Adriano Serckumecka
aserckumecka@fc.ul.pt
Wire Data

Log Files

Metrics

Windows Events

Visualize

elasticsearch

kibana

logstash

Ship

Parse & Transform

PACKETBEAT

FILEBEAT

METRICBEAT

WINLOGBEAT
<table>
<thead>
<tr>
<th>Company</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>eBay</td>
<td>Searching across 800 million listings in subseconds</td>
</tr>
<tr>
<td>Goldman Sachs</td>
<td>Tracking and analyzing stock trades to provide better financial guidance</td>
</tr>
<tr>
<td>Verizon</td>
<td>Reducing MTTR 10x, while cutting costs and increasing customer satisfaction</td>
</tr>
<tr>
<td>Symantec</td>
<td>Successfully switched from Solr to Elasticsearch with Elastic Support</td>
</tr>
<tr>
<td>Netflix</td>
<td>Ensuring message delivery and operational excellence</td>
</tr>
<tr>
<td>USAA</td>
<td>Securing USAA's entire internal network and application portfolio</td>
</tr>
<tr>
<td>Cisco</td>
<td>Leveraging data to detect and defeat hackers</td>
</tr>
<tr>
<td>The New York Times</td>
<td>Providing search for all 164 years of published articles</td>
</tr>
<tr>
<td>Facebook</td>
<td>Delivering a better help experience for over a billion users</td>
</tr>
<tr>
<td>FICO</td>
<td>Helping determine and protect your credit score</td>
</tr>
<tr>
<td>Salesforce</td>
<td>Empowering businesses with log analysis for usage trends</td>
</tr>
<tr>
<td>Thomson Reuters</td>
<td>Driving better research, analysis, and journalism</td>
</tr>
<tr>
<td>Company</td>
<td>Use Case</td>
</tr>
<tr>
<td>------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>CERN</td>
<td>Making history-changing discovery possible</td>
</tr>
<tr>
<td>Microsoft</td>
<td>Providing search on Azure and powering Social Dynamics</td>
</tr>
<tr>
<td>Kaidee</td>
<td>Scaling search and simplifying access across an online marketplace</td>
</tr>
<tr>
<td>Mozilla</td>
<td>Powering MozDef's efforts to secure your systems</td>
</tr>
<tr>
<td>Influence Health</td>
<td>Helping 1,100+ Hospitals Provide Better Healthcare</td>
</tr>
<tr>
<td>The Guardian</td>
<td>Revitalizing the newspaper industry with real-time readership data</td>
</tr>
<tr>
<td>VANDIS</td>
<td>Keeping its customers’ infrastructures secure with X-Pack</td>
</tr>
<tr>
<td>Jobrapido</td>
<td>Finding the right jobs for 60 million users across 58 countries</td>
</tr>
<tr>
<td>Adobe</td>
<td>Delivering a better user interface experience</td>
</tr>
<tr>
<td>SwissLife</td>
<td>Creating a 360 Customer Vision powered by Elasticsearch</td>
</tr>
<tr>
<td>Hotel Tonight</td>
<td>Finding you the right hotel, in the right place, and in the right moment</td>
</tr>
<tr>
<td>Greenman Gaming</td>
<td>Leveling up the video game industry</td>
</tr>
</tbody>
</table>
### Where it is used?

<table>
<thead>
<tr>
<th>Company</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPAREGROUP EU</td>
<td>Creating a memorable shopping experience</td>
</tr>
<tr>
<td>The Home Depot</td>
<td>Finding the right tool for your job</td>
</tr>
<tr>
<td>AUTO SCOUT 24</td>
<td>Helping millions find their dream car</td>
</tr>
<tr>
<td>ACTIVISION</td>
<td>Leveraging social data to improve video game customer support</td>
</tr>
<tr>
<td>DigitalOcean</td>
<td>Powering operations across the entire DigitalOcean environment</td>
</tr>
<tr>
<td>IBM</td>
<td>Providing the operational log analysis engine for Bluemix Apps</td>
</tr>
<tr>
<td>Tango</td>
<td>Connecting 250 million people around the world</td>
</tr>
<tr>
<td>docker</td>
<td>Making it easy to find the right container for running distributed applications</td>
</tr>
<tr>
<td>zendesk</td>
<td>Streamlining customer care by delivering a better search experience</td>
</tr>
<tr>
<td>orange</td>
<td>Delivering a better site search experience</td>
</tr>
<tr>
<td>MAYO CLINIC</td>
<td>Improving patient care by streamlining the analysis of medical information</td>
</tr>
<tr>
<td>FireEye</td>
<td>Protecting your most sensitive network information</td>
</tr>
</tbody>
</table>
Packetbeat

Packetbeat is
• a real-time network packet analyzer,
• sniffs the traffic between the servers,
• parses the application-level protocols on the fly, and
• correlates the messages into transactions.
Packetbeat

- ICMP (v4 and v6)
- DNS
- HTTP
- AMQP 0.9.1
- Cassandra
- Mysql
- PostgreSQL
- Redis
- Thrift-RPC
- MongoDB
- Memcache
Filebeat

Filebeat is a log data shipper. Installed as an agent on servers, Filebeat monitors the log directories or specific log files, tails the files, and forwards them either to Elasticsearch or Logstash for indexing.
Filebeat

Prospector 1: 
/var/log/*.log

Prospector 2: 
/var/log/apache2/*

Spooler

Elasticsearch

Logstash

Kafka

Redis

Filebeat

Wire Data

Log Files

Metrics

Windows Events

Ship
Metricbeat

Metricbeat is a lightweight shipper installed on servers to periodically collect metrics from the operating system and from services running on the server.
Metricbeat

- Apache
- HAPerxy
- MongoDB
- MySQL
- Nginx
- PostgreSQL
- Redis
- System
- Zookeeper
Metricbeat

Winlogbeat ships Windows event logs to Elasticsearch or Logstash. You can install it as a Windows service on Windows XP or later.
Metricbeat

It can capture events such as:

• application events
• hardware events
• security events
• system events
Another?

You can create a customized beat.
### Community Beats

The open source community has been hard at work developing new Beats. You can check out some of them here.

<table>
<thead>
<tr>
<th>Beat</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>amazonbeat</td>
<td>Reads data from a specified Amazon product.</td>
</tr>
<tr>
<td>apachebeat</td>
<td>Reads status from Apache HTTPD server-status.</td>
</tr>
<tr>
<td>burrowbeat</td>
<td>Monitors Kafka consumer lag using Burrow.</td>
</tr>
<tr>
<td>cassandrabeat</td>
<td>Uses Cassandra's nodetool cffstats utility to monitor Cassandra database nodes and lag.</td>
</tr>
<tr>
<td>cloudflarebeat</td>
<td>Indexes log entries from the Cloudflare Enterprise Log Share API.</td>
</tr>
<tr>
<td>cloudtrailbeat</td>
<td>Reads events from Amazon Web Services' CloudTrail.</td>
</tr>
<tr>
<td>consulbeat</td>
<td>Reads services health checks from consul and pushes them to elastic.</td>
</tr>
<tr>
<td>dockbeat</td>
<td>Reads Docker container statistics and indexes them in Elasticsearch.</td>
</tr>
<tr>
<td>elasticbeat</td>
<td>Reads status from an Elasticsearch cluster and indexes them in Elasticsearch.</td>
</tr>
<tr>
<td>execbeat</td>
<td>Periodically executes shell commands and sends the standard output and standard error to Logstash or Elasticsearch.</td>
</tr>
<tr>
<td>factbeat</td>
<td>Collects facts from Facter.</td>
</tr>
<tr>
<td>flowbeat</td>
<td>Collects, parses, and indexes sflow samples.</td>
</tr>
<tr>
<td>hsbeat</td>
<td>Reads all performance counters in Java HotSpot VM.</td>
</tr>
</tbody>
</table>
Logstash

Logstash is an open source data collection engine with real-time pipelining capabilities. Logstash can dynamically unify data from disparate sources and normalize the data into destinations of your choice.
Logstash

Data Source → Logstash Instance → Data Destination

- Input plugin
- Filter plugin
- Output plugin
Logstash

- Beats
- File
- Kafka
- Rackspace
- RabbitMQ
- Syslog
- Twitter
- ...

Diagram:

Data Source → Logstash Instance → Data Destination
- Input plugin
- Filter plugin
- Output plugin
Logstash

- Alter
- Anonymize
- GeoIP
- Grok
- Mutate
- Prune
- Useragent
- ...

Data Source → Logstash Instance → Data Destination

- Input plugin
- Filter plugin
- Output plugin
Logstash

Data Source → Logstash Instance → Data Destination

- Input plugin
- Filter plugin
- Output plugin

File
Kafka
Nagios
S3
TCP
UDP
Zabbix
...

23
Wire Data
- PACKETBEAT

Log Files
- FILEBEAT

Metrics
- METRICBEAT

Windows Events
- WINLOGBEAT

Ship

Parse & Transform

logstash

Visualize

kibana
Elasticsearch

Elasticsearch is a highly scalable open-source full-text search and analytics engine. It allows us to store, search, and analyze big volumes of data quickly and in near real time.

It works on Apache Lucene.
Lucene

• Developed in 1998
• Lucene is a search engine
  • Build index on a given document content
  • Answer to queries through its built index
• It offers
  • Scalability
  • Relevance ranking
  • Integration between different data sources
• Freely available
• Open Source
  • Apache Lucene
Lucene

• Features
  • Rich query syntax
  • Ranked search
    • Sorts by relevance or other customizable rule
  • Flexible queries
    • Phrases, text, numeric, ...
  • Field specific queries
    • Name, price, location, ....
  • Fast indexing
  • Fast searching
  • Apache License 2.0
Lucene

• Which kind of searches can Lucene perform?
  • If the field is indexed: any

• By default
  • Term
    • Docs which contains word “world”
  • Phrase
    • Docs which contains phrase “hello world”
  • Prefix
    • Docs which contains words begining with "wo"
  • Boolean
    • SHOULD, MUST and MUST NOT
  • Wildcard
    • Docs which contains word “w***d”
  • Fuzzy
    • Docs which contains word “world” within a distance 2, e.g. Wor, Word, ...
  • Range
  • MatchAll
  • ...
Elasticsearch – Inverted Index

**Stopword list**
- a
- and
- around
- every
- for
- from
- is
- it
- not
- on
- one
- the
- to
- under

**Inverted index**

<table>
<thead>
<tr>
<th>ID</th>
<th>Term</th>
<th>Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>best</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>blue</td>
<td>1, 3</td>
</tr>
<tr>
<td>3</td>
<td>bright</td>
<td>1, 3</td>
</tr>
<tr>
<td>4</td>
<td>butterfly</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>breeze</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>forget</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>great</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>hangs</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>need</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>retire</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>search</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>sky</td>
<td>2, 3</td>
</tr>
<tr>
<td>13</td>
<td>wind</td>
<td>2</td>
</tr>
</tbody>
</table>

**Documents**

- **Document 1**: The bright blue butterfly hangs on the breeze.
- **Document 2**: It’s best to forget the great sky and to retire from every wind.
- **Document 3**: Under blue sky, in bright sunlight, one need not search around.
Elasticsearch

Inverted Index

Query

Q: “blue sky”

ID | Term | Document : position
---|------|---------------------
1  | best | 2 : 3
2  | blue | 1 : 3, 3 : 2
3  | bright | 1 : 2, 3 : 5
4  | butterfly | 1 : 4
5  | breeze | 1 : 8
6  | forget | 2 : 5
7  | great | 2 : 7
8  | hangs | 1 : 5
9  | needs | 3 : 8
10 | retire | 2 : 11
11 | search | 3 : 10
12 | sky | 2 : 8, 3 : 3
13 | wind | 2 : 14

Match on sequential terms

blue - 3 : 2
sky - 3 : 3

Search object

<table>
<thead>
<tr>
<th>Document reference</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>100%</td>
</tr>
</tbody>
</table>
Elasticsearch

• Architecture
  • Master Node
    • Controls cluster-wide operations
    • Responsible to create/delete indexes
    • Assign shards to nodes
    • Keep track of nodes states
      • Forwards them to all nodes
Elasticsearch

• Architecture
  • Data Node
    • Holds data and inverted index
  • Ingest Node
    • Pre-process documents
  • Coordinating Node
    • Responsible to execute/forward queries
Elasticsearch
Elasticsearch
Elasticsearch
Elasticsearch
Elasticsearch
Elasticsearch - Interfaces

Beats
Logstash
Restful API (JSON over HTTP)
Clients (Java, C++, Python, Go, etc.)
Kibana
Elasticsearch Query

This example returns the account numbered 20:

```bash
curl -XPOST 'localhost:9200/bank/_search?pretty' -d ' {
    "query": { "match": { "account_number": 20 } }
}'
```

This example returns all accounts containing the term "mill" in the address:

```bash
curl -XPOST 'localhost:9200/bank/_search?pretty' -d ' {
    "query": { "match": { "address": "mill" } }
}'
```

This example returns all accounts containing the term "mill" or "lane" in the address:

```bash
curl -XPOST 'localhost:9200/bank/_search?pretty' -d ' {
    "query": { "match": { "address": "mill lane" } }
}'
```

This example is a variant of match (match_phrase) that returns all accounts containing the phrase "mill lane" in the address:

```bash
curl -XPOST 'localhost:9200/bank/_search?pretty' -d ' {
    "query": { "match_phrase": { "address": "mill lane" } }
}'
```
Elasticsearch - Query

GET /twitter/_search?q=user:kimchy
Elasticsearch - Query

GET /twitter/_search?q=user:kimchy
Elasticsearch for Hadoop
Kibana

Kibana is an open source analytics and visualization platform designed to work with Elasticsearch. You can perform data analysis and visualize your data in a variety of charts, tables, and maps.
ES at MEO

- 3000+ sources
- 250k+ queries daily
- 1.8B documents
- ~1.5B documents presented daily
- 475 indexes
- ~750Gb for data per node

**Commodity hardware (no SSD)**
- Data nodes: 2x 16Gb Quad Xeon
- Master nodes: 2x 16Gb Dual Core x64
- Query nodes: 2x 8Gb Quad i5
- Logstash: 2 inputs; DNS round-robin
ES at MEO

• Currently working
  • ML
    • Near automatic root cause analysis
    • Open Spark access to work data
    • Recomendation engine
  • Event producers
    • Nagios
    • Applications (log entries)
    • Agent to report metrics
    • ...

Kibana Demo

http://demo.elastic.co/packetbeat
Questions?

Thanks!

aserckumecka@lasige.di.fc.ul.pt