UNDERSTANDING THE DIGITAL AUDIENCE

How Splunk Software is Used to Find the Needle and See the Whole Haystack

Use Cases

• Marketing Analytics & Reporting
• Improving Operational Efficiencies
Executive Summary

In the early days of Splunk® Enterprise, its major appeal was delivering centralized searches for log and event data. The initial benefits were powerful because manual processes and tedious procedures were eliminated. With access to such data from a single location, users were empowered to find the proverbial “needle in the haystack.” Virtually all of Splunk’s early customers were network, security or application developers desperate for a tool that enabled “Google for their logs” to find error messages, failed logins or the thrown exception.

As Splunk software evolved beyond a search engine for machine data into a platform for operational intelligence, it has begun to attract the attention of people interested in looking at the “whole haystack.” Data scientists, business analysts, digital marketers and others who analyze large amounts of data have begun to incorporate Splunk software into their organizations’ data management architecture.

One Splunk customer—a business analyst for a national media company—discovered the solution when searching for a way to report on the company’s “whole haystack” of digital audio and video distribution. Once Splunk software was in the company’s environment, the analyst found that it enabled her to address many critical challenges faced by other analysts:

• **Track usage of new platforms.** Ten years ago, a digital presence simply meant a website; now it means a website, a mobile site, mobile applications, partnerships with other sites and a presence on social media platforms. Traditional tools, which are designed on legacy technology frameworks to track and analyze each of these different platforms, haven’t kept pace. Because Splunk software can flexibly collect and index any type of machine data, it is perfectly suited to track new platforms as soon as they come online.

• **Present the big picture.** Traditional tools are designed with databases that require data normalization to conform to a certain schema. This means that they can report on some types of audience activity (such as page views to a website), but not easily on others (such as tweets). Splunk software imposes no schema and doesn’t rely on a database, so it can index, correlate and report on any type of user activity.

• **Answer unexpected questions.** Traditional solutions offer a collection of pre-defined or out-of-the-box reports, possibly including a limited search and correlation capability. For many users, a select few of these out-of-the-box reports are useful, but many are irrelevant. The few out-of-the-box reports that are useful address very specific questions (e.g., “How did that story do?”). Unfortunately, they are limited in supporting iterative analytics or answering subsequent questions (e.g., “Did it fare differently on different platforms in different cities?”). Thanks to its flexibility, Splunk software can easily adapt and apply correlation and statistical analysis to the data, enabling analysts to answer unforeseen questions.

## Business Benefits at a Glance

<table>
<thead>
<tr>
<th>Challenges</th>
<th>How Value Is Measured</th>
<th>Business Impact</th>
</tr>
</thead>
</table>
| Required timely visibility into all operational data | • Better business intelligence  
• More informed decisions  
• Increased customer engagement | Grew its audience by 17% within a single year             |
| Wanted to aggregate metrics from various sources with a single platform | Reduce the use of multiple solutions to track digital audio and video traffic | Saving approximately $100,000 each year in log tracking service costs |
| Needed deeper insight into system performance | Optimize API performance | • Sped API performance by 50%  
• Reduced infrastructure upgrades  
• Enhanced user experience |
**Many Lenses, Little Insight**

As a business analyst, the customer was exposed to solutions from many vendors. There was no shortage of business intelligence platforms, dashboards and tools designed to meet one or two of the company’s needs. The customer regularly got emails and phone calls from firms selling products for tracking the company’s main site, its mobile applications, its overall mobile site traffic, brand sentiment as expressed in social media sites, and the audio and video it distributed on its site. These vendors offered historical analysis, real-time dashboards and everything in between.

**Enter Splunk**

The company had invested in a traditional web analytics platform that relied on a JavaScript-based, client-side tracking methodology. This was sufficient for tracking traffic to the website, where the company could include the requisite JavaScript into its web pages. However, a critical component of the company’s digital strategy included distributing audio and video through channels it couldn’t “tag” such as iTunes.

To report on that kind of traffic, the analyst needed insight into the access logs where each download request was recorded. After much research, the analyst concluded that Splunk Enterprise was the best option to parse and analyze that data quickly.

However, once Splunk software was acquired to solve the problem of reporting on audio and video traffic, the company found that the solution helped to solve other issues by empowering its stakeholders with the data they need to make critical decisions.

**A single analytics tool couldn’t paint the whole picture.** The company had a website, three different mobile applications, a mobile site, and audio and video content distributed through third-party apps including iTunes. It used three different types of tracking tools to record and report on traffic to each of these various channels.

Thus, when executives asked for a dashboard that provided a holistic view, analysts would have to extract reports from each tool, normalize the data in Excel and then craft a dashboard that summarized the data into high-level numbers that stakeholders could understand at a glance.

This process was tedious, time-consuming and prone to errors. As such, reports couldn’t be created more frequently than once a month. In the news business, content strategy discussions are not monthly—they happen at least once a day. Because the handcrafted reports lacked up-to-the-moment insights, they ultimately served as historical records.

**The tracking tools couldn’t keep pace with technology.** Virtually all digital intelligence tools rely on client-side tracking to gather user information. This approach allows the tool to record a user’s every point and click—very useful for user interface designers.

However, the predominant requirement for web traffic analysis is to understand how a user engages with the content. For this type of analysis, server-side tracking is just as useful as client-side tracking, if not more so.

Client-side tracking has some disadvantages: it can only track traffic to platforms the customer can “tag,” which means that third-party apps, social media sites, and direct audio and video downloads are excluded. Also, for those platforms that can be tagged, it’s a time-consuming process to code everything properly and start recording the right data in the most useful way. If a platform is launched—a new app for an Android tablet, for example—and the tagging wasn’t done properly, the information is lost.

**Company stakeholders couldn’t ask complex questions.** One of the selling points frequently heard from vendors was the large number of pre-designed reports their tool would provide. Indeed, most analytics tools came with an excess of 100 reports. Unfortunately, most of these pre-designed reports weren’t useful to the product team.

For example, they could generate a report that showed traffic broken down by operating system (Macintosh, iOS, Windows XP, Android, etc.), but at one point the product team needed to break iOS traffic down by version—something traditional tools just couldn’t do.
In other words, traditional tools could give stakeholders enough basic information to approximate the size of their audience and how they engage with the content, but these systems were limited in their ability to help the team to thoroughly understand the complexities of the audience, which impeded innovation.

The Splunk Solution: Answering Key Questions

Once the Splunk platform was indexing data, the customer was able to generate the basic reports to meet stakeholders’ needs. All stakeholders—from the content producers to the media sales department to the executive suite—wanted answers to several very key questions:

“How did my programs do last week?” Before having direct access to the raw logs, answering even this basic question was challenging. Since programs were distributed on a variety of platforms—including websites accessed on desktops, mobile sites and mobile apps—stakeholders usually had to piece together reports from different tracking tools by cutting and pasting numbers into Excel. This meant that reports were laborious, delayed and not generated very frequently.

Once they had designed some basic dashboards in Splunk Enterprise, stakeholders were able to automate the delivery of weekly reports for anyone who requested one. For example, reports on audio and video downloads were based on standard access logs (see Figure 1).

They associated each program ID with the program name (in this case “Products”) using a lookup table (see Figure 2), so that the final report would be easily readable by anyone.

Lookup Tables

Lookup tables allow the user to expand on raw data by associating a value within the event with other information. To accomplish this, the user creates a comma-separated file and delineates the column headers on the first line of the file. Each subsequent line’s first value matches the value in the raw data.

ProgramID,ProgramName,ProgramCategory
pum,Products,Long Form
des,Design,Short Form
hum,Humor,Short Form

Once a lookup table is configured through Splunk’s web interface, anytime the raw value is found in the data, the other values on the corresponding line of the lookup table are added as if they, too, were part of that event.

Figure 1. A standard access log highlighting the status field and program ID.

Audio File Download Event

Figure 2. Lookup table to associate program ID with program name.

First, the query filtered out events that didn’t represent completed downloads:

> sourcetype="access_combined" status<300

Then, they added a timechart command to group downloads into increments of one day, broken down by program name:

> sourcetype="access_combined" status<300 | timechart span=1day count ProgramName

They used the visualization tool in Splunk Enterprise to create a stacked column graph that displayed both overall traffic and each program’s relative contribution to the total (see Figure 3).

What programs are trending on what platforms? All media organizations were investing heavily in multi-platform distribution, but few were able to make truly informed decisions about how each platform could best be used.
The stakeholders for the national media company could only suspect that certain types of content were more popular on certain platforms—long-form music programs fared better on the iPad, for example, while short news segments were more popular on desktop browsers. They were basing their decisions mostly on gut instinct—until they had a platform for machine data.

To track which audio files were accessed on which platforms, they used the section of each event that detailed the user agent, in this case, "CompanyApp/2.3.3 (iPhone; U; CPU OS 6_0 like Mac OS X; en_us)" (see Figure 4).

![Figure 3. Visualization in Splunk showing total traffic and each program's contribution to the overall traffic.](image)

Figure 3. Visualization in Splunk showing total traffic and each program’s contribution to the overall traffic.

However, because the user agent is defined by the software running on the user’s device, it tends to be very detailed and variable. In this sample line, for example, the user agent includes the software title and version, the device, operating system, and language. This represented a challenge for the customer, because a simple report showing events by user agent would have generated thousands of results.

To solve this problem, stakeholders created event types (see Figure 5) to group similar user agents into broader categories. For example, to learn when a user

### Event Types

Event types allow the user to group individual events into categories using standard Splunk search language. For example, the user could create an event type called “Bad Request” and define the search string as:

```
status>=300 OR error* OR fail*
```

Then, a search for

```
eventtype="Bad Request"
```

would match any event that fit the described search string. Even more useful is events can be grouped by event types. Thus,

```
* | stats count by eventtype
```

would group all events by their associated event types. For example, if a user wanted to understand what percent of his events were good requests and what percent were bad, he would create an event type describing “Good Request” and one describing “Bad Request” and then apply the “top” function, as in:

```
* | top eventtype
```

The result would look like:

![Figure 5. Event types group similar users agents into broader categories.](image)
downloaded a file from the company's mobile application using an iPhone, they created an event type called "Mobile App—iPhone" that filtered for any event where the user agent contained the text "CompanyApp" and "iPhone." The end result was a tidy, readable chart that enabled stakeholders to understand how users were engaging on different device/application combinations (see Figure 6).

**Figure 6.** Chart that shows how users engage on different platforms.

**What percentage of users have upgraded?**
Although stakeholders found it useful to group user agent data into larger categories for general reports, there were times when being able to filter events to answer specific strategic questions was also very useful.

For example, at one point the company’s mobile product team was considering adding a new feature to its iPad application. This feature would only work for users who had upgraded their iOS operating system to the latest version, and adding the feature entailed a significant investment of development time. Would that investment be worth it? Could most of their users actually take advantage of the new feature?

The product team wondered if there was any way it could estimate what percentage of users had upgraded. Fortunately, although that type of analysis required filtering events by an obscure substring within the user agent field (see Figure 7), Splunk’s approach to indexing the entire raw event and extracting subsets of the data into fields only at search time enabled the customer to perform the analysis.

**Audio File Download Event**
```
69.133.25.145 - - [04/Oct/2012:07:49:14 -0500] "GET /pum/pum092112pod.mp3 HTTP/1.1" 206 68056 ""-" CompanyApp/2.3.3 (iPhone; U; CPU OS 6_0 like Mac OS X; en_us)
```

**Figure 7.** A standard access log highlighting a specific substring.

The customer used the rex command in the Splunk software to extract the portion of the user agent field relevant to the analysis:
```
* | rex field=useragent "OS (?<version>\d)" | stats count by version
```

This query generated results that indicated that more than 70% of users had, indeed, upgraded and could enjoy the new feature. By being able to perform such a specific analysis quickly, the customer enabled stakeholders to make tactical and strategic decisions informed by timely operational data.

**How many people have seen my story on Twitter?**
Social media platforms are powerful promotional tools for news organizations and have become a major driver of traffic. While the customer had a wide variety of social media tracking tools to choose from, most offered metrics that weren’t valuable—in particular, they focused on mysteriously calculated sentiment scores, while failing to give access to more basic indicators.

Using a scripted input (see Figure 8), the business analyst indexed tweets containing hashtags and keywords related to the company. When tweets included a link to a particular story on the company’s site, they were able to extract the URL to create a report that listed the day’s most tweeted stories. They were even able to calculate a reach for each story by adding the number of followers of each tweeter at the time they tweeted the story link.

**Finding the Needle While Seeing the Whole Haystack**
As with so many customers, once this news organization’s business analyst brought in Splunk software, its usage spread throughout the enterprise. While the
Splunk software can index data one of three ways: by monitoring a file or folder, by monitoring the output of a port, or by indexing the output of a script. When indexing the output of a script, Splunk can be configured to trigger the script at a frequency determined by the user.

Among other things, the scripted input is useful for indexing data “scraped” from a website or an API, such as the Twitter API.

Figure 8. Scripted inputs allow users to index valuable data.

production team was using Splunk Enterprise to better understand the digital audience, application development teams were using the software to diagnose errors in their custom-built content management system and still others were using Splunk software to measure usage of the company’s publicly available API.

The results were forthcoming and substantial. First, the company grew its overall audience. By using Splunk software to better understand how audiences engage in different platforms with different content, it was able to optimize the overall experience. In the year after implementing Splunk Enterprise, its digital listening grew 17%.

Second, by using Splunk Enterprise to track digital audio and video traffic, it saved approximately $100,000 per year in log tracking service costs.

Moreover, the company used Splunk software to accelerate API performance by 50%, reducing the need for infrastructure upgrades, improving its audience’s overall digital experience and increasing user satisfaction.

In other words, rather than simply searching for a needle in the haystack, this company uses Splunk Enterprise to see the whole haystack and, ultimately, to improve its operations and efficiencies.

With Splunk, They Captured the Big Picture

In this use case, we explored how Splunk software can enable operational intelligence for a classic business analytics challenge: “How do I extract value from the information across a variety of disparate sources and third-party applications?” This use case demonstrated:

• **Elimination of data silos.** Because Splunk software indexes all kinds of data, the customer’s reliance on multiple reports and formats from varied systems was eliminated and the customer was able provide real-time information to stakeholders.

• **Correlations drive analytics.** Because Splunk Enterprise correlates different types of data, including structured and unstructured data, the customer was able to perform critical analytics and gain business intelligence such as determining whether users had upgraded their devices.

• **Flexible analytics powered by a read-time schema.** Because Splunk software collects data in full fidelity without any filtering, the customer doesn’t lose any potential value by making its data fit in a schema. This means the customer can engage in on-the-fly analysis to determine the success of its programming.

• **Value generation across multiple use cases.** In addition to gaining insights into its digital audience, the company accelerated API performance by 50%, curtailing the need for infrastructure upgrades and providing a better user experience.

About Splunk

Splunk Inc. (NASDAQ: SPLK) is the pioneer in analyzing machine data to deliver Operational Intelligence for security, IT and the business. Splunk provides the enterprise machine data fabric that drives digital transformation. More than 12,000 customers in over 110 countries use Splunk in the cloud and on-premises. Join millions of passionate users by trying a free trial of Splunk.