write search queries and perform various operations on the data. You can set up knowledge objects to make operations smarter and to bring intelligence to your systems. These knowledge objects will monitor your events and give notifications when certain conditions occur. These results can be collated and visualized by creating reports and timecharts. Summing it all up, knowledge objects are at the heart of enriching your data and creating intelligence. Knowledge objects are user-defined entities used for extracting knowledge from existing or run-time data in order to achieve data enrichment. So, let's get started with the first knowledge object i.e Splunk Timechart.

Splunk Timechart
Let me explain all about Splunk Timechart and where they can be used. As an example, assume you have a large amount of data and you need to measure the number of sales and revenue generated on a monthly basis for an international apparel chain.
Splunk Timechart can be used to analyze if the performance metrics (sales and revenue in this case) has had an increasing or a decreasing trend over time.
Splunk Timechart refers to visualization of any data with respect to time.
In Timechart, data is represented in the form of line, area or column charts which is plotted against x-axis that is always a time field, whereas y-axis is the variable field.

For the above example, if we had to create the Timechart for both sales and revenue numbers of an apparel chain on a monthly basis, we can plot sales and revenue on the y-axis and time on x-axis. The Splunk Timechart visualization would look like:



Splunk Knowledge Objects

Splunk Timechart is often compared to Stats and Chart commands. The underlying structure between the three commands are quite different, you can refer to the table which explains the difference between them.

Stats	Chart	Timechart
Stats is a reporting command which is used to present data in a tabular format.	Chart displays the data in the form of a bar, line or area graph. It also gives the capability of generating a pie chart.	Timechart allows you to look at bar and line graphs. However, pie charts are not possible.
In Stats command, you can use multiple fields to build a table.	In Chart, it takes only 2 fields, each field on X and Y axis respectively.	In Timechart, it takes only 1 field since the X-axis is fixed as the time field.

Now, you know how data can be visualized using Splunk Timechart. Next, let's learn another knowledge object- Splunk Data Models. Let me help you to understand it properly.

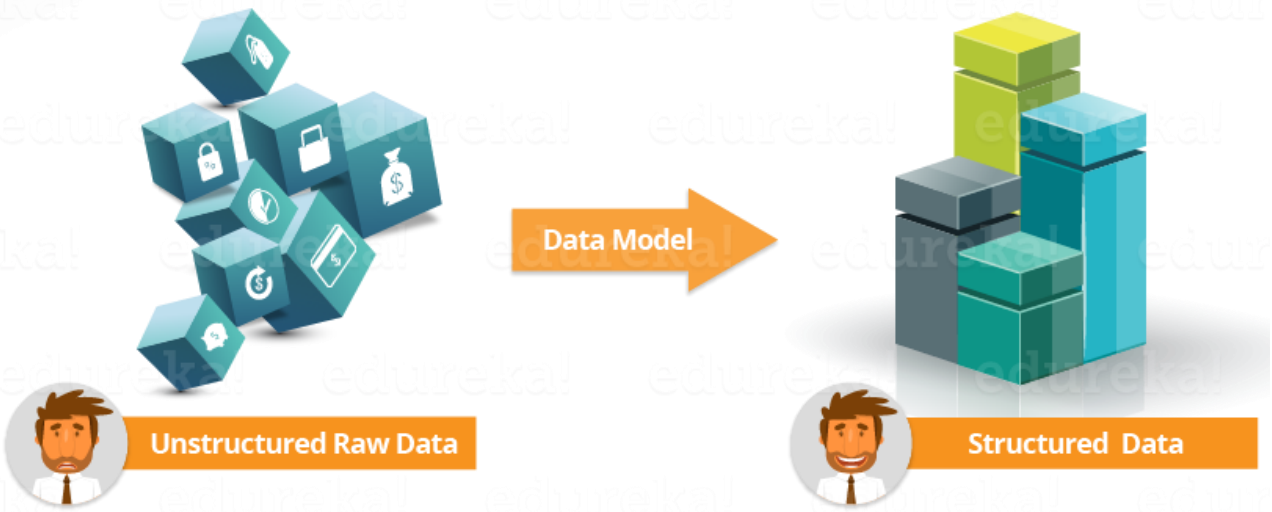
Splunk Data Models

Imagine you have a large amount of unstructured data which is critical to your business, and you want an easy way out to access that information without using complex search queries. This can be done using data models as it presents your data in a sorted and hierarchical manner. The key benefits of data models are:

It helps non-technical users to interact with data via pivot UI as they don't have to indulge in writing complex search queries. Pivots are a representation of the dataset in the form of tables, charts or any kind of visualization. Once you create a pivot, you can save it as a report or a dashboard panel.

Data models make it easy to re-use domain knowledge.

As Splunk Machine Analytics is difficult to understand, Data models help in providing a structured way to map complex data.



Splunk Knowledge Objects

Data models, as the name suggests, are models comprising of one or more datasets. Data models help in providing structure to your complex data and gives a broader view to understand the source, logic and domain knowledge. This can generate specialized searches based on the datasets. Data models is one of the major knowledge objects in Splunk as it combines other knowledge objects to provide a meaningful representation of your data. Data models are a combination of multiple knowledge objects such as Lookups, Event types, Field and more (refer to the below image).

By now, you might have understood what data models are and how they are useful. You must be also wondering whether you can generate your own data model. The answer is yes, you can design a new data model and also edit existing models. This can be done using the Data Model Editor. However, only users who have been assigned the Admin or Power role can create data models. Other users have to first manage their permissions in order to create data models.

Let me take you through the steps involved in creating data models:
Step 1: Go to Settings-> Data Models.

Step 2: Click 'New Data Model' to create a new data model.

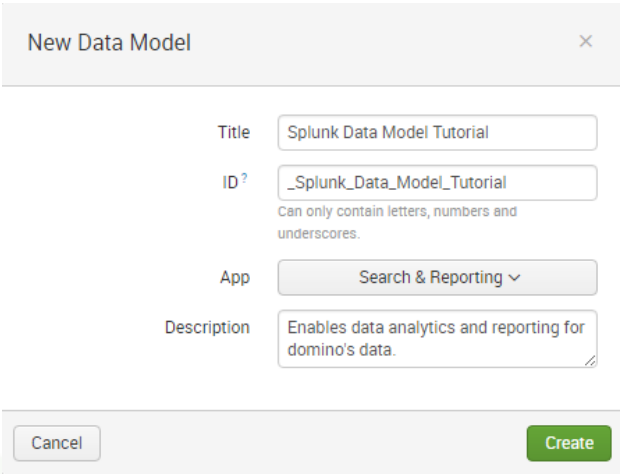
Step 3: Specify a 'Title' to your data model. You can use any character in the title, except an asterisk. The data model 'ID' field will get filled automatically as it is a unique identifier. It can only contain letters, numbers, and underscores. Spaces between characters are not allowed.

Step 4: Choose the 'App' you are working on currently. By default, it will be 'home'.

Step 5: Add a 'Description' to your data model.

Step 6: Click 'Create' and open the new data model in the Data Model Editor. Below, I have attached a screenshot that will help you understand the procedure for creating a data model:

Splunk Knowledge Objects



Let us now understand how data models can be used:
Splunk Use Case

Project Statement: Create data Models to solve the Big Data challenge of Domino's Pizza.
All of us are familiar with Domino's Pizza. With outlets in 81 countries, it is one of the largest pizza chains in the world. First of all, do you know how they collect data in real time from several touch points? Secondly, how do they examine real-time data globally to improve their customer performance?
Data models are ideal in such a scenario since they help in organizing and managing huge data in a structured manner.

For the Domino's example, it will return a JSON file for the "Domino's Data" data model. It has the model ID "Splunk Data Model Tutorial". Now, let us have a look at how data models structure the data:

```
| Splunk Data Model Tutorial
```


```
{[-]
  description:Enables data analytics and reporting for domino's data.
  displayName:Domino's Data
  modelName:Splunk Data Model Tutorial
  objectNameList:[[-]
    Customer_errors
    failed_orders
    telephone_orders
    website_orders
    promotional_offers
  ]
  objectsSummary:{[+]
  }
  objects:[[+]
  ]
}
Show as raw text
```

Splunk Knowledge Objects

*Note: The example of promotional data used is representative in nature and the data present might not be accurate.

In this example, if you send raw data to Splunk, the Data Model helps you create the structure by representing it in a JSON.

As you can see from the above image, there is an Object Name List that contains five subsets: Customer errors, failed orders, telephone orders, website orders and promotional offers.
The first subset, 'Customer errors' contains all the error data that customers face while processing an order.
The second subset, 'failed order' contains all the relevant data that deals with failed orders.
The third subset, 'telephone order' contains data that are processed via telephone.
'Website orders' will collect data ordered via domino's website and the fifth subset 'promotional offers' deals with all the coupons and offers from Domino's.

As data model divides the data into various subsets, it brings clarity to your data to help you analyze it in a hierarchical format, thereby solving Domino's Big Data Challenge.
By now, you would have understood how data can be visualized using Splunk Timechart and managed using Data Models. Next, let me explain another knowledge object i.e Splunk Alert, and how it can be used.

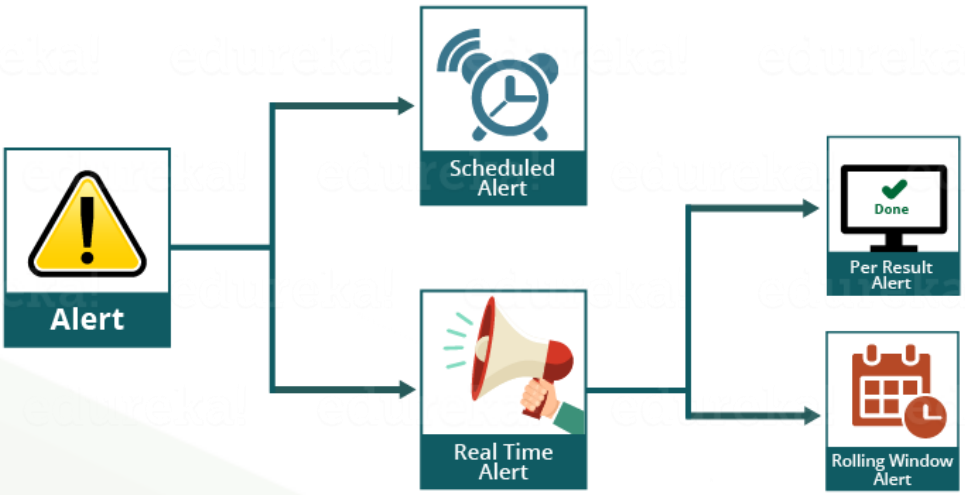
Splunk Alert

Let's consider a situation where you have a real-time application that needs to be up and running all the time. If it crashes or an error occurs during the operation, the problem needs to be identified and fixed immediately. But how will you know when something goes wrong? You cannot manually sit in front your system and monitor its state 24×7. The convenient way out would be to get notified instantly when something goes wrong. This is where Splunk Alerts can help you.
Alerts are used to monitor your events and perform actions when pre-defined conditions occur.

- Alerts are triggered based on search results and user-defined conditions.
- Alerts use saved searches to look for events in real time or at a scheduled time.
- There is a set of alert actions that helps you get notifications about a particular event. Alert action refers to the response that occurs when an alert is triggered. Some of the main alert actions are:
    - Email notifications: Send an email notification to specified recipients
    - Run scripts: Invoke a custom script
    - Send log events: Send log event to Splunk receiver endpoint
    - Webhook: Generic HTTP POST to a specified URL.

Now that you have a basic idea of what a Splunk alert is and how it works, let me go on further and list down the different types of alerts and when they are used:

Splunk Knowledge Objects



In the above image, you can see that there are two types of alerts: Scheduled alert and Real-time alert. Also, real-time alerts are further classified into per-result and rolling window alerts. Don't worry, I will explain each one of them in detail. First, let's start with scheduled alert:

Scheduled Alert: Suppose you belong to a retail firm and need to know the sales status at the end of every day, you can create a scheduled alert which will notify the total sales at 12 AM every day. This type of an alert can be used whenever an immediate response to an event is not required.

Real-time Alert: This type of an alert is used whenever an immediate response to an event is required. As I have mentioned earlier, real-time alerts are further classified into per-result alert and rolling window alert which are explained below:

Per-result Alert: Let us take a scenario where a networking website admin wants to know whenever the website is down with error '500'. Here, the admin can choose the per-result trigger condition so that every failed attempt can be tracked.
Per-result alert type can be used when you want a notification in real-time whenever the search returns a result that matches the search condition.

Rolling window Alert: Imagine you need to create an alert which notifies if the person has 5 consecutive failed login attempts in a span of 15 minutes. This can be done using this real-time alert with rolling time window triggering. It is used to monitor the results in a specific time interval like every 30 minutes or 1 hour whenever it matches the search condition.

Now that you know the different types of alerts, you must be wondering how an alert can be created.

Splunk Knowledge Objects

Problem Statement: Let's say you want alerts every time you experience an unusually high number of server errors on a web host.

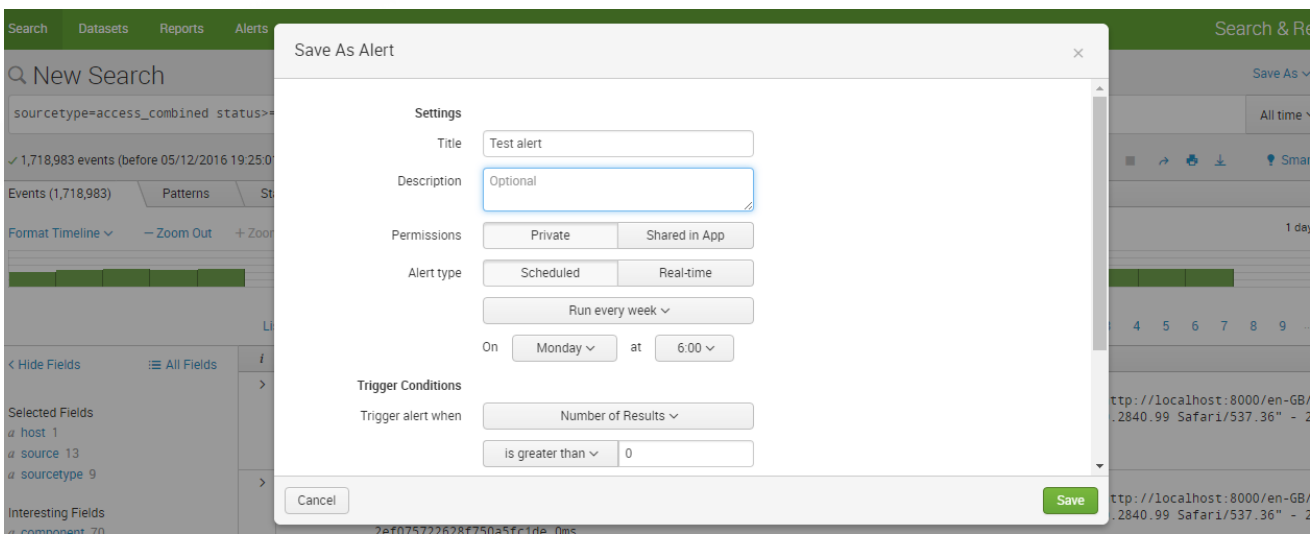
For these types of scenario, you can follow the below steps to create an alert.

Step 1: Open your Splunk Enterprise and write a search condition which you want as an alert.
In the above scenario, you can save the below search query for setting up an alert:
sourcetype=access_combined status>=500

Step 2: After writing the search query, when you click on the 'Save As' button, you will be asked a set of questions like alert title, description, type, permission and many more. You can use the cron schedule which simplifies this process and brings flexibility to perform the task at any point of time. It can be done by using some cron parameters corresponding to minute, hour, day of month, month and day of week. We use the cron expression to customize a schedule.

For example:
*/5**** – for every 5 minutes
*/30**** – for every 30 minutes
0 */12 *** – Every 12 hours, on the hour
0 */20 ***- Every 20 minutes, Monday through Friday.
0 9 1-7 * 1- First Monday of each month, at 9 AM.

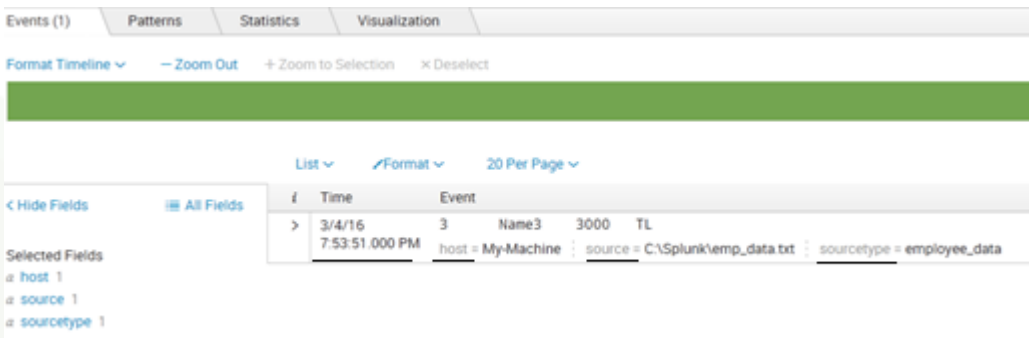
Below, I have attached a screenshot that will help you understand the procedure for creating an alert:



Splunk Knowledge Objects

Splunk Events

An event refers to any individual piece of data. The custom data that has been forwarded to Splunk Server are called Splunk Events. This data can be in any format, for example: a string, a number or a JSON object.
Let me show you how events look in Splunk:



As you can see in the above screenshot, there are default fields (Host, Source, Sourcetype and Time) which gets added after indexing. Let us understand these default fields:

1. Host: Host is a machine or an appliance IP address name from where the data comes. In the above screenshot, My-Machine is the host.
2. Source: Source is where the host data comes from. It is the full pathname or a file or directory within a machine.
   For example: C:\Splunk\emp_data.txt
3. Sourcetype: Sourcetype identifies the format of the data, whether it is a log file, XML, CSV or a thread field. It contains the data structure of the event.
   For example: employee_data
4. Index: It is the name of the index where the raw data is indexed. If you don't specify anything, it goes into a default index.
5. Time: It is a field which displays the time at which the event was generated. It is barcoded with every event and cannot be changed. You can rename or slice it for a period of time in order to change its presentation.
   For example: 3/4/16 7:53:51 represents the timestamp of a particular event.

Now, let us learn how Splunk Event types help you to group similar events.

Splunk Event Types

Assume you have a string containing the employee name and employee ID and you want to search the string using a single search query rather than searching them individually. Splunk Event types can help you here. They group these two separate events and you can save this string as a single event type (Employee_Detail).

Splunk Knowledge Objects

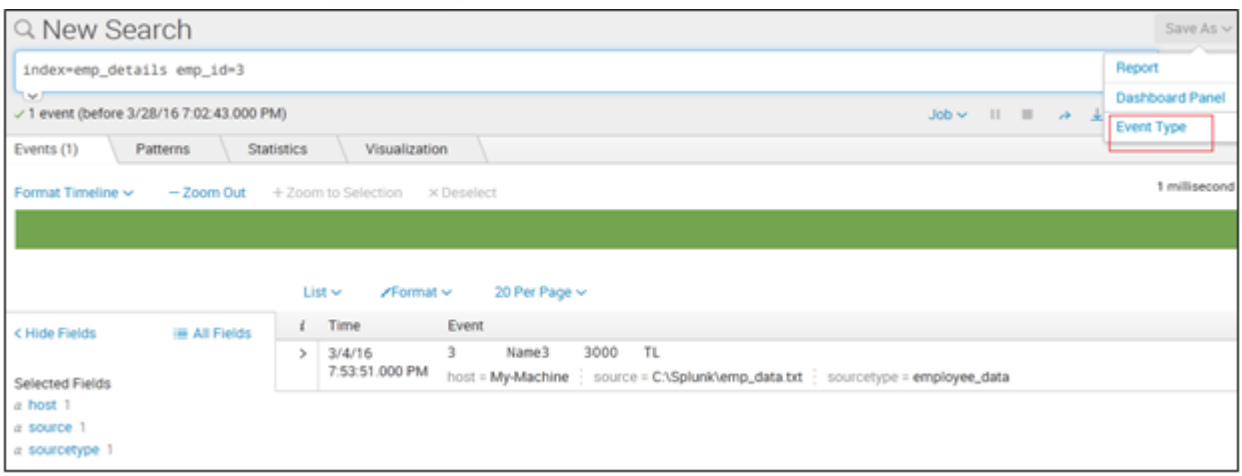
- Splunk event type refers to a collection of data which helps in categorizing events based on common characteristics.
- It is a user-defined field which scans through huge amount of data and returns the search results in the form of dashboards. You can also create alerts based on the search results.

Do note that you cannot use a pipe character or a sub search while defining an event type. But, you can associate one or more tags with an event type. Now, let us learn how these Splunk event types are created.
There are multiple ways to create an event type:

1. Using Search
2. Using Build Event Type Utility
3. Using Splunk Web
4. Configuration files (eventtypes.conf)

Let us go into more detail to understand it properly:

1. **Using Search:** We can create an event type by writing a simple search query.
Go through the below steps to create one:
> Run a search with the search string
For Example: index=emp_details emp_id=3;
> Click Save As and select Event Type.
You can refer to the below screenshot to get a better understanding:



2. **Using Build Event Type Utility:** The Build Event Type utility enables you to dynamically create event types based on events returned by searches. This utility also enables you to assign specific colors to event types.

You can find this utility in your search results. Let's go through the below steps:

Step1: Open the dropdown event menu
Step2: Find the down arrow next to the event timestamp
Step3: Click Build event type

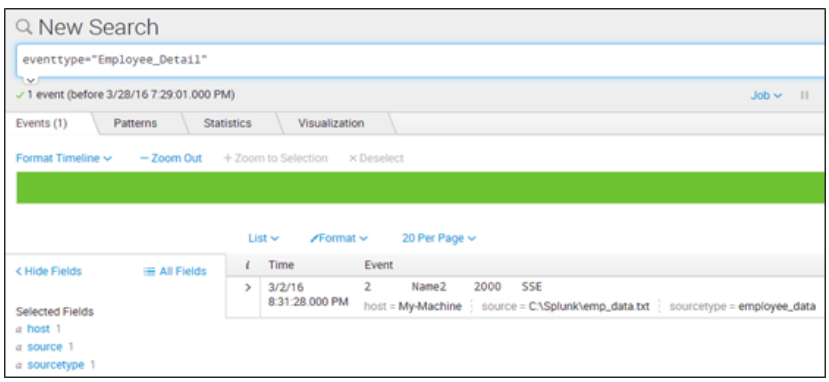
Once you click on 'Build Event Type' displayed in the above screenshot, it will return the selected set of events based on a particular search.

3. **Using Splunk Web:** This is the easiest way to create an event type.
For this, you can follow these steps:

» Go to Settings
» Navigate to Event Types
» Click New

Let me take the same employee example to make it easy.
Search query would be same in this case:
index=emp_details emp_id=3
Refer to the below screenshot to get a better understanding:

4. **Configuration files (eventtypes.conf):** You can create event types by directly editing eventtypes.conf configuration file in $SPLUNK_HOME/etc/system/local
For Example: "Employee_Detail"
Refer to the below screenshot to get a better understanding:



Splunk Knowledge Objects

Splunk Tags

You must be aware of what a tag means in general. Most of us use the tagging feature in Facebook to tag friends in a post or photo. Even in Splunk, tagging works in a similar fashion. Let's understand this with an example. We have an emp_id field for a Splunk index. Now, you want to provide a tag (Employee2) to emp_id=2 field/value pair. We can create a tag for emp_id=2 which can now be searched using Employee2.

- Splunk tags are used to assign names to specific fields and value combinations.
- It is the simplest method to get the results in pair while searching. Any event type can have multiple tags to get quick results.
- It helps to search groups of event data more efficiently.
- Tagging is done on the key value pair which helps to get information related to a particular event, whereas an event type provides the information of all the events associated with it.
- You can also assign multiple tags to a single value.

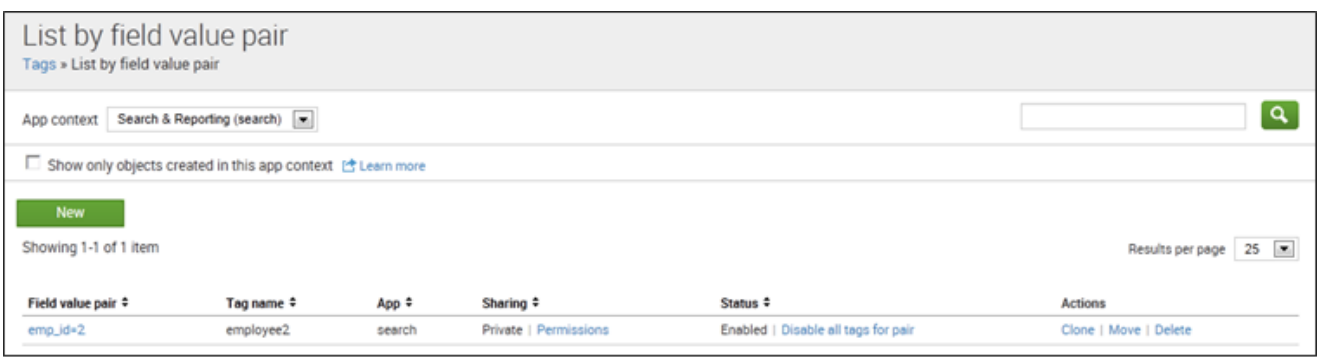
Look at the screenshot on right side to create a Splunk tag.
Go to Settings -> Tags

Now, you might have understood how a tag is created. Let us now understand how Splunk tags are managed. There are three views in Tag Page under Settings:

1. List by field value pair
2. List by tag name
3. All unique tag objects

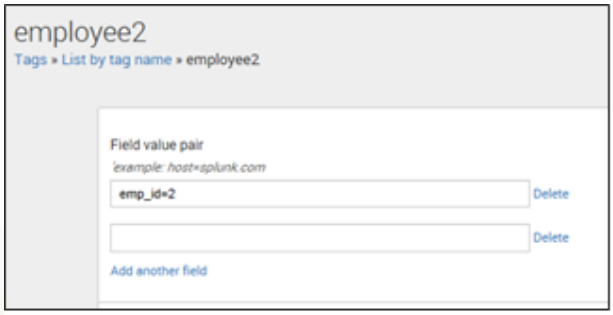
Let us get into more details and understand different ways to manage and get quick access to associations that are made between tags and field/value pairs.

1. **List by field value pair:** This helps you to review or define a set of tags for a field/value pair. You can see the list of such pairings for a particular tag.
Refer to the below screenshot to get a better understanding:

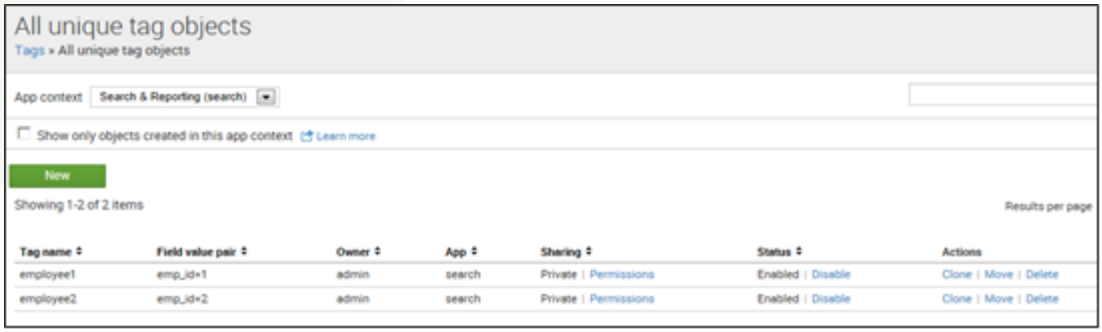


Splunk Knowledge Objects

2. **List by tag name:** It helps you to review and edit the sets of field/value pairs. You can find the list of field/value pairing for a particular tag by going to 'list by tag name' view and then click on the tag name. This takes you to the detail page of the tag.
Example: Open the detail page of employee 2 tag.
Refer to the below screenshot to get a better understanding:



3. **All unique tag objects:** It helps you to provide all the unique tag names and field/value pairings in your system. You can search a particular tag to quickly see all the field/value pairs with which it's associated. You can easily maintain the permissions, to enable or disable a particular tag.
Refer to the below screenshot to get a better understanding:



Now, there are 2 ways to search tags:

- If we need to search a tag associated with a value in any field, we can use:
  tag=<tagname>
  In the above example, it would be: tag=employee2
- If we are looking for a tag associated with a value in a specified field, we can use:
  tag::<field>=<tagname>
  In the above example, it would be: tag::emp_id=employee2

3 Concepts You Must Know As A Splunk Administrator

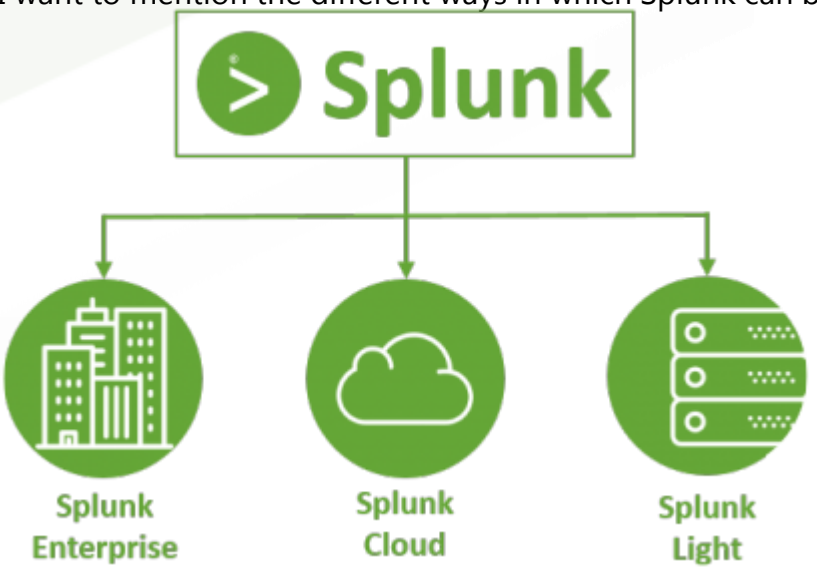
To become a seasoned Splunker, you need to be aware of certain administration aspects of Splunk. By learning these aspects you can manage the end to end working of Splunk and overcome the tag of 'only a Splunk user'. In this Splunk tutorial blog, I will cover the below mentioned concepts which are essential from a Splunk admin perspective:

- Licensing in Splunk
- Data ageing
- Configuration files in Splunk

Besides these concepts, I will also talk about the roles and responsibilities of a Splunk Admin.

Licensing In Splunk

First of all, Splunk licensing is the most important aspect in a Splunk environment because it is the license which controls how much data comes into your system and gets indexed on a daily basis. So it is important that you choose the best license plan be it for your organization or for yourself. Before I talk about the types of licenses available, I want to mention the different ways in which Splunk can be used.



Splunk Enterprise is an on-premise enterprise grade software license which you can buy and deploy locally in your infrastructure. With built in premium apps for security and operations, Splunk Enterprise is the most preferred model for medium to large scale infrastructure and businesses. Enterprise version can be purchased with either an annual subscription or a perpetual subscription and either way you can avail technical support from Splunk.

Splunk Tutorial

Splunk Cloud is the SaaS version of Splunk enterprise which is hosted and managed by Splunk themselves. Splunk cloud is similar to Splunk Enterprise from a feature perspective and has added advantages like ease of scalability, clustering, and zero down time. What this means is that, if you go for Splunk cloud you don't have to worry about updating your Splunk version or about data clustering and you can simply start off by mentioning the amount of data that you want to store and its retention period.

Splunk Light however is a smaller and a cheaper version of Splunk enterprise. In Splunk light, there is a cap of 20 GB that can be indexed in a day and it is limited to 5 native users. You cannot make use of premium apps and clustering is not possible. You have flexibility with respect to user roles, user interface is the same and even the commands are same except for enterprise class features and capabilities. You can also collect data from active directories, sensors and mobile apps.

When it comes to different types of licenses available, you can go for either:

- **Enterprise License**, which is the standard commercial/ corporate license or
- **Free License**, which is free, but has limitations with respect to features and functionality or
- **Forwarder License**(heavy forwarder/ universal forwarder), which can only forward the data to another instance/ host or
- **Enterprise Trial License**, which is a 60 day free trial of Splunk enterprise

License Master

Now that you have a fair idea of licensing in Splunk, it is time for me to introduce you to a very important Splunk role called **License master** which is the next section of this Splunk tutorial blog. License master is responsible for making sure that the Splunk environment always adheres to its license i.e the right amount of data gets indexed by Splunk.

Consider a scenario where you need to index 250 GB of data on one day, 500 GB of data the next day and 1 TB of data on another day and then it suddenly drops to 100 GB on some other day. Then, you should ideally purchase a 1 TB/day license, so that you can index upto 1 TB of data on any given day.

The license master thus ensures that you do not index more than the agreed volume of data as per your license. If you violate your licensing terms, then you will be given a notification on your Search head or Web UI saying that you have exceeded the indexing volume and you either need to reduce the amount of data coming in or you need to buy a higher capacity of license.

In case the license master is unreachable, then it is not possible to search the data. However, the data coming in to the Indexer will not be affected. The data will continue to flow into your Splunk deployment and the Indexers will continue to index the data as usual.

License Slave And License Pool

License slave and license pool are the next topics of discussion in this Splunk tutorial. **License slave** is that Splunk role which **reports back to the master**. Important point to note is that license slave is another name for an indexer. Search peer is also a name given to the Indexer.

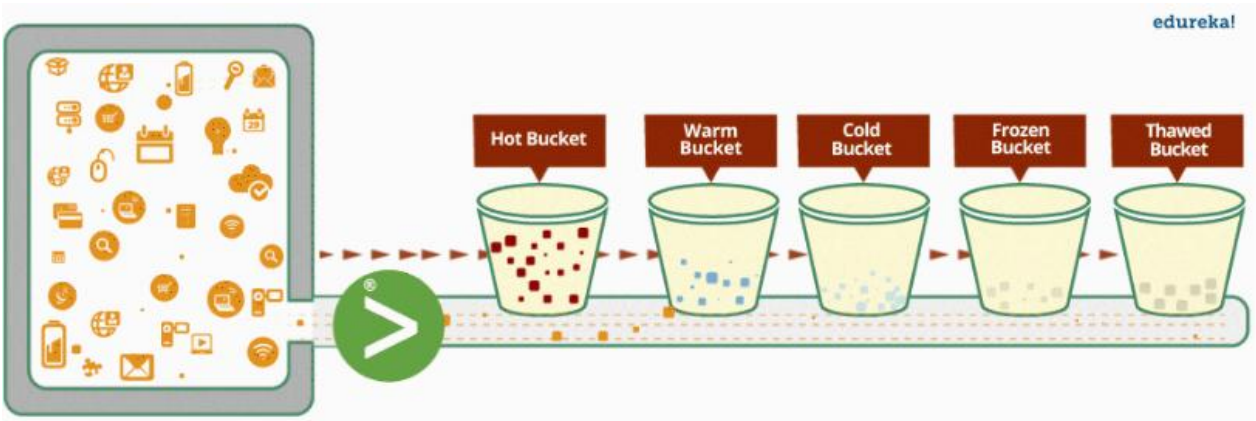
License master will determine the amount of data indexed by sending a message to the license slave. The slave would reply by saying that in 'x' minutes, 'y' amount of data has been ingested and cumulatively, the master will be collecting responses from all the slaves and be in sync.

The master and slaves together form a license pool. If any slave has exceeded the data limit or is about to exceed the limit, a warning message will be displayed across the pool. You will be notified which pool is facing an issue, so that you can re-arrange the licenses within your pool.

Consider a scenario where all three Indexers within your pool have a license of indexing only 20 GB of data/day. On a random day if Indexer-1 ingests 30 GB data, which is more than its licensed volume, then you can re-arrange the licenses within your pool by allocating additional resources from Indexer-2 and Indexer-3 to Indexer-1. Thus by sharing the additional volume from other Indexers, you can balance the licenses within your pool.

Data Ageing In Splunk

Licensing takes care of the data coming into an indexer. But, what happens next? Data is surely not going to stay in the there forever. So, in this section of Splunk tutorial blog, I am going to discuss how data in Splunk ages throughout its lifetime.



When data enters an Indexer, it passes through different stages called **buckets**. Data in each bucket is stored in a different directory. The different bucket stages are: **hot**, **warm**, **cold**, **frozen** and **thawed**. Over time, buckets 'roll' from one stage to the next stage in the above mentioned order.

- Data goes into a **hot bucket** the first time it gets indexed. Data in hot buckets are both searchable and are actively being written to. An index can have several hot buckets open at a time
- When the hot bucket reaches a certain size or attains a certain age or if *splunkd* gets restarted, the hot bucket rolls to **warm bucket**. At this point, a new hot bucket is created. Warm buckets are searchable, but are not actively written to. There can be many warm buckets
- When the indexer reaches a maximum number of warm buckets or after certain age, they begin to roll to**cold buckets**. Do note that it always selects the oldest warm bucket to roll to cold
- After a set period of time or when data in these cold buckets grow less important, they roll to either **frozen buckets** or **thawed buckets**. Data in these buckets are either archived or deleted

The bucket ageing policy, which determines when a bucket moves from one stage to the next, can be modified by editing the attributes in indexes.conf file. If you are wondering what .conf or configuration files are, then don't worry. That is the next topic of discussion in this Splunk tutorial blog. Alternatively, you can also set the path of these bucket directories from Splunk web.

Configuration Files

Configuration files play a very important role in the functioning of your Splunk environment. These configuration files contain Splunk system settings, configuration settings and app configuration settings. You can edit these files and accordingly changes will be reflected in your Splunk environment. However, the changes made to configuration files will be taken into effect only if the Splunk instance is restarted. These configuration files can be found in the below places:

- $SPLUNK_HOME/etc/system/default
- $SPLUNK_HOME/etc/system/local
- $SPLUNK_HOME/etc/apps/

Path where these configuration files are stored is consistent in all operating systems. They are always stored in $SPLUNK_HOME, the directory where Splunk is installed.

- By default, all the settings are stored in **$SPLUNK_HOME/etc/system/default**. Configuration files in this folder will contain default system settings such as host information, source, source type, timestamp, server properties and alerts.

- If you make any changes to the default configurations, then it is always advisable to put those files in**$SPLUNK_HOME/etc/system/local**. The reason is when you upgrade your Splunk to the next version, then the files in $SPLUNK_HOME/etc/system/default will be overwritten but Splunk will not overwrite files in the local directory. These files will be saved and can be used during your next restart.

- However, all your customizations, personalized configurations, built apps and saved searches will be stored in**$SPLUNK_HOME/etc/apps/default**. For example, if you create a search app, then it will be stored in this path: $SPLUNK_HOME/etc/apps/search/default. Similarly, it is highly advisable that Splunk admins create a backup in the local folder in apps directory: $SPLUNK_HOME/etc/apps/search/local.

There is another path where configuration files are stored: **$SPLUNK_HOME/etc/users**. In this folder, user specific settings in UI, user specific configurations and preference will be stored. As an administrator you can also store user specific settings for multiple Splunk users.

Everything you see in the UI is configurable/ modifiable via the configuration file. In fact there are a lot of options that cannot be edited via UI, but it is possible via CLI or by directly editing a configuration file. I prefer to use CLI over Web UI because it is easier at times. For example, you can restart Splunk from server settings or server controls, but I find it easier to go to command prompt and run a command. This is just one of the many things for which I prefer to use CLI over Web UI.

As I mentioned earlier, all configuration files have .conf extension and they store various settings and some of the most commonly stored settings are:

- System Settings
- Authentication
- Authorization
- Indexes
- Deployment Configurations
- Cluster Configurations
- Saved Searches

Basically, whatever you run and save would be stored in a file. For example, saved searches would be stored in a folder called savedsearches.conf and indexer location will be stored in indexes.conf and so on.

Configuration File Structure

Each configuration file consists of one or more stanzas. Each stanza begins with a stanza header enclosed in square brackets followed by file settings. Each setting is an attribute value pair that specifies particular configuration settings. Below is the basic pattern of a configuration file containing more than one stanza.

```
[stanza1]
<attr1> = <value>
<attr2> = <value> ...
[stanza2]
<attr1> = <value>
<attr2> = <value> ...
```

Important point to note is that attributes are in camel case, where the first word is completely in small letters and subsequent words will have first letter as capital and other letters will be small. This nomenclature should be followed accurately in case of troubleshooting and that is why this topic is very important for a Splunk administrator.

Configuration File Precedence

Configuration file precedence is the next topic in this Splunk tutorial blog. It is the most important concept from troubleshooting perspective and it is an interviewer's favourite question.

System local directory — ***highest priority***
App local directories
App default directories
System default directory — ***lowest priority***

It is important that you always remember, files in ***System local*** directory have the highest priority. ***App local***directories have the next priority and then comes ***App default*** directories. So, if there are no settings in App local directory, it will look for settings in App default directory and if there is no setting there, it will look for it in the next precedence level i.e ***System default*** directory.

If you put any settings or configurations in default directory and you want it to be respected then, it would not happen because if there is any stanza which Is overlapping or present in any of the above directories, then configurations in those files will be preferred.

Ways To Edit/ Set Your Configurations

You can edit your configurations either via Web UI or through configuration files. When you do it through web, Splunk will do a validation right away and tell you if the changes are a good move or a bad move. It will highlight fields and suggest which fields you need to fill. Thus, there is less chance of human errors while editing via Web UI.

When we change configurations in Splunk Web, that change will be written to a copy of the configuration file for that setting. Splunk software creates a copy of this configuration file, writes the change to that copy, and adds it to a directory under $SPLUNK_HOME/etc/.

We can also edit the configuration files directly for changing any setting. For some customizations that Splunk Web does not support, we must edit the configuration files directly. The direct configuration change however, requires frequent restart of Splunk Services.

One of the cons of editing configuration files directly is that you might end up breaking your Splunk instance. So it is important that you modify the configurations correctly. The best practice is to read out the documentation clearly, look up an example and setup a test environment before you make any changes directly into the configuration files of a production environment.

Common Configuration Files

This is the final topic under configuration files. Here, I will tell you what are the most common configuration files and and what settings and stanzas they contain.
Most commonly used configuration files in Splunk are:

- inputs.conf
- outputs.conf
- props.conf
- savedsearches.conf
- indexes.conf
- authentication.conf
- authorize.conf

Inputs.conf

This configuration file is used to capture data. For example, if you want to monitor a file coming in, or if you want to monitor certain ports then, you can do that by writing settings in this file. These can be set on Forwarders and Indexers. Commonly used attributes in inputs.conf are:

1. Host: This is used to identify the host from where the data is coming
2. Index: This is used to set the path where you want the data to be stored
3. Source: This is used to identify the path from where you are collecting data
4. Sourcetype: This is used to determine the format of incoming data.

```
[monitor://$SPLUNK_HOME\var\log\splunk]
index = _internal

[monitor://$SPLUNK_HOME\etc\splunk.version]
_TCP_ROUTING = *
index = _internal
sourcetype=splunk_version
```

Above is a screenshot of inputs.conf file whose stanza header is *monitor*. I have set the path of*index* to *_internal*, which is a default location where data is stored and *sourcetype=splunk_version*, which is the format of data in which I want to read.

Outputs.conf

This file is mostly set on forwarders or set on instances that forward data. The settings in this file tell Splunk instance where to send the data. For example, it can contain a string of IP addresses or machines names of your Indexers. You can also write settings here to perform load balancing, where the data will be routed to another Indexer if forwarding to one Indexer fails. You can also set the frequency of data forwarding and configure TCP acknowledgements. The levels of TCP output stanzas in outputs.conf are:

1. Global: [tcpout]
2. Target group
3. Single server

```
[tcpout]
maxQueueSize = auto
forwardedindex.0.whitelist = .*
forwardedindex.1.blacklist = _.*
forwardedindex.2.whitelist = (_audit|_internal|_introspection)
forwardedindex.filter.disable = false
indexAndForward = false
autoLBFrequency = 30
```

Above is a screenshot of outputs.conf file where I am *blacklisting* and *whitelisting data* forwarded to certain indexers. I have also set the f*requency of data forwarding* to *30*.

Props.conf

Props.conf is used to maintain and track the properties of every event that comes in. Props.conf applies "rules" while the data is getting parsed. For example, route specific events to a particular index, create index-time field extraction, you can specify how events gets timestamped, the format of the timestamp and how the events should break. Basically, props.conf will apply your configuration settings to your data while it is indexed. Props.conf can be written on indexer, heavy forwarder, and/or search head.

```
[postfix_syslog]
MAX_TIMESTAMP_LOOKAHEAD = 32
TIME_FORMAT = %b %d %H:%M:%S
TRANSFORMS-host = syslog-host
REPORT-syslog = syslog-extractions
SHOULD_LINEMERGE = False

[sendmail_syslog]
MAX_TIMESTAMP_LOOKAHEAD = 32
SHOULD_LINEMERGE = False
TIME_FORMAT = %b %d %H:%M:%S
TRANSFORMS = syslog-host
REPORT-syslog = sendmail-extractions
```

Above is a screenshot of props.conf file whose stanza header is *postfix_syslog*. *MAX_TIMESTAMP_LOOKAHEAD = 32* specifies to look 32 characters ahead into an event and identify the time stamp in the format of month, day, hour, minute and second.

Savedsearches.conf

Savedsearches.conf file is used to store pre-defined actions which need to be performed. For example, rules to perform saved searches will be present here. You can also define scheduled reports and alerts here. These files can be replicated in different instances to perform searches across different environments. Commonly used attributes in Savedsearches.conf are:

1. Scheduling options
2. Notification options
3. Settings of Email action
4. Settings of Script action

```
[Chart_Report]
action.email.reportServerEnabled = 0
action.email.useNSSubject = 1
alert.track = 0
display.general.type = visualizations
display.visualizations.charting.chart = bar
request.ui_dispatch_app = search
request.ui_dispatch_view = search
search = index=test_index | chart count by host
```

Above is a screenshot of savedsearches.conf file with a stanza header *chat_report*. I have set various attributes here to either true or false which is represented by boolean values. I have set the type of visualization chart as*bar* and I have defined search commands at the bottom of the image with certain parameters.

Indexes.conf

This file is written to manage and configure indexer settings. Indexes can be created by directly editing this file. You can set the homepath, where the indexer should store the data by default. You can also set cold, frozen and thawed paths for the indexer to store the aged out data. Commonly used attributes in indexes.conf are:

1. Per Index Options
2. Home Path
3. Cold Path
4. Thawed Path

```
[main]
homePath   = $SPLUNK_DB\defaultdb\db
coldPath   = $SPLUNK_DB\defaultdb\colddb
thawedPath = $SPLUNK_DB\defaultdb\thaweddb
tstatsHomePath = volume:_splunk_summaries\defaultdb\datamodel_summary
maxMemMB = 20
maxConcurrentOptimizes = 6
maxHotIdleSecs = 86400
maxHotBuckets = 10
maxDataSize = auto_high_volume
```

Authenticate.conf

This is the file where you have to store your LDAP authentication settings. If you are connecting to an active directory or LDAP directory protocol, then those attributes need to be setup in this file. Rules present here are also used to map users to various roles. Commonly used attributes in authentication.conf are:

- LDAP settings
- Map Roles

```
[OpenLDAP]
bindDN = uid=directory_bind,cn=users,dc=osx,dc=company,dc=com
bindDNpassword = directory_bind_account_password
groupBaseFilter =
groupNameAttribute = cn
SSLEnabled = 0
port = 389
userBaseDN = cn=users,dc=osx,dc=company,dc=com
host = hostname_OR_IP
userBaseFilter =
userNameAttribute = uid
groupMappingAttribute = uid
groupBaseDN = dc=osx,dc=company,dc=com
groupMemberAttribute = memberUid
realNameAttribute = cn
```

Authorize.conf

This file is used for providing the user interface and configuring their roles and capabilities. Basically, user specific preferences and differences are stored in this file. You can set the different levels of access for each user over here and control what each user views and accesses.

```
[role_admin]
accelerate_datamodel   = enabled
admin_all_objects      = enabled
change_authentication  = enabled
edit_deployment_client = enabled
list_deployment_client = enabled
edit_deployment_server = enabled
list_deployment_server = enabled
edit_dist_peer         = enabled
edit_forwarders        = enabled
edit_httpauths         = enabled
edit_input_defaults    = enabled
edit_monitor           = enabled
edit_roles             = enabled
edit_scripted          = enabled
edit_search_server     = enabled
edit_server            = enabled
edit_splunktcp         = enabled
edit_splunktcp_ssl     = enabled
edit_tcp               = enabled
edit_udp               = enabled
edit_user              = enabled
edit_view_html         = enabled
edit_web_settings      = enabled
```

In the next section of this Splunk tutorial blog, I will talk about the two different times at which processing happens. They are Index-time and Search-time.

Index-Time vs. Search-Time

Index-time processing is the processing of data that happens before the event is actually indexed. Examples of this are data fields which get extracted as and when the data comes into the index like source, host and timestamp.

Following are the processes that occur during index time:
1. Default field Source type customization
2. Index-time field extraction
3. Event timestamping
4. Event line breaking
5. Event segmentation

Search-time processing is the processing of data that happens while a search is running. Examples of this are any kind of searches or alerts or reminders or lookups.

Following are the processes which occur during search time:
1. Event segmentation (also happens at index time)
2. Event type matching
3. Search-time field extraction
4. Field aliasing
5. Field lookups from external data sources
6. Source type renaming
7. Tagging

I am pretty sure that by now, most of you have understood what factors affect a Splunk deployment and as Splunk users what operations are a must know other than the basic functionality of searching, analyzing, reporting and visualizing. Hence, I would like to conclude this blog by summarizing the roles and responsibilities of a Splunk administrator.

Splunk Admin Roles And Responsibilities

First thing that you should be aware of is that a Splunk Admin is not responsible for creating any reports, dashboards, knowledge objects or saved searches because they are user based roles.

But, Splunk admins are responsible for the following:

- Performing the Installation
- Managing the licenses. Creating the license master, connecting the license slaves to them and managing the license pools
- Managing the authentication method responsible for the environment
- Creating users and assigning them access
- Managing the user's roles and capabilities so that they perform their jobs and operations smoothly
- Setting up Search heads and Indexers and make sure they are connected and are able to communicate properly
- Ensuring that Search head is able to search the data from Indexer
- Ensuring that the Indexer is receiving the data properly from the forwarder
- Ensuring that forwarders and deployment servers are configured correctly, they are efficient, and making sure that time synchronization between them is happening correctly
- Ensuring that forwarder has right set of data and nomenclature is correct
- Ensuring that deployment servers managing the forwarders are collecting data timely and reporting happens in order
- Maintaining the clustering between search head clustering and index clusters
- Configuring data inputs for collecting data from forwarders and from UI and ensuring that data is coming in correctly is being read by Splunk
- Coordinating troubleshooting activities with Splunk Technical team.

Splunk Interview Questions

One thing is for certain: Implementing Splunk will transform your business and take it to the next level. But, the question is do you posses the skills to be a Splunker? If yes, then prepare yourselves for the most gruesome job interview because the competition is intense. You can start by going through the most common Splunk interview questions which are mentioned in this blog.

Splunk Interview Questions

The questions covered in this blog post have been shortlisted after collecting inputs from many industry experts to help you ace your interview. In case you want to learn the basics of Splunk then, you can start off by reading the first blog in my Splunk tutorial series: What Is Splunk? All the best!

Q1. What is Splunk? Why is Splunk used for analyzing machine data?

This question will most likely be the first question you will be asked in any Splunk interview. You need to start by saying that:
Splunk is a platform which allows people to get visibility into machine data, that is generated from hardware devices, networks, servers, IoT devices and other sources
Splunk is used for analyzing machine data because it can give insights into application management, IT operations, security, compliance, fraud detection, threat visibility etc
To learn more about this topic, you can read this blog: What Is Splunk?

Q2. Explain how Splunk works.

This is a sure-shot question because your interviewer will judge this answer of yours to understand how well you know the concept. The Forwarder acts like a dumb agent which will collect the data from the source and forward it to the Indexer. The Indexer will store the data locally in a host machine or on cloud. The Search Head is then used for searching, analyzing, visualizing and performing various other functions on the data stored in the Indexer.
You can find more details about the working of Splunk here: Splunk Architecture: Tutorial On Forwarder, Indexer And Search Head.

Q3. What are the components of Splunk?

Splunk Architecture is a topic which will make its way into any set of Splunk interview questions. As explained in the previous question, the main components of Splunk are **Forwarders**, **Indexers** and **Search Heads**. You can then mention that another component called **Deployment Server**(or **Management Console Host**) will come into the picture in case of a larger environment. Deployment servers:

- Act like an antivirus policy server for setting up Exceptions and Groups, so that you can map and create different set of data collection policies each for either a windows based server or a linux based server or a solaris based server
- Can be used to control different applications running in different operating systems from a central location
- Can be used to deploy the configurations and set policies for different applications from a central location.

Making use of deployment servers is an advantage because connotations, path naming conventions and machine naming conventions which are independent of every host/machine can be easily controlled using the deployment server.

Q4. Why use only Splunk? Why can't I go for something that is open source?

This kind of question is asked to understand the scope of your knowledge. You can answer that question by saying that Splunk has a lot of competition in the market for analyzing machine logs, doing business intelligence, for performing IT operations and providing security. But, there is no one single tool other than Splunk that can do all of these operations and that is where Splunk comes out of the box and makes a difference. With Splunk you can easily scale up your infrastructure and get professional support from a company backing the platform. Some of its competitors are Sumo Logic in the cloud space of log management and ELK in the open source category. You can refer to the below table to understand how Splunk fares against other popular tools feature-wise. The detailed differences between these tools are covered in this blog: Splunk vs ELK vs Sumo Logic.

Q5. Which Splunk Roles can share the same machine?

This is another frequently asked Splunk interview question which will test the candidate's hands-on knowledge. In case of small deployments, most of the roles can be shared on the same machine which includes **Indexer**, **Search Head** and **License Master.** However, in case of larger deployments the preferred practice is to host each role on stand alone hosts. Details about roles that can be shared even in case of larger deployments are mentioned below:

- Strategically, **Indexers** and **Search Heads** should have physically dedicated machines. Using Virtual Machines for running the instances separately is not the solution because there are certain guidelines that need to be followed for using computer resources and spinning multiple virtual machines on the same physical hardware can cause performance degradation.
- However, a **License master** and **Deployment server** can be implemented on the same virtual box, in the same instance by spinning different Virtual machines.

- You can spin another virtual machine on the same instance for hosting the **Cluster master** as long as the**Deployment master** is not hosted on a parallel virtual machine on that same instance because the number of connections coming to the **Deployment server** will be very high.

- This is because the **Deployment server** not only caters to the requests coming from the **Deployment master**, but also to the requests coming from the **Forwarders**.

Q6. What are the unique benefits of getting data into a Splunk instance via Forwarders?

You can say that the benefits of getting data into Splunk via forwarders are **bandwidth throttling**, **TCP connection** and an **encrypted SSL connection** for transferring data from a forwarder to an indexer. The data forwarded to the indexer is also load balanced by default and even if one indexer is down due to network outage or maintenance purpose, that data can always be routed to another indexer instance in a very short time. Also, the forwarder caches the events locally before forwarding it, thus creating a temporary backup of that data.

Q7. Briefly explain the Splunk Architecture

You can find the detailed explanation in this link: Splunk Architecture: Tutorial On Forwarder, Indexer And Search Head.

Q8. What is the use of License Master in Splunk?

License master in Splunk is responsible for making sure that the right amount of data gets indexed. Splunk license is based on the data volume that comes to the platform within a 24hr window and thus, it is important to make sure that the environment stays within the limits of the purchased volume.

Consider a scenario where you get 300 GB of data on day one, 500 GB of data the next day and 1 terabyte of data some other day and then it suddenly drops to 100 GB on some other day. Then, you should ideally have a 1 terabyte/day licensing model. The license master thus makes sure that the indexers within the Splunk deployment have sufficient capacity and are licensing the right amount of data.

Q9. What happens if the License Master is unreachable?

In case the license master is unreachable, then it is just not possible to search the data. However, the data coming in to the Indexer will not be affected. The data will continue to flow into your Splunk deployment, the Indexers will continue to index the data as usual however, you will get a warning message on top your Search head or web UI saying that you have exceeded the indexing volume and you either need to reduce the amount of data coming in or you need to buy a higher capacity of license.

Basically, the candidate is expected to answer that the indexing does not stop; only searching is halted.

Q10. Explain 'license violation' from Splunk perspective.

If you exceed the data limit, then you will be shown a 'license violation' error. The license warning that is thrown up, will persist for 14 days. In a commercial license you can have 5 warnings within a 30 day rolling window before which your Indexer's search results and reports stop triggering. In a free version however, it will show only 3 counts of warning.

Q11. Give a few use cases of Knowledge objects.

Knowledge objects can be used in many domains. Few examples are:

Physical Security: If your organization deals with physical security, then you can leverage data containing information about earthquakes, volcanoes, flooding, etc to gain valuable insights

Application Monitoring: By using knowledge objects, you can monitor your applications in real-time and configure alerts which will notify you when your application crashes or any downtime occurs

Network Security: You can increase security in your systems by blacklisting certain IPs from getting into your network. This can be done by using the Knowledge object called lookups

Employee Management: If you want to monitor the activity of people who are serving their notice period, then you can create a list of those people and create a rule preventing them from copying data and using them outside

Easier Searching Of Data: With knowledge objects, you can tag information, create event types and create search constraints right at the start and shorten them so that they are easy to remember, correlate and understand rather than writing long searches queries. Those constraints where you put your search conditions, and shorten them are called event types.

These are some of the operations that can be done from a non-technical perspective by using knowledge objects. Knowledge objects are the actual application in business, which means Splunk interview questions are incomplete without Knowledge objects. In case you want to read more about the different knowledge objects available and how they can be used, read this blog: Splunk Tutorial On Knowledge Objects

Q12. Why should we use Splunk Alert? What are the different options while setting up Alerts?

This is a common question aimed at candidates appearing for the role of a Splunk Administrator. Alerts can be used when you want to be notified of an erroneous condition in your system. For example, send an email notification to the admin when there are more than three failed login attempts in a twenty-four hour period. Another example is when you want to run the same search query every day at a specific time to give a notification about the system status.

Different options that are available while setting up alerts are:

- You can create a web hook, so that you can write to hipchat or github. Here, you can write an email to a group of machines with all your subject, priorities, and body of the message
- You can add results, .csv or pdf or inline with the body of the message to make sure that the recipient understands where this alert has been fired, at what conditions and what is the action he has taken
- You can also create tickets and throttle alerts based on certain conditions like a machine name or an IP address. For example, if there is a virus outbreak, you do not want every alert to be triggered because it will lead to many tickets being created in your system which will be an overload. You can control such alerts from the alert window.

You can find more details about this topic in this blog: Splunk alerts.

Q13. Explain Workflow Actions

Workflow actions is one such topic that will make a presence in any set of Splunk Interview questions. Workflow actions is not common to an average Splunk user and can be answered by only those who understand it completely. So it is important that you answer this question aptly.

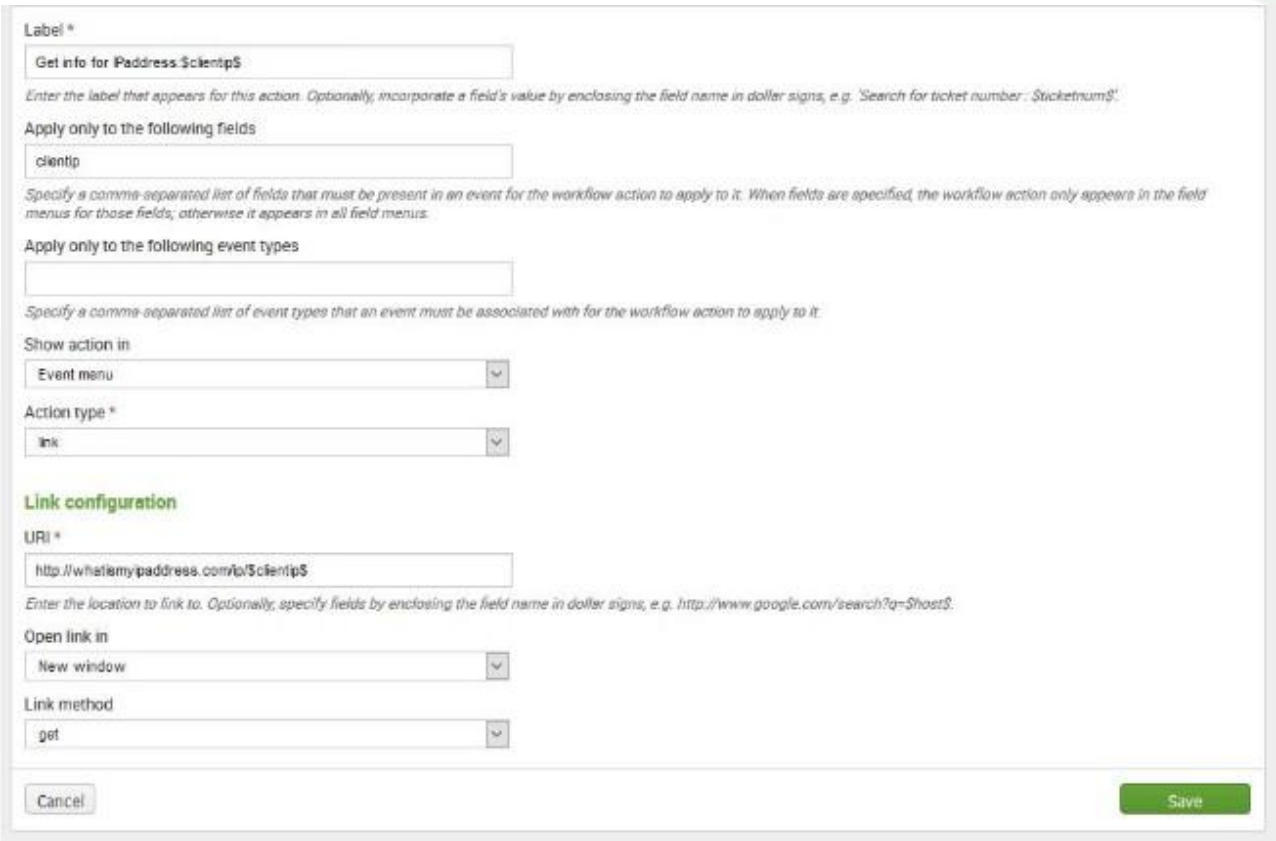
You can start explaining Workflow actions by first telling why it should be used.

Once you have assigned rules, created reports and schedules then what? It is not the end of the road! You can create workflow actions which will automate certain tasks. For example:

- You can do a double click, which will perform a drill down into a particular list containing user names and their IP addresses and you can perform further search into that list
- You can do a double click to retrieve a user name from a report and then pass that as a parameter to the next report
- You can use the workflow actions to retrieve some data and also send some data to other fields. A use case of that is, you can pass latitude and longitude details to google maps and then you can find where an IP address or location exists.

The screenshot below shows the window where you can set the workflow actions.

Splunk Interview Questions



Q14. Explain Data Models and Pivot

Data models are used for creating a structured hierarchical model of your data. It can be used when you have a large amount of unstructured data, and when you want to make use of that information without using complex search queries.
A few use cases of Data models are:

Create Sales Reports: If you have a sales report, then you can easily create the total number of successful purchases, below that you can create a child object containing the list of failed purchases and other views

Set Access Levels: If you want a structured view of users and their various access levels, you can use a data model

Enable Authentication: If you want structure in the authentication, you can create a model around VPN, root access, admin access, non-root admin access, authentication on various different applications to create a structure around it in a way that normalizes the way you look at data.
So when you look at a data model called authentication, it will not matter to Splunk what the source is, and from a user perspective it becomes extremely simple because as and when new data sources are added or when old one's are deprecated, you do not have to rewrite all your searches and that is the biggest benefit of using data models and pivots.

On the other hand with pivots, you have the flexibility to create the front views of your results and then pick and choose the most appropriate filter for a better view of results. Both these options are useful for managers from a non-technical or semi-technical background. You can find more details about this topic in this blog: Splunk Data Models.

Q15. Explain Search Factor (SF) & Replication Factor (RF)

Questions regarding Search Factor and Replication Factor are most likely asked when you are interviewing for the role of a Splunk Architect. SF & RF are terminologies related to Clustering techniques (Search head clustering & Indexer clustering).

- The search factor determines the number of searchable copies of data maintained by the indexer cluster. The default value of search factor is 2. However, the Replication Factor in case of Indexer cluster, is the number of copies of data the cluster maintains and in case of a search head cluster, it is the minimum number of copies of each search artifact, the cluster maintains
- Search head cluster has only a Search Factor whereas an Indexer cluster has both a Search Factor and a Replication Factor
- Important point to note is that the search factor must be less than or equal to the replication factor

Q16. Which commands are included in 'filtering results' category?

There will be a great deal of events coming to Splunk in a short time. Thus it is a little complicated task to search and filter data. But, thankfully there are commands like 'search', 'where', 'sort' and 'rex' that come to the rescue. That is why, filtering commands are also among the most commonly asked Splunk interview questions.

Search: The 'search' command is used to retrieve events from indexes or filter the results of a previous search command in the pipeline. You can retrieve events from your indexes using keywords, quoted phrases, wildcards, and key/value expressions. The 'search' command is implied at the beginning of any and every search operation.

Where: The 'where' command however uses 'eval' expressions to filter search results. While the 'search' command keeps only the results for which the evaluation was successful, the 'where' command is used to drill down further into those search results. For example, a 'search' can be used to find the total number of nodes that are active but it is the 'where' command which will return a matching condition of an active node which is running a particular application.

Sort: The 'sort' command is used to sort the results by specified fields. It can sort the results in a reverse order, ascending or descending order. Apart from that, the sort command also has the capability to limit the results while sorting. For example, you can execute commands which will return only the top 5 revenue generating products in your business.

Rex: The 'rex' command basically allows you to extract data or particular fields from your events. For example if you want to identify certain fields in an email id: *abc@edureka.co*, the 'rex' command allows you to break down the results as *abc* being the user id, *edureka.co* being the domain name and *edureka* as the company name. You can use rex to breakdown, slice your events and parts of each of your event record the way you want.

Q17. What is a *lookup* command? Differentiate between *inputlookup* & outputlookup commands

Lookup command is that topic into which most interview questions dive into, with questions like: Can you enrich the data? How do you enrich the raw data with external lookup?

You will be given a use case scenario, where you have a csv file and you are asked to do lookups for certain product catalogs and asked to compare the raw data & structured csv or json data. So you should be prepared to answer such questions confidently.

Lookup commands are used when you want to receive some fields from an external file (such as CSV file or any python based script) to get some value of an event. It is used to narrow the search results as it helps to reference fields in an external CSV file that match fields in your event data.

An ***inputlookup*** basically takes an input as the name suggests. For example, it would take the product price, product name as input and then match it with an internal field like a product id or an item id. Whereas, an***outputlookup*** is used to generate an output from an existing field list. Basically, inputlookup is used to enrich the data and outputlookup is used to build their information.

Q18. What is the difference between 'eval', 'stats', 'charts' and 'timecharts' command?

'Eval' and 'stats' are among the most common as well as the most important commands within the Splunk SPL language and they are used interchangeably in the same way as 'search' and 'where' commands.

- At times 'eval' and 'stats' are used interchangeably however, there is a subtle difference between the two. While 'stats' command is used for computing statistics on a set of events, 'eval' command allows you to create a new field altogether and then use that field in subsequent parts for searching the data.
- Another frequently asked question is the difference between 'stats', 'charts' and 'timecharts' commands. The difference between them is mentioned in the table below.

Splunk Interview Questions

Stats	Chart	Timechart
Stats is a reporting command which is used to present data in a tabular format.	Chart displays the data in the form of a bar, line or area graph. It also gives the capability of generating a pie chart.	Timechart allows you to look at bar and line graphs. However, pie charts are not possible.
In Stats command, you can use multiple fields to build a table.	In Chart, it takes only 2 fields, each field on X and Y axis respectively.	In Timechart, it takes only 1 field since the X-axis is fixed as the time field.

Q19. What are the different types of Data Inputs in Splunk?

This is the kind of question which only somebody who has worked as a Splunk administrator can answer. The answer to the question is below.
The obvious and the easiest way would be by using files and directories as input Configuring Network ports to receive inputs automatically and writing scripts such that the output of these scripts is pushed into Splunk is another common way
But a seasoned Splunk administrator, would be expected to add another option called windows inputs. These windows inputs are of 4 types: registry inputs monitor, printer monitor, network monitor and active directory monitor.

Q20. What are the defaults fields for every event in Splunk?

There are about 5 fields that are default and they are barcoded with every event into Splunk.
They are host, source, source type, index and timestamp.

Q21. Explain file precedence in Splunk.

File precedence is an important aspect of troubleshooting in Splunk for an administrator, developer, as well as an architect. All of Splunk's configurations are written within plain text .conf files. There can be multiple copies present for each of these files, and thus it is important to know the role these files play when a Splunk instance is running or restarted. File precedence is an important concept to understand for a number of reasons:

- To be able to plan Splunk upgrades
- To be able to plan app upgrades
- To be able to provide different data inputs and
- To distribute the configurations to your splunk deployments.

To determine the priority among copies of a configuration file, Splunk software first determines the directory scheme. The directory schemes are either a) Global or b) App/user.

When the context is global (that is, where there's no app/user context), directory priority descends in this order:

1. System local directory — *highest priority*
2. App local directories
3. App default directories
4. System default directory — *lowest priority*

When the context is app/user, directory priority descends from user to app to system:

1. User directories for current user — *highest priority*
2. App directories for currently running app (local, followed by default)
3. App directories for all other apps (local, followed by default) — for exported settings only
4. System directories (local, followed by default) — *lowest priority*

Q22. How can we extract fields?

You can extract fields from either event lists, sidebar or from the settings menu via the UI.
The other way is to write your own regular expressions in props.conf configuration file.

Q23. What is the difference between Search time and Index time field extractions?

As the name suggests, Search time field extraction refers to the fields extracted while performing searches whereas, fields extracted when the data comes to the indexer are referred to as Index time field extraction. You can set up the indexer time field extraction either at the forwarder level or at the indexer level.

Another difference is that Search time field extraction's extracted fields are not part of the metadata, so they do not consume disk space. Whereas index time field extraction's extracted fields are a part of metadata and hence consume disk space.

Q24. Explain how data ages in Splunk?

Data coming in to the indexer is stored in directories called buckets. A bucket moves through several stages as data ages: **hot**, **warm**, **cold**, **frozen** and **thawed**. Over time, buckets 'roll' from one stage to the next stage.

- The first time when data gets indexed, it goes into a **hot bucket**. Hot buckets are both searchable and are actively being written to. An index can have several hot buckets open at a time
- When certain conditions occur (for example, the hot bucket reaches a certain size or *splunkd* gets restarted), the hot bucket becomes a **warm bucket** ("rolls to warm"), and a new hot bucket is created in its place. Warm buckets are searchable, but are not actively written to. There can be many warm buckets

- Once further conditions are met (for example, the index reaches some maximum number of warm buckets), the indexer begins to roll the warm buckets to **cold** based on their age. It always selects the oldest warm bucket to roll to cold. Buckets continue to roll to cold as they age in this manner
- After a set period of time, cold buckets roll to **frozen**, at which point they are either archived or deleted.

The bucket aging policy, which determines when a bucket moves from one stage to the next, can be modified by editing the attributes in indexes.conf.

Q25. What is summary index in Splunk?
Summary index is another important Splunk interview question from an administrative perspective. You will be asked this question to find out if you know how to store your analytical data, reports and summaries. The answer to this question is below.

The **biggest advantage** of having a summary index is that you can retain the analytics and reports even after your data has aged out. For example:
- Assume that your data retention policy is only for 6 months but, your data has aged out and is older than a few months. If you still want to do your own calculation or dig out some statistical value, then during that time, summary index is useful
- For example, you can store the summary and statistics of the percentage growth of sale that took place in each of the last 6 months and you can pull the average revenue from that. That average value is stored inside summary index.

But the **limitations** with summary index are:
- You cannot do a needle in the haystack kind of a search
- You cannot drill down and find out which products contributed to the revenue
- You cannot find out the top product from your statistics
- You cannot drill down and nail which was the maximum contribution to that summary.

That is the use of Summary indexing and in an interview, you are expected to answer both these aspects of benefit and limitation.

Q26. How to exclude some events from being indexed by Splunk?

You might not want to index all your events in Splunk instance. In that case, how will you exclude the entry of events to Splunk.
An example of this is the debug messages in your application development cycle. You can exclude such debug messages by putting those events in the null queue. These null queues are put into **transforms.conf** at the forwarder level itself.

If a candidate can answer this question, then he is most likely to get hired.

Q27. What is the use of Time Zone property in Splunk? When is it required the most?

Time zone is extremely important when you are searching for events from a security or fraud perspective. If you search your events with the wrong time zone then you will end up not being able to find that particular event altogether. Splunk picks up the default time zone from your browser settings. The browser in turn picks up the current time zone from the machine you are using. Splunk picks up that timezone when the data is input, and it is required the most when you are searching and correlating data coming from different sources. For example, you can search for events that came in at 4:00 PM IST, in your London data center or Singapore data center and so on. The timezone property is thus very important to correlate such events.

Q28. What is Splunk App? What is the difference between Splunk App and Add-on?

Splunk Apps are considered to be the entire collection of reports, dashboards, alerts, field extractions and lookups.
Splunk Apps minus the visual components of a report or a dashboard are Splunk Add-ons. Lookups, field extractions, etc are examples of Splunk Add-on.

Any candidate knowing this answer will be the one questioned more about the developer aspects of Splunk.

Q29. How to assign colors in a chart based on field names in Splunk UI?

You need to assign colors to charts while creating reports and presenting results. Most of the time the colors are picked by default. But what if you want to assign your own colors? For example, if your sales numbers fall below a threshold, then you might need that chart to display the graph in red color. Then, how will you be able to change the color in a Splunk Web UI?

You will have to first edit the panels built on top of a dashboard and then modify the panel settings from the UI. You can then pick and choose the colors. You can also write commands to choose the colors from a palette by inputting hexadecimal values or by writing code. But, Splunk UI is the preferred way because you have the flexibility to assign colors easily to different values based on their types in the bar chart or line chart. You can also give different gradients and set your values into a radial gauge or water gauge.

Q30. What is sourcetype in Splunk?

Now this question may feature at the bottom of the list, but that doesn't mean it is the least important among other Splunk interview questions.

Sourcetype is a default field which is used to identify the data structure of an incoming event. Sourcetype determines how Splunk Enterprise formats the data during the indexing process. Source type can be set at the forwarder level for indexer extraction to identify different data formats. Because the source type controls how Splunk software formats incoming data, it is important that you assign the correct source type to your data. It is important that even the indexed version of the data (the event data) also looks the way you want, with appropriate timestamps and event breaks. This facilitates easier searching of data later.

For example, the data maybe coming in the form of a csv, such that the first line is a header, the second line is a blank line and then from the next line comes the actual data. Another example where you need to use sourcetype is if you want to break down date field into 3 different columns of a csv, each for day, month, year and then index it. Your answer to this question will be a decisive factor in you getting recruited.

I hope this set of Splunk interview questions will help you in preparing for your interview. You can check out the different job roles, a Splunk skilled professional is liable for by reading this blog on Splunk Careers.

Thank you!

For Queries during the session and class recording:
Post on Twitter @edurekaIN: #askEdureka
Post on Facebook /edurekaIN

For more details please contact us:
US : 1800 275 9730 (toll free)
INDIA : +91 88808 62004
Email us : sales@edureka.co

edureka!