# 第四部分： 利用 ELK 做大数据分析

极客时间

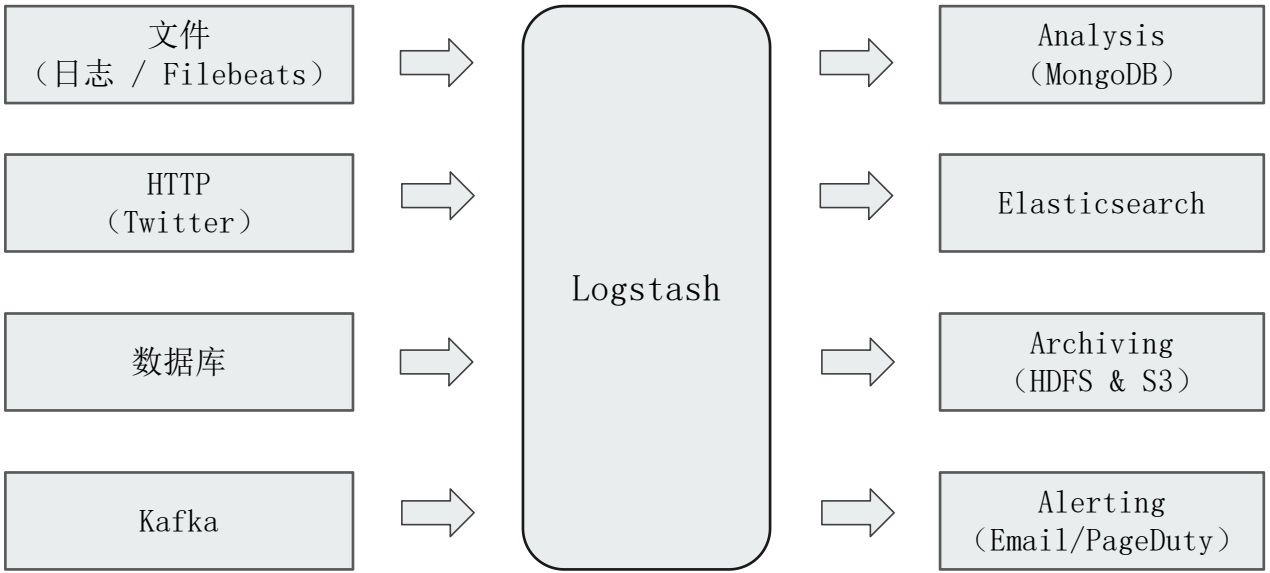
扫码试看/订阅极客时间

《Elasticsearch核心技术与实战》视频课程

# Logstash 入门及架构介绍

# Logstash

- ELT 工具 / 数据搜集处理引擎。支持 200 多个插件

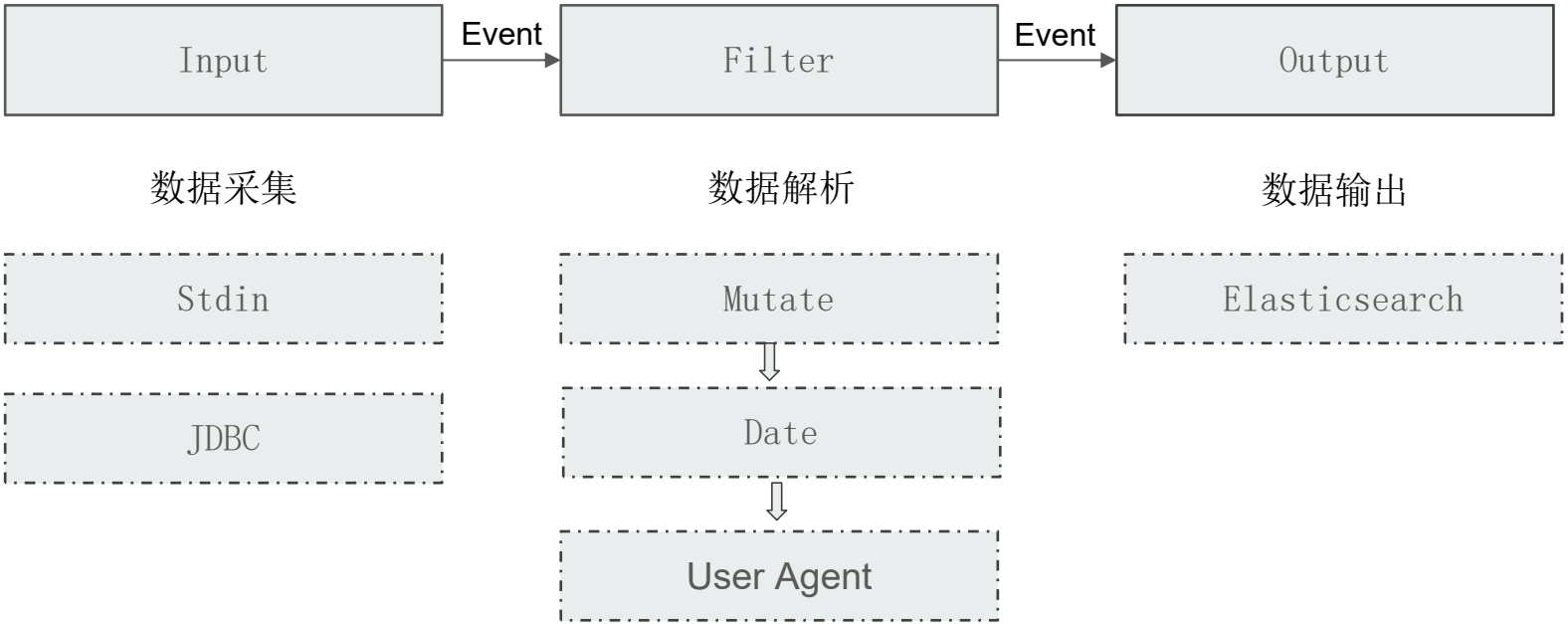


# Logstash Concepts

- Pipeline
  - 包含了 input-filter-output 三个阶段的处理流程
  - 插件生命周期管理
  - 队列管理
- Logstash Event
  - 数据在内部流转时的具体表现形式。数据在 input 阶段被转换为 Event，在 output 被转化成目标格式数据
  - Event 其实是一个 Java Object，在配置文件中，对 Event 的属性进行增删改查

# Logstash 架构简介

Codec（Code / Decode）：将原始数据 decode 成 Event；将 Event encode 成目标数据



# Logstash 配置文件结构

```
input { stdin { } }

filter {
  grok {
    match => { "message" => "%{COMBINEDAPACHELOG}" }
  }
  date {
    match => [ "timestamp" , "dd/MMM/yyyy:HH:mm:ss Z" ]
  }
}

output {
  elasticsearch { hosts => ["localhost:9200"] }
  stdout { codec => rubydebug }
}
```

- Bin/logstash -f demo.conf
- Pipeline
  - Input / Filter / Output
- Codec
  - Line / json

# Input Plugins

- 一个 Pipeline 可以有多个 input 插件
  - Stdin / File
  - Beats / Log4J / Elasticsearch / JDBC / Kafka / Rabbitmq / Redis
  - JMX / HTTP / Websocket / UDP / TCP
  - Google Cloud Storage / S3
  - Github / Twitter

# Output Plugins

- 将 Event 发送到特定的目的地，是 Pipeline 的最后一个阶段
- 常见 Output Plugins (https://www.elastic.co/guide/en/logstash/7.1/output-plugins.html)
  - Elasticsearch
  - Email / Pageduty
  - Influxdb / Kafka / Mongodb / Opentsdb / Zabbix
  - Http / TCP / Websocket

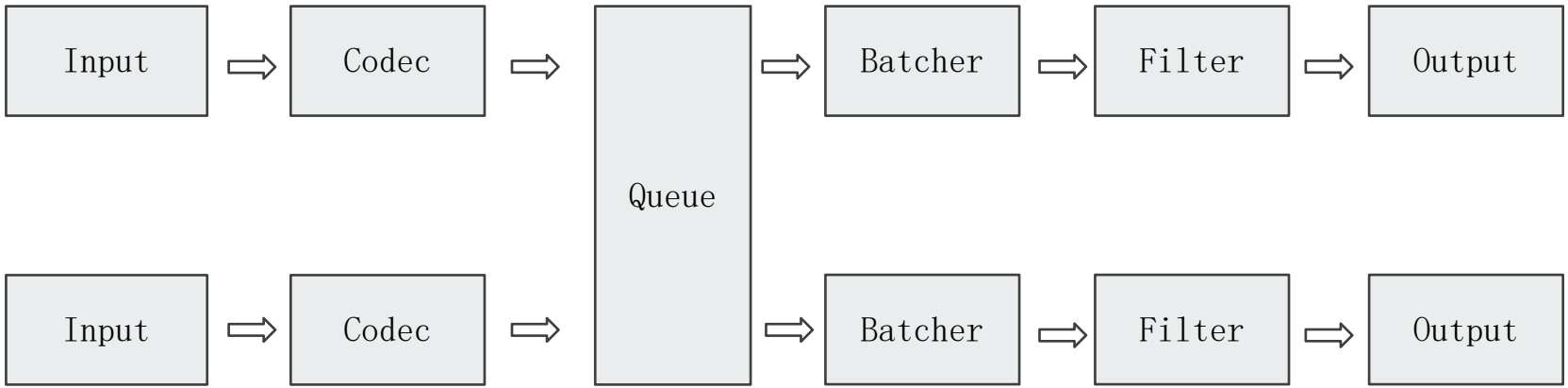
# Codec Plugins

- 将原始数据 decode 成 Event；将 Event encode 成目标数据
- 内置的 Codec Plugins (https://www.elastic.co/guide/en/logstash/7.1/codec-plugins.html)
  - Line / Multiline
  - JSON / Avro / Cef (ArcSight Common Event Format)
  - Dots / Rubydebug

# Filter Plugins

- 处理 Event
- 内置的 Filter Plugins (https://www.elastic.co/guide/en/logstash/7.1/filter-plugins.html)
  - Mutate - 操作 Event 的字段
  - Metrics - Aggregate metrics
  - Ruby - 执行 Ruby 代码

# Queue


Input
Codec
Queue
Batcher
Filter
Output
Input
Codec
Batcher
Filter
Output

# 多 Pipelines 实例

```
- pipeline.id: my-pipeline_1
  path.config: "/etc/path/to/p1.config"
  pipeline.workers: 3
- pipeline.id: my-other-pipeline
  path.config: "/etc/different/path/p2.cfg"
  queue.type: persisted
```

- Pipeline.works：Pipeline 线程数，默认是 CPU 核数
- Pipeline.batch.size：Batcher 一次批量获取等待处理的文档数，默认 125。需结合 jvm.options 调节
- Pipeline.batch.delay：Batcher 等待时间

# Logstash Queue

- In Memory Queue
  - 进程 Crash，机器当机，都会引起数据的丢失
- Persistent Queue
  - Queue.type.persisted (默认是 memory)
    - Queue.max_bytes: 4gb
  - 机器当机，数据也不会丢失；数据保证会被消费；可以替代 Kafka 等消息队列缓冲区的作用
  - https://www.elastic.co/guide/en/logstash/7.1/persistent-queues.html

# Codec Plugin - Single Line

```
sudo bin/logstash -e "input{stdin{codec=>line}}output{stdout{codec=> rubydebug}}"
sudo bin/logstash -e "input{stdin{codec=>json}}output{stdout{codec=> rubydebug}}"
sudo bin/logstash -e "input{stdin{codec=>line}}output{stdout{codec=> dots}}"
```

# Codec Plugin - Multiline

- 设置参数
  - Pattern： 设置行匹配的正则表达式
  - What ：如果匹配成功，那么匹配行属于上一个事件还是下一个事件
    - Previous / Next
  - Negate true / false：是否对 pattern 结果取反
    - True / False

# Codec Plugin - Multiline(异常日志)

```
Exception in thread "main" java.lang.NullPointerException
        at com.example.myproject.Book.getTitle(Book.java:16)
        at com.example.myproject.Author.getBookTitles(Author.java:25)
        at com.example.myproject.Bootstrap.main(Bootstrap.java:14)
```

```
input {
  stdin {
    codec => multiline {
      pattern => "^\s"
      what => "previous"
    }
  }
}
```

# Input Plugin - File

- 支持从文件中读取数据，如日志文件
- 文件读取需要解决的问题
  - 只被读取一次。重启后需要从上次读取的位置继续（通过 sincedb 实现）
- 读取到文件新内容，发现新文件
- 文件发生归档操作（文档位置发生变化，日志 rotation），不能影响当前的内容读取

# Filter Plugin

- Filter Plugin 可以对 Logstash Event 进行各种处理，例如解析，删除字段，类型转换
  - Date：日期解析
  - Dissect：分割符解析
  - Grok：正则匹配解析
  - Mutate：处理字段。重命名，删除，替换
  - Ruby：利用 Ruby 代码来动态修改 Event

# Filter Plugin - Mutate

- 对字段做各种操作
  - Convert 类型转换
  - Gsub 字符串替换
  - Split / Join / Merge 字符串切割，数组合并字符串，数组合并数组
  - Rename 字段重命名
  - Update / Replace 字段内容更新替换
  - Remove_field 字段删除

# 利用 JDBC 插件导入数据到 Elasticsearch

# 同步数据库数据到 Elasticsearch

- 需求 - 将数据库中的数据同步到 ES，借助 ES 的全文搜索，提高搜索速度
  - 需要把新增用户信息同步到 Elasticsearch 中
  - 用户信息 Update 后，需要能被更新到 Elasticsearch
  - 支持增量更新
  - 用户注销后，不能被 ES 所搜索到

# JDBC Input Plugin & 设计实现思路

```
input {
  jdbc {
    jdbc_driver_class => "com.mysql.jdbc.Driver"
    jdbc_connection_string => "jdbc:mysql://localhost:3306/db_example"
    jdbc_user => root
    jdbc_password => ymruan123
    #启用追踪，如果为true，则需要指定tracking_column
    use_column_value => true
    #指定追踪的字段，
    tracking_column => "last_updated"
    #追踪字段的类型，目前只有数字(numeric)和时间类型(timestamp)，默认是数字类型
    tracking_column_type => "numeric"
    #记录最后一次运行的结果
    record_last_run => true
    #上面运行结果的保存位置
    last_run_metadata_path => "jdbc-position.txt"
    statement => "SELECT * FROM user where last_updated >:sql_last_value;"
    schedule => " * * * * * *"
  }
}
```

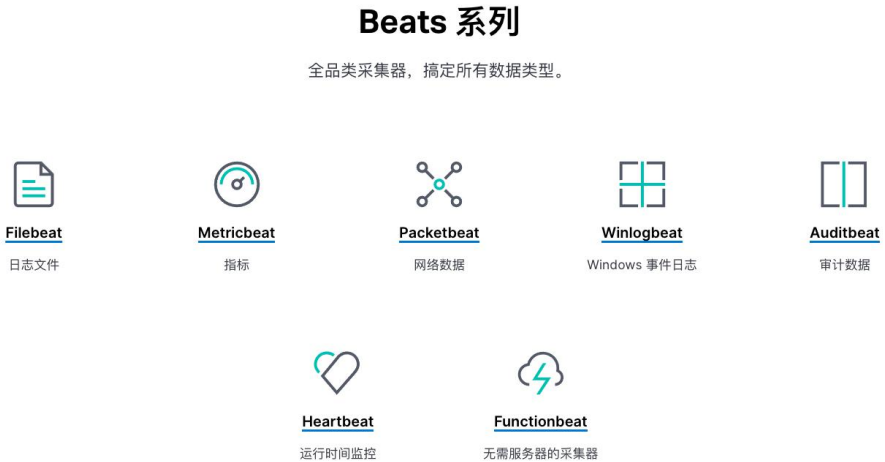
- 支持通过 JDBC Input Plugin 将数据从数据库从读到 Logstash
  - 需要自己提供所需的 JDBC Driver
- Scheduling
  - 语法来自 Rufus-scheduler
  - 扩展了 Cron，支持时区
- State
  - Tracking_column / sql_last_value

# Demo

- https://spring.io/guides/gs/accessing-data-mysql/
- 支持 新增 / 更新 / 删除 三种 API
- User 字段包含了一个 last update 的字段
- User 表包含了一个 is deleted 字段

# Beats 介绍

# 什么是 Beats




- Light weight data shippers
  - 以搜集数据为主
  - 支持与 Logstash 或 ES 集成
- 全品类 / 轻量级 / 开箱即用 / 可插拔 / 可扩展 / 可视化

# Metricbeat 简介

- 用来定期搜集操作系统，软件的指标数据
  - Metric v.s Logs
    - Metric - 可聚合的数据，定期搜集
    - Logs 文本数据，随机搜集
- 指标存储在 Elasticsearch 中，可以通过 Kibana 进行实时的数据分析

# Metricbeat 组成

- Module
  - 搜集的指标对象，例如不同的操作系统，不同的数据库，不同的应用系统
- Metricset
  - 一个 Module可以有多个 metricset
  - 具体的指标集合。以减少调用次数为原则进行划分
    - 不同的 metricset 可以设置不同的抓取时长

# Module

- Metricbeat 提供了大量的开箱即用的 Module
  - https://www.elastic.co/guide/en/beats/metricbeat/7.1/index.html
- 通过执行 metricbeat module list 查看
- 通过执行 metricbeat moudle enable module_name 定制

# Metricsets

- 每个 Module 都有自己的 metricsets，以 System Module 为例
  - core
  - CPU
  - disk IO
  - filesystem
  - load
  - memory

# Metricbeat Event

```
{
  "@timestamp": "2016-06-22T22:05:53.291Z",
  "agent": {
    "type": "metricbeat"
  },
  "host": {
     "hostname": "host.example.com",
   },
  "event": {
    "dataset": "system.process",
    "module": process
  },
  .
  .
  .

  "type": "metricsets"
}
```

# Metricbeat Demo

- 下载安装并配置 Metricbeat
  - Enable 查看 system & Mysql
  - 定义 interval
- 配置 Kibana Dashboard
- 运行
- 查看 索引 / Dashboard

# Packetbeat

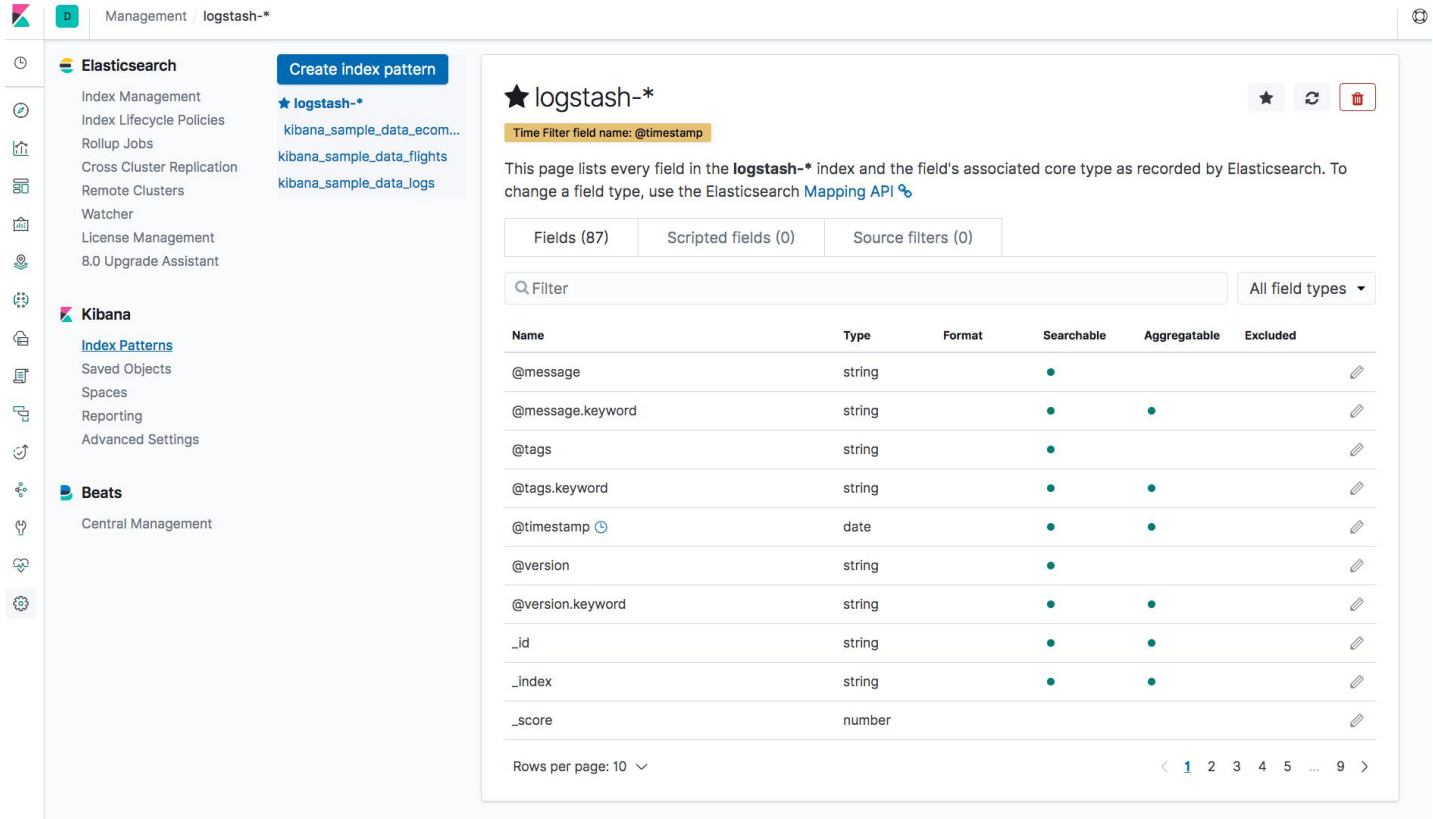
- Packetbeat - 实时网络数据分析，监控应用服务器之间的网络流量
  - 常见抓包工具 - Tcpdump /wireshark
  - 常见抓包配置 - Pcap 基于 libpcap，跨平台 / Af_packet 仅支持 Linux，基于内存映射嗅探，高性能
- Packetbeat 支持的协议
  - ICMP / DHCP / DNS / HTTP / Cassandra / Mysql / PostgresSQL / Redis / MongoDB / Memcache / TLS
- Network flows：抓取记录网络流量数据，不涉及协议解析

# Packetbeat Demo

- https://www.elastic.co/guide/en/beats/packetbeat/7.1/packetbeat-getting-started.html
- 安装配置
- 配置 Kibana Dashboard
  - Packetbeat setup --dashboards
- 运行 Packetbeat
- 查看 Dashboard

# 使用 Index Pattern 配置数据

# Index Pattern

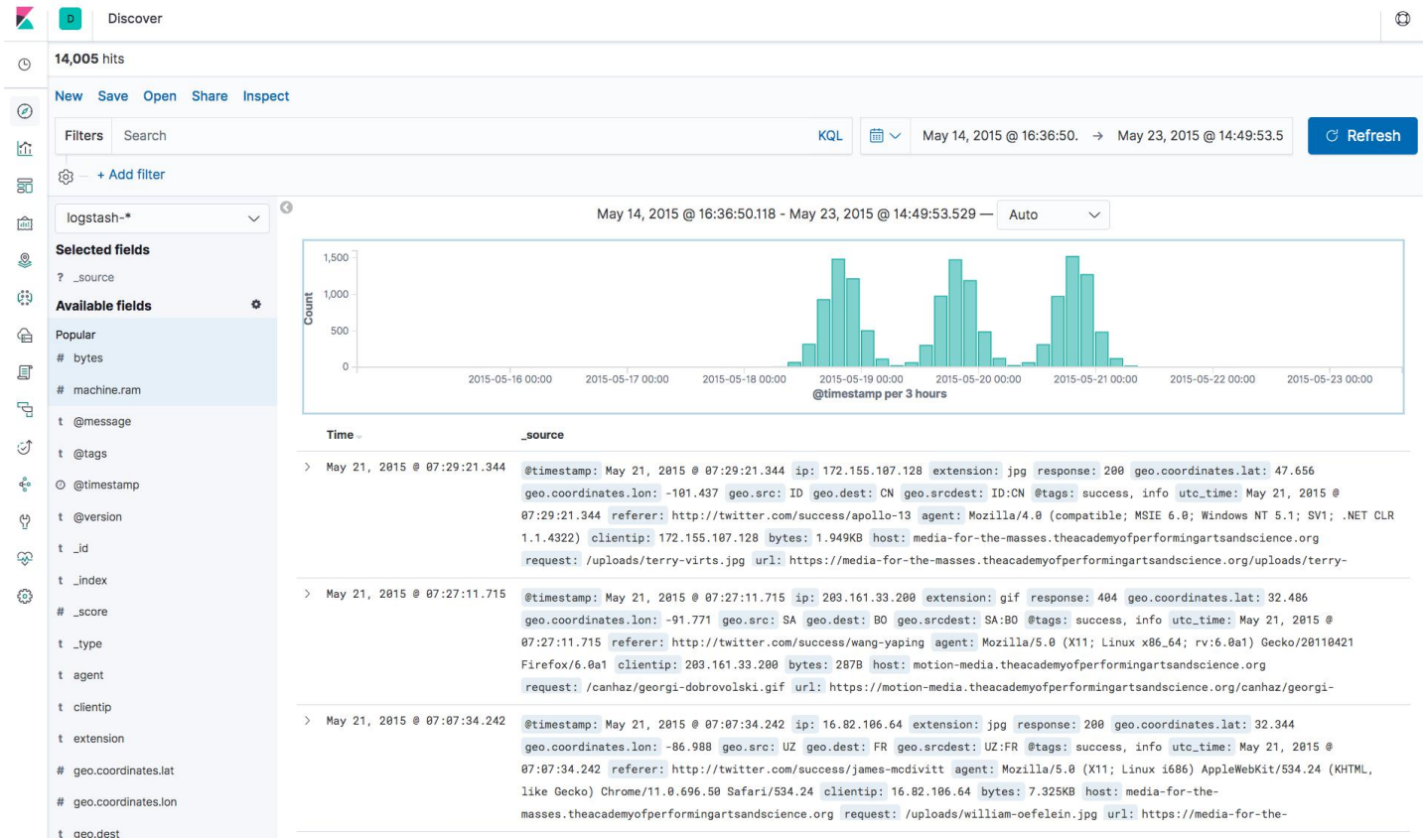


# Demo

- Management
  - Index Patterns / Saved Objects （导出备份）
- Created Index Pattern （Save As default Pattern）
  - Fields：Type，Searchable/Aggregation / Format（Number / Image）
  - Script Fields

# 使用 Kibana Discovery 探索数据

# Discovery


Discover
14,005 hits
New Save Open Share Inspect
Filters Search KQL May 14, 2015 @ 16:36:50. → May 23, 2015 @ 14:49:53.5 Refresh
+ Add filter
logstash-*
Selected fields
? _source
Available fields
Popular
# bytes
# machine.ram
t @message
t @tags
@timestamp
t @version
t _id
t _index
# _score
t _type
t agent
t clientip
t extension
# geo.coordinates.lat
# geo.coordinates.lon
t geo.dest
May 14, 2015 @ 16:36:50.118 - May 23, 2015 @ 14:49:53.529 — Auto
Count
1,500
1,000
500
0
2015-05-16 00:00
2015-05-17 00:00
2015-05-18 00:00
2015-05-19 00:00
2015-05-20 00:00
2015-05-21 00:00
2015-05-22 00:00
2015-05-23 00:00
@timestamp per 3 hours
Time
_source
May 21, 2015 @ 07:29:21.344
@timestamp: May 21, 2015 @ 07:29:21.344 ip: 172.155.107.128 extension: jpg response: 200 geo.coordinates.lat: 47.656 geo.coordinates.lon: -101.437 geo.src: ID geo.dest: CN geo.srcdest: ID:CN @tags: success, info utc_time: May 21, 2015 @ 07:29:21.344 referer: http://twitter.com/success/apollo-13 agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322) clientip: 172.155.107.128 bytes: 1.949KB host: media-for-the-masses.theacademyofperformingartsandscience.org request: /uploads/terry-virts.jpg url: https://media-for-the-masses.theacademyofperformingartsandscience.org/uploads/terry-
May 21, 2015 @ 07:27:11.715
@timestamp: May 21, 2015 @ 07:27:11.715 ip: 203.161.33.200 extension: gif response: 404 geo.coordinates.lat: 32.486 geo.coordinates.lon: -91.771 geo.src: SA geo.dest: BO geo.srcdest: SA:BO @tags: success, info utc_time: May 21, 2015 @ 07:27:11.715 referer: http://twitter.com/success/wang-yaping agent: Mozilla/5.0 (X11; Linux x86_64; rv:6.0a1) Gecko/20110421 Firefox/6.0a1 clientip: 203.161.33.200 bytes: 287B host: motion-media.theacademyofperformingartsandscience.org request: /canhaz/georgi-dobrovolski.gif url: https://motion-media.theacademyofperformingartsandscience.org/canhaz/georgi-
May 21, 2015 @ 07:07:34.242
@timestamp: May 21, 2015 @ 07:07:34.242 ip: 16.82.106.64 extension: jpg response: 200 geo.coordinates.lat: 32.344 geo.coordinates.lon: -86.988 geo.src: UZ geo.dest: FR geo.srcdest: UZ:FR @tags: success, info utc_time: May 21, 2015 @ 07:07:34.242 referer: http://twitter.com/success/james-mcdivitt agent: Mozilla/5.0 (X11; Linux i686) AppleWebKit/534.24 (KHTML, like Gecko) Chrome/11.0.696.50 Safari/534.24 clientip: 16.82.106.64 bytes: 7.325KB host: media-for-the-masses.theacademyofperformingartsandscience.org request: /uploads/william-oefelein.jpg url: https://media-for-the-

# Demo

- 设置时间过滤器
- 写入条件进行过滤
- 根据字段过滤
- 查看字段数据统计
- 文档上下文

# 基本可视化组件介绍

# Demo

- 账户存款
  - Pie Chart （Inspector）
- 日志相关
  - Area Chart （X 轴 Y 轴， 顺序，etc）
  - Bar Chart

# 构建 Dashboard

# Demo

- 安装配置
- 配置 Kibana Dashboard
- 运行
- 查看 Dashboard
  - Dashboard 是一组相关主题的可视化组件
  - Dashboard 的设计准则 （Resize）

# 用 Monitoring 和 Alerting 监控 Elasticsearch 集群

# X-Pack Monitoring

- X-Pack 提供了免费集群监控的功能
- 使用 Elasticsearch 监控 Elasticsearch
  - Xpack.monitoring.collection.interval 默认设置 10 秒
- 在生产环境中，建议搭建 dedicated 集群用于 ES 集群的监控。有以下几个好处
  - 减少负载和数据
  - 当被监控集群出现问题，还能看到监控相关的数据

# 配置 Monitoring

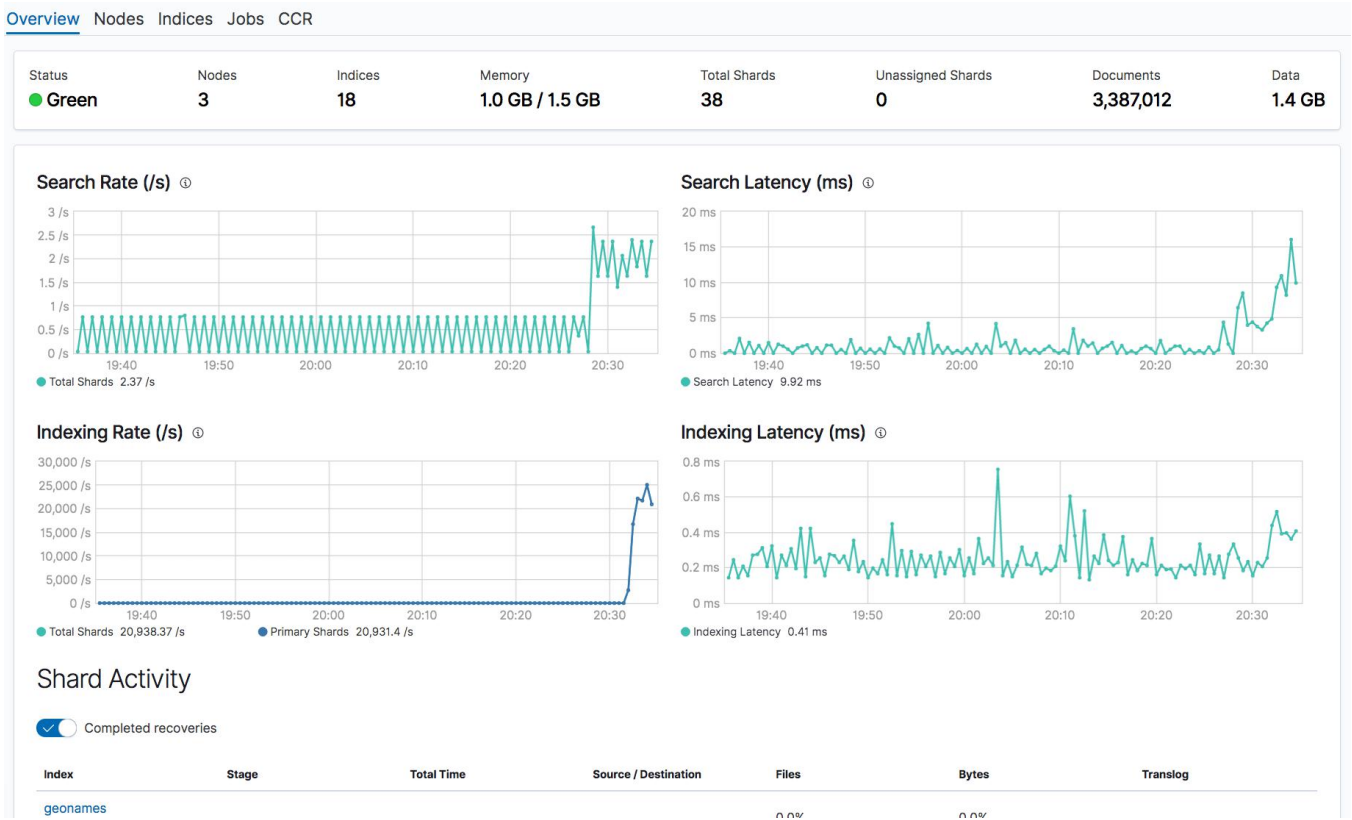
| 配置项 | 含义 |
| --- | --- |
| xpack.monitoring.collection.indices | 默认监控所有索引。支持配置索引列表（逗号间隔） |
| xpack.monitoring.collection.interval | 搜集数据的时间间隔，默认 10 秒 |
| xpack.monitoring.history.duration | 数据保留的时间，默认 7 天 |

https://www.elastic.co/guide/en/x-pack/current/monitoring-settings.html

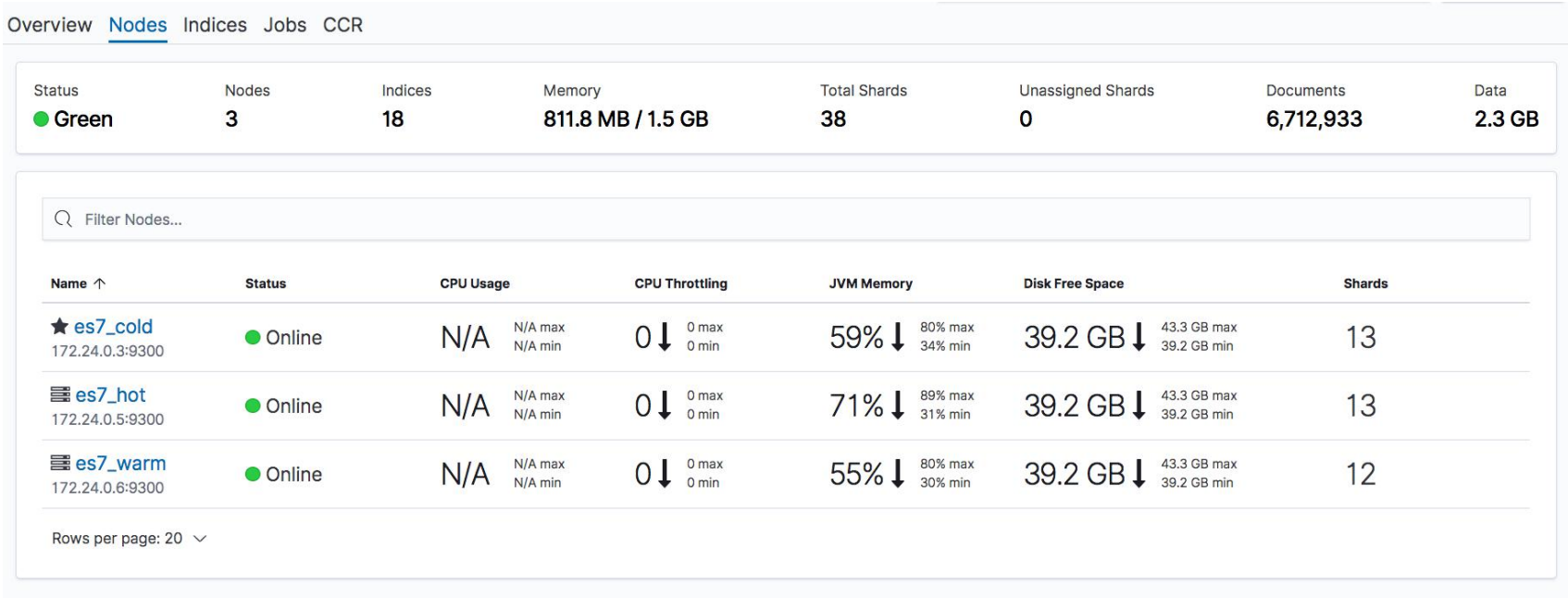
# Watcher for Alerting

- 需要 Gold 账户
- 一个 Watcher 由 5个部分组成
  - Trigger - 多久被触发一次（例如：5 分钟触发一次）
  - Input - 查询条件（在所有日志索引中查看 “ERROR” 相关）
  - Condition - 查询是否满足条件（例如：大于 1000 条返回）
  - Actions - 执行相关操作（例如：发送邮件）

# Overall


Overview
Nodes
Indices
Jobs
CCR
Status
Green
Nodes
3
Indices
18
Memory
1.0 GB / 1.5 GB
Total Shards
38
Unassigned Shards
0
Documents
3,387,012
Data
1.4 GB
Search Rate (/s)
3 /s
2.5 /s
2 /s
1.5 /s
1 /s
0.5 /s
0 /s
19:40
19:50
20:00
20:10
20:20
20:30
Total Shards 2.37 /s
Search Latency (ms)
20 ms
15 ms
10 ms
5 ms
0 ms
Search Latency 9.92 ms
Indexing Rate (/s)
30,000 /s
25,000 /s
20,000 /s
15,000 /s
10,000 /s
5,000 /s
0 /s
Total Shards 20,938.37 /s
Primary Shards 20,931.4 /s
Indexing Latency (ms)
0.8 ms
0.6 ms
0.4 ms
0.2 ms
0 ms
Indexing Latency 0.41 ms
Shard Activity
Completed recoveries
Index
Stage
Total Time
Source / Destination
Files
Bytes
Translog
geonames

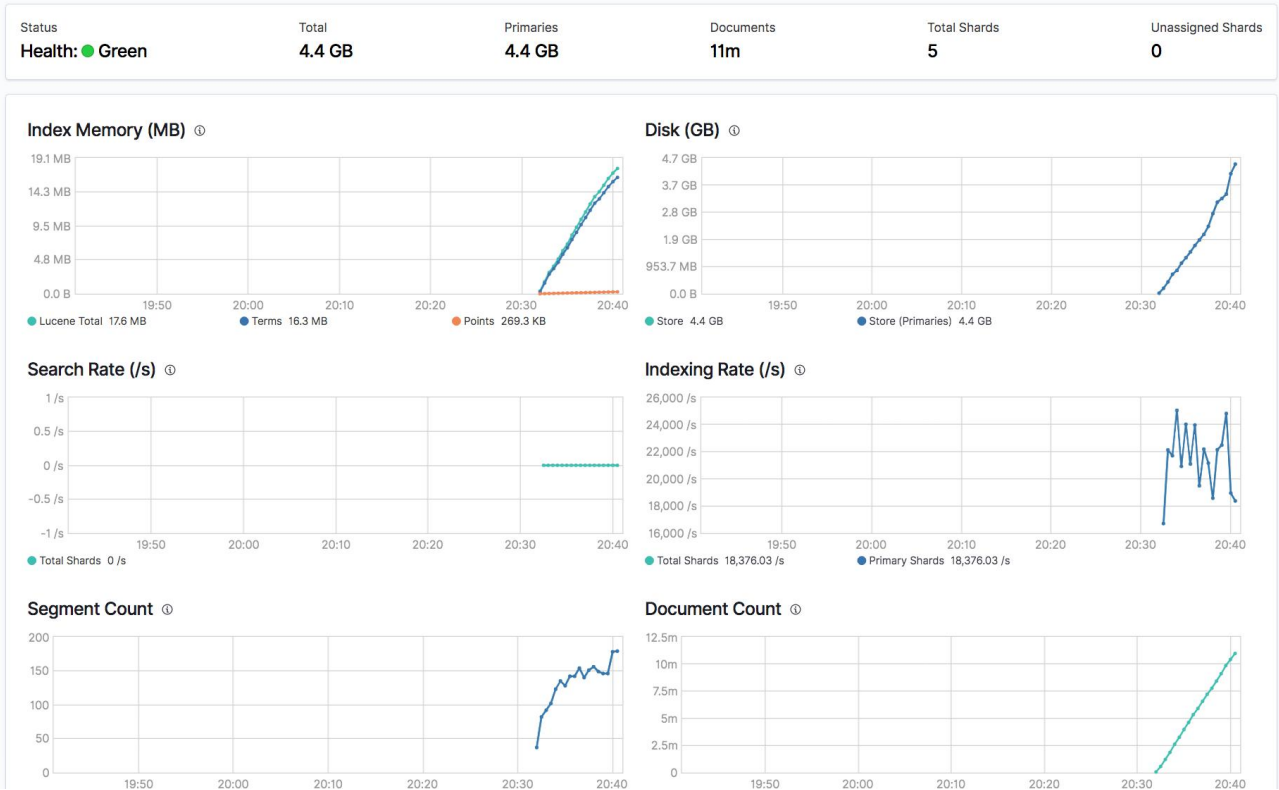
# Nodes


Overview
Nodes
Indices
Jobs
CCR
Status
Green
Nodes
3
Indices
18
Memory
811.8 MB / 1.5 GB
Total Shards
38
Unassigned Shards
0
Documents
6,712,933
Data
2.3 GB
Filter Nodes...
Name
Status
CPU Usage
CPU Throttling
JVM Memory
Disk Free Space
Shards
es7_cold
172.24.0.3:9300
Online
N/A
N/A max
N/A min
0
0 max
0 min
59%
80% max
34% min
39.2 GB
43.3 GB max
39.2 GB min
13
es7_hot
172.24.0.5:9300
Online
N/A
N/A max
N/A min
0
0 max
0 min
71%
89% max
31% min
39.2 GB
43.3 GB max
39.2 GB min
13
es7_warm
172.24.0.6:9300
Online
N/A
N/A max
N/A min
0
0 max
0 min
55%
80% max
30% min
39.2 GB
43.3 GB max
39.2 GB min
12
Rows per page: 20

# Index

| Name ↑ | Status | Document Count | Data | Index Rate | Search Rate | Unassigned Shards |
|---|---|---|---|---|---|---|
| .kibana_1 | Green | 140 | 2.1 MB | 0 /s | 0.04 /s | 0 |
| .monitoring-es-7-2019.08.26 | Green | 19.6k | 28.5 MB | 0 /s | 0.18 /s | 0 |
| .monitoring-es-7-2019.08.27 | Green | 67k | 96.9 MB | 5.75 /s | 0.18 /s | 0 |
| .monitoring-kibana-7-2019.08.26 | Green | 1.4k | 961.6 KB | 0 /s | 0.12 /s | 0 |
| .monitoring-kibana-7-2019.08.27 | Green | 3.3k | 1.7 MB | 0.1 /s | 0.12 /s | 0 |
| .triggered_watches | Green | 2 | 94.4 KB | 0.1 /s | 0 /s | 0 |
| .watcher-history-9-2019.08.27 | Green | 702 | 2.0 MB | 0.1 /s | 0 /s | 0 |
| .watches | Green | 6 | 1.6 MB | 0.1 /s | 0 /s | 0 |
| geonames | Green | 8m | 2.9 GB | 2,224.75 /s | 0 /s | 0 |
| ilm_index-000015 | Green | 4 | 3.7 KB | 0 /s | 0 /s | 0 |
| kibana_sample_data_ecommerce | Green | 4.7k | 10.3 MB | 0 /s | 0 /s | 0 |
| kibana_sample_data_flights | Green | 13.1k | 13.4 MB | 0 /s | 0 /s | 0 |
| kibana_sample_data_logs | Green | 14k | 22.5 MB | 0 /s | 0 /s | 0 |
| logstash-2015.05.18 | Green | 4.6k | 35.9 MB | 0 /s | 0 /s | 0 |
| logstash-2015.05.19 | Green | 4.6k | 37.4 MB | 0 /s | 0 /s | 0 |
| logstash-2015.05.20 | Green | 4.8k | 37.9 MB | 0 /s | 0 /s | 0 |
| my_index | Green | 0 | 566.0 B | 0 /s | 0 /s | 0 |
| test | Green | 95 | 49.6 KB | 0 /s | 0 /s | 0 |

# Index Level

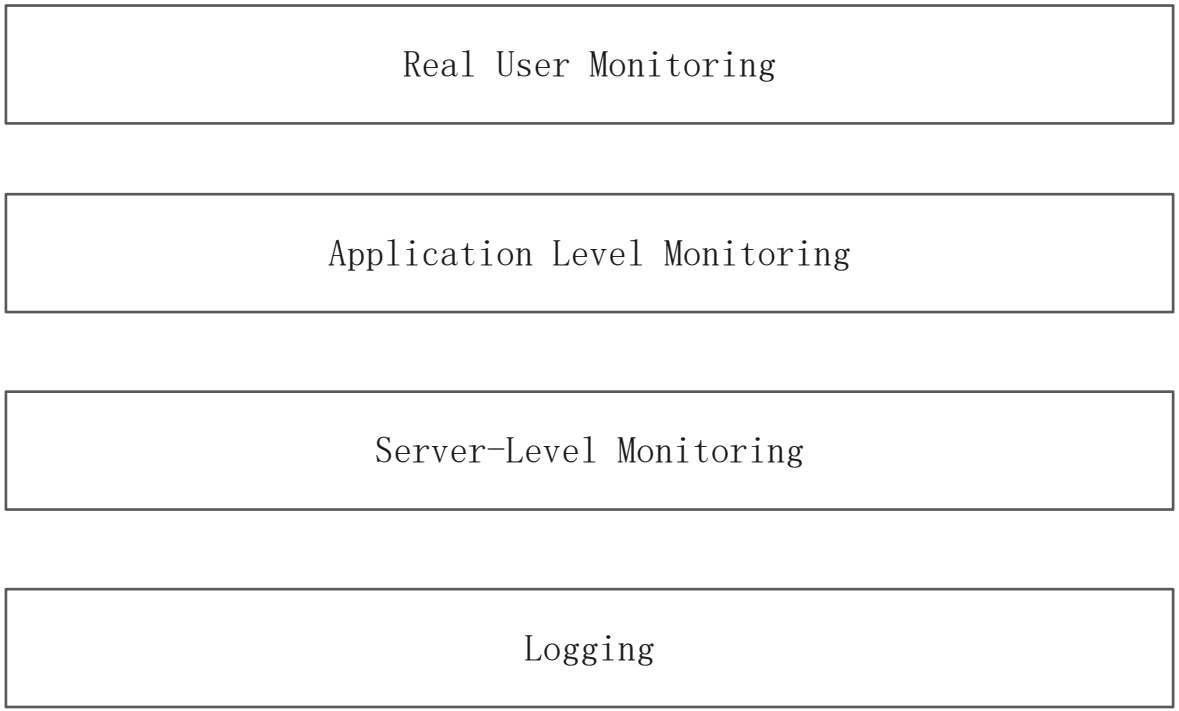

Status
Health: Green
Total
4.4 GB
Primaries
4.4 GB
Documents
11m
Total Shards
5
Unassigned Shards
0
Index Memory (MB)
19.1 MB
14.3 MB
9.5 MB
4.8 MB
0.0 B
19:50
20:00
20:10
20:20
20:30
20:40
Lucene Total 17.6 MB
Terms 16.3 MB
Points 269.3 KB
Disk (GB)
4.7 GB
3.7 GB
2.8 GB
1.9 GB
953.7 MB
0.0 B
Store 4.4 GB
Store (Primaries) 4.4 GB
Search Rate (/s)
1 /s
0.5 /s
0 /s
-0.5 /s
-1 /s
Total Shards 0 /s
Indexing Rate (/s)
26,000 /s
24,000 /s
22,000 /s
20,000 /s
18,000 /s
16,000 /s
Total Shards 18,376.03 /s
Primary Shards 18,376.03 /s
Segment Count
200
150
100
50
0
Document Count
12.5m
10m
7.5m
5m
2.5m
0

# X-Pack Monitoring Demo

- Clusters Dashboard
- Nodes Dashboard
- Individual Node dashboard
- Indices Dashboard

# 用 APM 进行程序性能监控

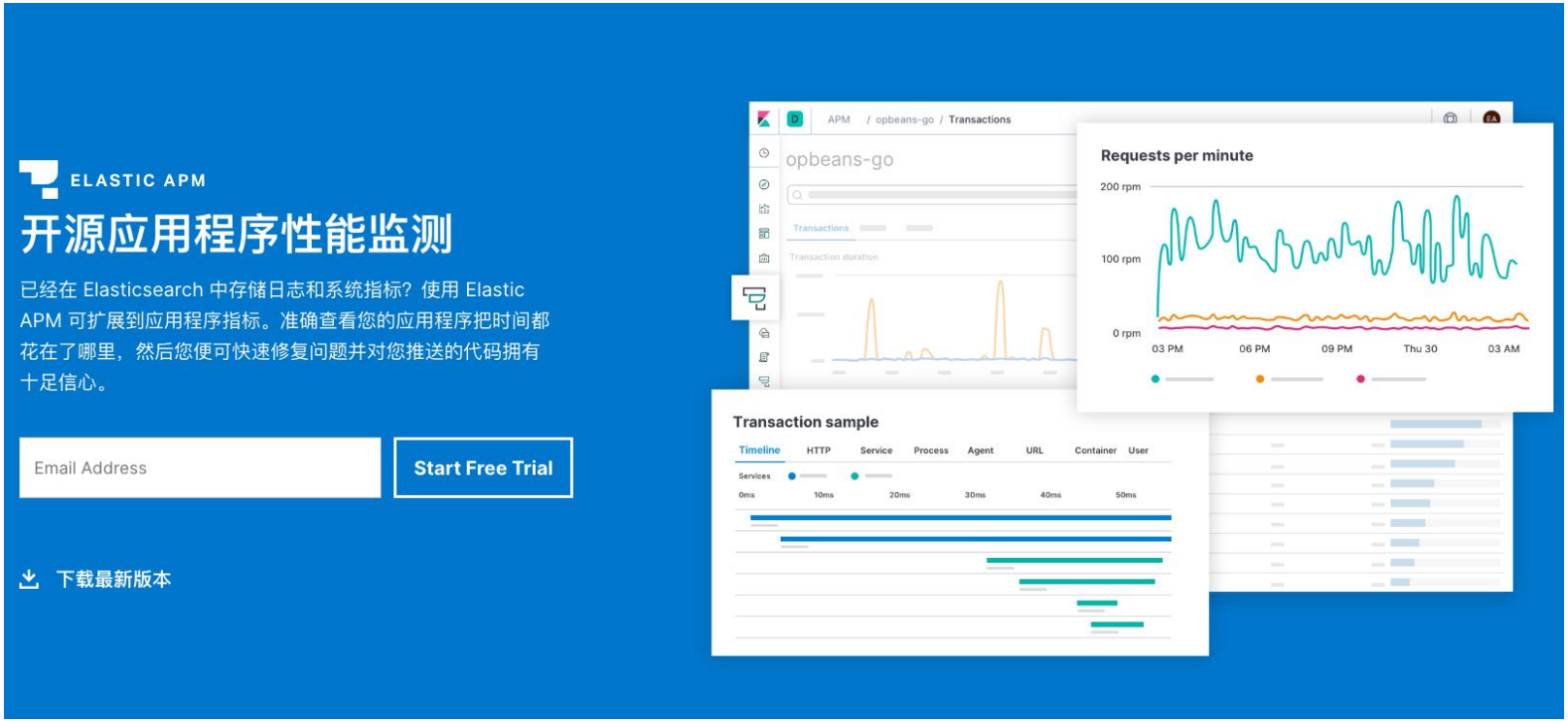
# Elastic 全栈监控



# 核心应用指标

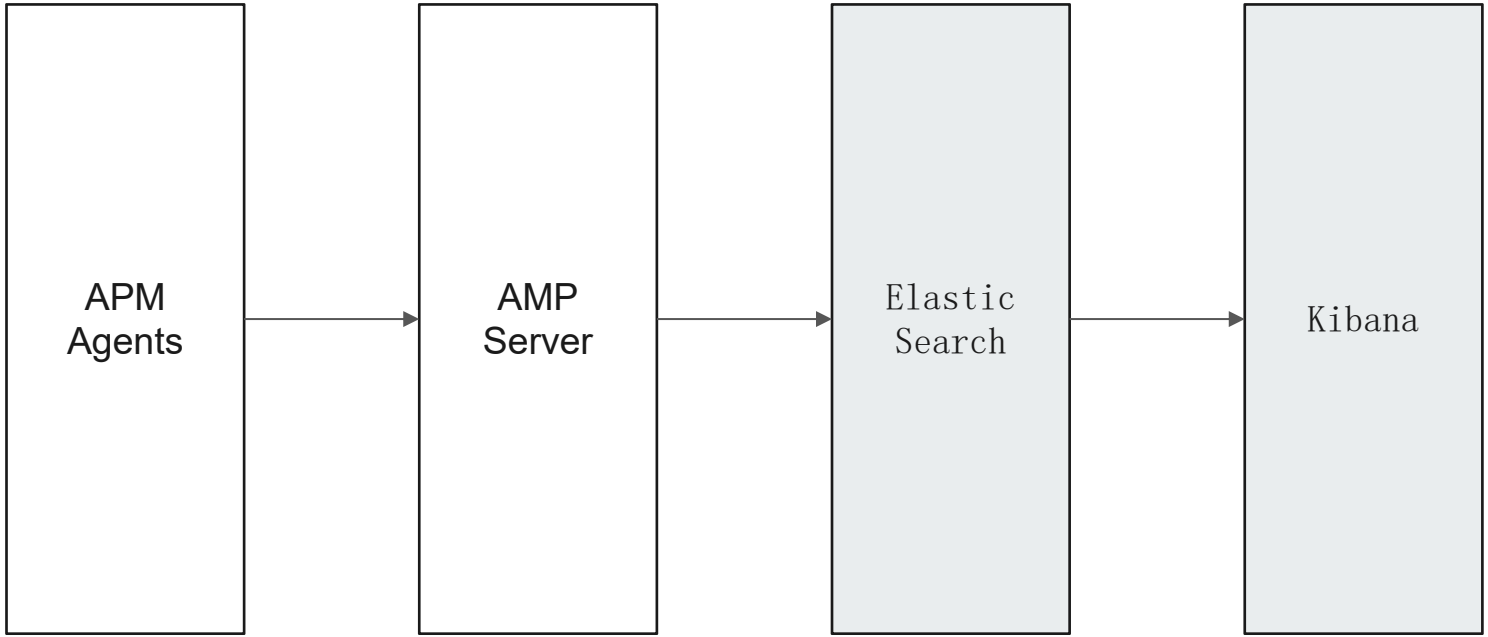
- 请求响应时间
- 未处理的错误及异常
- 可视化调用关系
- 发现性能瓶颈
- 代码下钻

# APM




https://www.elastic.co/cn/products/apm

# APM 如何整合到 Elastic Stack



# Demo

- 安装配置
  - https://www.elastic.co/cn/downloads/past-releases/apm-server-7-1-0
- 运行 Spring + MySQL 程序
- 运行性能测试脚本
- 查看结果 Dashboard

# 用机器学习实现异常检测（上）

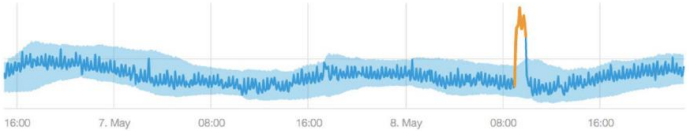
# 异常检测所解决的问题

- 解决一些基于规则或者 Dashboard 难以实时发现的问题
- IT 运维
  - 如何知道系统正常运行 / 如何调节阈值触发合适的报警 / 如何进行归因分析
- 信息安全
  - 哪些用户构成了内部威胁 / 系统是否感染了病毒
- 物联网 / 数据采集监控
  - 工厂和设备是否正常运营

# 什么是正常

- 什么是正常
  - 随着时间的推移，某个个体一直表现出一致的行为
  - 某个个体和他的同类比较，一直表现出和其他个体一致的行为

# 什么是异常




- 什么是异常
  - **和自己比** – 个体的行为发生了急剧的变化
  - **和他人比** – 个体明显区别于其他的个体

# 判定异常需要一定的指导

# 相关术语

- Elastic 平台的机器学习功能
  - Elastic 的ML，主要针对**时序数据**的**异常检测和预测**
- 非监督机器学习
  - 不需要使用人工标签的数据来学习，仅仅依靠历史数据自动学习
- 贝叶斯统计
  - 一种概率计算方法，使用先验结果来计算现值或者预测未来的数值
- 异常检测
  - 异常代表的是不同的，但未必代表的是坏的 / 定义异常需要一些指导，从哪个方面去看

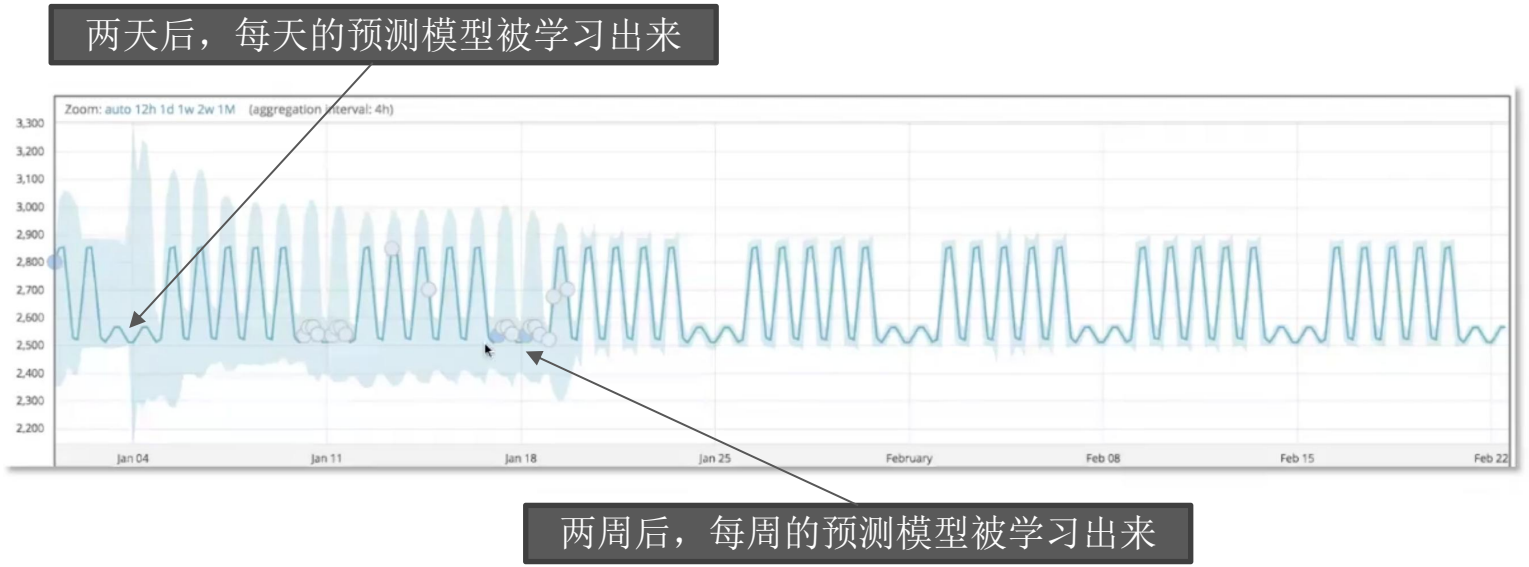
# 如何学习“正常”

- 观察不同的人每天走路的步数，由此预测明天他会走多少步
- 需要观察不同的人，需要观察多久？
  - 一天 / 一周 / 一个月 / 一年 / 十年
- 直觉： 观察的数据多，你的预测越准确
- 使用这些观察来创建一个模型
  - 概率分布函数；使用这个模型找出什么事几乎不可能的事件

# 机器学习帮你自动挑选模型

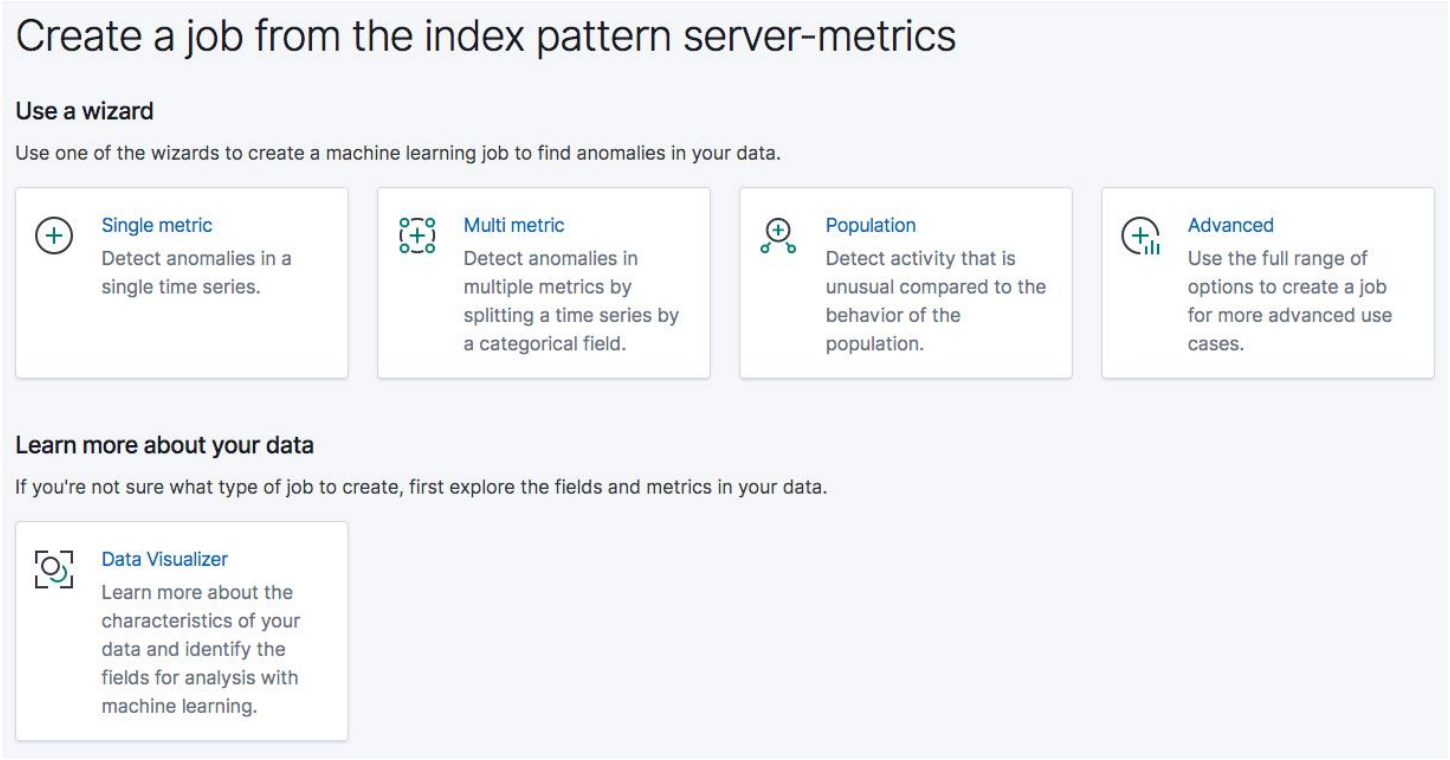
- 使用成熟的机器学习技术，挑选适合数据的正确的统计模型
- 更好的模型 = 更好的异常检测 = 更少的误报和漏报
- 出现在低概率区域，发现异常

# 模型与需要考虑任何的周期

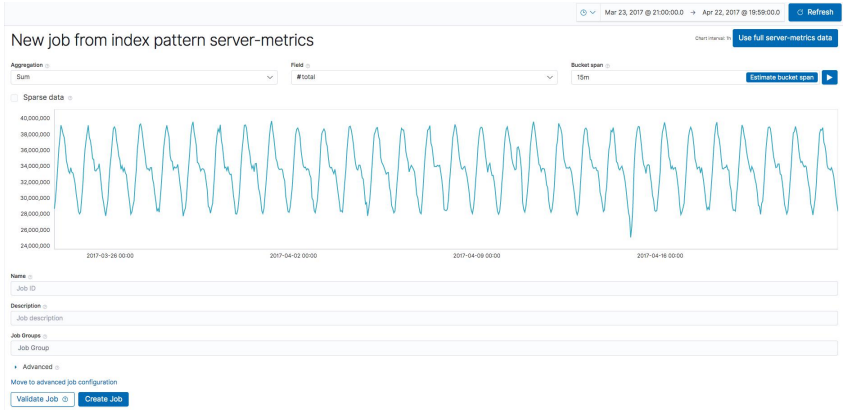



周期选择
需要一定周期的学习，才能使的置信区间的范围更小
时间太长：影响因素太多，导致随机分布
时间太短：完全是随机波动

# ES ML：单指标 / 多指标 / 种群分析



# 单指标任务



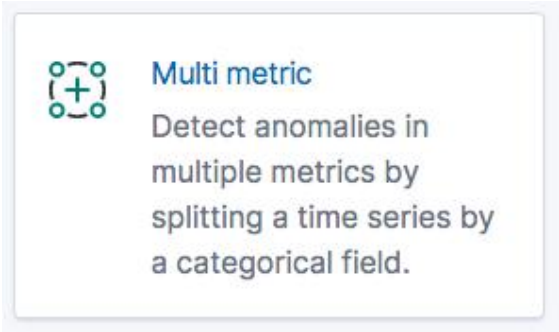

- Create New Job
- 选择 Index Pattern，选择 Fare quote
- 选择单指标任务
  - Aggregation
  - Field
  - Bucket Span （取决于用户的数据业务 / 数据的采样时间 / 波动情况 / 需要预测的频率）
  - Use full data

# Demo

- Create New Single Metric Job
  - 基于 Index Pattern 和 Full data set
- Create Customer URL
  - Saved Query
- 对数据实现预测

# 用机器学习实现异常检测（下）

# 多指标检测




- 多个指标
- 按照某个字段进行分类
- 什么是影响因子
  - 影响因子是一个字段
  - 这个字段从逻辑分析上看，是可能造成异常的原因
  - 不一定必须是一个检测器，但是这个字段通常能够被用于区分数据
  - 对影响因子也可以打分，打分是基于其多大程度会影响到异常

# 多指标预测 Demo

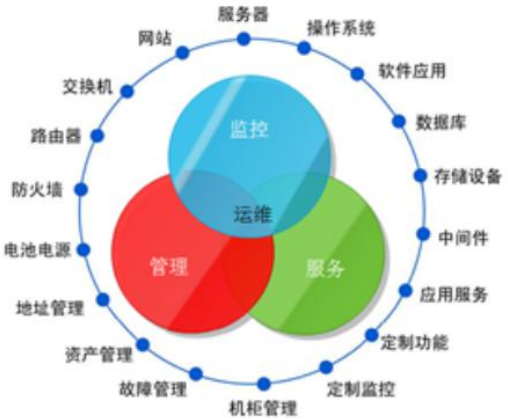
# 种群分析

- 如何分析
  - 种群之间比较
  - 不和自己的过去比较


- 适用情景
  - 个体具备高 Cardinality / 少数个体从时间上是稀疏的 / 作为整体看，群体行为是均匀的
- 不适用的场景
  - 个体的行为各自都不相同

# 用 Elastic Stack 进行日志管理

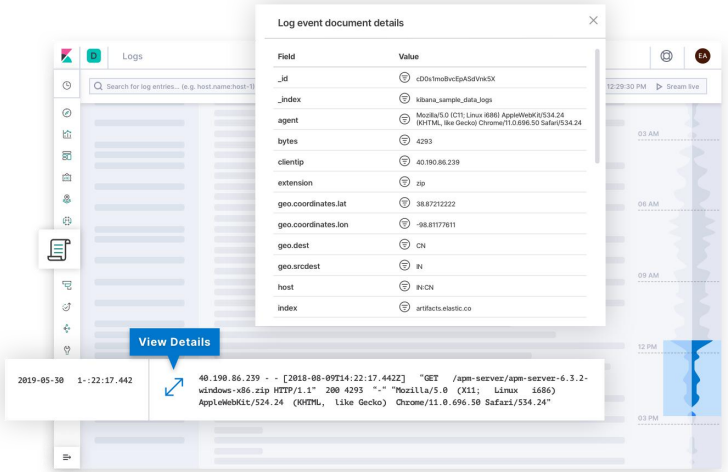
# 日志的重要性

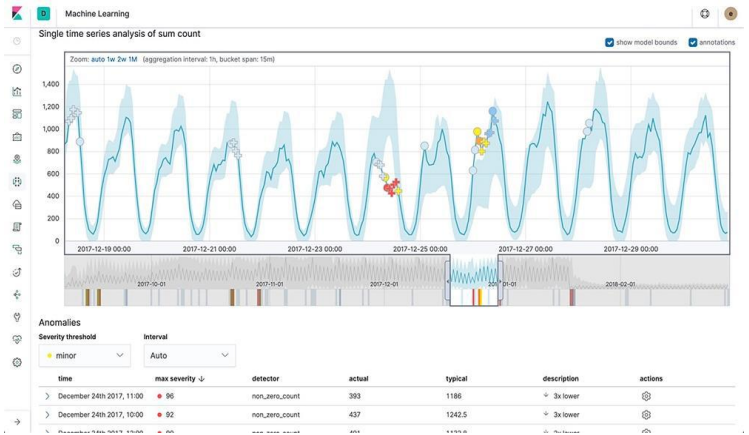



- 为什么重要
  - 运维：医生给病人看病。日志就是病人对自己的陈述
  - 恶意攻击，恶意注册，刷单，恶意密码猜测
- 挑战
  - 关注点很多，任何一个点都有可能引起问题
  - 日志分散在很多机器，出了问题时，才发现日志被删了
  - 很多运维人员是消防员，哪里有问题去哪里

# 集中化日志管理

日志搜集 → 格式化分析 → 检索与可视化 → 风险告警

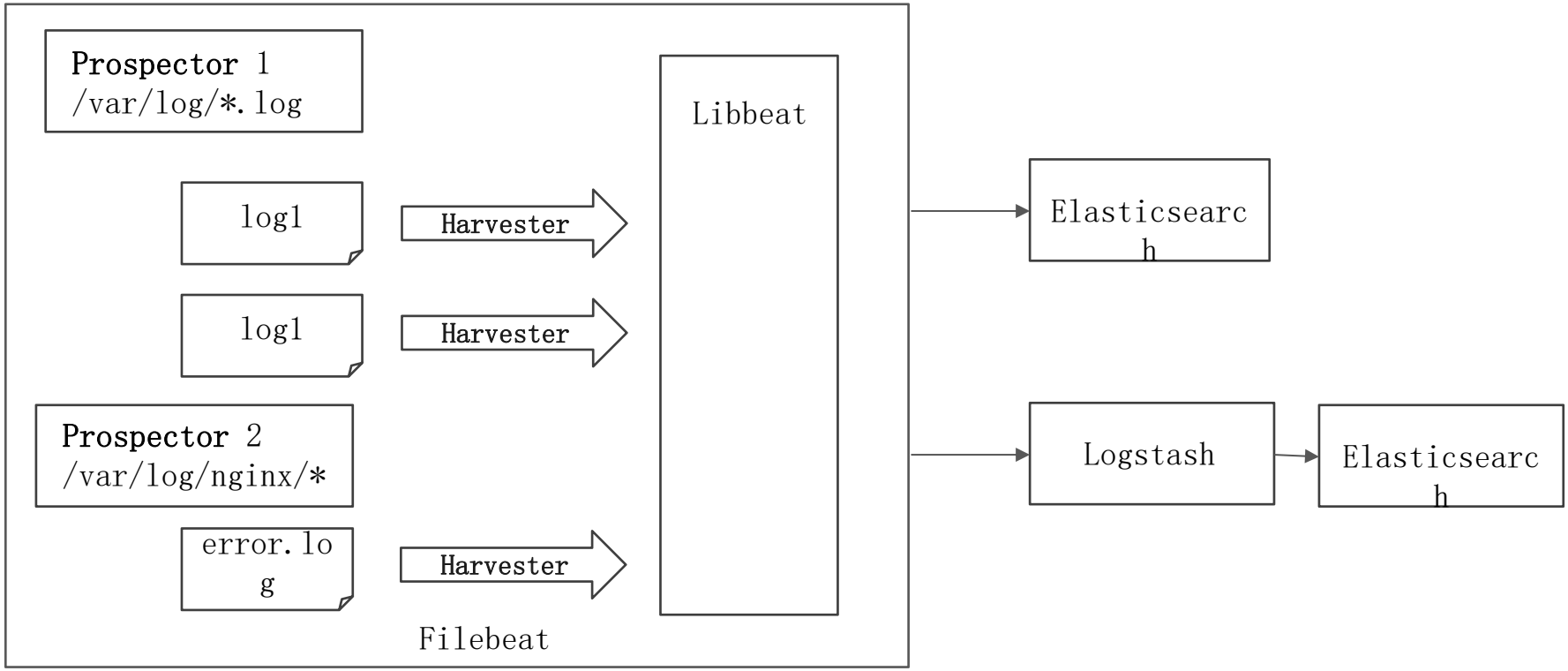


# Filebeat 简介

- 简介
  - A log data shipper for local files
  - 读取日志文件，Filebeat 不做数据的解析，加工处理
    - 日志是非结构化数据
    - 需要进行处理后，以结构化的方式保存到 Elasticsearch
  - 保证数据至少被读取一次
  - 处理多行数据，解析 JSON 格式 ，简单的过滤

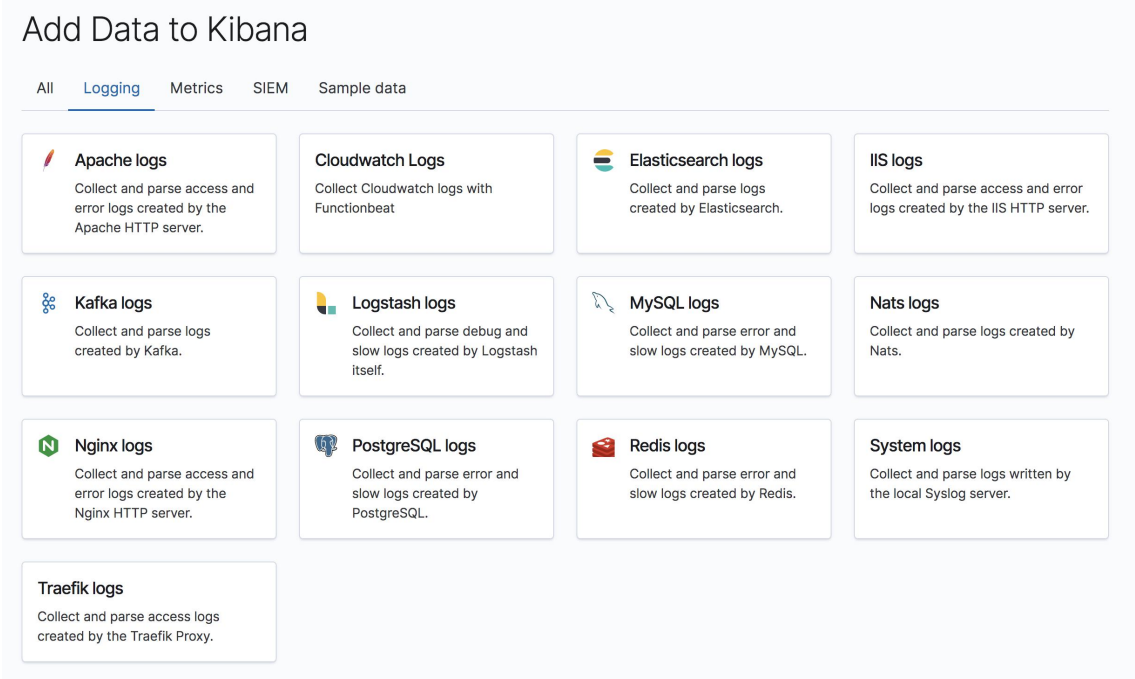
# Filebeat 简介


Prospector 1
/var/log/*.log
log1
Harvester
log1
Harvester
Prospector 2
/var/log/nginx/*
error.log
Harvester
Filebeat
Libbeat
Elasticsearch
Logstash
Elasticsearch

# Filebeat 执行流程

- 定义数据采集：Prospector 配置。通过 filebeat.yml
- 建立数据模型： Index Template
- 建立数据处理流程：Ingest Pipeline
- 存储并提供可视化分析：ES + Kibana Dashboard

# Modules 开箱即用




- 大量开箱即用的日志模块
  - 简化使用流程
  - 减少开发的投入
  - 最佳参考实践
- 一些命令
  - `./filebeat modules list`
  - `./Filebeat modules enable nginx`

极客时间

# Demo

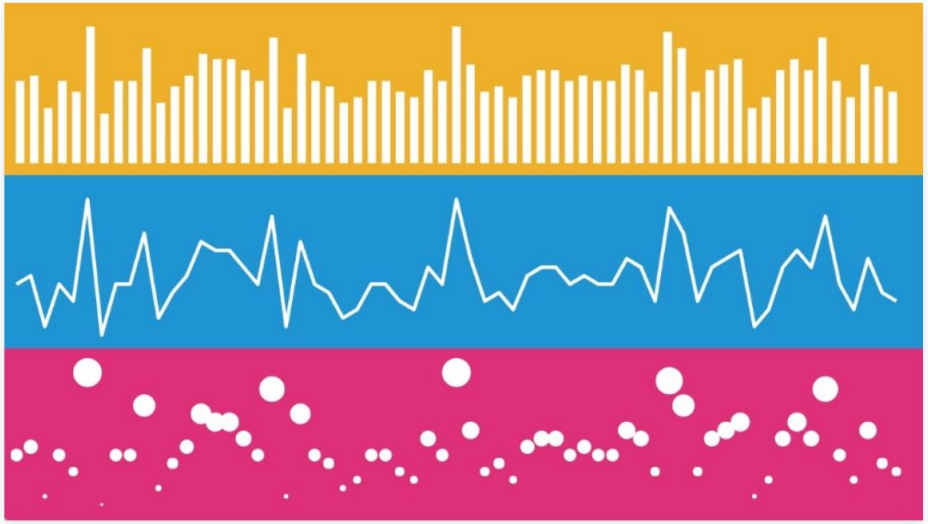
- Filebeat modules list
- Filebeat modules enable system
- Module 和 module.d 目录
- Filebeat export template
- Filebeat setup -dashboards
- ./filebeat -e

# 用 Canvas 进行数据的实时展示

# 实时展示数据，并且达到完美像素级要求

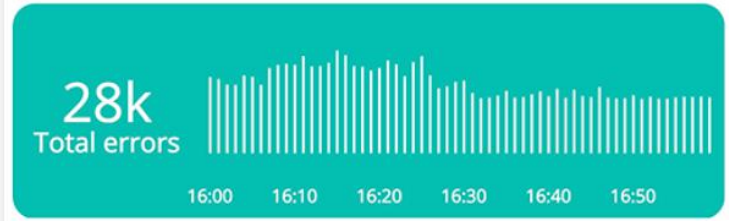
- 用更加酷炫的方式，演绎你的数据
  - 基于 ES 实现准实时的数据分析
- 更好的想法，更大的屏幕
  - 品牌宣传，会议大屏
- 高度定制化
  - 调色板 / CSS / 拖放元素

# 个性化方式展现你的数据



- 公司的 Logo
- 符合公司的配色方案以及设计元素
- Kibana 中免费提供

# 日志分析



```
[2019-02-22 09:59:59,965] DEBUG [ReplicaFetcher replicaId=1,
leaderId=2, fetcherId=0] Built incremental fetch
(sessionId=2101881119, epoch=503272) for node 2. Added 0
partition(s), altered 0 partition(s), removed 0 partition(s)
out of 52 partition(s)
(org.apache.kafka.clients.FetchSessionHandler)
```

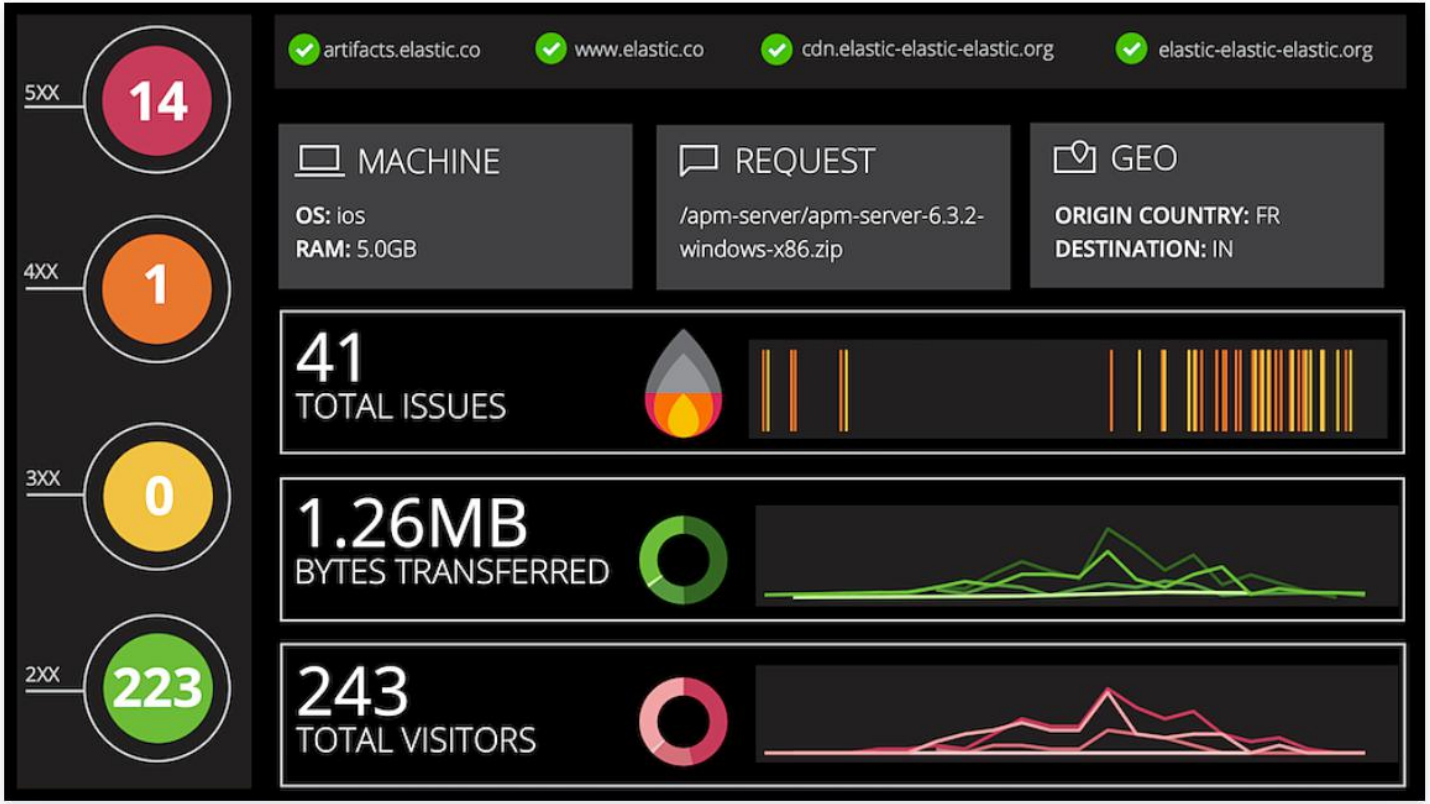
```
[2019-02-22 09:59:59,965] DEBUG [ReplicaFetcher replicaId=1,
leaderId=2, fetcherId=0] Node 2 sent an incremental fetch
response for session 2101881119 with 0 response partition(s),
52 implied partition(s)
(org.apache.kafka.clients.FetchSessionHandler)
```

```
[2019-02-22 09:59:59,961] DEBUG [ReplicaFetcher replicaId=0,
leaderId=1, fetcherId=0] Node 1 sent an incremental fetch
response for session 1732296219 with 0 response partition(s),
54 implied partition(s)
(org.apache.kafka.clients.FetchSessionHandler)
```

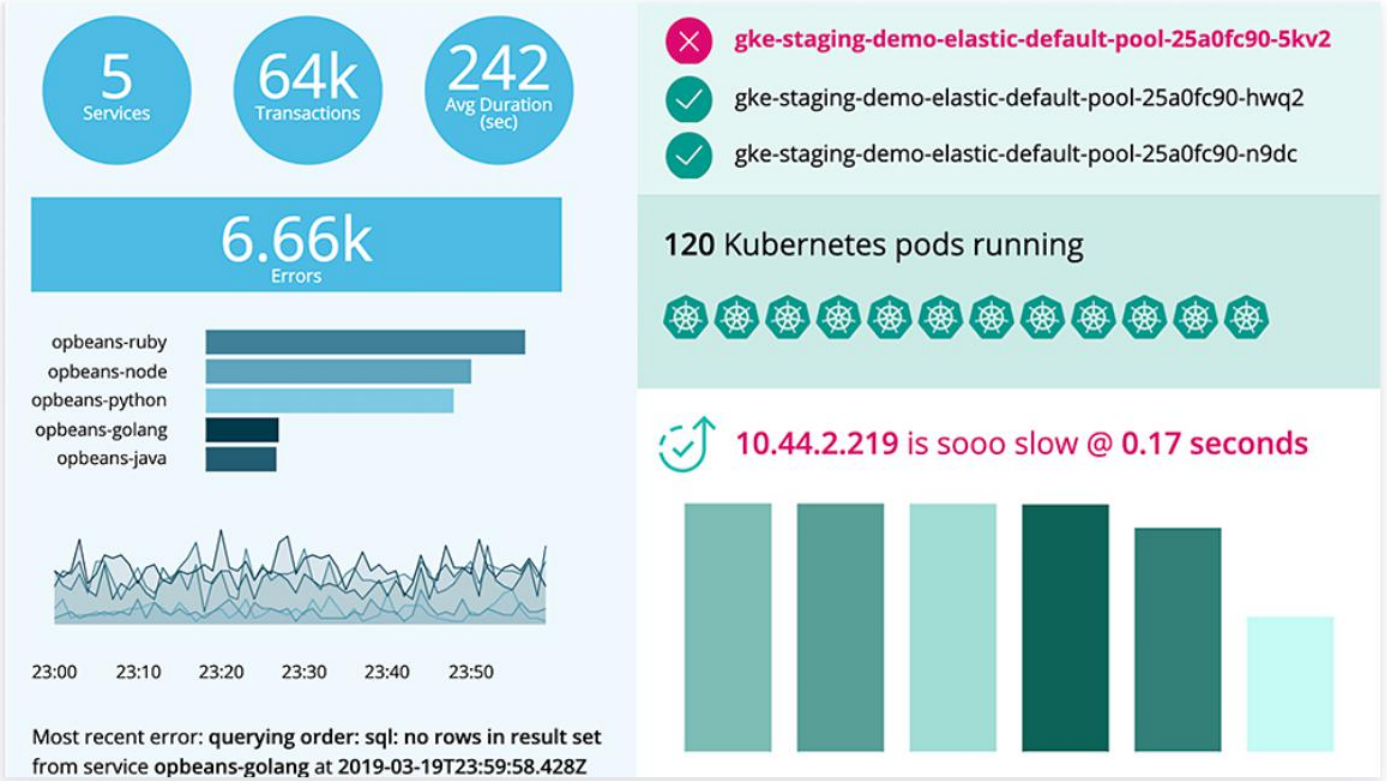
```
[2019-02-22 09:59:59,961] DEBUG [ReplicaFetcher replicaId=0,
leaderId=1, fetcherId=0] Built incremental fetch
(sessionId=1732296219, epoch=4764391) for node 1. Added 0
partition(s), altered 0 partition(s), removed 0 partition(s)
out of 54 partition(s)
(org.apache.kafka.clients.FetchSessionHandler)
```

```
2019-02-22 09:59:59.927 UTC [37283571] elastic@opbeans LOG:
```

# 基础设施监控


5XX
14
4XX
1
3XX
0
2XX
223
artifacts.elastic.co
www.elastic.co
cdn.elastic-elastic-elastic.org
elastic-elastic-elastic.org
MACHINE
OS: ios
RAM: 5.0GB
REQUEST
/apm-server/apm-server-6.3.2-windows-x86.zip
GEO
ORIGIN COUNTRY: FR
DESTINATION: IN
41
TOTAL ISSUES
1.26MB
BYTES TRANSFERRED
243
TOTAL VISITORS

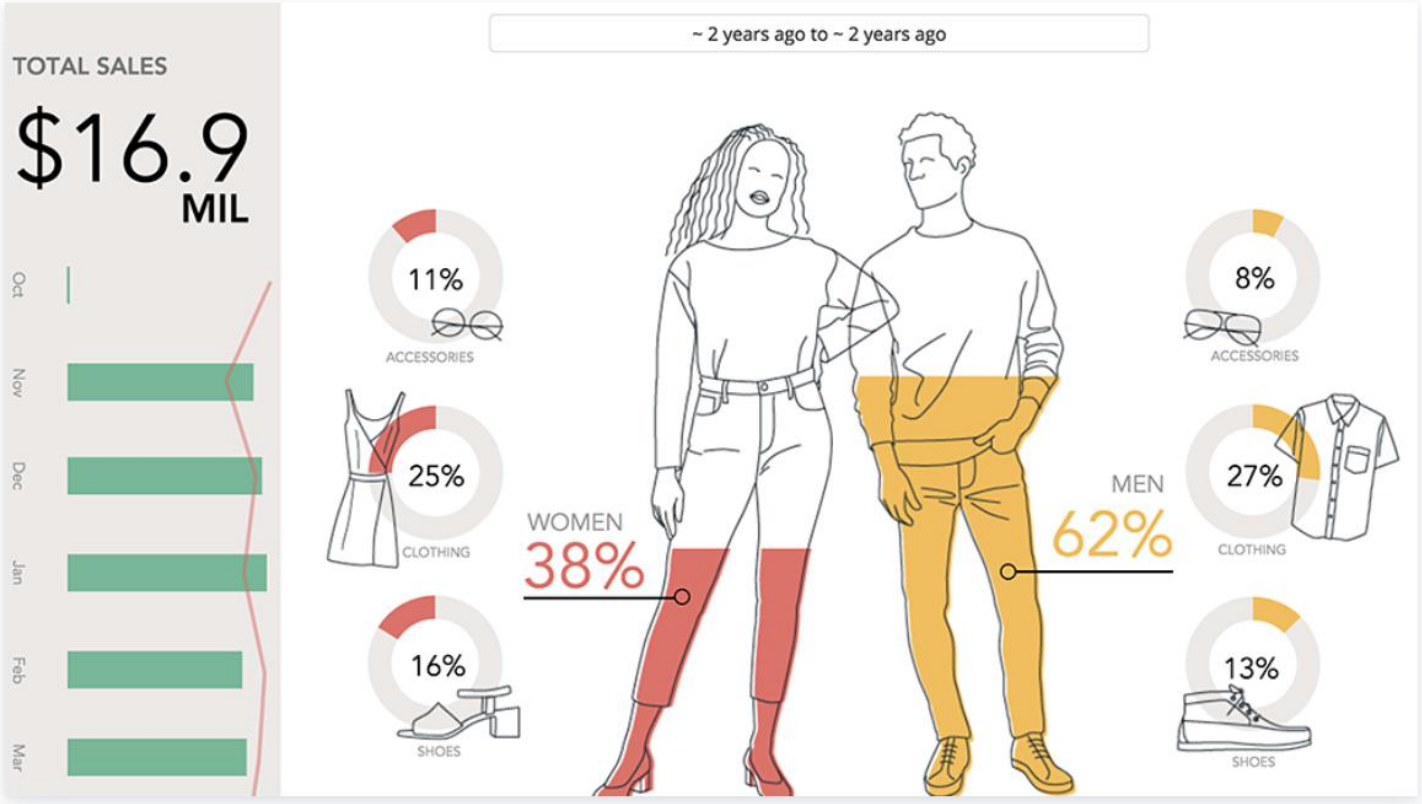
# APM


5
Services
64k
Transactions
242
Avg Duration (sec)
6.66k
Errors
opbeans-ruby
opbeans-node
opbeans-python
opbeans-golang
opbeans-java
23:00
23:10
23:20
23:30
23:40
23:50
Most recent error: querying order: sql: no rows in result set
from service opbeans-golang at 2019-03-19T23:59:58.428Z
gke-staging-demo-elastic-default-pool-25a0fc90-5kv2
gke-staging-demo-elastic-default-pool-25a0fc90-hwq2
gke-staging-demo-elastic-default-pool-25a0fc90-n9dc
120 Kubernetes pods running
10.44.2.219 is sooo slow @ 0.17 seconds

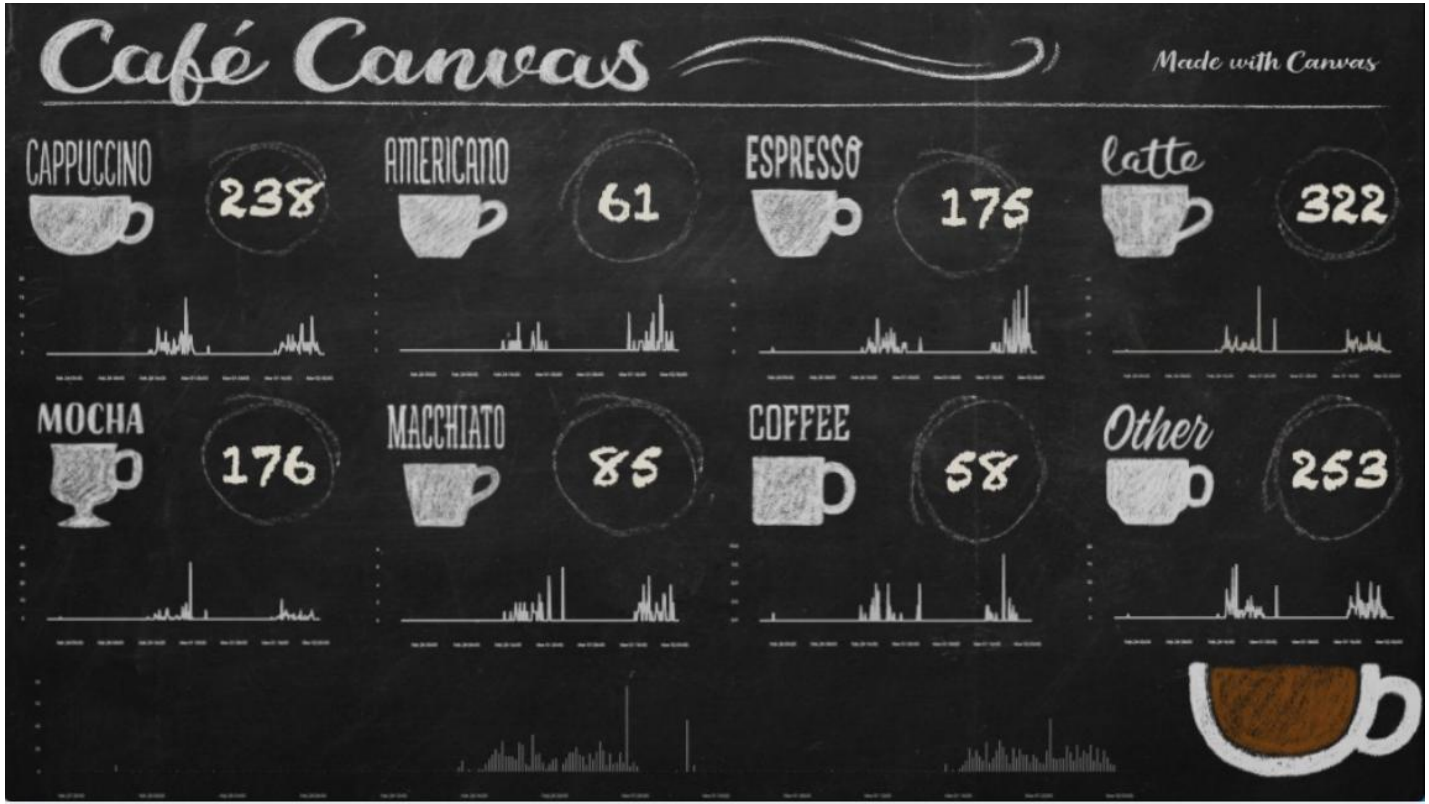
# 安全运维


Security Operations Team Report.pdf
1 / 1
Security Operations Team
Security Operations Team - April 9, 2019
Operational Activity
Here's some visibility into the day to day operational activities performed by the Security Operations Team team. If you'd like to inspect the metrics in more detail, feel free view our secure Kibana Dashboard.
Events Breakdown
4.9k
Events
9
Event types
17
Hosts effected
1k
SEV 1's

# 业务分析


TOTAL SALES
$16.9
MIL
Oct
Nov
Dec
Jan
Feb
Mar
~ 2 years ago to ~ 2 years ago
11%
ACCESSORIES
25%
CLOTHING
16%
SHOES
WOMEN
38%
MEN
62%
8%
ACCESSORIES
27%
CLOTHING
13%
SHOES

# Demo


Café Canvas
Made with Canvas
CAPPUCCINO
238
AMERICANO
61
ESPRESSO
175
latte
322
MOCHA
176
MACCHIATO
85
COFFEE
58
Other
253

极客时间

扫码试看/订阅极客时间

《Elasticsearch核心技术与实战》视频课程