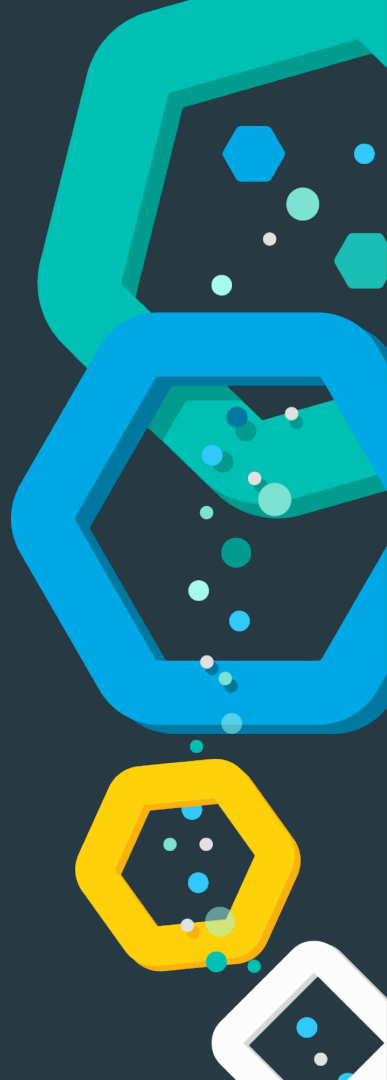
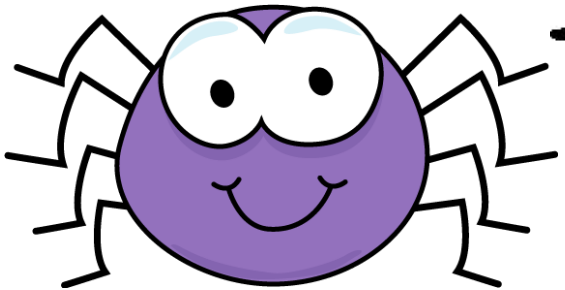
Yet Another Spider

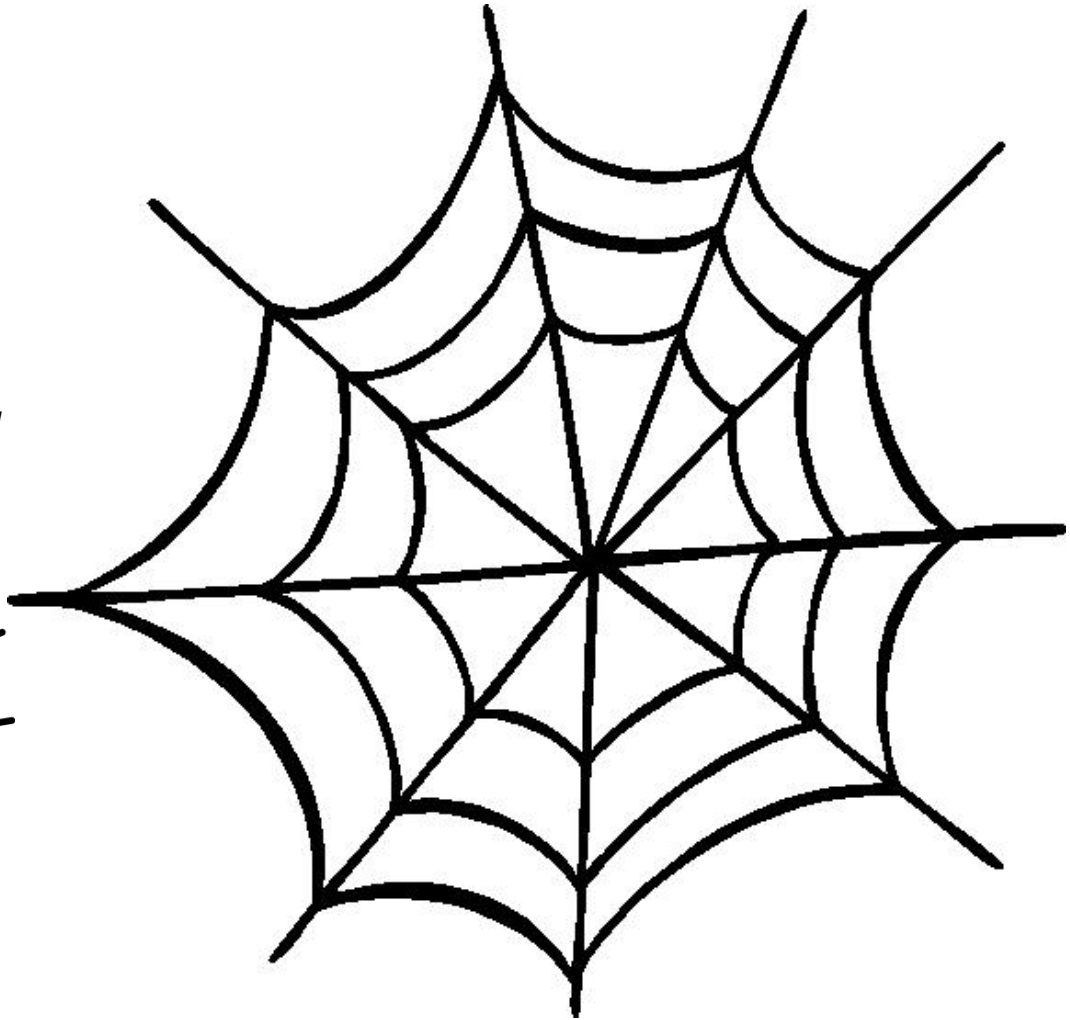
Medcl



What is a spider?

“Hey there ~ , i catch bugs and enjoy them”





Not this one, just kidding …

I think you already know

- Also known as Robot, Bot or Crawler
- It automatically discovery website
- Visit the whole website for you
- Collect web information for you
- Keeps a eye on the web and update it
- Store and Index web content for further process
- Every Search Engine have spider

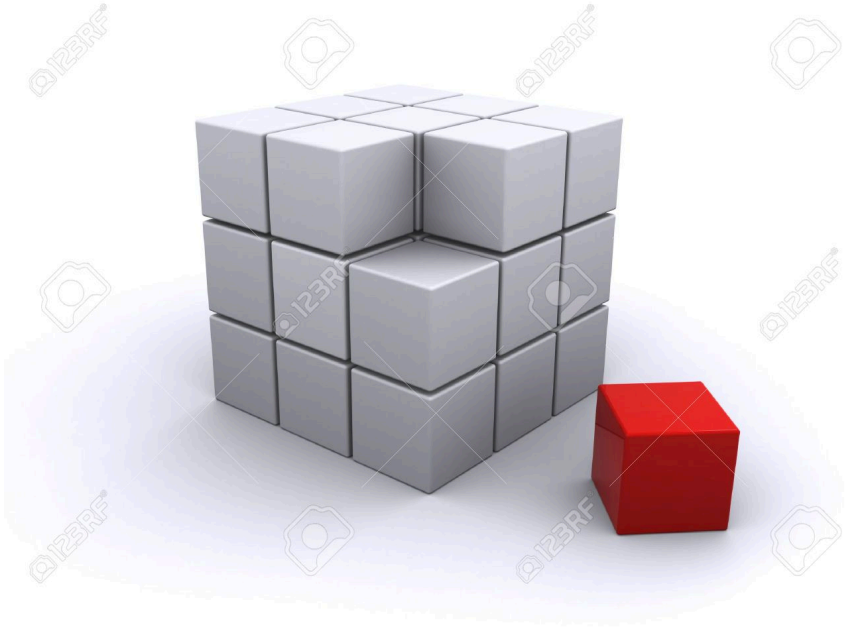
So, why reinvent a wheel?

There are so many OSS crawlers already, like: Scrapy, Nutch, Heritrix, etc. [1,2]

They are good for expert to use!

Just with a lot of "before" or "after" pain, generally they are good framework, but not good enough, not in a "elastic" way! — Medcl

Why not extend Logstash or Beats?



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

Yet another spider

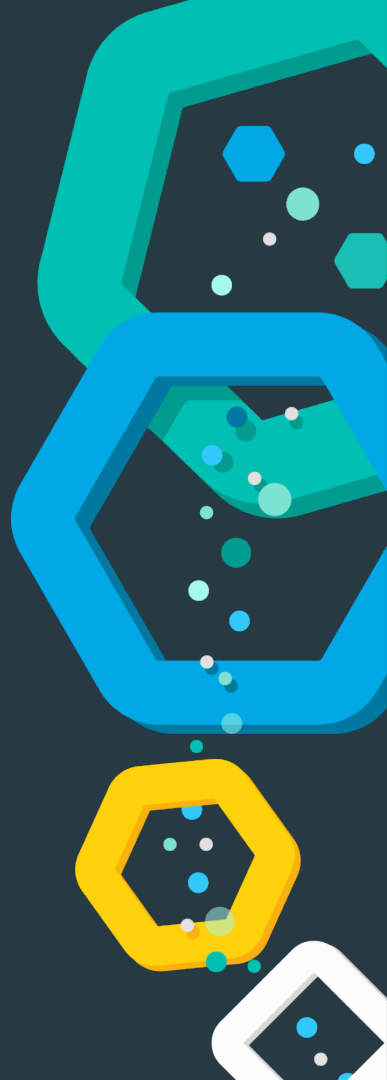
- Gopa

  Golang + pá chóng（爬虫）
- https://github.com/infinitbyte/gopa

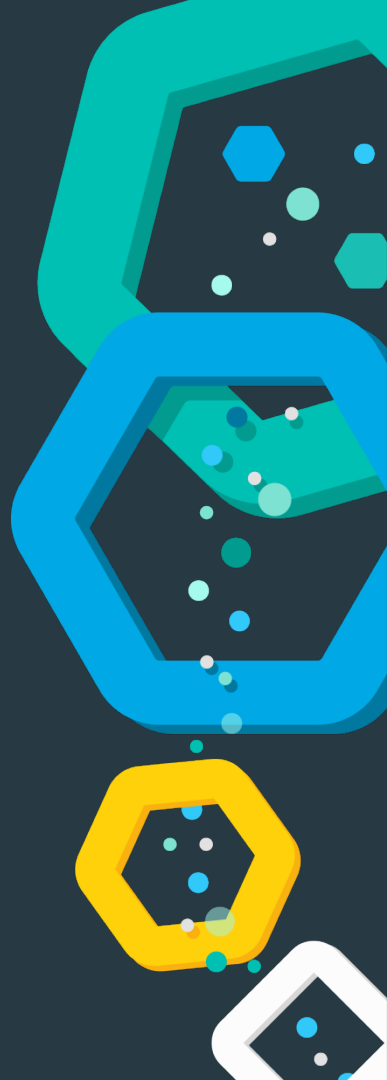
Goal of this project

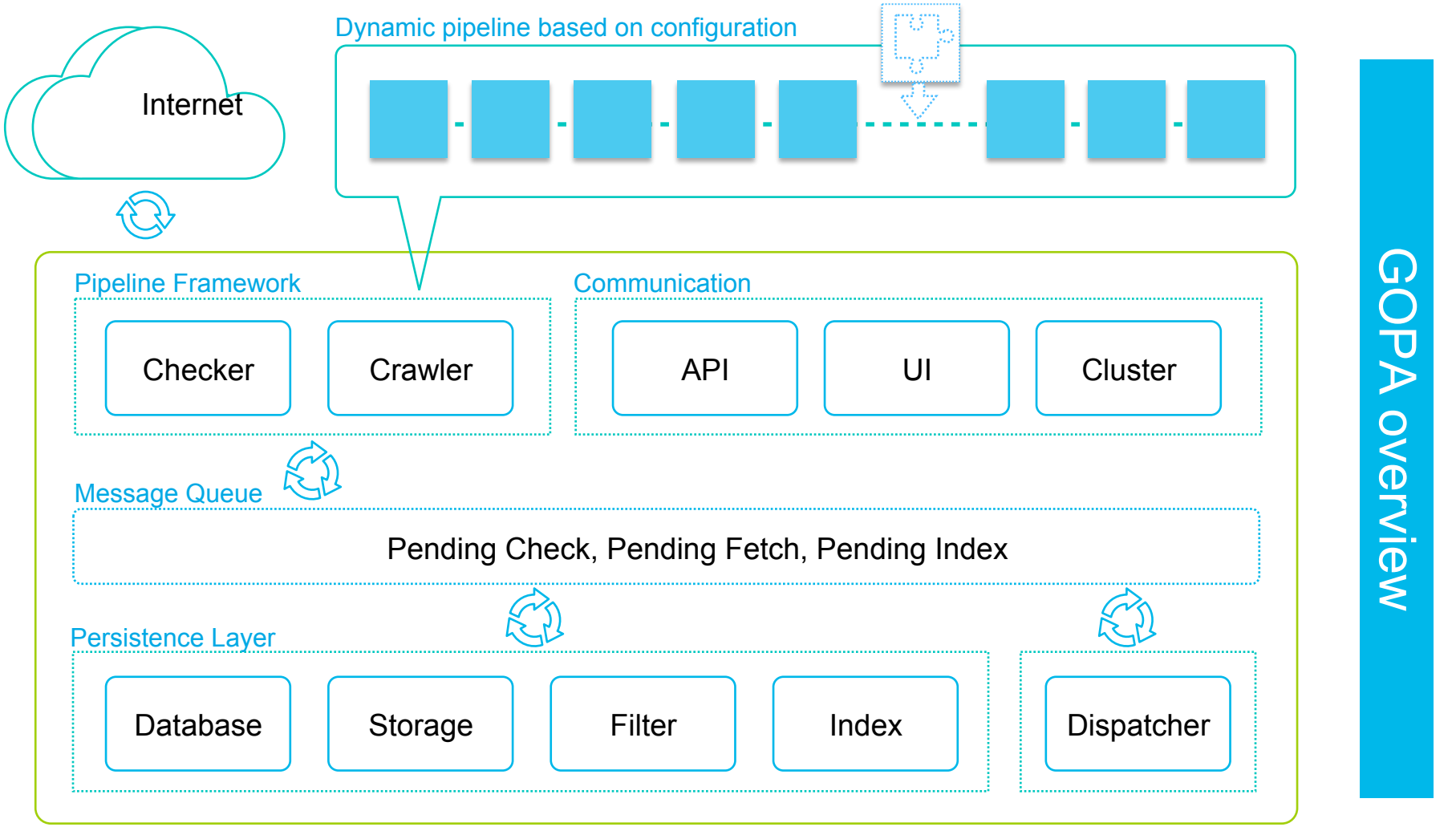
- Light weight, low footprint, memory requirement should < 100MB
- Easy to deploy, no runtime or dependency required
- Easy to use, no programming or scripts ability needed, out of box features
- Scalable and extensible in a easy way

- Demo

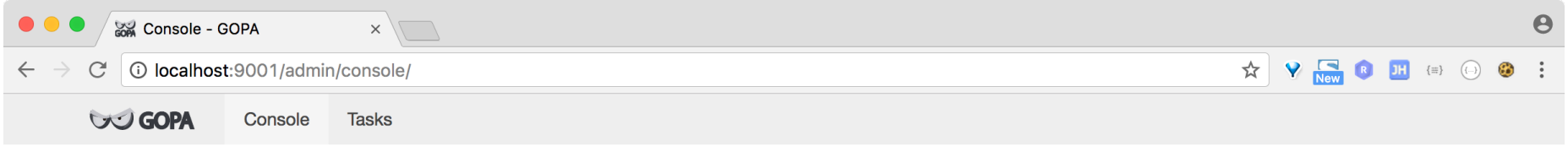


- Architecture





GOPA 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 4

elasticsearch.cn
conf.elasticsearch.cn es-guide-preview.elasticsearch.cn
grok.elasticsearch.cn

Total 4167

URL	LastUpdate	NextCheck	Status
https://elasticsearch.cn/topic/%E5%8C%97%E4%BA%AC%...	N/A	N/A	created
https://elasticsearch.cn/topic/ES%E6%95%B0%E6%8D%A...	N/A	N/A	created
https://elasticsearch.cn/crond/run/1513393168	N/A	N/A	created
https://elasticsearch.cn/topic/%E8%AE%A1%E7%AE%97%...	N/A	N/A	created
https://elasticsearch.cn/topic/%E6%A8%A1%E7%B3%8A%...	N/A	N/A	created
https://elasticsearch.cn/crond/run/1513393167	N/A	N/A	created
https://elasticsearch.cn/topic/%E4%B8%AD%E6%96%87%...	N/A	N/A	created
https://elasticsearch.cn/topic/%E5%85%B3%E4%BA%8EE...	N/A	N/A	created
https://elasticsearch.cn/people/2483	N/A	N/A	created
https://elasticsearch.cn/crond/run/1513393166	N/A	N/A	created
https://elasticsearch.cn/people/2484	N/A	N/A	created
https://elasticsearch.cn/people/2561	N/A	N/A	created
https://elasticsearch.cn/question/1269	N/A	N/A	created

Task

LastFetch: 2017-12-16 02:53:07.19264 +0000 UTC
LastCheck: 2017-12-16 02:53:07.977905 +0000 UTC
NextCheck: 2017-12-16 03:03:07.977905 +0000 UTC
ID: b8q8gnaaukimb59ob78g
Url: https://elasticsearch.cn/article/406
Reference: http://elasticsearch.cn
Reference: created
Depth: 0
Breadth: 0
Host: elasticsearch.cn
OriginalUrl:
Message:
Created: 2017-12-16 02:43:41.939546 +0000 UTC
Updated: 2017-12-16 02:53:07.984709 +0000 UTC
PipelineConfigID:

Screenshot

Snapshot(2)

SnapshotVersion: 0
SnapshotID: b8q8l4qaukimb59okvr0
SnapshotHash: 5f7c490b55fa0b852d9170953a63cdbbc4ab4041
SnapshotSimHash:
SnapshotCreated: 2017-12-16 02:53:07.192602 +0000 UTC

#	Created	Size	Version	Hash	Action
0	2017-12-16 02:53:07.192602 +0000 UTC	55349	1	5f7c490b55fa0b852d9170953a63cdbbc4ab4041	View

GOPA

*

Found about 1006 results (7ms)

Kibana - Elastic中文社区

Kibana - Elastic中文社区 输入关键字进行搜索 搜索: 发起问题 发现 话题 文章 活动 帮助 · · · 登录 注册 全部 Elasticsearch Logstash Kibana Beats 求职招聘 资讯动态 活动 Elastic日报 通知设置 新通知 我知道了 查看所有 等待回复 热门 推荐 最新 Kibana kibana可视化图表的那部分是用的echarts还是他自己集成的方法，急求大佬们解答，谢谢了? 贡献 puyunjiafly 回复了问题 • 2 人关注 • ...

https://elasticsearch.cn/sort_type-new__category-4__day-0__is_recommend-0__page-...

Kibana - Elastic中文社区

Kibana - Elastic中文社区 输入关键字进行搜索 搜索: 发起问题 发现 话题 文章 活动 帮助 · · · 登录 注册 全部 Elasticsearch Logstash Kibana Beats 求职招聘 资讯动态 活动 Elastic日报 通知设置 新通知 我知道了 查看所有 等待回复 热门 推荐 最新 Kibana 为什么kibana的visualize中Average计算数值有问题? 贡献 kennywu76 回复了问题 • 4 人关注 • 2 个回复 • 107 次浏览...

https://elasticsearch.cn/sort_type-new__category-4__day-0__is_recommend-0__page-...

Kibana - Elastic中文社区

Kibana - Elastic中文社区 输入关键字进行搜索 搜索: 发起问题 发现 话题 文章 活动 帮助 · · · 登录 注册 全部 Elasticsearch Logstash Kibana Beats 求职招聘 资讯动态 活动 Elastic日报 通知设置 新通知 我知道了 查看所有 等待回复 热门 推荐 最新 Kibana kibana monitoring 不更新状态了 回复 klause 发起了问题 • 1 人关注 • 0 个回复 • 96 次浏览 • 2017-11-20 1...

https://elasticsearch.cn/sort_type-new__category-4__day-0__is_recommend-0__page-...

Kibana - Elastic中文社区

File Ext

- .html(981)
- .md(8)
- .0(3)
- .1(2)
- .0++java(1)
- .0+ik分词器如何设置成默认分词器(1)
- .1导入oracle数据(1)
- .6(1)
- .7(1)
- .du(1)

Content Type

- text/html(1006)

Language

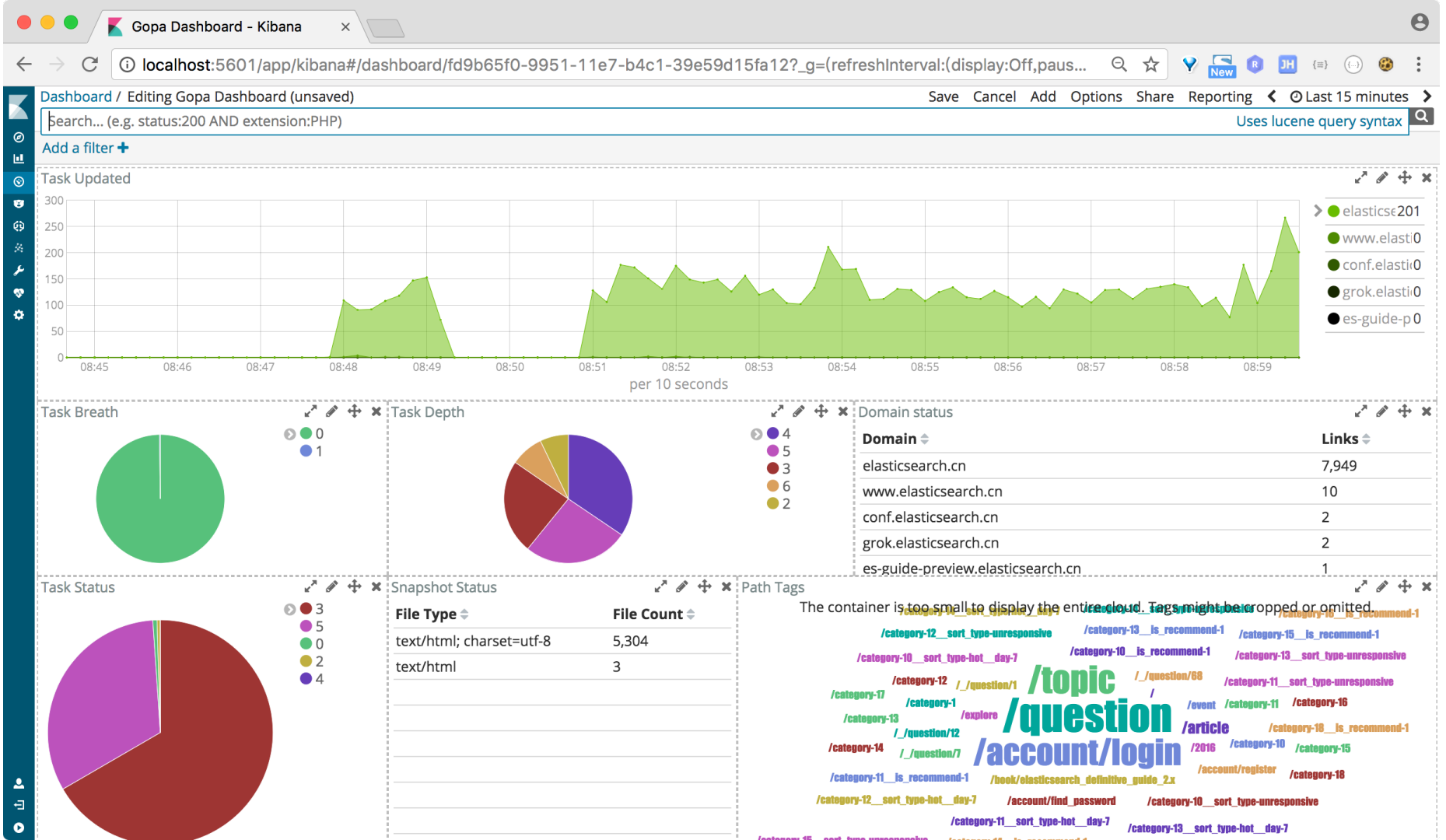
- zh(992)
- en(14)

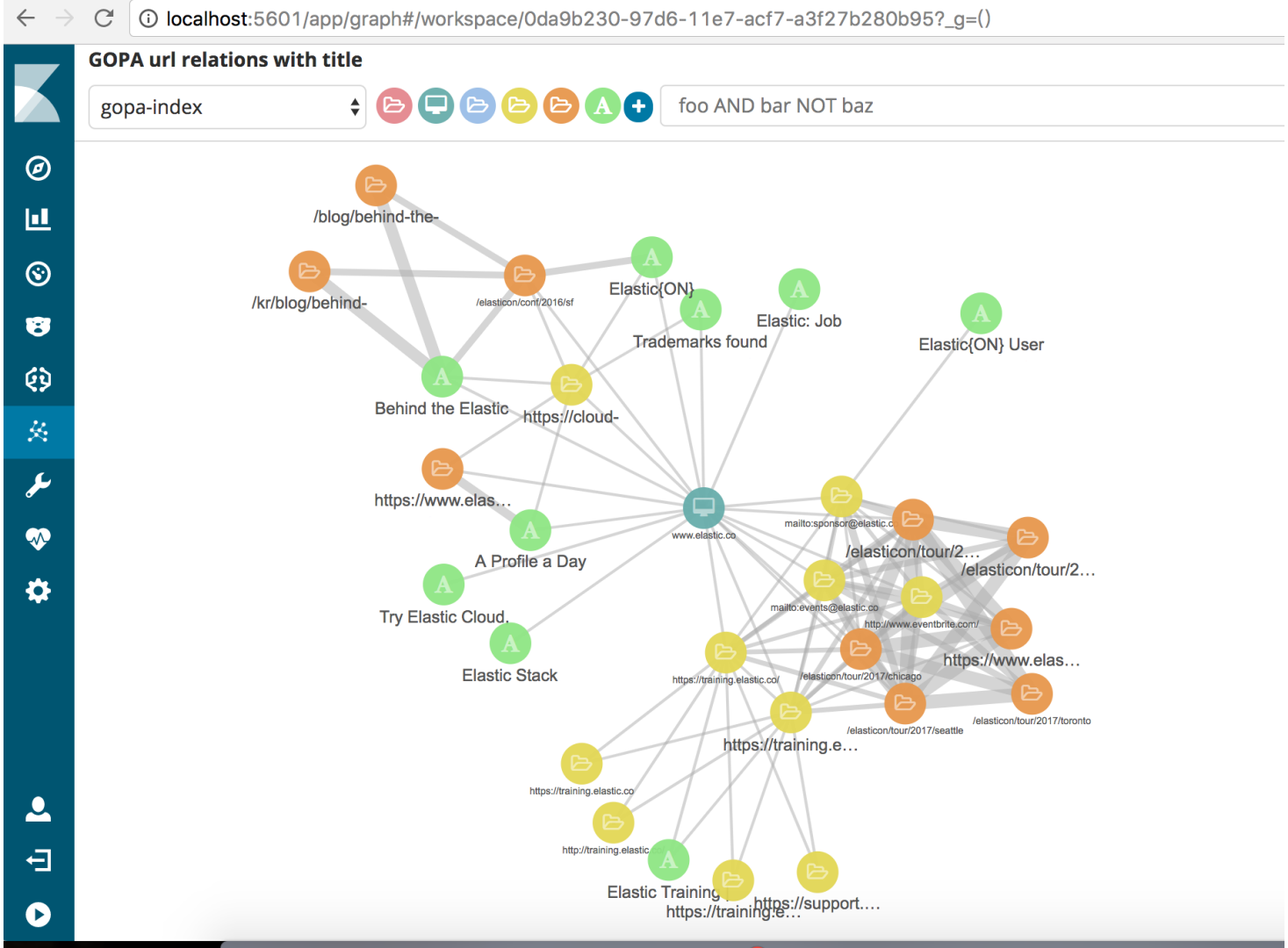
Host

- elasticsearch.cn(1004)
- conf.elasticsearch.cn(1)
- grok.elasticsearch.cn(1)

Protocol

- https(999)


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.…

Thank You