

# Scrapyd监控系统之SpiderKeeper和Scrapydweb详解

Zarten Python中文社区 昨天



zarten, 互联网一线工作者。

博客地址: [zhihu.com/people/zarten](https://zhihu.com/people/zarten)

## 概述

我们的scrapy爬虫项目可以部署在scrapyd服务器中, 可以通过scrapyd提供的接口访问web主页, 但这个页面比较简陋且一台scrapyd服务器提供一个主页, 若多台的话, 就要访问多个主页, 这样会比较麻烦。

有没有开源的库可以统一管理, 实现一键部署, 定时任务等功能呢? 下面将介绍2款比较有名的开源库: spiderkeeper和scrapydweb

这些库的共同点是: 底部调用scrapyd提供的接口, 二次封装成功能更加强大的系统。不同之处在于: scrapydweb相比spiderkeeper功能更加丰富一些, 且有异常邮件通知功能。spiderkeeper功能简陋些, 但足以应付大规模的机器, 更加成熟稳定些。

scrapydweb是最近新开发的比较大的监控系统, 可能会有不完善的地方, 以后会更加稳定丰富。

## SpiderKeeper详解

官网:

```
https://github.com/DormyMo/SpiderKeeper
```

采用一台机器提供2个scrapyd端口6800和6801作测试来模拟分布式，spiderkeeper与scrapyd服务器装在同一台机器上。

## 安装

命令:

```
pip3 install spiderkeeper
```

也可以制作成docker镜像，这里将不作具体的介绍，不会的看前面我写的文章。

## 启动

启动的前提是必须启动了scrapyd服务（scrapyd不懂的可以看前面写的文章）。

命令:

若scrapyd在本地机器:

```
spiderkeeper --server=http://localhost:6800
```

若scrapyd在远程机器或多端口:

提示: scrapyd的ip地址最好写真实的ip，虽然在本机也最好写真实ip

```
spiderkeeper --server=http://你的scrapyd的ip:6800 --server=http://你的scrapyd的ip:6801
```

访问

```
http://你的spiderkeeper的ip:5000
```

登录时需要授权，用户名和密码默认都是admin。若需要修改密码等配置信息，找到spiderkeeper安装包下的config.py文件

若不知道包所在的位置，可以使用如下命令:

pip3 show 包名

```
# Secret key for signing cookies
SECRET_KEY = "secret"

# log
LOG_LEVEL = 'INFO'

# spider services
SERVER_TYPE = 'scrapyd'
SERVERS = ['http://localhost:6800']

# basic auth
NO_AUTH = False
BASIC_AUTH_USERNAME = 'admin'
BASIC_AUTH_PASSWORD = 'admin'
BASIC_AUTH_FORCE = True
```

知乎 @Zarten

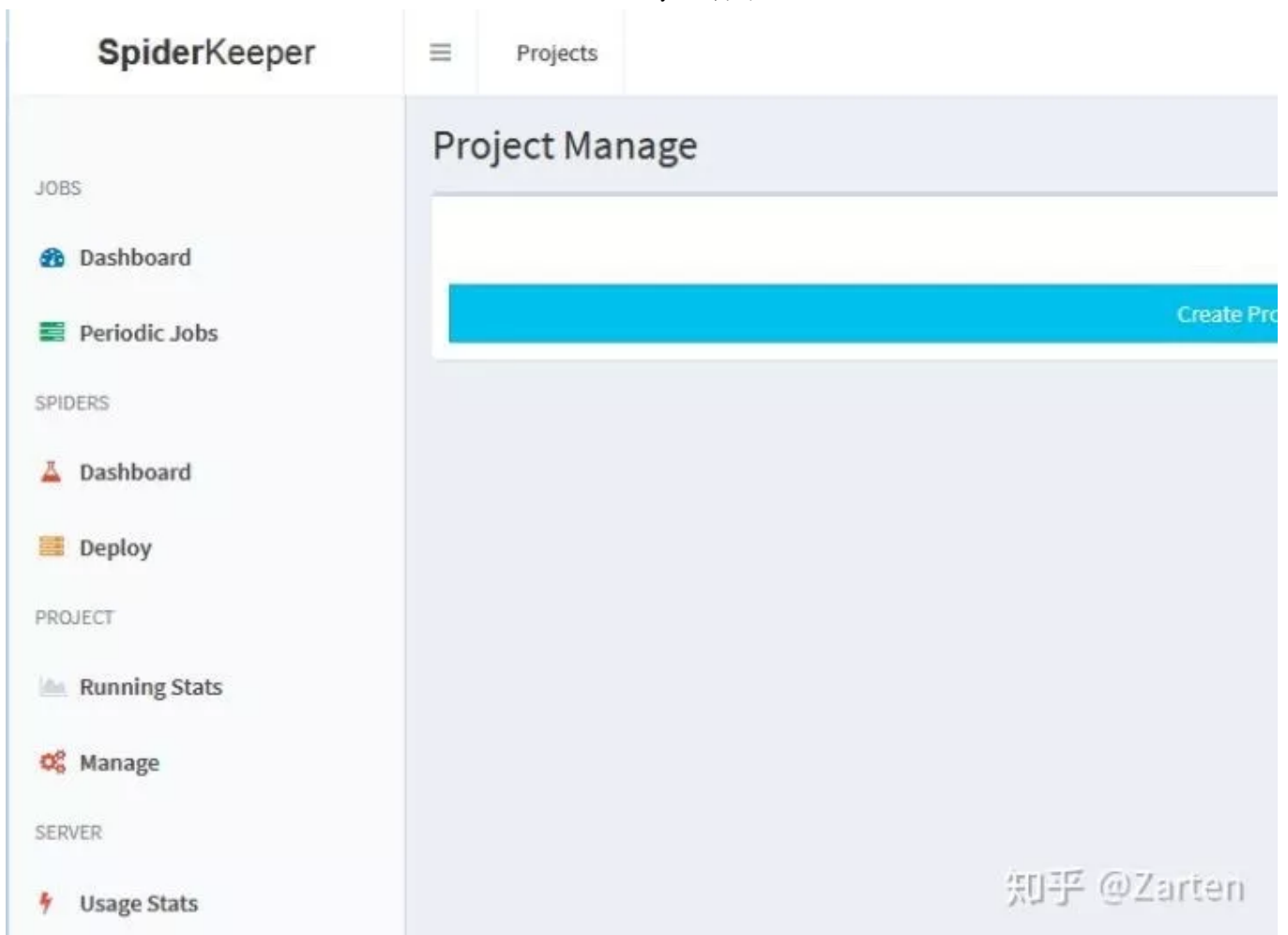
若有特殊需求，需要改默认的5000端口，找到安装包目录下的run.py文件进行修改

```
default= 0.0.0.0 )
parser.add_option("--port",
                  help="port, default:5000",
                  dest='port',
                  type="int",
                  default=5000)
parser.add_option("--username",
```

知乎 @Zarten

修改完成重启即可

界面如下所示：



## scrapy项目部署

若使用scrapyd-deploy工具部署后，spiderkeeper无法自动识别出部署的项目，必须在网页中手动部署，首先通过命令生成egg文件再上传egg文件即可。这点个人觉得挺麻烦的。

### 步骤：

- 1.在scrapyd项目中scrapy.cfg文件中写好scrapyd服务器信息
- 2.生成egg文件命令：

```
scrapyd-deploy --build-egg output.egg
```

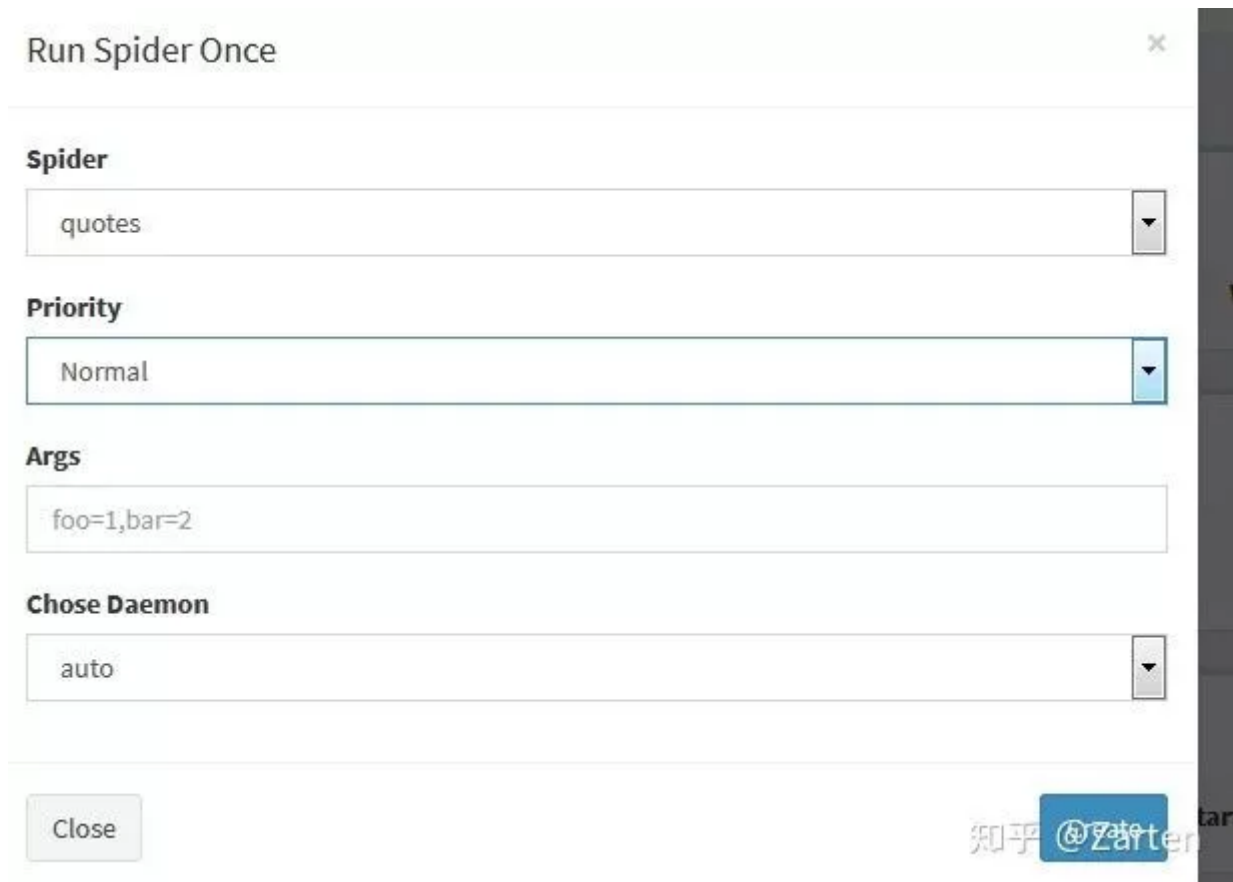
- 3.上传文件



## 运行项目

可以一次运行和定时周期运行。有个不足之处是：每次只能运行一台scrapyd机器上的项目。

- 一次运行



如下图为运行状态：

Running Jobs									
#	Job	Spider	Args	Priority	Runtime	Started	Log	Running On	Action
1	1	quotes		NORMAL	0 h 0 m	2019-03-29 08:00:43	<a href="#">Log</a>	<a href="#">http://...-0001</a>	<a href="#">Stop</a>

- 定时周期运行

### Add Periodic Job

#### Spider

quotes

#### Choose Month

Every Month

#### Priority

Normal

#### Choose Day of Week

Every day

#### Args

foo=1,bar=2

#### Choose Day of Month

Every day

#### Choose Hour

Every Hour

#### Choose Minutes

0

### Advanced Options

#### Chose Daemon

auto

#### Cron Expressions (m h dom mon dow)

0 \* \* \* \*

Close

Create

知乎 @Zarten

## 总结

其他功能将不作介绍，总体看来界面比较简洁清爽，功能比较简单

## scrapydweb详解

官网：

<https://github.com/my8100/scrapydweb>

采用一台机器提供2个scrapyd端口6800和6801作测试来模拟分布式，scrapydweb与scrapyd服务器装在同一台机器上。

这个库内容更丰富一些，并且做了大量的日志记录功能

## 安装

```
pip3 install scrapydweb
```

也可制作docker镜像。由于官方需要第一次自动生成配置文件scrapydweb\_settings\_v8.py，修改配置文件后，再次启动，这里制作镜像时事先提取出此文件了。

Docker构建的原文件：

```
https://pan.baidu.com/s/1viPC-Zx1\_ZAGxYQAwBF0zQ
```

提取码： pgg0

构建前修改 SCRAPYD\_SERVERS的信息和用户名密码信息，也可构建完成后进入容器修改

## 启动

启动前必须确保对应的scrapyd服务已经启动

命令：

```
scrapydweb
```

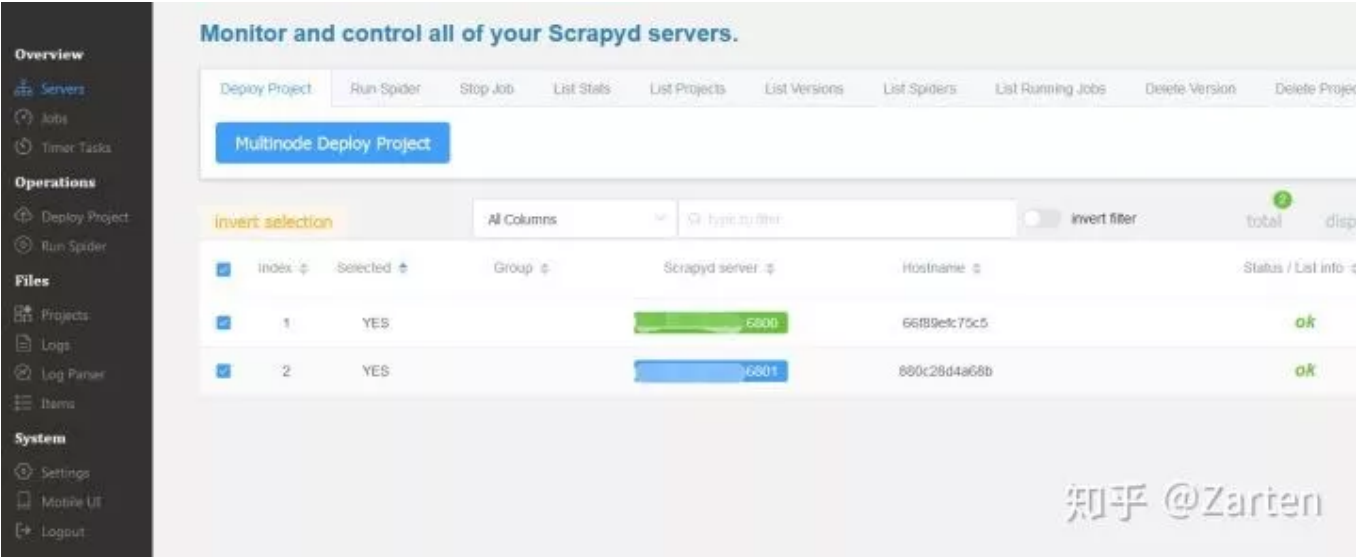
首次启动后生成配置文件，需要修改scrapyd服务器信息。

若通过上面的docker方式启动，若构建前已经配置好，则不需修改；若没有配置好，也可进入容器内修改。

## 访问

```
http://你的scrapydweb的ip:5000
```

界面



scrapy项目部署

scrapydw可以使用scrapydw-deploy工具直接部署， scrapydw会自动识别项目，也可在主页中Deploy Project菜单来部署。不过本人更青睐使用scrapydw-deploy直接部署，方便快捷。

命令部署

scrapydw-deploy 别名

网页部署界面：



Auto packaging

Upload file

folder

ScrapydWeb\_demo

▼

(1 project)

\* project

ScrapydWeb\_demo

version

2019-03-29T09\_44\_32

To select nodes in the Servers page »

\* multinodes

47.106.107.159:6801

▼

Package & Deploy

知乎 @Zarten

## 运行项目

运行项目分为一次运行和定时周期运行。但是scrapydweb就可以通过multinodes同时启动多个机器的scrapy项目

- 一次运行

可以设置settings文件和传入的参数

Scrapyd server

:6800

\* project

scrapy\_redis\_test

\* \_version

r1

\* spider

quotes

settings & arguments

timer task

\* curl command

Click the "Check CMD" button below to preview

To select nodes in the Servers page »

\* multinodes

All 2 nodes selected

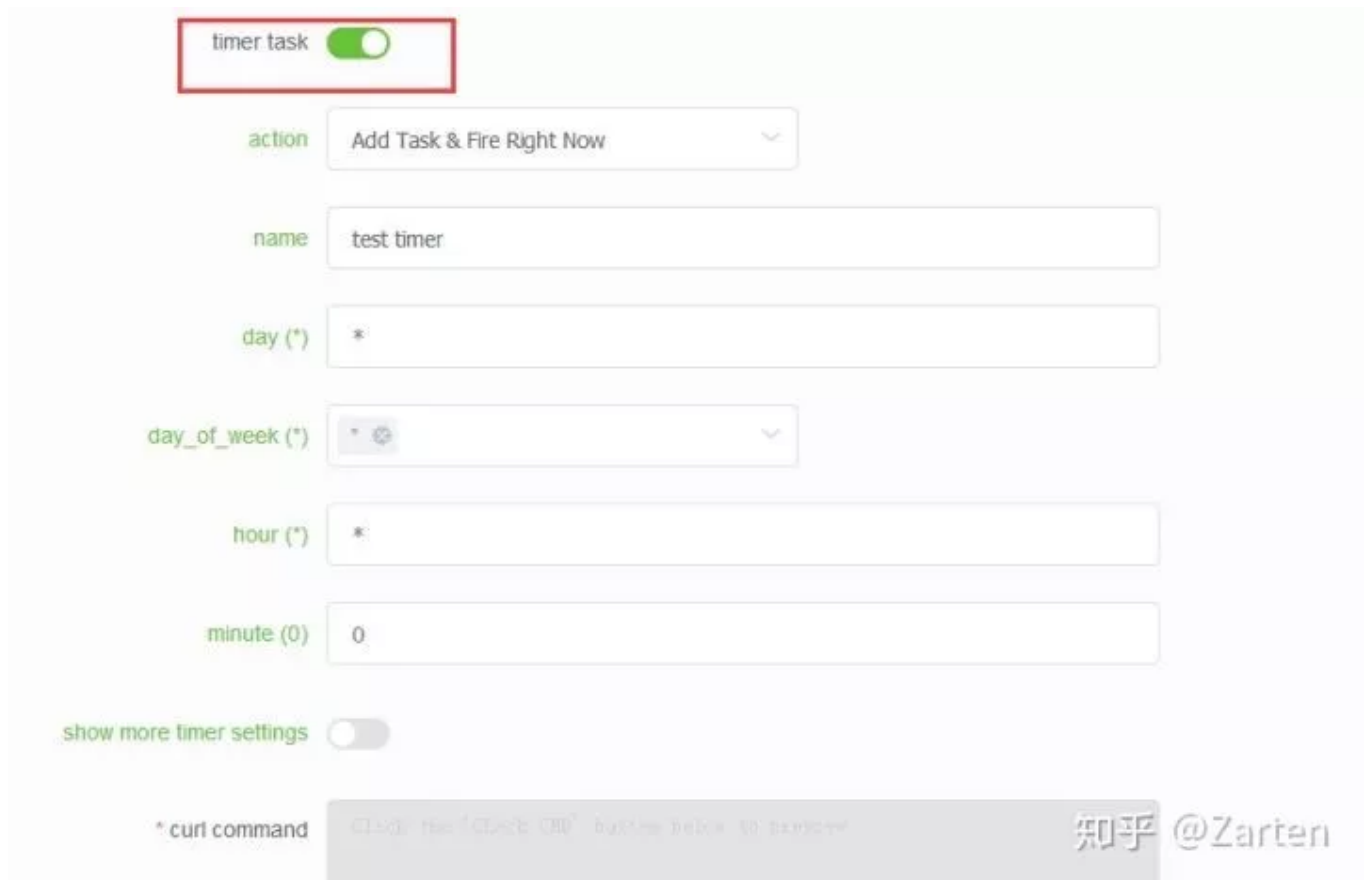
Check CMD

Run Spider

知乎 @Zarten

- 定时周期运行

从下图可以看到，通过timer task可以设置定时周期运行



timer task ☒

action Add Task & Fire Right Now

name test timer

day (\*) \*

day\_of\_week (\*) \*

hour (\*) \*

minute (0) 0

show more timer settings ☐

\* curl command

知乎 @Zarten

这个定时任务内部使用的是apscheduler库实现， 定时的参数设置可以参考这里

例如我要设置每2分钟爬一次，其他都设置为‘\*’，*minute*那里设置为‘/2’

## 总结

spiderkeeper更加简洁清爽一些，若要求不高，可以使用这个。

scrapydweb库更加丰富强大，相比spiderkeeper，优势是可以：

- 1.使用工具scrapyd-deploy部署，scrapydweb能自动识别
- 2.可以同时启动多台机器的项目
- 3.定时功能采用apscheduler库，定时更加强大
- 4.采用日志记录系统及历史情况
- 5.有邮件通知功能

## 热门推荐

用Python创建微信机器人

用Python机器人监听微信群聊

用Python获取摄像头并实时控制人脸

开源项目 | 用Python美化LeetCode仓库

推荐Python中文社区旗下的几个服务类公众号

征稿启事 | Python中文社区有奖征文



▼ 点击成为[社区注册会员](#)

「在看」一下，一起PY！

[阅读原文](#)