2025-07-04周报

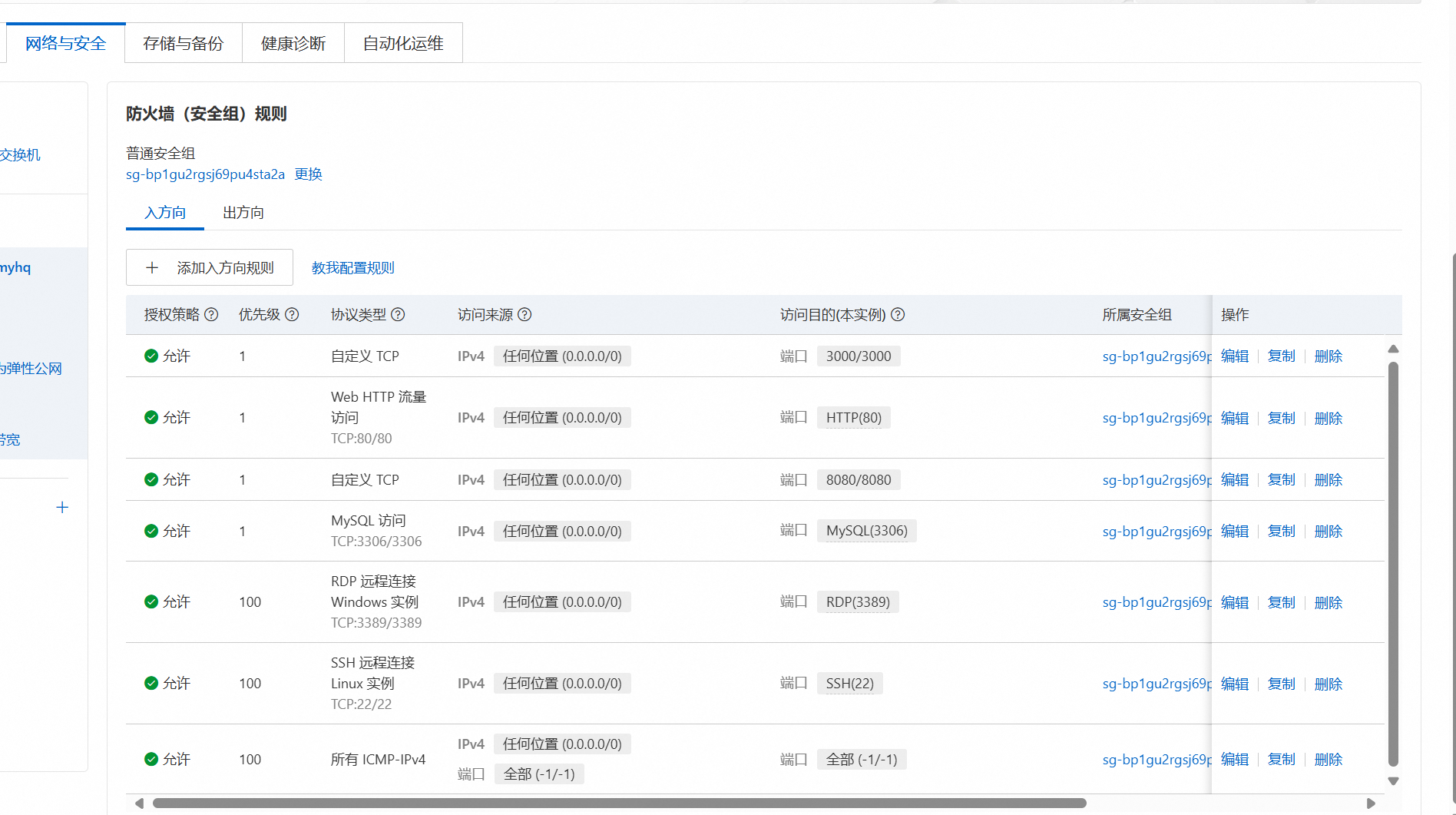
马嘉润

1. 爬虫脚本的自动化和前后端的部署

本周主要工作是将服务器上爬虫脚本进行优化，添加了原始数据的自动化入库的操作，目前在本地对整个爬虫系统进行测试，看能否完成每日更新，数据入库和上传obs的整体流程。

结构化数据爬取时（github），会在本地cache文件夹中存入github的信息，这是结构化数据更新的一个依据，cache文件夹很大，目前服务器存储空间不够，我们在看可以怎么优化存储空间。

服务器的防火墙需要开放8080，3000，3306这几个端口。



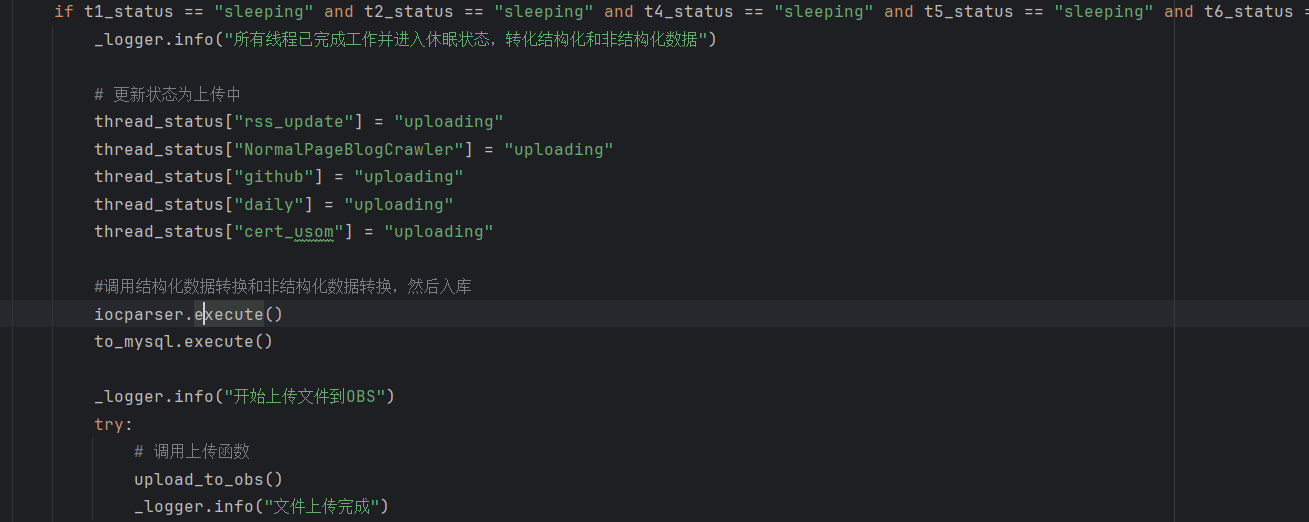
曾家豪-本周工作

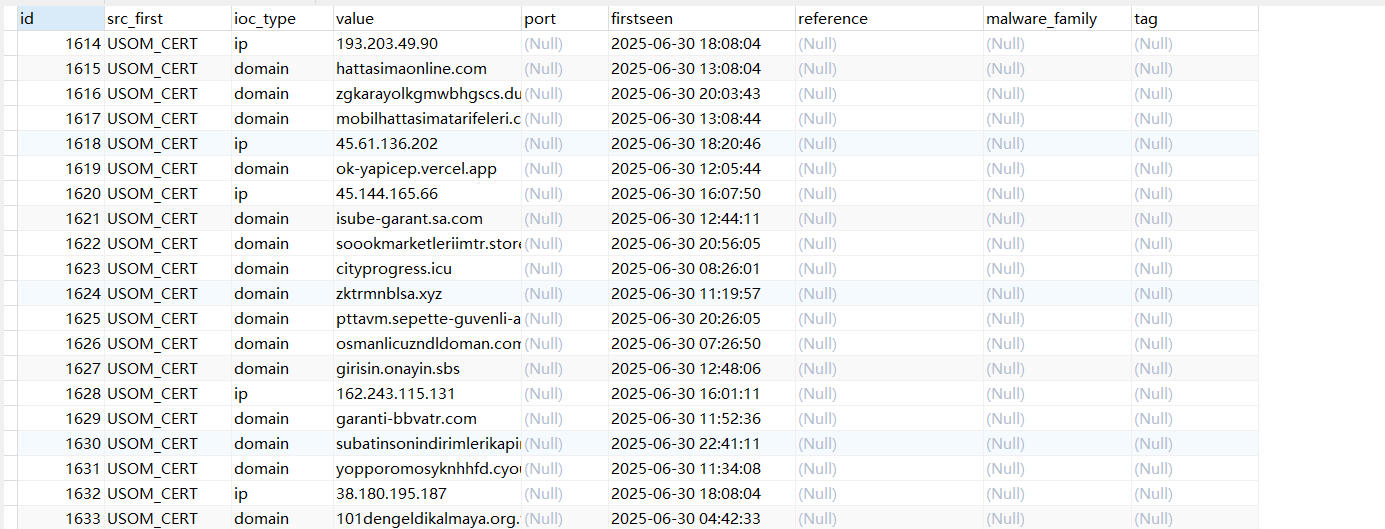
1. 负责结构化数据转化后自动化入库的工作

具体的逻辑是，在爬虫线程全部休眠后，处理结构化数据，将结构化数据按对应的字段，统一转化为csv文件，然后自动化存储入库，然后将转化后的csv文件删除，最后调用obs上传原始文件。

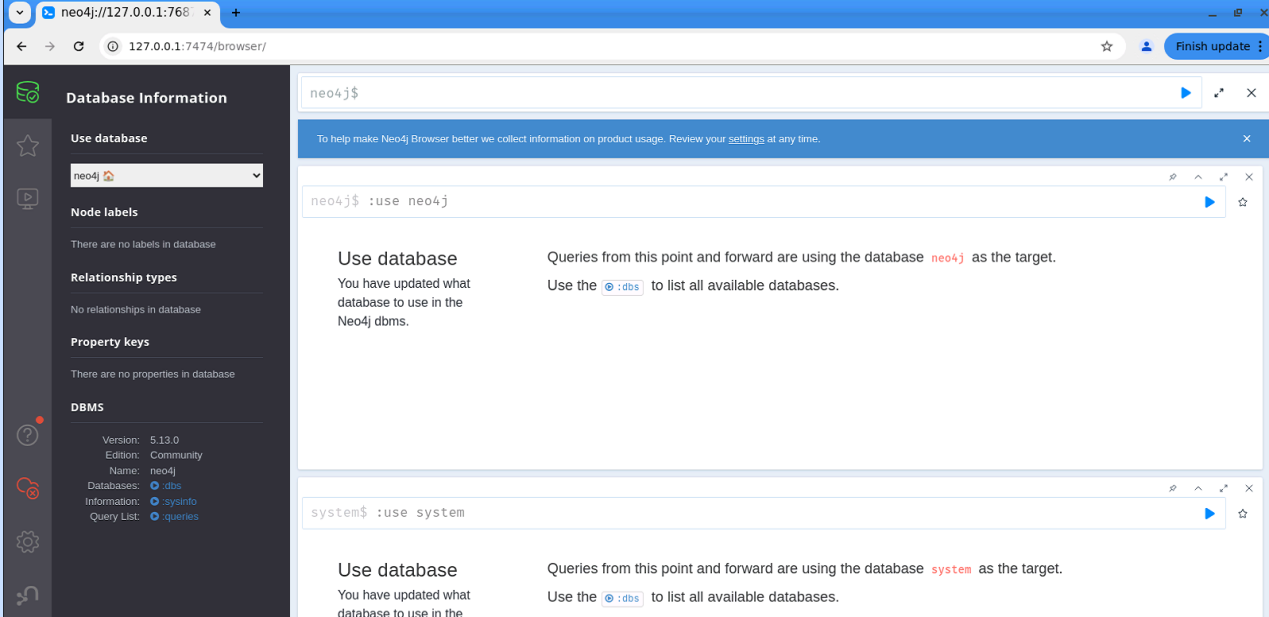
Iocparser.execute执行结构化数据的转换工作

to\_mysql.execute执行数据库自动化入库的工作





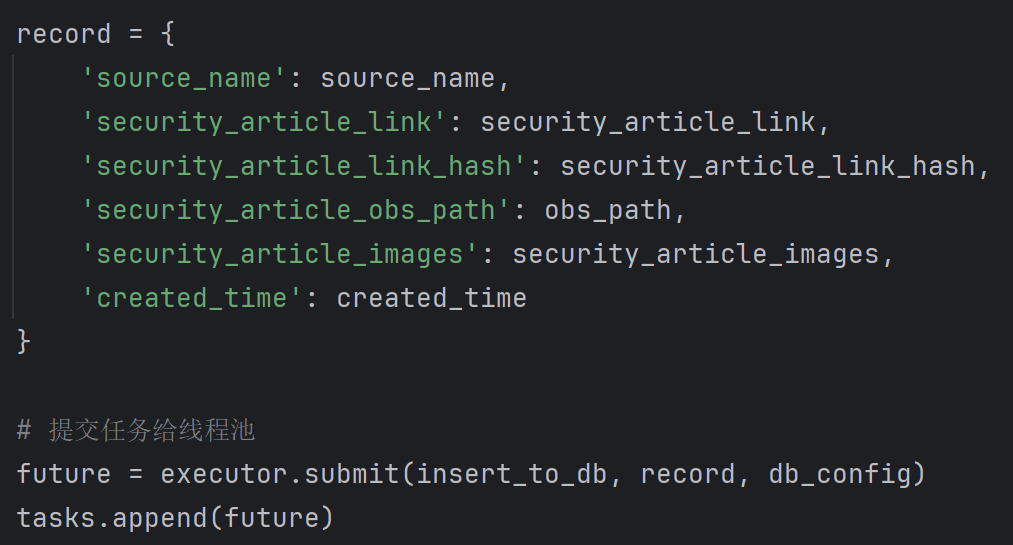
1. 完成了neo4j图形数据库的安装



冯世玲-本周工作

负责非结构化数据转化后自动化入库的工作

遍历以日期为单位划分的origin\_link路径，逐层读取每个来源（source）的哈希命名文件，提取其中的安全文章链接信息，并根据路径规则查找对应的图片资源文件名列表；随后结合文章哈希值生成对应的OBS路径信息，统一封装为数据库插入记录。该过程通过process\_data()主函数完成任务调度，并使用线程池（ThreadPoolExecutor）并发调用insert\_to\_db()将记录写入MySQL数据库。



为保证时间信息的准确性，修改了原始时间字段的处理逻辑：在遍历到每个日期文件夹时，通过读取当前系统时间生成完整的“年月日 + 时分秒”格式时间戳（created\_time），再将其转为datetime对象并存入数据库。

