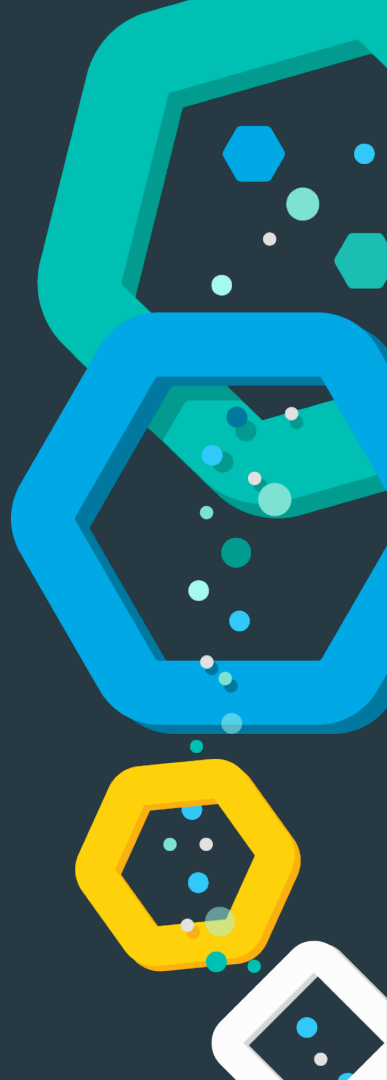
又一个爬虫

曾勇（Medcl）



什么是爬虫

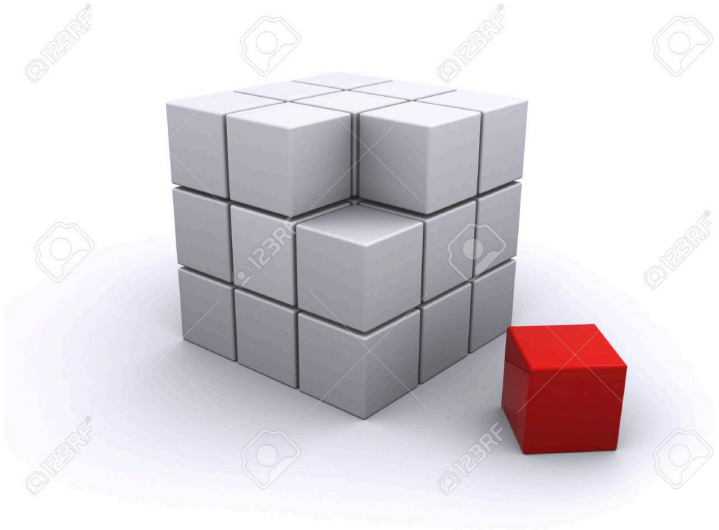
- 又叫 Robot、Bot 或 Crawler
- 简单来说
  - 自动探索网站
  - 帮你访问整个站点
  - 自动为你收集网站信息
  - 自动更新
  - 抽取处理网页内容
  - 存储索引和快照
  - 。。。

为什么造这个轮子?

现在已经有很多开源的爬虫了: *Scrapy*，*Nutch*，*Heritrix* 等等．*[1,2,3]*

但是!

- 大多仅仅是一个爬虫框架！*ES* 与 *Lucene*
- 需要熟悉各种与爬取任务本身无关的知识!
- 开发和部署的环境复杂，痛苦!
- 太重了!
- 分布式、管理、监控、扩展复杂!
- 结果就是一大堆凌乱无法维护的脚本，或是一大堆技术拼凑的大杂烩。



1.http://bigdata-madesimple.com/top-50-open-source-web-crawlers-for-data-mining/
2.https://github.com/BruceDone/awesome-crawler
3.https://gitee.com/explore/starred/spider

所以

- 狗爬，Gopa

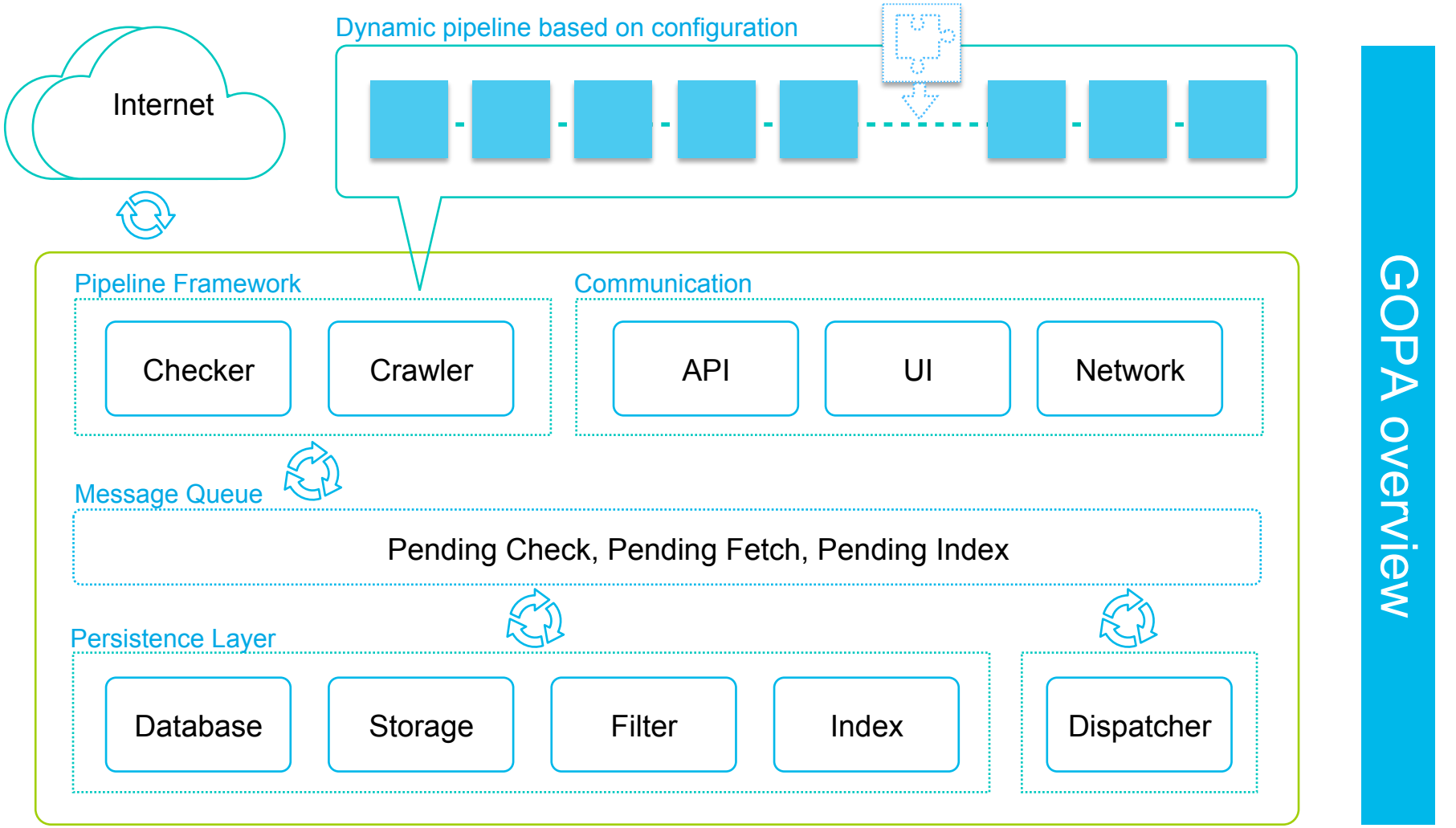
  Golang + pá chóng（爬虫）

https://github.com/infinitbyte/gopa

目标

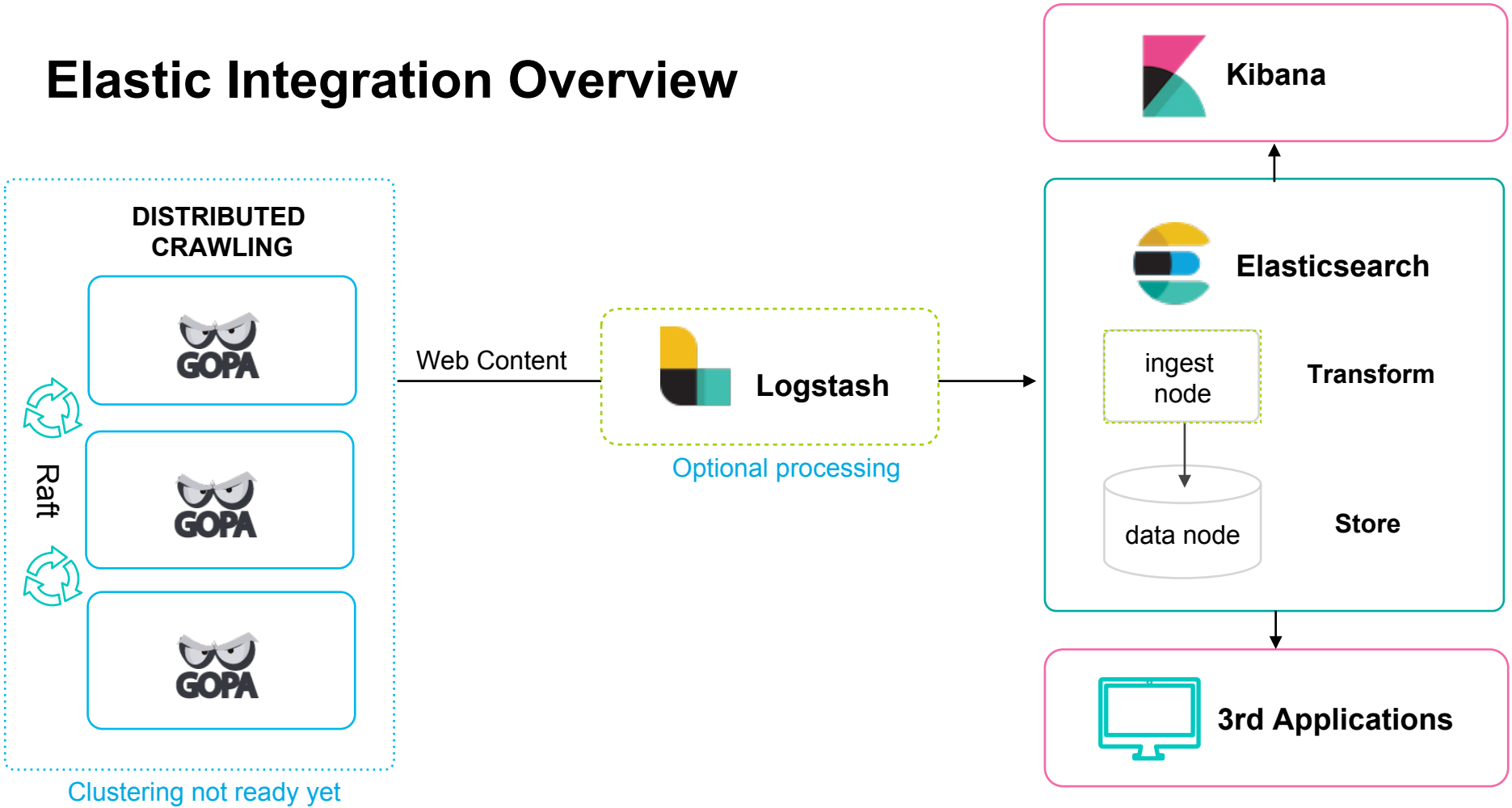
- 轻量级，内存占用 < 100MB
- 易于部署，无运行时和环境依赖
- 方便使用，无需编程和脚本技能
- 开箱即用
- 提供 RESTful API 和 UI
- 简单，可伸缩，可扩展

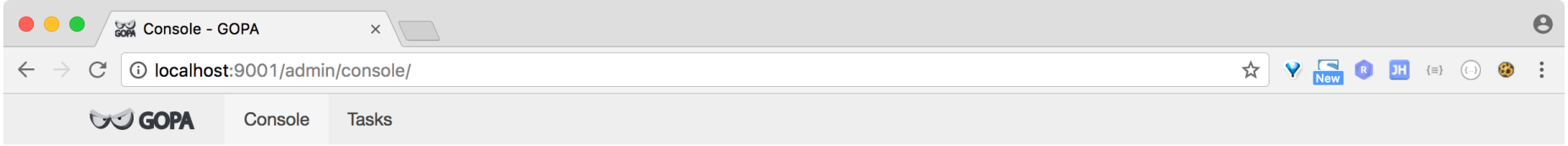
DEMO



GOPA overview

Elastic Integration Overview





Connection established.

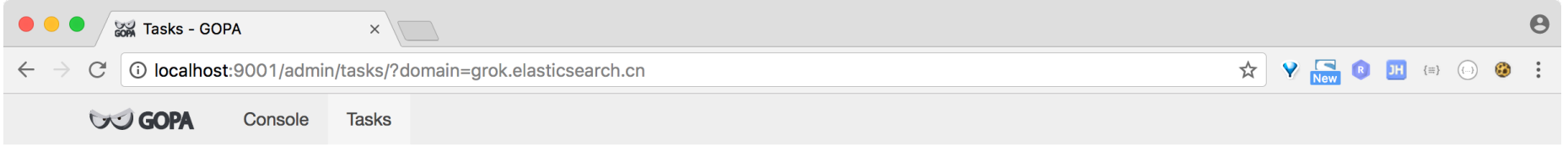
```
  _______ _______ _________  ____
 /  _____/ \_____  \\_____   \/  _  \
/   \  ___  /   |  \|     __/  /_\  \
\    \_\  \/    |   \    |  /    |   \
 \_____  /\_______  /____|  \____|__  /
       \/         \/               \/
[gopa] 0.9.0_SNAPSHOT
///last commit: 08232b4, Sat Sep 23 19:11:37 2017 +0800, medcl, normalize content-type ///
```

Press <s> to quick input; Type HELP for more details.

Realtime logging

DEBUG | FilePattern, eg: crawler*.go | FuncPattern, eg: runPipeline* | MessagePattern, eg: *timeout | ▶ START | ■ STOP

```
[08:51:48] [DEBUG] [url_normalization.go:321] [Process] finished normalization,, https://elasticsearch.cn/topic/%E7%A4%BE%E5%8C%BA%E6%B4%BB%E5%8A%A8, /topic, /社区活动.htm
[08:51:48] [DEBUG] [url_normalization.go:188] [Process] domain mismatch,elasticsearch.cn vs www.elastic.co
[08:51:48] [DEBUG] [checker.go:181] [execute] ignored url, https://elasticsearch.cn/article/189
[08:51:48] [DEBUG] [url_normalization.go:321] [Process] finished normalization,, https://elasticsearch.cn/topic/Elastic%7BON%7D17, /topic, /Elastic{ON}17.html
[08:51:48] [DEBUG] [index.go:270] [Search] search: http://dev:9200/gopa-task/_search
[08:51:48] [DEBUG] [webhunter.go:255] [ExecuteRequest] let's: POST, http://dev:9200/gopa-task/_search
[08:51:48] [DEBUG] [save_snapshot.go:72] [Process] save snapshot to db, url:https://elasticsearch.cn/article/254,domain:elasticsearch.cn,path:/article,file:/254.html,save
```



Total 5

www.elasticsearch.cn(10) es-guide-preview.elasticsearch.cn(1) elasticsearch.cn(2497) conf.elasticsearch.cn(2) grok.elasticsearch.cn(1)

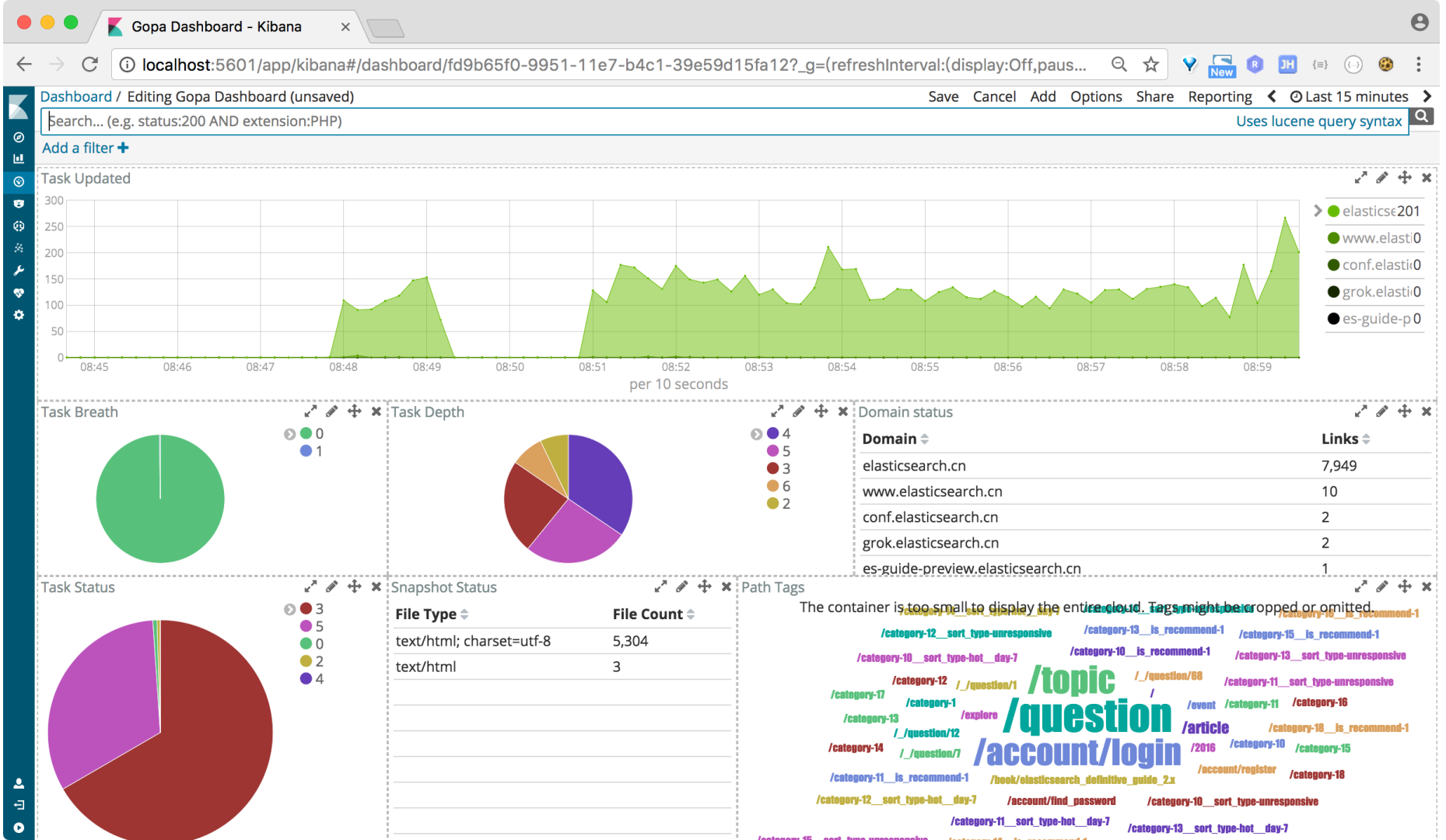
Total 1

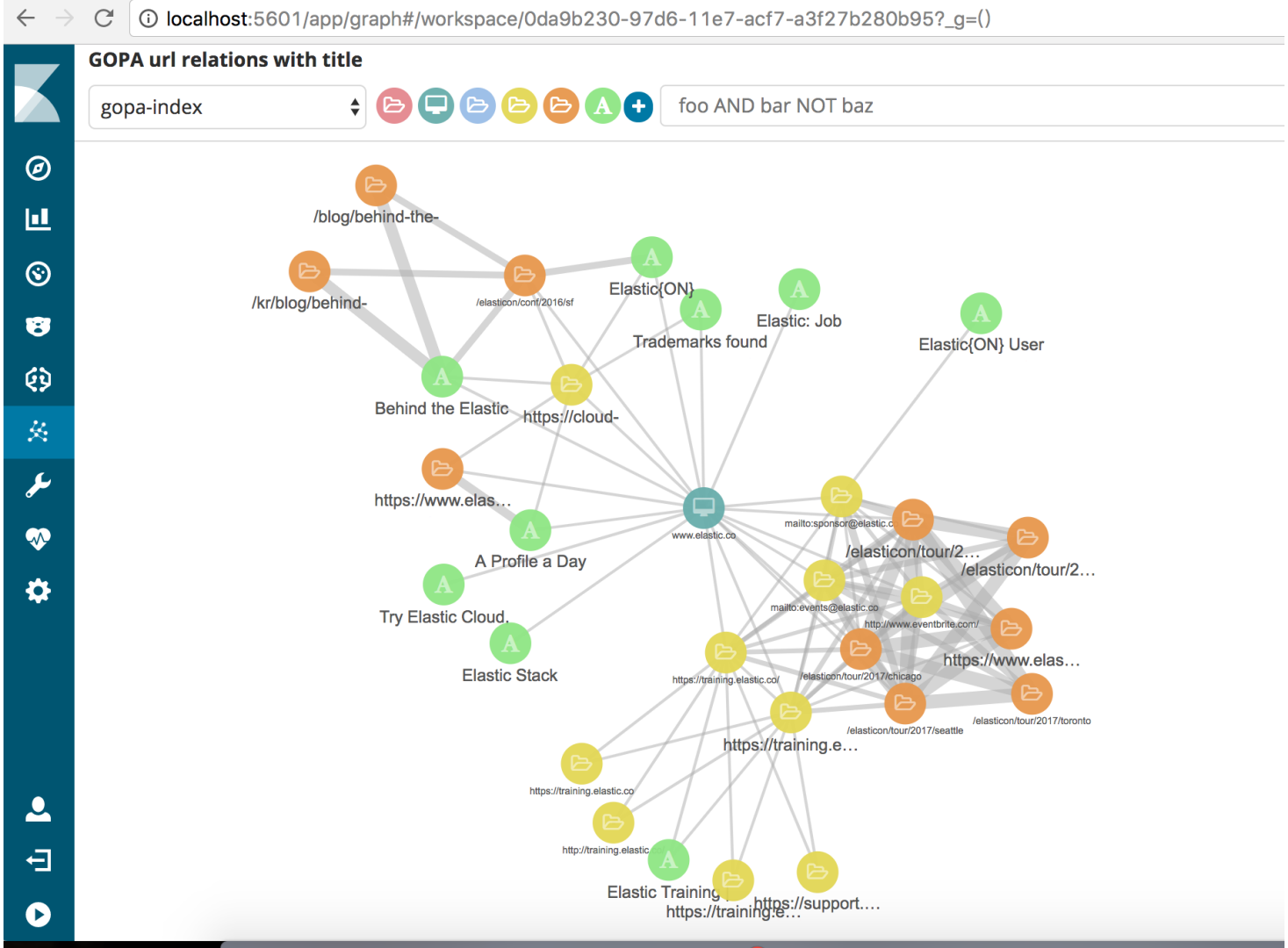
URL	LastUpdate	NextCheck	Status
http://grok.elasticsearch.cn	~1 min ago	~24 hours later	success

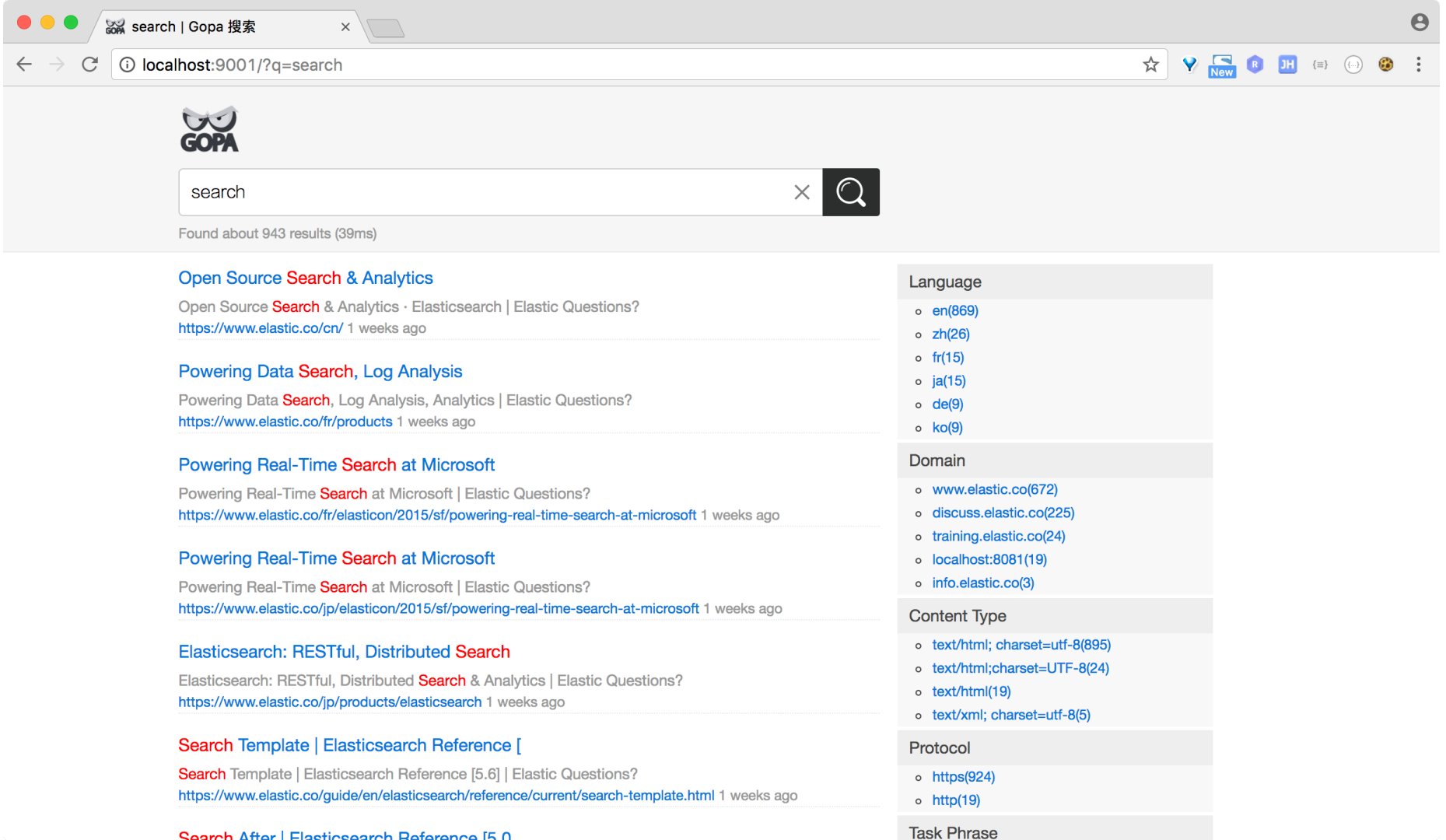
1

GitHub | Issues | Releases | Changelog

GOPA


Gopa Dashboard - Kibana
localhost:5601/app/kibana#/dashboard/fd9b65f0-9951-11e7-b4c1-39e59d15fa12?_g=(refreshInterval:(display:Off,paus...
Dashboard / Editing Gopa Dashboard (unsaved)
Save Cancel Add Options Share Reporting Last 15 minutes
Search... (e.g. status:200 AND extension:PHP)
Uses lucene query syntax
Add a filter
Task Updated
300 250 200 150 100 50 0
08:45 08:46 08:47 08:48 08:49 08:50 08:51 08:52 08:53 08:54 08:55 08:56 08:57 08:58 08:59
per 10 seconds
elasticse 201
www.elasti 0
conf.elasti 0
grok.elasti 0
es-guide-p 0
Task Breath
0
1
Task Depth
4
5
3
6
2
Domain status
Domain
Links
elasticsearch.cn 7,949
www.elasticsearch.cn 10
conf.elasticsearch.cn 2
grok.elasticsearch.cn 2
es-guide-preview.elasticsearch.cn 1
Task Status
3
5
0
2
4
Snapshot Status
File Type
File Count
text/html; charset=utf-8 5,304
text/html 3
Path Tags
The container is too small to display the entire cloud. Tags might be cropped or omitted.
/topic
/question
/account/login
/article
/explore
/event
/2016
/book/elasticsearch_definitive_guide_2.x
/account/register
/account/find_password
/question/68
/question/1
/question/12
/question/7
/category-1
/category-10
/category-11
/category-12
/category-13
/category-14
/category-15
/category-16
/category-17
/category-18
/category-10__sort_type-hot__day-7
/category-10__is_recommend-1
/category-10__sort_type-unresponsive
/category-11__is_recommend-1
/category-11__sort_type-hot__day-7
/category-11__sort_type-unresponsive
/category-12__sort_type-hot__day-7
/category-12__sort_type-unresponsive
/category-13__is_recommend-1
/category-13__sort_type-hot__day-7
/category-13__sort_type-unresponsive
/category-15__is_recommend-1
/category-18__is_recommend-1


localhost:5601/app/graph#/workspace/0da9b230-97d6-11e7-acf7-a3f27b280b95?_g=()
GOPA url relations with title
gopa-index
foo AND bar NOT baz
/blog/behind-the-
/kr/blog/behind-
/elasticon/conf/2016/sf
Elastic{ON}
Trademarks found
Elastic: Job
Elastic{ON} User
Behind the Elastic
https://cloud-
https://www.elas…
A Profile a Day
Try Elastic Cloud.
Elastic Stack
www.elastic.co
mailto:sponsor@elastic.c
/elasticon/tour/2…
/elasticon/tour/2…
mailto:events@elastic.co
http://www.eventbrite.com/
https://www.elas…
https://training.elastic.co/
/elasticon/tour/2017/chicago
/elasticon/tour/2017/seattle
/elasticon/tour/2017/toronto
https://training.e…
https://training.elastic.co
http://training.elastic.
Elastic Training
https://training.e…
https://support.…


search | Gopa 搜索
localhost:9001/?q=search
GOPA
search
Found about 943 results (39ms)
Open Source Search & Analytics
Open Source Search & Analytics · Elasticsearch | Elastic Questions?
https://www.elastic.co/cn/ 1 weeks ago
Powering Data Search, Log Analysis
Powering Data Search, Log Analysis, Analytics | Elastic Questions?
https://www.elastic.co/fr/products 1 weeks ago
Powering Real-Time Search at Microsoft
Powering Real-Time Search at Microsoft | Elastic Questions?
https://www.elastic.co/fr/elasticon/2015/sf/powering-real-time-search-at-microsoft 1 weeks ago
Powering Real-Time Search at Microsoft
Powering Real-Time Search at Microsoft | Elastic Questions?
https://www.elastic.co/jp/elasticon/2015/sf/powering-real-time-search-at-microsoft 1 weeks ago
Elasticsearch: RESTful, Distributed Search
Elasticsearch: RESTful, Distributed Search & Analytics | Elastic Questions?
https://www.elastic.co/jp/products/elasticsearch 1 weeks ago
Search Template | Elasticsearch Reference [
Search Template | Elasticsearch Reference [5.6] | Elastic Questions?
https://www.elastic.co/guide/en/elasticsearch/reference/current/search-template.html 1 weeks ago
Language
en(869)
zh(26)
fr(15)
ja(15)
de(9)
ko(9)
Domain
www.elastic.co(672)
discuss.elastic.co(225)
training.elastic.co(24)
localhost:8081(19)
info.elastic.co(3)
Content Type
text/html; charset=utf-8(895)
text/html;charset=UTF-8(24)
text/html(19)
text/xml; charset=utf-8(5)
Protocol
https(924)
http(19)
Task Phrase

谢谢！

https://github.com/infinitbyte/gopa