Splunk Essentials

Splunk is a powerful tool that is used extensively to search, monitor, and analyze any machine data.

This book is designed to introduce you quickly to the benefits of using the Splunk Enterprise system. Understanding this technology will allow you to engage with your important data and ensure that it is collected, stored, managed, reported on, and utilized well to enable you to make better business decisions. By equipping you with this knowledge, you will be better prepared to tackle data issues in the fast-paced business world of today. You will learn about various vital topics such as data collection, managing apps, creating reports, and analyzing data using Splunk. You will also be equipped with skills to help you obtain a Twitter API key for use with the Twitter app for Splunk.

Who this book is written for

This book is intended for a business person, analyst, or student who wants to quickly learn how to use Splunk to manage data.

It would be helpful to have a bit of familiarity with basic computer concepts, but no prior experience of Splunk is required.

What you will learn from this book

- Use Splunk to collect all types of data
- Create reports in Splunk to help understand what your data says and develop business insights
- Design dashboards in Splunk to present data in many useful ways on one interface
- Set up alerts to communicate problems or issues with a system
- Utilize the Report Builder to create values over time, top values, and rare values reports
- Understand how to bring in and analyze live streaming data
- Familiarize yourself with data storage in Splunk

In this package, you will find:

- The author biography
- A preview chapter from the book, Chapter 2 “An Introduction to Indexing and Searching”
- A synopsis of the book’s content
- More information on Splunk Essentials

About the Author

Betsy Page Sigman is a distinguished professor at the McDonough School of Business at Georgetown University in Washington, D.C. She has taught courses in statistics, project management, databases, and electronic commerce for the last 16 years, and has been recognized with awards for teaching and service. Before arriving at Georgetown, she worked at George Mason University, the U.S. Bureau of the Census, Decision/Making/Information, the American Enterprise Institute, and the Social Science Data Center (now Roper Center) at the University of Connecticut.

Recent publications include a Harvard Business case study and a Harvard Business review article, articles in the Decision Sciences Journal of Innovative Education and Decision Line, and a case study in Educause Review Online. Additionally, she is a frequent media commentator on technological issues and big data.
A big thank you to Richard Harvey, Mohammed Fahad, Utkarsha S. Kadam and the other editors and staff at Packt Publishing for your help in every step along the way to finishing this book. Thanks also to my colleagues and students at the McDonough School of Business at Georgetown University. Thanks especially to Bill Garr, Rob Pongsajapan, Marie Selvanandin, and Kristin Bolling, and the Center for New Designs in Learning and Scholarship (CNDLS), for exploring the exciting world of big data and Splunk together. It has been a wonderful place to learn, grow, and serve for the last 16 years. I need to thank my brothers, Tim and Rick Page, for being there to challenge and encourage me throughout my life. Most of all, I want to thank my brilliant and wonderful husband, Chuck, my astonishing daughter and son-in-law, Page and Daniel Thies, and my three sons. Johnny, thanks for always inspiring me technologically; Richard, thanks for your sense of humor that keeps us all laughing; and James, thanks for always being there for all of us. Edward, the grandson who lights up all our lives, is too young to read this now. He was born into an extraordinary world—one that I hope and pray technology will continue to improve.
Splunk Essentials

Splunk Enterprise Software, or Splunk, is an extremely powerful tool for searching, exploring, and visualizing data of all types. Splunk is becoming increasingly popular, as more and more businesses, both large and small, discover its ease and usefulness. Analysts, managers, students, and others can quickly learn how to use the data from their systems, networks, web traffic, and social media to make attractive and informative reports.

This is a straightforward, practical, and quick introduction to Splunk that should have you making reports and gaining insights from your data in no time. Throughout the book, we have provided step-by-step instructions, pointers, and illustrations to help you on your way.

What This Book Covers

Chapter 1, Introducing Splunk, introduces you to Splunk Enterprise Software and its powerful capabilities.

Chapter 2, An Introduction to Indexing and Searching, explains indexing in Splunk and shows you how to do a simple search.

Chapter 3, More on Using Search, further develops your skills in using Splunk's search command.

Chapter 4, Reports in Splunk, shows you how to create reports and dashboards.

Chapter 5, Splunk Applications, explores the wide variety of Splunk apps and add-ons.

Chapter 6, Using the Twitter App, illustrates how to use the Twitter app for analyzing live Twitter data streams.

Chapter 7, Monitoring and Creating Alerts in Splunk, instructs you on how to monitor systems and create useful alerts that can help control processes and prevent problems.
An Introduction to Indexing and Searching

In the previous chapter, we showed you how to bring in data from different sources and index it. Data must be turned into information and made relevant before we can use it successfully, as raw data in files or streams won't help us answer the questions that arise while analyzing the data for our businesses or organizations. We need to collect the data that we are interested in before we can analyze it. And this is where Splunk’s capabilities shine.

In this chapter, we will cover these important next steps for using Splunk:

- Collecting data to search
- How Splunk indexes data
- Using indexed data
- Specifying a sourcetype
- SPL and what it is
- How to perform your own simple search
Collecting data to search

In the previous chapter, we showed you how to bring data from a file into Splunk. We also discussed how data from virtually any source can be brought into Splunk. The following diagram shows the various types of data (such as Twitter, Facebook, RSS, network, and many others that are pictured) that can be easily integrated into Splunk, then searched, added to other data, monitored, and then used for creating dashboards, reports, and other kinds of analyses. Notice that the storage capabilities of Splunk are also included in the screenshot:

Many types of data can be used with Splunk

(Source: http://www.businessinsider.com/investors-are-eating-up-these-two-enterprise-tech-ipsos-heres-why-2012-4)

Almost any kind of data can be entered into Splunk, and then stored, searched, analyzed, and reported on. Additionally, you will also see a logo labeled Hadoop. You may have even heard this term before, in connection to big data. Hadoop, an Apache open source software package, is a method of storing and analyzing big data that has a lot in common with Splunk. Hadoop and Splunk can work together with the application called Hunk, which we'll talk about later in Chapter 5, Splunk Applications.
Before going on, it is important to note that one of the most important
capabilities of Splunk is that you can bring in large amounts of data from
several different sources and easily store it and analyze it in one location.

Indexing data with Splunk

When we processed the data file in the previous chapter, we uploaded the data and
Splunk processed and indexed the data. It is worthwhile to examine a bit further
what happens when indexing takes place:

1. To create an index actually requires two steps: parsing and indexing. The
   parsing part includes the adding of metadata that always includes the host,
   source, and sourcetype. The indexing portion takes the events, splits them
   into searchable segments, and finally creates the index and raw data files.
2. After this happens, the data can then be easily searched through Splunk.
   The following screenshot shows how the data is brought into Splunk by
   forwarders. A forwarder takes data from a source, such as a web server,
   and then sends it to a full instance of Splunk:

   ![Machine-generated IT Data Contains A Categorical Record of Activity and Behavior](source)

   This diagram shows how Splunk uses forwarders to take data from complex IT infrastructures
   and then sends it to be indexed and searched.

   (Source: [http://www.businessinsider.com/investors-are-eating-up-these-two-enterprise-tech-ipo-s-heres-why-2012-4](http://www.businessinsider.com/investors-are-eating-up-these-two-enterprise-tech-ipo-s-heres-why-2012-4))
Using indexed data

Once you have indexed a file successfully, as we did in Chapter 1, Introducing Splunk, it will be listed with any other indexes that have already been created, and you can now do searches on it.

Viewing a list of indexes

To see a list of your indexes, follow the steps given next:

1. First, visit the home page (a quick way is to just click on the Splunk icon).
2. Find the Settings drop-down menu.
3. To finally view your indexes, under Data, select Indexes.

You will see a screen like the one shown here:
You will see a number of internal indexes, which are preceded by an underscore. These indexes include the logs and metrics that record Splunk’s internal processing. Notice that the non-internal indexes listed here are history, main, splunklogger, summary, and Twitter. The main index is often selected as a default index. The history and splunklogger indexes were used for previous versions, but are not generally used now. The summary index stores events that have been aggregated using a transforming command to set up searches over long time periods. And the Twitter index is created when you use the Twitter app, as we will do in Chapter 6, Using the Twitter App.

Bringing in indexed data
We need to bring in the indexed data before we can search it. If we do not specify an index, index=main, which is set here to be searched by default via the indexes searched by the default setting, will be assumed. To bring in all the indexed data, we could specify index=* If we want to bring in the Twitter index (which you will create in Chapter 6, Using the Twitter App), we can just specify index=twitter.

When we processed the data from our file in Chapter 1, Introducing Splunk, it was indexed by default. So we do not have to specify this index when we use it as our data source and go on to learn more about how to search in Splunk.

Specifying a sourcetype
Identifying a sourcetype for data is important because it tells Splunk how to format the data. The sourcetype is one of the default fields assigned to each event that is processed. Splunk uses it to decide how it is going to process your data. The correct sourcetype is usually assigned automatically when indexing data, for Splunk comes with many predefined sourcetypes.

One such sourcetype is access_combined. Using this, Splunk can analyze combined access log files, the types that are part of the massive amount of data exhaust created by web servers such as Microsoft IIS or Apache. Some common sourcetypes include the following:

<table>
<thead>
<tr>
<th>Sourcetype</th>
<th>Used for</th>
</tr>
</thead>
<tbody>
<tr>
<td>access_combined</td>
<td>A standardized format for text files used by HTTP web servers when generating server log files</td>
</tr>
<tr>
<td>cisco_syslog</td>
<td>Cisco standard system logs</td>
</tr>
<tr>
<td>apache_error</td>
<td>Errors</td>
</tr>
</tbody>
</table>
Sometimes the access_combined sourcetype specifies _wcookie, which indicates that each cookie set during an HTTP request is logged. The data we brought in and indexed in Chapter 1, Introducing Splunk, was specified access_combined_wcookie. To specify this particular sourcetype, type the following into the search bar:

\texttt{sourcetype=access\_combined\_wcookie}

This will pull up the web server logs with this sourcetype so you can then use them for analysis.

When adding custom data formats, such as logs from applications built in-house, you can specify a descriptive sourcetype for the technology as the sourcetype is what is being used to differentiate the data type. For Cisco iOS devices, you can use \texttt{sourcetype=cisco:ios}.

**What is Search Processing Language (SPL)?**

After we have our data indexed, we can begin to search. The default application for Splunk is the search application. It is assumed that you are doing a search unless you indicate otherwise. Searches are made using the Search Processing Language (SPL). Through search, Splunk lets the user comb through the indexed data to find what he or she needs for answering questions.

In the simplest of terms, if you only put the term failed in the search box, for instance, it knows you want to do a search and will automatically search for \texttt{failed} anywhere in the data, and will return each event that fits with failed highlighted.

**Using pipes when processing data with Splunk**

However, SPL can be used to do much more advanced searches and analyses as well. Pipes are a way to do this. The pipe character (\texttt{|}) can be used to chain together different commands in a search. In the previous simple search and in our following search, a search is implied in the first pipe, but the term \texttt{search} itself is left out. In other words, in the following search, we could say \texttt{buttercupgames} or search \texttt{buttercupgames} and it means the same thing. There are many other commands that can be used as well, and they are listed and discussed as follows. Consider the following piped command:

\texttt{buttercupgames | timechart count by itemId limit=10}
The command following the pipe character acts on the data after it comes from the previous pipe. Hence, a pipe can refer to either the pipe character or the command between pipes. So, as our first pipe is the term buttercupgames, with the search term implied, all the events containing the word buttercupgames will be gathered; then the second pipe’s instructions about creating a timechart showing the count by itemId will be carried out on that gathered data. We’ll cover these more advanced processes in the chapters ahead.

**Types of SPL commands**

SPL commands can be organized into groups as shown in the following table. We will now go through each of these groups:

<table>
<thead>
<tr>
<th>Purpose of Command</th>
<th>What it Does</th>
<th>Actual Commands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter</td>
<td>Reduces results to a smaller set.</td>
<td>search/ where/ dedup/ head/ tail</td>
</tr>
<tr>
<td>Sort</td>
<td>Orders the results and can also be used to limit the number of results.</td>
<td>sort</td>
</tr>
<tr>
<td>Group</td>
<td>Puts those results like members together in groups to better see patterns in the data.</td>
<td>transaction</td>
</tr>
<tr>
<td>Report</td>
<td>Takes results of a search and summarizes them for a report.</td>
<td>top/rare/ stats/ chart/ timechart</td>
</tr>
<tr>
<td>Other</td>
<td>Included in this group are those that allow you to filter out fields, modify fields, or add fields to your results.</td>
<td>fields/ replace/ eval/ rex/ lookup</td>
</tr>
</tbody>
</table>

In the following tables, we discuss each type of command, what it does, and give examples of how it is used.
Filter commands

Search, of course, is included as a filter command as it results in a smaller data set. The other filter commands take the results from a search and then further reduce them based on the commands you use:

<table>
<thead>
<tr>
<th>Command</th>
<th>What it Does</th>
</tr>
</thead>
<tbody>
<tr>
<td>search</td>
<td>This is the most important command Splunk has. It is the default command as well, so there is no need for you to type it in the search box. However, if you do another search after one or more pipes, you do need to include the word search in the command. We'll learn more about search in the section How to perform simple searches.</td>
</tr>
<tr>
<td>where</td>
<td>This command takes an expression, such as where monthly_sales &gt; avg_mon_sales, and evaluates it. If it is TRUE, it is kept in the search results.</td>
</tr>
<tr>
<td>dedup</td>
<td>This command only keeps the first x results for each search. dedup source returns only the first result for each source. Building on this, dedup 3 source returns only the first three results for each source.</td>
</tr>
<tr>
<td>head/tail</td>
<td>These commands look for a specified number of searched terms, counting from the top or bottom of the list of events. The head command returns the first x results. head 10 returns the first ten results. The tail command returns the last x results. tail 10 returns the last ten results.</td>
</tr>
</tbody>
</table>
# The sort command

This group contains just the sort command. Here are some examples of sorts and what they do:

<table>
<thead>
<tr>
<th>Command</th>
<th>What it Does</th>
</tr>
</thead>
<tbody>
<tr>
<td>sort 0 anyfield</td>
<td>This command sorts in ascending order by user ID (A to Z, 1 to infinity, depending on whether the anyfield field is a number or name). The 0 means that all results are sorted, not just the default 10,000.</td>
</tr>
<tr>
<td>sort 1000 fieldone -fieldtwo</td>
<td>Sorts by fieldone in ascending order, then by fieldtwo in descending order, and returns up to 1,000 results.</td>
</tr>
<tr>
<td>sort -fieldone, +fieldtwo</td>
<td>Sorts by fieldone in descending order, and fieldtwo in ascending order. This command will return 10,000 results (the default).</td>
</tr>
</tbody>
</table>

# The grouping command

There is only one grouping command that we describe in the following table, although there are others. It is an important one, as it allows you to group selected events together (note that grouping can also be done through the stats command):

<table>
<thead>
<tr>
<th>Command</th>
<th>What it Does</th>
</tr>
</thead>
<tbody>
<tr>
<td>transaction</td>
<td>A transaction takes selected events and groups them together.</td>
</tr>
<tr>
<td>transaction ipaddress host maxspan=60s</td>
<td>groups together all events with the same combination of ipaddress and host, where the first and last event are no more than 60 seconds apart.</td>
</tr>
</tbody>
</table>
Reporting commands

The reporting commands listed here are the most important ones. They are vital to performing analyses in Splunk and also aid in creating dashboards and reports:

<table>
<thead>
<tr>
<th>Command</th>
<th>What it Does</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>top/rare</strong></td>
<td>The <em>top</em> command returns the values that occur most often, as well as their counts and percentages. The default is 10. <em>top source</em> returns a list of the top 10 sources, including their counts and percentages. <em>top 15 source, host</em> returns a list of the 15 most frequent source-host combinations.</td>
</tr>
<tr>
<td><strong>stats</strong></td>
<td>The <em>stats</em> command returns the results of statistical calculations. It can return a single number, as in <em>stats dc(source)</em>, which gives a distinct count that includes each different source. Or it can return a table, as in <em>stats max(kbps) by host</em>, which gives the maximum speed for each host.</td>
</tr>
<tr>
<td><strong>chart</strong></td>
<td>The <em>chart</em> command is used for creating tables of data. In each chart, the x-axis labels are indicated by either <em>over</em> or <em>by</em>. <em>chart count(fail</em>) over host* creates a chart showing the count of events that include the phrase &quot;fail&quot; plus anything after that (for example, &quot;failed&quot;, &quot;failure&quot;, and the like) for each value of <em>host</em>. For more on the <em>chart</em> command, go to <a href="http://docs.splunk.com/Documentation/Splunk/6.1.3/SearchReference/chart">http://docs.splunk.com/Documentation/Splunk/6.1.3/SearchReference/chart</a>.</td>
</tr>
<tr>
<td><strong>timechart</strong></td>
<td>The <em>timechart</em> command produces a chart with time as the x-axis. <em>timechart span=1d avg(delay) by host</em> creates a chart showing the average delay by each host during a 1 day period.</td>
</tr>
</tbody>
</table>
Other commands

These commands are also commonly used for analysis in Splunk. Several of those listed help subset and modify fields for targeted analyses. The lookup command links a field to a lookup table, from which results can be identified and output:

<table>
<thead>
<tr>
<th>Command</th>
<th>What it Does</th>
</tr>
</thead>
<tbody>
<tr>
<td>fields</td>
<td>The <code>fields</code> command is used to remove fields from a search. Thus, the command <code>fields field1 field3</code> keeps only the fields labeled <code>field1</code> and <code>field3</code>.</td>
</tr>
<tr>
<td>replace</td>
<td>The <code>replace</code> command substitutes one value for another. In the statement <code>replace 0 with Check, 9 with Warning in Status</code>, status values of 0 are replaced with <code>Check</code> and status values of 9 are replaced with <code>Warning</code>.</td>
</tr>
<tr>
<td>eval</td>
<td>The <code>eval</code> command makes calculations and puts them into a new field. This code, <code>eval Depth=case(depth&lt;=3, &quot;Low&quot;, depth&gt;3 AND depth&lt;=10, &quot;Medium&quot;, depth&gt;10, &quot;High&quot;)</code>, creates a new field, <code>Depth</code>, and uses the <code>case</code> function to assign the labels <code>Low</code>, <code>Medium</code>, or <code>High</code>, depending on the value.</td>
</tr>
<tr>
<td>lookup</td>
<td>The <code>lookup</code> command calls up a lookup table that lets you add new field values. In the statement, <code>lookup status_desc status OUTPUT description</code>, the field, <code>status</code>, is looked up in the <code>status_desc</code> lookup table and the corresponding description is output.</td>
</tr>
</tbody>
</table>

How to perform simple searches

Now we'll go on to do a couple of simple searches. In Chapter 1, Introducing Splunk, we brought in data from a file. This data included information on events that Splunk created for a fictional online store that sells games. It includes logs from the web server as well as MySQL, a backend database system. We'll do a simple search of these logs here, and will do more advanced searches in the chapters ahead. The steps and screenshots for this process are presented as follows:

1. First, you need to start up Splunk. Go to your start menu and activate Splunk. Notice that on the right, you should have a substantial number of events listed. These are the events that were indexed when you read in the file in Chapter 1, Introducing Splunk.
2. You are interested in looking at all the events involving Buttercup Games, one of the games you have at this fictional online store. Type **buttercupgames** into the **Search** box shown in the following screenshot:

![Search box screenshot](image)

Enter buttercupgames into the search box

3. You will see something like the events listed in the following screenshot. The actual events will be different, as Splunk updates the fictional data on this site. But the events you see will have a similar structure. Incidences of the search term are highlighted in yellow. Events are listed in descending order by time, with the highest (newest) timestamp first:
Notice that when you search, the search term will be highlighted in yellow in each event and the events are listed in descending order by time, or with the highest (newest) timestamp first.

4. Now add to the search term itself. Next to buttercupgames, type date_wday="wednesday". Your results will look similar to the screenshot for just buttercupgames, but you’ll notice that each of the events shows date_wday=wednesday.

Use quotes when searching for a specific value in a specific field:
When we used the search term buttercupgames, we did not specify the field in which we were looking, so everywhere that buttercupgames occurred was picked up. When we look for date_wday="wednesday", we are looking for a specific value in a specific field, so we need to specify the field we are looking for as well as the value. It is a good idea to put the search term in quotes, but this is only required if the text you are searching for contains whitespaces or special characters.
The next search will show the difference between using the implied AND and specifying OR in a search. This is important to understand as you continue to learn about searching in Splunk:

1. Suppose that you want to try to track down all instances of failed passwords that were coming into the system.

2. Click on the Splunk icon in the top left-hand corner of the screen to go back to the home page.

3. If you type in the word *fail*, you might be surprised when you get no results. The reason for this is that if you just type in *fail*, it looks only for that, and if it does not find those specific letters, followed by a space, it will not return anything. So, it will miss *failed* or any other version of *fail* that you might think it would pick up.

4. Now type in *fail* and search and you will get a different result. This time, you'll see thousands of events that show *failed*. Since you are interested specifically in failed passwords, you decide to search on the term *failed password*. Note the number of events in the upper left-hand corner.

There is an implied AND when you do a search in Splunk. To get results for two different terms, be sure to use OR.

5. Imagine that you want to look at the events where there was a failed password for users *myuan* and *harrison*. If you put in *failed password myuan harrison*, you will get no results because of the implied AND (you cannot have a user who is both myuan and harrison at the same time). But if you put an OR between myuan and harrison, that is, *failed password myuan OR harrison*, Splunk returns all results for failed passwords for either user.

If, for some reason, you get no results for either of these users, it probably means that the fictional users had no events. In this case, just do a simple search on the term *user* and select two other users from the events you see and go through Step 5 given previously.
Summary

In this chapter, we have covered the way data is collected by Splunk, indexed, and prepared for searching. We've learned about the different commands that make up the Search Processing Language (SPL) and the way commands can be piped together. Lastly, we've learned how to do some simple searches, which prepares us to do more advanced analysis in the chapters ahead.

In the next chapter, you will go on to sort, filter, change, and create fields to do more advanced analysis in Splunk.
Where to buy this book

You can buy Splunk Essentials from the Packt Publishing website. Alternatively, you can buy the book from Amazon, BN.com, Computer Manuals and most internet book retailers. Click here for ordering and shipping details.