

ElasticSearch分析与实践



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

The screenshot shows the GitHub search interface with the query 'elasticsearch'. The search results are sorted by 'Best match' and show 317 repository results. The top results are:

- elasticsearch/elasticsearch**: Open Source, Distributed, RESTful Search Engine. Last updated 2 hours ago. 4,683 stars, 1,097 forks.
- richardwilly98/elasticsearch-river-mongodb**: MongoDB River Plugin for ElasticSearch. Last updated 2 minutes ago. 308 stars, 48 forks.
- jprante/elasticsearch-river-jdbc**: JDBC river for Elasticsearch. Last updated 12 days ago. 170 stars, 70 forks.
- elasticsearch/elasticsearch-hadoop**: Read and write data to/from ElasticSearch within Hadoop. Last updated 3 days ago. 79 stars, 28 forks.

On the left side, there are navigation links for Repositories (317), Code (17,981), Issues (2,008), and Users (2). Below that is a 'Languages' section with a list of programming languages and their respective repository counts:

Language	Count
Java	317
Ruby	167
JavaScript	139
Python	117
PHP	69
Shell	49
Puppet	40
Perl	38
Scala	16
C#	13

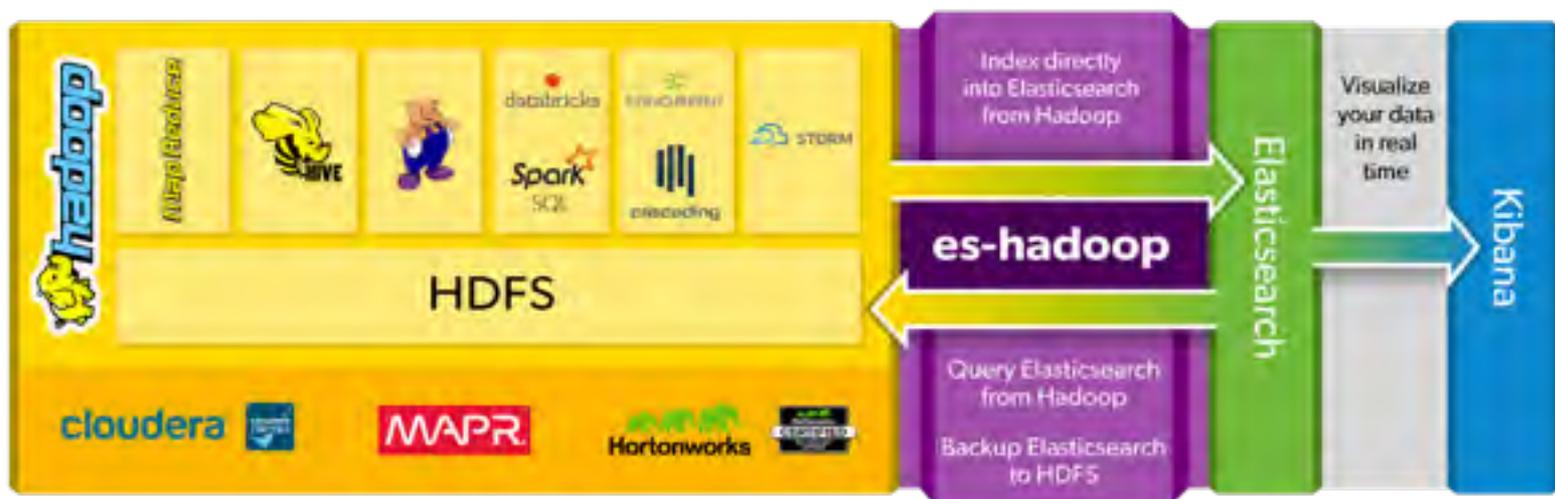
ElasticSearch特点及生态圈

——ELK



ElasticSearch特点及生态圈

——ES-Hadoop



ElasticSearch特点及生态圈

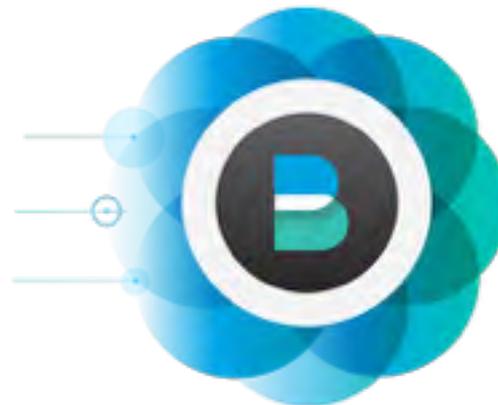
——ES-Beats

Packetbeat

Topbeat

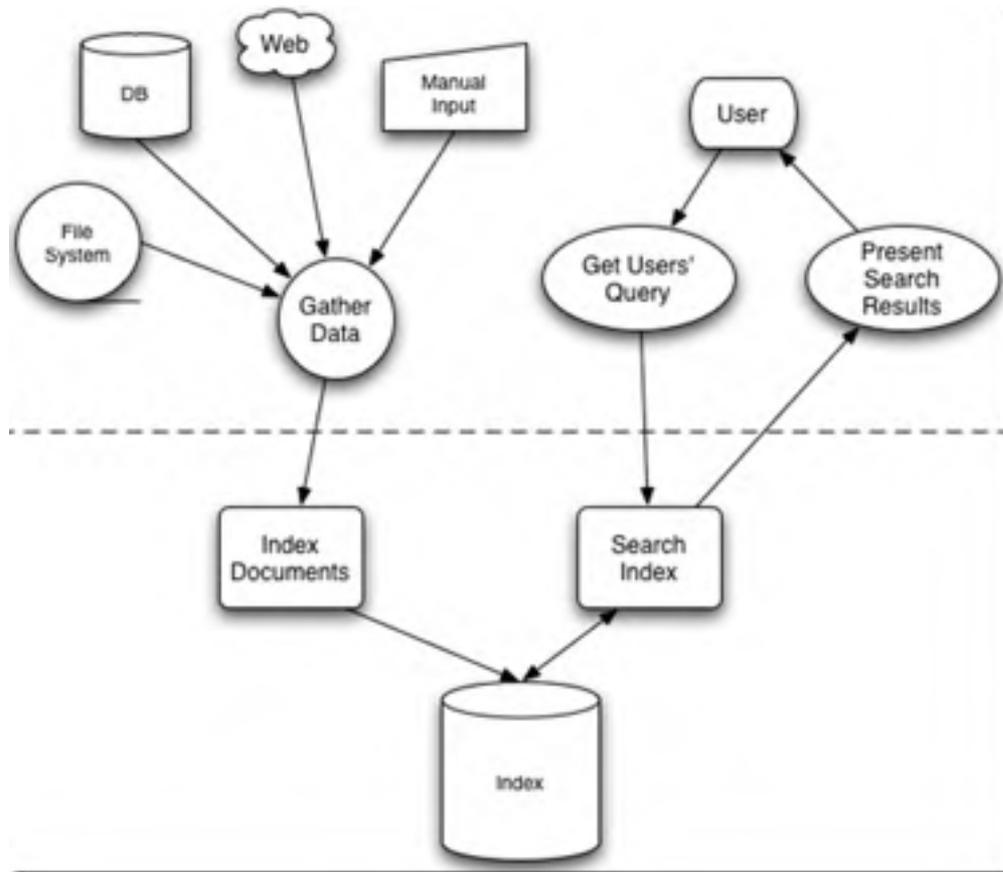
Filebeat

Winlogbeat



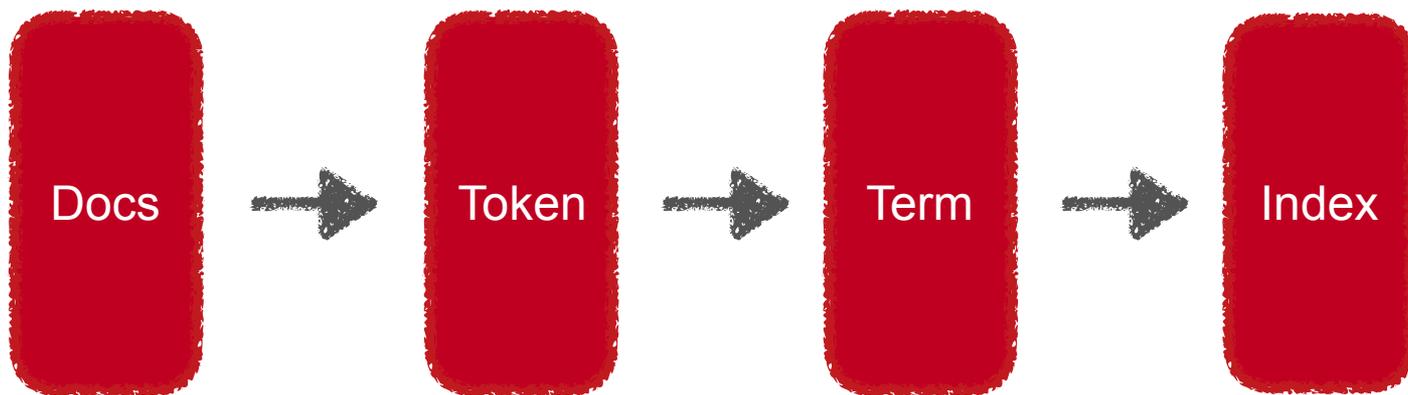
Lucene原理

- 索引创建-Indexing
- 索引查询-Search index



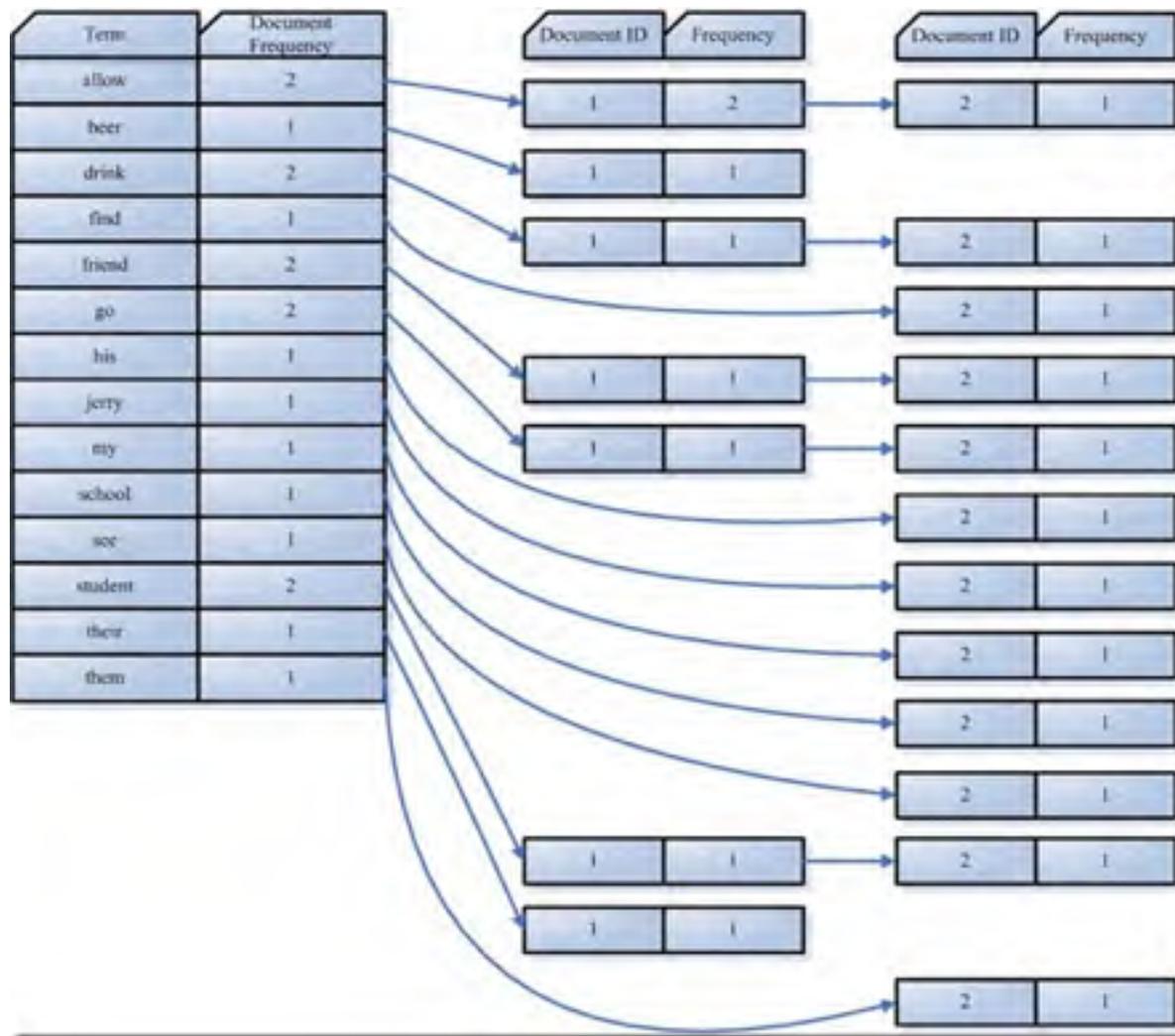
Lucene原理

- 索引创建indexing



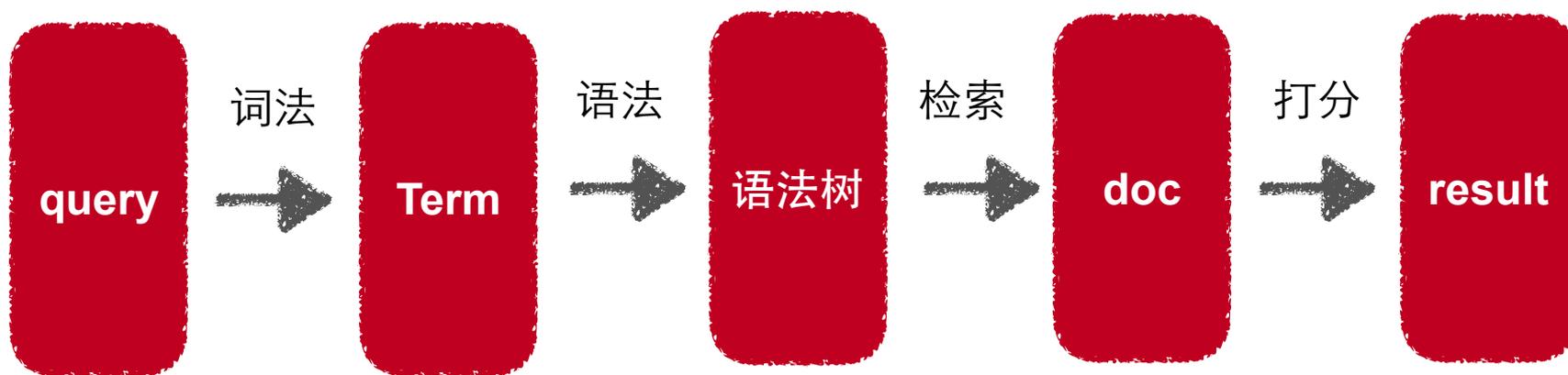
Lucene原理

- 倒排索引表



Lucene原理

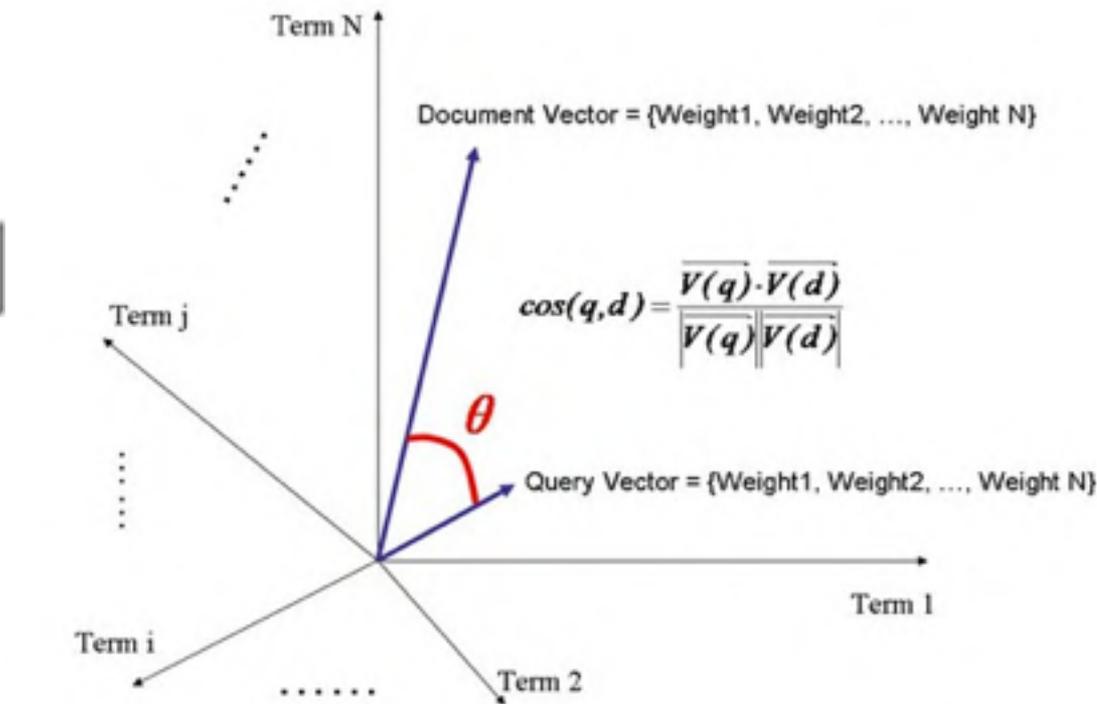
- 索引查询



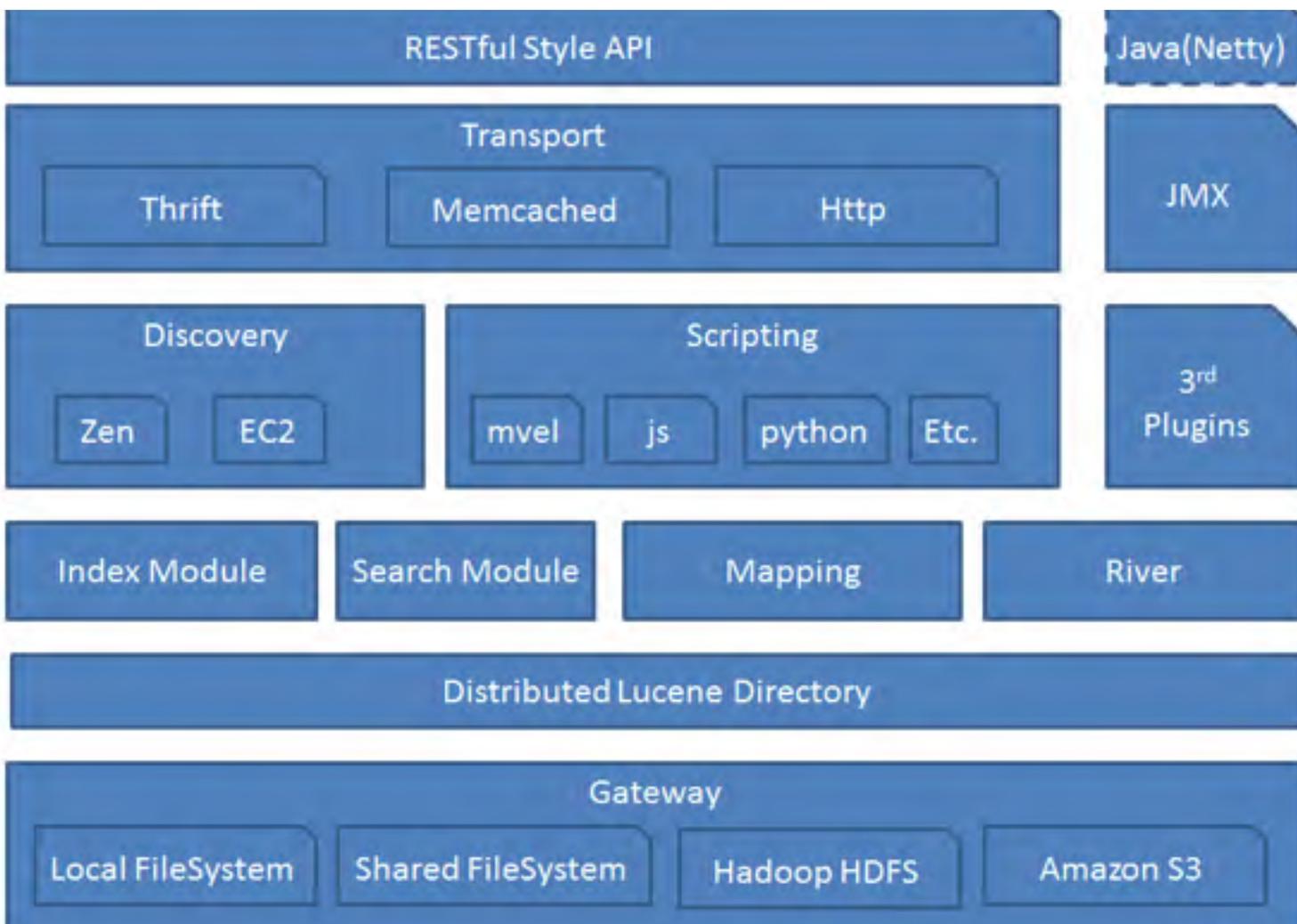
Lucene原理

- 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

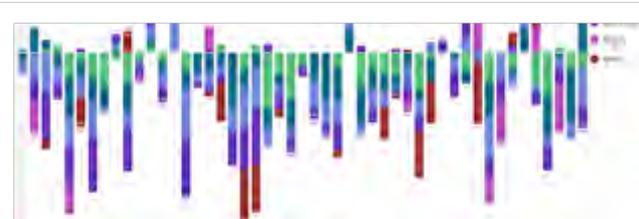
- 分词插件
- 同步插件
- 数据传输插件
- 脚本插件
- Alert
- Shield

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.



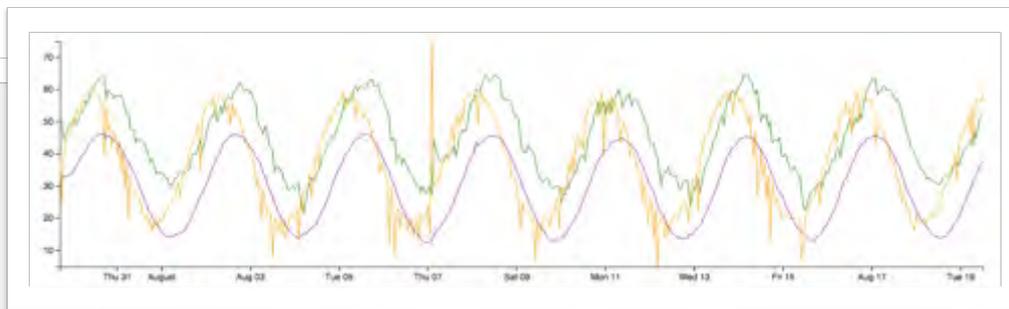
ElasticSearch-Pipeline Aggregations

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

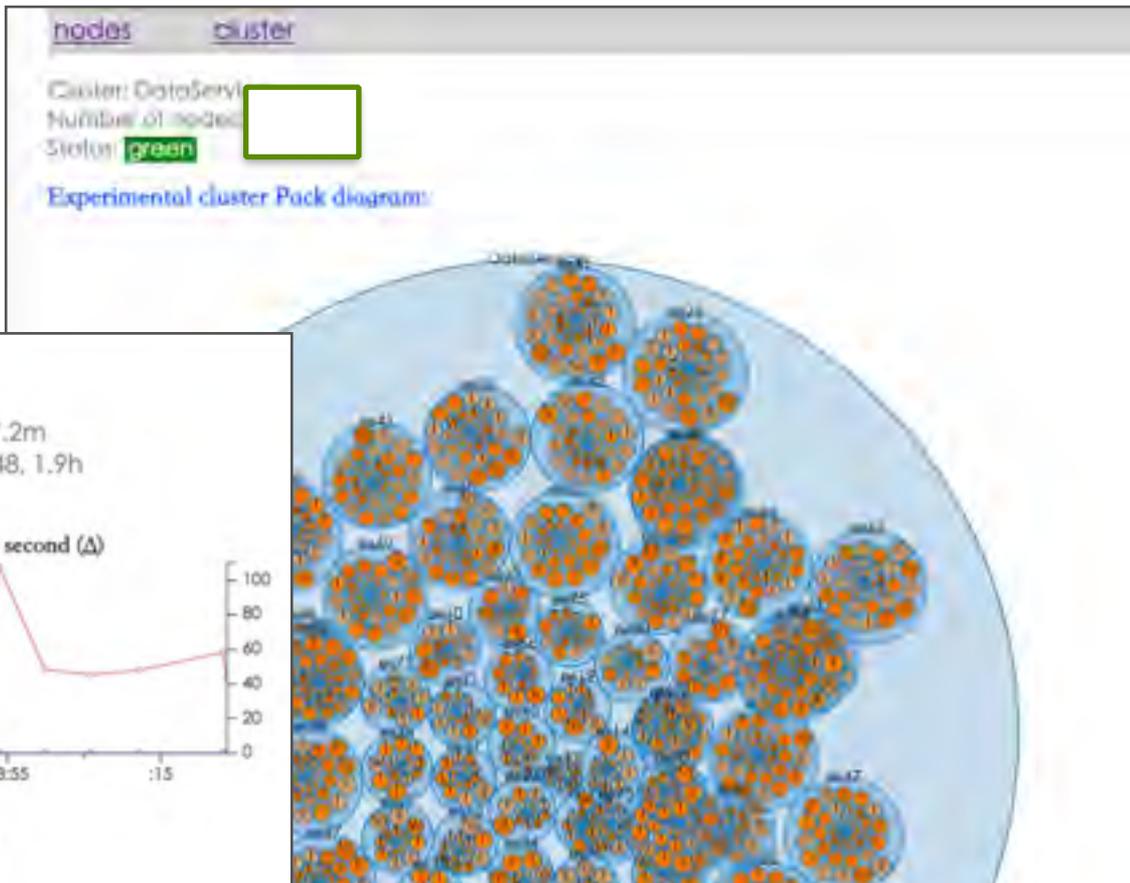
The screenshot displays the ElasticSearch Kopf interface. At the top, there are filters for indices and nodes. Below the filters, a table lists indices: 'dataservice', 'followers_v2', and 'influence_v1'. Each index row includes details like shards, docs, and size. Below the index rows, a list of nodes (es10 to es16) is shown. Each node row includes a star icon, the node name, and links for heap, disk, cpu, and load. The 'followers_v2' index column shows green boxes with numbers (2, 18, 17, 10) indicating the number of shards on each node.

Index	Shards	Docs	Size
dataservice	5 * 2	940,573	2.45GB
followers_v2	30 * 2	1,088,755,175	1.11TB
influence_v1	5 * 2	208,000	122.73GB

Node	Heap	Disk	CPU	Load	Shards
es10	heap	disk	cpu	load	2
es11	heap	disk	cpu	load	18
es12	heap	disk	cpu	load	17
es13	heap	disk	cpu	load	
es14	heap	disk	cpu	load	
es15	heap	disk	cpu	load	10
es16	heap	disk	cpu	load	

ElasticSearch管理和监控工具

- big desk
- 集群整体和流量情况

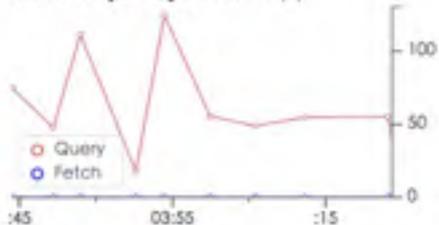


Indices

Docs count: 1 [redacted]
Docs deleted: [redacted] 6

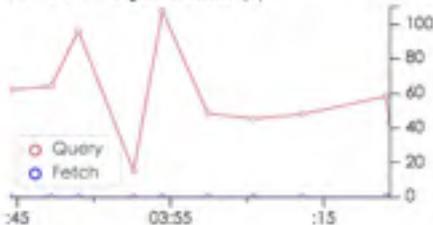
Flush: 2839, 17.2m
Refresh: 268648, 1.9h

Search requests per second (Δ)



Query: 4759877
Fetch: 7450

Search time per second (Δ)



Query: 17.4h
Fetch: 1.9m

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 gender > 0 and os > "500" and sc in ("0001009","0002036","0016030","...") or bs>585 and gender > 0`
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

5'

1000亿

单次

并发5个

并发10个

ElasticSearch

19005ms

21005ms

27736ms

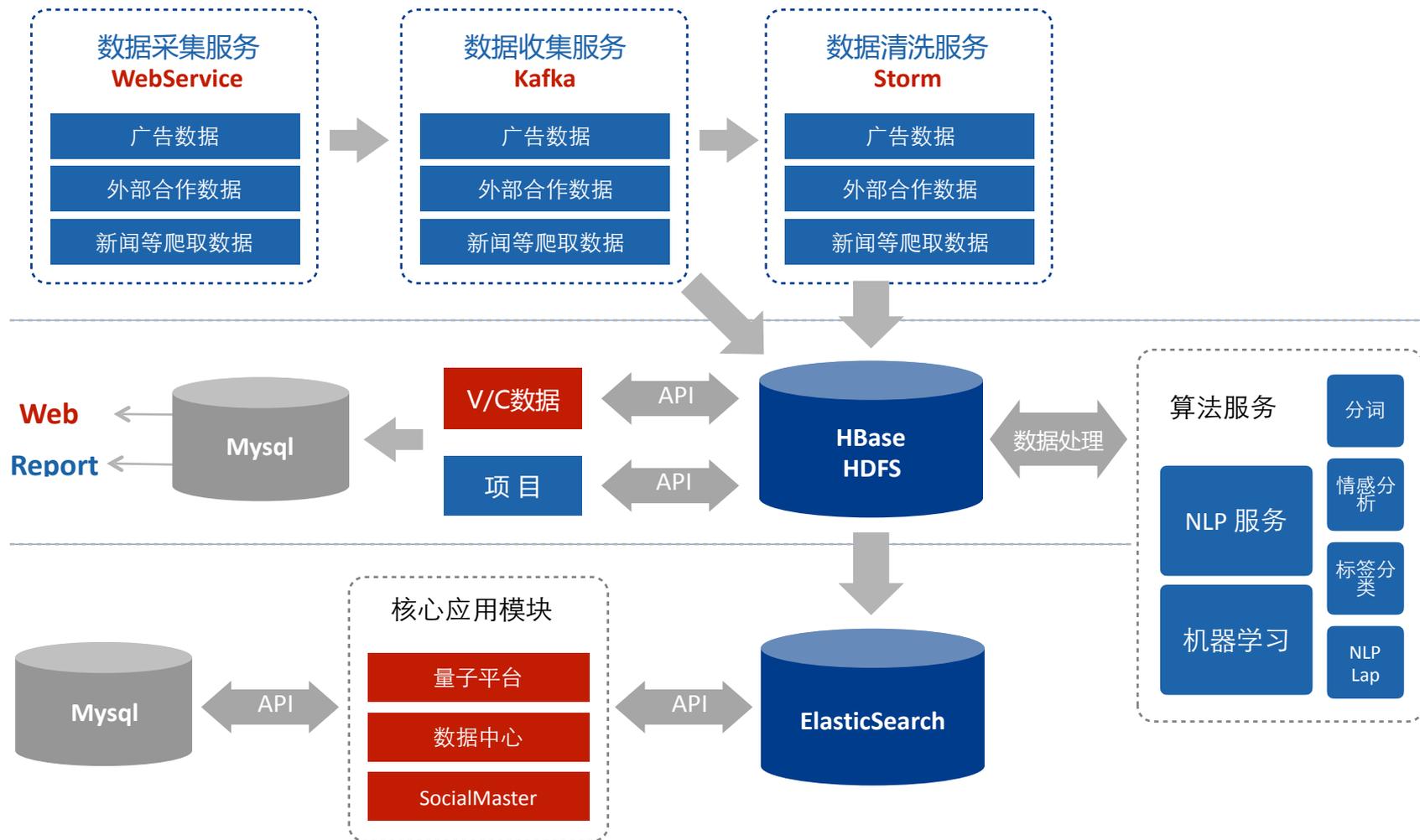
Pinot

19019ms

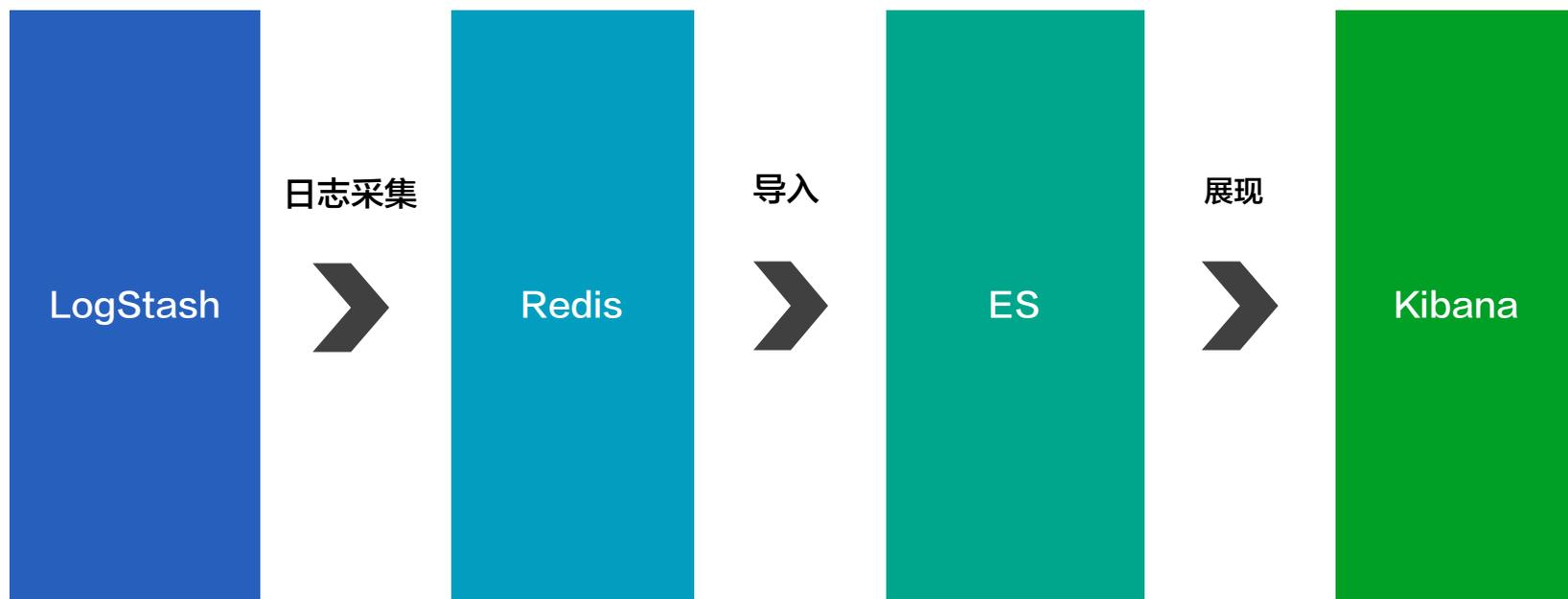
failed

failed

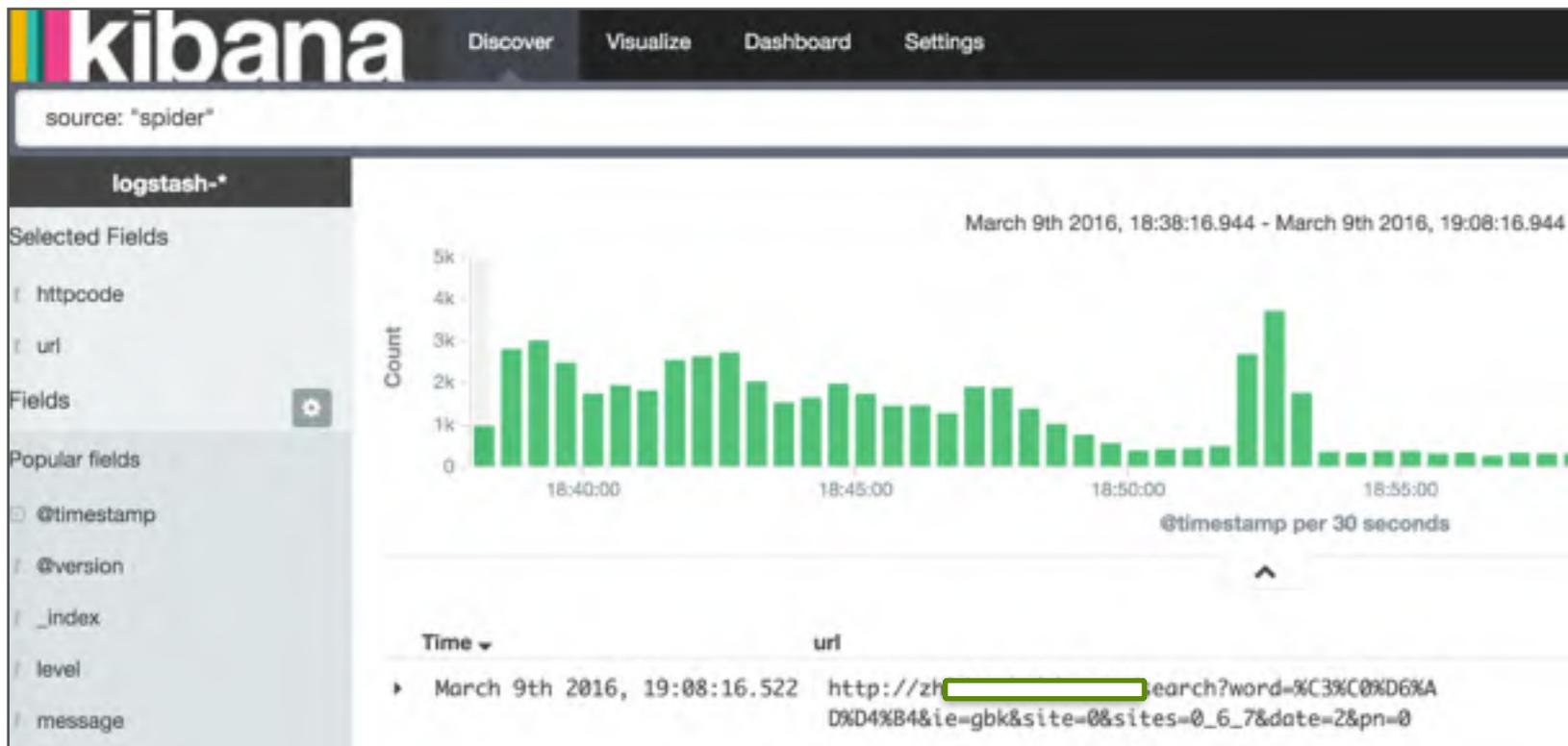
ElasticSearch数字营销案例



ELK实践



ELK实践



Kibana-Discover

- 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

