splunk> The search engine for IT data.

Deployment Guide

Whether you're a small IT shop with a few servers or a large data center with thousands of machines and terabytes of data, this Deployment Guide will show you how to scale Splunk for your environment including:

- Single host deployments
- Using Splunk for distributed data access
- Multiple datastore clustering
- Multiple datastore peering
- High availability deployments

Introduction

Splunk is a high performance software server that indexes and securely manages all your logs and IT data. The Splunk Server indexes IT data from ANY source. There is no need to configure it for specific formats, write regular expressions or change your logging output. It's easy to download, install and use and it's very powerful. And unlike inflexible hardware-based approaches, you can run Splunk on any size server and almost any operating system.

Splunk's versatile distributed architecture can be formulated to meet specific deployment needs. We'll discuss a few of the major deployment models here, but as you learn more about Splunk, you'll be able to determine the right model for your situation.

Before settling on a particular deployment model, you'll want to put together a plan. Start by gathering information to answer the following questions.

- What data sources do you want to access, index and manage?
- How much data volume do these sources generate in a typical 24-hour period?
- What type of server hardware will you use to deploy Splunk?
- Who do you want to have access to which data sources?
- How long will you want to retain your indexed data for search and for longer term storage?

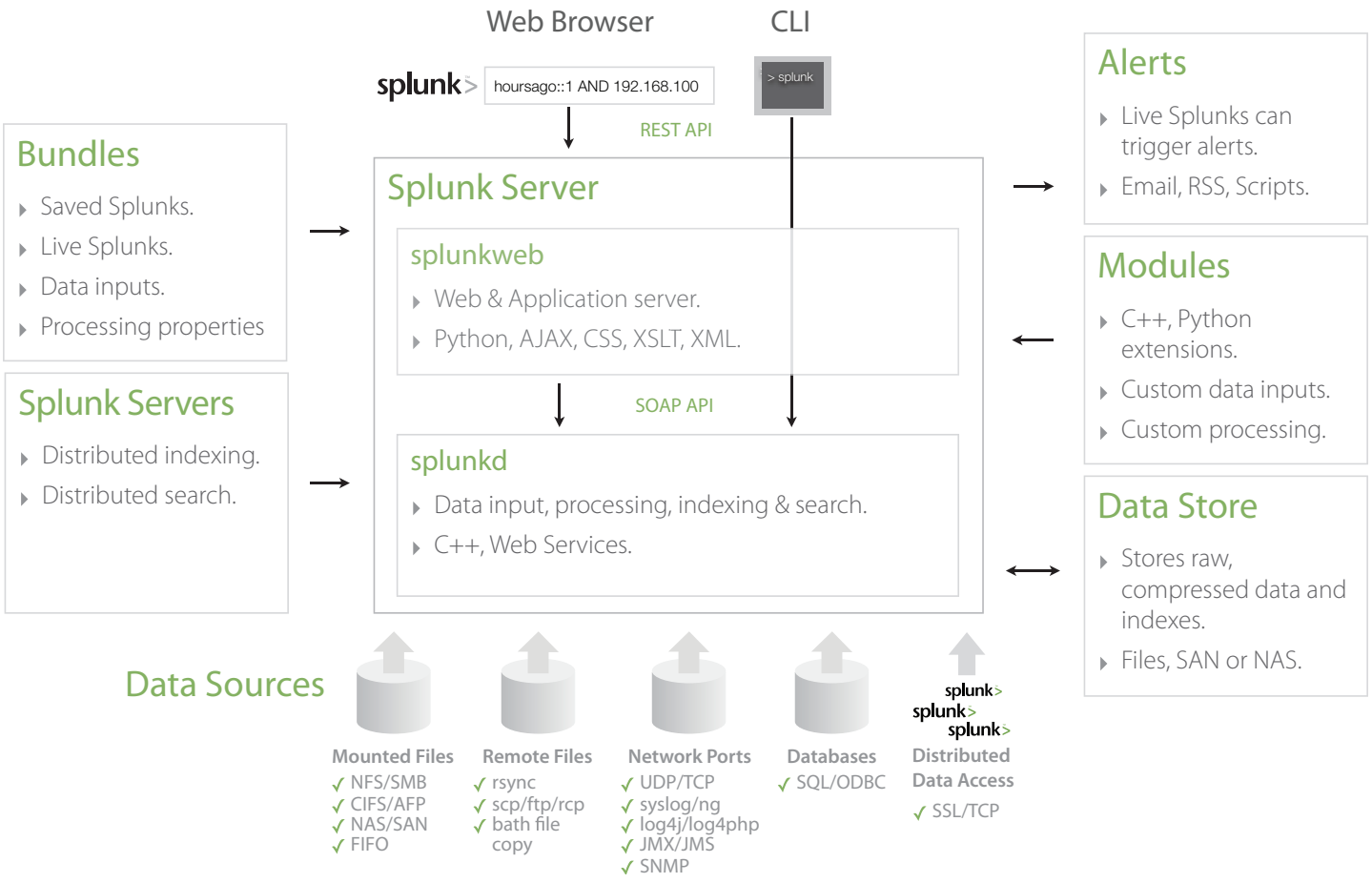
Once you have an idea of the data sources and volume, you can decide on the best way to access your data. Check out our Data Access Guide for detailed information on different ways to get to your data. The volume of your data and what type of server hardware you plan to use will impact the type of set-up you'll need to index your data quickly. Detailed performance benchmarks can be found in the Splunk Performance Guide. Finally, who you want to access which data will determine whether you'll want to use one or multiple Splunk Servers to control user access.

A Bit About Architecture

Splunk is a high performance, scalable software server written in C/C++ and Python. It indexes and searches logs and other IT data in real time. Splunk works with data generated by any application, server or device. The Splunk Developer API is accessible via REST, SOAP or the command line. After downloading, installing and starting Splunk, you'll find two Splunk Server processes running on your host, splunkd and splunkweb.

- **splunkd** is a distributed C/C++ server that accesses, processes and indexes streaming IT data and also handles search requests. splunkd processes and indexes your data by streaming it through a series of pipelines, each made up of a series of processors. Pipelines are single threads inside the splunkd process, each configured with a single snippet of XML. Processors are individual, reusable C/C++ or Python functions that act on the stream of IT data passing through a pipeline. Pipelines can pass data to one another via queues. splunkd supports a command line interface for searching and viewing results.
- **splunkweb** is a Python-based application server providing the Splunk web user interface. It allows users to search and navigate IT data stored by Splunk servers and to manage your Splunk deployment through the browser interface. splunkweb communicates with your web browser via REST and communicates with splunkd via SOAP.

Splunk Architecture



Single Host Deployment

The simplest deployment model is to run all the components of Splunk on a single server. If your data volume does not exceed the capacity of a single server this may be the right deployment for you.

A single server deployment includes the splunkWeb and splunkd processes and the Splunk datastore on a single server machine. In this model, splunkd can access, process, index and search data on a single Splunk Server. In an alternative single server configuration the Splunk datastore can reside on a NAS, SAN or other host across the network.

The single server deployment allows for multiple logical indexes to be stored within a single physical Splunk datastore. In this way, data can be kept logically separate for searching and is useful when source or source type alone doesn't provide enough data

separation. In addition, separate user accounts are available in a single server deployment. Users can have their own personalized settings including role-based access controls, Saved Splunks and notifications through Live Splunks.

If your data volume is greater than the capacity of a single server, distributed deployment models provide the ability to scale Splunk to large, multi-terabyte daily data volumes. Distributed deployments also offer greater control over data access, processing, indexing and user access controls.

Distributed Data Access

Splunk can be configured to access and forward data to another Splunk Server. The local data can be forwarded over TCP to another splunkd instance to process and index the data. Splunk is a more flexible alternative than syslog and syslog-ng.

Distributed data access provides the best control over data access for a diverse infrastructure. You can install Splunk on any source host and configure it to use any Splunk input module to access data from files, FIFO queues and network ports on that host.

Each instance of splunkd can forward its data to another instance of splunkd using Splunk-2-Splunk over TCP.

Distributing the data access function to local splunkd input configurations provides more machine resources for the splunkd index configuration by offloading the data access function. See the Installation Manual and the Administrator Manual for more information on configuring splunkd for input only.

Multiple Datastore Clustering

For large data volume environments additional data indexing capacity can be achieved with a multiple datastore clustering deployment. In this model, several Splunk Servers can share the task of indexing data from one or more sources.

Individual data sources can be handled by separate datastores or a single data source can be forked during data access or processing to different datastores.

Splunk-2-Splunk provides a single logical search environment through a distributed search capability to search across multiple datastores in real-time via SOAP/TCP.

Multiple Datastore Peering

Multiple datastore peering allows for data level access controls in large, multi-application and/or multi-data center environments. Similar to multiple datastore clustering, multiple datastore peering also offers additional data indexing capacity.

Having separate Splunk datastores makes it possible to define users and/or groups with access to specific data sources and originating hosts. Each instance of splunkd configured to index data has its own set of managed users, groups and access controls.

Peering allows those users with access to more than one Splunk datastore to search across peered datastores using the Splunk-2-Splunk interface running SOAP over TCP.

High Availability

In a single server deployment, data is sent to a single datastore via a single path. Failure of any server or storage component will mean some or all the indexed data can't be searched and may cause data to be lost. High availability deployments ensures that the routine failure of any component does not impact continued indexing and search access to any data.

Splunk achieves high availability in a multiple Splunk datastore configuration from which data that is indexed in one Splunk Server is forwarded to another Splunk Server.

Alternatively, data sources can be forked at the data access layer for lower level replication of the IT data.