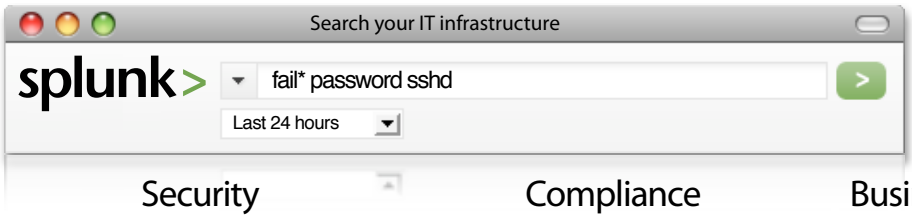
splunk>

The Guide to IT Search

IT infrastructures are too complex, virtual, service oriented and mission critical. Search is a scalable, versatile and dynamic way to manage the modern datacenter.

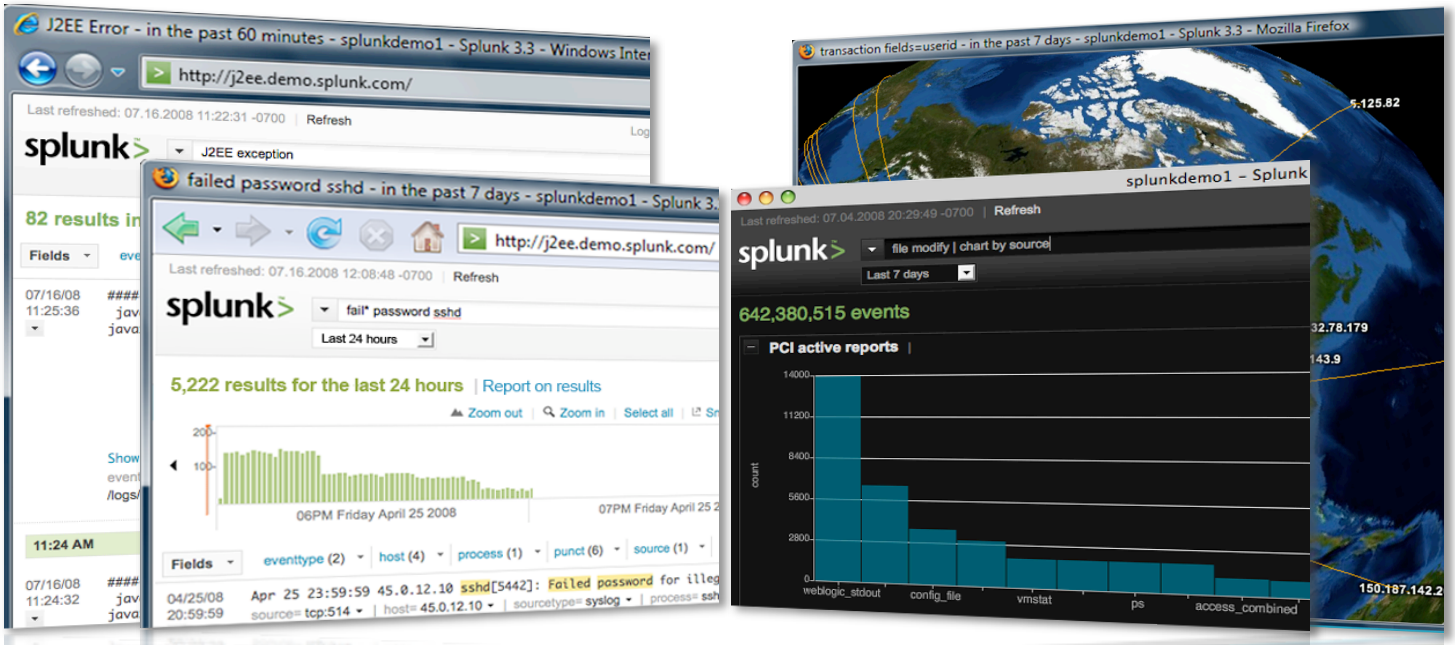


Operations
Troubleshoot Problems

Security
Investigate attacks

Compliance
Reporting and controls

Business Intelligence
Analyst transactions



What is IT Search?

IT management isn't new. We've been trying to keep complex systems under control for a long time. Search isn't new either. Every day millions of people search and navigate billions of web pages served by computers all over the world.

IT Search is new. It's the ability to instantly search all the data generated by applications, servers and network devices in your IT infrastructure. IT Search can help with:

- **Operations:** pinpoint and recover from problems quickly, keep applications, servers and networks running
- **Security:** fast, in-depth incident response to lower exposure and risk
- **Compliance:** meet mandates and requirements for IT data management without disrupting operations
- **Business Intelligence:** real-time visibility into user and business activities

Finding and fixing problems, following the trail of an attacker or tracing transactions requires a holistic view across all components.

Troubleshooting problems often means correlating web server logs, SOA messages, database transactions and configuration changes.

Investigating security incidents demands both the analysis of security events from server logs, firewalls and IDS scans, in addition to application events, configurations and scripts to understand what really happened.

On one hand, compliance mandates require systematic review and long term retention of IT data from components all over your IT infrastructure, but at the same time put up more barriers to accessing IT data for day-to-day operations.

The fact is IT infrastructures are far more mission critical, far more open to security threats and far more scrutinized for compliance than ever before. They're also far more complicated.

IT Search arms your network engineers, system administrators, security and compliance analysts, developers, customer support, help desk staff, and even business users with an up to the moment understanding of what's happening in your IT infrastructure.

How is it different?

IT Search is different from previous approaches to IT management. Here's how.

Everything in one place. It used to make sense to manage data centers in silos. But with today's distributed, scale out computing, and the proliferation of complex web-based applications and virtualization this just doesn't work anymore. IT Search breaks down these silos, indexing data from every component. Search, alert and report on all your IT data from every application, server and device — all in one place. Finding and fixing problems, following the trail of an attacker and tracing transactions is suddenly a whole lot faster and incredibly easy.

Scales to 100% of IT Data. A typical data center can generate more than a terabyte of IT data a day including:

- logs
- configurations
- traps and alerts
- messages
- scripts and code
- and performance data and statistics

Traditional IT tools leave you unable to perform comprehensive incident response, threat analysis or compliance audits. IT Search manages and gives you instant access to 100% of your IT data enabling faster problem resolution, threat response and insight into user and system activities. You don't need special agents, adapters or parsers for specific data formats and you get the correlation you need without writing lots of elaborate rules.

Keeps up with change. With constantly changing dependencies and thousands of continually evolving IT components, static assumptions about your environment are way too brittle. It's too easy for a new dependency to be missed and for a critical piece of data to fall through the cracks. IT Search employs a schema-less design and its continuous indexing keeps up with change so you don't have to. Search continuously indexes all your IT data against time so you can see changes in action. And it dynamically interprets the data when you perform a search, eliminating the need to keep up with ever-changing data formats.

Controlled access. IT Search was designed for serious security. Keep your valuable IT data protected with secure data handling, granular access controls, auditability, assurance of data integrity and integration with existing authentication systems. Finally you can control access to all your IT data and eliminate the need for system administrators, security and compliance people to touch production systems.

Plays nice with others. IT Search integrates with your existing enterprise management, security and compliance tools right out of the box. It's simple to launch searches from other tools and send alerts to any existing consoles. Index the data already collected by existing management tools to extend the life of your investments.

Fast implementation and ROI. How much time have you spent implementing IT solutions only to find out they don't do what you need? You never get time to install all the different adapters or generate all the complex rules, and once you build a schema to integrate the data sources it becomes stale the moment your environment changes. IT Search delivers immediate value by indexing and letting you search all your data, from any source, right away without investing time interpreting and integrating multiple, different data sources.

You'll do more with less. Remember when you just couldn't keep up? With IT Search you and your team will do a lot more in less time with fewer resources. You can each add your own knowledge to your IT data as you search. No need to have 20 people in a room or on a conference call. Find problems faster, investigate security incidents before attackers cover their tracks and generate those compliance reports in no time.

How do you get there?

The Guide to IT Search is your first step to choosing an IT Search solution. In it, you'll find questions to ask yourself before deciding on criteria and priorities. We do have a suggestion - consult with an expert. We've talked with hundreds of enterprises, service providers and government organizations, big and small, about IT Search. Whatever the scale or size of budget, IT Search can help.

Some basic questions

What should a good IT Search product do? Most organizations don't understand what they need to leverage large amounts of IT data for operations, security, compliance and business intelligence.

Sit down for a moment and think about what you really want. Is it a full-scale system to search, alert, report and secure all your IT data across multiple data centers? Or is it a simple collection and ad-hoc search for a single application?

If you're like most of us, you're probably aiming to start with something between these extremes. Time and money are not limitless, but compromised availability, security and the distraction of compliance typical in complex infrastructures just isn't an option anymore.

What to look for

IT Search delivers several key features:

- It **indexes** all your IT data
- It lets you **search**, **alert** and **report** on it
- It brings powerful **visualization** of your data
- It enables you to **share knowledge** about it within your organization and across the IT community
- It **scales** to any topology and volume
- It **secures** your data as a critical information asset

Read on to learn more.

Index

You'll probably begin wondering how you're going to access all your IT data in all the different formats and locations across all your data center components. IT Search offers a variety of flexible input methods and doesn't need any special adapters or parsers for specific data formats. So you can immediately index logs, configurations, traps and alerts, messages, scripts, performance data and statistics from all your applications, servers and network devices.

Flexible data input. You can monitor file systems for scripts and configuration changes, capture archive files, find and tail live application logs according to policy, connect to network ports to receive syslog, SNMP and other network-based instrumentation.

You'll also want an extensible scripted input capability to execute and capture the output of system status commands, connect to event APIs, query databases, subscribe to message queues and call remote resources. Look for prepackaged scripts to query database tables via DBI, connect to the Windows event API, capture Windows events remotely via WMI, monitor the Windows registry, connect to common secure event APIs like OPSEC LEA, access the output of common Unix / Linux system status commands like ps, top and vmstat, and remotely copy files via scp, rsync, ftp and sftp.

Index everything. Remember you'll want to index all the content in the data itself, not just a few predetermined fields or events. Dense indexing of every byte in the original data is essential if you're going to keep up with your growing and changing environment.

Keep track. With all this data coming from so many different sources, it's critical to capture the location (host and the source file, port or script) where the event originated, and index that too so you can later find data from particular locations.

Learn any source. Next the challenge turns to understanding all the different data formats. A web server access log has little in common with SNMP traps, application server stack traces or database audit tables. Working with all your sources in any format requires the flexibility to learn new formats as they appear.

IT Search can figure out data formats by sampling data and learning how to identify events, not just from simple one-line-per-event sources, but from complicated multi-line and XML sources too. It indexes the structure of the events themselves — pattern of punctuation, line breaks, etc. — so events of common formats can be found later. And it classifies sources with similar patterns with the same source type, so it's easy to find events from like sources across hosts.

No longer do you have to write complex regular expressions to parse clumsy, constantly changing formats. No need to manage different agents and adapters for each vendor's data formats.

For all time. All this streaming data means extracting, and normalizing timestamps is very important. IT Search can automatically figure out the time of any event — even with the most bizarre formats. Data missing timestamps can be handled by inferring timestamps based on context.

Real time. It's vital that all this happens fast, fast, fast. IT people depend on real-time information. Within seconds after a component generates data you'll want it available for search, alerting and reporting.

No structure. IT Search persists the original, unaltered data and a highly efficient, unstructured index into every byte of the original data. Shy away from products that persist your data in a fixed schema adding significant storage overhead, restricting your ability to search, and limiting the type of data you can index.

Search

Individual IT components can generate hundreds of events per second. A data center can log more than a terabyte a day. IT Search makes it easy to quickly search and navigate all of this data.

Fast, freeform search on anything. IT Search makes it possible to search quickly on anything in your data, not just a few pre-determined fields. It supports intuitive Boolean, nested, quoted string and wildcard searches familiar to anyone comfortable on the Web. This allows users to quickly iterate and refine their searches without knowing anything about specific data formats.

Time search. Given the large volume and repetitive nature of IT data, users often need to start by narrowing their search to a specific time range. Web and document search engines search terms or keywords within the data. They have little notion of time-based searching. With a focus on when events happen, IT Search lets users combine time and term searches. This ability to search across every tier of your infrastructure for errors and configuration changes in the seconds before a system failure occurs is incredibly powerful.

Knowledge extraction. Freeform and time-based search is just part of the picture. Leveraging dense indexing and clever data mining techniques, IT Search automatically extracts and names fields in your IT data as you search. It also lets you identify and rename fields interactively.

Fields provide the common metadata to perform structured search, alerting, reporting and analysis. This approach is much quicker to implement and more flexible than imposing a rigid set of field mapping rules ahead of time. It also supports different views into the same data, eliminating redundant storage of the same data to serve different user communities, such as security and operations.

Interactive results. Compared to command line scripts and tools, an interactive IT Search interface dramatically improves the user's experience and the speed with which tasks can be accomplished. Modern browser-based technologies like AJAX bring advanced features — type ahead, quick navigation through time, and the ability to quickly refine searches by clicking on fields or terms within the search results by using interactive histograms and filters. Best of all, no browser plug-ins or clumsy application frameworks are required.

Navigate relationships. Today you probably attempt to understand your IT system's behavior by manually looking at and trying to piece together individual events from different technologies. Usually this involves more than one person, each with specific domain expertise like networking, operating systems, databases, web servers or security.

Much like hyperlinked documents and sites on the world wide web, IT Search lets you navigate a series of related events across the infrastructure. Now a system administrator working on a web-based application can follow the sequence of activities from the web server to the application and eventually the database in order to locate the source of a failure or security breach. With a simple starting search and a few clicks you can verify a customer's problem report, establish the exact time of the error and get to the root cause quickly.

Transaction search. Sending an email, placing an order on a website or connecting a VOIP call will create a number of events across different IT components. Often you'll want to search for these collections of events that are all part of a transaction. For example, find all the sendmail events with the same userid, between a login and a logout, that occur within 10 minutes.

IT Search lets you correlate events by finding common characteristics and then saving that search as a transaction so you can find the same type of transactions again for different search parameters.

Knowledge types. As you navigate your data, you'll identify groups of events and transactions that you would like to be able to refer to by a single name. With IT Search you can discover these "like" events and transactions, name and save them to reference in other searches, alerts and reports.

Keep watch. Systems and security administrators are often firefighting a problem that's unfolding in real-time. IT Search gives you the ability to search a live stream of incoming events, much like tailing a live file, but with the power of looking across your entire infrastructure from a single place.

Alert

Undoubtedly you want to be proactive — not just search your IT data on an ad-hoc basis. IT Search provides flexible alerting capabilities that naturally improve your monitoring coverage over time. And because it works across different components and technologies, it's the most flexible monitoring tool in your arsenal.

Turn searches into alerts. Any search can be run on a schedule and trigger notifications or actions based on the search results. Notifications can be sent via email, RSS or SNMP to other management consoles. Actions trigger scripts performing user-described activities like restarting an application, server or network device.

Scheduling alerts is a great way to complete the investigation of a problem or security incident by proactively looking for similar occurrences in the future.

Get sophisticated. Don't be deceived by the simplicity of search-based alert set up. With powerful search-based correlation you can find patterns like IP addresses that hit your firewall frequently and initiate a service sending traffic back outside the firewall indicating a potential attack.

Report

Sometimes you need a birds-eye view provided by summary reports, tables and charts. The cutting edge of IT Search marries powerful reporting with the speed, flexibility and scale of IT Search.

Report on search results. Search results can be easily summarized as reports with interactive charts, graphs and tables. The simplicity of analyzing massive amounts of data will amaze you (and your boss). For example, a report can show the total bytes sent by IP address from firewall activity events; a table showing bytes per protocol per IP address; or a chart illustrating firewall traffic by hour for a specific employee's laptop. Any field can be used as reporting criteria. And remember, because fields are identified as your search you can specify new fields without re-indexing your data.

Front and center. Collect useful reports and add them to dashboards for different types of users and situations for at-a-glance reference. Reports can also be scheduled as alerts and run on a periodic basis. Send the results to team members by email or RSS feed.

Share

Everyone knows IT data is generally poorly documented by vendors, developers and operations staff. IT Search is the first technology to break through this problem.

Get smarter. With IT Search every user can add their own knowledge as they go. As you're saving searches, identifying different types of fields, events and transactions you make the whole system smarter for everyone else. And that knowledge doesn't walk out the door when someone leaves.

Plays well with others. The knowledge doesn't stop at your organization's boundary — it lets you tap into knowledge built by a larger community of users to share generally useful searches, alerts and reports or other knowledge about common IT data sources and problems.

Scale

With IT Search you can scale your installation from a single application and just a few data sources to one or more data centers and thousands of sources. You'll find a wide range of options to access data, store it, search it and route it to other systems.

Easy installation. A self-contained software server with no dependencies on third-party programs makes the right IT Search solution easy-to-install and get running. You'll also want to make sure the vendor supports a wide variety of operating system and hardware platforms. Because IT Search is software it can exist within a virtualized infrastructure rather than requiring dedicated hardware, power and rack space.

Lots of data. Your data center generates more IT data than you probably ever imagined. A single production server can generate hundreds of megabytes of data a day. Firewalls and web servers can each generate many times that amount. If you're managing a hundred components, expect 10 to 100 gigabytes a day. As you approach a thousand components they'll generate hundreds of gigabytes a day. More than a thousand components and you've got 1 TB to 10 TB every day.

This volume of data is also subject to retention requirements ranging from a few days for incident response, to months and years for compliance.

IT Search must deliver high performance to keep up. Unfortunately, you're in a particularly difficult position to judge what to buy. That's because most vendors rate performance capabilities differently. Comparison is almost impossible.

Here are some things to look for and consider:

- **Indexing throughput.** Don't look at event-per-second (EPS) ratings. Event sizes can vary from a few hundred bytes to a megabyte or more. EPS ratings are usually calculated at whatever size is optimal for one specific vendor's solution. Instead, look at throughput — megabytes per second (MBps) on a modern CPU processor. If the vendor is unable or unwilling to quote you throughput versus event rates, move on and find someone who will.
- **Search speed.** Searches of any type should return results in seconds, not minutes or hours.
- **Storage efficiency.** Measured as a percentage of the original data stream size, storage efficiency determines the amount of storage capacity you'll need to retain your data and the associated indexes. A good solution will require 25% to 50% of the original data size to retain your data and a useful set of indexes. Beware of solutions that claim 10% or less of original data size. That indicates just the storage of compressed data and no indexing.
- **Archiving.** Eventually you may decide to tier the storage of your IT data. Tiered storage can provide lower cost and better redundancy. Archiving data based on disk utilization or age will come in handy for building a multi-tiered data store. Make sure your solution lets you restore your archives on demand.
- **Linear scalability.** IT Search scales linearly by adding more computing power. You may want to run multiple servers on different machines to linearly scale your deployment or run multiple servers on a large multi-core, multi-processor machine.
- **Distributed search.** Often it won't be feasible to physically centralize all your data in one place. You will likely need to search across multiple IT Search installations and data stores in different technology or geographic silos.
- **Data routing and cloning.** With all the data streams to manage, you'll want the ability to route data based on characteristics and content. This will be important to scale and secure your IT Search installation. And as you come to depend on IT Search as a mission critical part of your IT infrastructure you'll probably want to clone important data to multiple IT Search servers for high availability.
- **Integration.** If you're like most IT shops, you've made significant investments in other management tools. Wouldn't it be great if you could integrate IT Search with those tools? Imagine launching in-context searches from your network management console, sending IT Search alerts to your system management console, or automating trouble ticket creation when unusual activity occurs. Be sure your IT Search solution provides multiple integration points and a robust API.

Secure

You'll need to keep your IT data secure. Especially as you realize what a valuable information asset you have. Look for secure data handling, access controls, auditability, assurance of data integrity and integration with existing authentication systems.

Real-time capture. The biggest threat to the integrity of IT data arises when it lingers on systems that are potentially compromised. That's why real-time data capture and forwarding to a central, hardened server is possibly the most critical requirement to ensure chain-of-evidence. Beware of legacy log appliances that can only

retrieve file-based logs or configurations via batch copies that can be tampered with.

Secure data access and transport. IT data can be sensitive. Private consumer or corporate information requires secure access, transport and storage. You should evaluate potential solutions for encrypted access to data streams using something like TCP/SSL. It's also more secure and convenient if all data streams can be multiplexed and transported over a single network port. And make sure user access is secured using something like HTTPS or SSH for command line access.

Granular access control. Of course you also need the ability to control the actions users can take and what data they can access. You don't necessarily want to allow the application development team access to your IDS scans, alerts and firewall logs. You'll likely need a flexible, role-based system that lets you build your own roles to map to your organization's policies for different classes of users. In some environments, like multi-tenant services, you may need to physically control access to data. The ability to route select data to distinct IT Search installations will let you physically separate data in different data stores.

Single sign-on. If you're using access controls internally and have organizational access control policies, you'll want to make sure you can integrate your IT Search solution with your authentication system whether it's LDAP, Active Directory, e-Directory or another authentication system.

Auditability. Once you have your access controls set-up, monitor who's doing what. Any administrative and user activities should be logged and accessible so you can audit who's accessing what data and when.

Data integrity. You'll also need to ensure the integrity of your data. How do you know the search results or report you're viewing is based on data that hasn't been tampered with? Look to make sure individual events are being signed and that streams of events are being block signed. This helps prove that nobody has inserted or deleted events from the original stream.

Hardened deployment. Keeping an audit trail and signing events is worthless if the IT Search server can be compromised. Be sure your vendor provides hardening guidelines.

The Bottom Line

The point is there's a lot to consider when comparing IT Search solutions. The vendor who gives you all the information to make an informed decision is probably the one who spent the time designing a good solution in the first place.

Get Started Today !

- Download your own free copy of Splunk today at www.splunk.com/download.
- Visit www.splunk.com/applications for more information on using Splunk IT Search to power your data center and integrate Operations, Security, Compliance and Business Intelligence.

IT Search Requirements - Outline

1 Index

a	Indexes all IT data from any application, server or network device including logs, configurations, messages, traps and alerts, metrics and performance data without any custom adapters for specific formats.
b	Flexible real-time and on-demand access to data from files, network ports and databases and custom APIs and interfaces.
	Listens to TCP and UDP network ports to receive syslog, syslog-ng and other network inputs.
	Consumes archive files.
	Captures new events in live log files in real-time.
	Monitors files for changes.
	Queries database tables via DBI.
	Monitors Windows events remotely via WMI.
	Natively accesses the Windows event API.
	Monitors the Windows registry for changes.
	Connects to OPSEC LEA and other key security event protocols.
	Subscribes to message queues like JMS.
	Captures the output of Unix / Linux system status commands like ps, top, and vmstat.
	Remotely copies files via scp, rsync, ftp and sftp.
	Extensible via scripted inputs to capture the output of new status commands, connect to new event APIs and subscribe to different kinds of message queues.
c	Universally indexes data in any format without adapters or parsers for specific data formats.
	Identifies events in single line, multi-line and complex XML structures.
	Recognizes and normalizes timestamps in any format. Handles bad or missing timestamps through contextual inference.
	Captures and indexes the structure of each event.
	Tracks and indexes the host and source of each event.
	Classifies sources dynamically.
d	Densely indexes every term in the original data.
e	Retains original, unaltered IT data.
f	Builds an unstructured index without any persistent schema.

2 Search (contin.)

a	Search across all events from all components in all formats at once.
b	Fast results for any search on any term instead of query optimization for specific fields/columns in a persistent schema.
c	Free form ad-hoc search on any term in the original events with support for Booleans, nesting, quoted strings and wildcards.
d	Precise searches using fields identified within the data at search time. Supports multiple schema views into the same data without redundant storage or re-indexing.

2 Search (contin.)

e	Type ahead suggestions make it easy to discover what to search.
f	Navigate related events and refine searches by clicking on fields or terms within the search results.
g	Search by time across all data formats.
h	Visualize trends and navigate results using time-based histograms and summaries.
i	Search for transactions across different components.
j	Persist searches as event and transaction types and search, filter and summarize by event and transaction type.
k	Discover fields, event types and transactions interactively at search time.
l	Save and use form searches to simplify routine search scenarios.
m	Search streams of incoming events via live tailing.
n	Browser based, interactive AJAX user interface. No plug-ins required.
o	Optional scriptable CLI interface for both live tail and ad-hoc search.

3 Alert

a	Run any search on a schedule and set alerting conditions based on thresholds and deltas in the number and distribution of results.
b	Trigger alerts via email, RSS, SNMP or scripts.
c	Take automated corrective or follow-on actions via scripted alerts.
d	Embed sophisticated correlation rules in alerts via subsearches.

4 Report

a	Build summary reports based on the results of any search interactively by pointing and clicking on available fields and statistics.
b	Reporting uses fields and schemas identified at search time. Supports multiple schema views into the same data without redundant storage or re-indexing.
c	Supports sophisticated statistical and summary analysis by pipelining advanced search commands together in a single search.
d	View report results in tabular form.
e	View report results as interactive line, bar, pie, scatterplot and heat map charts.
f	Pivot or drill down on any field or term.
g	Schedule any search or report for automated delivery via email or RSS.
e	Cache the results of scheduled reports for re-use.
f	Create personalized dashboards including searches, reports and alerts.

5 Share (contin.)

a	Supports many ways in which users can enrich the view onto their IT data provided by the IT Search solution.
b	Save searches, form searches, alerts, reports and dashboards and share them by role.
c	Save searches as event and transaction types. All users can search, alert and report on event and transaction types identified by other users.
d	Apply common tags across different event types, transaction types and field values. All users can search, alert and report on tags added by other users.

5 Share (contin.)

e	Define new fields interactively within search results. All users can use fields identified by other users in searches and reports.
f	Save any knowledge configuration as an add-on. Import add-ons into other installations.
g	Online knowledge base enables global knowledge sharing and look-up right from search results.
h	Online community site allows you to find add-ons created by other organizations and share your add-ons with the community.

6 Scale

a	The IT Search server is a self-contained software package. No dependencies on 3rd-party programs. Runs in virtualized server and storage environments.
b	Native packages (rpm, deb, pkg, dmg, msi, etc.) and archive format distributions (.tgz., .zip, .tar.Z) are available for most widely-deployed operating systems including Linux, Windows, Solaris, Free BSD, Mac OSX and AIX.
c	Servers work together to scale deployments. Supports both centralized and decentralized models for IT data management across the organization.
d	Provides real-time centralization of IT data from production servers with reliable data transport over TCP.
e	Policy-based data routing among IT Search servers and to 3rd party systems.
f	Linear scaling to terabytes per day via distributed search and data balancing.
g	Single view across silos via distributed search.
h	High availability via data cloning.
i	Centralized, policy-based configuration management across servers in a distributed deployment.
j	REST API enables quick integration with other IT management tools and systems.
k	Mutli-core, multi-processor support.
l	High throughput. 3.3-22 MBps (22,000 to 160,000 events per second at 150 bytes per event) on an Intel dual-core processor.
m	Tunable indexing levels can be set for different sources or events.
n	Fast search speed. Results in seconds, not minutes or hours.
o	Highly efficient compressed storage - 12-48% of the original data size typical for syslog depending on indexing level.
p	Datastore uses local or network storage and is compatible with incremental file system back-up utilities.
q	Index is segregated by time to support extended retention times without impact to search performance.
r	Configurable archiving and data retirement policy by age or size.
s	Archive and restore compressed or fully indexed data on demand. Facilitates maintaining oldest data using lower cost nearline storage for extended retention times.

7 Secure

a	Flexible roles for controlled user and API access. Supports granular data access and capabilities by role.
b	Authentication and authorization integration with Active Directory, eDirectory and other LDAP implementations.
c	Authentication API to work with web-based and passthrough single sign-on technologies such as Kerberos and Radius.
d	Real-time remote indexing of data sources to minimize the opportunity for alteration of audit trails on compromised hosts.
e	Secure data stream access and distributed functionality via SSL/TCP. Secure user access via HTTPS.

7 Secure

f	Available as a hardened virtual appliance.
g	Block-signs events to demonstrate data integrity.
h	Maintains a complete, signed audit trail of administrative actions and search history.
f	Monitors its own configurations for unauthorized change.