



How to use Elastic Stack?

白凡@Sunlands

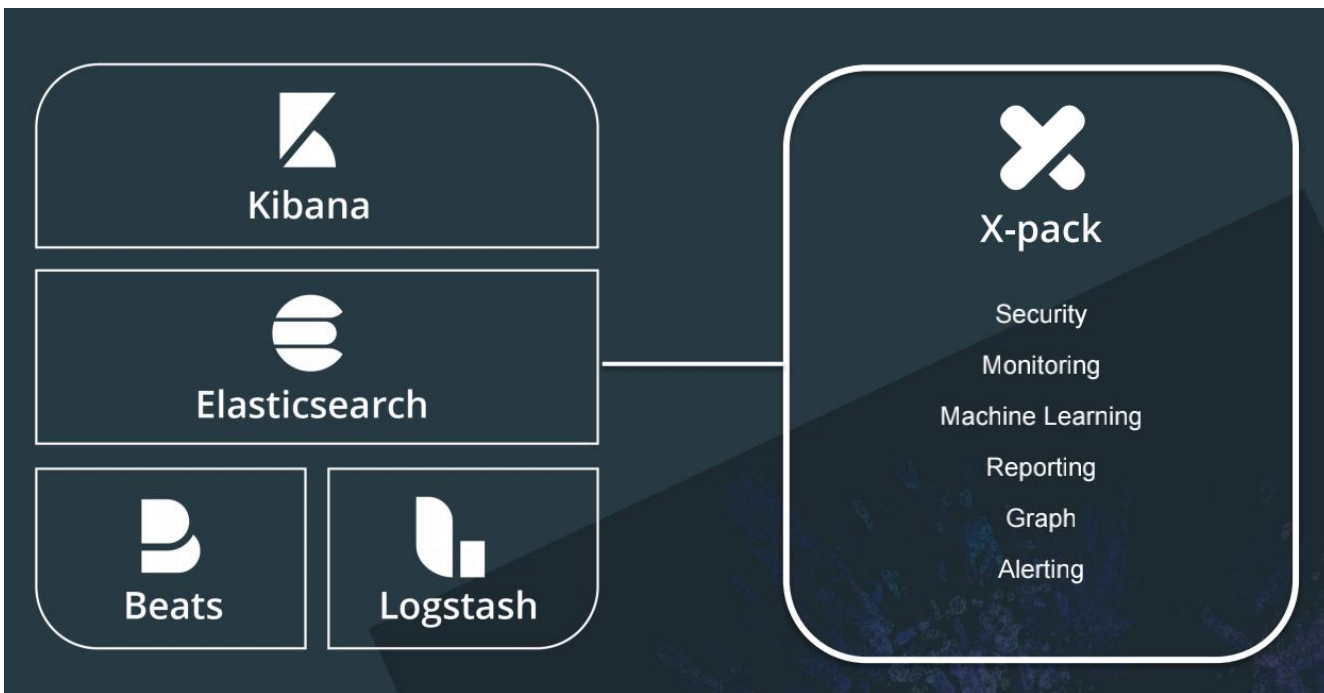
Nov 4, 2017

主要内容

- What is Elastic Stack?
- How to use ES?
- OneIndex
- Sunlands ES Roadmap

What is Elastic Stack?

Elastic Stack是整个Elastic框架集合，其中主要成员包含：
Elasticsearch, Beats, Logstash, Kibana以及X-pack（收费产品，不开源）



What is Elastic Stack?



Elasticsearch

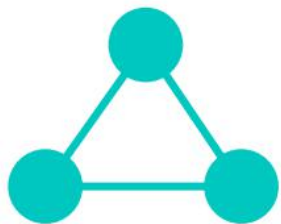
Elasticsearch全文检索功能基于Apache Lucence，其相关度计算算法依据：
词条频度/倒排文档频度(TF/IDF)

词条频度(Term Frequency)词条在当前文档中越频繁的话，那么权重就越高。在一个字段中出现了 5 次的词条应该比只出现了 1 次的文档更加相关。

字段长度归约(Field-length Norm)字段越短，那么其权重就越高。如果一个词条出现在较短的字段，如 **title** 字段中，那么该字段的内容相比更长的 **body** 字段而言，更有可能是关于该词条的。

Elasticsearch

Platform around a **distributed** data store



Distributed
Scaleable
Highly Available

```
{  
  job: "developer",  
  like: "elasticsearch"  
}
```

API
JSON
Language clients



Search
Aggregations
Geospatial

How to use ES?

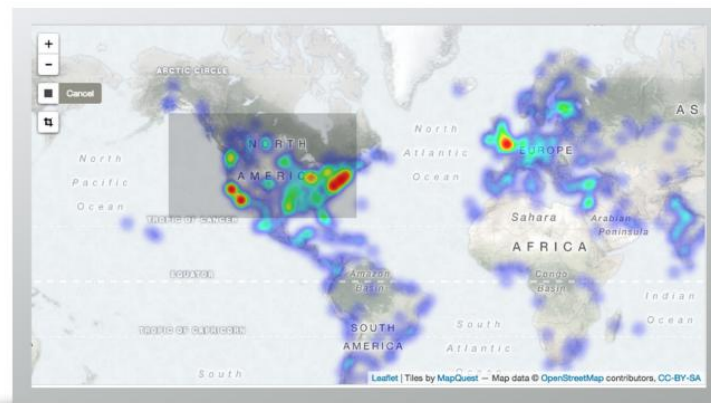
ELK

- 目前Elastic Stack最广泛的应用场景之一，用于实时收集监控日志数据。
- Beats/Logstash: 对服务器性能日志，业务日志等进行采集。
- Elasticsearch: 接收采集来的日志数据，并以文件形式写入磁盘存储。
- Kibana: 数据可视化平台，对ES数据进行查询展示，具有丰富的Dashboard模板。

How to use ES?

ELK

Visualize and Explore



Easily create bar charts, line and scatter plots, histograms, pie charts, and maps

How to use ES?

Search

- 基于Elasticsearch的全文检索系统，也是ES目前重要的应用场景。
- 丰富的查询API，支持RESTful web接口，Java Client API
- 安装搭建方便，配置相对简单
- 分布式存储结构，集群可轻松水平拓展
- 功能全面的分词系统，可满足大部分检索需求
- 倒排索引存储，保证搜索性能要求
- head,cebro等插件齐全
- 社区完善，有问必答

How to use ES?

Search

The screenshot shows the GitHub repository page for Elasticsearch. The top navigation bar includes links for Pull requests, Issues, Marketplace, and Explore. Below the repository list, there are filters for Repositories (16K), Code (1M), Commits (207K), Issues (102K), Wikis (9K), and Users (175). The search results are sorted by 'Best match' and show 16,175 repository results. The first result is 'elastic/elasticsearch', an Open Source, Distributed, RESTful Search Engine, written in Java, with 26.3k stars. The second result is 'dockerfile/elasticsearch', an ElasticSearch Dockerfile for trusted automated Docker builds, with 417 stars. The third result is 'docker-library/elasticsearch', a DEPRECATED repository, with 284 stars. The fourth result is 'mispnp/elasticsearch', Supporting assets for 'Run Elasticsearch on Azure' content, with 418 stars.

elasticsearch Pull requests Issues Marketplace Explore

Repositories 16K Code 1M Commits 207K Issues 102K Wikis 9K Users 175

16,175 repository results Sort: Best match

elastic/elasticsearch Java ★ 26.3k
Open Source, Distributed, RESTful Search Engine
java search-engine elasticsearch
Apache-2.0 license Updated 41 minutes ago

dockerfile/elasticsearch ★ 417
ElasticSearch Dockerfile for trusted automated Docker builds.
MIT license Updated on 15 Nov 2016

docker-library/elasticsearch Shell ★ 284
DEPRECATED; see <https://www.elastic.co/guide/en/elasticsearch/reference/current/docker.html>
Apache-2.0 license Updated 14 days ago

mispnp/elasticsearch Java ★ 418
Supporting assets for "Run Elasticsearch on Azure" content

The screenshot shows the Bilibili website interface. The top navigation bar includes links for 首页 (Home), 直播 (Live), 分类 (Categories), 游戏 (Games), 鱼吧 (Fish Bar), and 鱼乐盛典 (Fish Joy Festival). The search bar is set to '绝地求生' (PUBG). The main content area shows search results for '绝地求生'. The '综合' (General) tab is selected, showing a list of related videos. The first video is '绝地求生 1950 直播', featuring a character in a white suit. Below it, there are several '正在直播' (Live Now) thumbnails for various PUBG-related content, including '绝地求生赛事', '绝地求生教程', '绝地求生直播', '绝地求生大神', '绝地求生新手', '绝地求生老手', and '绝地求生大神'. The '相关视频' (Related Videos) section shows a list of related videos, including '绝地求生 11923', '绝地求生 5631', '绝地求生 12003', '绝地求生 8530', and '绝地求生 12000'.

首页 直播 分类 游戏 鱼吧 鱼乐盛典

绝地求生 下载 12

综合 直播 主播 视频

相关分类

绝地求生 1950 直播

相关直播

绝地求生赛事 4.8万 人正在观看

绝地求生教程 2271 人正在观看

绝地求生直播 1734 人正在观看

绝地求生大神 24 人正在观看

绝地求生新手 66 人正在观看

绝地求生老手 1503 人正在观看

绝地求生大神 26 人正在观看

绝地求生大神 0 人正在观看

相关视频

绝地求生 11923 47

绝地求生 5631 41

绝地求生 12003 60

绝地求生 8530 118

绝地求生 12000 51

How to use ES?

数据可视化

- Elasticsearch具有良好的可读性，足以应付绝大多数的高并发请求，并能保持较好的接口性能
- 查询API丰富，可通过简单逻辑运算，实现多维度高复杂的数据查询
- Kibana具有丰富的Dashboard模板，简单配置，即可实现。也可通过ECharts等插件，自定义Dashboard
- 轻松实现数据的环比，同比，实时计算，趋势图等方案

How to use ES?

数据可视化

查询日期: 20170223 到 20170225

查询分区: 选择分区

筛选条件: 与 有效观看人数 大于 100 与 UV 大于 100 添加条件

刷新 重置

视频ID	直播间名称	日期	开播时长 (分钟)	UV	有效观看人数	人为观看时长 (分钟)	取消人数	新增关注用户数	鱼翅充值 (元)	鱼翅ARPU	付费率
32167	临空世界王豪!!!	20170223									
68172	斗鱼直播王豪, 三天不直播!	20170223									
308211	LOL精彩团战, 团战直播!	20170223									
507204	英雄联盟! 英雄联盟!	20170223									
354337	D1王者荣耀! 王者荣耀!	20170223									
62164	斗鱼第一主播! 斗鱼第一主播!	20170223									
38956	斗鱼第一主播! 斗鱼第一主播!	20170223									
521158	英雄联盟! 英雄联盟!	20170223									
1423120	英雄联盟! 英雄联盟!	20170223									



How to use ES?

风控系统

- 风险控制是所有公司系统不可忽视的环节
- 平台用户或咨询师作弊行为监控
- 平台系统对可能的存在风险问题预警
- 平台系统风险出现后损失挽回及进一步扩散控制
- 平台系统风险保留，记录留存

How to use ES?

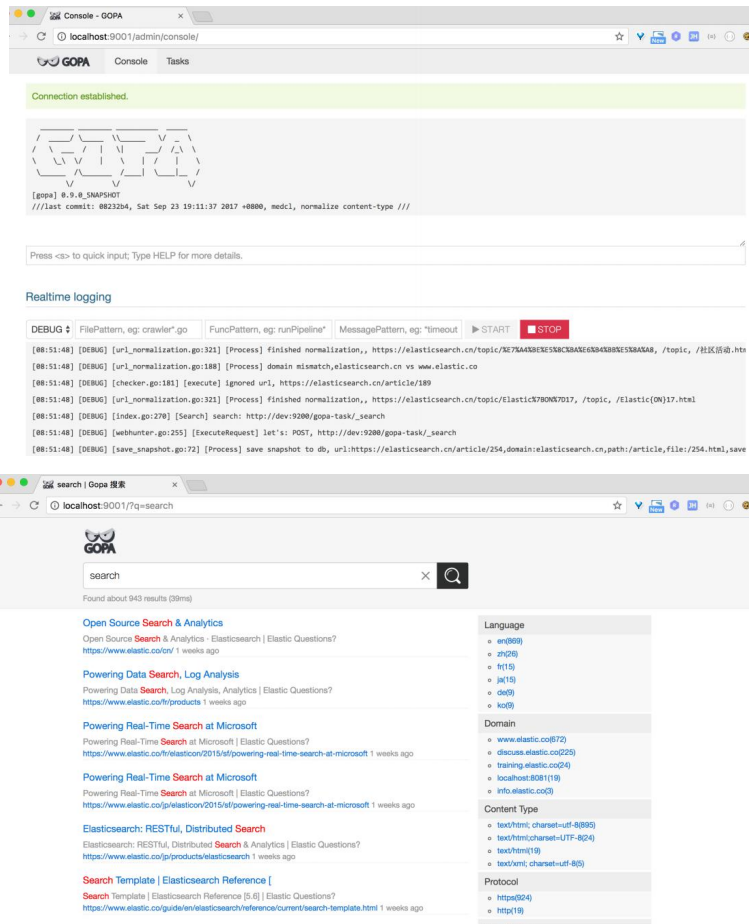
风控系统



How to use ES?

舆情系统

- 舆情系统数据可采用ES做中间存储查询
- GOPA, 相比于Nutch,Scrapy更加轻量级, 搭建使用简单, 无缝连接ES
- 亦可使用Nutch等爬虫工具, 写入ES



How to use ES?

自定义词库

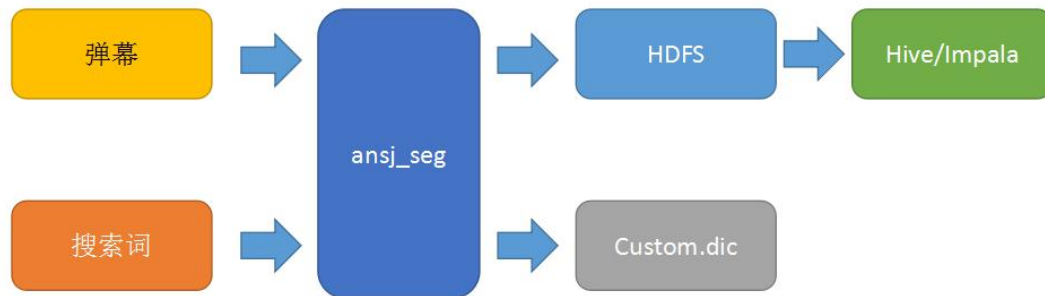
- 生成符合自己生产环境的自定义词库
- 自定义词库可用于全站关键词分析，用户关键词分析以及全站热门词汇分析等
- 生成的自定义词库用于ES分词，提高检索系统可用性
- 用户画像



How to use ES?

自定义词库

- 采用Ansj中文分词器
- HDFS，离线统计分析
- Custom.dic，供索引分词



How to use ES?

推荐系统

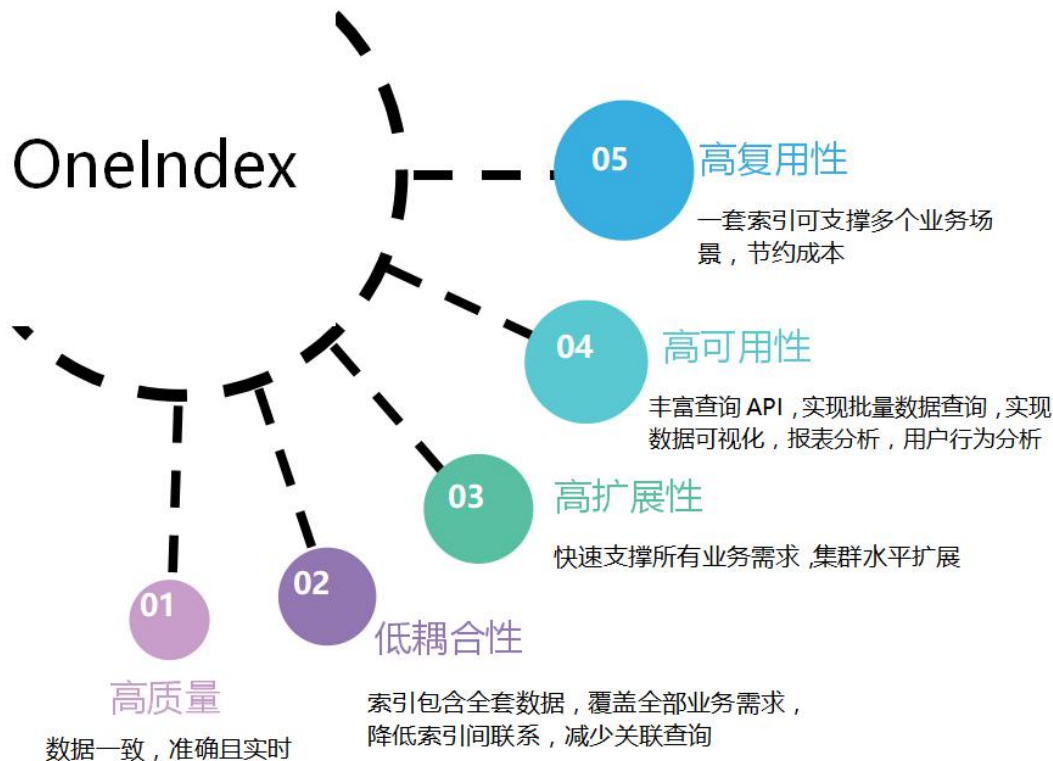
- 制定有效推荐规则
- 用户画像数据写入ES
- 通过ES boost权重配置，可以轻松实现推荐业务
- 无需重写大量逻辑代码
- 高并发环境下，能保证系统的可用性，可读性



OneIndex?!

OneIndex?!

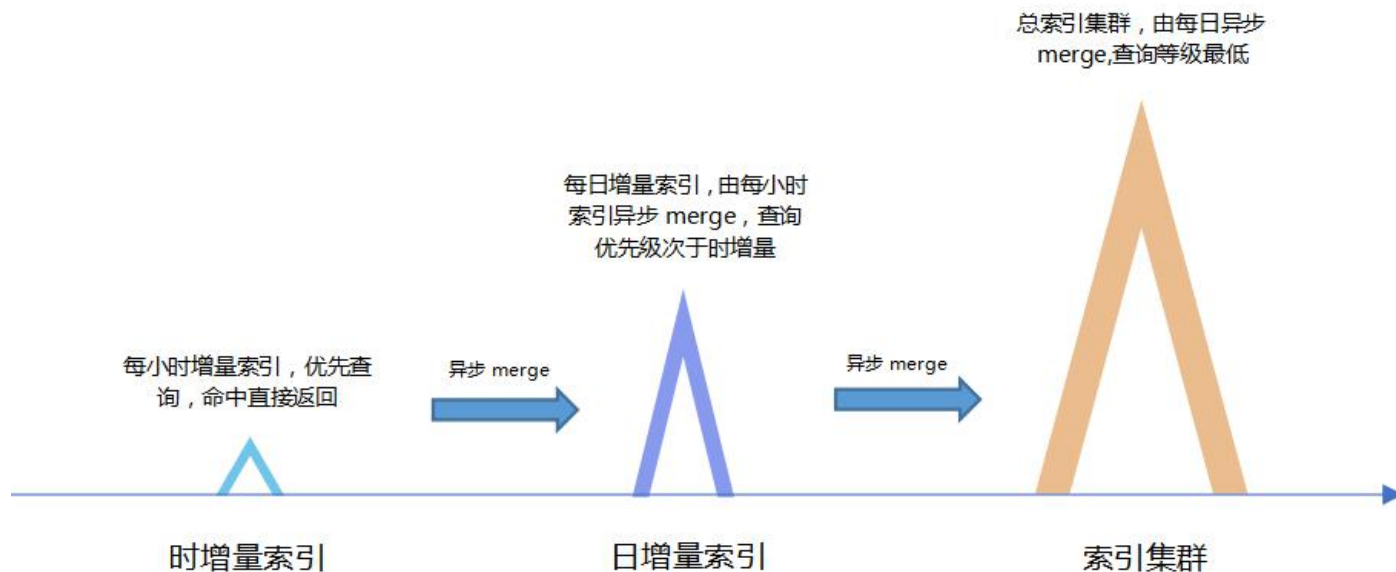
- 概念来源于阿里“OneData”
- 大数据之痛：
- 高人力成本、数据错误、浪费资源、杂乱无章、效率低下、批量多维度查询问题
- 基于Elasticsearch的“OneIndex”





OneIndex?!

- 海量数据检索方案



Sunlands ES Roadmap



Thanks!