



高效号码自动采集平台 开发使用文档

版本	1.0
作者	商业安全研发部
发布日期	-/-

修订记录

版本	修订内容	修订人	修改日期
1.0	文档撰写	傅正(fuzheng@baidu.com)	2017.07.05
1.1	文档升级	傅正(fuzheng@baidu.com)	2018.05.15

目录

1 项目背景	4
2 流程架构设计	5
2.1 实现方案	5
2.2 架构设计图	5
3 项目使用说明	6

1 项目背景

原有号码采集流程如下：

- (1) 写 Xpath 模板，配置 IP 代理，启动爬虫
- (2) 根据模板写数据清洗脚本
- (3) 启动多进程执行脚本处理(单独运行另外脚本监控)
- (4) 运行脚本进行数据抽样审核
- (5) 手动切分数据，执行脚本进行入库

由于服务器计算能力有限，并且常见的大型网站，如美团，大众点评，都有完备的反爬虫系统(如验证码，IP 屏蔽，Session 验证，日志分析等)，常见的爬虫技术(Scrapy + PhantomJs + Selenium)应对需要消耗巨大的资源。而且对于美团这类大型网站，及时顺利爬取，以单台服务器计算也需要耗费数十天，爬取下来后，上亿原始网页的存储都存在严峻的挑战。

这仅仅是在爬取阶段，在后续的数据清洗阶段，对每一个网页，因为结构不同，都需要重新改写脚本，还需要启用额外的监控进程对发现的问题进行更改，即使提取成功，面对上百 GB 甚至 PB 的文件，还需要做到随机抽样，手动切分数据，这些操作更是步履维艰。而且对于一个网站，往往需要重复上述流程，这对人力资源、计算资源，都是巨大的浪费。

摒弃传统数据采集方法(如 Scrapy、PhantomJs、Selenium)，通过百度中文缓存库获取数据源，并提供集数据清洗自定义配置、OCR 文本识别、周期采集、任务进度实时跟踪、自定义语法解析、超大数据(百 GB 以上)随机评审等功能的高效率自动采集平台(最高支持 6000 万/每小时数据处理)，实现一键采集、一键入库，任何人都可 0 门槛实现号码数据采集。

通过对上述流程进行模块化封装，简化流程如下：

- (1) 选择配置项
- (2) 一键采集/一键评审/一键重启/一键入库

2 流程架构设计

2.1 实现原理

一、整体架构:采用 Django + gunicorn(+gevent) + MongoDB 作为平台框架，其中 Django 支持 APP 可插拔便于项目快速迭代，gunicorn 支持协程，结合支持数据结构模式自由的 MongoDB，对外提供 RESTful API，提供多进程、高性能服务；

二、数据爬取:以百度中文缓存库来源，通过 Hadoop 获得满足条件的数据(基于百度内 PIE 结构化数据抽取平台提供的 API 接口),以 FTP 数据流的形式返回提取数据供后续处理；

三、数据处理:

- (1) 使用 WebSocket 协议，实现全双工通信，实时监控数据处理进程；
- (2) 结合 Celery + Redis，实现任务异步化，队列化，定时执行，以支持重复采集与高耗时任务后台执行；
- (3) 通过 mmap(内存映射文件)，实现对超大数据的随机访问，进而实现号码数据随机抽样评审；
- (4) 为解决 Python 中 GIL 造成单任务 CPU 使用率不高问题，使用多进程池，并结合共享内存实现 Queue 队列，最终实现对数据提取任务，单进程(FTP 数据流)生产数据，多进程消费模型；
- (5) 通过百度 IDL 提供的文字识别接口，对指定数据提供 OCR 识别能力；
- (6) 支持通过上下文对号码进行自动分类；

2.2 实现原理

由于架构图过大，架构图单独放在文件夹内。

3 项目使用说明

以下将展示平台使用，数据处理端完全无需用户处理，只需几个按钮即可0门槛完成数据采集任务。

(1) 平台监控:

- 1、近 21 天（包含全部时间）总体数据提取入库趋势图，如图 1
- 2、近 7 天（包含全部时间）每个任务数据提取入库趋势图，如图 2

网页通用集成平台



图 1



图 2

(2) 新增任务

点击主页面列表框上方“新增任务”按钮，进行新增任务操作，“新增任务”页面如图 3 所示。

新建任务

若想要新建模板，请先[点击这里](#)

选择模板: 请选择下列模板

新增模板: 请输入新增模板名(可选)

请填写以下属性与模板中对应字段

结构名: 默认为'默认类别，可改名'

号码: 请完整输入号码对应字段 | 指定值(可选) | OCR ☐

名字: 请完整输入名字对应字段 | 指定值(可选) | OCR ☐

类别: 请选择号码所属类别

[更多映射配置](#)
[字段自定义组合](#)

任务名: 请自定义任务名

URLpattern: 请输入目标URL模式,如http://(w*).anjuke.com/prop/view/A(id*)

关闭 启动

图 3

图中红框为必填项，绿框部分可展开是可选项，绿框展开后如图 4。

任务详情

以下内容为任务配置与数据抽取日志

原始文件
入库文件
日志文件

任务配置信息

结构名:

号码:

指定值(可选)

OCR ☐

名字:

指定值(可选)

OCR ☐

类别:

城市:

指定值(可选)

OCR ☐

地址:

指定值(可选)

OCR ☐

介绍:

指定值(可选)

OCR ☐

标签:

指定值(可选)

OCR ☐

组合句柄:

保存
取消

任务名

模板名称

URLpattern

数据提取入库趋势图

图 4

图 4 绿框部分为可选项，新建任务中各输入框功能将在下述介绍。点击蓝框按钮，如数据是图片，则可支持 OCR 识别，未配置模块，请先点击图 3 中红色背景框中提示的“这里”进入 PIE 的可视化模块配置页面,如图 5



图 5

图 5 为数据爬取可视化配置界面，在 Url 输入之后，会展示抓取界面，然后点击右侧 界面中需要抓取的元素之后，为其取别名，如图 5 中红框部分 name,phone,addresss 分别填至图 4 中的名字，号码，地址。图 5 绿框中是类别结构名，可自定义。图 5 蓝框部分是 模板名称对应图 4 中新增模板（若选择模板下拉框有则无需填写），配置完后点击“创建模板”，返回号码自动采集平台。

以下介绍采集平台中新增任务各字段对应功能（对照图 4）

选择模板:选择对应要抓取的网页，若没有可点击上方新建。

新增模板（选择模板和新增模板至少填写一个）:无期望模板，在新建之后，填写新建的模板名。

结构名（选填）:提取结构 key，即对应应在图 6 中绿框部分。

类别（必填）:为当前任务提取的数据的类别行业，下拉选取。此外，若选择“自动分类”，则将会根据号码数据上下文进行自动分类。若选择“暂不分类”，则号码类别为空。

号码（必填），名字（必填），城市（选填），地址（选填），介绍（选填），标签（选填）均为图 6 中红框对应别名字段，按照自身命名对应填写即可，这些字段后都有“指定值（可选）”输入框，都是可选项，填写后，产出的对应字段数据即为输入的值，而与原始数据无关。

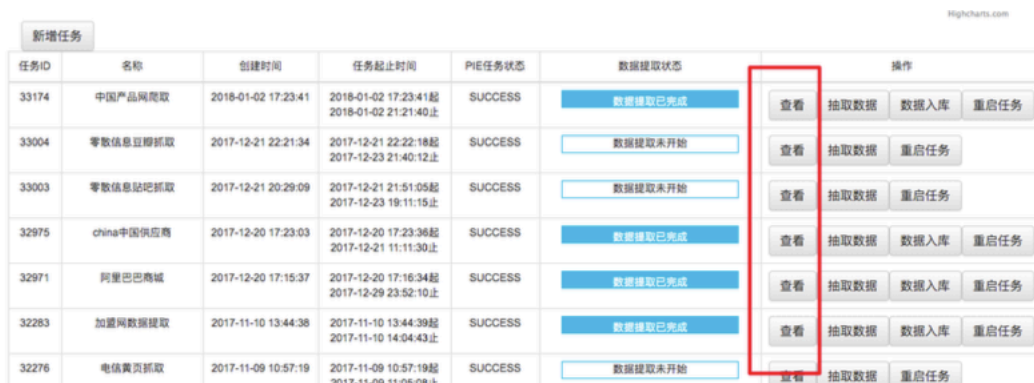
任务名（必填）:可自己填写喜欢的任务名。

组合句柄（选填）:如果想要对产出数据进行随意组合，如用户想让最后得到的地址是地址与名字的组合，即 地址=地址+名字，那么只需要输入 address=address+city (其中的单词为图 5 中绿框的所填字段)，再如用户想在所有的名字前加上“中国”，那么 只需输入 name="中国"+name。

URLpattern (必填):输入期望爬取的网站网址正则表达式，如提取太平洋汽车网，我们找到的一个页面为
<http://price.pcauto.com.cn/70907/contact.html>，则该网站的网 址集合
http://price.pcauto.com.cn/\d*/contact.html

(3) 查看功能

如图 6，在主菜单界面点击展看后，会展示任务详情悬浮层，如图 7



任务ID	名称	创建时间	任务起止时间	PVE任务状态	数据提取状态	操作
33174	中国产品网抓取	2018-01-02 17:23:41	2018-01-02 17:23:41起 2018-01-02 21:21:40止	SUCCESS	数据提取已完成	查看 抽取数据 数据入库 重启任务
33004	零散信息豆瓣抓取	2017-12-21 22:21:34	2017-12-21 22:22:18起 2017-12-23 21:40:12止	SUCCESS	数据提取未开始	查看 抽取数据 重启任务
33003	零散信息贴吧抓取	2017-12-21 20:29:09	2017-12-21 21:51:05起 2017-12-23 19:11:15止	SUCCESS	数据提取未开始	查看 抽取数据 重启任务
32975	china中国供应商	2017-12-20 17:23:03	2017-12-20 17:23:36起 2017-12-21 11:11:30止	SUCCESS	数据提取已完成	查看 抽取数据 数据入库 重启任务
32971	阿里巴巴商城	2017-12-20 17:15:37	2017-12-20 17:16:34起 2017-12-29 23:52:10止	SUCCESS	数据提取已完成	查看 抽取数据 数据入库 重启任务
32283	加置网数据提取	2017-11-10 13:44:38	2017-11-10 13:44:39起 2017-11-10 14:04:43止	SUCCESS	数据提取已完成	查看 抽取数据 数据入库 重启任务
32276	电信黄页抓取	2017-11-09 10:57:19	2017-11-09 10:57:19起 2017-11-09 11:05:08止	SUCCESS	数据提取未开始	查看 抽取数据 重启任务

图 6



图 7

任务详情中，最上方“原始文件”，“入库文件”，“日志文件”按钮，可分别查看原始数据（超大文件，图8）、按照用户个性配置生成的结果文件（超大文件，图9）、操作日志，支持超大文件随机抽取。

```
http://bd.meituan.com/shop/51590136 {"template_name": "美团-酒店客栈-模板2.xml", "update_time": "1499986153", "user_id": "美团-酒店客栈-模板2",
"@has_userresult": "false", "url_md5": "31aa523fa85db007d9ad748e0c1cd54a", "pagecrawl_time": "1499986151", "to_rawbase": "false", "meituan_hotel_2": {"url":
"http://bd.meituan.com/shop/51590136", "pagecrawl_time": "1499986151", "update_time": "1499986153", "url_md5": "31aa523fa85db007d9ad748e0c1cd54a", "url":
"http://bd.meituan.com/shop/51590136", "create_time": "1499986153", "source": "pie", "pattern": "http://.*\\meituan.com/shop/\\d+", "is_uniform": "true", "@id":
"http://bd.meituan.com/shop/51590136", "template_id": "20647"}

http://bj.meituan.com/shop/119606044 {"template_name": "美团-酒店客栈-模板2.xml", "update_time": "1500400325", "user_id": "美团-酒店客栈-模板2",
"@has_userresult": "false", "url_md5": "af95cff4be61c3374f45c025ed0c9647", "pagecrawl_time": "1482679764", "to_rawbase": "false", "meituan_hotel_2": {"url":
"http://bj.meituan.com/shop/119606044", "pagecrawl_time": "1482679764", "update_time": "1500400325", "url_md5": "af95cff4be61c3374f45c025ed0c9647", "url":
"http://bj.meituan.com/shop/119606044", "create_time": "1482679825", "source": "pie", "pattern": "http://.*\\meituan.com/shop/\\d+", "is_uniform": "true", "@id":
"http://bj.meituan.com/shop/119606044", "template_id": "20647"}

http://bj.meituan.com/shop/1381512 {"template_name": "美团-酒店客栈-模板2.xml", "update_time": "1500375198", "user_id": "美团-酒店客栈-模板2", "@has_userresult":
"false", "url_md5": "375ab2041752e744f99a3ef236f81180", "pagecrawl_time": "1482964287", "to_rawbase": "false", "meituan_hotel_2": {"url":
"http://bj.meituan.com/shop/1381512", "pagecrawl_time": "1482964287", "update_time": "1500375198", "url_md5": "375ab2041752e744f99a3ef236f81180", "url":
"http://bj.meituan.com/shop/1381512", "create_time": "1482964288", "source": "pie", "pattern": "http://.*\\meituan.com/shop/\\d+", "is_uniform": "true", "@id":
"http://bj.meituan.com/shop/1381512", "template_id": "20647"}
```

图 8

```
{"category": {"$id": "10", "$ref": "Category"}, "city": "万宁市", "sourceurl": "http://aks.58.com/zufang/30101594651726x.shtml", "name": "董斌", "source": "spider_pie", "phone": [{"number": "18139102888"}], "address": "万宁市", "op": "UPDATE"}

{"category": {"$id": "10", "$ref": "Category"}, "city": "万宁市", "sourceurl": "http://ankang.58.com/zufang/17245050209285x.shtml", "name": "好房", "source": "spider_pie", "phone": [{"number": "18992512856"}], "address": "万宁市", "op": "UPDATE"}

{"category": {"$id": "10", "$ref": "Category"}, "city": "万宁市", "sourceurl": "http://as.58.com/zufang/29260852336437x.shtml", "name": "女士", "source": "spider_pie", "phone": [{"number": "13674226559"}], "address": "万宁市", "op": "UPDATE"}

{"category": {"$id": "10", "$ref": "Category"}, "city": "万宁市", "sourceurl": "http://ay.58.com/zufang/19184738649606x.shtml", "name": "广厦中介", "source": "spider_pie", "phone": [{"number": "13323622593"}], "address": "万宁市", "op": "UPDATE"}

{"category": {"$id": "10", "$ref": "Category"}, "city": "万宁市", "sourceurl": "http://ay.58.com/zufang/19759768090757x.shtml", "name": "史女士", "source": "spider_pie", "phone": [{"number": "13353722651"}], "address": "万宁市", "op": "UPDATE"}
```

图 9

(4) 数据处理实时监控

在爬虫任务主界面，通过 WebSocket 实时展现数据爬取任务进展，并实时更新爬取的号码数量，如图 10。

新增任务						
任务ID	名称	创建时间	任务起止时间	PIE任务状态	数据提取状态	操作
31496	天颐堂	2017-09-18 11:52:01	2017-09-18 11:52:01起 2017-12-20 22:33:53止	SUCCESS	数据提取已完成	查看 抽取数据 数据入库 重启任务
31495	慧聪网公司	2017-09-18 11:37:31	2017-09-18 11:37:31起 2017-11-15 12:17:35止	SUCCESS	数据提取已完成	查看 抽取数据 数据入库 重启任务
31224	点评亲子摄影类	2017-09-01 16:35:48	2017-09-01 16:35:48起 2017-09-04 09:43:18止	SUCCESS	数据提取已完成	查看 抽取数据 数据入库 重启任务
31066	快速100网点	2017-08-25 11:45:13	2017-08-25 11:45:13起 2018-01-02 17:34:25止	SUCCESS	7% 提取9649条数据	查看 抽取数据 重启任务
31062	糯米商铺	2017-08-24 22:11:22	2017-08-24 22:11:22起 2017-10-10 00:00:00止	SUCCESS	数据提取已完成	查看 抽取数据 数据入库 重启任务

图 10

(5) 数据入库

如图 6 当数据提取状态栏显示数据提取已完成时，点击数据入库即可

(6) 任务重启

如图 6，点击重启任务即可