Mastering Splunk

Optimize your machine-generated data effectively by developing advanced analytics with Splunk

James Miller
In this package, you will find:

- The author biography
- A preview chapter from the book, Chapter 1 “The Application of Splunk”
- A synopsis of the book’s content
- More information on Mastering Splunk

About the Author

James Miller is an IBM certified and accomplished senior project leader, application/system architect, developer, and integrator with over 35 years of extensive applications and system design and development experience. He has held various positions such as National FPM Practice Leader, Microsoft Certified Solutions Expert, technical leader, technical instructor, and best practice evangelist. His experience includes working on business intelligence, predictive analytics, web architecture and design, business process analysis, GUI design and testing, data and database modeling and systems analysis, and the design and development of client-based, server-based, web-based and mainframe-based applications, systems, and models.

His responsibilities included all the aspects of solution design and development, including business process analysis and re-engineering, requirement documentation, estimation and project planning/management, architectural evaluation and optimization, test preparation, and management of resources. His other work experience includes the development of ETL infrastructures, such as data transfer automation between mainframe systems (DB2, Lawson, Great Plains, and more) and the client/server or between SQL servers and web-based applications. It also includes the integration of enterprise applications and data sources.

In addition, James has acted as Internet Applications Development Manager, responsible for the design, development, QA, and delivery of multiple websites, including online trading applications, warehouse process control and scheduling systems, and administrative and control applications. He was also responsible for the design, development, and administration of a web-based financial reporting system for a $450 million organization, reporting directly to the CFO and his executive team.
In various other leadership roles, such as project and team leader, lead developer, and applications development director, James has managed and directed multiple resources, using a variety of technologies and platforms.

He has authored the book *IBM Cognos TM1 Developer's Certification Guide, Packt Publishing*, and a number of whitepapers on best practices, such as *Establishing a Center of Excellence*. Also, he continues to post blogs on a number of relevant topics based on personal experiences and industry best practices.

He currently holds the following technical certifications:

- IBM Certified Developer – Cognos TM1 (perfect score—100 percent in exam)
- IBM Certified Business Analyst – Cognos TM1
- IBM Cognos TM1 Master 385 Certification (perfect score—100 percent in exam)
- IBM Certified Advanced Solution Expert – Cognos TM1
- IBM Certified TM1 Administrator (perfect score—100 percent in exam)

He has technical expertise in IBM Cognos BI and TM1, SPSS, Splunk, dynaSight/arcplan, ASP, DHTML, XML, IIS, MS Visual Basic and VBA, Visual Studio, Perl, WebSuite, MS SQL Server, Oracle, Sybase SQL Server, miscellaneous OLAP tools, and more.

I would like to thank my wife and soul mate, Nanette L. Miller, who has given me her everything always.
Mastering Splunk

This book is designed to go beyond the introductory topics of Splunk, introducing more advanced concepts (with examples) from an enterprise architectural perspective. This book is practical yet introduces a thought leadership mindset, which all Splunk masters should possess.

This book walks you through all of the critical features of Splunk and makes it easy to help you understand the syntax and working examples for each feature. It also introduces key concepts for approaching Splunk's knowledge development from an enterprise perspective.

What This Book Covers

Chapter 1, The Application of Splunk, provides an explanation of what Splunk is all about and how it can fit into an organization's architectural roadmap. The evolution aspect is also discussed along with what might be considered standard or typical use cases for this technology. Finally, some more out-of-the-box uses for Splunk are given.

Chapter 2, Advanced Searching, demonstrates advanced searching topics and techniques, providing meaningful examples as we go along. It focuses on searching operators, command formats and tags, subsearching, searching with parameters, efficient searching with macros, and search results.

Chapter 3, Mastering Tables, Charts, and Fields, provides in-depth methods to leverage Splunk tables, charts, and fields. It also provides working examples.

Chapter 4, Lookups, covers Splunk lookups and workflows and discusses more on the value and designing aspect of lookups, including file and script lookups.

Chapter 5, Progressive Dashboards, explains the default Splunk dashboard and then expands into the advanced features offered by Splunk for making business-effective dashboards.

Chapter 6, Indexes and Indexing, defines the idea of indexing, explaining its functioning and its importance and goes through the basic to advanced concepts of indexing step by step.

Chapter 7, Evolving Your Apps, discusses advanced topics of Splunk applications and add-ons, such as navigation, searching, and sharing. Sources to find additional application examples are also provided.

Chapter 8, Monitoring and Alerting, explains monitoring as well as the alerting capabilities of the Splunk technology and compares Splunk with other monitoring tools.
Chapter 9, Transactional Splunk, defines and describes Splunk transactions from an enterprise perspective. This chapter covers transactions and transaction types, advanced use of transactions, configuration of types of transactions, grouping events, concurrent events in Splunk, what to avoid during transactions, and so on.

Chapter 10, Splunk – Meet the Enterprise, introduces the idea of Splunk from an enterprise perspective. Best practices on important developments, such as naming, testing, documentation, and developing a vision are covered in detail.

Appendix, Quick Start, gives examples of the many resources one can use to become a Splunk master (from certification tracks to the company's website, and support portal, and everything in between). The process to obtain a copy of the latest version of Splunk and the default installation of Splunk is also covered.
The Application of Splunk

In this chapter, we will provide an explanation of what Splunk is and how it might fit into an organization's architectural roadmap. The evolution of this technology will also be discussed along with what might be considered standard or typical use cases for the technology. Finally, some more out-of-the-box uses for Splunk will be given.

The following topics will be covered in this chapter:

• The definition of Splunk
• The evolution of Splunk
• The conventional uses of Splunk
• Splunk — outside the box

The definition of Splunk

"Splunk is an American multinational corporation headquartered in San Francisco, California, which produces software for searching, monitoring, and analyzing machine-generated big data, via a web-style interface."


The company Splunk (which is a reference to cave exploration) was started in 2003 by Michael Baum, Rob Das, and Erik Swan, and was founded to pursue a disruptive new vision of making machine-generated data easily accessible, usable, and valuable to everyone.

Machine data (one of the fastest growing segments of big data) is defined as any information that is automatically created without human intervention. This data can be from a wide range of sources, including websites, servers, applications, networks, mobile devices, and so on, and can span multiple environments and can even be Cloud-based.
The Application of Splunk

Splunk (the product) runs from both a standard command line as well as from an interface that is totally web-based (which means that no thick client application needs to be installed to access and use the tool) and performs large-scale, high-speed indexing on both historical and real-time data.

Splunk does not require a restore of any of the original data but stores a compressed copy of the original data (along with its indexing information), allowing you to delete or otherwise move (or remove) the original data. Splunk then utilizes this searchable repository from which it efficiently creates graphs, reports, alerts, dashboards, and detailed visualizations.

Splunk's main product is Splunk Enterprise, or simply Splunk, which was developed using C/C++ and Python for maximum performance and which utilizes its own Search Processing Language (SPL) for maximum functionality and efficiency.

The Splunk documentation describes SPL as follows:

"SPL is the search processing language designed by Splunk® for use with Splunk software. SPL encompasses all the search commands and their functions, arguments, and clauses. Its syntax was originally based upon the UNIX pipeline and SQL. The scope of SPL includes data searching, filtering, modification, manipulation, insertion, and deletion."

Keeping it simple

You can literally install Splunk—on a developer laptop or enterprise server and (almost) everything in between—in minutes using standard installers. It doesn't require any external packages and drops cleanly into its own directory (usually into c:\Program Files\Splunk). Once it is installed, you can check out the readme—splunk.txt—file (found in that folder) to verify the version number of the build you just installed and where to find the latest online documentation.

Note that at the time of writing this book, simply going to the website http://docs.splunk.com will provide you with more than enough documentation to get you started with any of the Splunk products, and all of the information is available to be read online or to be downloaded in the PDF format in order to print or read offline. In addition, it is a good idea to bookmark Splunk's Splexicon for further reference. Splexicon is a cool online portal of technical terms that are specific to Splunk, and all the definitions include links to related information from the Splunk documentation.
After installation, Splunk is ready to be used. There are no additional integration steps required for Splunk to handle data from particular products. To date, Splunk simply works on almost any kind of data or data source that you might have access to, but should you actually require some assistance, there is a Splunk professional services team that can answer your questions or even deliver specific integration services. This team has reported to have helped customers integrate with technologies such as Tivoli, Netcool, HP OpenView, BMC PATROL, and Nagios.

Single machine deployments of Splunk (where a single instance or the Splunk server handles everything, including data input, indexing, searching, reporting, and so on) are generally used for testing and evaluations. Even when Splunk is to serve a single group or department, it is far more common to distribute functionalities across multiple Splunk servers.

For example, you might have one or more Splunk instance(s) to read input/data, one or more for indexing, and others for searching and reporting. There are many more methodologies for determining the uses and number of Splunk instances implemented such as the following:

- Applicable purpose
- Type of data
- Specific activity focus
- Work team or group to serve
- Group a set of knowledge objects (note that the definition of knowledge objects can vary greatly and is the subject of multiple discussions throughout this book)
- Security
- Environmental uses (testing, developing, and production)

In an enterprise environment, Splunk doesn't have to be (and wouldn't be) deployed directly on a production server. For information's sake, if you do choose to install Splunk on a server to read local files or files from local data sources, the CPU and network footprints are typically the same as if you were tailing those same files and piping the output to Netcat (or reading from the same data sources). The Splunk server's memory footprint for just tailing files and forwarding them over the network can be less than 30 MB of the resident memory (to be complete; you should know that there are some installations based on expected usage, perhaps, which will require more resources).

In medium- to large-scale Splunk implementations, it is common to find multiple instances (or servers) of Splunk, perhaps grouped and categorized by a specific purpose or need (as mentioned earlier).
These different deployment configurations of Splunk can completely alter the look, feel, and behavior of that Splunk installation. These deployments or groups of configurations might be referred to as Splunk apps; however, one might have the opinion that Splunk apps have much more ready-to-use configurations than deployments that you have configured based on your requirements.

**Universal file handling**

Splunk has the ability to read all kinds of data—in any format—from any device or application. Its power lies in its ability to turn this data into operational intelligence (OI), typically out of the box and without the need for any special parsers or adapters to deal with particular data formats.

Splunk uses internal algorithms to process new data and new data sources automatically and efficiently. Once Splunk is aware of a new data type, you don't have to reintroduce it again, saving time.

Since Splunk can work with both local and remote data, it is almost infinitely scalable. What this means is that the data that you are interested in can be on the same (physical or virtual) machine as the Splunk instance (meaning Splunk's local data) or on an entirely different machine, practically anywhere in the world (meaning it is remote data). Splunk can even take advantage of Cloud-based data.

Generally speaking, when you are thinking about Splunk and data, it is useful to categorize your data into one of the four types of data sources.

In general, one can categorize Splunk data (or input) sources as follows:

- **Files and/or directories:** This is the data that exists as physical files or locations where files will exist (directories or folders).
- **Network events:** This will be the data recorded as part of a machine or environment event.
- **Windows sources:** This will be the data pertaining to MS Windows' specific inputs, including event logs, registry changes, Windows Management Instrumentation, Active Directory, exchange messaging, and performance monitoring information.
- **Other sources:** This data source type covers pretty much everything else, such as mainframe logs, FIFO queues, and scripted inputs to get data from APIs and other remote data interfaces.
Confidentiality and security

Splunk uses a typical role-based security model to provide flexible and effective ways to protect all the data indexed by Splunk, by controlling the searches and results in the presentation layer.

More creative methods of implementing access control can also be employed, such as:

- Installing and configuring more than one instance of Splunk, where each is configured for only the data intended for an appropriate audience
- Separating indexes by Splunk role (privileged and public roles as a simple example)
- The use of Splunk apps such as configuring each app appropriately for a specific use, objective, or perhaps for a Splunk security role

More advanced methods of implementing access control are field encryptions, searching exclusion, and field aliasing to censored data. (You might want to research these topics independent of this book's discussions.)

The evolution of Splunk

The term big data is used to define information that is so large and complex that it becomes nearly impossible to process using traditional means. Because of the volume and/or unstructured nature of this data, making it useful or turning it into what the industry calls OI is very difficult.

According to the information provided by the International Data Corporation (IDC), unstructured data (generated by machines) might account for more than 90 percent of the data held by organizations today.

This type of data (usually found in massive and ever-growing volumes) chronicles an activity of some sort, a behavior, or a measurement of performance. Today, organizations are missing opportunities that big data can provide them since they are focused on structured data using traditional tools for business intelligence (BI) and data warehousing.

Mainstream methods such as relational or multidimensional databases used in an effort to understand an organization's big data are challenging at best.
Approaching big data solution development in this manner requires serious experience and usually results in the delivery of overly complex solutions that seldom allow enough flexibility to ask any questions or get answers to those questions in real time, which is not the requirement and not a nice-to-have feature.

The Splunk approach

"Splunk software provides a unified way to organize and to extract actionable insights from the massive amounts of machine data generated across diverse sources."


Splunk started with information technology (IT) monitoring servers, messaging queues, websites, and more. Now, Splunk is recognized for its innate ability to solve the specific challenges (and opportunities) of effectively organizing and managing enormous amounts of (virtually any kind) machine-generated big data.

What Splunk does, and does well, is to read all sorts (almost any type, even in real time) of data into what is referred to as Splunk’s internal repository and add indexes, making it available for immediate analytical analysis and reporting. Users can then easily set up metrics and dashboards (using Splunk) that support basic business intelligence, analytics, and reporting on key performance indicators (KPIs), and use them to better understand their information and the environment.

Understanding this information requires the ability to quickly search through large amounts of data, sometimes in an unstructured or semi-unstructured way. Conventional query languages (such as SQL or MDX) do not provide the flexibility required for the effective searching of big data.

These query languages depend on schemas. A (database) schema is how the data is to be systematized or structured. This structure is based on the familiarity of the possible applications that will consume the data, the facts or type of information that will be loaded into the database, or the (identified) interests of the potential end users.
A NoSQL query approach method is used by Splunk that is reportedly based on the Unix command’s pipelining concepts and does not involve or impose any predefined schema. Splunk’s search processing language (SPL) encompasses Splunk’s search commands (and their functions, arguments, and clauses).

Search commands tell Splunk what to do with the information retrieved from its indexed data. An example of some Splunk search commands include stats, abstract, accum, crawl, delta, and diff. (Note that there are many more search commands available in Splunk, and the Splunk documentation provides working examples of each!)

“You can point Splunk at anything because it doesn’t impose a schema when you capture the data; it creates schemas on the fly as you run queries” explained Sanjay Meta, Splunk’s senior director of product marketing.

– InformationWeek 1/11/2012.

The correlation of information

A Splunk search gives the user the ability to effortlessly recognize relationships and patterns in data and data sources based on the following factors:

- Time, proximity, and distance
- Transactions (single or a series)
- Subsearches (searches that actually take the results of one search and then use them as input or to affect other searches)
- Lookups to external data and data sources
- SQL-like joins

Flexible searching and correlating are not Splunk’s only magic. Using Splunk, users can also rapidly construct reports and dashboards, and using visualizations (charts, histograms, trend lines, and so on), they can understand and leverage their data without the cost associated with the formal structuring or modeling of the data first.
The Application of Splunk

Conventional use cases
To understand where Splunk has been conventionally leveraged, you'll see that the applicable areas have generally fallen into the categories, as shown in the following screenshot. The areas where Splunk is conventionally used are:

- Investigational searching
- Monitoring and alerting
- Decision support analysis

Investigational searching
The practice of investigational searching usually refers to the processes of scrutinizing an environment, infrastructure, or large accumulation of data to look for an occurrence of specific events, errors, or incidents. In addition, this process might include locating information that indicates the potential for an event, error, or incident.

As mentioned, Splunk indexes and makes it possible to search and navigate through data and data sources from any application, server, or network device in real time. This includes logs, configurations, messages, traps and alerts, scripts, and almost any kind of metric, in almost any location.

"If a machine can generate it - Splunk can index it…"

– www.Splunk.com

Splunk's powerful searching functionality can be accessed through its Search & Reporting app. (This is also the interface that you used to create and edit reports.)

A Splunk app (or application) can be a simple search collecting events, a group of alerts categorized for efficiency (or for many other reasons), or an entire program developed using the Splunk's REST API.
The apps are either:

- Organized collections of configurations
- Sets of objects that contain programs designed to add to or supplement Splunk's basic functionalities
- Completely separate deployments of Splunk itself

The Search & Reporting app provides you with a search bar, time range picker, and a summary of the data previously read into and indexed by Splunk. In addition, there is a dashboard of information that includes quick action icons, a mode selector, event statuses, and several tabs to show various event results.

Splunk search provides you with the ability to:

- Locate the existence of almost anything (not just a short list of predetermined fields)
- Create searches that combine time and terms
- Find errors that cross multiple tiers of an infrastructure (and even access Cloud-based environments)
- Locate and track configuration changes

Users are also allowed to accelerate their searches by shifting search modes:

- They can use the fast mode to quickly locate just the search pattern
- They can use the verbose mode to locate the search pattern and also return related pertinent information to help with problem resolution
- The smart mode (more on this mode later)

A more advanced feature of Splunk is its ability to create and run automated searches through the **command-line interface (CLI)** and the even more advanced, Splunk's **REST API**.

Splunk searches initiated using these advanced features do not go through Splunk Web; therefore, they are much more efficient (more efficient because in these search types, Splunk does not calculate or generate the event timeline, which saves processing time).
Searching with pivot

In addition to the previously mentioned searching options, Splunk's pivot tool is a drag-and-drop interface that enables you to report on a specific dataset without using SPL (mentioned earlier in this chapter).

The pivot tool uses data model objects (designed and built using the data model editor (which is, discussed later in this book) to arrange and filter the data into more manageable segments, allowing more focused analysis and reporting.

The event timeline

The Splunk event timeline is a visual representation of the number of events that occur at each point in time; it is used to highlight the patterns of events or investigate the highs and lows in event activity.

Calculating the Splunk search event timeline can be very resource expensive and intensive because it needs to create links and folders in order to keep the statistics for the events referenced in the search in a dispatch directory such that this information is available when the user clicks on a bar in the timeline.

Splunk search makes it possible for an organization to efficiently identify and resolve issues faster than with most other search tools and simply obsoletes any form of manual research of this information.

Monitoring

Monitoring numerous applications and environments is a typical requirement of any organization's data or support center. The ability to monitor any infrastructure in real time is essential to identify issues, problems, and attacks before they can impact customers, services, and ultimately profitability.

With Splunk's monitoring abilities, specific patterns, trends and thresholds, and so on can be established as events for Splunk to keep an alert for, so that specific individuals don't have to.
Splunk can also trigger notifications (discussed later in this chapter) in real time so that appropriate actions can be taken to follow up on an event or even avoid it as well as avoid the downtime and the expense potentially caused by an event.

Splunk also has the power to execute actions based on certain events or conditions. These actions can include activities such as:

- Sending an e-mail
- Running a program or script
- Creating an organizational support or action ticket

For all events, all of this event information is tracked by Splunk in the form of its internal (Splunk) tickets that can be easily reported at a future date.

Typical Splunk monitoring marks might include the following:

- **Active Directory**: Splunk can watch for changes to an Active Directory environment and collect user and machine metadata.
- **MS Windows event logs and Windows printer information**: Splunk has the ability to locate problems within MS Windows systems and printers located anywhere within the infrastructure.
- **Files and directories**: With Splunk, you can literally monitor all your data sources within your infrastructure, including viewing new data when it arrives.
- **Windows performance**: Windows generates enormous amounts of data that indicates a system's health. A proper analysis of this data can make the difference between a healthy, well-functioning system and a system that suffers from poor performance or downtime. Splunk supports the monitoring of all the Windows performance counters available to the system in real time, and it includes support for both local and remote collections of performance data.
- **WMI-based data**: You can pull event logs from all the Windows servers and desktops in your environment without having to install anything on those machines.
- **Windows registry information**: A registry's health is also very important. Splunk not only tells you when changes to the registry are made but also tells you whether or not those changes were successful.
The Application of Splunk

Alerting

In addition to searching and monitoring your big data, Splunk can be configured to alert anyone within an organization as to when an event occurs or when a search result meets specific circumstances. You can have both your real-time and historical searches run automatically on a regular schedule for a variety of alerting scenarios.

You can base your Splunk alerts on a wide range of threshold and trend-based situations, for example:

- Empty or null conditions
- About to exceed conditions
- Events that might precede environmental attacks
- Server or application errors
- Utilizations

All alerts in Splunk are based on timing, meaning that you can configure an alert as:

- **Real-time alerts**: These are alerts that are triggered every time a search returns a specific result, such as when the available disk space reaches a certain level. This kind of alert will give an administrator time to react to the situation before the available space reaches its capacity.

- **Historical alerts**: These are alerts based on scheduled searches to run on a regular basis. These alerts are triggered when the number of events of a certain kind exceed a certain threshold. For example, if a particular application logs errors that exceed a predetermined average.

- **Rolling time-frame alerts**: These alerts can be configured to alert you when a specific condition occurs within a moving time frame. For example, if the number of acceptable failed login attempts exceed 3 in the last 10 minutes (the last 10 minutes based on the time for which a search runs).

Splunk also allows you to create scheduled reports that trigger alerts to perform an action each time the report runs and completes. The alert can be in the form of a message or provide someone with the actual results of the report. (These alert reports might also be set up to alert individuals regardless of whether they are actually set up to receive the actual reports!)
Reporting
Alerts create records when they are triggered (by the designated event occurrence or when the search result meets the specific circumstances). Alert trigger records can be reviewed easily in Splunk, using the Splunk alert manager (if they have been enabled to take advantage of this feature).

The Splunk alert manager can be used to filter trigger records (alert results) by application, the alert severity, and the alert type. You can also search for specific keywords within the alert output. Alert/trigger records can be set up to automatically expire, or you can use the alert manager to manually delete individual alert records as desired.

Reports can also be created when you create a search (or a pivot) that you would like to run in the future (or share with another Splunk user).

Visibility in the operational world
In the world of IT service-level agreement (SLA), a support organization's ability to visualize operational data in real time is vital. This visibility needs to be present across every component of their application's architecture.

IT environments generate overwhelming amounts of information based on:

- Configuration changes
- User activities
- User requests
- Operational events
- Incidents
- Deployments
- Streaming events

Additionally, as the world digitizes the volume, the velocity and variety of additional types of data becoming available for analysis increases.

The ability to actually gain (and maintain) visibility in this operationally vital information is referred to as gaining operational intelligence.
Operational intelligence

Operational intelligence (OI) is a category of real-time, dynamic, business analytics that can deliver key insights and actually drive (manual or automated) actions (specific operational instructions) from the information consumed.

A great majority of IT operations struggle today to access and view operational data, especially in a timely and cost-efficient manner.

Today, the industry has established an organization's ability to evaluate and visualize (the volumes of operational information) in real time as the key metric (or KPI) to evaluate an organization's operational ability to monitor, support, and sustain itself.

At all levels of business and information technology, professionals have begun to realize how IT service quality can impact their revenue and profitability; therefore, they are looking for OI solutions that can run realistic queries against this information to view their operational data and understand what is occurring or is about to occur, in real time.

Having the ability to access and understand this information, operations can:

- Automate the validation of a release or deployment
- Identify changes when an incident occurs
- Quickly identify the root cause of an incident
- Automate environment consistency checking
- Monitor user transactions
- Empower support staff to find answers (significantly reducing escalations)
- Give developers self-service to access application or server logs
- Create real-time views of data, highlighting the key application performance metrics
- Leverage user preferences and usage trends
- Identify security breaches
- Measure performance

Traditional monitoring tools are inadequate to monitor large-scale distributed custom applications, because they typically don't span all the technologies in an organization's infrastructure and cannot serve the multiple analytic needs effectively. These tools are usually more focused on a particular technology and/or a particular metric and don't provide a complete picture that integrates the data across all application components and infrastructures.
A technology-agnostic approach

Splunk can index and harness all the operational data of an organization and deliver true service-level reporting, providing a centralized view across all of the interconnected application components and the infrastructures—all without spending millions of dollars in instrumenting the infrastructure with multiple technologies and/or tools (and having to support and maintain them).

No matter how increasingly complex, modular, or distributed and dynamic systems have become, the Splunk technology continues to make it possible to understand these system topologies and to visualize how these systems change in response to changes in the environment or the isolated (related) actions of users or events.

Splunk can be used to link events or transactions (even across multiple technology tiers), put together the entire picture, track performance, visualize usage trends, support better planning for capacity, spot SLA infractions, and even track how the support team is doing, based on how they are being measured.

Splunk enables new levels of visibility with actionable insights to an organization's operational information, which helps in making better decisions.

Decision support – analysis in real time

How will an organization do its analysis? The difference between profits and loss (or even survival and extinction) might depend on an organization's ability to make good decisions.

A Decision Support System (DSS) can support an organization's key individuals (management, operations, planners, and so on) to effectively measure the predictors (which can be rapidly fluctuating and not easily specified in advance) and make the best decisions, decreasing the risk.

There are numerous advantages to successfully implemented organizational decision support systems (those that are successfully implemented). Some of them include:

- Increased productivity
- Higher efficiency
- Better communication
- Cost reduction
- Time savings
- Gaining operational intelligence (described earlier in this chapter)
- Supportive education
The Application of Splunk

- Enhancing the ability to control processes and processing
- Trend/pattern identification
- Measuring the results of services by channel, location, season, demographic, or a number of other parameters
- The reconciliation of fees
- Finding the heaviest users (or abusers)
- Many more…

Can you use Splunk as a real-time decision support system? Of course, you can! Splunk becomes your DSS by providing the following abilities for users:

- Splunk is adaptable, flexible, interactive, and easy to learn and use
- Splunk can be used to answer both structured and unstructured questions based on data
- Splunk can produce responses efficiently and quickly
- Splunk supports individuals and groups at all levels within an organization
- Splunk permits a scheduled-control of developed processes
- Splunk supports the development of Splunk configurations, apps, and so on (by all the levels of end users)
- Splunk provides access to all forms of data in a universal fashion
- Splunk is available in both standalone and web-based integrations
- Splunk possess the ability to collect real-time data with details of this data (collected in an organization's master or other data) and so much more

ETL analytics and preconceptions

Typically, your average analytical project will begin with requirements: a predetermined set of questions to be answered based on the available data. Requirements will then evolve into a data modeling effort, with the objective of producing a model developed specifically to allow users to answer defined questions, over and over again (based on different parameters, such as customer, period, or product).

Limitations (of this approach to analytics) are imposed to analytics because the use of formal data models requires structured schemas to use (access or query) the data. However, the data indexed in Splunk doesn't have these limitations because the schema is applied at the time of searching, allowing you to come up with and ask different questions while they continue to explore and get to know the data.
Another significant feature of Splunk is that it does not require data to be specifically extracted, transformed, and then (re)loaded (ETL'ed) into an accessible model for Splunk to get started. Splunk just needs to be pointed to the data for it to index the data and be ready to go.

These capabilities (along with the ability to easily create dashboards and applications based on specific objectives), empower the Splunk user (and the business) with key insights—all in real time.

**The complements of Splunk**

Today, organizations have implemented analytical BI tools and (in some cases) even enterprise data warehouses (EDW).

You might think that Splunk will have to compete with these tools, but Splunk's goal is to not replace the existing tools and work with the existing tools, essentially complimenting them by giving users the ability to integrate understandings from available machine data sources with any of their organized or structured data. This kind of integrated intelligence can be established quickly (usually in a matter of hours, not days or months).

Using the compliment (not to replace) methodology:

- Data architects can expand the scope of the data being used in their other analytical tools
- Developers can use **software development kits (SDKs)** and **application program interfaces (APIs)** to directly access Splunk data from within their applications (making it available in the existing data visualization tools)
- Business analysts can take advantage of Splunk's easy-to-use interface in order to create a wide range of searches and alerts, dashboards, and perform in-depth data analytics

Splunk can also be the *engine* behind applications by exploiting the Splunk ODBC connector to connect to and access any data already read into and indexed by Splunk, harnessing the power and capabilities of the data, perhaps through an interface more familiar to a business analyst and not requiring specific programming to access the data.
The Application of Splunk

ODBC

An analyst can leverage expertise in technologies such as MS Excel or Tableau to perform actions that might otherwise require a Splunk administrator using the Splunk ODBC driver to connect to Splunk data. The analyst can then create specific queries on the Splunk-indexed data, using the interface (for example, the query wizard in Excel), and then the Splunk ODBC driver will transform these requests into effectual Splunk searches (behind the scenes).

Splunk – outside the box

Splunk has been emerging as a definitive leader to collect, analyze, and visualize machine big data. Its universal method of organizing and extracting information from massive amounts of data, from virtually any source of data, has opened up and will continue to open up new opportunities for itself in unconventional areas.

Once data is in Splunk, the sky is the limit. The Splunk software is scalable (datacenters, Cloud infrastructures, and even commodity hardware) to do the following:

"Collect and index terabytes of data, across multi-geography, multi-datacenter and hybrid cloud infrastructures"

– Splunk.com

From a development perspective, Splunk includes a built-in software REST API as well as development kits (or SDKs) for JavaScript and JSON, with additional downloadable SDKs for Java, Python, PHP, C#, and Ruby and JavaScript. This supports the development of custom "big apps" for big data by making the power of Splunk the "engine" of a developed custom application.

The following areas might be considered as perhaps unconventional candidates to leverage Splunk technologies and applications due to their need to work with enormous amounts of unstructured or otherwise unconventional data.

Customer Relationship Management

Customer Relationship Management (CRM) is a method to manage a company’s interactions with current and future customers. It involves using technology to organize, automate, and synchronize sales, marketing, customer service, and technical support information—all ever-changing and evolving—in real time.
Emerging technologies
Emerging technologies include the technical innovations that represent progressive
developments within a field such as agriculture, biomed, electronic, energy,
manufacturing, and materials science to name a few. All these areas typically
deal with a large amount of research and/or test data.

Knowledge discovery and data mining
Knowledge discovery and data mining is the process of collecting, searching, and
analyzing a large amount of data in a database (or elsewhere) to identify patterns
or relationships in order to drive better decision making or new discoveries.

Disaster recovery
Disaster recovery (DR) refers to the process, policies, and procedures that are related
to preparing for recovery or the continuation of technology infrastructure, which
are vital to an organization after a natural or human-induced disaster. All types of
information is continually examined to help put control measures in place, which
can reduce or eliminate various threats for organizations. Different types of data
measures can be included in disaster recovery, control measures, and strategies.

Virus protection
The business of virus protection involves the ability to detect known threats and
identify new and unknown threats through the analysis of massive volumes of
activity data. In addition, it is important to strive to keep up with the ever-evolving
security threats by identifying new attacks or threat profiles before conventional
methods can.

The enhancement of structured data
As discussed earlier in this chapter, this is the concept of connecting machine
generated big data with an organization's enterprise or master data. Connecting
this data can have the effect of adding context to the information mined from
machine data, making it even more valuable. This "information in context" helps
you to establish an informational framework and can also mean the presentation of
a "latest image" (from real-time machine data) and the historic value of that image
(from historic data sources) at meaningful intervals.

There are virtually limitless opportunities for the investment of enrichment of data
by connecting it to a machine or other big data, such as data warehouses, general
ledger systems, point of sale, transactional communications, and so on.
**Project management**

Project management is another area that is always ripe for improvement by accessing project specifics across all the projects in all genres. Information generated by popular project management software systems (such as MS Project or JIRA, for example) can be accessed to predict project bottlenecks or failure points, risk areas, success factors, and profitability or to assist in resource planning as well as in sales and marketing programs.

The entire product development life cycle can be made more efficient, from monitoring code checkins and build servers to pinpointing production issues in real time and gaining a valuable awareness of application usage and user preferences.

**Firewall applications**

Software solutions that are firewall applications will be required to pour through the volumes of firewall-generated data to report on the top blocks and accesses (sources, services, and ports) and active firewall rules and to generally show traffic patterns and trends over time.

**Enterprise wireless solutions**

Enterprise wireless solutions refer to the process of monitoring all wireless activity within an organization for the maintenance and support of the wireless equipment as well as policy control, threat protection, and performance optimization.

**Hadoop technologies**

What is Hadoop anyway? The Hadoop technology is designed to be installed and run on a (sometimes) large number of machines (that is, in a cluster) that do not have to be high-end and share memory or storage.

The object is the distributed processing of large data sets across many severing Hadoop machines. This means that virtually unlimited amounts of big data can be loaded into Hadoop because it breaks up the data into segments or pieces and spreads it across the different Hadoop servers in the cluster.

There is no central entry point to the data; Hadoop keeps track of where the data resides. Because there are multiple copy stores, the data stored on a server that goes offline can be automatically replicated from a known good copy.
So, where does Splunk fit in with Hadoop? Splunk supports the searching of data stored in the Hadoop Distributed File System (HDFS) with Hunk (a Splunk app). Organizations can use this to enable Splunk to work with existing big data investments.

**Media measurement**
This is an exciting area. Media measurement can refer to the ability to measure program popularity or mouse clicks, views, and plays by device and over a period of time. An example of this is the ever-improving recommendations that are made based on individual interests—derived from automated big data analysis and relationship identification.

**Social media**
Today's social media technologies are vast and include ever-changing content. This media is beginning to be actively monitored for specific information or search criteria.

This supports the ability to extract insights, measure performance, identify opportunities and infractions, and assess competitor activities or the ability to be alerted to impending crises or conditions. The results of this effort serve market researchers, PR staff, marketing teams, social engagement and community staff, agencies, and sales teams.

Splunk can be the tool to facilitate the monitoring and organizing of this data into valuable intelligence.

**Geographical Information Systems**
Geographical Information Systems (GIS) are designed to capture, store, manipulate, analyze, manage, and present all types of geographical data intended to support analysis and decision making. A GIS application requires the ability to create real-time queries (user-created searches), analyze spatial data in maps, and present the results of all these operations in an organized manner.
Mobile Device Management

Mobile devices are commonplace in our world today. The term mobile device management typically refers to the monitoring and controlling of all wireless activities, such as the distribution of applications, data, and configuration settings for all types of mobile devices, including smartphones, tablet computers, ruggedized mobile computers, mobile printers, mobile POS devices, and so on. By controlling and protecting this big data for all mobile devices in the network, Mobile Device Management (MDM) can reduce support costs and risks to the organization and the individual consumer. The intent of using MDM is to optimize the functionality and security of a mobile communications network while minimizing cost and downtime.

Splunk in action

Today, it is reported that over 6,400 customers across the world rely on the Splunk technology in some way to support their operational intelligence initiatives. They have learned that big data can provide them with a real-time, 360-degree view of their business environments.

Summary

In this chapter, we provided you with an explanation of what Splunk is, where it was started, and what its initial focus was. We also discussed the evolution of the technology, giving the conventional use cases as well as some more advanced, forward-thinking, or out-of-the-box type opportunities to leverage the technology in the future.

In the next chapter, we will explore advanced searching topics and provide practical examples.
Where to buy this book

You can buy Mastering Splunk 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.