Let's chat about Splunk and ELK...

Kate Lawrence-Gupta – Platform Architect
Splunk

klawrencegupta@splunk.com

October 2018 | Version 1.0

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> .conf18

Kate

- 15+ years experience in infrastructure management, systems operations, security & big data architecture.
- Spent the past 6 years with Comcast
  - Principal Engineer (Splunk)
  - Senior Manager of Engineering & Software Development
    - Focus on open source integrations with existing data platforms
- Inaugural SplunkTrust member & 2013 Revolution Award Winner (Innovation)
- Joined Splunk ~6 months ago as Platform Architect in the Global Engineering team.

Data is Critical

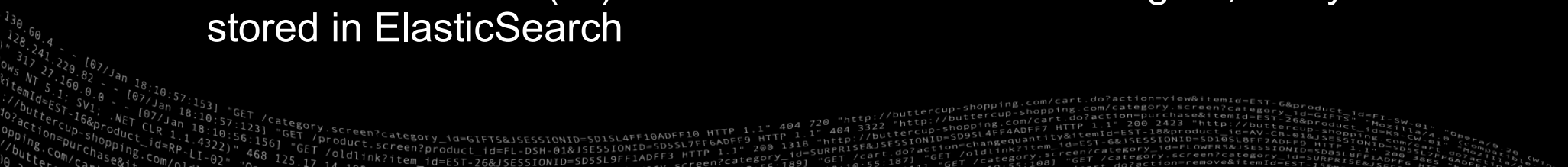
Extracting Value

Which path...

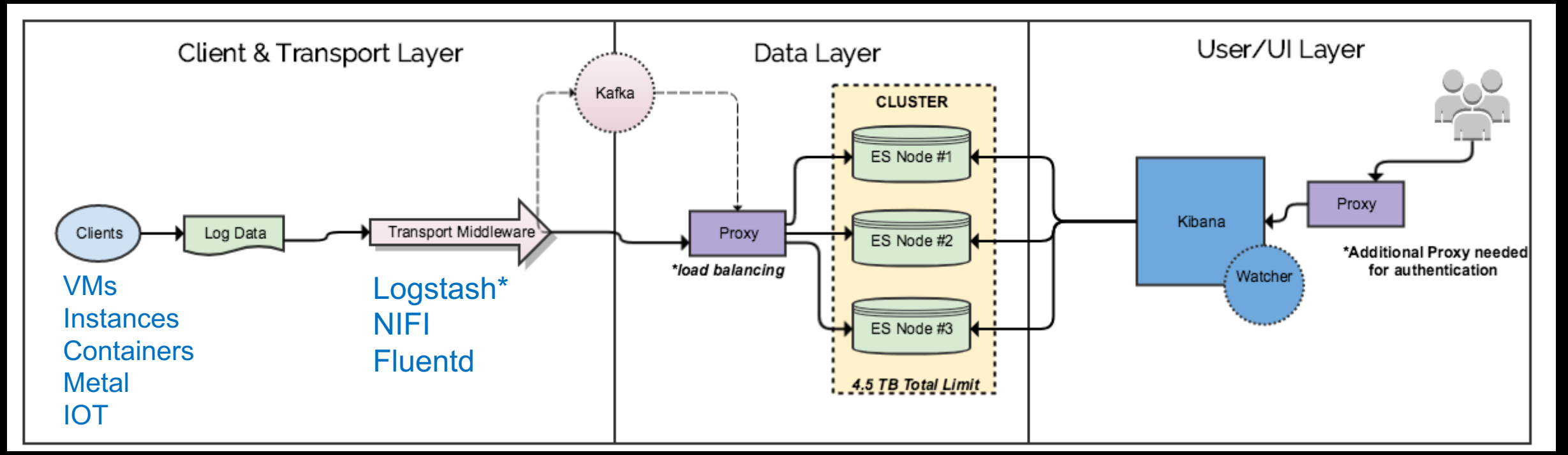
Splunk or ELK...?

What is ELK

- ELK represents a suite of open source tools that work together in a stack to provide a complete experience to the end user for managing log data.
  - ElasticSearch
    - The data layer where the log data is physically stored on disk in indices.
  - Logstash
    - The transport or middleware layer that allows the log data to be sent from clients to ElasticSearch
    - This is also where the schema or format of your data is defined that allows for analysis
      - This is commonly referred to as schema-on-write methodology
  - Kibana
    - The user interface (UI) that allows the user to investigate, analyze & visualize the data stored in ElasticSearch



ELK Ecosystem – High Level Architecture

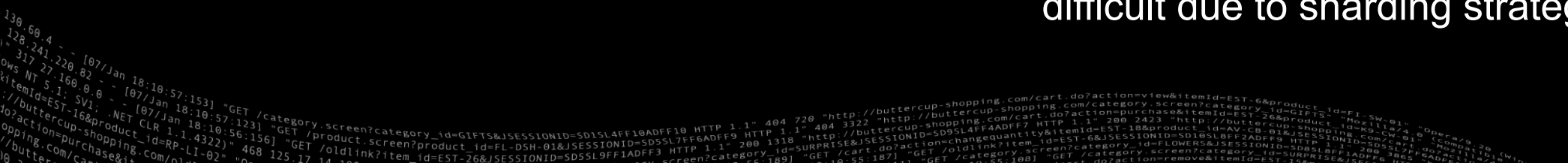


Pros

- Open Source
- Active Development Community
- Robust Query Language
- Hosted Solutions are available (AWS, Elastic Cloud, GCP)
- Additional support is available from vendors for on-premise or cloud based deployments
- Flexible integration models available
- Learning curve is relatively low

Cons

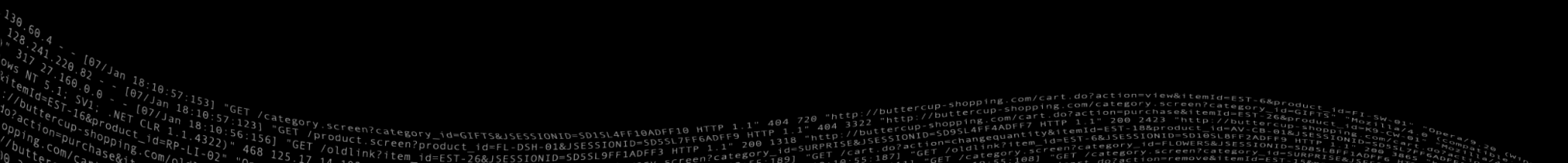
- Schema-on-write methodology can be difficult to manage and does not work well for unstructured data sets
- Scaling into TBs/Day of ingest can be challenging
- X-pack plugins are needed for SAML, alerting, graphing & internal monitoring of deployment
- Kibana performance issues with large datasets & proximity searches (*Elasticsearch aggregations can assist here)
- Flexibility in design increases development and support-time.
- Managing large deployments can be difficult due to sharding strategies



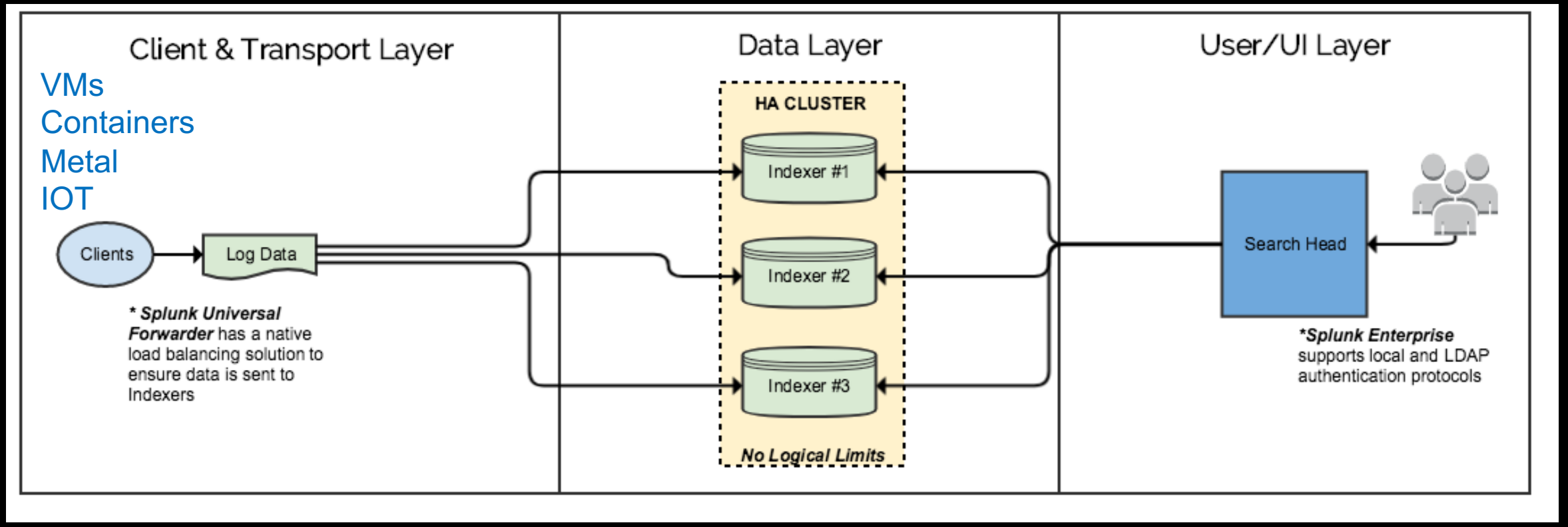
What is Splunk

- Splunk is a an enterprise commercial solution designed for log data search and analysis. It has 3 major components:

  - Universal Forwarder
    - This is the client layer that with the Splunk forwarding agent deployed will tail logs, monitor TCP ports, or run custom scripts and is designed to send data to Splunk indexers.
  - Indexers & Cluster Master
    - The is the data layer where log data is stored & aggregated for search & other analysis.
  - Search Head
    - The user interface (UI) that allows the user to investigate, aggregate & visualize the data stored in Splunk



Splunk Ecosystem



Pros

- Simplified stack requires less moving pieces
- Hosted solutions are available with Splunk Cloud (AWS)
- Schema-on-demand design allows for greater flexibility for data ingest
- Scaling to PB/day is quite doable
- Built-in user management, LDAP & SAML integrations
- Distributed map reduction capabilities will process 100's of millions of data points
- Data compression of 50%+ allows for more data in a smaller storage footprint
- Minimal logical limits on per cluster data storage.
- 1000's of 3rd party apps and plugins
- Strong user-driven support community

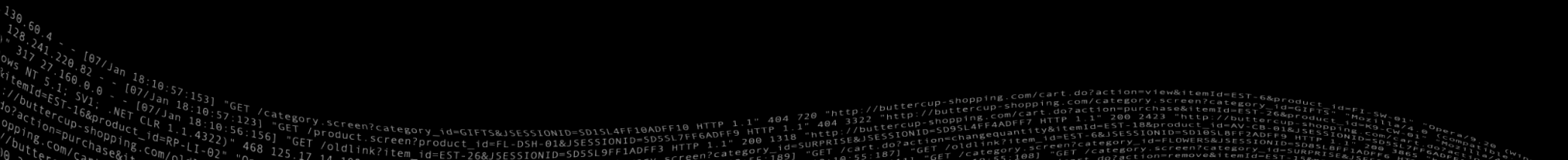
Cons

- Cost can be perceived as high
- Moderate learning curve to enable advanced analysis and features
- Logical limits on clustering/bucket replication
- Needs more nuanced documentation & reference implementation guides.
- Very large scale architecture can also be challenging in terms of design

Use Case

SSH Monitoring Use Case

- ***Stark Industries*** has deployed 5000 Linux hosts and wants to monitor log activity from both the VMs & the deployed applications. They also want to be able to alert on SSH login failures.



splunk> .conf18

SSH Monitoring Use Case

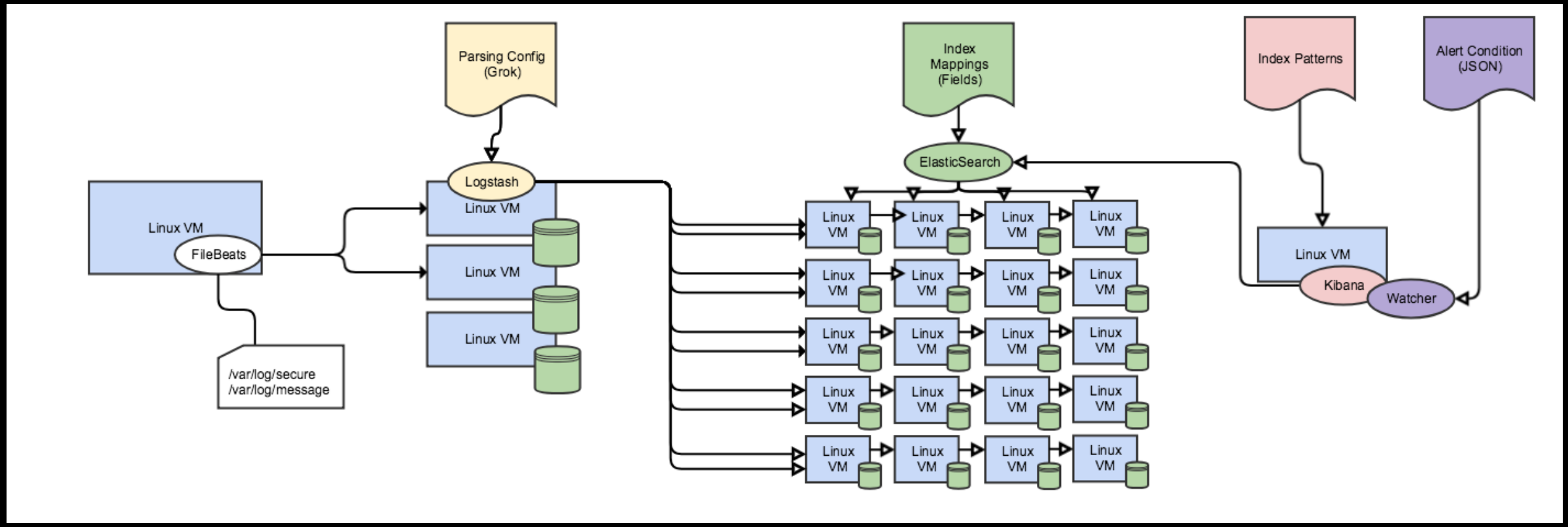
In this ELK scenario the company *Stark Industries* would need to deploy the following high-level components (* we will assume this is all net-new deployment)

1. Deploy a client agent (filebeats is a popular choice) to each of the 100 hosts
2. Build out a filebeat.yml configuration for input of SSH related logs & output to Logstash
3. Build out a Grok parsing config that matches the data to be ingested
4. Build a transport layer (logstash-server) & deploy the logstash.yml
5. Provision an ElasticSearch cluster that matches your retention and ingest needs.
6. Define an index & field pattern mapping for ElasticSearch ingest
7. Stand-up Kibana instance with default index patterns & enable the separate Watcher plugin to allow alerting on a search condition.
8. Define the alerting syntax needed (Lucene) and create the alert configuration file (JSON).

Assumptions

- All deployments are hosted in AWS for equal pricing comparisons
- Pricing & Deployment scale are based on 1TB/day of Ingest with a 30-day retention period
- Default replication will be used (5 primary shards and 1 replica per node)
- Follow default AWS sizing recommendations for Elastic
- Logstash priced configuration is designed to keep 3-days of pre-processed raw data available for re-indexing as needed.
- All costs reflected are for 1-month of continuous operation at On-Demand pricing
- An existing automation system for supported operations
- An existing monitoring system is in place for supported operations

ELK Ecosystem - SSH Monitoring Use Case



1. Filebeat.yml – stores the client's configuration of what files to monitor and where to output the log data
2. Logstash.yml - stores the parsing configuration and Grok patterns needed to parse the stream of log data into fields that ElasticSearch can store logically
3. ElasticSearch index configuration – this is not stored on disk and can only be managed through the API
4. Index Patterns – these are stored within the .kibana index provisioned (typically) on the same ElasticSearch cluster and is defined with a JSON object
5. Alert Condition – this is stored within the .watcher index provisioned (typically) on the same ElasticSearch cluster and is defined with a JSON object

ELK Monthly Cost Estimate

ELK Estimate				
Service Type	Components	Region	Component Price	Service Price
Amazon EC2 Service (US East (N. Virginia))				$21,678.38
20 x I3.4xlarge + 3 t2.xlarge	Compute:	US East (N. Virginia)	$18,678.38	
GP SSD – 4500 IOPS	EBS Volumes:	US East (N. Virginia)	$3,000	
	EBS IOPS:	US East (N. Virginia)	$0	
AWS Support (Business)				$1,817.28
	Support for all AWS services:		$1,817.28	
		Free Tier Discount:		($3)
		Total Monthly Payment:		**$23,492.66**

splunk> .conf18

SSH Monitoring Use Case

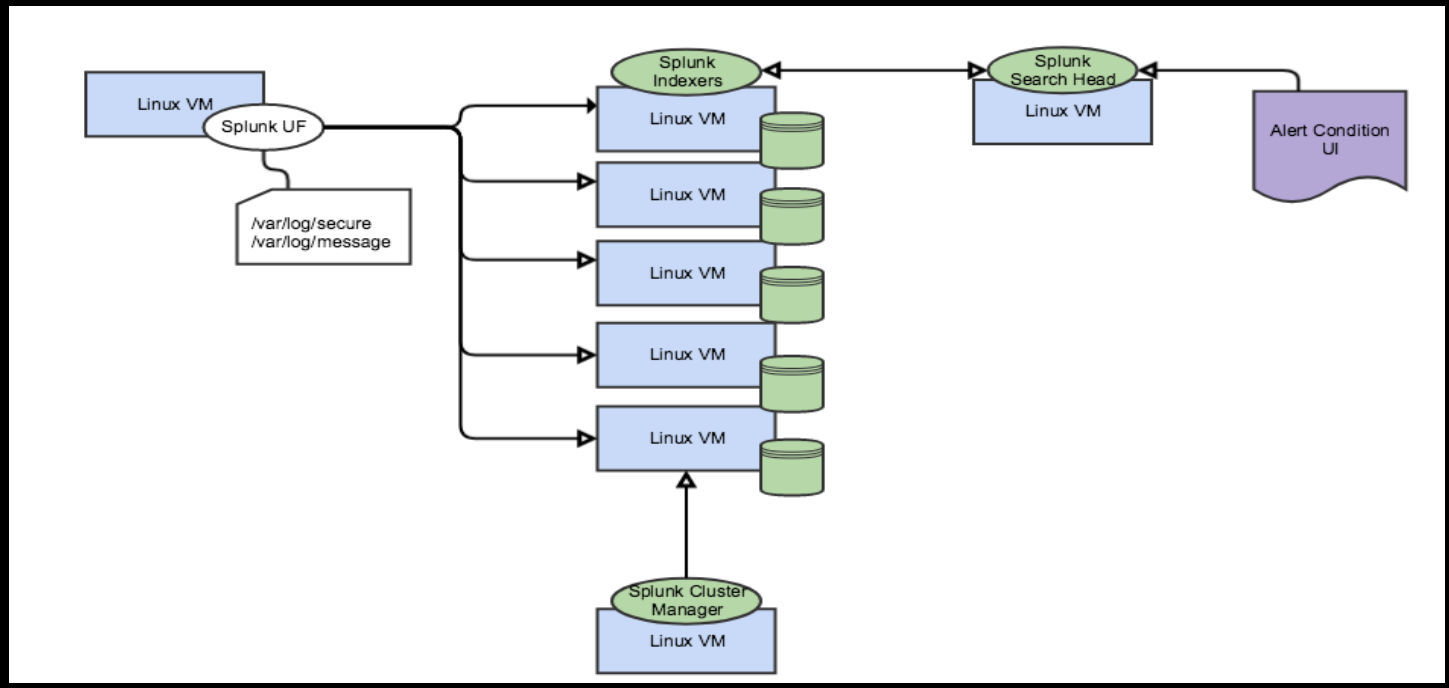
- Scaling:
  - Logstash & ElasticSearch scale horizontally
  - As throughput of the deployment goes up:
    - Additional Logstash nodes & storage will be needed to process the data before it's indexed.
      - This layer also requires a back-pressure memory configuration to handle persistent queuing. The default is 4Gb but may need to be increased to 8GB as deployment throughput goes up.
  - This can lead to escalating infrastructure costs as the deployment grows and additional capacity is needed for retention, increased ingest capacity & High Availability.
  - Logging also tends to be spiky meaning that capacity will need to be provisioned for peak ingest times.
- Operating
  - In an ELK deployment you are looking at a minimum of 4 tools (Beat, ElasticSearch, Logstash & Kibana) in the stack.
    - This can require expertise across multiple toolsets that can be difficult to find and/or find training available.
    - Large time commitment to creating in-house knowledge, documentation and training to use and maintain a customized system
    - Additional configuration management & monitoring tools are needed.
- Support
  - Vendor support & consulting is available at additional cost

SSH Monitoring Use Case

In this Splunk scenario the company *Stark Industries* would need to deploy the following high-level components (* we will assume this is all net-new deployment)

1. Deploy the Universal Splunk forwarder to each of the 5000 hosts
2. Build out an inputs configuration to capture SSH related data
3. Build an outputs configuration to send data to Splunk indexers
4. Provision a cluster of Splunk indexers
5. Provision a cluster manager to manage indices on Splunk indexers
6. Stand up a dedicated Splunk search head & peer to the provisioned Splunk indexers
7. Verify the data & setup the alert with the Splunk UI

Splunk – SSH Use Case w/Splunk Classic Architecture



This deployment needs the following independent configurations:

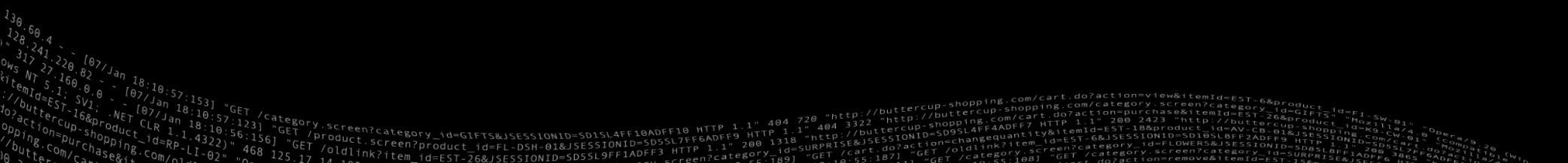
1. Inputs.conf– stores the client's configuration of what files to monitor and metadata (index, sourcetype, etc)
2. Server.conf – stores the cluster manager specifications for replication & search factor
3. Outputs.conf- stores the clients configuration of where to send the monitored data
4. SavedSearch.conf – where the alert configuration is actively stored. However the alert itself is defined through the UI

Assumptions

- All deployments are hosted in AWS for equal pricing comparisons
- Pricing & Deployment scale are based on 1TB/day of Ingest with a 30-day retention period
- Splunk is configured for a RF:2 SF:2 factor for HA
- All costs reflected are for 1-month of continuous operation at On-Demand pricing
- An existing automation system for supported operations
- An existing monitoring system is in place for supported operations

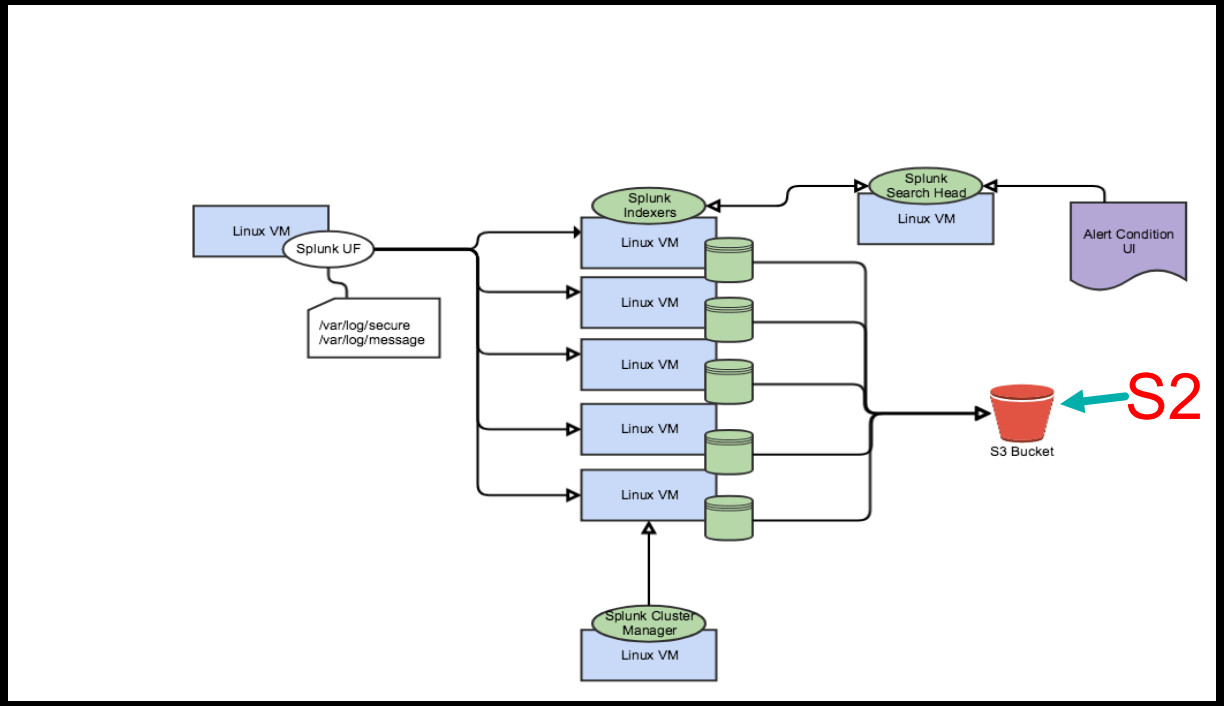
Splunk Monthly Cost Estimate – “Classic” Architecture

Splunk Estimate				
Service Type	Components	Region	Component Price	Service Price
Amazon EC2 Service (US East (N. Virginia))				$9,394.78
7 x i3.4xlarge	Compute:	US East (N. Virginia)	$6,394.78	
5x 6TB GP SSD Volumes	EBS Volumes:	US East (N. Virginia)	$3,000	
	EBS IOPS:	US East (N. Virginia)	$0	
AWS Support (Business)				$939.18
	Support for all AWS services:		$939.18	
		Free Tier Discount:		($3)
		Total Monthly Payment:		**$10,330.96**



splunk> .conf18

Splunk – SSH Use Case w/S2 option



This deployment needs the following independent configurations:

1. Inputs.conf– stores the client's configuration of what files to monitor and metadata (index, sourcetype, etc)
2. Server.conf – stores the cluster manager specifications for replication & search factor
3. Outputs.conf- stores the clients configuration of where to send the monitored data
4. SavedSearch.conf – where the alert configuration is actively stored. However the alert itself is defined through the UI
5. S2 Implementation added where S3 object store is now used & disk allocated to indexing tier is a hot/warm cache layer instead.

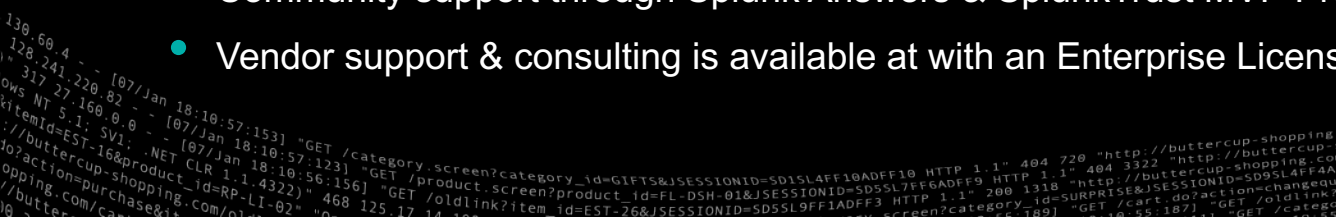
Splunk Monthly Cost Estimate – S2 Architecture

Splunk S2 Estimate				
Service Type	Components	Region	Component Price	Service Price
Amazon EC2 Service (US East (N. Virginia))				$7,394.78
7 x i3.4xlarge	Compute:	US East (N. Virginia)	$6,394.78	
5x 2TB GP SSD Volumes (Hot/Warm Cache)	EBS Volumes:	US East (N. Virginia)	$1,000	
	EBS IOPS:	US East (N. Virginia)	$0	
Amazon S3 Service (US East (N. Virginia))				$706.56
30 TB S3 Storage	S3 Standard Storage:	US East (N. Virginia)	$706.56	
AWS Support (Business)				$809.83
	Support for all AWS services:		$809.83	
		Free Tier Discount:		($3.12)
		Total Monthly Payment:		**$8,908.05**

Estimated %25 cost reduction/month

Business Considerations

- Scaling
  - Splunk will also scale horizontally
    - Indexing & Search layers will scale independently based on search need and indexing need
  - As throughput of the deployment goes up:
    - Additional indexers will be needed to accommodate more ingest but with the compression ratio of 50% you can keep more log data in a smaller footprint.
  - No hard disk size limits on Splunk indexers (can exceed 1.5TB of available data) but some limitations on bucket replications.
  - Deployments have been able to scale to several PBs per day
  - Splunk has a built-in queueing mechanism that allows for more flexibility over peak ingest times.
  - Splunk integration with S3 objects stores can assist in storage costs & retention mandates
- Operating
  - Splunk is a single eco-system that has a consistent framework throughout helping make overall administration more manageable by fewer staff
  - Splunk has a built-in configuration management system (Deployment Server & Deployers) to help with consistent configuration, but also easily integrates with Ansible, Chef, Jenkins and other frameworks.
- Support
  - Professional Training & Certification paths available
  - Community support through Splunk Answers & SplunkTrust MVP Program available.
  - Vendor support & consulting is available at with an Enterprise License purchase



splunk> .conf18

Q & A

Links:

https://www.elastic.co/guide/en/logstash/current/deploying-and-scaling.html
https://thoughts.t37.net/designing-the-perfect-elasticsearch-cluster-the-almost-definitive-guide-e614eabc1a87
https://calculator.s3.amazonaws.com/index.html
https://www.elastic.co/guide/en/logstash/current/tuning-logstash.html
http://splunk-sizing.appspot.com/
https://logz.io/blog/elastic-stack-6-new/
https://www.elastic.co/guide/en/elasticsearch/hadoop/current/mapreduce.html
https://www.elastic.co/blog/how-to-enable-saml-authentication-in-kibana-and-elasticsearch
https://www.datadoghq.com/blog/elasticsearch-performance-scaling-problems/
https://www.elastic.co/guide/en/x-pack/current/xpack-introduction.html

Thank You

Don't forget to rate this session in the .conf18 mobile app