Log all the things!

Honza Král
@honzakral
Logs?
Events!

Log lines

Twitter feed

Invoices

Metrics
Why?
What happened last Tuesday?
Multiple machines
Multiple logs
Analysis/Discovery
Time period
Time? Time?! Time!

apache

unix timestamp

log4j

postfix.log

ISO 8601
Correlate events

Web Server logs VS Load Balancer
  see immediately that caching is off
  static files leaking to gunicorn

Web Server VS Database

500s VS Deploys
  new version has a bug

Traffic VS Ad Campaigns
Central storage
   Even for data from different systems

Enriched data
   IP -> location, hostname
   URL -> author, product, category

Search
   user:honza status:404

Analysis
   Visualisations for easy pattern discovery
Centralised Logging
Steps

Collect data
Parse data
Enrich data
Store data
Search and aggregate
Visualize data
Elastic Stack
Steps in Elastic Stack

Collect data
Parse data
Enrich data
Store data
Search and aggregate
Visualize data
Steps in Elastic Stack

Collect data
Parse data
Enrich data
Store data
Search and aggregate
Visualize data
metricbeat:
  modules:
    - module: redis
      metricsets: ["info"]
      hosts: ["host1"]
      period: 1s
      enabled: true
    - module: apache
      metricsets: ["info"]
      hosts: ["host1"]
      period: 30s
      enabled: true

filebeat:
  prospectors:
    - paths:
      - "logs/access.log"
    document_type: access
    multiline:
      pattern: ^#
      negate: true
      match: after

protocols:
  http:
    ports: [80, 8000]
  mysql:
    ports: [3306]
  redis:
    ports: [6379]
  psql:
    ports: [5432]
  thrift:
    ports: [9090]

output:
  logstash:
    hosts: ["localhost:5044"]
<table>
<thead>
<tr>
<th><strong>Inputs</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Monitoring</strong></td>
<td>collectd, graphite, ganglia, snmptrap, zenoss</td>
</tr>
<tr>
<td><strong>Datastores</strong></td>
<td>elasticsearch, redis, sqlite, s3</td>
</tr>
<tr>
<td><strong>Queues</strong></td>
<td>kafka, rabbitmq, zeromq</td>
</tr>
<tr>
<td><strong>Logging</strong></td>
<td>beats, eventlog, gelf, log4j, relp, syslog, varnish log</td>
</tr>
<tr>
<td><strong>Platforms</strong></td>
<td>drupal_dblog, gemfire, heroku, sqs, s3, twitter</td>
</tr>
<tr>
<td><strong>Local</strong></td>
<td>exec, generator, file, stdin, pipe, unix</td>
</tr>
<tr>
<td><strong>Protocol</strong></td>
<td>imap, irc, stomp, tcp, udp, websocket, wmi, xmpp</td>
</tr>
</tbody>
</table>
aggregate alter anonymize collate csv
cidr clone cipher checksum
date
dns
drop elasticsearch extractnumbers
environment elapsed fingerprint
grok geoip
i18n json json_encode kv mutate
metrics multiline metaevent prune punct
ruby range syslog_pri sleep split throttle
translate uuid urldencode xml
zeromq ...
### Outputs

<table>
<thead>
<tr>
<th>Store</th>
<th>elasticsearch, gemfire, mongodb, redis, riak, rabbitmq, solr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitoring</td>
<td>ganglia, graphite, graphtastic, nagios, opentsdb, statsd, zabbix</td>
</tr>
<tr>
<td>Notification</td>
<td>email, hipchat, irc, pagerduty, sns</td>
</tr>
<tr>
<td>Protocol</td>
<td>gelf, http, lumberjack, metriccatcher, stomp, tcp, udp, websocket, xmpp</td>
</tr>
<tr>
<td>External service</td>
<td>google big query, google cloud storage, jira, loggly, riemann, s3, sqs, syslog, datadog</td>
</tr>
<tr>
<td>External monitoring</td>
<td>boundary, circonus, cloudwatch, librato</td>
</tr>
<tr>
<td>Local</td>
<td>csv, dots, exec, file, pipe, stdout, null</td>
</tr>
</tbody>
</table>
Distributed Search Engine

Open Source

Document-based

Based on Lucene

JSON over HTTP
Data Management

Cluster
Collection of Nodes

Index
Collection of Shards

Shard
Unit of scale
Distributed across cluster
Primary and replica
### Time based data flow

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td>replicas to speed up search on stronger boxes</td>
</tr>
<tr>
<td>Week old</td>
<td>snapshot keep only 1 replica</td>
</tr>
<tr>
<td>Month old</td>
<td>move to weaker boxes</td>
</tr>
<tr>
<td>2 months</td>
<td>close the indices</td>
</tr>
<tr>
<td>3 months</td>
<td>delete</td>
</tr>
</tbody>
</table>
Architecture

Collect
beats

Enrich
logstash

Store
elasticsearch

Visualize
kibana
Logging and Python
Enhance your logs

Track metrics
- execution time
- query time
- # of queries

Include metadata
- user_id
- content

Log as JSON
Add structured info

Track info through services

Log to file

Add filebeat to read the file
Thanks!

Honza Král
@honzakral