scrapydd Documentation

Release 0.4.23

kevenli

Sep 26, 2017

Contents

Scrapydd (Scrapy Distributed Deamon) is a distributed scrapy spiders scheduling system. On the scrapydd system, you can hold and control versions of spider project eggs, schedule spiders, watch job history and logs. It can be also scale out easily.

Contents:

CHAPTER 1

Installation

Requirements

- **tornado** For async programming and the web server.
- **apscheduler** Internal scheduling engine.
- **scrapyd** Project egg storage.
- **sqlalchemy** Data accessing
- **sqlalchemy-migrate** Database migrations.

Installing Scrapydd

By pip:

```
pip install scrapydd
```

You can also install scrapydd manually:

1. Download compressed package from github releases.
2. Decompress the package
3. Run `python setup.py install`

CHAPTER 2

Webhook

Webhook help to support system integrations. When a spider job is completed, the server will start to send crawled data to a customized url.

The webhook post data to `payload_url`, each key/value field is urlencoded before post, unicode data will be treated as UTF8 encoding, and if the value is dict/tuple/list, it will be json enconded. One request for each crawled item.

The frequency of posting data would be no more than 1 request/second.

You can modify spider's webhook settings list this:

```
curl -XPOST http://localhost:6800/projects/{projectname}/spiders/{spidername}/webhook␣
↪-d payload_url = {address}
```

Or to delete an existing webhook:

```
curl -XDELETE http://localhost:6800/projects/{projectname}/spiders/{spidername}/
↪webhook
```

CHAPTER 3

Configuration

Both server and agent use the `scrapydd.conf` file for system configuration. The file will be looked up in the following locations:

- /etc/scrapydd/scrapydd.conf
- /etc/scrapyd/conf.d/*
- scrapydd.conf
- ~/.scrapydd.conf

Server

Server configurations should appears under the `[server]` section.

bind_address

The ipaddress which web server bind on. Default: `0.0.0.0`

bind_port

The port web server running on. Default: `6800`

client_validation

Whether validate client's certificate on SSL, Default: `false`

debug

Whether run server on debug mode. Debug mode will set logging level to DEBUG. Default: `false`.

https_port

HTTPS port to listen on, specify this key will enable SSL mode.

Default: None

server_name

Server's hostname. When SSL enabled, the public certificate will be loaded as filename *server_name*.crt and private certificate will be loaded as filename *server_name*.key in the `keys` folder. Default: `localhost`

Agent

Agent configurations should appears under the `[agent]` section.

debug

Whether run agent on debug mode. Debug mode will set logging level to DEBUG. Default: `false`

server

The IP address or hostname of the server which this agent connect to. Default: `localhost`

server_port

The port of server. Default: `6800`

slots

How many concurrent jobs the agent would run. Default: `1`

request_timeout

Request timeout in seconds when communicating to server. Default: `60`

Example

Here is an example configuration file with all the defaults:

```
[server]
bind_address = 0.0.0.0
bind_port = 6800
debug = false

[agent]
server = localhost
server_port = 6800
debug = false
slots = 1
```