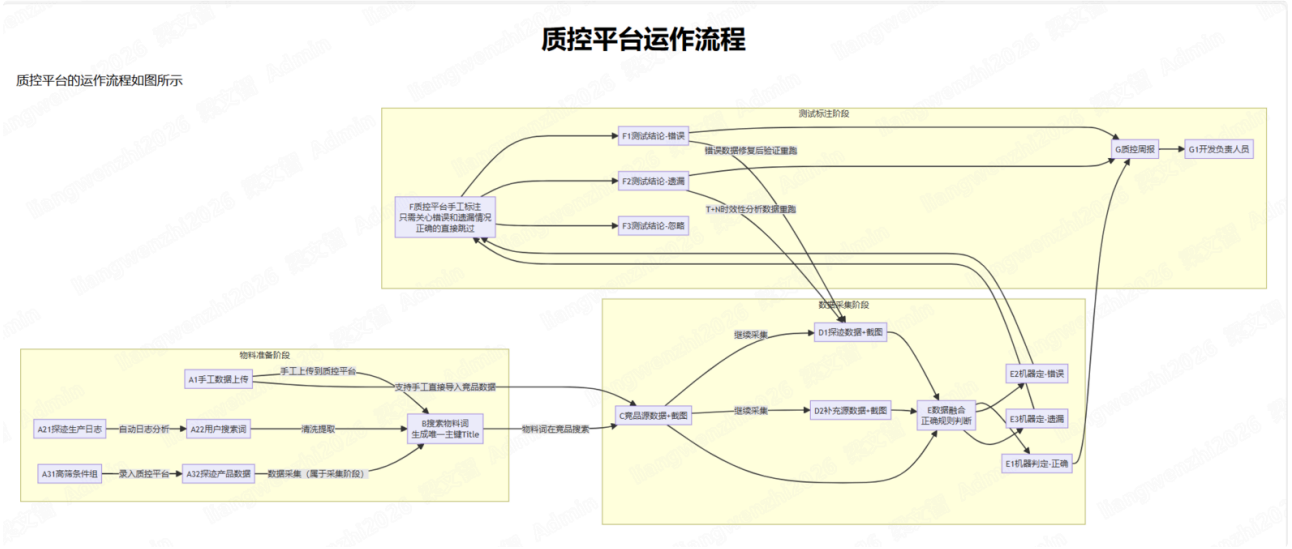
AISpider 质控平台

点此进入质控平台,外网不可访问

质控平台原理和使用说明

https://alidocs.dingtalk.com/i/nodes/m9bN7RYPWdlyezMbFD1QPdGoWZd1wyK0?doc_type=wiki_doc&utm_medium=main_vertical&utm_scene=team_space&utm_source=search

质控平台运作流程



一 物料准备阶段 （用于生成 spider 库中的数据的种子）

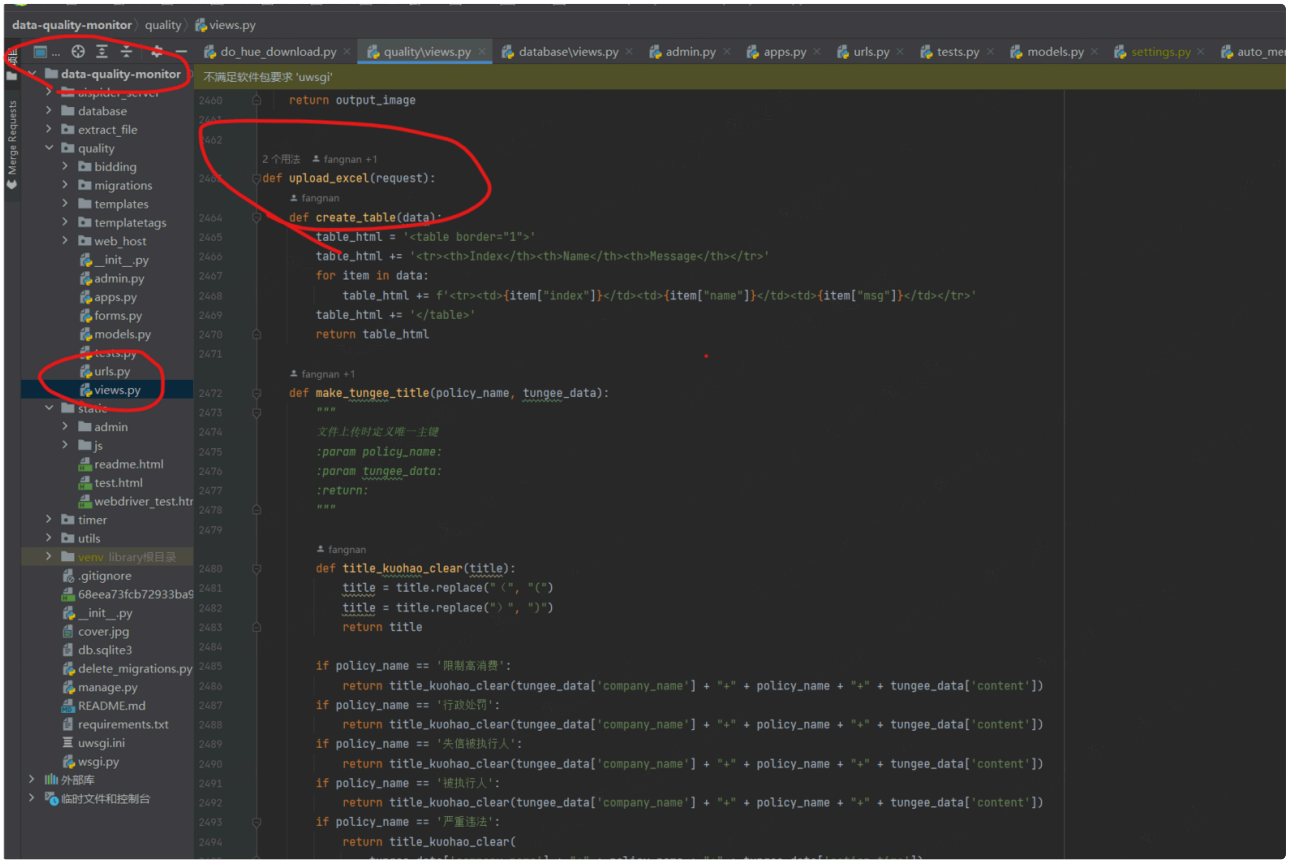
1 A1 手工数据上传-->竞品源数据 + 截图

应用场景：不需要去竞品网站爬取数据，竞品数据已经保存在文件中

具体如下：

1 如果搜索词不是简单类型的 如 excel 表格或者其他，和传入的这些数据不用去网站采集数据而是所有数据都在文件中则选择物料文件上传和修改 data-quality-monitor 文件夹里面的 quality 里面的 view.py 文件中的 def upload_excel 函数 在里面根据传入的文件构造 spider 库中的数据，状态码设置为 0.

- selenium爬虫资源集群
- [用户数据]LQJ的数据图
- [用户数据]LQJ的数据二图
- [物料]物料文件上传
- [物料]名单配置上传
- [工具]OSS文件上传
- [统计]运行情况和统计数据
- [标注]工商数据
- [标注]工商企查查数据
- [标注]行政处罚数据

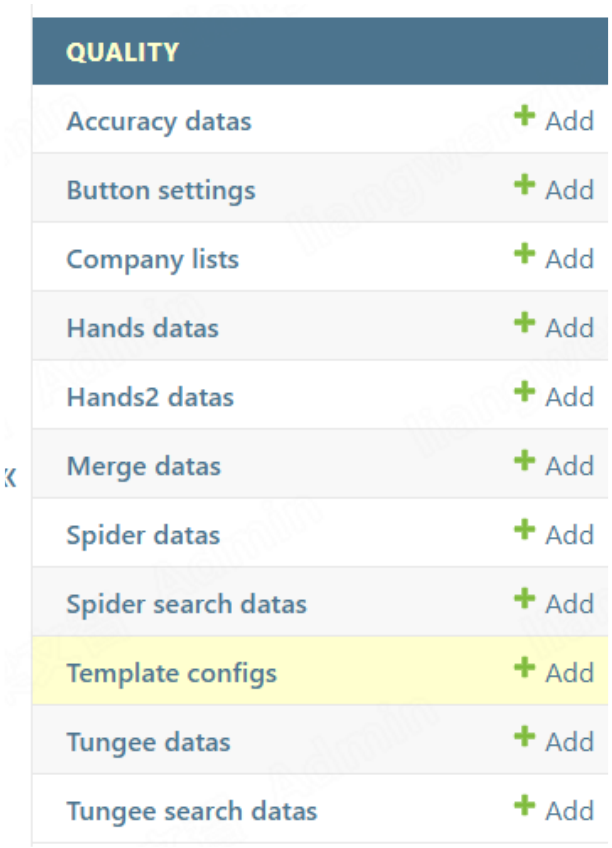


2 A21 探迹生产日志-->竞品源数据 + 截图

应用场景：需要去网站爬取数据，只有公司名等不知道详细数据

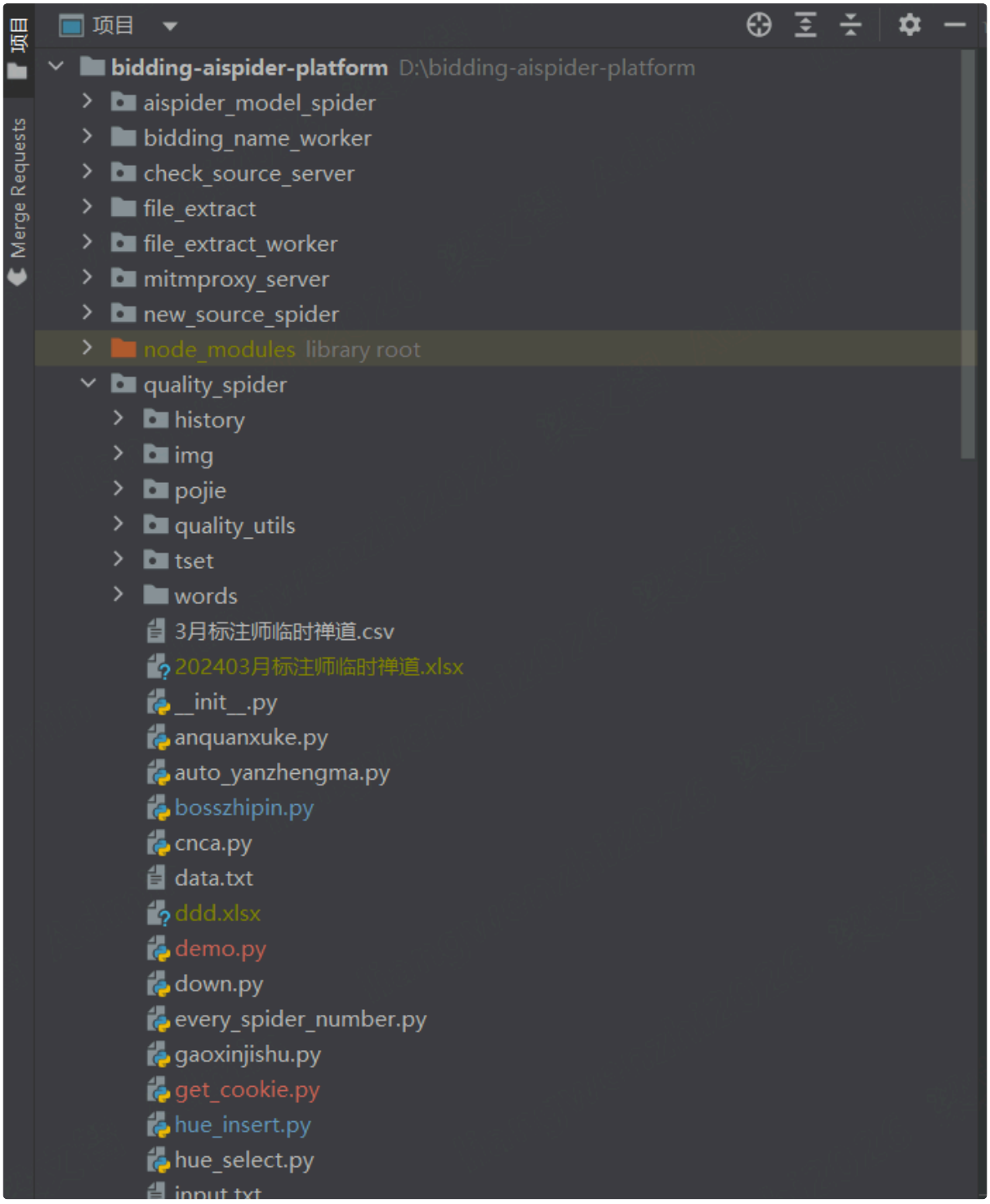
具体：

1：如果搜索词比较简便如（公司名 +XXX+XXX）和需要该搜索词去某某网站上去搜索，可以往 Template configs 添加新的种子源 再到质控平台欢迎页在名单配置上传就可以上传搜索词了



2 编写 spider 爬虫脚本

1 在 bidding-aispider-platform 中的 quality_spider 中添加新的.py 爬虫脚本，根据搜索词去网站搜索数据，构建数据，然后存储成下面 data 的这种数据结构 title 为匹配探迹数据的主键（重要） source 为网站的名称 type 为该爬虫爬取数据的类型 data 为爬取的数据 url 为爬取网站的 url screenshot_url 为截图的 url week_id 为数据存储时的时间 存储到 spider 库中，状态码更新为 1.


项目
项目
Merge Requests
bidding-aispider-platform D:\bidding-aispider-platform
aispider_model_spider
bidding_name_worker
check_source_server
file_extract
file_extract_worker
mitmproxy_server
new_source_spider
node_modules library root
quality_spider
history
img
pojie
quality_utils
tset
words
3月标注师临时禅道.csv
202403月标注师临时禅道.xlsx
__init__.py
anquanxuke.py
auto_yanzhengma.py
bosszhipin.py
cnca.py
data.txt
ddd.xlsx
demo.py
down.py
every_spider_number.py
gaoxinjishu.py
get_cookie.py
hue_insert.py
hue_select.py
input.txt

3 A31 高筛条件组（咨询方楠）

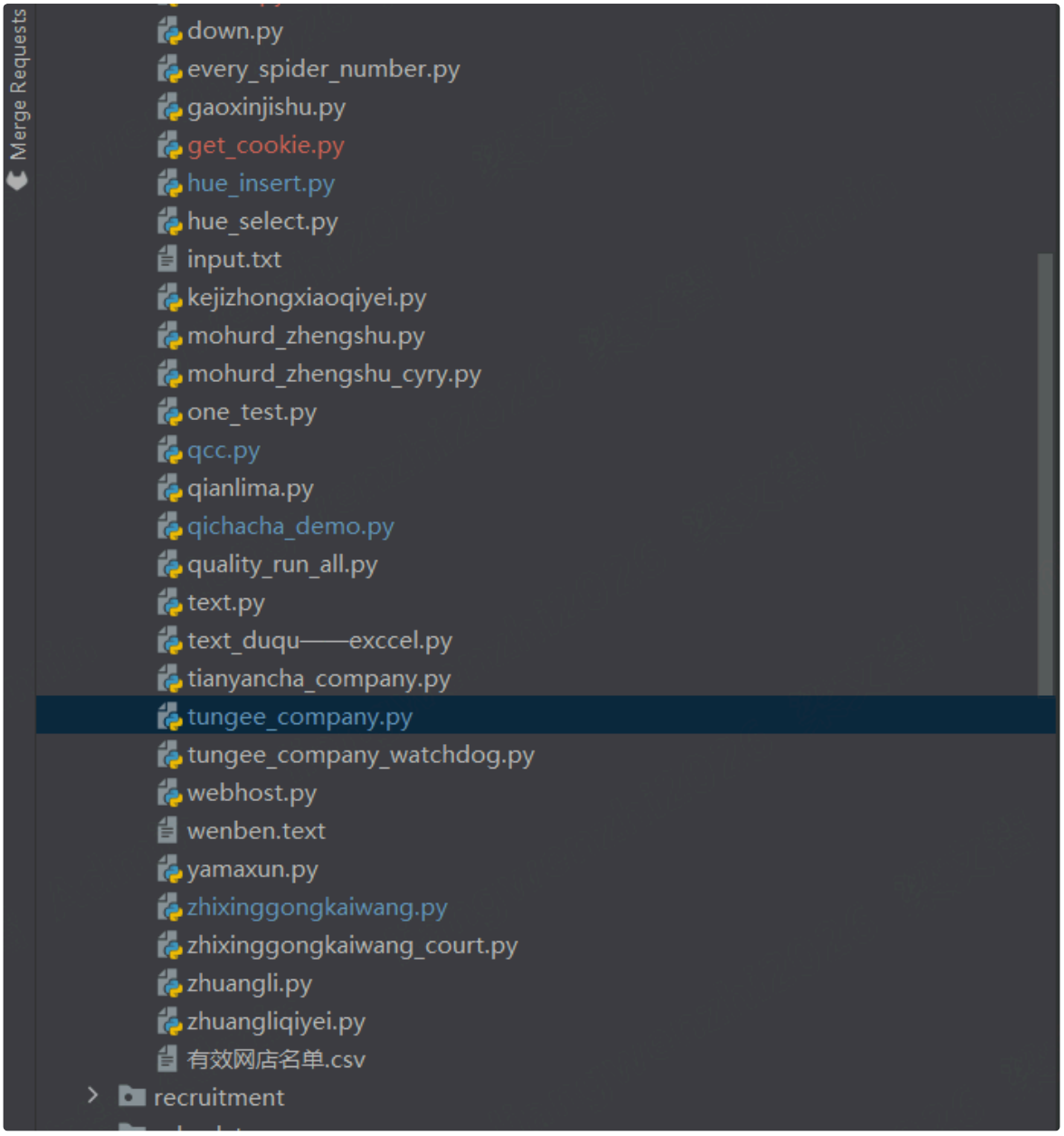
二 数据采集阶段

1 D1 探迹数据 + 截图

应用场景：只有探迹和竞品数据比对

具体：

1 编写 tungee_company.py 数据采集脚本，在已有的框架下编写脚本 ， 首先从 spider 库中取出数据并根据数据的公司名或者其他在探迹对应的版本下去搜索获取想要的数据 data_msg title 为匹配 spider 库的主键 spider_id 为从 spider 库中数据的 id source 为探迹 type 为数据的类型（如 限制高消费）data 为在探迹的数据 ， 状态码更新为 2.



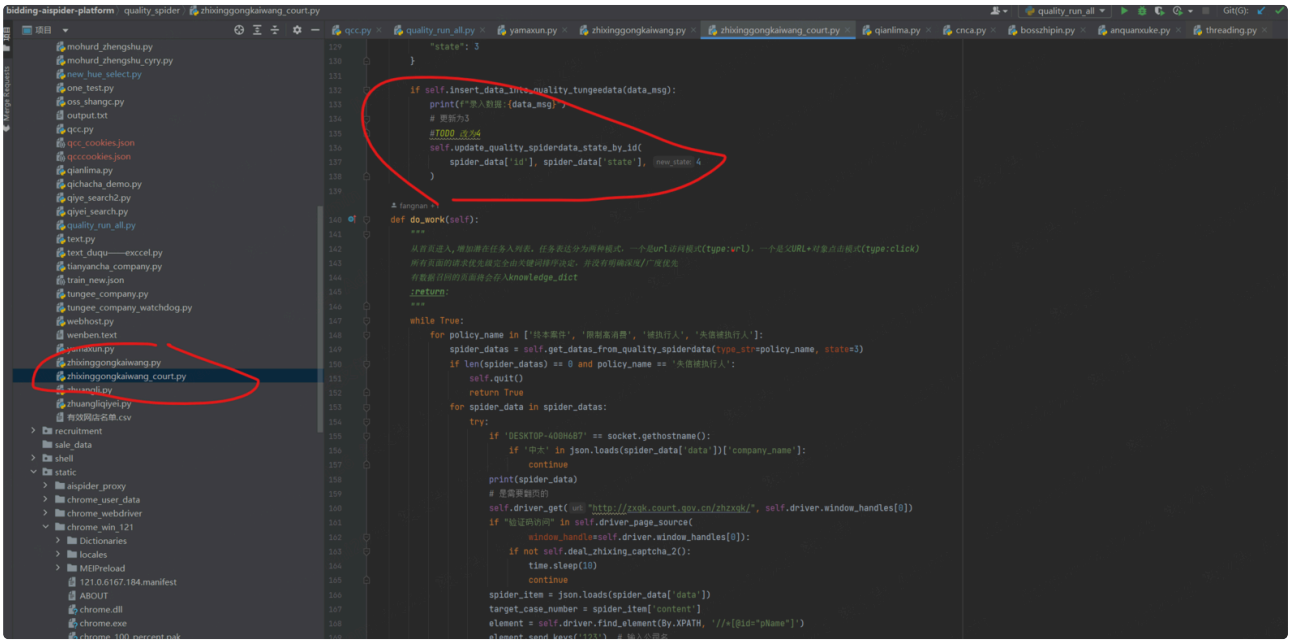

```
        datas[key[0]] = value[0]
    data_msg = {
        "title": self.make_tungee_title(policy_name, datas),
        "spider_id": word['id'],
        "source": "探迹",
        "type": policy_name,
        "data": json.dumps(datas, separators=(',', ':'),
                           ensure_ascii=False),
        "url": self.driver_current_url(window_handle=window_handle),
        "screenshot_url": self.take_screenshot_and_upload_to_oss(
            window_handle=window_handle),
        "week_id": word['week_id'],
    }
```

2 D2 补充源数据 + 截图

应用场景：存在除了探迹和竞品的第三方网站（限制高消费）

具体：

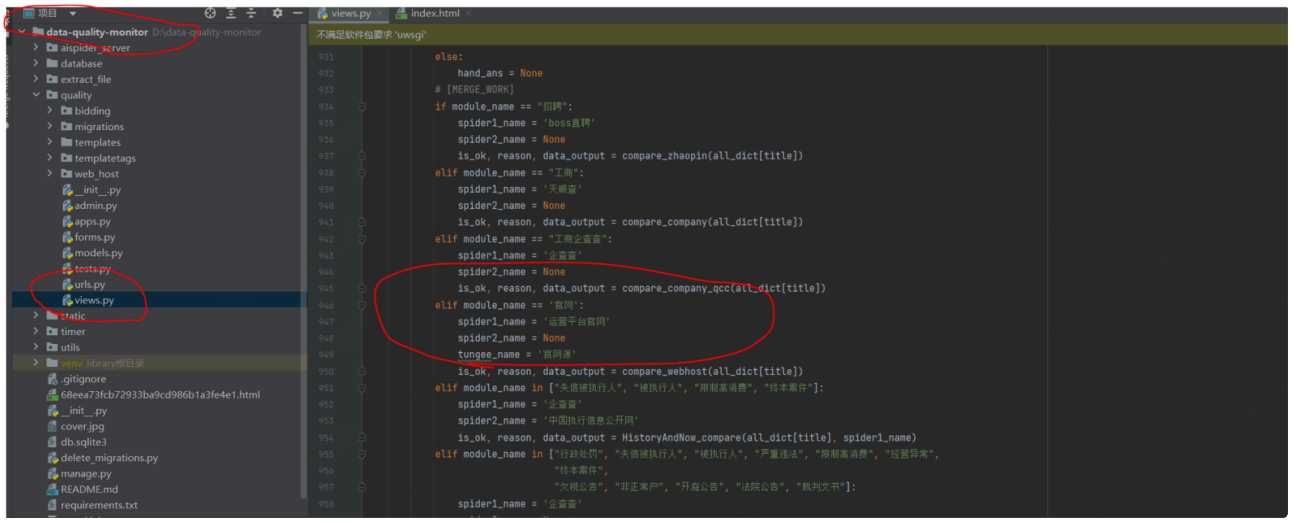
如需要到第三方平台进行比较的数据，如（限制高消费）状态码设置为 4



3 E 数据融合正确规则判断

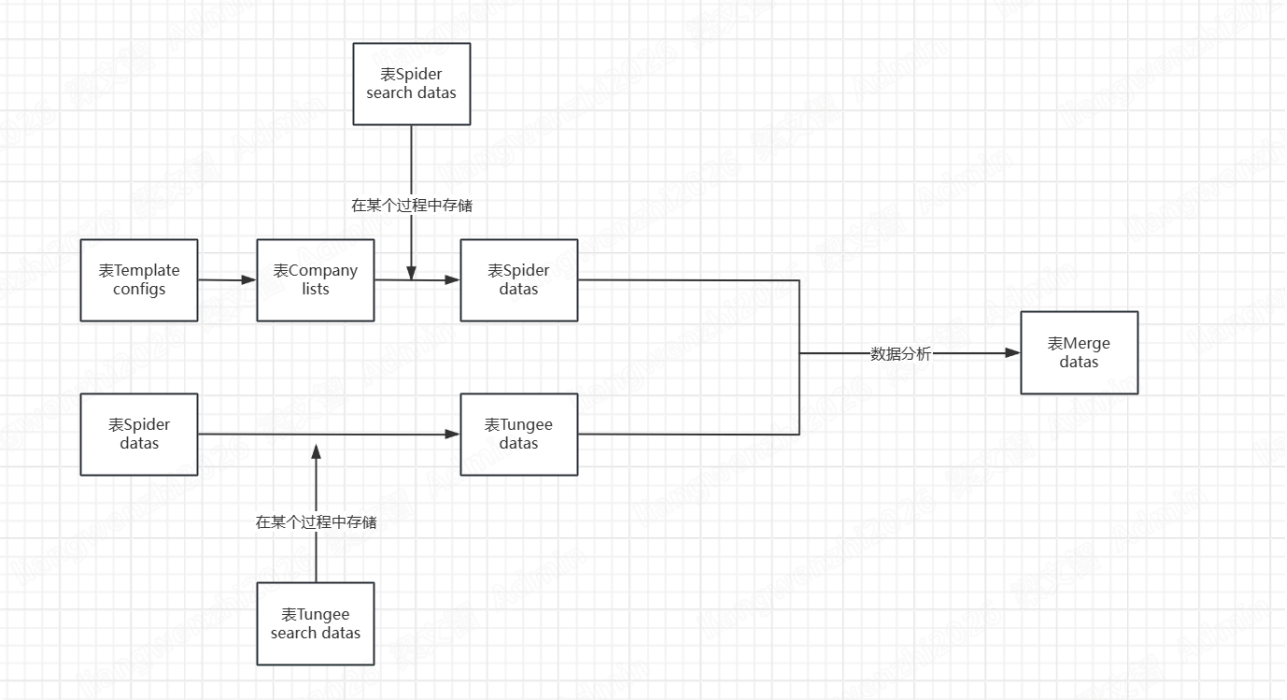
具体：

最后，在 data-quality-monitor 这个文件的 views 然后在标红的代码块附近添加新源的新函数，构建比对规则。



三 数据库表功能解释：

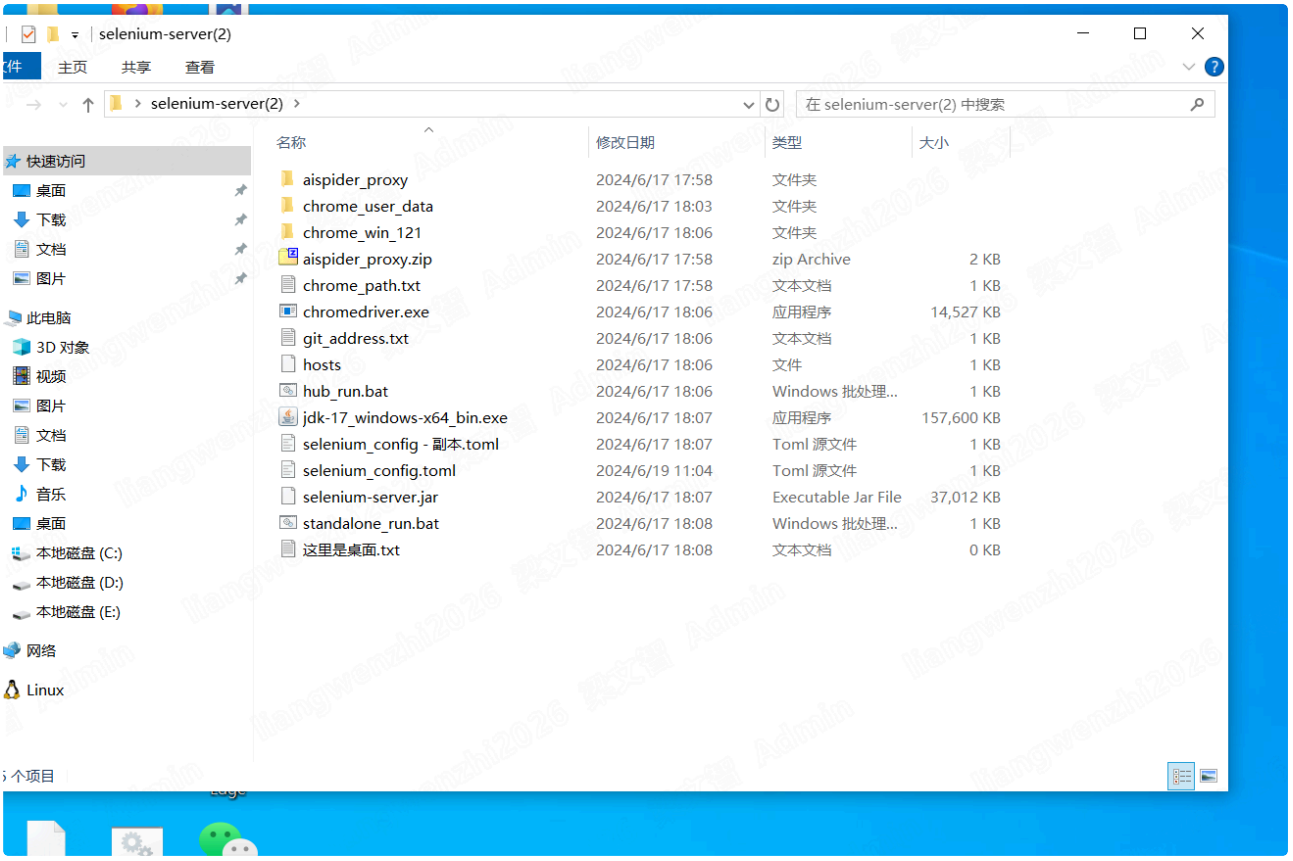
前面是从业务角度进行梳理的，但往往直接操作表效率更高



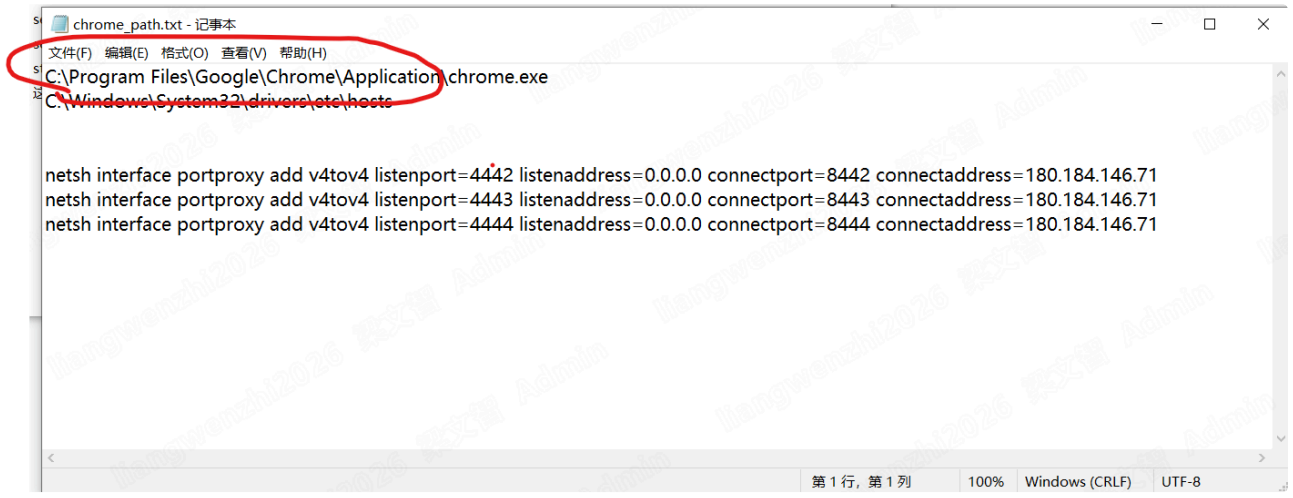
质控平台的大体工作流程是根据种子爬取数据存储到 spiderdatas 状态码设置为 0，在采集过程中把每一个数据采集的最后一步的截图存储到 spider search datas,然后从 spiderdatas 表里面取出数据在探迹产品端进行数据采集，将采集的数据存储到 Tungeedatas 里面，状态码更新为 2，在采集过程中把每一个数据采集的最后一步的截图存储到 Tungee search datas 里面，然后根据一定的比对方式，把结果存储到 Merge datas 里面

四 selenium 集群启动流程

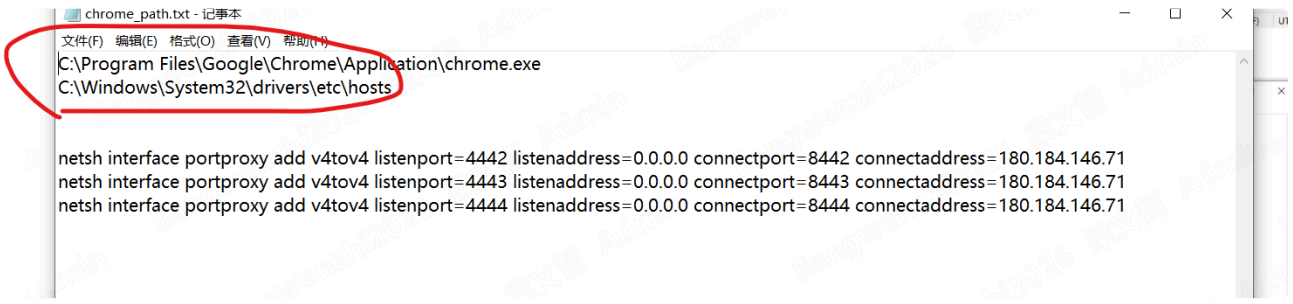
1 首先，需要检查每一个机器是否是 121 版本，和是否下载了 jdk_17-windows-x64_bin.exe 和是否把 selenium-server 这个文件夹放在桌面上。(只限在同一个私域里面运行，如探迹公司的内部网络)



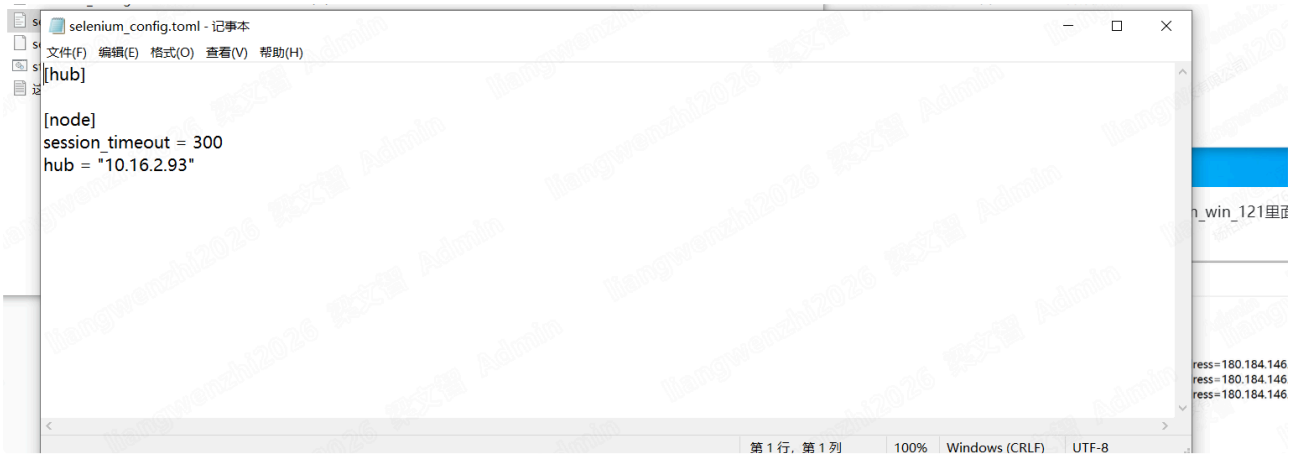
怎么・检查是否是 121 版本呢，打开以下这个路径把里面所有文件都删除掉，把上面 chorm_win_121 里面的所有文件复制到删除的文件夹里面



防止浏览器自动更新 找到以下路径把 hosts 文件替换成文件夹下的 hosts 文件



2 配置 selenium_config.toml 文件



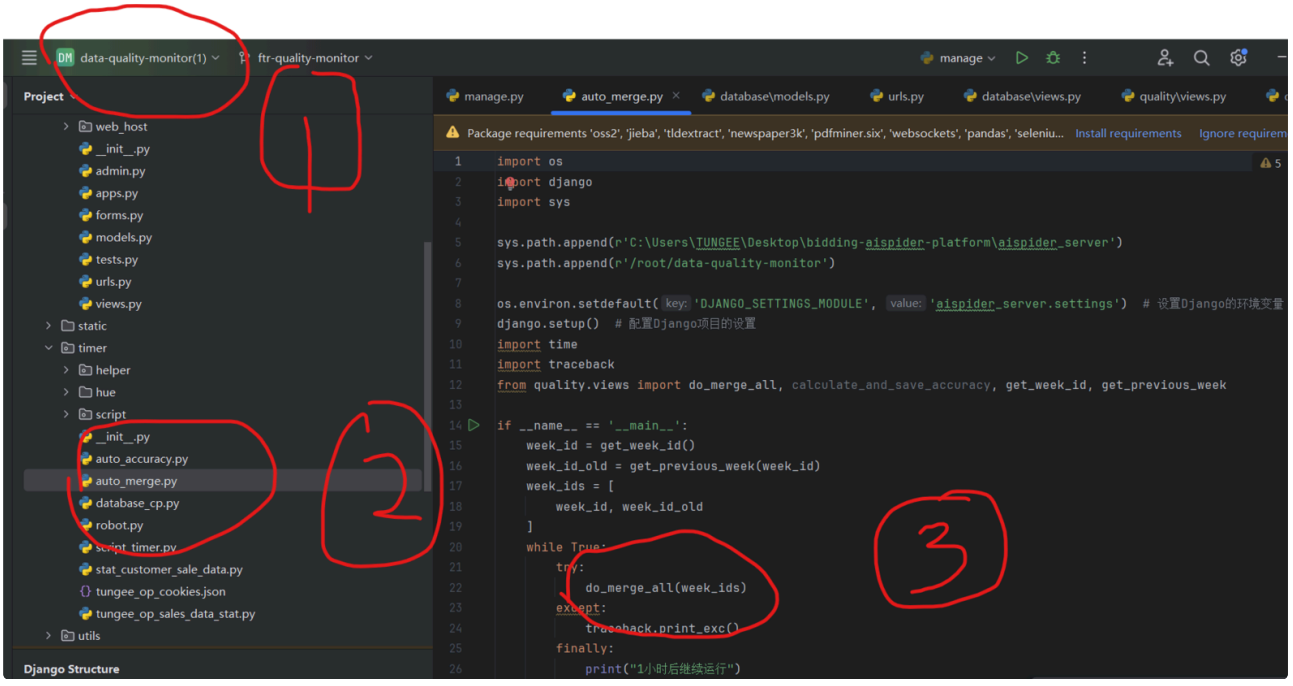
把所有主机的这个文件的 hub 改为运行程序的主机的 ip (运行程序的主机改为 http://+ 主机 ip+:4444)

3 最后，启动子节点运行 standalone_run.bat 主节点运行 hub_run.bat

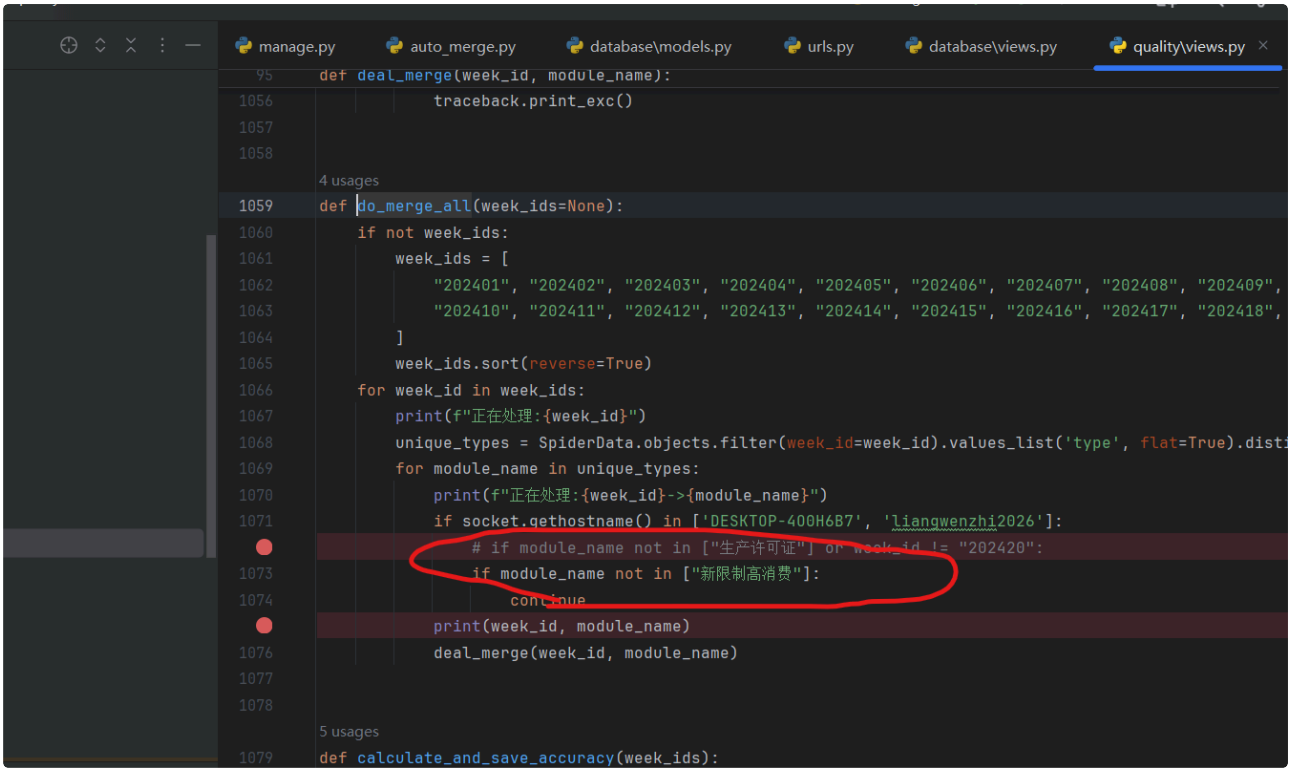
五：某个维度的调试方法：

如图所示：

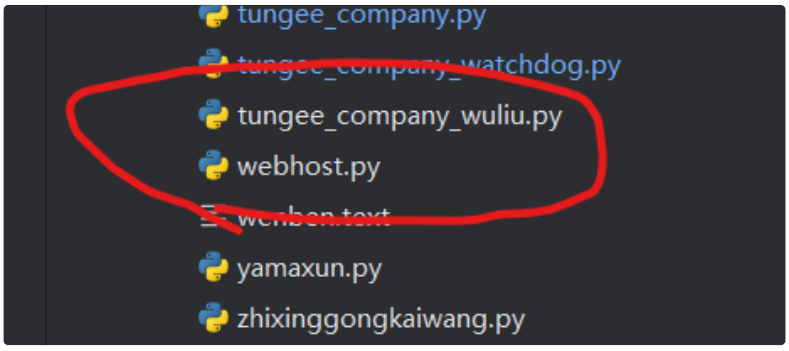
找到 data-quality-monitor 文件下的 auto_merge.py 里面的 d_merge_all()的函数：导航到这个函数里面



如果你想调式某个维度，可以在下图里面进行修改，然后就可以调试了



官网的启动在 在 bidding-aispider-platform 中的 quality_spider 中



admin 后台看任务状态可以在 180.184.146.71:8000/admin 账号密码为 liqiuju666 147258369Asd