splunk> .conf2017

Shrinking the Elephant in the Room

Maximizing logs' business value with AWS

Chris Gordon | Software Engineer, Yelp

Zach Musgrave | Technical Lead, Yelp

Patrick Shumate | Solutions Architect, Amazon Web Services

September 27, 2017 | 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.

Introduction

DevOps, Splunk, Storage, and You

splunk> .conf2017

Our Splunk Cluster

Multi-site, multi-cluster, 4x replication

Storage

~1 petabyte

Daily Searches

Tens of thousands

Daily Ingestion

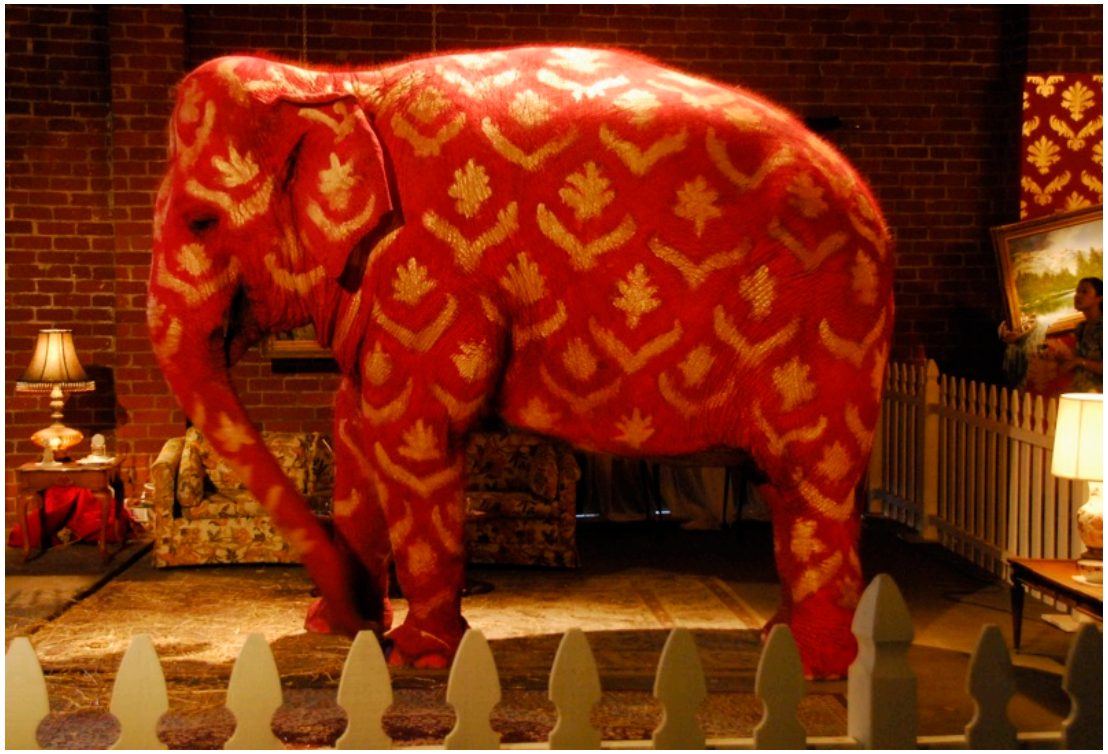
Tens of terabytes

splunk> .conf2017

The Elephant in the Room

…

- Initial budget set at $X
- One year later, costs at 150% of $X
  - Users want to ingest *all the things*
- After *Shrinking the Elephant…*
  - Storage costs **down** 15%
  - Headroom for new data **up** 40%
  - Logical retention **down** 20%
- Business effects
  - Users lost no insight
  - Business value left unchanged



https://www.flickr.com/photos/44124323641@N01/246805948

It Begins

Business Cat pulls the trigger

https://www.pinterest.com/pin/298715387758064097

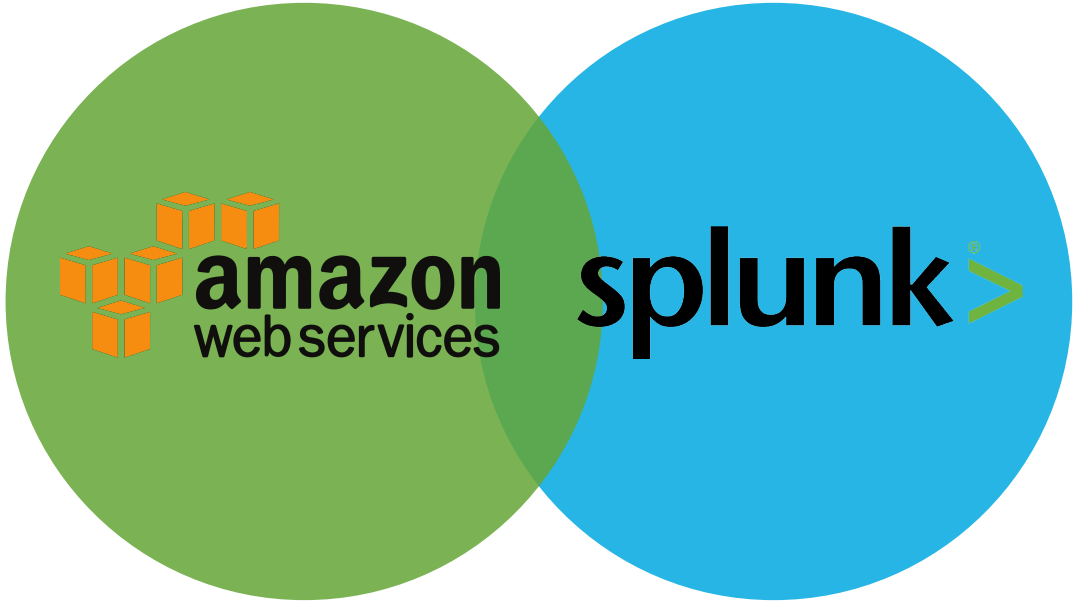
splunk> .conf2017

Vendors

Presupposition is the root of all conference talks

Infrastructure

Public cloud



License

Provisioned based on daily ingestion

AWS?

Cloud?

Amazon Web Services (AWS) is a secure cloud services platform, offering compute power, database storage, content delivery and other functionality to help businesses scale and grow.

A Broad IT Infrastructure Platform

The AWS Cloud provides a broad set of infrastructure services, such as computing power, storage options, networking and databases, delivered as a utility: on-demand, available in seconds, with pay-as-you-go pricing.

https://aws.amazon.com/what-is-aws/

Budgeting

Several sources of cost

Splunk License

Essentials

Storage

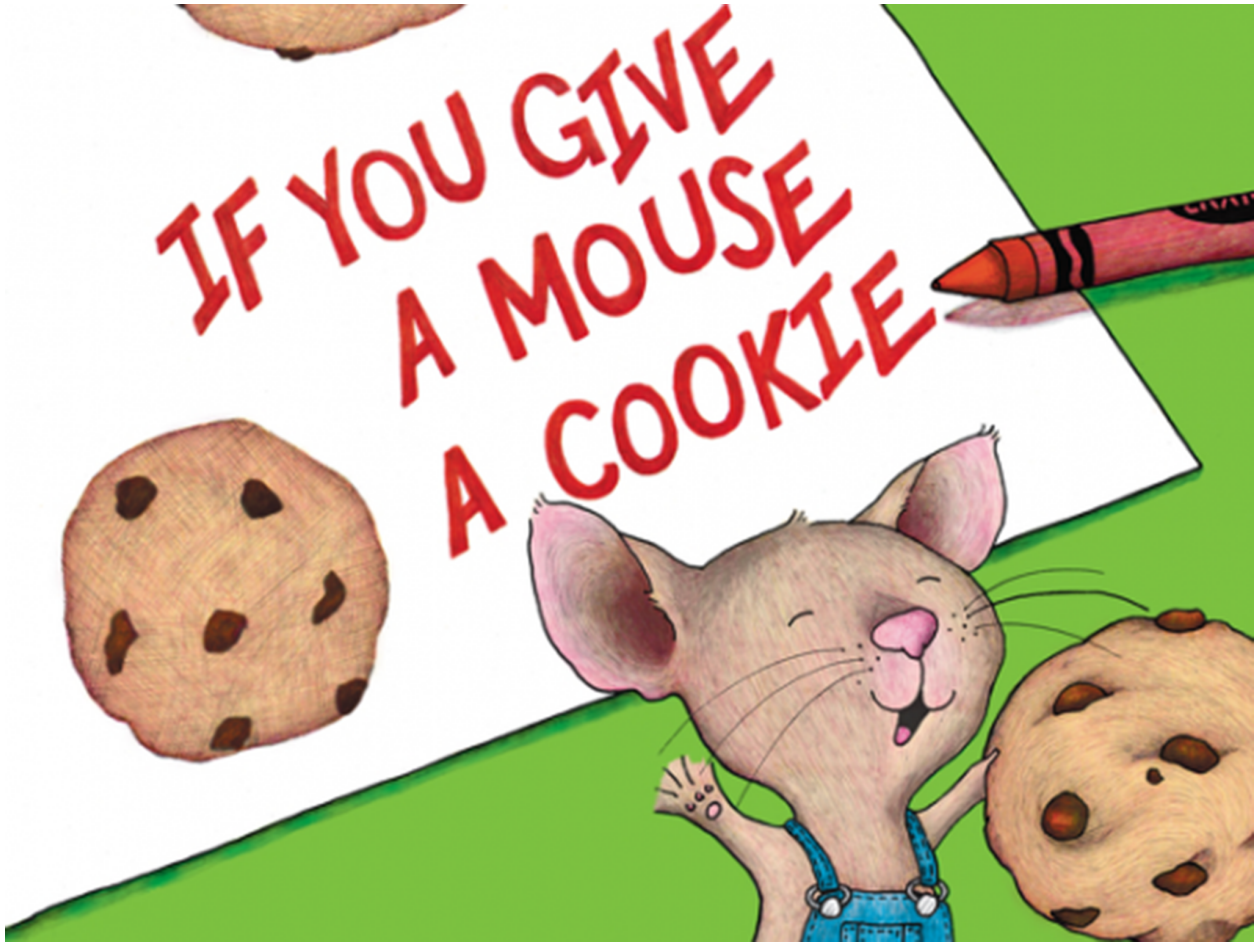
Petabytes aren't free

Indexers

Throughput to storage

Budgeting

One big source of cost



http://www.independent.co.uk/arts-entertainment/books/if-you-give-a-mouse-a-cookie-childrens-book-has-a-secret-political-message-about-helping-yourself-a6782616.html

“Can we ingest ten terabytes a day? And retain it forever?”

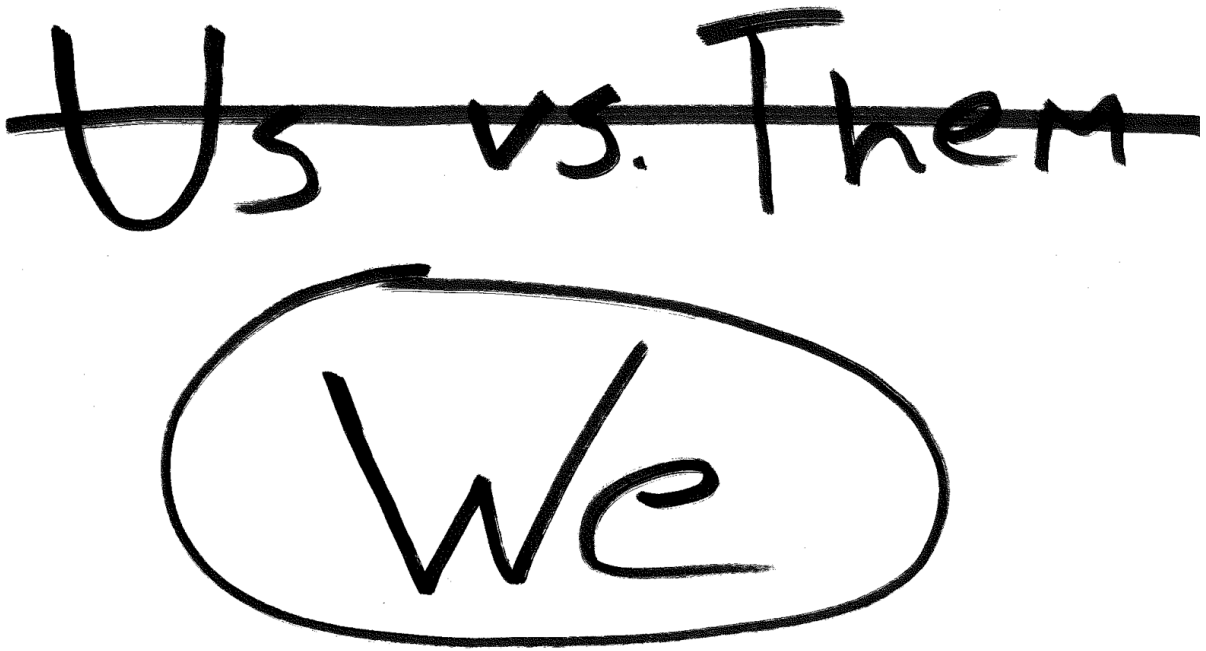
An engineer, who shall remain nameless

splunk> .conf2017

DevOps

Or, *we're all in this together*

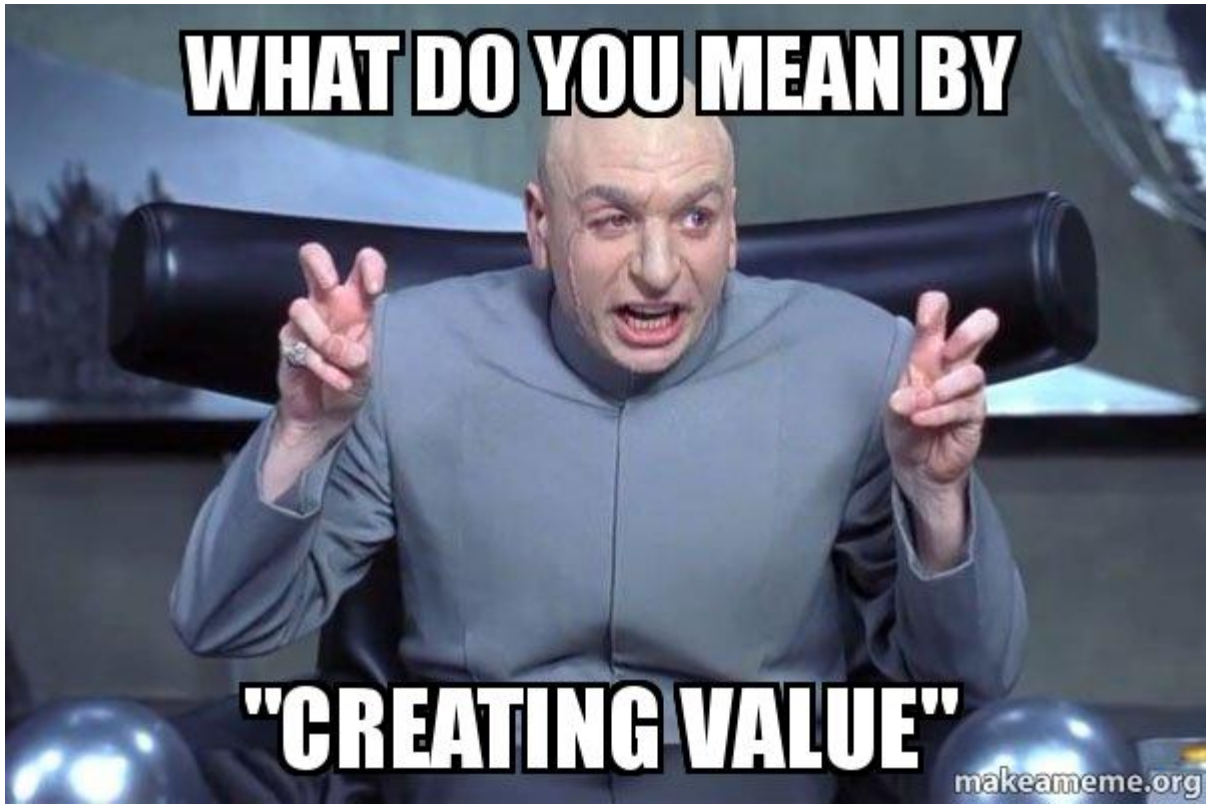
- Retain all data forever…
  - Users: Happy
  - Finance: Livid
- Retain all data for one day…
  - Users: Frustrated
  - Finance: Thrilled
- Retain data based on value…
  - Users: *Mostly* Happy
  - Finance: Placated



What's Value, Anyway?

...

- Benefit the business
  - Close more and larger deals faster
  - Retain existing accounts
  - Limit resource loss (right-sizing)
- Improve employee productivity
  - Find insights faster
  - Discover correlations across verticals
  - Automate rote tasks
- Something else?
  - That depends on you!



https://makeameme.org/meme/what-do-you-3kq4js

Value is Intrinsic

…

- Your users interact with data…
  - Therefore, it's valuable
  - Trust that they know best
- How valuable?
  - Frequency of access
  - Depth of interaction
  - Consequences if it disappeared

https://www.gulosolutions.com/2015/02/secret-making-uxd-feel-right/

How To Quantify Value

And engineer for flexibility

- Splunk isn't your data's final resting place
  - Next up: Next-generation ingestion from Amazon S3
- Splunk tells you **a lot** about access patterns
  - But only if you ask nicely: We'll show you how
- Merely shutting off retention isn't acceptable
  - Business is agile: You have to hedge your bets
- Pre-processing is your friend
  - Not just map/reduce
  - More like map/count/reduce/reduce/reduce…

http://trekcore.com/blog/

We Can All Get Along

It’s the DevOps mantra

http://if-you-give-a-mouse-a-cookie.wikia.com/wiki/File:If_you_give_a_mouse_a_cookie.jpg

splunk> .conf2017

Getting Your Data into Splunk

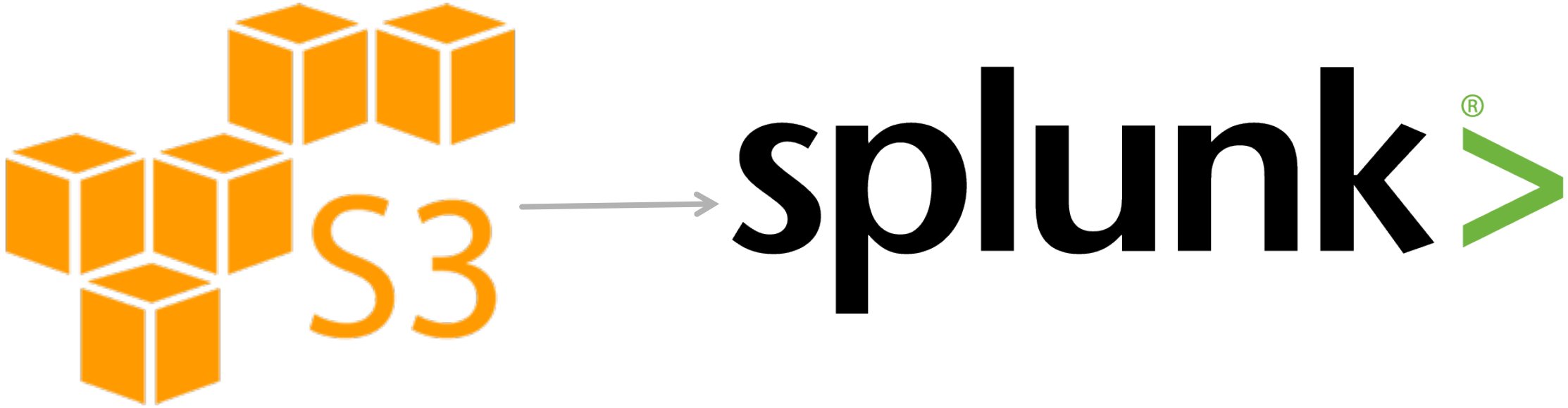
Amazon S3, SQS, and more

splunk> .conf2017

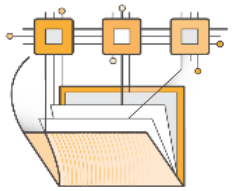
Getting Your Data into Splunk

Leveraging S3 as your data source

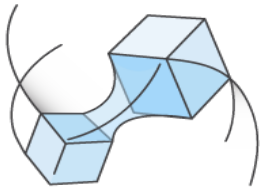
- Where do the logs come from? Amazon S3!
  - Cheap, always available storage
  - Useful API for retrieving and manipulating your data



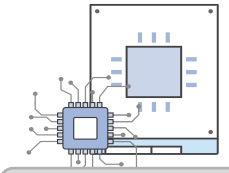
Storage Is A Platform: AWS Storage



Amazon EFS

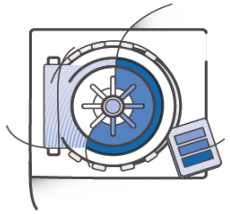


Amazon EBS



Amazon EC2 Instance Store

Amazon S3



Amazon Glacier

AWS Direct Connect

AWS Snowball

ISV Connectors

Amazon Kinesis Firehose

S3 Transfer Acceleration

Storage Gateway

splunk> .conf2017

Getting Your Data into Splunk

Transferring data via SQS and SNS

- How do the logs get into Splunk? Amazon SQS and SNS
  - SQS and S3 APIs do all the hard work
  - Scripted input: Read from SQS queues
    - A config file can inform the script about what to consume and where to get it
      - Auto-generate **`props.conf`** and **`inputs.conf`**
  - Scaling up is easy…
    - Just add more forwarders!

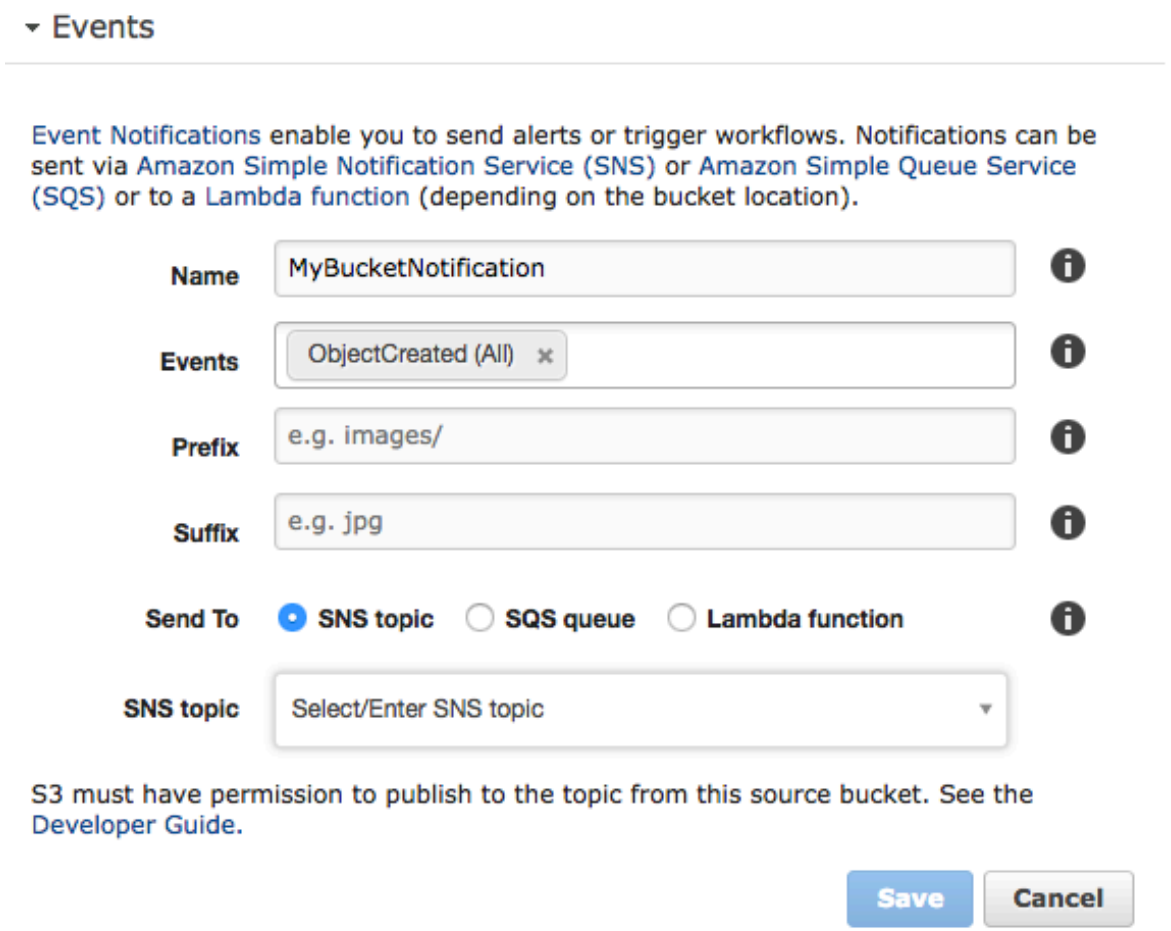
```
sqs:
  my-sqs-queue:
    aws_region: "us-west-2"

logs:
  my-sourcetype:
    sqs_name: "my-sqs-queue"
    s3_prefix:
     - "s3://yelp-logs-us-west-2/logs/mylog/"
    index: "my index"
    time_prefix: "\"timestamp\":"
    time_format: "%+"
```

Getting Your Data into Splunk

Putting data on SQS

- Set up a bucket notification
  - Simple case: Send straight to SQS queue
  - Less simple case: Send to SNS topic
    - Why?
      - Allows for multiple consumers of the notifications
    - How?
      - Send to SNS topic
      - SNS topic feeds into your SQS queue



splunk> .conf2017

Getting Your Data into Splunk

What does all this get you?

- The obvious
  - You automatically ingest whatever S3 data you want!
- The less obvious
  - You can **backfill** whatever data you want with little to no effort
    - Simple: A script generates a *bucket notification* for anything you want to ingest
    - Add a new log to Splunk? Instantly ingest any historical data
- The least obvious
  - You can **re-ingest** data that's rolled out of retention on a whim
    - Really make use of that unlimited license

http://www.backtothefuture.com/

Getting Your Data into Splunk

Retroactive Ingest

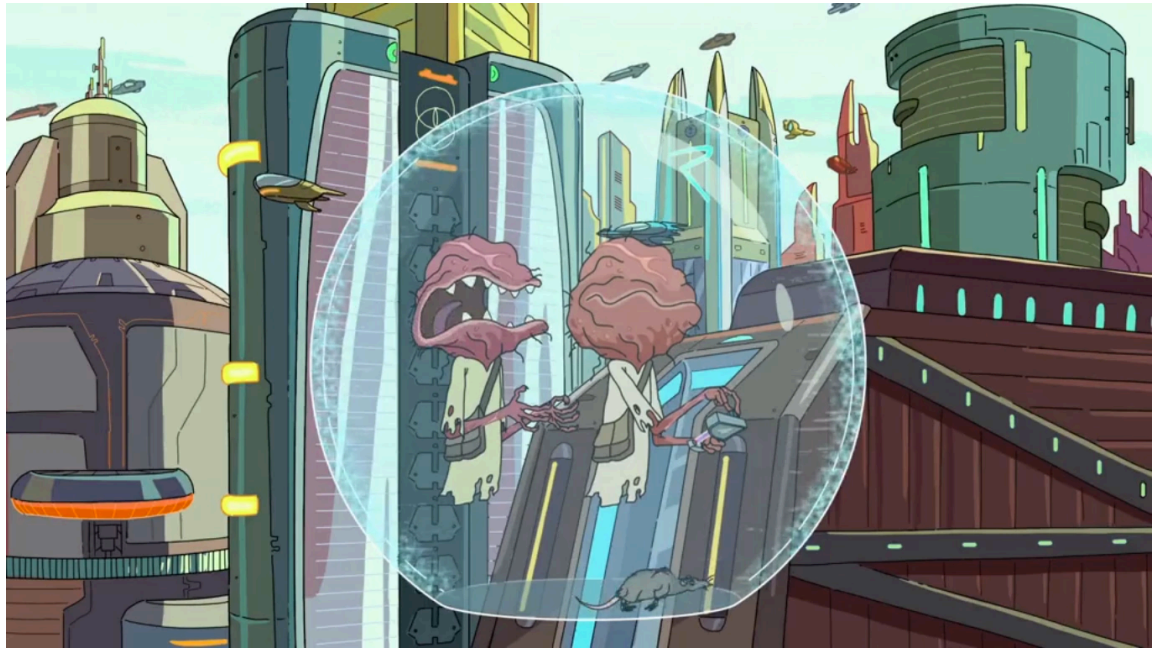
- Allocate a short retention index for ad-hoc ingestion and search
  - Ingest the data, do what you want with it, then let it roll out again soon after
    - Ingest the data, run a summary report to collect a bunch of stats to be stored long-term
      - More on this later
  - **Don't keep data around longer than you need it**

http://btulp.com/12208/cake-graphic/

Getting Your Data into Splunk

Retroactive Ingest (cont)

- Are there any downsides? *Sort of…*
  - Can't think of index retention windows the same
    - Data rolls out based on index time, not time of the event
      - You've broken time
    - What does this mean?
      - You will need to control retention by setting the appropriate size limit
      - You can set up alerts to let you know when data is starting to roll out too early



http://rickandmorty.wikia.com/wiki/File:Screenshot_2015-09-29_at_11.41.47_PM.png

Minimizing Storage Costs without Losing Value

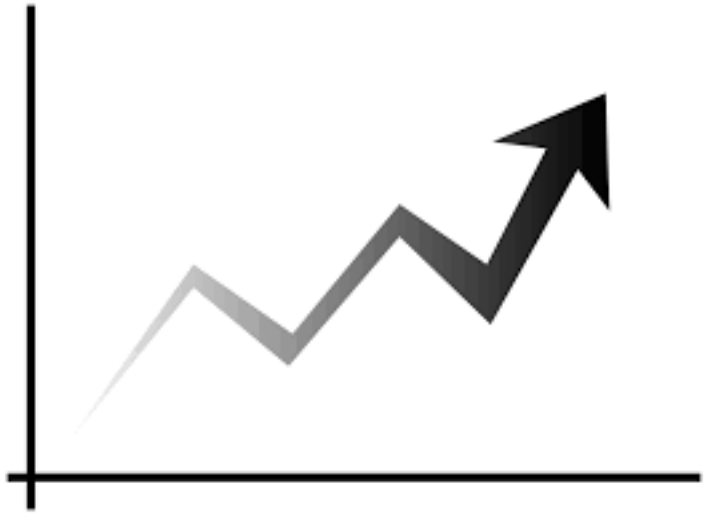
Or, Summary Indexing 101

splunk> .conf2017

Minimizing Storage Costs

What is Summary Indexing?

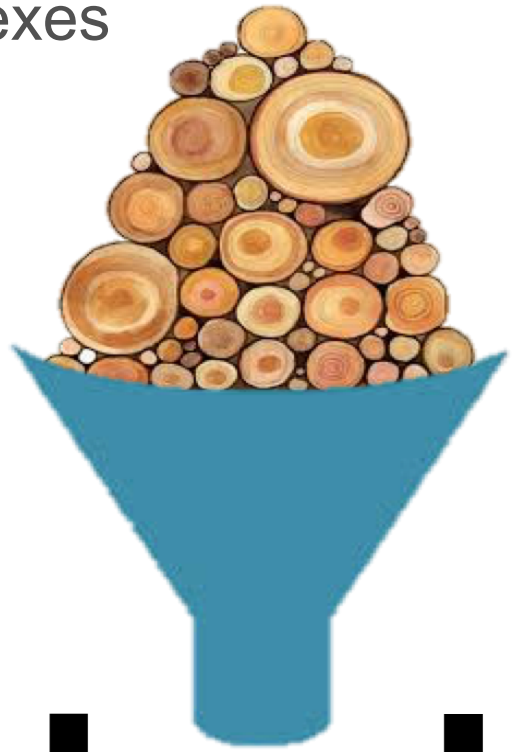
- What is summary indexing?
  - Think of it as computing statistical rollups of your logs
  - As data gets older, *you don't care about individual logs*
    - Trends become more important
  - These stats have insignificant storage cost
    - You can essentially store them forever
  - Dashboards based on summary indexes load *extremely quickly*



splunk> .conf2017

Minimizing Storage Costs

Leveraging Summary Indexes



- Take summary indexes a step further
  - Use them to enable storing super verbose logs in Splunk
  - Lots of DevOps tools produce insane amounts of logs
    - Puppet, CloudTrail, HAProxy, NGINX, etc…

- How to ingest them cost effectively?
  - You guessed it, summary indexes!

splunk>

Minimizing Storage Costs

Leveraging Summary Indexes (cont)

- These individual logs typically stop being useful after a few days
  - So, only store them for a few days!
- Perform daily (or weekly) summary reports to persist any statistics or trends
  - These can be stored in a summary index at insignificant cost

- How long should retention be?
  - We'll come back to this
- What if someone needs the logs longer?
  - No problem, just re-ingest them!

splunk> .conf2017

Measuring the Business Value of Your Logs

Analyzing the cost/value of your data

splunk> .conf2017

Measuring Cost/Value

Mapping log ingest and usage

- So we've covered how to minimize cost, but…
  - How to attribute a cost to each log?
    - **`license_usage`** logs!
  - How to determine a log's value relative to its cost?
    - **`audit`** logs!

- You still likely need to talk to stakeholders, but it's a starting point
  - Who are the stakeholders?
    - We'll come back to this

Measuring Cost/Value

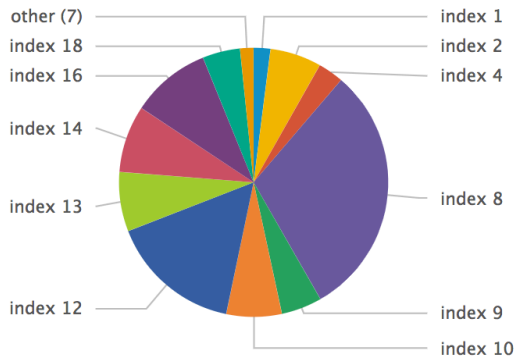
How much does each log cost?

- How to attribute a cost to each log?
  - Splunk indexer deployments: Relatively static
    - Most of the cost is here
    - For on-site deployments, costs don't change as ingestion goes up or down
      - Costs change only when you scale storage/indexer count
  - Solution: Attribute a percentage of cost to each log based on their ingestion ratios

Total Cost

$100,000

Index Costs



Note: Numbers have been changed to protect the innocent…

splunk> .conf2017

Measuring Cost/Value

Which logs are being used?

- How to determine a log's value?
  - Count how often logs are searched
  - Re-use the cost per log data
  - Result is a *Cost per Search* metric
    - Useful for finding candidates for removal
- Types of high CPS logs
  - Unsearched: Can be removed easily
  - Seldom searched: Require further investigation

stream	stSize_GB	costPerMonth	SearchCount	costPerSearch
stream 379	3,506	$ 582.30	5	$ 116.46
stream 181	988	$ 164.16	2	$ 82.08
stream 56	6,050	$ 1,004.91	19	$ 52.89
stream 170	485	$ 80.50	2	$ 40.25
stream 304	3,204	$ 532.12	21	$ 25.34
stream 129	1,593	$ 264.59	18	$ 14.70
stream 166	126	$ 20.91	2	$ 10.45
stream 362	122	$ 20.31	2	$ 10.16
stream 180	290	$ 48.25	5	$ 9.65
stream 19	112	$ 18.52	2	$ 9.26
stream 385	8,596	$ 1,427.67	156	$ 9.15
stream 377	26,172	$ 4,346.96	533	$ 8.16
stream 346	339	$ 56.33	7	$ 8.05
stream 76	2,100	$ 348.80	49	$ 7.12
stream 6	160	$ 26.56	5	$ 5.31
stream 192	334	$ 55.54	14	$ 3.97
stream 55	96	$ 16.00	5	$ 3.20
stream 358	82	$ 13.65	5	$ 2.73
stream 159	214	$ 35.48	13	$ 2.73
stream 184	180	$ 29.91	14	$ 2.14
stream 384	1,407	$ 233.71	129	$ 1.81
stream 175	180	$ 29.84	17	$ 1.76
stream 307	830	$ 137.92	94	$ 1.47
stream 147	394	$ 65.43	51	$ 1.28
stream 183	3,240	$ 538.13	427	$ 1.26

« prev 1 2 3 4 5 6 7 8 9

Note: Numbers have been changed to protect the innocent…

splunk> .conf2017

Measuring Cost/Value

Identify where to reduce costs

- Cost Per Search
  - Useful for finding high cost/low value logs
  - **But…** once identified, what do you do with them?
    - Ask stakeholders if logs can just be removed
- Which brings us back to the question:
  - Who are the stakeholders?
    - Instead of just counting searches, count by user!

Identify Users of a Stream

Stream

nginx_access.nginx_acce...

	user	SearchCount
1	user 1	464
2	user 2	187
3	user 3	126
4	user 4	94
5	user 5	78
6	user 6	48
7	user 7	46
8	user 8	43
9	user 9	39
10	user 10	24
11	user 11	17
12	user 12	17
13	user 13	16
14	user 14	10

http://img.taopic.com/uploads/allimg/120209/2195-12020Z9542514.jpg

splunk> .conf2017

Measuring Cost/Value

Identify where to reduce costs

- We've identified the users of our removal candidate
  - What if they push back?
    - This log is useful to somebody
    - Put it on **cheaper hardware** instead of removing completely
    - Lower QOS while still supporting developer needs
- Where else can you reduce costs without impacting users?
  - Drop your unused retention!
    - But how to identify unused retention? Back to the **`audit`** logs!

https://imgur.com/gallery/MeuZf

Measuring Cost/Value

Determine the useful retention window

- How far into the past is a log useful?
  - The **`audit`** log has the answer!
  - Each search has a start time
    - *Retention window = time of search – start time*
  - Count the number of searches in **tiered** windows
    - One day, three days, one week, two weeks, three weeks…

Total Searches

Total Searches Past One Day

Total Cost Minus Cost of Storing One Day

stream	TotalCount	costPerMonth	OneDayCount	CostofStoringPastOneDay	ThreeDayCount	CostofStoringPastThreeDays	OneWeekCount	CostofStoringPastOneWeek	TwoWeekCount	CostofStoringPastTwoWeeks	Three
nginx_access.nginx_access	2563	$ 1,539.66	120	$ 1,488.34	52	$ 1,385.69	46	$ 1,154.75	10	$ 769.83	

Note: Numbers have been changed to protect the innocent…

stream = index.sourcetype

Total Cost

splunk> .conf2017

Measuring Cost/Value

Drop Retention Without Losing Value

- Many logs' search counts will drop off after a few days
  - But not completely
  - How do we lower retention without also removing value?
    - Remember summary indexes?

Identify Streams with the worst Retention Utilization

Percentile of CPS to show: 95

Minimum Cost per Month: 50

Finding owners
- ○ Only show users as owners
- ● Show apps and users

Display the logs in the Nth percentile of CPS for any retention window

Note: Numbers have been changed to protect the innocent…

	stream	Users	SearchCounts	costPerMonth	TotalCount	OneDayCount	CostofStoringPastOneDay	CPS_OneDay	ThreeDayCount
1	stream 1	user_288 user_299	261 176	$ 2,693.46	878	78	$ 2,603.68	33.38	68
2	stream 2	user_82	15910	$ 1,865.24	16625	136	$ 1,803.07	13.26	46
3	stream 3	user_64	21690	$ 1,578.43	22201	20	$ 1,525.82	76.29	2
4	stream 4	user_154	5789	$ 1,292.75	5890	41	$ 1,249.66	30.48	37
5	stream 5	user_238	1238	$ 954.00	2563	120	$ 922.20	7.68	52
6	stream 6	user_332 user_355	75 46	$ 884.61	178	54	$ 855.13	15.84	40
7	stream 7	user_514 user_532	10 8	$ 622.66	44	18	$ 601.91	33.44	18

Measuring Cost/Value

How to think about savings

- Remember: Attributed costs are just a **percentage** of total infrastructure costs
  - Costs don't actually go down until you scale down the cluster
    - Use cost as a proxy for expected value
  - Instead of thinking about savings, **think about what else you can ingest**
    - Provide more value
- What if you do need to reduce costs?
  - Keep a spreadsheet to inform your decisions
    - How many indexers you need to drop to meet a goal
    - How much data you need to drop to fit into a smaller cluster

Space Required on each Indexer (GB)	Free Space Left on each Indexer (GB)
6,695.73	471.11
Number of Indexers that Should be Added	Percentage of space left as headroom on each indexer
-2	6.57%

Index	Current Total Size (all indexers)	Current Median Oldest Age on Indexers (days)	Desired Retention (days)	Desired Headroom (protection from log explosions/ indexer losses)	Estimated total size (when at full retention)	Space needed(total size + headroom)	Space required per indexer (GB)	Percentage of Whole	Cost (per month)	Space to Allocate MB (indexes.conf)
nginx_access	60,061.00	30	7	5.00%	14,014.23	14,714.95	294.30	4.10%	$2,225.67	303,128.00

Note: Numbers have been changed to protect the innocent…

splunk> .conf2017

Wrapping Up

The Elephant Has Been Shrunk

splunk> .conf2017

Final Results

The elephant has been shrunk

1. Our cluster is far more efficient
   - 40% more headroom for future growth
   - In line with company budget expectations
2. Our users retain all valuable data
   - We know what that is now
   - We can leverage it to further optimize
3. Our ingestion scales along with us
   - We add, remove, *re-add* data easily
   - Special cases are easy: *Put it in S3!*

Key Takeaways

Hope you're still awake!

1. **Don't be an Admin. Be an Owner.**
   - Take responsibility for your Splunk!
2. Make cluster decisions **with data**
   - Measure real use of your cluster
   - *If they use it, then it's important.*
3. Ingest intelligently
   - Use Amazon S3 to do this cheaply and flexibly

Thank You

Don't forget to rate this session in the .conf2017 mobile app

splunk> .conf2017