

scrapyd 服务部署爬虫项目

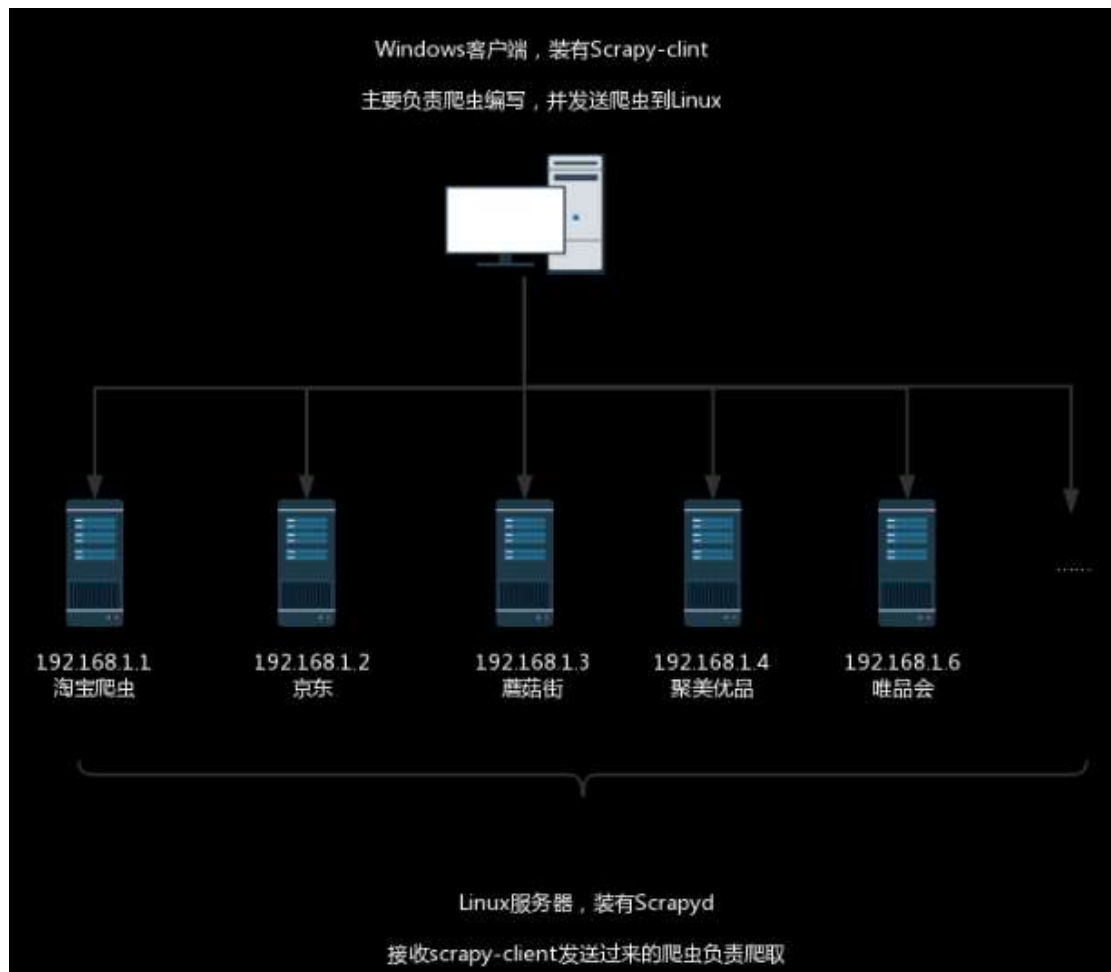
应用场景

- 只有一台开发主机
- 能够通过 Scrapyd-client 打包和部署 Scrapy 爬虫项目，以及通过 Scrapyd JSON API 来控制爬虫

缺点：命令行操作太麻烦

工作原理

Scrapyd 是一个部署和运行 Scrapy 爬虫的应用程序。它能够通过 JSON API 部署（上传）工程，并且控制工程中爬虫地启动、停止、暂停，修改。



在每台服务器上安装 scrapyd, 并开启 scrapyd 服务

在开发爬虫的客户端安装上 scrapyd 的客户端 scrapyd-client

通过 scrapyd-client 把不同网站的爬虫发送到不同的服务器

环境搭建

依赖类库

Python 2.6 or above

Twisted 8.0 or above

Scrapy 0.17 or above

Scrapyd 安装

pip install scrapyd

```
C:\Users\admin>workon env2
(env2) C:\Users\admin>pip install scrapyd
Collecting scrapyd
  Downloading https://files.pythonhosted.org/packages/dc/71/5a029b124477ab4402b9d401925e52db243819faae2befe542235921da5e/scrapyd-1.2.0-py2.py3-none-any.whl
Requirement already satisfied: Twisted>=8.0 in d:\test\virtualenv\env2\lib\site-packages (from scrapyd) (18.4.0)
Requirement already satisfied: six in d:\test\virtualenv\env2\lib\site-packages (from scrapyd) (1.11.0)
Requirement already satisfied: Scrapy>=1.0 in d:\test\virtualenv\env2\lib\site-packages (from scrapyd) (1.5.0)
Requirement already satisfied: zope.interface>=4.4.2 in d:\test\virtualenv\env2\lib\site-packages (from Twisted>=8.0->scrapyd) (4.5.0)
Requirement already satisfied: hyperlink>=17.1.1 in d:\test\virtualenv\env2\lib\site-packages (from Twisted>=8.0->scrapyd) (18.0.0)
Requirement already satisfied: incremental>=16.10.1 in d:\test\virtualenv\env2\lib\site-packages (from Twisted>=8.0->scrapyd) (17.5.0)
Requirement already satisfied: constantly>=15.1 in d:\test\virtualenv\env2\lib\site-packages (from Twisted>=8.0->scrapyd) (15.1.0)
Requirement already satisfied: Automat>=0.3.0 in d:\test\virtualenv\env2\lib\site-packages (from Twisted>=8.0->scrapyd) (0.7.0)
Requirement already satisfied: pyOpenSSL in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=1.0->scrapyd) (18.0.0)
Requirement already satisfied: queuelib in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=1.0->scrapyd) (1.5.0)
Requirement already satisfied: cssselect>=0.9 in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=1.0->scrapyd) (1.0.3)
Requirement already satisfied: service-identity in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=1.0->scrapyd) (17.0.0)
```

```

Requirement already satisfied: PyDispatcher>=2.0.5 in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=1.0->scrapyd) (2.0.5)
Requirement already satisfied: lxml in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=1.0->scrapyd) (4.2.3)
Requirement already satisfied: w3lib>=1.17.0 in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=1.0->scrapyd) (1.19.0)
Requirement already satisfied: parsel>=1.1 in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=1.0->scrapyd) (1.4.0)
Requirement already satisfied: setuptools in d:\test\virtualenv\env2\lib\site-packages (from zope.interface>=4.4.2->Twisted>=8.0->scrapyd) (39.2.0)
Requirement already satisfied: idna>=2.5 in d:\test\virtualenv\env2\lib\site-packages (from hyperlink>=17.1.1->Twisted>=8.0->scrapyd) (2.7)
Requirement already satisfied: attrs>=16.1.0 in d:\test\virtualenv\env2\lib\site-packages (from Automat>=0.3.0->Twisted>=8.0->scrapyd) (18.1.0)
Requirement already satisfied: cryptography>=2.2.1 in d:\test\virtualenv\env2\lib\site-packages (from pyOpenSSL->Scrapy>=1.0->scrapyd) (2.2.2)
Requirement already satisfied: pyasn1-modules in d:\test\virtualenv\env2\lib\site-packages (from service-identity->Scrapy>=1.0->scrapyd) (0.2.2)
Requirement already satisfied: pyasn1 in d:\test\virtualenv\env2\lib\site-packages (from service-identity->Scrapy>=1.0->scrapyd) (0.4.3)

Requirement already satisfied: cffi>=1.7; platform_python_implementation != "PyPy" in d:\test\virtualenv\env2\lib\site-packages (from cryptography>=2.2.1->pyOpenSSL->Scrapy>=1.0->scrapyd) (1.11.5)
Requirement already satisfied: asn1crypto>=0.21.0 in d:\test\virtualenv\env2\lib\site-packages (from cryptography>=2.2.1->pyOpenSSL->Scrapy>=1.0->scrapyd) (0.24.0)
Requirement already satisfied: pycparser in d:\test\virtualenv\env2\lib\site-packages (from cffi>=1.7; platform_python_implementation != "PyPy"->cryptography>=2.2.1->pyOpenSSL->Scrapy>=1.0->scrapyd) (2.18.1)

Installing collected packages: scrapyd
Successfully installed scrapyd-1.2.0

(env2) C:\Users\admin>

```

Scrapyd-Client 安装

pip install scrapyd-client


```
(env2) C:\Users\admin>pip install scrapyd-client
Collecting scrapyd-client
  Downloading https://files.pythonhosted.org/packages/e1/76/3dd5f5be5c98436bb71d7212bf18233ae8519da0a7d03f541d4fe0f91be6b/scrapyd_client-1.1.0-py2.py3-none-any.whl
Requirement already satisfied: six in d:\test\virtualenv\env2\lib\site-packages (from scrapyd-client) (1.11.0)
Requirement already satisfied: Scrapy>=0.17 in d:\test\virtualenv\env2\lib\site-packages (from scrapyd-client) (1.5.0)
Requirement already satisfied: parsel>=1.1 in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=0.17->scrapyd-client) (1.4.0)
Requirement already satisfied: cssselect>=0.9 in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=0.17->scrapyd-client) (1.0.3)
Requirement already satisfied: Twisted>=13.1.0 in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=0.17->scrapyd-client) (18.4.0)

Requirement already satisfied: pyOpenSSL in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=0.17->scrapyd-client) (18.0.0)
Requirement already satisfied: queuelib in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=0.17->scrapyd-client) (1.5.0)
Requirement already satisfied: PyDispatcher>=2.0.5 in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=0.17->scrapyd-client) (2.0.5)
Requirement already satisfied: service-identity in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=0.17->scrapyd-client) (17.0.0)
Requirement already satisfied: lxml in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=0.17->scrapyd-client) (4.2.3)
Requirement already satisfied: w3lib>=1.17.0 in d:\test\virtualenv\env2\lib\site-packages (from Scrapy>=0.17->scrapyd-client) (1.19.0)
Requirement already satisfied: Automat>=0.3.0 in d:\test\virtualenv\env2\lib\site-packages (from Twisted>=13.1.0->Scrapy>=0.17->scrapyd-client) (0.7.0)
Requirement already satisfied: incremental>=16.10.1 in d:\test\virtualenv\env2\lib\site-packages (from Twisted>=13.1.0->Scrapy>=0.17->scrapyd-client) (17.5.0)
Requirement already satisfied: zope.interface>=4.4.2 in d:\test\virtualenv\env2\lib\site-packages (from Twisted>=13.1.0->Scrapy>=0.17->scrapyd-client) (4.5.0)
Requirement already satisfied: constantly>=15.1 in d:\test\virtualenv\env2\lib\site-packages (from Twisted>=13.1.0->Scrapy>=0.17->scrapyd-client) (15.1.0)
Requirement already satisfied: hyperlink>=17.1.1 in d:\test\virtualenv\env2\lib\site-packages (from Twisted>=13.1.0->Scrapy>=0.17->scrapyd-client) (18.0.0)
Requirement already satisfied: cryptography>=2.2.1 in d:\test\virtualenv\env2\lib\site-packages (from pyOpenSSL->Scrapy>=0.17->scrapyd-client) (2.2.2)
Requirement already satisfied: attrs in d:\test\virtualenv\env2\lib\site-packages (from service-identity->Scrapy>=0.17->scrapyd-client) (18.1.0)
Requirement already satisfied: pyasn1-modules in d:\test\virtualenv\env2\lib\site-packages (from service-identity->Scrapy>=0.17->scrapyd-client) (0.2.2)
```

```

Requirement already satisfied: pyasn1 in d:\test\virtualenv\env2\lib\
\site-packages (from service-identity->Scrapy>=0.17->scrapyd-client)
(0.4.3)
Requirement already satisfied: setuptools in d:\test\virtualenv\env2\
\lib\site-packages (from zope.interface>=4.4.2->Twisted>=13.1.0->Scr
apy>=0.17->scrapyd-client) (39.2.0)
Requirement already satisfied: idna>=2.5 in d:\test\virtualenv\env2\
lib\site-packages (from hyperlink>=17.1.1->Twisted>=13.1.0->Scrapy>=
0.17->scrapyd-client) (2.7)
Requirement already satisfied: asn1crypto>=0.21.0 in d:\test\virtual
env\env2\lib\site-packages (from cryptography>=2.2.1->pyOpenSSL->Scr
apy>=0.17->scrapyd-client) (0.24.0)
Requirement already satisfied: cffi>=1.7; platform_python_implementation
!= "PyPy" in d:\test\virtualenv\env2\lib\site-packages (from cr
yptography>=2.2.1->pyOpenSSL->Scrapy>=0.17->scrapyd-client) (1.11.5)

Requirement already satisfied: pycparser in d:\test\virtualenv\env2\
lib\site-packages (from cffi>=1.7; platform_python_implementation !=
t) (2.18)
Installing collected packages: scrapyd-client
Successfully installed scrapyd-client-1.1.0

(env2) C:\Users\admin>

```

验证安装

scrapyd-deploy -h

出现如上信息表示 scrapy-client 安装成功

```

(env2) C:\Users\admin>scrapyd-deploy -h
Usage: scrapyd-deploy [options] [ [target] | -l | -L <target> ]

Deploy Scrapy project to Scrapyd server

Options:
  -h, --help                show this help message and exit
  -p PROJECT, --project=PROJECT
                           the project name in the target
  -v VERSION, --version=VERSION
                           the version to deploy. Defaults to current t
                           imestamp
  -l, --list-targets        list available targets
  -a, --deploy-all-targets
                           deploy all targets
  -d, --debug                debug mode (do not remove build dir)
  -L TARGET, --list-projects=TARGET
                           list available projects on TARGET
  --egg=FILE                use the given egg, instead of building it
  --build-egg=FILE          only build the egg, don't deploy it

```


window 下 scrapyd-client 安装的问题

在 windows 上使用 scrapyd-client 安装后, 并不能使用相应的命令'scrapyd-deploy'

(env2) C:\Users\admin>scrapyd-deploy -h
'scrapyd-deploy' 不是内部或外部命令, 也不是可运行的程序
或批处理文件。
<https://blog.csdn.net/Kxyky>

需要在"D:\test\virtualenv\env2\Scripts" 目录下新建批处理文件 scrapyd-deploy.bat:

scrapyd.exe	2018/9/21 15:06	应用程序	101 KB
scrapyd-deploy	2018/9/21 15:44	文件	10 KB
scrapyd-deploy.bat	2018/9/21 16:42	Windows 批处理	1 KB
scrapyrt.exe	2018/9/21 16:35	应用程序	101 KB

内容:

```
@echo off
"d:\test\virtualenv\env2\scripts\python.exe"
"d:\test\virtualenv\env2\Scripts\scrapyd-deploy" %1 %2 %3 %4 %5 %6 %7 %8 %9
```



Scrapy API 安装

通过 API 可以获取当前 scrapy 任务运行状况

pip install python-scrapy-api

```
(env2) C:\Users\admin>pip install python-scrapy-api
Collecting python-scrapy-api
  Downloading https://files.pythonhosted.org/packages/13/13/cf8bbd7a6462a805c26bfd8eb92a0fc1f5e69a13fa2e5fbd87360943cacc/python_scrapy_api-2.1.2-py2.py3-none-any.whl
Requirement already satisfied: requests in d:\test\virtualenv\env2\lib\site-packages (from python-scrapy-api) (2.19.1)
Requirement already satisfied: idna<2.8,>=2.5 in d:\test\virtualenv\env2\lib\site-packages (from requests->python-scrapy-api) (2.7)
Requirement already satisfied: urllib3<1.24,>=1.21.1 in d:\test\virtualenv\env2\lib\site-packages (from requests->python-scrapy-api) (1.23)
Requirement already satisfied: certifi>=2017.4.17 in d:\test\virtualenv\env2\lib\site-packages (from requests->python-scrapy-api) (2018.4.16)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in d:\test\virtual
```

```

ualenv\env2\lib\site-packages (from requests->python-scrapyd-api) (3
.0.4)
Installing collected packages: python-scrapyd-api
Successfully installed python-scrapyd-api-2.1.2

(env2) C:\Users\admin>

```

Scrapyrt 的安装

scrapyrt 为 scrapy 提供了一个调度的 HTTP 接口，有了它，就不需要再执行 scrapy 命令而是通过一个 http 接口来调度 scrapy 任务了。

scrapyrt 比 scrapyd 更轻量，如果不需要分布式多任务的话，可以简单实用 scrapyrt 实现远程 scrapy 任务的调度

pip install scrapyrt

```

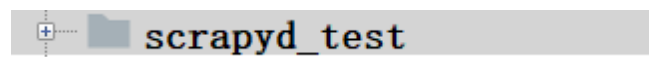
(env2) C:\Users\admin>pip install scrapyrt
Collecting scrapyrt
  Downloading https://files.pythonhosted.org/packages/cc/ca/e04513d8
a7f900ed09d5558b7ef0768197ecd519c7359815b4b744ff4bab/scrapyrt-0.10.
tar.gz
Requirement already satisfied: Twisted>=14.0.0 in d:\test\virtualenv
\env2\lib\site-packages (from scrapyrt) (18.4.0)
Requirement already satisfied: Scrapy>=1.0.0 in d:\test\virtualenv\env2\lib\site-packages (from scrapyrt) (1.5.0)
Collecting demjson (from scrapyrt)
  Downloading https://files.pythonhosted.org/packages/96/67/6db789e2
533158963d4af689f961b644ddd9200615b8ce92d6cad695c65a/demjson-2.2.4.
tar.gz (131kB)
    7% |██████████| 10kB 22kB/s eta 0:00:06

```

向 scrapyd 中部署项目

新建一个空的 python 工程

新建一个空的 python 工程，并在本工程下启动 scrapyd



启动 scrapyd 服务

启动后不要关闭，要基于 scrapyd 服务来运行后面的部署和爬虫。

scrapyd

```
E:\Python\代码\scrapy\scrapyd_test>workon env2
(env2) E:\Python\代码\scrapy\scrapyd_test>scrapyd
2018-10-16T21:01:58+0800 [-] Loading d:\test\virtualenv\env2\lib\site-
packages\scrapyd\txapp.py...
2018-10-16T21:02:01+0800 [-] Scrapyd web console available at http://1
27.0.0.1:6800/
2018-10-16T21:02:01+0800 [-] Loaded.
2018-10-16T21:02:01+0800 [twisted.application.app.AppLogger#info] twis
td 18.4.0 (d:\test\virtualenv\env2\scripts\python.exe 3.5.3) starting
up.
2018-10-16T21:02:01+0800 [twisted.application.app.AppLogger#info] reac
tor class: twisted.internet.selectreactor.SelectReactor.
2018-10-16T21:02:01+0800 [-] Site starting on 6800
2018-10-16T21:02:01+0800 [twisted.web.server.Site#info] Starting facto
ry <twisted.web.server.Site object at 0x000002214D453128>
2018-10-16T21:02:01+0800 [Launcher] Scrapyd 1.2.0 started: max_proc=32
, runner='scrapyd.runner'
```

配置爬虫项目

编辑 scrapy 项目目录下的 scrapy.cfg 文件：

```
scrapy.cfg ×
1 # Automatically created by: scrapy startproject
2 #
3 # For more information about the [deploy] section see:
4 # https://scrapyd.readthedocs.io/en/latest/deploy.html
5
6 [settings]
7 default = JobSpider.settings
8
9 [deploy:51Job] → 项目部署到scrapyd上的名称
10 #url = http://localhost:6800/ → scrapy项目的地址
11 project = JobSpider → 项目名称
12
```

进入需要部署的项目根目录

```
(env2) e:\>cd E:\Python\代码\scrapy\JobSpider
(env2) E:\Python\代码\scrapy\JobSpider>■
```

查看当前可用于部署到 scrapyd 服务中的爬虫有哪些。

scrapyd-deploy -l

```
(env2) E:\Python\代码\scrapy\JobSpider>scrapyd-deploy -l
51Job http://localhost:6800/
(env2) E:\Python\代码\scrapy\JobSpider>
```

参数 1: [deploy: 51job]

参数 2: scrapy.cfg 文中的 url

查看当前项目中可用的爬虫

scrapy list

```
(env2) E:\Python\代码\scrapy\JobSpider>scrapy list
pythonPosition
```

列举 scrapyd 服务中已经部署的爬虫项目

curl http://localhost:6800/listprojects.json

```
(env2) E:\Python\代码\scrapy\JobSpider>curl http://localhost:6800/li
stprojects.json
{"projects": [], "status": "ok", "node_name": "QIKUWW-JJ3BDIB"}
```

当前 projects 中为空，没有已部署项目

注意：curl 不是内部文件，需要提前安装，见附录

上传项目

scrapyd-deploy 51Job -p JobSpider

```
(env2) E:\Python\代码\scrapy\JobSpider>scrapyd-deploy 51Job -p JobSpider
Packing version 1537523150  ← 版本
Deploying to project "JobSpider" in http://localhost:6800/addversion.json
Server response (200):
{"spiders": 1, "version": "1537523150", "status": "ok", "node_name": "QIKUWW-
JJ3BDIB", "project": "JobSpider"}
                                部署名称 项目名称

(env2) E:\Python\代码\scrapy\JobSpider>
```

status: 上传至 scrapyd 的状态

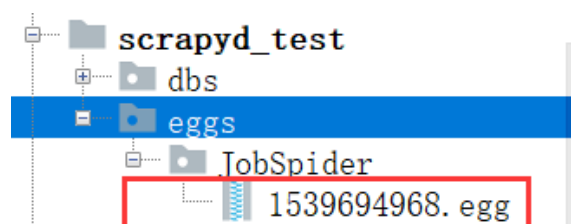
project: 上传的项目名称

version: 此次上传项目的版本号，项目可以多次打包上传，每次上传都会有不同的版本号

spiders: 该项目中包含的爬虫个数

部署完成之后，在创建的 python 工程里可以看到多了一个 eggs 的文件夹，以及里面的一

些内容，里面所存放的就是 scrapyd-deploy 的工程打包成.egg 的文件，可以看到 version 就是文件的名称，每当我们执行一次 scrapyd-deploy 就会生成一个新的 egg，如图所示：



查询当前项目中的爬虫个数

curl http://localhost:6800/listspiders.json?project=JobSpider

```
(env2) E:\Python\代码\scrapy\JobSpider>curl http://localhost:6800/listspider
s.json?project=JobSpider
{"spiders": [{"pythonPosition": "pythonPosition", "status": "ok", "node_name": "QIKUWWW-JJ3BDIB"}]}
```

爬虫名称 状态 分布式部署中主机名称

启动爬虫

curl http://localhost:6800/schedule.json -d project=JobSpider -d spider=pythonPosition

```
C:\Users\admin>workon env2
(env2) C:\Users\admin>curl http://localhost:6800/schedule.json -d pr
oject=JobSpider -d spider=pythonPosition
{"jobid": "af5d0792be4811e89bd9d8cb8af1ef5c", "node_name": "QIKUWWW-
JJ3BDIB", "status": "ok"}
```

项目名称 爬虫名称

关闭（取消）爬虫

curl http://localhost:6800/cancel.json -d project=JobSpider -d job=4c89cda2d12511e8adabd8cb8af1ef5c

```
(env2) C:\Users\admin>curl http://localhost:6800/cancel.json -d project=JobSpider -d job=8f5d0792be4811e89bd0d8cb8af1ef5c
{"prevstate": null, "node_name": "QIKUWWW-JJ3BDIB", "status": "ok"}
(env2) C:\Users\admin>
```

jobid

删除项目

```
curl http://localhost:6800/delproject.json -d project=JobSpider
```

```
(env2) C:\Users\admin>curl http://localhost:6800/delproject.json -d project=JobSpider
{"node_name": "QIKUWWW-JJ3BDIB", "status": "ok"}
```

爬虫项目名称

浏览器中查看

scrapyd 提供了一个查看界面

<http://localhost:6800/>



Scrapyd

Available projects: **JobSpider**

- [Jobs](#)
- [Logs](#)
- [Documentation](#)

How to schedule a spider?

To schedule a spider you need to use the API (this web UI is only for monitoring)

Example using [curl](#):

```
curl http://localhost:6800/schedule.json -d project=default -d spider=somespider
```

For more information about the API, see the [Scrapyd documentation](#)

Jobs

[Go back](#)

Project	Spider	Job	PID	Start	Runtime	Finish	Log
Pending							
Running							
Finished							
JobSpider	pythonPosition	8f5d0792be4811e89bd0d8cb8af1ef5c		2018-09-22 17:19:08	0:00:03	2018-09-22 17:19:11	Log