

Log Analysis as a Service using open source scalable systems

Gurvinder Singh Dahiya, Uninett AS
Belgrade Security Workshop, 20.03.2015



Motivation

- Distributed Systems
- Centralized interface to logs
- Easier access
- Detection of hidden pattern
- Access to logs across organization
- Cenralized alerts and anomaly detection across services



Challenges

- Different Formats and logging methods
- Different requirements for processing
- Different Dashboards
- Various Alerts requirements



What is Log ?

```
Jun  2 07:40:34 scintilla kernel: [77262.488918] hid-generic 0003:046D:0A15.001C: input,hidraw4: USB HID v1.00 Device [Logitech Logitech G35 Headset] on usb-0000:00:1d.0-1.5.4/input3
```

```
[2014-06-02 14:08:39,870][INFO ][cluster.metadata      ] [pltrd003] [mail-2014.06.02] update_mapping [mail] (dynamic)
```

```
2014-06-02T12:11:25.271Z 158.36.2.74 https://idp.feide.no feide:sso ntnu.no [u'urn:mace:feide.no:services:no.ntnu.ssowrapper'] 1401711085.27
```

```
158.38.213.3 - - [02/Jun/2014:14:12:48 +0200] "POST /__es/logstash-2014.05.26/_search HTTP/1.1" 200 329628 "https://logs.uninett.no/"; "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/34.0.1847.137 Safari/537.36"
```

```
DEBUG 2014-06-02 14:15:13,371 [Thread-110743] no.uninett.agora.user.UserRoleSiteUtil Querying for liferay sites for roles: [fs]
```

TIMESTAMP + DATA = LOG

Time Formats

PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27


THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13

20130227 2013.02.27 27.02.13 27-02-13

27.2.13 2013.II.27. $27\frac{1}{2}$ -13 2013.158904109

MMXIII-II-XXVII MMXIII $\frac{LVII}{CCCLXV}$ 1330300800

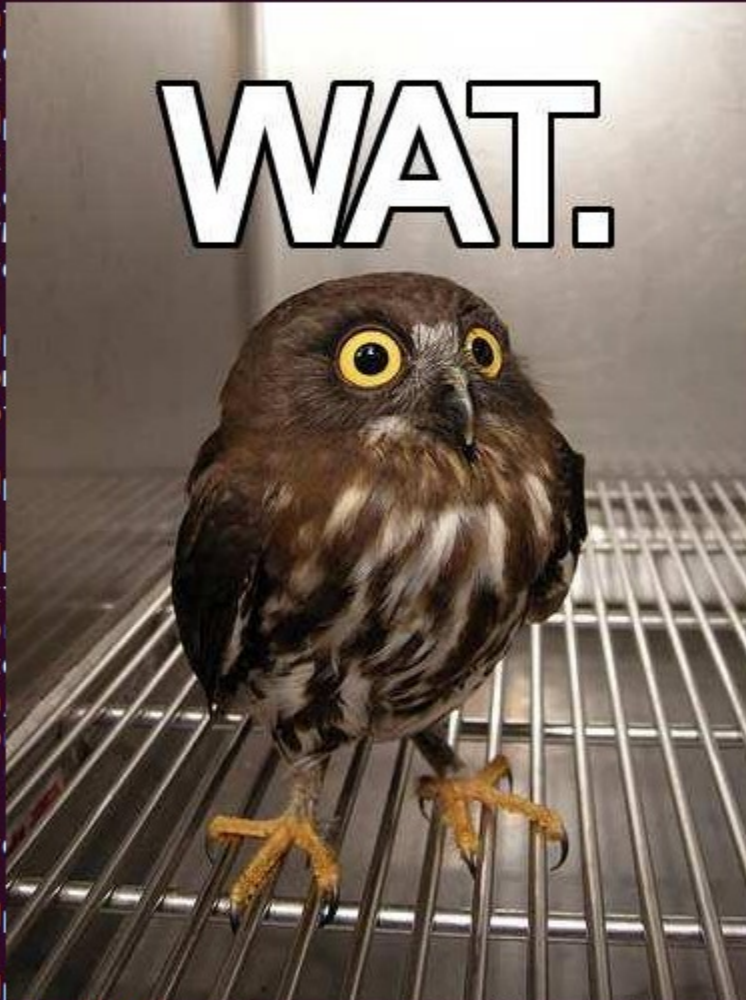
$((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$ ~~2013~~ 

10/11011/1101 02/27/20/13 $\begin{matrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ 5 & 6 & 7 & 8 \end{matrix}$

A log is human readable...

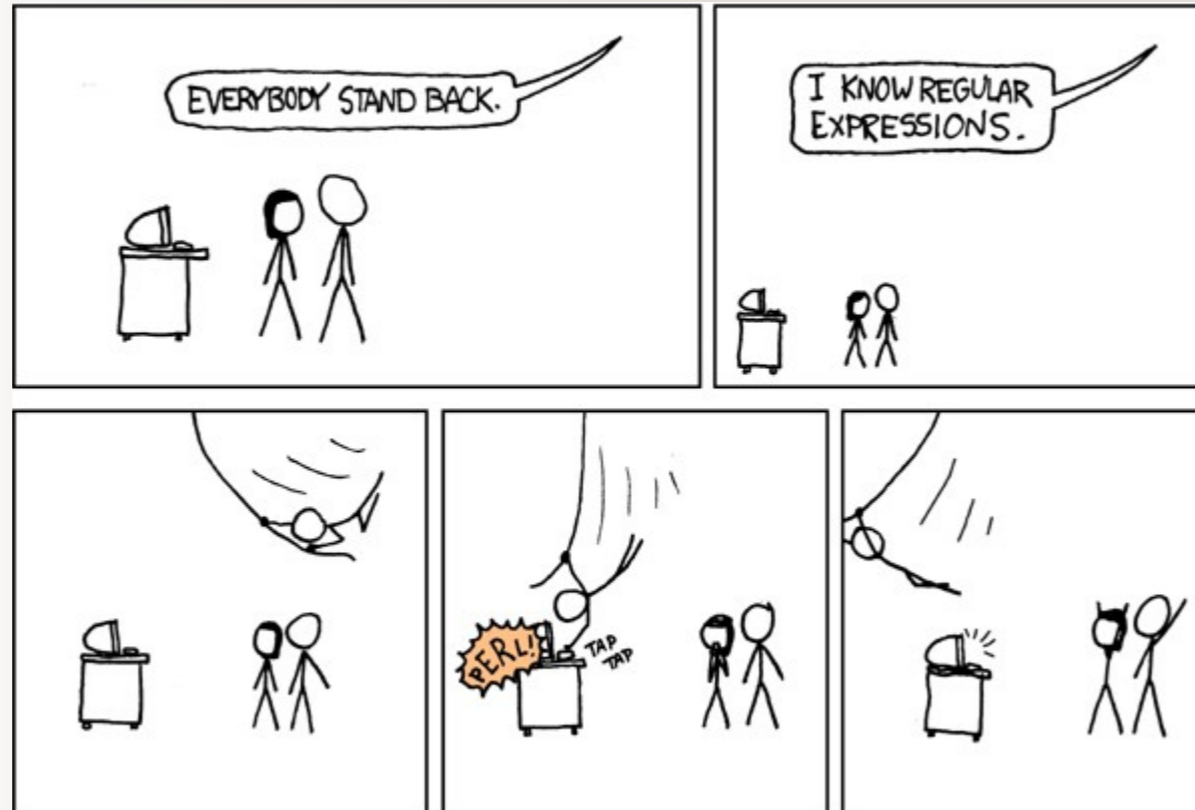
```
Jun  2 07:40:34 scintilla kernel: [77262.488918] hid-generic 0003:046D:0A15.001C: input,hidraw4:  
USB HID v1.00 Device [Logitech Logitech G35 Headset] on usb-0000:00:1d.0-1.5.4/input3
```

```
65.55.215.109 - - [18/Jan/2013:19:57:13 -0500] "GET /robots.txt HTTP/1.1" 301 303 "-" "msnbot-media/1.1 (+http://search.msn.com/msnbot.htm)"
65.55.215.109 - - [18/Jan/2013:19:57:13 -0500] "GET /gallery/gallery/Subaru/ormeau_-_25_April_2005/thumbs/IMG_2854.JPG HTTP/1.1" 301 303 "-" "msnbot-media/1.1 (+http://search.msn.com/msnbot.htm)"
178.255.215.69 - - [18/Jan/2013:20:02:41 -0500] "GET /robots.txt HTTP/1.1" 301 303 "-" "Mozilla/5.0 (compatible; Exabot/3.0; +http://www.exabot.com/go/robot)"
178.255.215.69 - - [18/Jan/2013:20:02:42 -0500] "GET /wordpress/tag/smoke/ HTTP/1.1" 301 303 "-" "Mozilla/5.0 (compatible; Exabot/3.0; +http://www.exabot.com/go/robot)"
66.249.73.60 - - [18/Jan/2013:20:08:19 -0500] "GET /wordpress/cooking/mozambique-style-piri-piri-chicken/feed/ HTTP/1.1" 301 351 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
24.225.83.49 - - [18/Jan/2013:20:17:53 -0500] "GET /favoritehen.html" "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; WOW64; Trident/5.0)"
66.249.73.60 - - [18/Jan/2013:20:21:19 -0500] "GET /wordpress/0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
205.169.30.2 - - [18/Jan/2013:20:23:44 -0500] "GET /favoritehen.html" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.17 (KHTML, like Gecko) Chrome/24.0.1313.107 Safari/537.17"
72.14.199.133 - - [18/Jan/2013:20:23:45 -0500] "GET /wordpress/r.html; 18 subscribers; feed-id=12636598490283692241)"
66.249.73.60 - - [18/Jan/2013:20:34:00 -0500] "GET /wordpress/la/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
157.55.33.80 - - [18/Jan/2013:20:48:51 -0500] "GET /robots.txt" "Mozilla/5.0 (compatible; Bingbot/2.0; +http://www.bing.com/bingbot.htm)"
157.55.33.80 - - [18/Jan/2013:20:50:50 -0500] "GET /wordpress/.bing.com/bingbot.htm)"
66.249.73.60 - - [18/Jan/2013:21:00:01 -0500] "GET /wordpress/301 350 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
58.22.10.92 - - [18/Jan/2013:21:00:04 -0500] "GET /wordpress/HTTP/1.0" 301 352 "http://xesla.ro/wordpress/cooking/past-and-pastrami-pizza-with-garlic-pizza-fritta/pizza-with-garlic-pizza-fritta/" "Opera/9.80 (Windows NT 6.1; Win64; x64; U; ru) Presto/2.10.289 Version 11.12"
24.220.125.231 - - [18/Jan/2013:21:06:00 -0500] "GET /wordpress/on-recipes/venison-stews/wild-game-gumbo/" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_7_4) AppleWebKit/536.26.17 (KHTML, like Gecko) Version/6.0.2 Safari/536.26.17"
72.14.199.133 - - [18/Jan/2013:21:23:45 -0500] "GET /wordpress/r.html; 18 subscribers; feed-id=12636598490283692241)"
66.249.73.60 - - [18/Jan/2013:21:26:07 -0500] "GET /wordpress/glebot/2.1; +http://www.google.com/bot.html)"
66.249.73.60 - - [18/Jan/2013:21:28:57 -0500] "GET /wordpress/01 332 "-" "DoCoMo/2.0 N905i(c100;TB;W24H16) (compatible; Googlebot-Mobile/2.1; +http://www.google.com/bot.html)"
95.108.150.235 - - [18/Jan/2013:21:29:07 -0500] "GET /robots.txt HTTP/1.1" 301 303 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
95.108.150.235 - - [18/Jan/2013:21:29:07 -0500] "GET /robots.txt HTTP/1.1" 301 307 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
95.108.150.235 - - [18/Jan/2013:21:29:08 -0500] "GET / HTTP/1.1" 301 303 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
95.108.150.235 - - [18/Jan/2013:21:29:08 -0500] "GET / HTTP/1.1" 301 307 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
94.228.34.233 - - [18/Jan/2013:21:30:08 -0500] "GET /wordpress/cooking/chocolate-bread/feed/ HTTP/1.1" 301 332 "-" "magpie-crawler/1.1 (U; Linux amd64; en-GB; +http://www.brandwatch.net)"
```



But machine parsable ... maybe ?

Jun 2 07:40:34 scintilla kernel: [77262.488918] hid-generic 0003:046D:0A15.001C: input,hidraw4: USB HID v1.00 Device [Logitech Logitech G35 Headset] on usb-0000:00:1d.0-1.5.4/input3



Apache Regex ..

```
((?:\b(?:[0-9A-Za-z][0-9A-Za-z-]{0,62})(?:\.(?:[0-9A-Za-z][0-9A-Za-z-]{0,62}))*(\.\b)|((?!<[0-9])(?:25[0-5]|2[0-4][0-9]|[0-1]?[0-9]{1,2})[.](?:25[0-5]|2[0-4][0-9]|[0-1]?[0-9]{1,2})[.](?:25[0-5]|2[0-4][0-9]|[0-1]?[0-9]{1,2})[.](?:25[0-5]|2[0-4][0-9]|[0-1]?[0-9]{1,2}))(![0-9])) ([a-zA-Z0-9_-]+) ([a-zA-Z0-9_-]+) \(((?:3[01]|[1-2]?[0-9]|0?[1-9]))/(\b(?:Jan(?:uary)?|Feb(?:ruary)?|Mar(?:ch)?|Apr(?:il)?|May|Jun(?:e)?|Jul(?:y)?|Aug(?:ust)?|Sep(?:tember)?|Oct(?:ober)?|Nov(?:ember)?|Dec(?:ember)?)\b)/([0-9]+):((?!<[0-9])(?:2[0123]|[01][0-9])):(?:[0-5][0-9]))(?:((?:[0-5][0-9]|60)(?:[.,][0-9]+)?))(![0-9]) ((?:[+-]?([0-9]+)))\ ("(\b\w+\b) (((?:/[A-Za-z0-9$.+!*'(),~:#%_-]*+)(?:\?[A-Za-z0-9$.+!*'(),~:#%&/=;:_-]*))?) HTTP/((?:((?!<[0-9.+])(>[+-]?((?:[0-9]+(?:\.[0-9]+)?)(?:\.[0-9]+))))))" ((?:((?!<[0-9.+])(>[+-]?((?:[0-9]+(?:\.[0-9]+)?)(?:\.[0-9]+))))))|(?:(?:((?!<[0-9.+])(>[+-]?((?:[0-9]+(?:\.[0-9]+)?)(?:\.[0-9]+))))))|(?:(?:([A-Za-z]+(\+[A-Za-z]+)?)://(?:((([a-zA-Z0-9_-]+))?:[^\@]*?)@)?(?:(((?:\b(?:[0-9A-Za-z][0-9A-Za-z-]{0,62})(?:\.(?:[0-9A-Za-z][0-9A-Za-z-]{0,62}))*(\.\b)|((?!<[0-9])(?:25[0-5]|2[0-4][0-9]|[0-1]?[0-9]{1,2})[.](?:25[0-5]|2[0-4][0-9]|[0-1]?[0-9]{1,2})[.](?:25[0-5]|2[0-4][0-9]|[0-1]?[0-9]{1,2}))(![0-9]))))?:\b(?:[0-9]+)\b))?)? (?:(((?:/[A-Za-z0-9$.+!*'(),~:#%_-]*+)(?:\?[A-Za-z0-9$.+!*'(),~:#%&/=;:_-]*))?)|) (((?:(?<!\|)(?:"(?:\.[^\|"])*"|(?:'(?:\.[^\|']*)'|(?:(?:\.[^\|])*)))
```

Evolution of log processing

- Stage 1 - Single Host

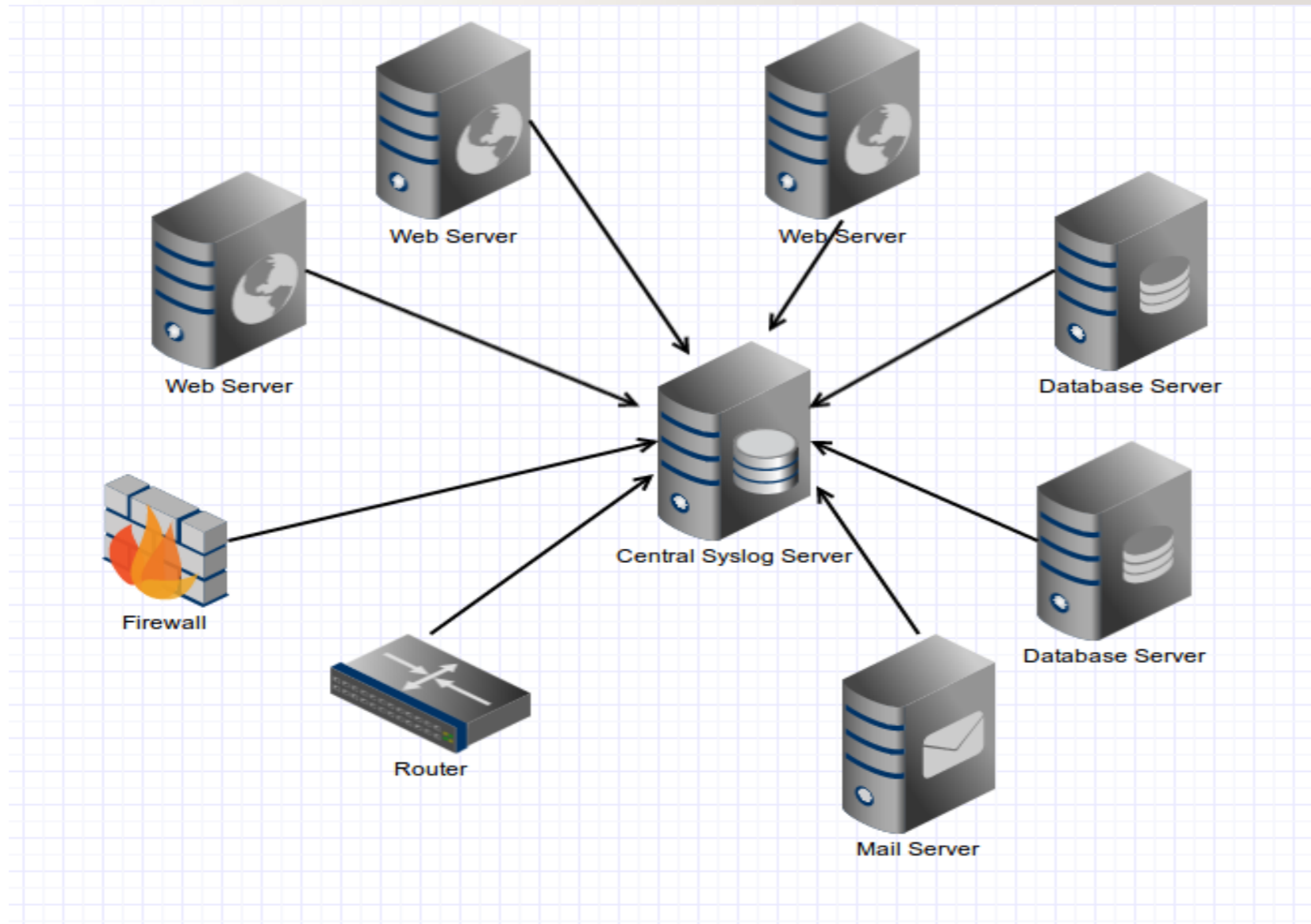
```
$ grep -I "Invalid user " /var/log/auth.log* | awk '{ print $10; }'
```

- Stage 2 - A handful of hosts

```
#!/bin/sh  
USER=root  
KEY=/root/public_key.pub  
for HOST in server1 server2 server3 server4  
do  
  ssh -I $USER -i $KEY $HOST grep -I "Invalid user " \  
    /var/log/auth.log* | awk '{ print $10; }'  
done
```

Evolution of log processing

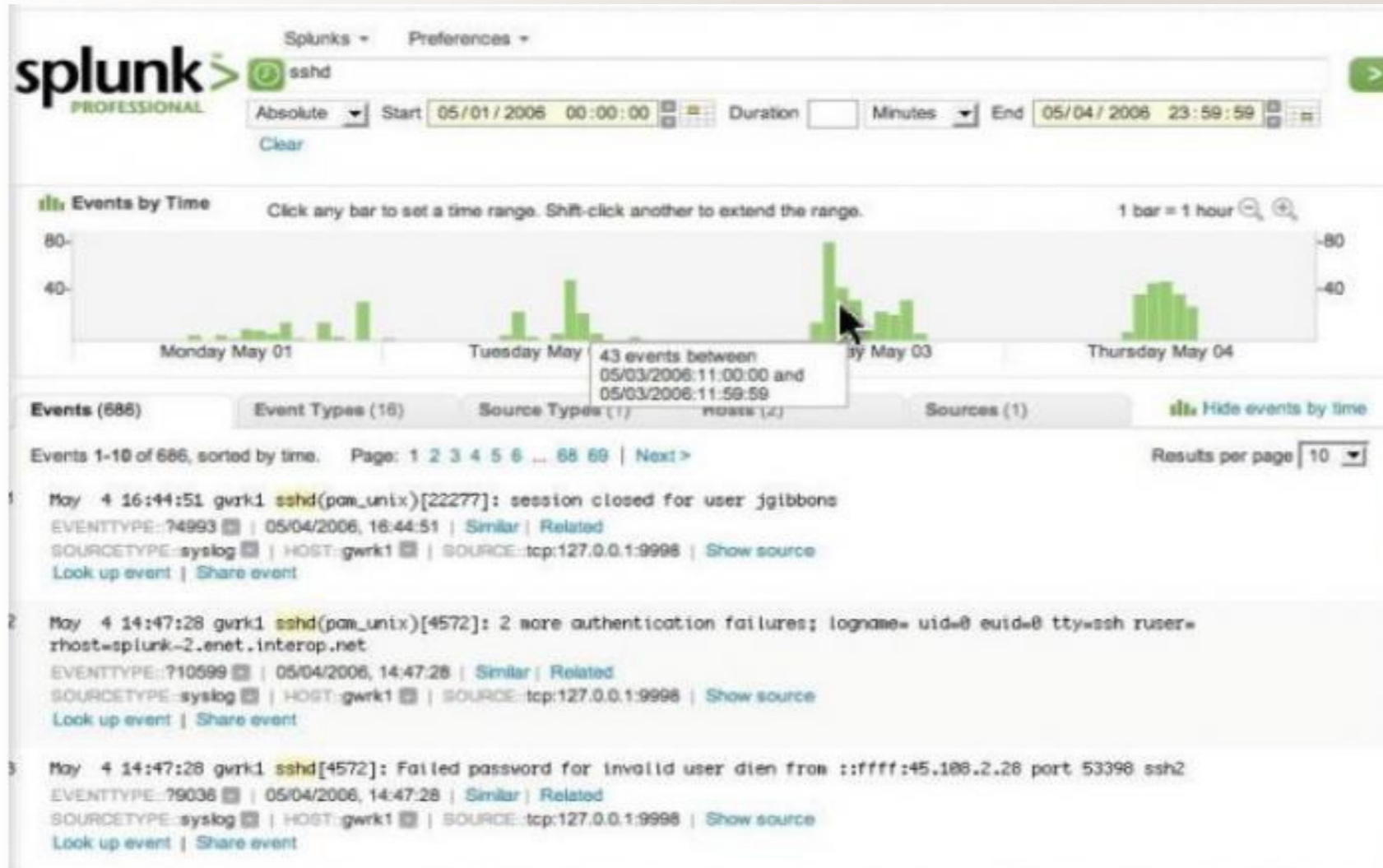
- Stage 3 - Lot of servers



```
$ grep -I "Invalid user " /rsyslog/*/auth.log* | awk '{ print $10; }'
```

Evolution of log processing

- Stage 4 - Start using splunk



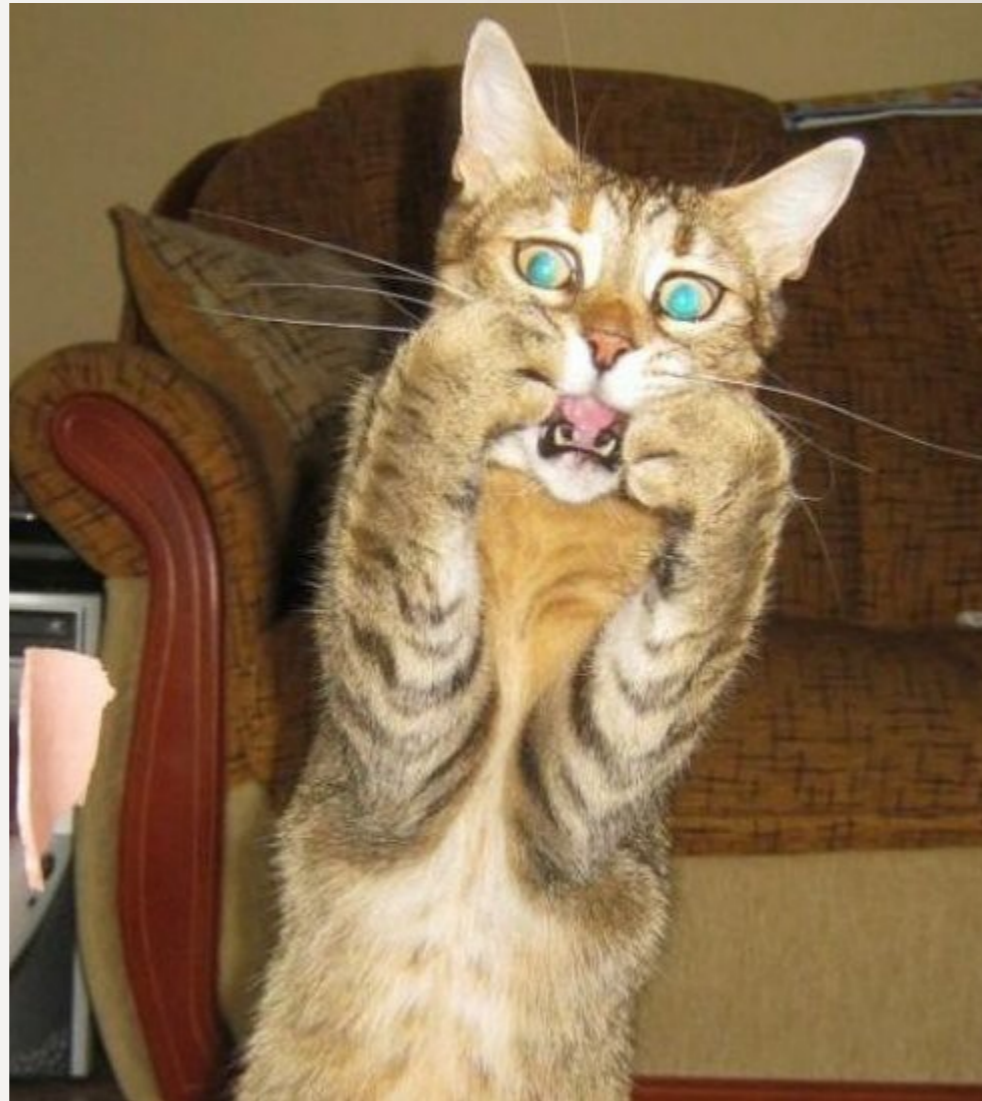
Evolution of log processing

- The first one is free - 500 MB per day



Evolution of log processing

- Incoming invoice from Splunk



Evolution of log processing

- Stage 5 - Open source scalable solutions
 - Logstash
 - Elasticsearch
 - Kibana
 - ZeroMQ
 - Logstash-forwarder
 - Rsyslog
 - Statsd

Logstash

- Turns this:
 - `«192.168.0.74 - - [13/May/2014:04:28:55 -0500] "GET /robots.txt HTTP/1.1" 301 303 "-" "Mozilla/5.0 (compatible; DSASE/1.0; bot@gmail.com)"»`

- Into:

```
{  
  "client address": "192.168.0.74",  
  "user": null,  
  "timestamp": "2014-05-13T14:04:28-0500",  
  "verb": "GET",  
  "path": "/robots.txt",  
  "query": null,  
  "http version": 1.1,  
  "response code": 301,  
  "bytes": 303,  
  "referrer": null  
  "user agent": "Mozilla/5.0 (compatible; DSASE/1.0; bot@gmail.com)"  
}
```


Elasticsearch

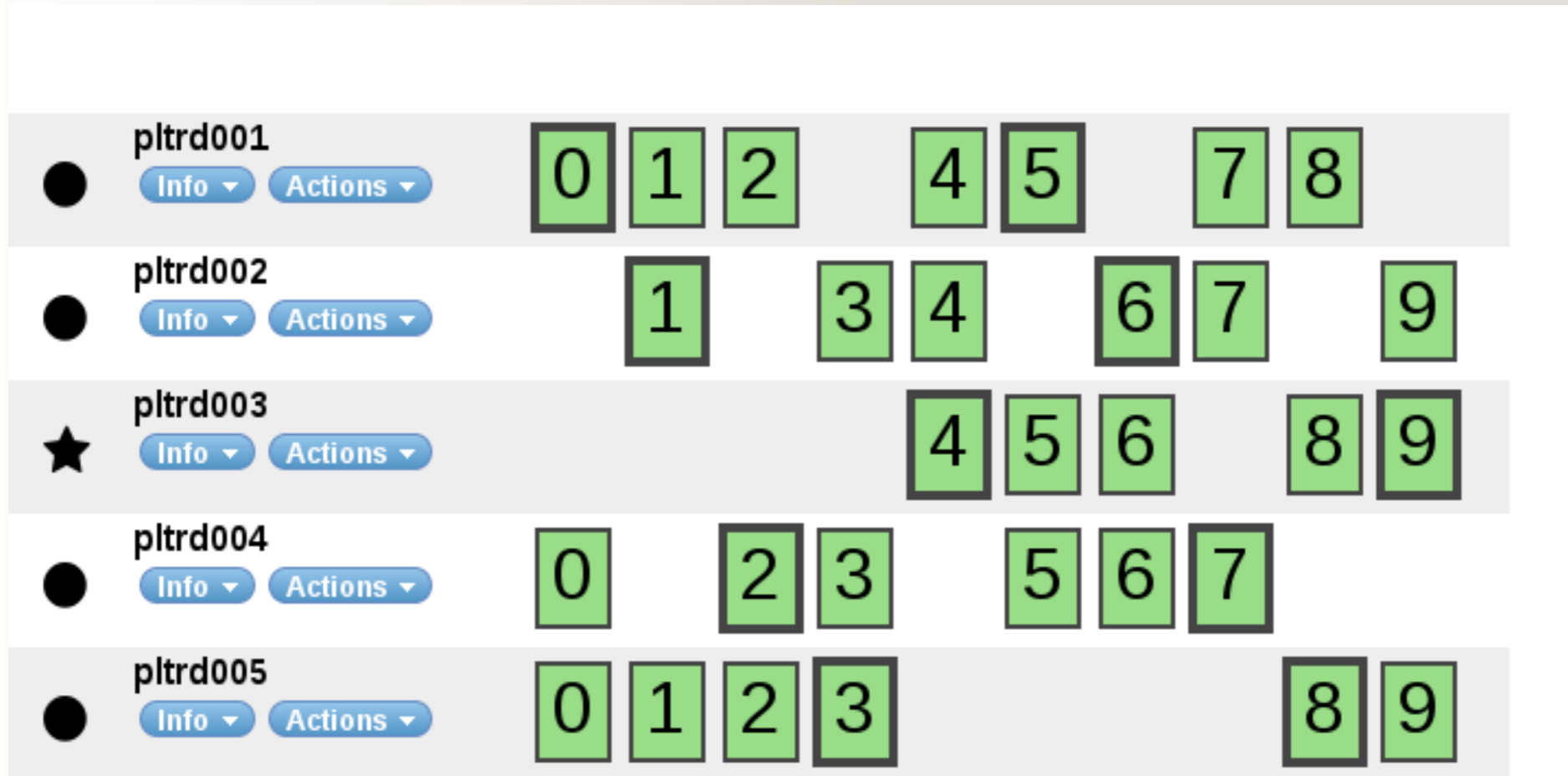
- Document - Oriented Free Text Search & Analytics Engine
- JSON
- Apache Lucene
- No Schema
- Mapping Types
- Horizontally Scaleable, Distributed
- REST API
- Vibrant Ecosystem
- Completely Open Source

Elasticsearch

- Index
 - Logical collection of data; might be time based
 - Analogous to a database
- Sharding
 - Split logical data over several machines
 - Write scalability
 - Control data flows
- Replication
 - Read scalability
 - Removing SPOF

Elasticsearch Shards and Replication

- Index allocation with 10 shards and 3 replication factors



Kibana

- Javascript based web application
- No dependency except a web server (Changing a bit in Kibana 4)
- Visualize data stored in Elasticsearch
- Dynamic panel and dashboard creation support
- Multiple panel types support
 - Tables
 - Pie charts
 - Maps
 - Text
 - Trends
 - Histograms

Kibana Visualization

Gun Death Map

Kibana 3 milestone 2

This dashboard is updated daily from data provided by slate.com. Every data point is a single human life taken by a gun in the United States since December 13th, 2012. Click on clusters to drill down. Hover over blue markers to see the names of the deceased.

Since

12/13/2012

00:00:00

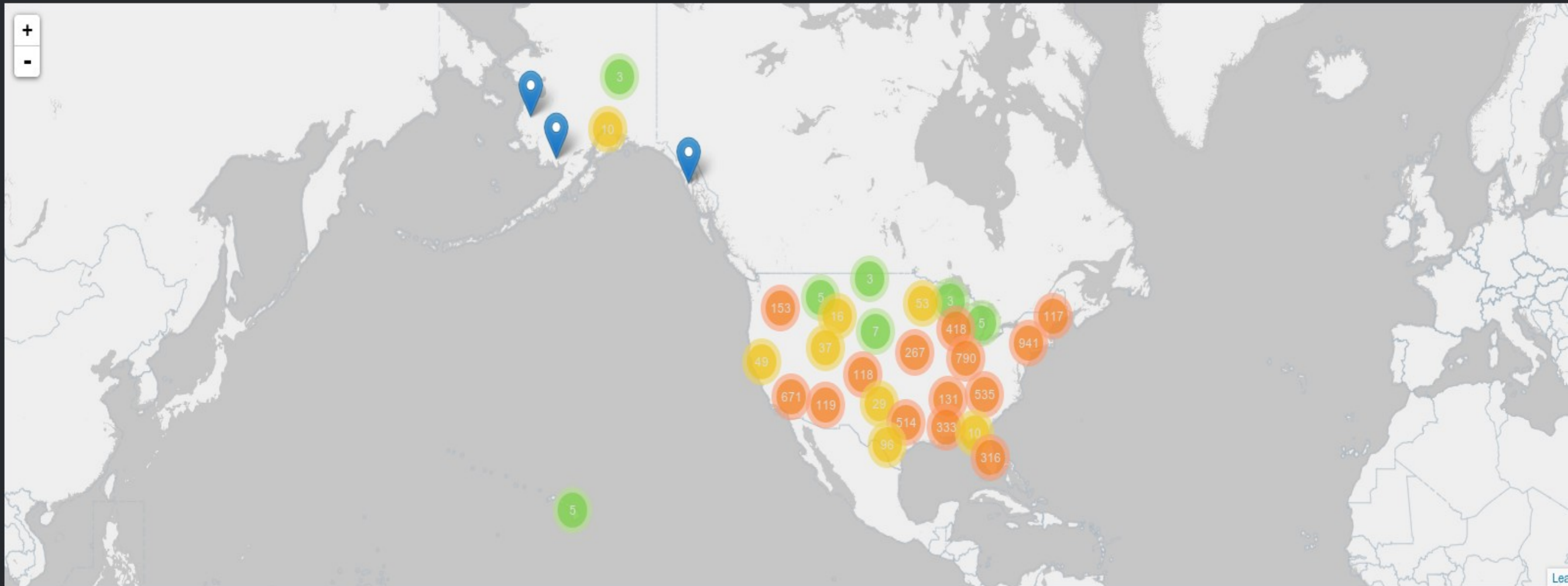


Relative | Absolute | **Since** | Auto-refresh

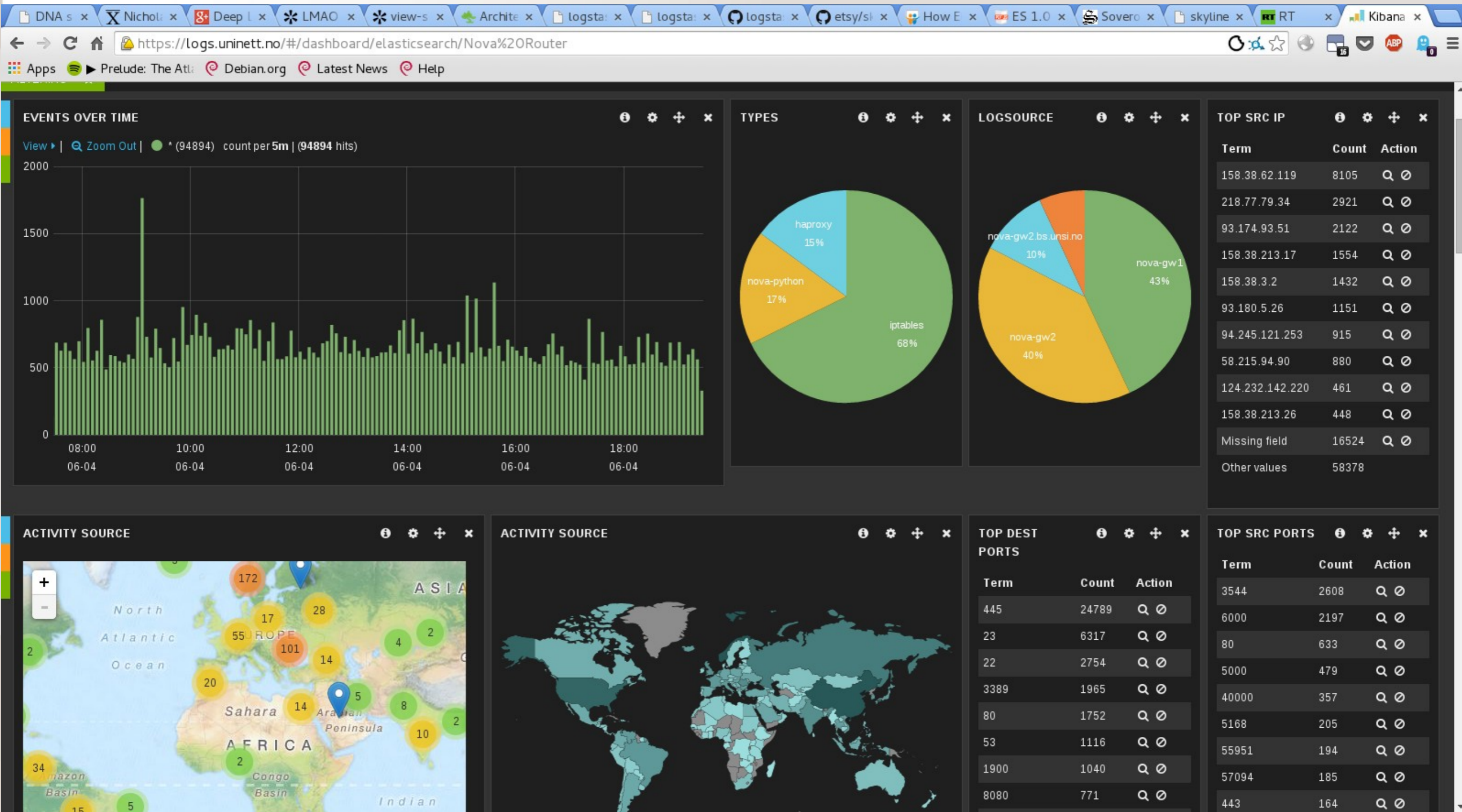
Dashboard Control



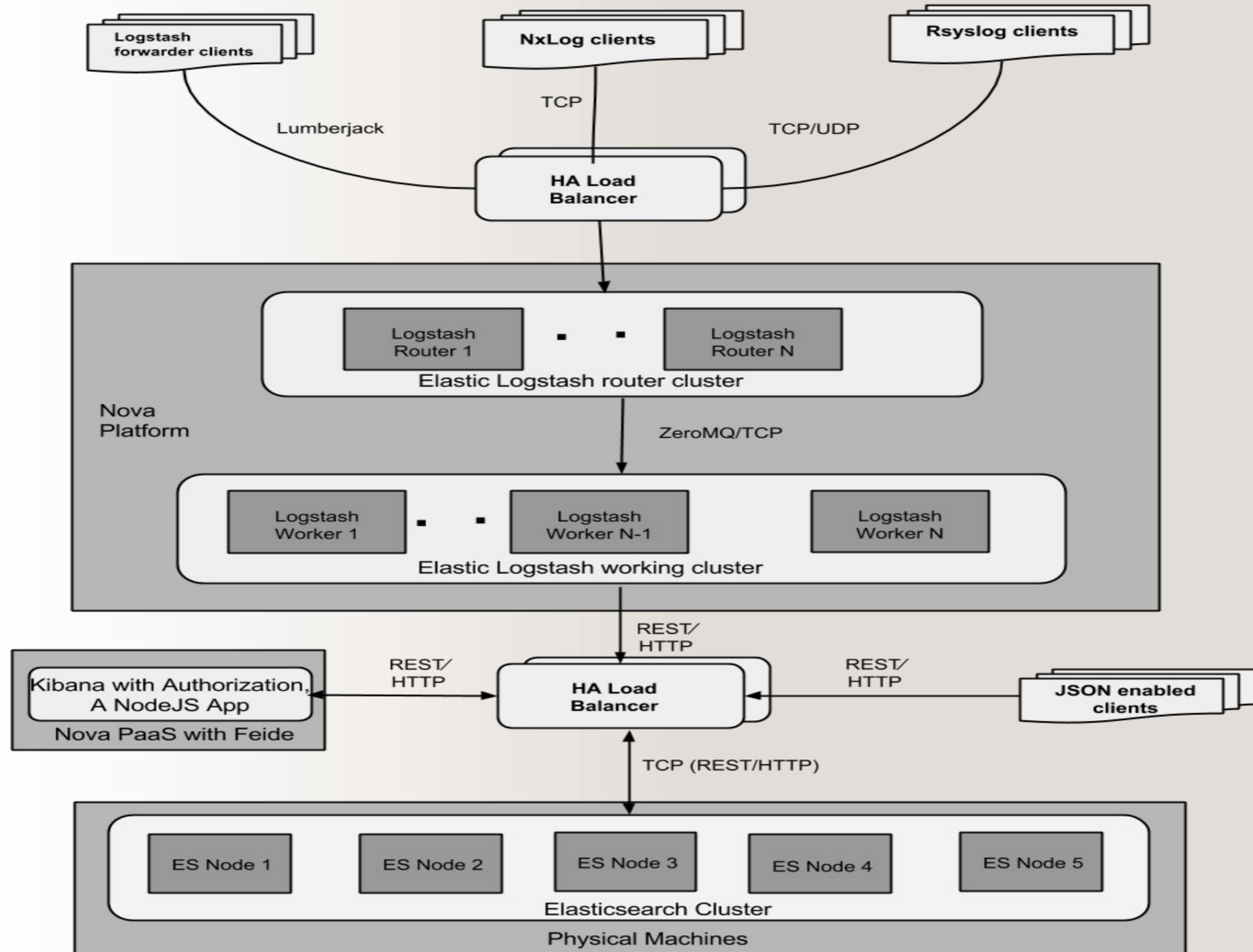
People killed by Guns in the United States



Kibana Visualization



Lego Brick-Built



Authentication & Authorization

- Sensitive information
- Privacy concerns
- Support LDAP groups
- Support Feide, Norwegian single-sign on solution
- A nodejs app, forked from another kibana-proxy
- Access is allowed on per service basis by service owner

Data Model

- Normalize the diversity of information
- Logsource field for host information
- Username
- src_ip, dest_ip, src_port, dest_port
- _service
- type

@timestamp	Q 0 𐀀	2014-06-04T19:04:28.000Z
@version	Q 0 𐀀	1
_id	Q 0 𐀀	87f2ed1bf7a0806f5ee6732a860bb1f4
_index	Q 0 𐀀	dns-2014.06.04
_service	Q 0 𐀀	dns
_type	Q 0 𐀀	dns
direction	Q 0 𐀀	IN
flags	Q 0 𐀀	+
geoip.asn	Q 0 𐀀	UNINETT, The Norwegian University & Research Network
geoip.continent_code	Q 0 𐀀	EU
geoip.country_code2	Q 0 𐀀	NO
geoip.country_code3	Q 0 𐀀	NOR
geoip.country_name	Q 0 𐀀	Norway
geoip.ip	Q 0 𐀀	158.38.100.171
geoip.latitude	Q 0 𐀀	62
geoip.location	Q 0 𐀀	10,62
geoip.longitude	Q 0 𐀀	10
geoip.number	Q 0 𐀀	AS224
geoip.timezone	Q 0 𐀀	Europe/Oslo
host	Q 0 𐀀	2001:700:1:8:0:0:180:57:58234
logsource	Q 0 𐀀	moholt
message	Q 0 𐀀	Jun 4 21:04:28 moholt named[2124]: client 158.38.100.171#45933: query: puppet.uninett.no.uninett.no IN AAAA + (158.38.180.57)
pid	Q 0 𐀀	2124
program	Q 0 𐀀	named
qtype	Q 0 𐀀	AAAA
query	Q 0 𐀀	puppet.uninett.no.uninett.no
src_ip	Q 0 𐀀	158.38.100.171
src_port	Q 0 𐀀	45933
timestamp	Q 0 𐀀	Jun 4 21:04:28
tld	Q 0 𐀀	no
type	Q 0 𐀀	dns



Log Analysis as Service (LAAS)

- It's up and running
- Pilot is offered till April 14, 2015 and production from April 15, 2015
- In pilot with Geant SA7-T1 activity
 - Collaboration with GarrNet, Heanet
- Following institutions are sending logs
 - Høgskolen i Oslo-Akerhus
 - Høgskolen i Hedemark
 - Høgskolen i Østfold
 - Høgskolen i Ålesund
 - Handelshøyskolen
 - Uninett AS
- Currently receiving 60 GB data per day
- Currently total sliding data size is 2TB

Log Analysis as Service (LAAS)

- It's becoming an eco system
 - Suricata IDS service coming up and can send logs from IDS service to LaaS
 - Edudbg service used LaaS as backend to provide debugging information to all eduroam users in Norway.
- Institutions can share data with each other without giving access to servers itself
- Can share data with external consultants as well
- Collaborate on making parsers which can help saving resources and share benefit
- Collaborate on dashboards as well for similar log services



