Master The Dark Arts
Demystifying Splunk Architecture

J. Cory Minton | Principal Systems Engineer @ Dell EMC

Date | Washington, DC
Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Splunk, Splunk>, Listen to Your Data, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2017 Splunk Inc. All rights reserved.
J. Cory Minton
Principal SE and Data Analytics Leader

- 7+ years at Dell EMC
- Founder: Dell EMC Splunk Ninjas
- Splunk SE Certified
- I ❤️ hardware!
- Oracle and SAP background
- BS Engineering and MBA
- www.BigDataBeard.com

www.GoWithDaddy.com
Key Takeaways

- Size the infrastructure for a Splunk deployment
- Understand infrastructure impacts from small changes in Splunk
- Learn design concepts that will scale
- Hear how Dell EMC is doing it internally
- An easier way…
Problem...
Provide Fundamentals For Sizing A Splunk Deployment And Share Learned Best Practices.
Assumption #1
General understanding of Splunk platform

Rich Ecosystem of Apps

splunk>enterprise  splunk>cloud  hadoop  Free Splunk>

splunk> Platform for machine data

Forwarders  Syslog / TCP / other  Stream  DB connect  Mobile  Sensors and control systems  Mainframe data
Assumption #2
General understanding of Splunk infrastructure

Search Heads
Query information across indexers and are usually CPU and memory intensive.

Indexers
Write data to disk and are both CPU and I/O intensive.

Forwarders
Collect and forward data; usually lightweight and not resource intensive.
Assumption #3
General understanding of Splunk data management.

- **HOT** – Newest buckets of data that are still open for write
- **WARM** – Recent data but closed for writing (read only)
- **COLD** – Oldest data, commonly on cheaper, slower storage
- **FROZEN** – No longer searchable, commonly archived or deleted data
90% empirical
+ 10% experience
≠ 100% perfect every time
Big & Fast
What makes Splunk grow?

Performance
✓ Volume Of Ingest
✓ Search Performance
✓ More Users
✓ Big Apps

Capacity
✓ Volume Of Ingest
✓ Index Retention Periods
✓ Indexer Clustering
✓ Big Apps
Sizing Fundamentals

How many servers do I need?
# Machine Requirements

<table>
<thead>
<tr>
<th>Indexers</th>
<th>Reference Minimum</th>
<th>Mid-Range</th>
<th>High-Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>▶ 12 cores</td>
<td>▶ 24 cores</td>
<td>▶ 48 cores</td>
</tr>
<tr>
<td></td>
<td>▶ 12GB RAM</td>
<td>▶ 64GB RAM</td>
<td>▶ 128GB RAM</td>
</tr>
<tr>
<td></td>
<td>▶ 800 IOPS</td>
<td>▶ 800 IOPS</td>
<td>▶ SSD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Others</th>
<th>Search Head</th>
<th>Heavy Forwarder</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>▶ 16 cores</td>
<td>▶ 16 cores</td>
<td>▶ 8 cores</td>
</tr>
<tr>
<td></td>
<td>▶ 12GB RAM</td>
<td>▶ 12GB RAM</td>
<td>▶ 8GB RAM</td>
</tr>
<tr>
<td></td>
<td>▶ 300 IOPS</td>
<td>▶ 300 IOPS</td>
<td>▶ 300 IOPS</td>
</tr>
</tbody>
</table>

*Dark truth: Choose wisely…or scalability will suffer later.*
Indexer Sizing

- vCPU = CPU
- Hyperthreading ≠ CPU
- When in doubt, 100
## Search Heads

- **Dedicate**
- **When in doubt, 1 per 8**
- **Indexers > Search**

<table>
<thead>
<tr>
<th>Total Users: less than 4</th>
<th>&lt; 2GB/day</th>
<th>2 to 300 GB/day</th>
<th>300 to 600 GB/day</th>
<th>600GB to 1TB/day</th>
<th>1 to 2TB/day</th>
<th>2 to 3TB/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 combined instance</td>
<td>1 Search Head, 2 Indexers</td>
<td>1 Search Head, 3 Indexers</td>
<td>1 Search Head, 7 Indexers</td>
<td>1 Search Head, 10 Indexers</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total Users: up to 8</th>
<th>&lt; 2GB/day</th>
<th>2 to 300 GB/day</th>
<th>300 to 600 GB/day</th>
<th>600GB to 1TB/day</th>
<th>1 to 2TB/day</th>
<th>2 to 3TB/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 combined instance</td>
<td>1 Search Head, 1 Indexers</td>
<td>1 Search Head, 2 Indexers</td>
<td>1 Search Head, 3 Indexers</td>
<td>1 Search Head, 8 Indexers</td>
<td>1 Search Head, 12 Indexers</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total Users: up to 16</th>
<th>&lt; 2GB/day</th>
<th>2 to 300 GB/day</th>
<th>300 to 600 GB/day</th>
<th>600GB to 1TB/day</th>
<th>1 to 2TB/day</th>
<th>2 to 3TB/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Search Head, 1 Indexers</td>
<td>1 Search Head, 1 Indexers</td>
<td>1 Search Head, 3 Indexers</td>
<td>2 Search Heads, 4 Indexers</td>
<td>2 Search Heads, 10 Indexers</td>
<td>2 Search Heads, 15 Indexers</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total Users: up to 24</th>
<th>&lt; 2GB/day</th>
<th>2 to 300 GB/day</th>
<th>300 to 600 GB/day</th>
<th>600GB to 1TB/day</th>
<th>1 to 2TB/day</th>
<th>2 to 3TB/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Search Head, 1 Indexers</td>
<td>1 Search Head, 2 Indexers</td>
<td>2 Search Heads, 3 Indexers</td>
<td>2 Search Heads, 6 Indexers</td>
<td>2 Search Heads, 12 Indexers</td>
<td>3 Search Heads, 18 Indexers</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total Users: up to 48</th>
<th>&lt; 2GB/day</th>
<th>2 to 300 GB/day</th>
<th>300 to 600 GB/day</th>
<th>600GB to 1TB/day</th>
<th>1 to 2TB/day</th>
<th>2 to 3TB/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Search Head, 2 Indexers</td>
<td>1 Search Head, 2 Indexers</td>
<td>2 Search Heads, 4 Indexers</td>
<td>2 Search Heads, 7 Indexers</td>
<td>3 Search Heads, 14 Indexers</td>
<td>3 Search Heads, 21 Indexers</td>
<td></td>
</tr>
</tbody>
</table>
Utility Servers
Handy helpers...

- Heavy Forwarder
- License Master
- DMC
- Cluster Master
- Deployment

1:3
Sizing Fundamentals

How much storage do I need?
Assumption #3
General understanding of Splunk data management.

HOT – can be DAS in server or SAN (Flash is best)
WARM – same as Hot
COLD – adds option for NAS
FROZEN – No longer searchable, so object stores are option here (last resort)
Myth About Bucket Sizing…

- # of buckets x bucket size
- Not days…

indexes.conf

# volume definitions

[volume:hotwarm_cold]
path = /mnt/fast_disk
maxVolumeDataSizeMB = 3984589

# index definition (calculation is based on a single index)

[main]
homePath = volume:hotwarm_cold/defaultdb/db
coldPath = volume:hotwarm_cold/defaultdb/colddb
thawedPath = $SPLUNK_DB/defaultdb/thaweddb
homePath.maxDataSizeMB = 512000
coldPath.maxDataSizeMB = 2560000
maxWarmDBCount = 4294967295
frozenTimePeriodInSecs = 2592000
maxDataSize = auto_high_volume
coldToFrozenDir = /mnt/big_disk/defaultdb/frozendb
Indexer Deployment Options

Distributed Deployment
Indexer data is stored once and distributed across available indexers

Clustered Deployment
A group of indexers are configured to replicate each other’s data
Distributed Deployment

- Single copy of data
- Small
- Starter
- Storage-bound
Indexer Storage Capacity

1TB Ingested Data

Written Data

Raw Data
- Compressed Raw data
  - 30% of written data
  - → 150GB

Indexes
- Uncompressed ‘indexes’
  - 70% of written data
  - → 350GB
How Much Storage You Need?

1TB Ingested Data

= Daily indexing rate
  \times \frac{1}{2}
  \times Retention policy

= 1TB \times \frac{1}{2} \times 60 \text{ days} = 30TB

Raw Data

9TB

Indexer

Indexes

21TB
Indexer Clustering

- High Availability for Indexes
- Indexer Clustering Settings
  - Replication Factor = copies of raw data
  - Search Factor = copies of indexes
Splunk Indexer Availability

Multiple copies of index and raw data

- Index → # copies of indexes → Search factor (SF)
- Raw Data → # of copies of raw data → Replication factor (RF)

Copy 1
500GB

Copy 2
500GB

Copy 2
500GB

1TB Ingested Data
SF=2 / RF=2

500GB written → 500GB replicated

1TB * 60 days x \( \frac{1}{2} \) x
= 60TB (RF/SF=2) ** doubled **

1TB * 60 days x \( \frac{1}{2} \) x 3
= 90TB (RF/SF=3) ** tripled **
Multisite Indexer Clustering

- Protects indexes across disparate locations
- Enables Search Affinity
- Site specific RF/SF settings
- Sizing = each site + site protected
Unofficial, But Really Helpful Tool

http://splunk-sizing.appspot.com/
Splunk Sizing Questionnaire

- What is the licensed daily ingest rate for Splunk (expressed in some amount of GB/Day or TB/day)?
- What is the retention period for Hot/Warm and Cold (days kept in each tier)?
- Any data being sent to frozen? If so, what is the retention period and requirement for doing so?
- Is indexer clustering being leveraged? If so, what are the settings for Replication and Search Factor?
- How many indexer and search servers are deployed? Do you have a visualization you can share of the deployment?
- Is Splunk being run as a single site or multiple sites? If multiple, is multi-site clustering being leveraged?
- Is the Enterprise Security App or ITSI for Splunk deployed?
The right solutions to optimize your Splunk deployment
The Ready Solutions formula

Dell EMC portfolio

Priorities
- HPC
- Hybrid Cloud
- Biz Apps

Services
- Compute
- Deploy
- Assess
- Support
- Financial
- Consult

Knowledge

Value-add IP

Ready Nodes
Ready Bundles
Ready Systems

© 2017 SPLUNK INC.
Dell EMC Ready Solutions for Splunk

Ready System

VxRack + Isilon

VxRail + Isilon

Ready Bundle

PowerEdge + Isilon

“Meets or EXCEEDS minimum hardware requirements”
Logistics Leader
Doug called them out on Q1 earnings call...

- Simplified acquisition
- Leveraged Ninjas
- Deployed apps for all Dell EMC platforms
- Replatforming HW in near future
Wholesale Club Retailer

- Flashed Splunk
- Bottomless cold with Isilon... over 1PB!
- Decreased floor space by 30%
- Growing to +3TB/day
Winter is coming…
Splunk at Dell EMC
Our defense against Black Friday...

- eCommerce IT services
- Marketing effectiveness
- Security and threats
- Replatforming now
Splunk Applications From Dell EMC
Extend the power of Splunk to Dell EMC Platforms

What are Splunk Apps?
Splunk applications and add-ons allow user to import data into Splunk from specific sources.

Splunk & its partners have created a rich community called SplunkBase that has 1000s+ applications.

Why are Splunk Apps important?
Splunk apps and add-ons allow customers to incorporate new use cases and extend their Splunk environment. This leads to increased Splunk License needs as well as additional Hardware.

Dell EMC has apps for the following:
- VMAX
- XtremIO
- Isilon
- VNX
Solution centers
Staffed with engineers and Blueprint solution experts

Engagements begin with your challenges
• Briefings with a team of experts
• Architectural design sessions
• Proofs of concept
Let our Splunk Ninjas help you!

- Trained by Splunk
- Splunk Architecture Experts
- Dell EMC Portfolio Experts
- Religious about Best Practices
- Available across the GLOBE!!!

Email Splunk.Ninjas@emc.com
Don't forget to rate this session in the .conf2017 mobile app.