ElasticSearch分析与实践

卢亿雷 from AdMaster(精硕科技)

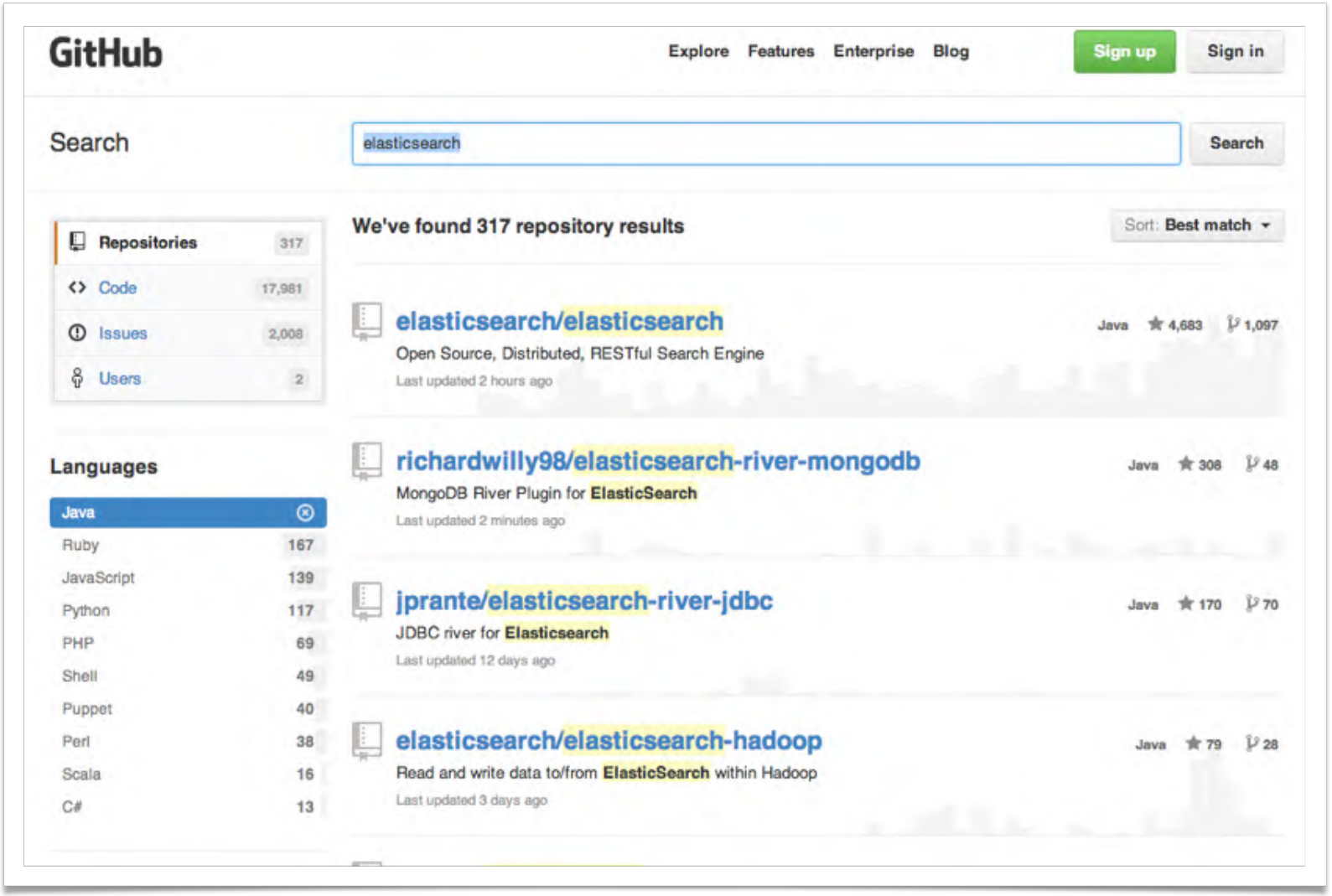
内容

- ElasticSearch特点及生态圈
- Lucene原理
- ElasticSearch架构和插件
- ElasticSearch管理和监控工具
- ElasticSearch应用案例
- ELK实践

ElasticSearch特点及生态圈

- 分布式实时分析与检索
- 高可用
- 多租户
- 全文搜索
- 面向文档
- 易用的Restful API
- 基于Apache Lucene

ElasticSearch特点及生态圈



ElasticSearch特点及生态圈

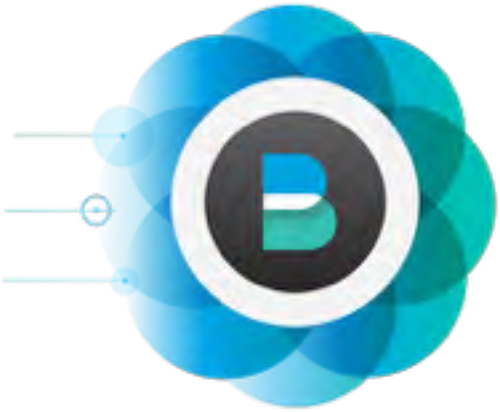
——ELK

ElasticSearch特点及生态圈

——ES-Hadoop

ElasticSearch特点及生态圈

——ES-Beats



Packetbeat

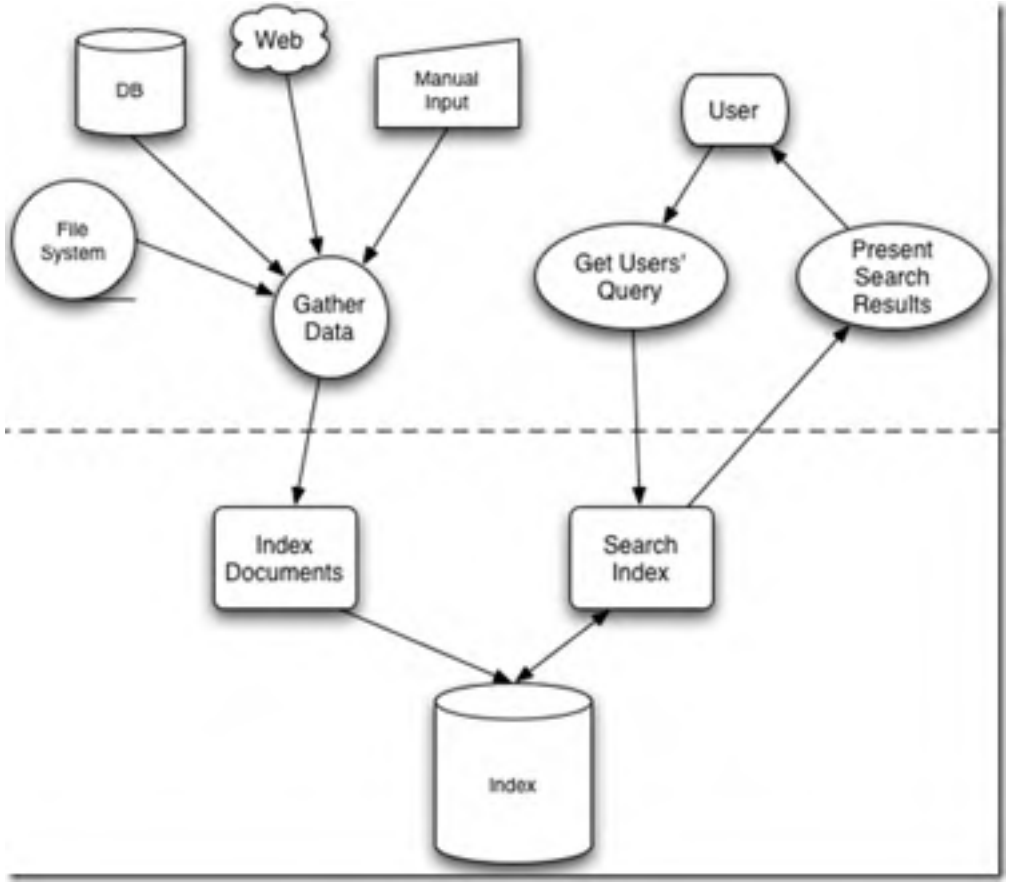
Topbeat

Filebeat

Winlogbeat

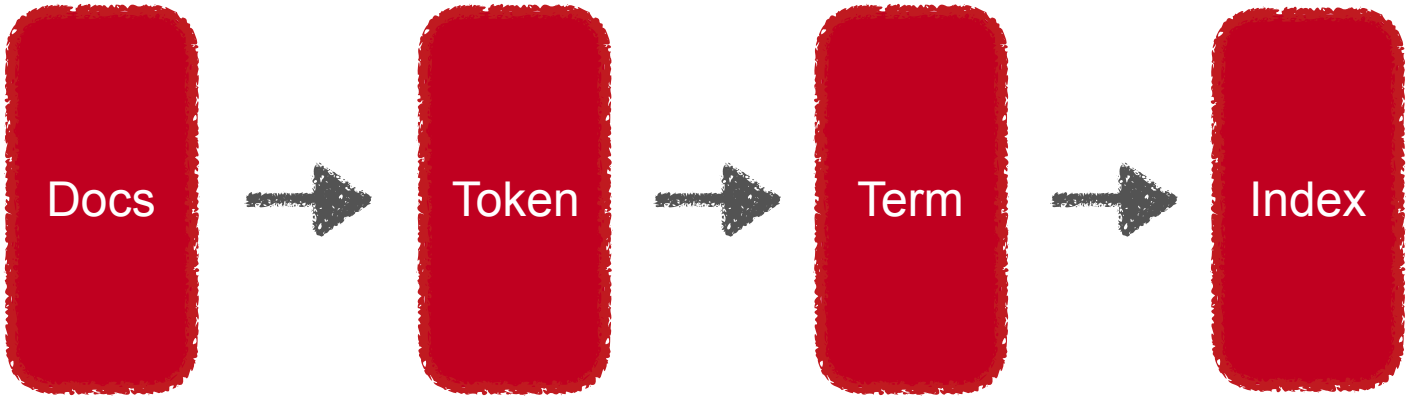
Lucene原理

- 索引创建-Indexing
- 索引查询-Search index



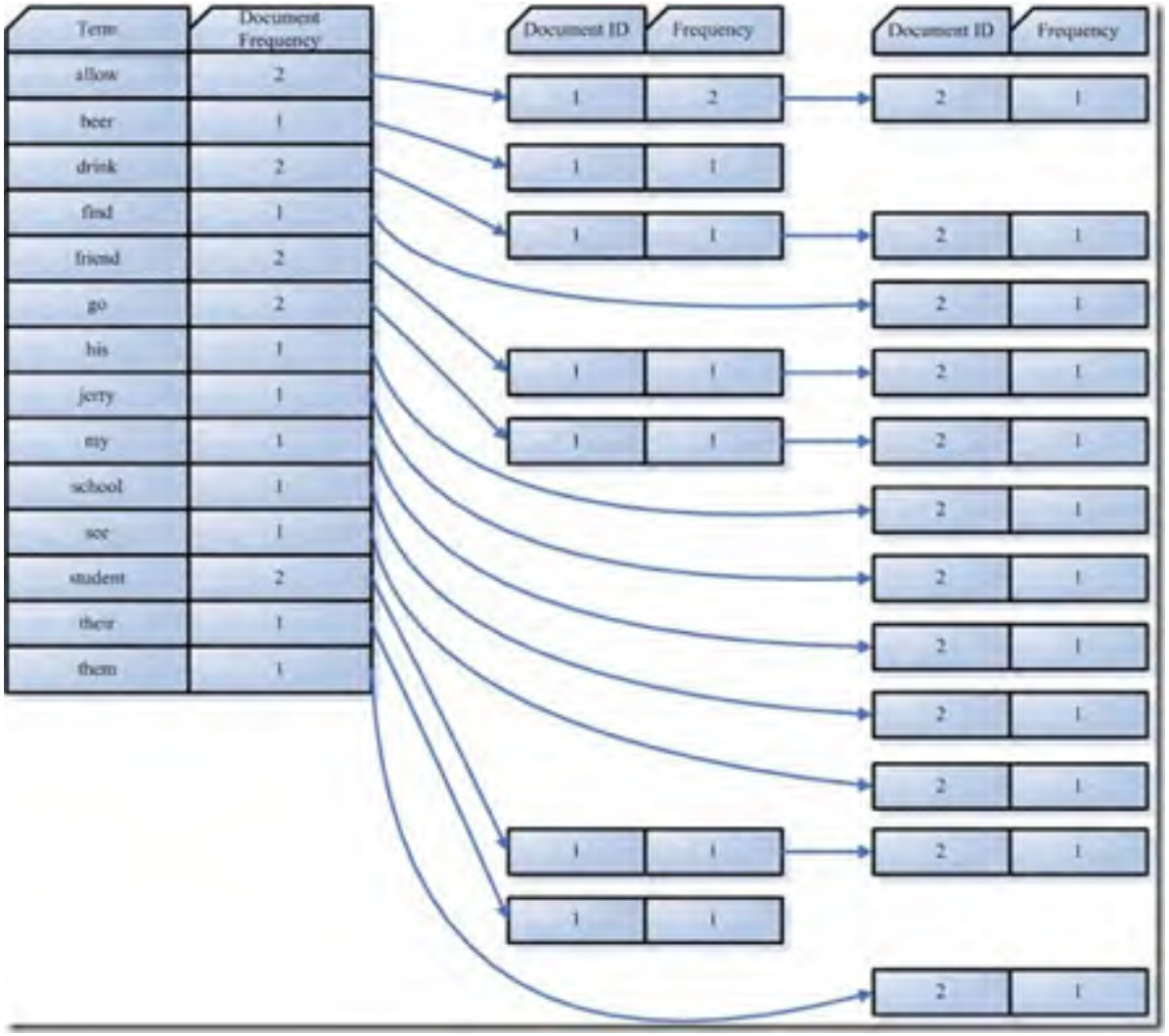
Lucene原理

- 索引创建indexing



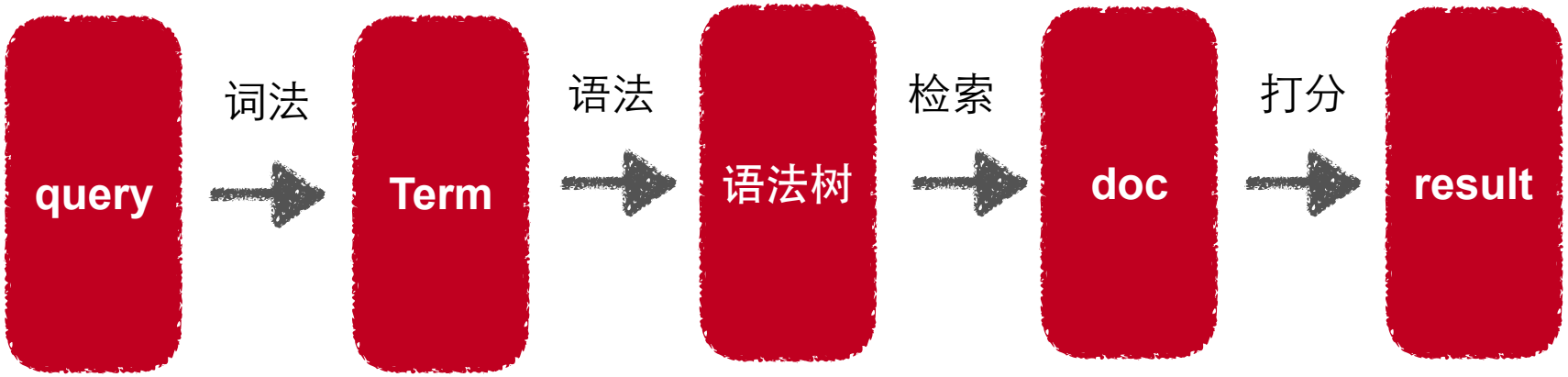
Lucene原理

- 倒排索引表



Lucene原理

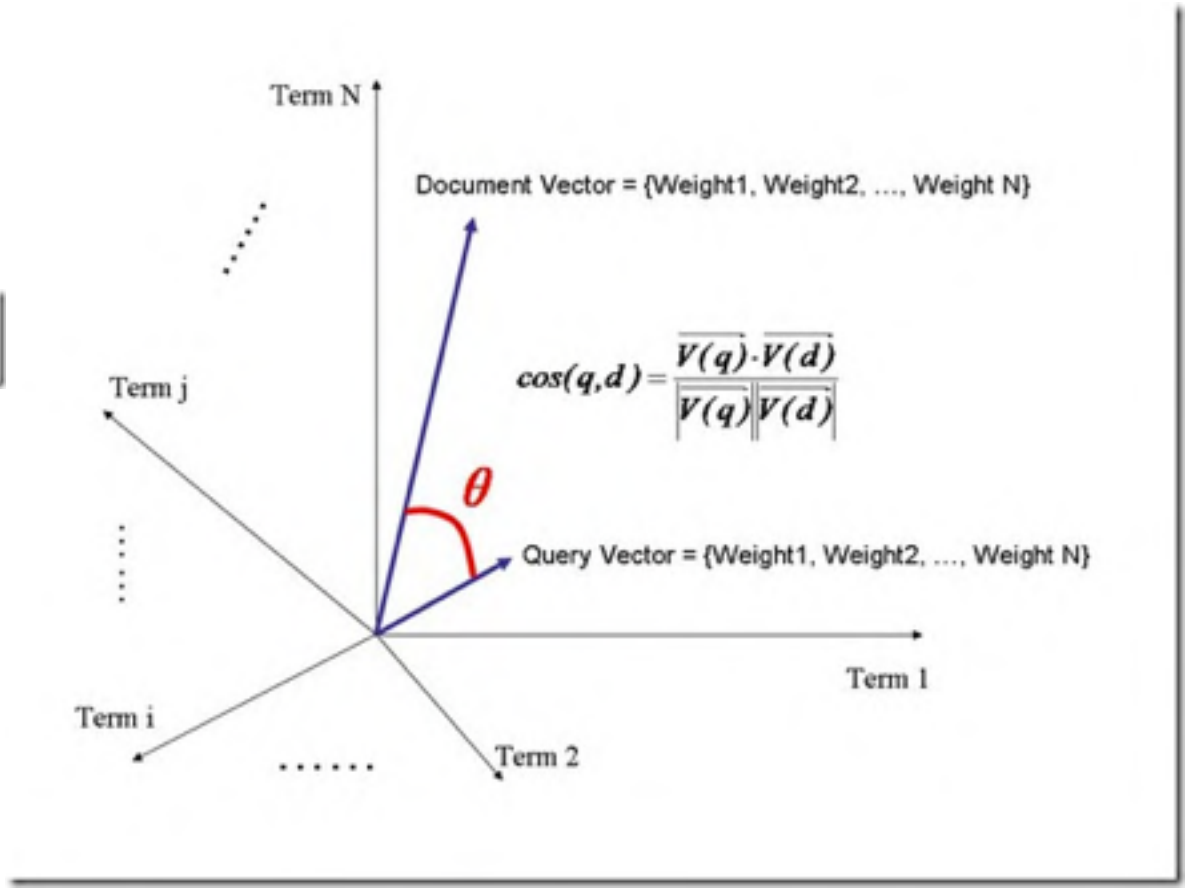
- 索引查询



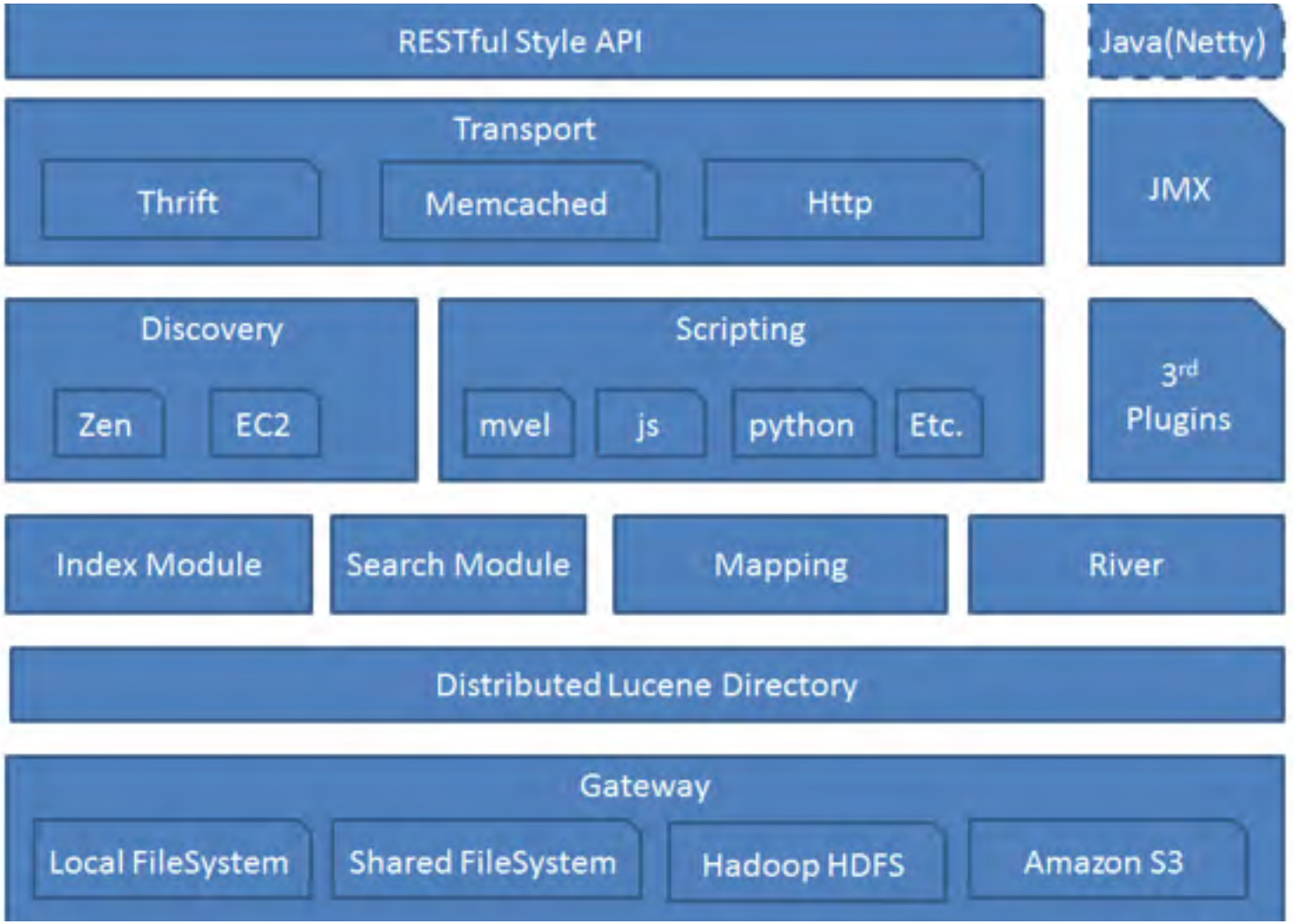
Lucene原理

- 索引查询，权重计算，相关性判断
- VSM向量空间模型

$$w_{t,d} = tf_{t,d} \times \log(n / df_t)$$



ElasticSearch架构和插件



ElasticSearch架构和插件

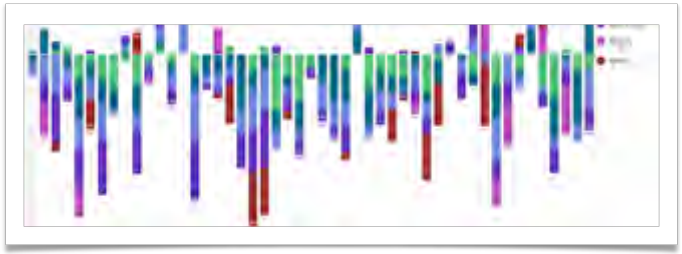
- 分词插件
- 同步插件
- 数据传输插件
- 脚本插件
- Alert
- Shield

ElasticSearch-Aggregations

```
GET /person/person/_search?search_type=count
{
    "aggs": {
        "by_country": {
            "terms": {
                "field": "address.country"
            }
        }
    }
}
```

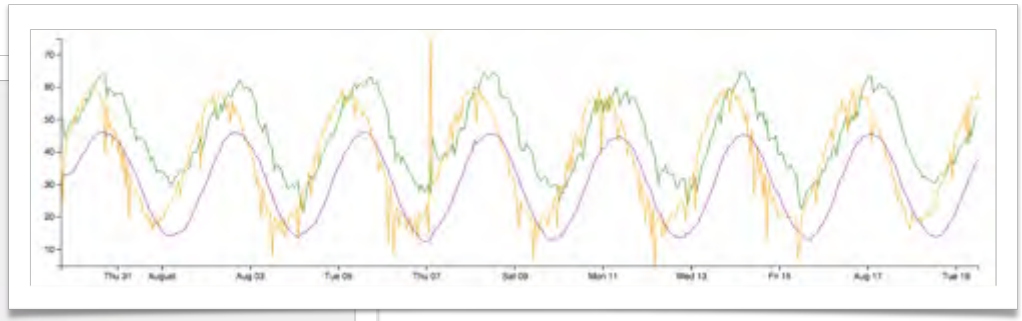
```
{ ..., "aggregations" : {
    "by_country" : {
      "buckets" : [ {
        "key" : "England",
        "doc_count" : 30051
      }, {
        "key" : "Germany",
        "doc_count" : 30004
      }, {
        "key" : "France",
        "doc_count" : 15034
      }, {
        "key" : "Spain",
        "doc_count" : 14912
      } ]}}}
```

Like facets but with more power
Can be nested to add additional dimensions
Give analytical insights into data
Allow complex visualizations
Major types: buckets and metrics
Types: terms, histogram, percentiles, etc.



ElasticSearch-Pipeline Aggregations

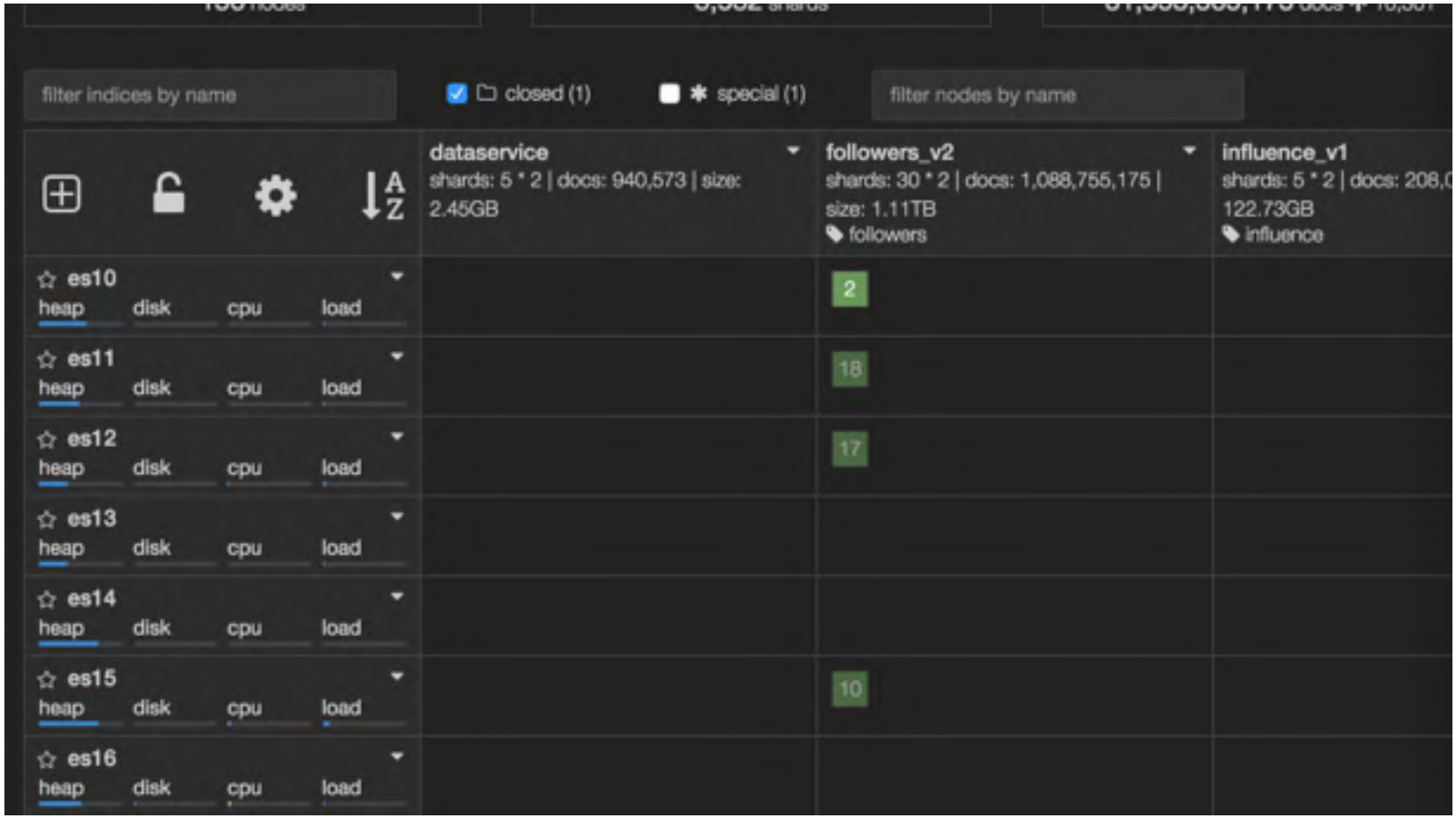
Work on outputs of other aggregations
Used for smoothing, prediction, etc.
Different types: avg, derivative, max, min, sum moving avg, cumulative sum, etc.



```
{
    "my_date_histo":{
        "date_histogram":{
            "field":"timestamp",
            "interval":"day"
        },
        "aggs":{
            "the_sum":{
                "sum":{ "field": "lemmings" }
            },
            "the_movavg":{
                "moving_avg":{ "buckets_path": "the_sum" }
            }
        }
    }
}
```

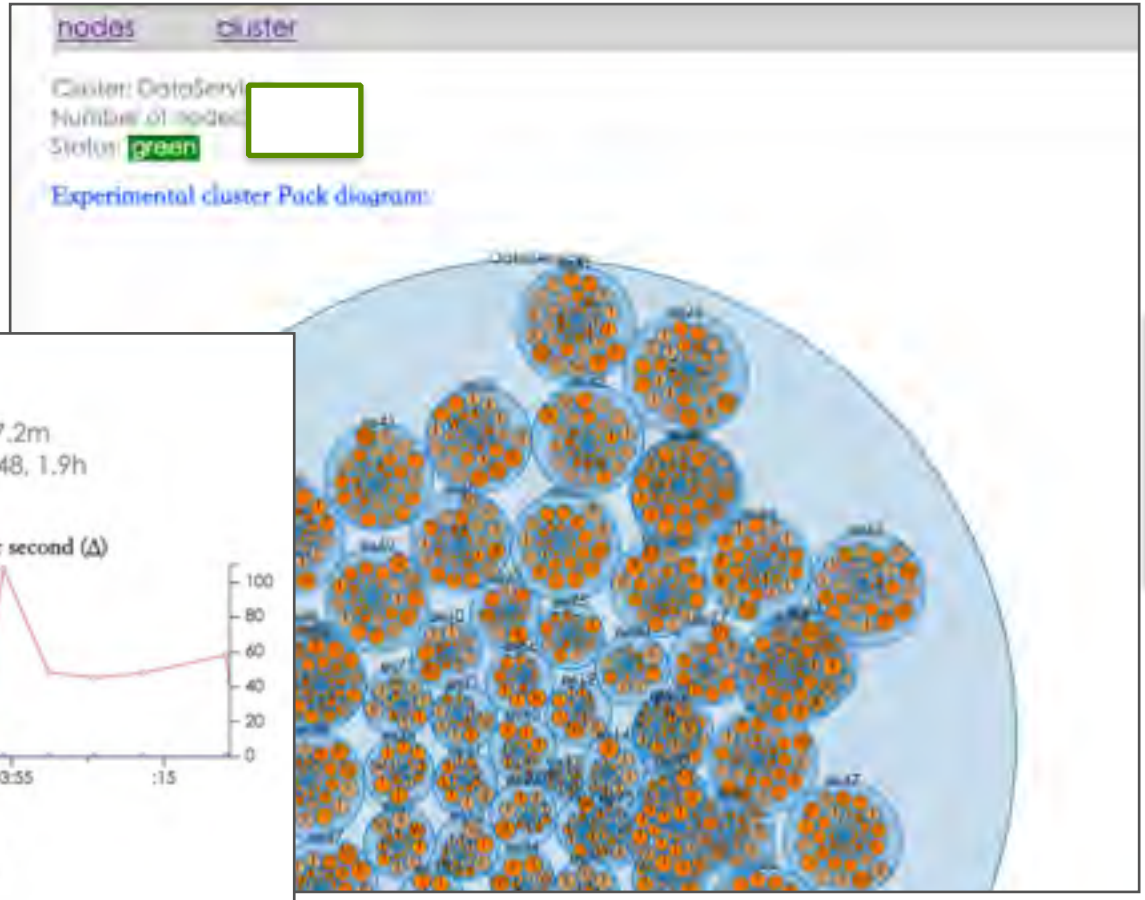
ElasticSearch管理和监控工具

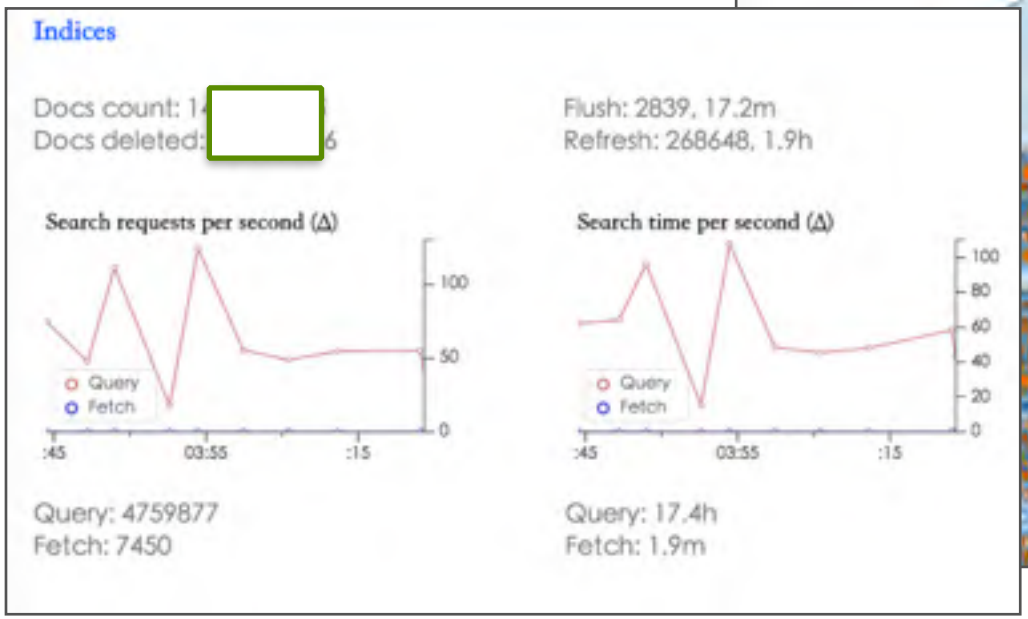
- kopf——优秀的监控和管理工具



ElasticSearch管理和监控工具

- big desk
- 集群整体和流量情况





ElasticSearch管理和监控工具

- Marvel, 官方监控插件

ElasticSearch应用案例

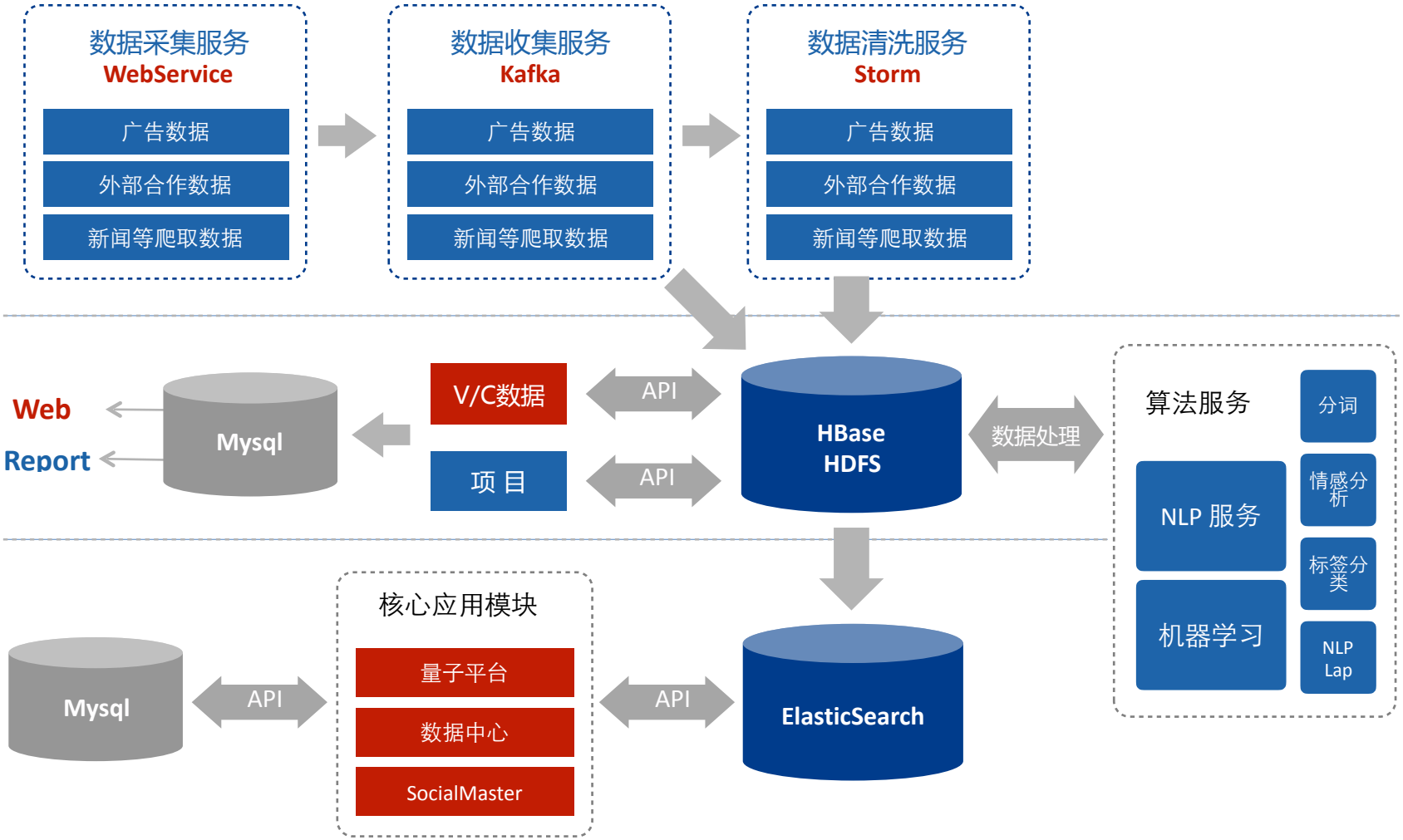
测试条件：

- 记录条数分为100亿以内和1000亿条
- 服务器数量为70台，配置为:CPU 12核，内存96G，硬盘48T
- 测试语句：**select count(*) from test where age > 25 and ge**
  0 and os > "500" and sc in
  ("0001009","0002036","0016030",”…") or bs>585 and gr
  by age,gender,os,bs
- 总共14列(200列）：动态列为3列（多值列），普通列为11列

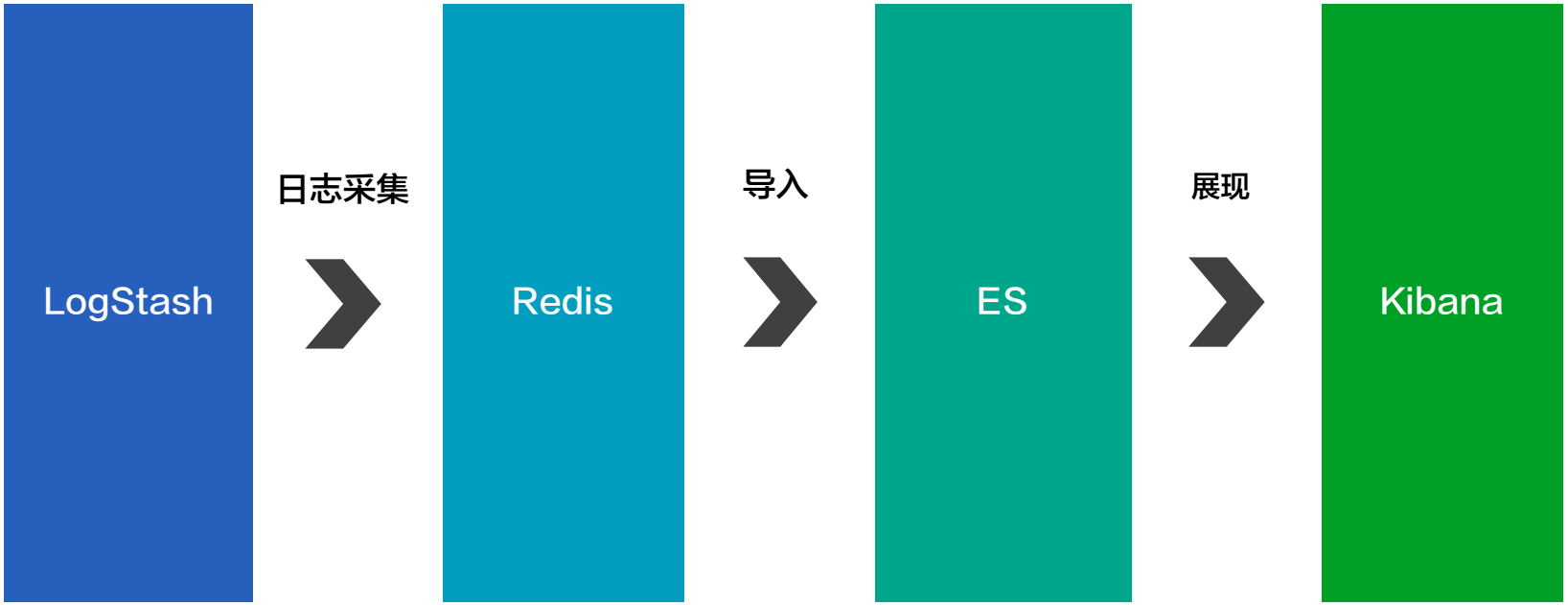
ElasticSearch应用案例

1000亿	单次	并发5个	并发10个
ElasticSearch	19005ms	21005ms	27736ms
Pinot	19019ms	failed	failed

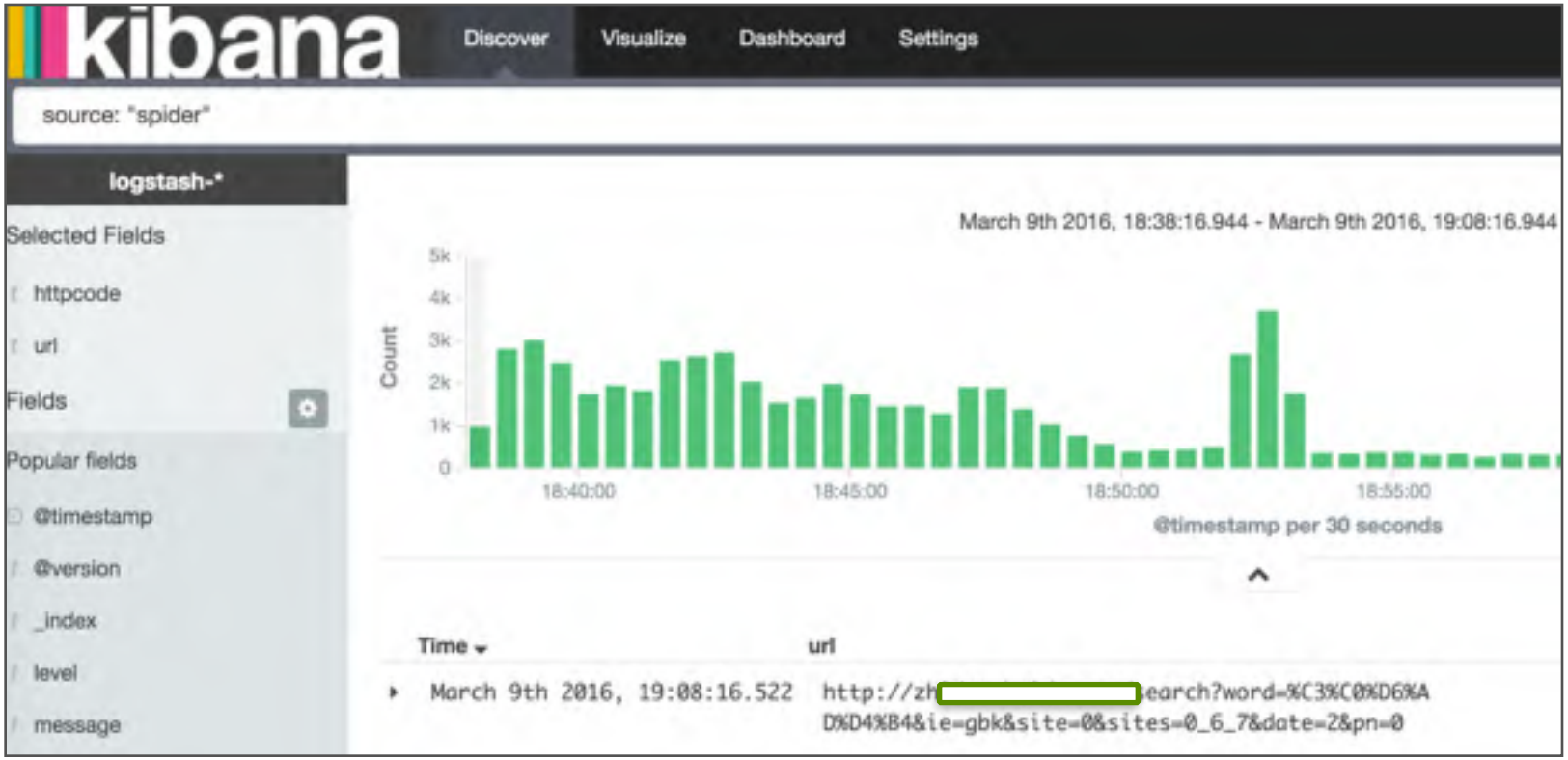
ElasticSearch数字营销案例



ELK实践



ELK实践


kibana
Discover
Visualize
Dashboard
Settings
source: "spider"
logstash-*
Selected Fields
httpcode
url
Fields
Popular fields
@timestamp
@version
_index
level
message
March 9th 2016, 18:38:16.944 - March 9th 2016, 19:08:16.944
Count
5k
4k
3k
2k
1k
0
18:40:00
18:45:00
18:50:00
18:55:00
@timestamp per 30 seconds
Time
url
March 9th 2016, 19:08:16.522
http://zh[redacted]earch?word=%C3%C0%D6%AD%D4%B4&ie=gbk&site=0&sites=0_6_7&date=2&pn=0

Kibana-Discover

- 设置时间过滤器
- 搜索并将搜索保存
- 页面自动刷新
- 按字段过滤
- 文档列表排序
- 查看字段数据统计

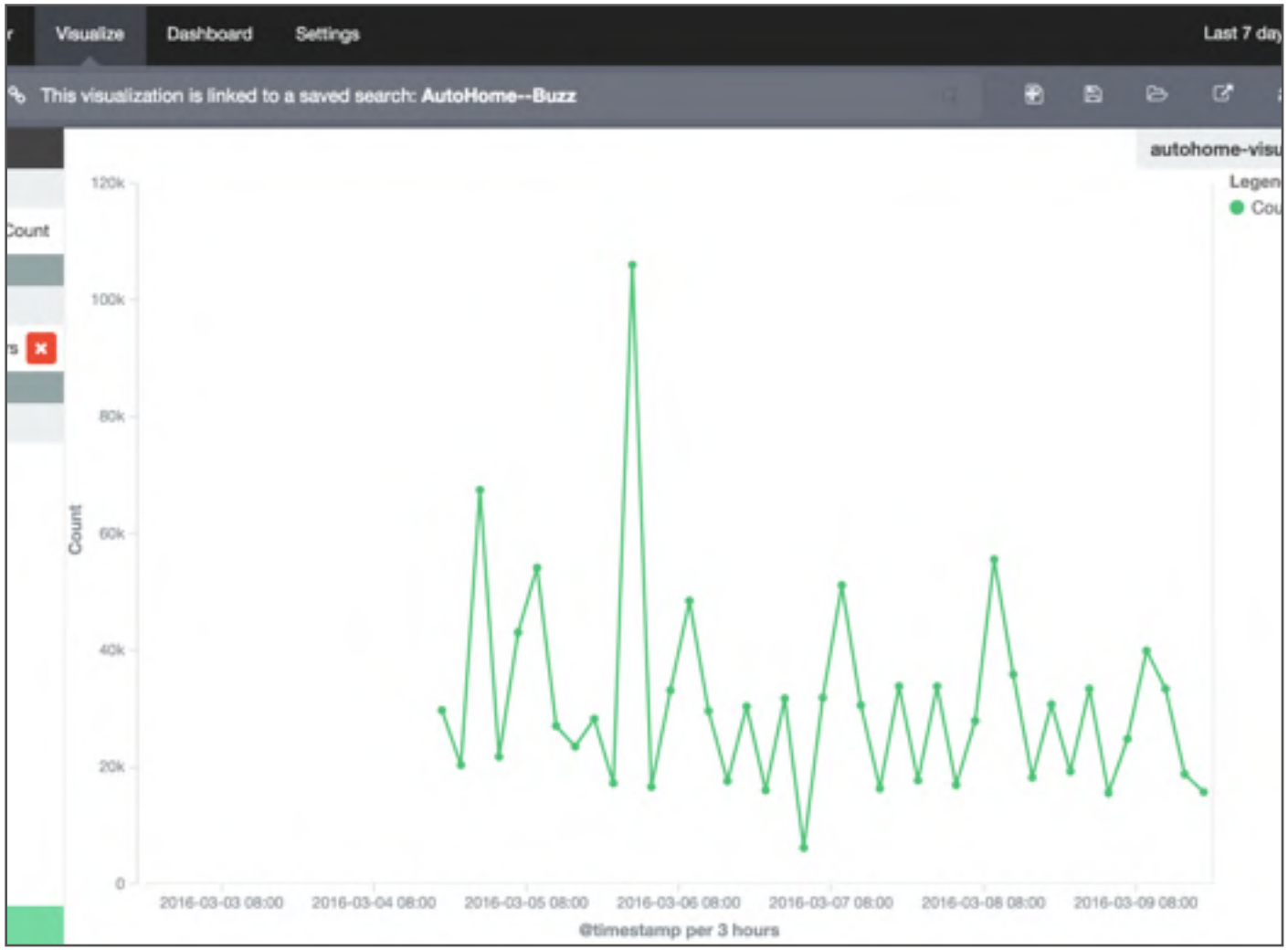
Kibana-Visualize

- 创建图表：
- 选择可视化图表类型
- 选择数据源（已保存的搜索或新的搜索）
- 配置

    Y轴的聚合类型：count, average, sum, min, max,

    cardinality(unique count)

ELK实践



Q & A

邮箱：johnlya@163.com
微信：johnlya

源码:
http://github.com/elastic
英文社区:
http://discuss.elastic.co
中文社区:
http://elasticsearch.cn