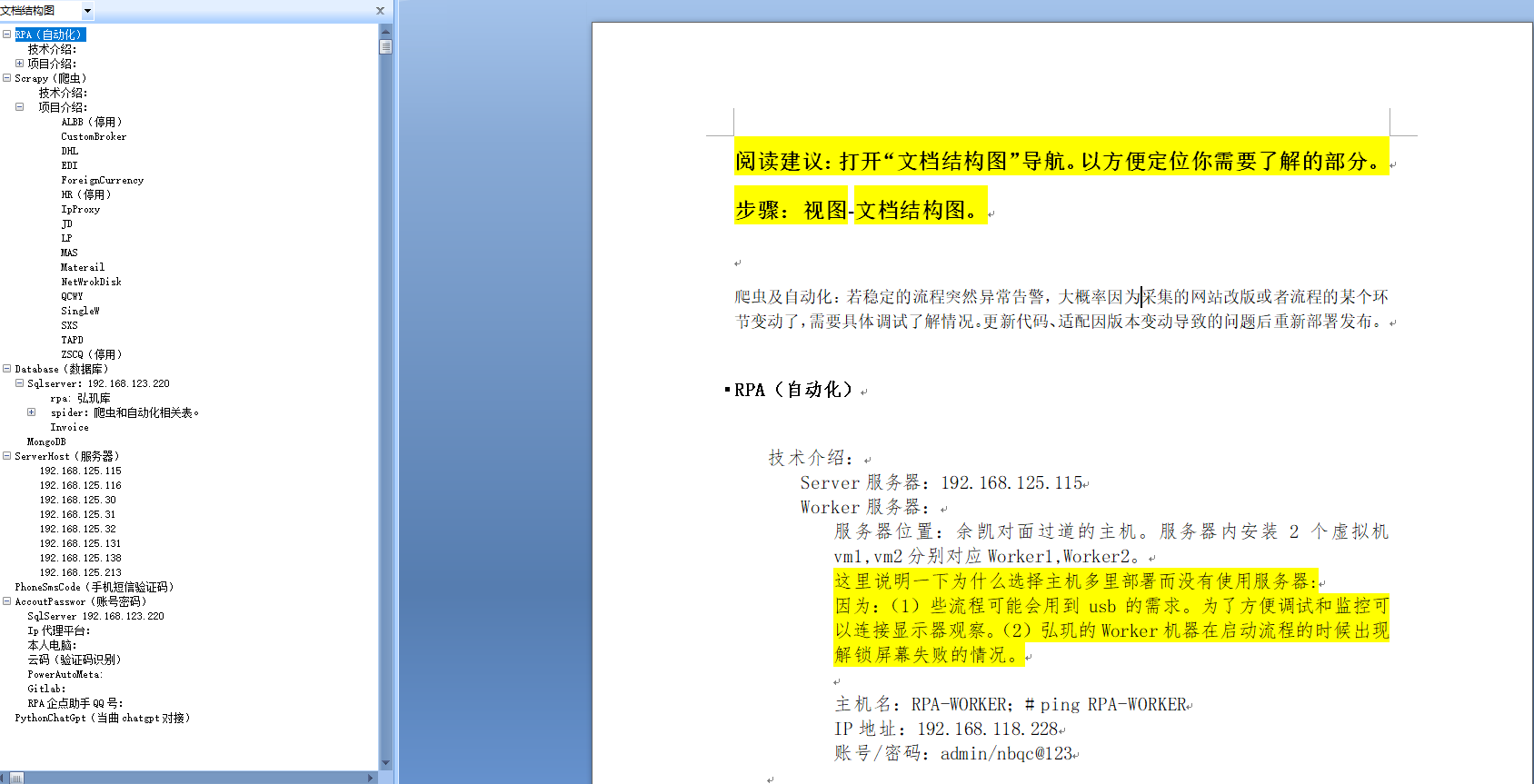
**阅读建议：打开“文档结构图”导航。以方便定位你需要了解的部分。**

**步骤：视图-文档结构图。**

****

爬虫及自动化：若稳定的流程突然异常告警，大概率因为采集的网站改版或者流程的某个环节变动了，需要具体调试了解情况。更新代码、适配因版本变动导致的问题后重新部署发布。

# RPA（自动化）

技术介绍：

Server服务器：192.168.125.116

Worker服务器：

服务器位置：余凯对面过道的主机。服务器内安装2个虚拟机vm1,vm2分别对应Worker1,Worker2。

这里说明一下为什么选择主机多里部署而没有使用服务器:

因为：（1）些流程可能会用到usb的需求。为了方便调试和监控可以连接显示器观察。（2）弘玑的Worker机器在启动流程的时候出现解锁屏幕失败的情况。

主机名：RPA-WORKER；# ping RPA-WORKER

IP地址：192.168.118.228

账号/密码：admin/nbqc@123

桌面存放弘玑5.0安装包。

管理后台：<http://cyclone.lumi.cn>

管理员账号密码：admin/nbqc@123

部署方式：docker

源码路径：/data

MinioBroket:

ciriservicefilesystem  
commonservicefilesystem  
orcserverfilesystem  
screencaptureorcfilesystem

Database:Sqlserver-192.168.123.220；rpa库

详情可以查看配置文件：

/data/c3po-5.0.RC/user\_param.yaml

cyclone.lumi.cn域名为内部dns定义。内部nginx做代理。

出现问题请联系弘玑人员处理。

项目介绍：

**舱单Npsel**

仓库地址：

<https://gitlab.powertradepro.com/zhaohui/npsel.git>

网站：<http://www.npsel.com/plan/manifest_goadd.do>

需要3个入参

Code：合同号，UserId：员工编号，UserName：员工姓名

启动方式：弘玑管理后台可以配置接口启动流程的方式。由ERP调用接口并传递参数。

流程中会调用这个接口获取需要填写的表单数据（冯之梁）

http://192.168.123.218:8000/api/services/Declaration/Declaration/GetWarehouseReceipt

**舱单Sino56**

仓库地址：

<https://gitlab.powertradepro.com/zhaohui/rpa-sina56.git>

同“舱单Npsel”一样的逻辑

网站：

http://www.sino56.com/sino-manifest/service/manifest/to-edit

**滴滴**

仓库地址：

<https://gitlab.powertradepro.com/zhaohui/rpa-dd.git>

登录滴滴企业版网站，操作新增，删除员工。更新员工部门属性。

读取SqlServer spider dd\_auto 表，根据type执行不同操作（新增，删除，变动），status字段是该记录的执行状态，是否操作成功。

操作流程：（1）登录网站（2）填写需要新增的员工（3）填写要删除的员工（4）更新部门部门变动的员工。每条记录执行成功都会将status设置为true。

dd\_auto表数据是由ERP写入（郭晨涛）

弘及管理后台可以查看定时任务。目前是每天18:30。

**数字发票**

仓库地址：

<https://gitlab.powertradepro.com/zhaohui/SZFP.git>

登录国家电子发票网站，遍历所有子公司，采集所有发票。上传到当曲云。

登录地址：

<https://tpass.ningbo.chinatax.gov.cn:8443/#/login?redirect_uri=https%3A%2F%2Fetax.ningbo.chinatax.gov.cn%2Flogin-web%2Fapi%2Fthird%2Fsso%2Flogin%2Fredirect%3Fqd%3Dkexin%26channelId%3Dweb%26goto%3D30010666&client_id=wwba3777ce294989b8226ccwb28w6a3d&response_type=code&state=prod>

账号/密码：9133021277231080XU/13454765982/nbqc2023

登录需要短信验证码：阅读文档“手机短信验证码”部分。

发票查询模块地址：

<https://dppt.ningbo.chinatax.gov.cn:8443/invoice-query/invoice-query>

发票类型：取得发票

发票种类：增值税专用发票，普通发票。

查询结果的所有发票 下载pdf，xml文件。Xml文件需要解压后上传。

调用平台附件上传接口（李伟）：/filemanage/api/app/file/upload，之后将返回minio路径上传到当曲云逻辑编排（余凯，邢婉秋）。

上传当曲云附件逻辑编排：6461c51dcd357e7012bbc5af。

下载前会判断当曲云是否已经存在此票。

下载前判断是否存在立即编排64643a06204768dae985b990

流程设置两个入参：IsAll，Env

Isall: 若为false表示只采集第一页的100条数据，不会翻页。若为true表示采集所有查询结果会翻页查到最后，但是查询条件中的时间区间默认的不会设置。

Env：若为test表示对接当曲测试环境，若为prod表示正式环境。

流程中设计了一些容错机制：正常情况报错不会影响这个程序的终止运行，程序会重新登录从头开始执行。但是若连续的3次报错会终止整个流程，并发送企业微信异常告警通知。

当xml和pdf都下载成功才会上传当曲云逻辑编排，若其中一个下载失败不会上传。（需求邢婉秋，余凯）。

**医社保**

仓库地址：

<https://gitlab.powertradepro.com/zhaohui/rpa-yishebao.git>

自动办理员工申报，停保医社保任务。

浙江省医社保现在可以一站办理，在一个页面填写员工信息表单数据。读取数据表SqlServer spider hr\_auto。

登录需要短信验证码：阅读“手机短信验证码”部分。

网址：https://oauth.zjzwfw.gov.cn/login.jsp

账号密码：偶尔HR会更换密码的情况，所以我就不写在文档中了。现在的账号密码可以在弘玑管理后台-数字资产中查看。及定时任务。

根据表中adjuse\_date字段：调整时间。

查询调整时间为当月并小于等于当天的type字段为社保基数status 字段为false的记录，再根据salary来执行对应的操作。

salary等于0 ：停保

salary不等于0：申报

操作成功后会更新status为true。若人事（蔡燕娜）手动处理了某记录，此时数据库status仍然为false，待任务执行次记录时会发送异常通知，跟蔡蔡沟通后可设置手动更新数据库status为true。

**智慧牙**

仓库地址：

<https://gitlab.powertradepro.com/zhaohui/rpa-zhy.git>

根据SqlServer spider zhihuiya\_search表中的关键词去智慧牙网站搜索并将结果保存数据库。

**资金流水的项目（停用）**

因为运行不稳定

**企点群助手**

仓库地址：

<https://gitlab.powertradepro.com/zhaohui/qidian.git>

遍历执行SqlServer spider qq\_group\_user表中的记录执行对应的操作（群组加人，群组删人）。

定时任务配置在弘玑管理后台。

注意：有个Bug：每次服务器重启后会出现。在群组加人的时候，会操作企点客户端，在群组对话框点击添加人员，搜索人员qq号的对话框元素定位失败。解决办法：打开设计器重新捕获元素。而且这个RPA流程开发是在worker服务器（vm1）开发的。因为在操作企点客户端的时候不同机器的元素定位会不一致。如果跨机器开发要注意开发环境腾讯企点的版本号。

**公积金**

公积金RPA使用的工具是PowerAutomate，企业版试用到期了，所以需要注册个人版是免费的。

读取的数据SqlServer spider hr\_auto表，type字段为“公积金基数”

可以将文件内容复制到power automate设计器中创建流程，流程名要和文件名一致。

公积金的流程文件位置：

[**https://gitlab.powertradepro.com/zhaohui/gjj.git**](https://gitlab.powertradepro.com/zhaohui/gjj.git)

# Scrapy（爬虫）

技术介绍：

**管理后台**：http://shuju.lumi.cn

服务器：192.168.125.115

路径：/root/spider-web

框架：Flask,Vue

Python环境：conda activate spider-admin

完全可以不用这个管理后台，因为有很多开源scrapy管理后台工具。因为开源的管理后台出现的比较晚，当初为了方便所开发，因个人习惯一直使用。

**爬虫开发**：

框架：scrapy

仓库地址：

<https://gitlab.powertradepro.com/zhaohui/LumiSpider.git>

部署：

代码映射到192.168.125.115服务器/root/LumiSpider路径

使用scrapyd命令部署scrapy项目。

例:

1. 进入python虚拟环境 conda activate spider
2. cd /root/LumiSpider/EDI & scrapyd-deploy edi-node1

说明：edi-node1 是scrapy.cfg中配置

提交之后可以再管理后台 Product模块中查看到。可能有1分钟的延迟。

启动方式：

(1) 可以在管理后台启动

(2) 通过http请求方式：192.168.125.115:6800。阅读scrapyd官方文档。

项目介绍：

ALBB（停用）

阿里巴巴国际站改价，对接国际站改价接口。

需要改价的记录读取CdcRecord表。

CustomBroker

采集[http://47.110.59.188:8080/?ac=%s](http://47.110.59.188:8080/?ac=%25s)站点。

遍历custom\_broker\_company中的公司名称ac表示报关行代码，ac\_code 列表代码中写死。每次采集3天的数据。下载Pdf，解析pdf数据保存数据库。如果出现某字段不正常为空，可能是因为pdf格式有变动解析失败。需要具体调试分析。解析工具： pdfplumber。Github开源

管理后台可查看定时任务。

DHL

填写中外运表单数据。

登录，验证码 云码 打码平台，关注费用

读取数据SqlServer spider dhl\_auto表，读取update\_datetime字段为空的记录，在任务执行完次记录之后将update\_dateime字段更新。

dhl\_auto数据由ERP写入（郭晨涛）

填写表单4步骤

Update\_zero

Update\_one

Update\_two

Update\_three

具体可以debug流程，查看是如何运行的。

DocxTokenReplace

.doc文档占位符工具

根据excel中的列名替换.doc模版中的占位符。具体使用方法可以找（施宇超）了解。

EDI

采集EDI （https://www.npedi.com/index）站点开截港时间。

账号可以在vessel\_account表维护。

程序每次随机使用一个账号登录。

遍历vessel\_notice表中的数据，查询没单的开截港时间。

查询到开截港时间则推送消息：

推送业务员 yw\_user字段，每天推送。

推送单证员 dz\_user 字段，节假日推送。

推送管理员 10867|10083|10072，节假日推送。

ForeignCurrency

汇率采集，数据保存foreign\_currency表。

采集中国银行和浦发银行的汇率数据。

定时任务可以在管理后台查看。

中国银行对IP的访问限制有点严格。

使用IP代理，防止公司IP被封。

Ip代理可以在mongodb spidre ip\_proxy表获取。

HR（停用）

原医社保自动化，已经被弘玑替换。

功能：

（1）读取sqlserver hr\_auto 表。登录社保网站填写停保，申报的表单。成功之后将记录status字段改为true。

（2）采集在保人员名单，保存hr\_auto\_user表

因社保网站改版较大，具体请调试修改。

因为网页经常改动，还会出现一些弹窗的情况，定时任务不稳定。

IpProxy

维护本地可用的Ip代理池。获取太阳代理我们已经购买的套餐

数据保存mongodb spider ip\_proxy 表。 mongodb://spider:bestadmin@mongo.mongo.svc.cluster.local:27017/spider

管理后台可查看定时任务，目前频率可保持全天表能有代理可用。

套餐是短效代理：5-25分钟。

JD

DisplayTop

采集显示器top数据，遍历尺寸列表。采集每种尺寸的top产品。

数据保存sqlserver tv\_params\_top500

TvTop

采集Tv top数据，遍历尺寸列表。采集每种尺寸Tv的top产品。

数据保存sqlserver tv\_params\_top500

watch\_stock

监控京东商品库存情况，并发送缺货通知。

已店铺为单位，采集店铺下所有商品，每个商品切换城市，判断每个城市下商品库存的状态。将缺货的商品保存到excel文件中。等全部采集完毕，企业微信发送通知给对应的运营人员。

管理后台查看定时任务。

LP

猎聘简历采集，通过邮件的方式上传到当曲云。

登录：猎聘网登录需要滑动验证验证。此验证码无法通过打码平台识别。因为没有底图，只有一张带缺口图。这个图片是后端生成直接返回，不能通过普通的缺口验证码比对方法计算出的缺口距离。所以目前只能人工配合滑动验证。（施宇超操作）

密码HR可能会变动，因为密码在代码写死。所以需要重新编译发布。

编译步骤：

cd LP/

pyinstaller –F main.py

需要注意：

chrome驱动是可执行程序的相对路径./drivers/chromedriver.exe

部署：堡垒机JumpServer对应的自动化/简历1服务器

MAS

中国移动手机号审核平台。

登录，数字验证码。云码识别。

通过6307411fc27cef8d5824249f逻辑编排获取要审核的手机号

提交 审核表单

再获取审核通过的手机号，通过6319a202f136adfb2d47ce66逻辑编排提交审核状态。

管理后台查看定时任务

Materail

原材料数据采集

一共采集5个网站

保太集团：

http://www.baotaigroup.com.cn/index/offer/hoffer?id=6

上海有色网：https://user.smm.cn/login

卓创网：https://prices.sci99.com/cn/

中塑网：https://quote.21cp.com/zs\_login

我的钢铁APP：

采集列表数据保存数据库 sqlserver materail表

注意：

（1）网页页面变动，需要调试适配解析规则

（2）如果账号密码变动，修改代码对应的密码变量值，重新发布打包，需改定时任务版本

NetWrokDisk

网盘日志同步到ElasticSearch数据库

记录每次采集的最有一条数据保存到Mongodb network\_disk表，之后下次采集的时候遇到这条数据则停止采集后面的数据。

去重用到Redis数据库，192.168.125.115

'ITEMS\_DROP\_DUPLICATES' : {

'NetworkDiskESItem': ['fp']

},

设置根据fp字段去重。

item['fp'] = log.get('ctime') + log.get('user\_name') + item['action'] + str(log.get('operand', ''))

fp字段是根据ctime，username，action， operand拼接。

详细可在代码中查看。

QCWY

前程无忧简历采集，通过邮件上传到当曲云平台。

登录需要验证码，与LP猎聘不同，它的验证码可以根据底图和缺口图进行比较计算滑动距离实现自动验证。

def get\_distance(img1, img2):

start\_x = 0

gap = 5

threhold = 60

for x in range(start\_x, img1.size[0]):

for y in range(img1.size[1]):

rgb1 = img1.load()[x, y]

rgb2 = img2.load()[x, y]

res1 = abs(rgb1[0] - rgb2[0])

res2 = abs(rgb1[1] - rgb2[1])

res3 = abs(rgb1[2] - rgb2[2])

if not (res1 < threhold and res2 < threhold and res3 < threhold):

return x – gap

这段是计算距离方法。

功能：（1）采集 未读 列表的简历。通过发送邮件传递给当曲云。记录保存sqlserver sync\_resume表，采集之后标记投递者为通过筛选。这样这条记录不会出现在未读列表中。（2）将当曲对投递者简历的标记反馈前程无忧，通过625fc6ad54625109b281eab6逻辑编排获取待反馈的数据，根据数据库保存的投递者详情页面连接，进入该投递者详情页并标记（合适/不合适）状态。之后调用625f581105758a24da0d811c逻辑编排通知当曲是否标记成功。

注意：通猎聘一样，密码变动需要修改代码Password值，重新编译部署。部署到 jumpServer（简历1）服务器exe放在桌面。

注意一下：简历1服务器桌面 drivers/chromedriver.exe驱动文件。

通过wineows 计划任务设置定时。

SingleW

单一窗口自动化工具。

细节流程通过调试代码或者运行exe了解。我会在类和函数写上相应的注释。

功能：

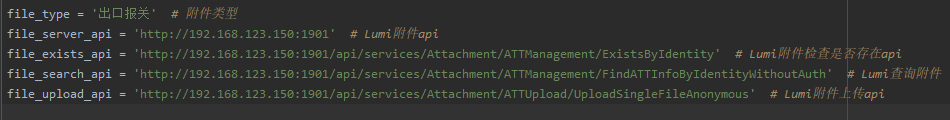
1. 采集报关单时间区间1个月内的数据。

Class Spider

('一般报关单', '转关提前报关单')

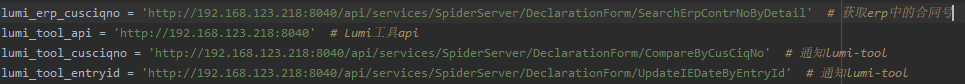
('结关', '放行', '退单', '删单')

预录单高级查询模块：报关单类型和报关单状态交叉遍历查询预录单并下载预录单附件和通知书附件。



调用ERP的附件接口上传附件。

采集完毕之后通知LumiTool工具端。



注意：合同号不规则，由于单一窗口采集的报关单合同号不是我们规定的合同号类型，需要嗲用lumi\_erp\_cusciqno接口纠正合同号，再保存数据库 invoice.spider.DeclarationForms表。

注意：单一窗口预录详情地址是动态的会定期更换，需要在每次在js文件中判断是否详情地址变动。并获取最新的URL地址。

注意：单一窗口预录单详情数据是加密的，但是有一个解密的js文件，可能也会变动，你需要抓包分析。并注入到爬虫程序中已获取数据并解密入库。

注意：在解密失败时，可以用ExtensionDataPDF字段，此字段是解析pdf的数据，经过解析已转换为ExtensionData字段结构化一样的格式。但是也需要关注解析的pdf格式是否发生变化。

意外情况：个别预录单采集的时候可能还没有附件，后续才有的所以附件的更新有问题。解决办法数据库删除此记录重新采集。情况很少

1. 签订委托协议。

Class AutoWT

读取sqlserver spider sw\_auto\_weituo 表。Status字段不为true并且 yld\_date字段为null的是需要发送委托的数据。

进入 签订委托协议 模块，填写委托表单。判断此报关行是否存在委托协议，不存在则先去签订报关行委托协议。

签订报关行委托协议勾选项：长期，取消自动确认，勾选辅助查验。并更新sw\_auto\_weituo表bgh\_status字段为True，

bgh\_status

True：申请中，

False：不存在委托协议还未申请。

None：可以发送委托，但是委托发送前要判断是否存在委托协议。

签订委托协议提交后会生成协议号，更新sw\_auto\_weituo表protocol\_code字段。

1. 续保报关行委托关系。

Class WtContinue

功能：（1）采集存在委托关系的报关行，保存到sqlserver spider sw\_wt\_deadline表。

（2）遍历sw\_wt\_deadline表中的报关行，判断过期时间字段是否到期，到期则重新续签委托关系。

编译&部署：

进入SingleW目录

pinstaller –F main.py

将dist/main.exe部署到RPA-DANZHENG服务器。

服务器IP地址：192.168.118.158

服务器账号/密码：admin/123

服务器路径：C:\Users\admin\spider\signle-automation

服务器位于19楼IT部交换机房。

服务器内创建3台虚拟机 vm1 vm2 vm3 对应插着3张U盾（金税盘）

每台vm桌面都有never\_lock.vbs脚本（双击执行防止vm虚拟机黑屏的情况。）

注意：告警

wechat.send\_message('10867|10124', '单一窗口自动化：登录失败\nIP：%s' % localaddr)

通知人员代码写死（赵辉|郭晨涛）

SXS

实习僧简历采集工具，通过邮件上传当曲云平台。同（前程无忧，猎聘）差不多，（1）采集简历，（2）反馈标记状态。

部署：简历1服务器。部署方式一样。

TAPD

目前只有一个任务还在使用 TapdErpSpider。

将Tapd和需求的状态更新ERP的IT建议状态。

调用ERP接口获取IT建议

<http://192.168.123.218:8000/api/services/Common/Common/GetVTaskItem>

sjjjsj ：tapd的开发完成时间。

releaseDate：tapd 预计发布时间。

调用更新接口更新字段

http://192.168.123.218:8000/api/services/Common/Common/UpdateERPJianYiChild

ZSCQ（停用）

因为网站改版，图片连接无法直接访问。

ZsqcSpider：采集知识产权网站上所有数据，每次采集前情况数据库中的数据。周期：每3个月

问题：采集开始的时候，老数据会情况。导致 "LUMI产权查询" 功能无法正常使用。避免影响：改为20:00之后采集。

ZscqCycleSpider：采集mssql的Zscq表中update\_datetime为3个月内的数据。周期：每1天。数据存在则更新。

# Database（数据库）

Sqlserver：192.168.123.220

**rpa**: 弘玑库

**spider**：爬虫和自动化相关表。

**albb\_product\_attr**: 阿里巴巴国际站商品表

存储阿里巴巴国际站商品数据，之前为国际站自动改价使用。现在已经停用。

**clendar**: 节假日表

节假日表。

数据有ERP写入（郭晨涛）。目的为爬虫/RPA根据节假日发送一些通知消息。

Flag = 是否节假日

**CdcRecord**: 阿里巴巴自动改价指令表

阿里巴巴国际站改价表。

数据由ERP写入（施军辉），根据记录改价。

因为改价功能已经又ERP实现（唐立星）。此表停用。

**custom\_broker**: 报关行数据表

报关行报关数据集表。

可查看Spider项目CustomBroker任务。

item['pre\_no'] = '' # 预录单编号

item['customs\_no'] = '' # 海关编号

item['consigner'] = '' # 境内发货人

item['exit\_category'] = '' # 出境类别

item['exit\_date'] = '' # 出口日期

item['declaration\_date'] = '' # 申报日期

item['internet\_content\_provider'] = '' # 备案号

item['transportation\_mode'] = '' # 运输方式

item['delivery\_code'] = '' # 提运单号

item['production\_unit'] = '' # 生产销售单位

item['supervise\_mode'] = '' # 监管方式

item['exemption\_nature'] = '' # 征免性质

item['licence\_code'] = '' # 许可证号

item['agreement\_no'] = '' # 合同协议号

item['trade\_country'] = '' # 贸易国（地区）

item['arrive\_country'] = '' # 运抵国（地区）

item['destination\_port'] = '' # 指运港

item['departure\_port'] = '' # 离境口岸

item['package\_kind'] = '' # 包装种类

item['packages\_number'] = '' # 件数

item['rough\_weight'] = '' # 毛重(千克)

item['net\_weight'] = '' # 净重(千克)

item['delivery\_terms'] = '' # 成交方式

item['freight'] = '' # 运费

item['premium'] = '' # 保费

item['incidentals'] = '' # 杂费

item['documents\_no'] = '' # 随附单证及编号

item['remark'] = '' # 标记唛码及备注

item['consignee'] = [] # 境外收货人

item['transport\_flight'] = [] # 运输工具名称及航次号

**dd\_auto**: 滴滴指令表

滴滴出行人员表。为弘玑RPA任务滴滴出行提供数据。

type = 操作类型

user\_id = 员工编号

user\_name = 员工姓名

phone = 手机号

department = 部门

conditions = 策略

status = 是否提交

**dhl\_auto**: 中外运指令表

自动DHL任务提供数据。

declareType = 申报类型

ownerCode = 海关注册编号

ownerCodeScc = 统一社会信用代码

ownerName = 经营单位公司名称

tradeModeName = 监管方式名称

transModeName = 成交方式名称

**financial\_flow**（停用）：

影刀RPA金融流水保存数据记录。

**foreign\_currency**：汇率数据表

实时汇率数据表，Spider.Forengin\_currency写入。



**hongbanzhuan**：（临时）

数据采集表（朱校良）需要采集的数据。

**hr\_auto**：公积金，医社保指令表

医社保任务表，Spider.HR 和 弘玑-医社保使用。

birth\_date = 出生日期

salary = 申报基数

status = 指令状态，是否完成

adjust\_date = 调整时间，指令的执行日期

completion\_datetime = 完成时间

update\_datetime = 指令上次执行时间

bread\_reason = 停保原因

quality = 户口性质

is\_marry = 是否结婚

**hr\_auto\_user**：公积金，医社保人员表

保存公积金，医社保在保人员名单。

type = （社保，医保，公积金）

**material**: 原材料数据表

原材料数据表。Spider.Materail项目写入数据。

rlease\_date = 官网发布时间

spider\_datetime = 采集时间

**postcode**：美国邮编

采集的美国邮编数据。

**qq\_group\_user**：企点群助手指令

企点群人员增删表。

group\_name = 企点群名称

action = 动作

user\_name = 员工姓名

user\_qq\_code = 员工企点号

update\_dateime = 上次执行时间

group\_code = 企点群号

status = 是否完成

execution\_dateime = 指令计划执行时间

**sw\_auto\_weituo**：单证自动委托

单一窗口委托数据表。

code = 合同号

bgh\_code = 报关行代码

bgh\_name = 报关行名称

hs\_code = 海关Hscode

source = 源国

unit = 单位

status = 状态，是否发送委托

protocol\_code = 协议号

hw\_name = 货物名称

card\_name = 抬头编号（单一窗口）

**sw\_wt\_deadline**：报关行委托续保

委托报关行数据表，记录报关行委托时间。

company = 抬头

company\_code = 抬头对应单一窗口的编码

agent\_name = 报关行名称

agent\_ent\_dateime = 委托过去时间

**sync\_resume**：招聘简历采集

招聘网简历采集数据表。

name = 投递者名称

link = 详情地址

position = 投递岗位

deliver\_date = 投递时间

source = 来源

uid = 爬虫id

lable = 标记反馈结果

**tv\_params\_top500**：京东商品Top

京东TV商品top表。

因为是top排行，所以采集下来要标记顺序。

page\_index = 排序号

page = 页码

**tyc\_search**：智慧牙查询词

天眼查关键词表，根据关键词查询天眼查结果

**verify\_code**：手机短信验证码

手机验证码记录表。

code = 验证码

create\_dateime= 创建时间

source = 来源

content = 短信内容

**vessel\_account**：海港EDI账号

EDI网站账号表，为爬虫使用。

**vessel\_notice**：海港EDI数据

EDL采集数据表。

box\_datetime = 开截时间

**zhihuiya\_data**：智慧牙

智慧牙数据表。

pn = 专利号

pdb时间

apd时间

status = 状态

patent\_type = 专利类型

fmr = 发明人

**zhihuiya\_search**：智慧牙关键词

爬虫采集的智慧牙关键词。

**zscq**：（停用）知识产权

知识产权数据表

**zscq\_brand**：（停用）知识产权关键词

爬虫采集知识产权的品牌。

**Invoice**

DeclarationForms

id = Column(Integer, primary\_key=True, comment='主键')

CusCiqNo = Column(String(128), comment='统一编号')

EntryId = Column(String(128), comment='海关编号')

UpdateTime = Column(DateTime, default=datetime.now(), comment='更新时间')

ConTrNo = Column(String(128), comment='合同编号')

CusDecStatusName = Column(String(128), comment='报关状态')

ReleasingNo = Column(String(128), comment='放行通知书编号')

InspectionNo = Column(String(128), comment='查验通知书编号')

HasUploaded = Column(Boolean, default=False, comment='文件上传标志')

CkBgdUploaded = Column(Boolean, default=False, comment='出口货物报关单附件')

ExtensionData = Column(Text, comment='详情数据')

ExtensionDataPDF = Column(Text, comment='详情数据')

Submitter = Column(String(20), comment='单证员，设置为None')

Archived = Column(Boolean, default=False, comment='归档标记，设置为False')

DeclDate = Column(DateTime, comment='报关时间')

SpiderIeDate = Column(String(50), comment='进/出口时间')

MongoDB

mongodb://192.168.125.115:27017/spider

这个库，爬虫管理后台的数据库。

mongodb://spider:bestadmin@mongo.mongo.svc.cluster.local:27017/spider

这个库，是存放一些爬虫数据。

# ServerHost（服务器）

192.168.125.115

Ssh 用户名/密码：root/bestadmin

爬虫服务器。部署爬虫服务，部署爬虫管理后台，接收短信验证码服务。

部署方式：supervisor

可用通过命令查看部署的服务

spervisorctl status

数据库：mongo, redis；已配置开机启动启动。

部署方式：docker

容器：redis, mongodb

使用命令查看容器详情及数据挂载路径：

doker inspect redis

doker inspect mongodb

192.168.125.116

Ssh 用户名/密码：root/bestadmin

192.168.125.30

Ssh 用户名/密码：root/nbqc@123

192.168.125.31

Ssh 用户名/密码：root/ nbqc@123

192.168.125.32

Ssh 用户名/密码：root/ nbqc@123

这三台搭建了集群K8s，部署了Selenium Grid服务，爬虫调用chrome浏览器。

部署过scrapyd的Pod。后续没用过。

kubesphere：192.168.125.30:30880

账号/密码：admin/Nbqc@123

192.168.125.33（停用）

发票认证RPA服务器（邢婉秋）。

网页大改版吧不能用了，邢婉秋说：先留着吧也不占地方。

192.168.125.131

简历采集服务器。

部署了是3个采集工具。前程无忧，猎聘，实习僧。

由windows计划任务配置定时。

（每天操作人：施宇超）

192.168.125.138

智联简历采集服务器。

部署Powerautomate工具。

账号：[zhaohui1995910@163.com](mailto:zhaohui1995910@163.com)

密码：1995910zhaohui

是我个人版账号。

每天定时任务由影刀配置定时任务启动 PowerAutoMate 流程。

影刀账号（施宇超）

影刀定时启动注意：服务器不能锁屏，所以链接方式用VNC。

192.168.125.213

老金融流水影刀工具，将流水导入ERP。（操作人：储红媛）

# PhoneSmsCode（手机短信验证码）

部署：192.168.125.115

启动工具：supervisor

代码路径：/root/spider-web/receive\_phone\_code.py

Python环境：conda activate spider-admin # 启动前在115服务器执行

@app.route('/receive/sms/code', methods=['post'])

def post():

content = str(request.form.to\_dict().get('content', ''))

只有这一个接口，接收手机短信验证码保存到数据库 verify\_code 表。

手机要求安装客户端“短信转发器”https://github.com/pppscn/SmsForwarder

并配置转发规则，可以对照老手机的配置方式。

# AccoutPasswor（账号密码）

SqlServer 192.168.123.220

账号/密码: spider/spider2020

Ip代理平台：

账号/密码：13454765982/ qucheng@2020

登录地址： <http://www.taiyanghttp.com/login/>

本人电脑：

10867/zhaohui@1995910

云码（验证码识别）

账号/密码：13454765982/ qucheng@2022

PowerAutoMeta:

招聘

hrzp@lumi.cn

Huc502352

单证

dzcy@lumi.cn

Dzam1354441

公积金

sbgjj@lumi.cn

CLuw5135212

微软PowerAutomate即将过期，这次企业版真的需要续费了。2023-6-2

Gitlab：

地址：https://gitlab.powertradepro.com/users/sign\_in

账号：3002943629@qq.com

密码：zhaohui@1995910

RPA企点助手QQ号：

账号/密码：3003855079 / Lumi@2022

# PythonChatGpt（当曲chatgpt对接）

技术介绍

框架Flask

仓库地址：<https://gitlab.powertradepro.com/apaas_group/dangqu.pt.ai.git>

# PowerAutomate（微软RPA）

公积金

sbgjj@lumi.cn

CLuw5135212

读取SqlServer spider hr\_auto表，type类型为“公积金基数”的指令，登录公积金网站完成申报，封存操作。

舱单

dzcy@lumi.cn

Dzam1354441

智联招聘

hrzp@lumi.cn

Huc502352