1. 下载镜像包：链接: https://pan.baidu.com/s/1pPc5UXSxfHxWEfYBomE93w 提取码: a5y4
2. 360解压一次
3. WinSCP 上传到linux

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Docker 部署\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

》安装docker：

(可选)Window安装docker： <https://blog.csdn.net/hunan961/article/details/79484098>

Centos7安装docker: [**https://www.cnblogs.com/yufeng218/p/8370670.html**](https://www.cnblogs.com/yufeng218/p/8370670.html)

》配置docker镜像加速

vim /etc/docker/daemon.json

{ "registry-mirrors": ["自己阿里云注册获取"] }

》加载docker镜像：

centos:python3 (包含:python3,scrapy,chrome)：

加载(耗时)：**docker load -i** centos:python3**.tar**

创建&运行：(创建两台爬虫机)

**docker run -itd -p 6801:6800 --privileged=true centos:python3 /usr/sbin/init**

**docker run -itd -p 6802:6800 --privileged=true centos:python3 /usr/sbin/init**

centos:nmdb(包含：Nginx，redis，mysql，mongodb)：

加载：**docker load -i** centos:nmdb**.tar**

创建&运行：

**docker run -itd -p 81:80 -p 8049:8049 -p 3307:3306 -p 27018:27017 -p 6380:6379 --privileged=true centos:nmdb /usr/sbin/init**

splash:d190402(scrapy-splash镜像)：

{ 二选一：耗时

拉取：**docker pull scrapinghub/splash**

创建&运行：(创建两台splash渲染机)

**docker run -itd -p 8050:8050 --privileged=true scrapinghub/splash /usr/sbin/init**

**docker run -itd -p 8051:8050 --privileged=true scrapinghub/splash /usr/sbin/init**

}

{ 二选一：耗时

加载：**docker load -i** splash:d190402**.tar**

创建&运行：(创建两台splash渲染机)

**docker run -itd -p 8050:8050 --privileged=true** splash:d190402 **/usr/sbin/init**

**docker run -itd -p 8051:8050 --privileged=true** splash:d190402 **/usr/sbin/init**

}

gerapy镜像：

拉取：**Git clone** <https://github.com/Gerapy/Gerapy.git>

创建&运行：

# 进入目录下执行（同 .git 同级目录）

**docker run -d -v /home/gerapy:/app/gerapy -p 8000:8000 thsheep/gerapy:master**

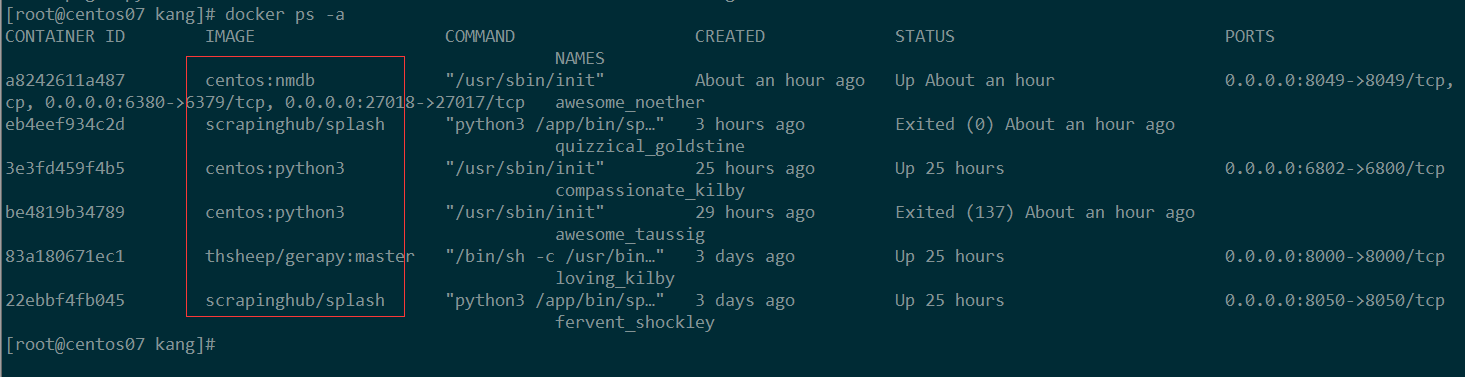
部署完成可见：

Docker\_数据库： centos:nmdb 一台

Docker\_爬虫机： centos:python3 两台

Docker\_splash渲染： splash 两台

Docker\_gerapy 管理平台：gerapy 一台



以下Host 或 IP === 宿主机IP(非Docker内ip)

Docker\_数据库\_账号密码：

MONGO\_PORT = '27018'  
MONGO\_USER = 'root'  
MONGO\_ADMIN\_PWD = '123456'  
  
MYSQL\_PORT = '3307'  
MYSQL\_USER = 'root'  
MYSQL\_ROOT\_PWD = '123456'  
  
DB\_NAME = 'spider'

# for redis  
REDIS\_HOST=''  
REDIS\_PORT='6380'  
REDIS\_PASSWORD='123456'

REDIS\_URL = 'redis://:{}@{}:{}'.format(REDIS\_PASSWORD,REDIS\_HOST,REDIS\_PORT) # RedisSpider

**Docker\_gerapy 管理平台:**

访问：<http://ip:8000/#/client/create>

如果访问失败检查 gerapy 服务是否启动：

docker exec -it docker\_gerapy\_容器ID /bin/bash

ps –ef |grep gerapy

# 启动gerapy服务

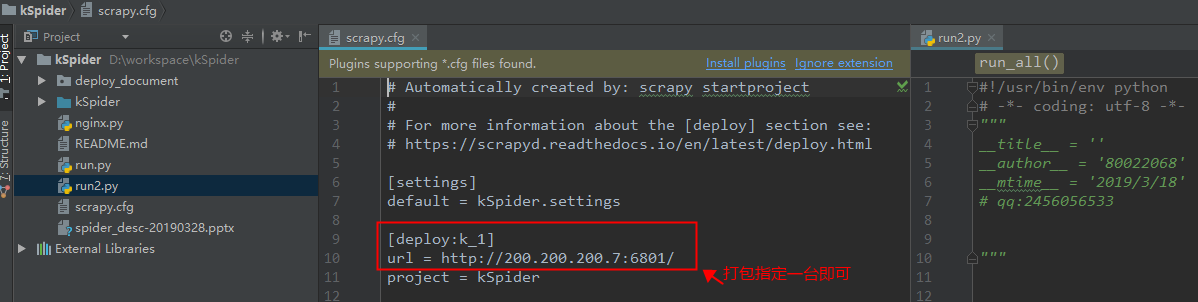
**cd /app/gerapy && nohup gerapy runserver 0.0.0.0:8000 &**



这里的端口 === 创建Docker\_爬虫机 指定的端口映射(**6801** or **6802**)

爬虫项目需要放在 **/home/gerapy/projects** 目录下(创建docker时映射的目录)

**关于项目配置scrayd 打包地址：**



**Docker\_splash渲染\_splash地址：**

# for splash  
SPLASH\_URL = 'http://ip:8050'  
  
  
# splash+nginx 负载均衡  
NGINX\_SPLASH\_URL='http://ip:8049'

**关于splash+nginx 负载均衡渲染：**

docker exec -it docker\_数据库\_容器ID /bin/bash

vim /usr/local/nginx/conf/nginx.conf

# 8050-8051 为之前创建splash容器映射端口

# 8049 为centos:nmdb创建时映射端口

upstream splash {

least\_conn;

server ip:**8050**;

server ip:**8051**;

}

server {

listen **8049**;

location / {

proxy\_pass http://splash;

}

}

( 通过 访问 docker\_centos:nmdb Nginx 服务代理转发 到 docker\_centos：splash )

目前已知的Bug：

**centos:nmdb 会遇到的问题**

问题1：centos:nmdb创建容器之后 mysql服务拉不起来 报：

Fatal error: Can't open and lock privilege tables: Table storage engine for 'user' doesn't have this option.

原因：

变动了hostname主机名，MySQL data文件需要初始化

查看：

**vim /etc/my.cof**

port = 3306

socket = /tmp/mysql.sock

basedir = /usr/local/mysql

datadir = /data/mysql

pid-file = /data/mysql/mysql.pid

user = mysql

解决：

# MySQL初始化（mysql数据会全丢！！要记录下随机密码）

**./mysqld --initialize --user=mysql --basedir=/usr/local/mysql --datadir=/data/mysql**

# 查看随机密码

**tail -100f /data/mysql/mysql-error.log |grep password**

# 赋权

**chmod -R 777 /data/mysql**

# 再次启动

**./mysqld\_safe --defaults-file=/etc/my.cnf &**

# 登录&&修改秘密&&配置远程连接

**./mysql -u root -p**

**set password=password("123456");**

**GRANT ALL PRIVILEGES ON \*.\* TO 'root'@'%' IDENTIFIED BY '123456' WITH GRANT OPTION;**

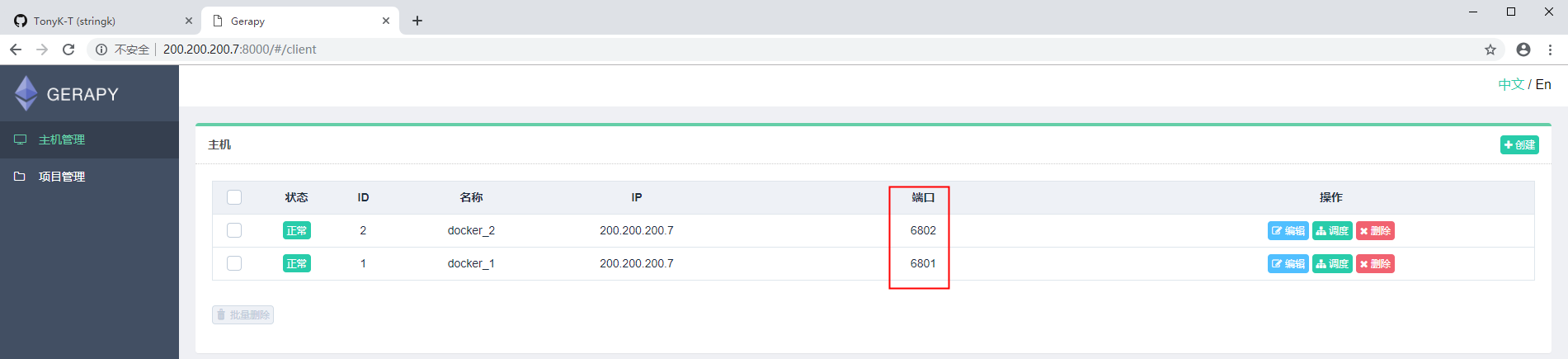
**flush privileges;**

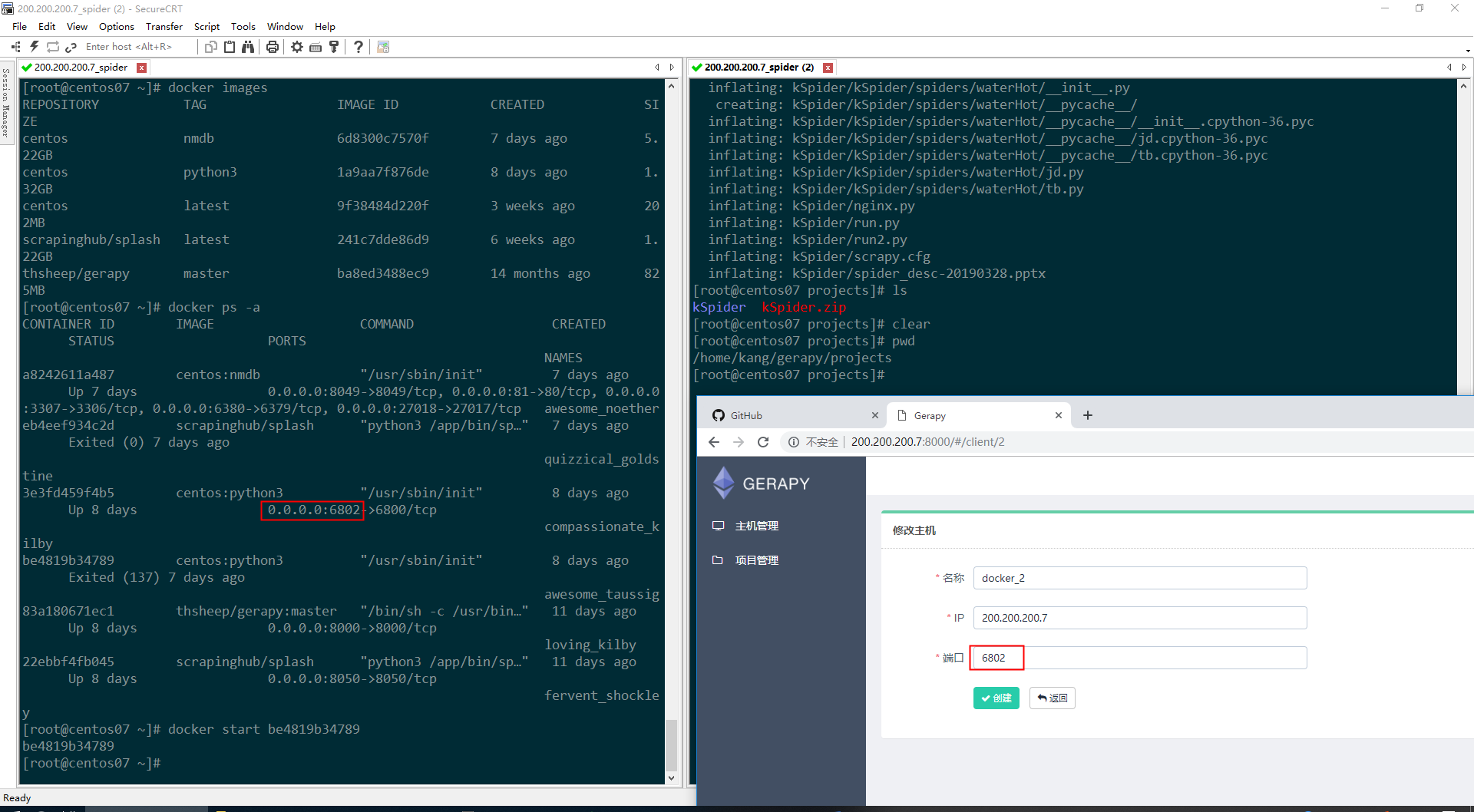
**use mysql;** # 这里切了库

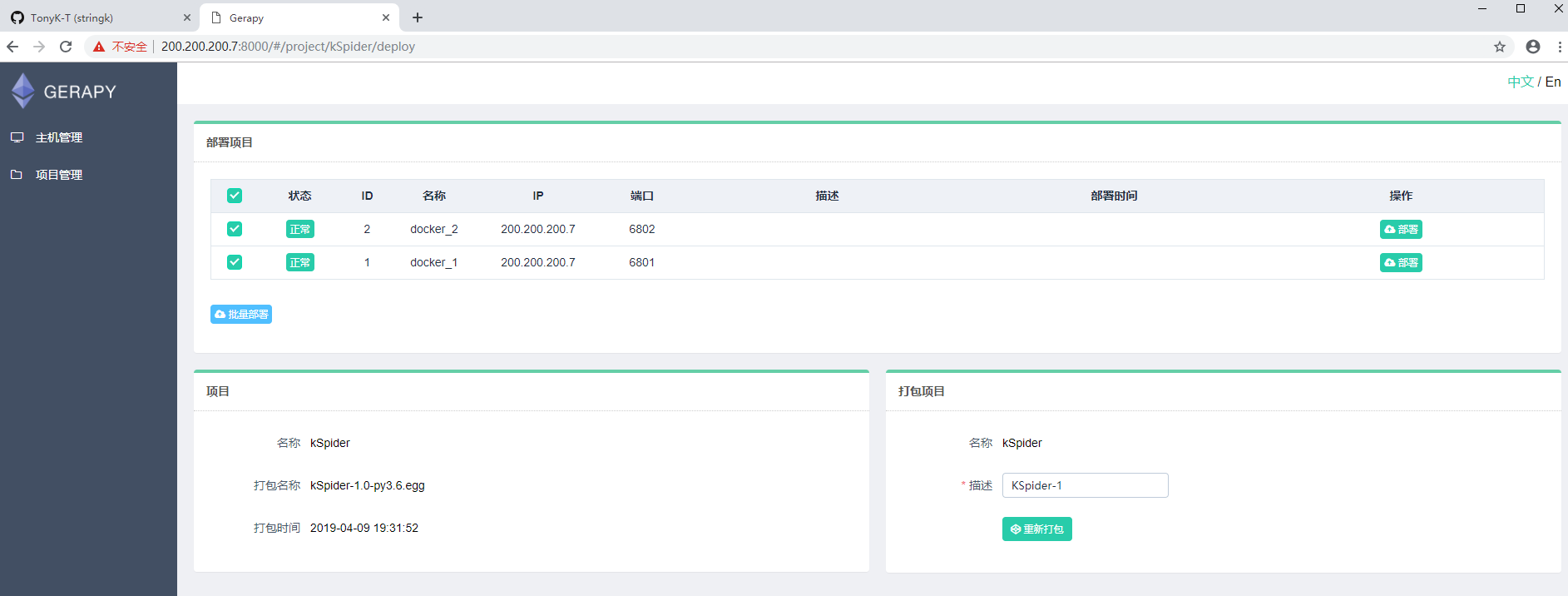
**alter user 'root'@'localhost' identified with mysql\_native\_password by '123456';**

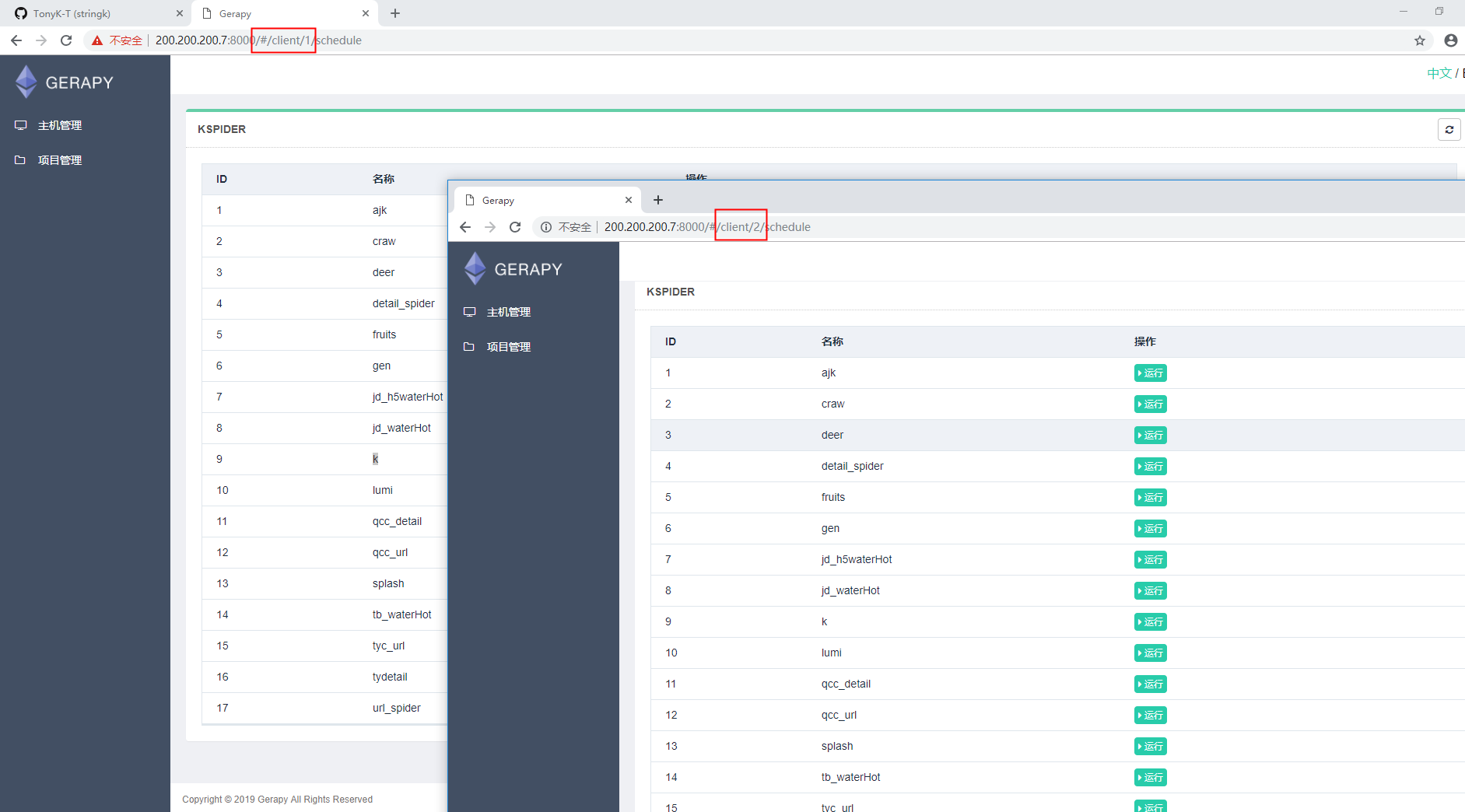
**flush privileges;**

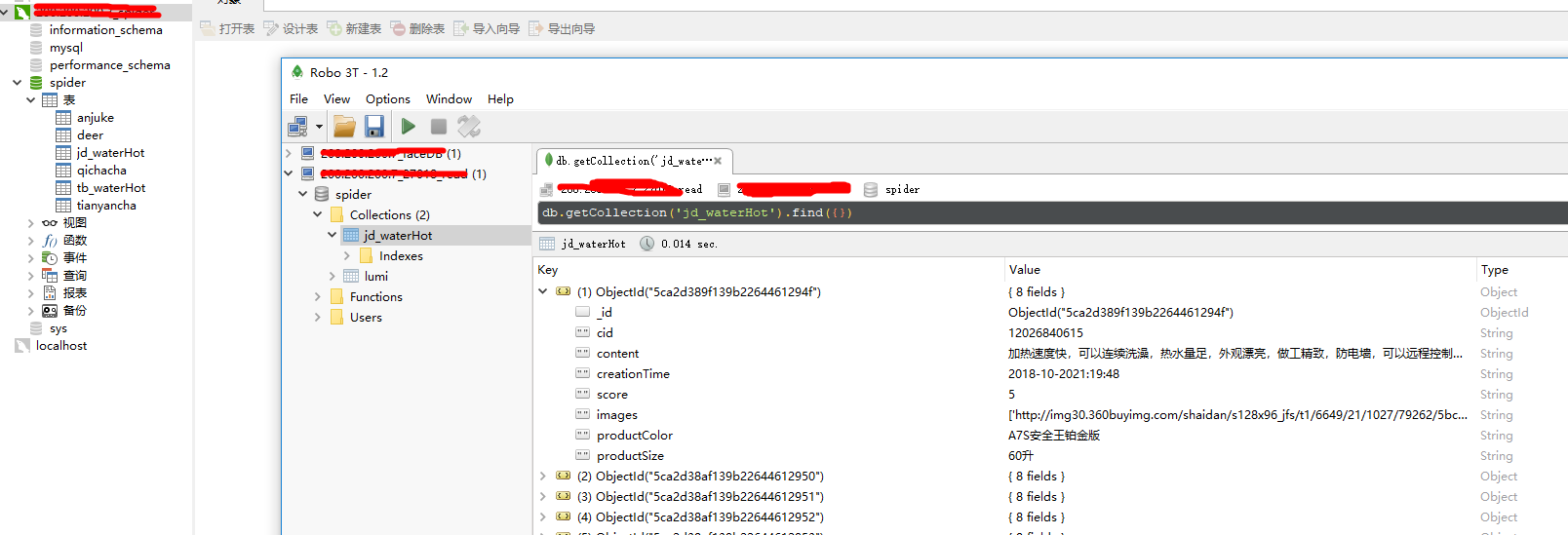
效果图展示：











----------------非Docker部署-----------------------

环境搭建：

1. 环境准备
   1. windows | Linux
   2. Python3 安装包
2. Python && Scrapy环境部署

2.1 安装Python-略

2.2 安装Scrapy以及其他库：

pip install –r packages.txt

2.3 安装Scrapy报Twsited 异常解决：

下载对应Python版本的Twisted：

下载URL： <https://www.lfd.uci.edu/~gohlke/pythonlibs/#twisted>

1. Redis && Mysql && Mongodb 安装

3.1 安装-略 （oneinstack一键安装，但是会比较占磁盘）

3.2 添加远程访问连接-略

1. (可选)安装docker && 部署 scrapy-splash

4.1 安装docker-略

4.2 部署scrapy-splash

作用：js渲染,js事件等

拉取镜像：**docker pull scrapinghub/splash**

创建并运行容器：**docker run -p 8050:8050 scrapinghub/splash**

查看容器运行状态：**docker ps –a**

五．(可选)安装chrome 或 Phantomjs

作用：模拟浏览器操作(滑块验证,js渲染等)

5.1 安装chrome-略

5.2 下载驱动并添加至环境变量(如果使用chrome则驱动需要同 chrome版本对应)

chromeDrive下载地址：<http://chromedriver.storage.googleapis.com/index.html>

PhantomJs下载地址：<http://phantomjs.org/download.html>

六．(可选)安装Gerapy 管理平台

平台作用：爬虫打包部署，运行状态控制等

(Ps: packages.txt 已包含gerapy 库以及其他依赖库）

6.1 添加 scrapyd 和 gerapy 到环境变量

6.2 创建 Gerapy目录并且初始化

**mkdir -p /home/gerapyScrapy**

**cd /home/gerapyScrapy && gerapy init**

**cd /home/gerapyscrapy/gerapy && migrate**

6.3 添加 **8001**防火墙规则

**/sbin/iptables -I INPUT -p tcp --dport 8001 -j ACCEPT**

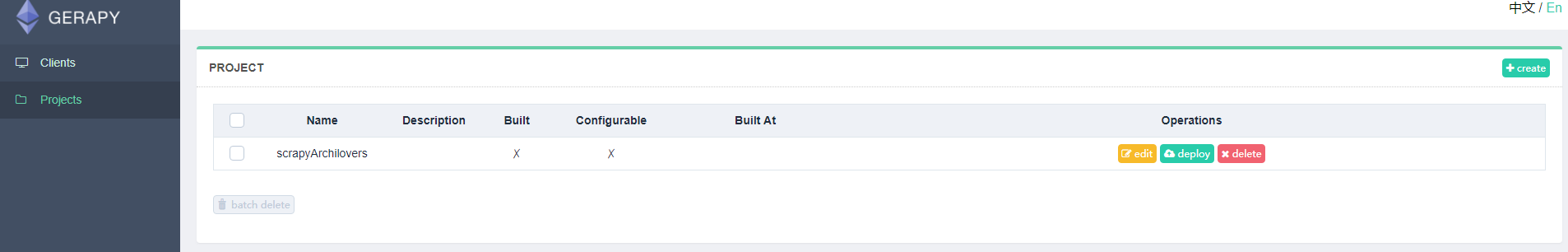
**service iptables save**

# 启动 gerapy 服务

**cd /home/gerapyscrapy/gerapy && gerapy runserver 0.0.0.0:8001**

# 访问 gerapy （此时只能访问gerapy,还不能打包部署，scrapyd服务没起）

<http://ip:8001/>



6.4 (可选)配置Nginx 代理访问 Gerapy

# 作用：通过auth\_basic 认证访问scrapyd,不配置则是公开的

6.4.1 生成auth认证密匙

**htpasswd -c .htpasswd root**

6.4.2 配置Nginx

**service nginx stop**

**vim /usr/local/nginx/conf/nginx.conf**

**server {**

**listen 6801;**

**location / {**

**proxy\_pass http://localhost:6800;**

**auth\_basic 'scrapyd test';**

**auth\_basic\_user\_file /conf/htpasswd;**

**}**

**}**

# 重启 Nginx

**service nginx reload**

6.4.3 修改 scrapyd.conf 文件

**vim /etc/scrapyd/scrapyd.conf**

修改 bind\_address = 0.0.0.0

6.4.4 启动 scrapyd 服务

scrapyd # 命令行执行,(此时Gerapy可以通过scrapyd打包部署爬虫)

七．(可选) WEB UI 可视化爬虫搭建

7.1 下载：git clone <https://github.com/scrapinghub/portia>

7.2 **docker run -v ~/portia\_projects:/app/data/projects:rw -p 9001:9001 scrapinghub/portia**

7.3 访问 ：<http://localhost:9001/#/projects>

7.4 详细使用参考：<http://www.aibbt.com/a/18091.html>

项目部署

八、上传爬虫项目至服务器

(如果部署了gerapy服务，则需要上传爬虫项目到指定目录**/home/ gerapyScrapy/gerapy/projects**)

8.1 爬虫项目修改数据库连接地址

hetSpider.db.config.py

8.2 settings 配置redis，chromeDrive，scrapy-splash等连接地址

8.3 运行 爬虫：

8.3.1 通过Gerapy打包部署运行

8.3.2 通过Python脚本方式运行