2025-03-28周报

陈宇航、杨启程、赵映雅 —— 漏洞页面前端编写和前后端交互

1. 整体框架

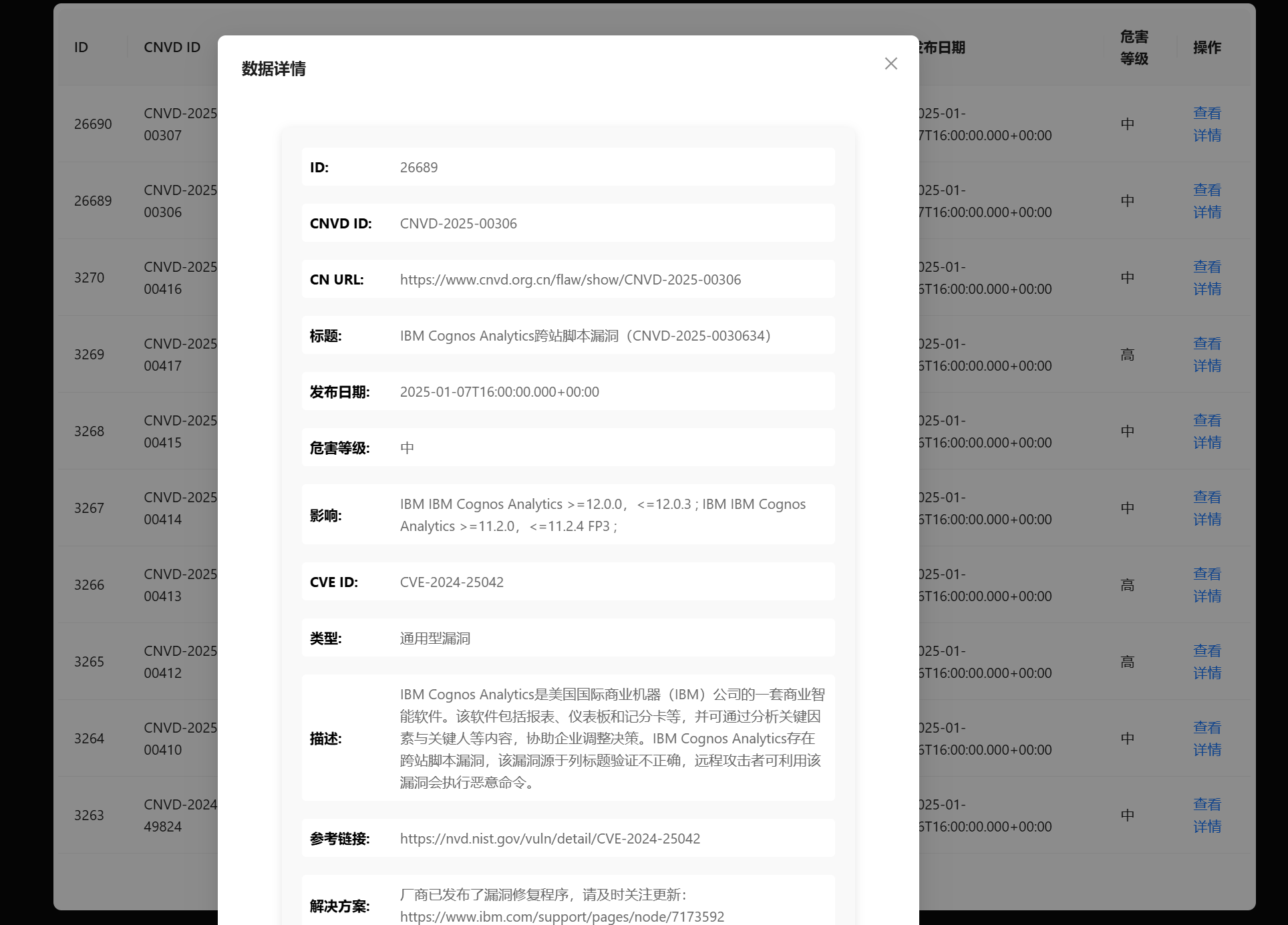
* 最新漏洞展示（概况+详情框）
* 查询漏洞（关键字：漏洞名、编号、日期）
* 漏洞标签查询及展示

2、页面展示

* LatestData展示最新1000条漏洞:



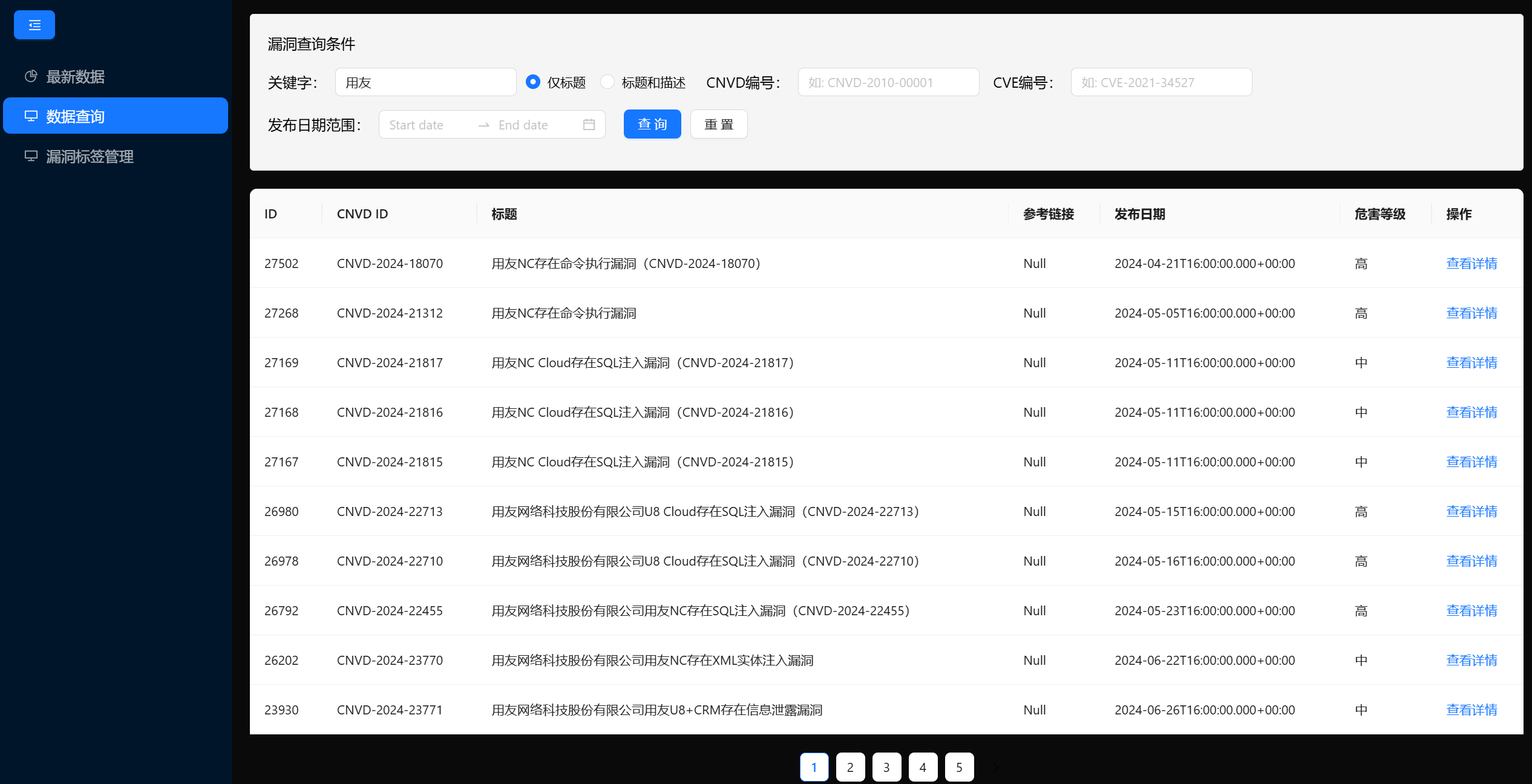
点击“查看详情”：



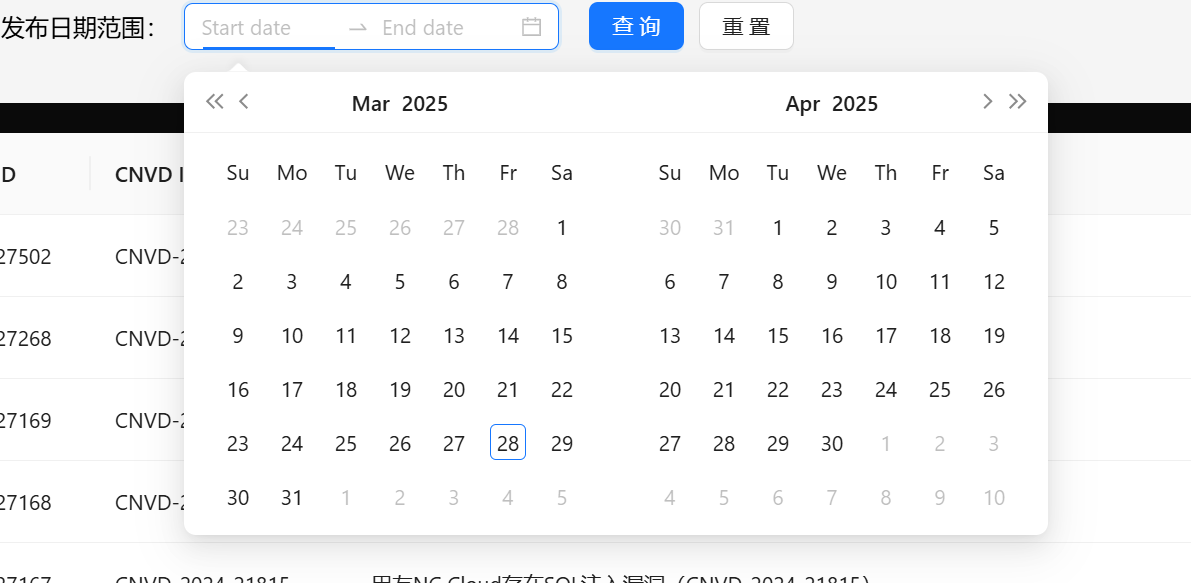
* QueryData:

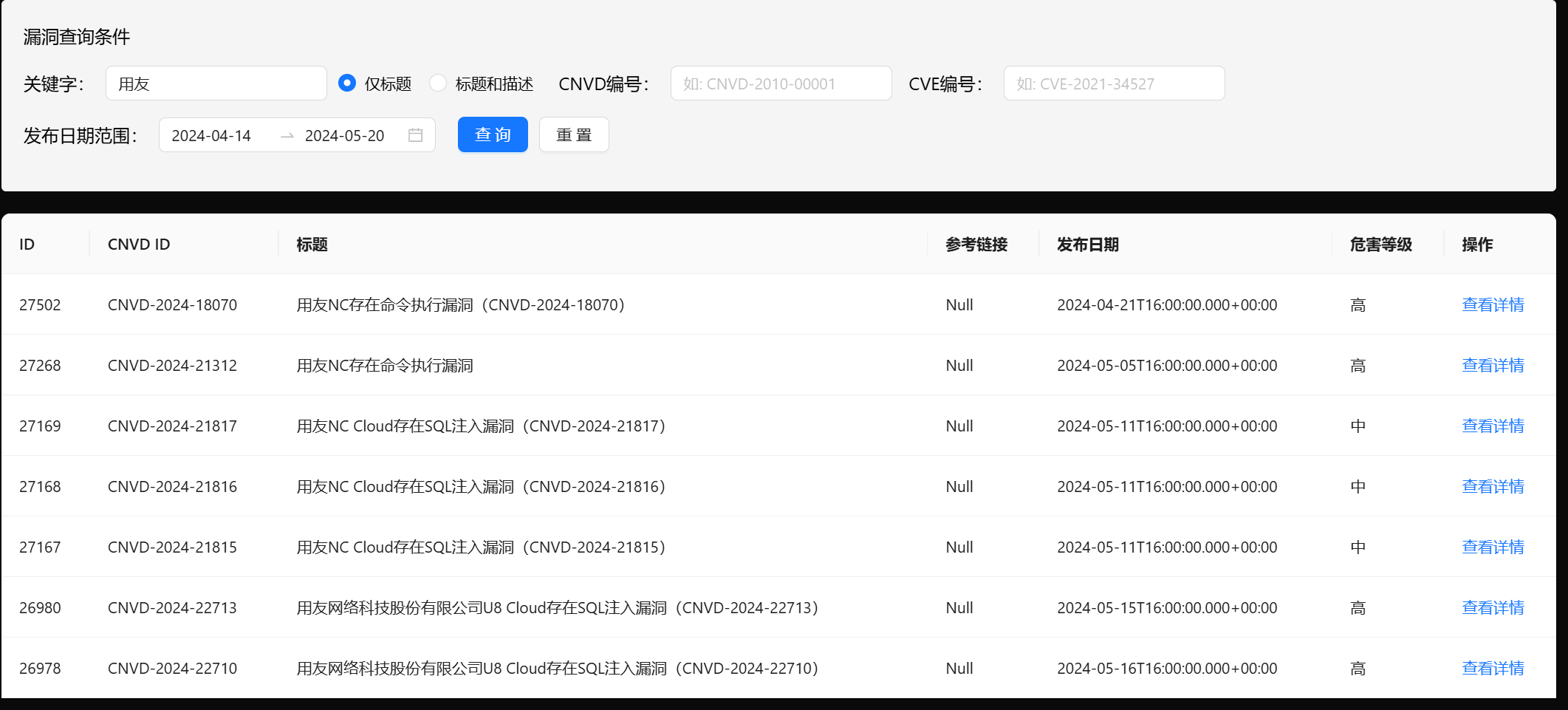


输入关键字“用友”：

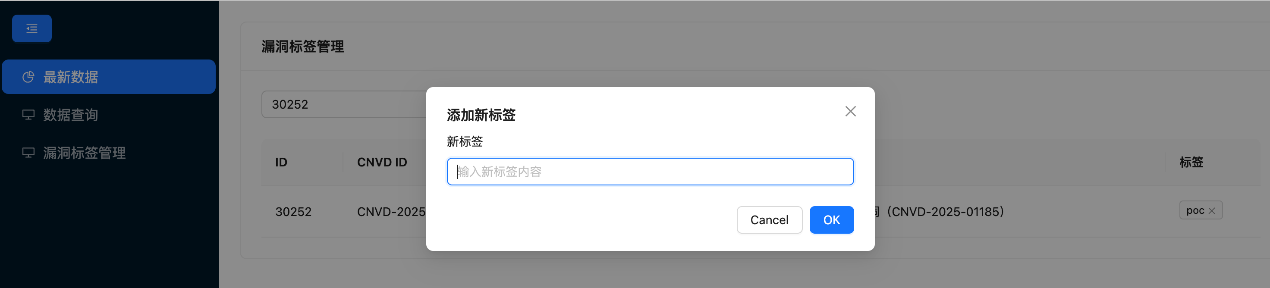


选择日期：2024-04-14-2024-05-20





* 漏洞标签



曾家豪、马嘉润、冯世玲 —— 采集系统测试和bug解决

1. 整体框架

* 采集系统的全量数据爬取和bug
* 微信公众号环境的配置和obs桶边更新边上传
* Ocr的测试

1. 详细内容
   1. 采集系统已爬取的全量数据

目前非结构化数据包含1996篇博客

结构化数据包含土耳其cert包含39w条数据，其它的包含12条数据源的信息，因为有些数据是将所有数据放在一个文件里的，这样的数据文件大约有400个。

遇到的问题：linux系统中webdriver加载页面好像总是会超时，然后抛出异常，导致一些需要模拟点击的数据源无法获取

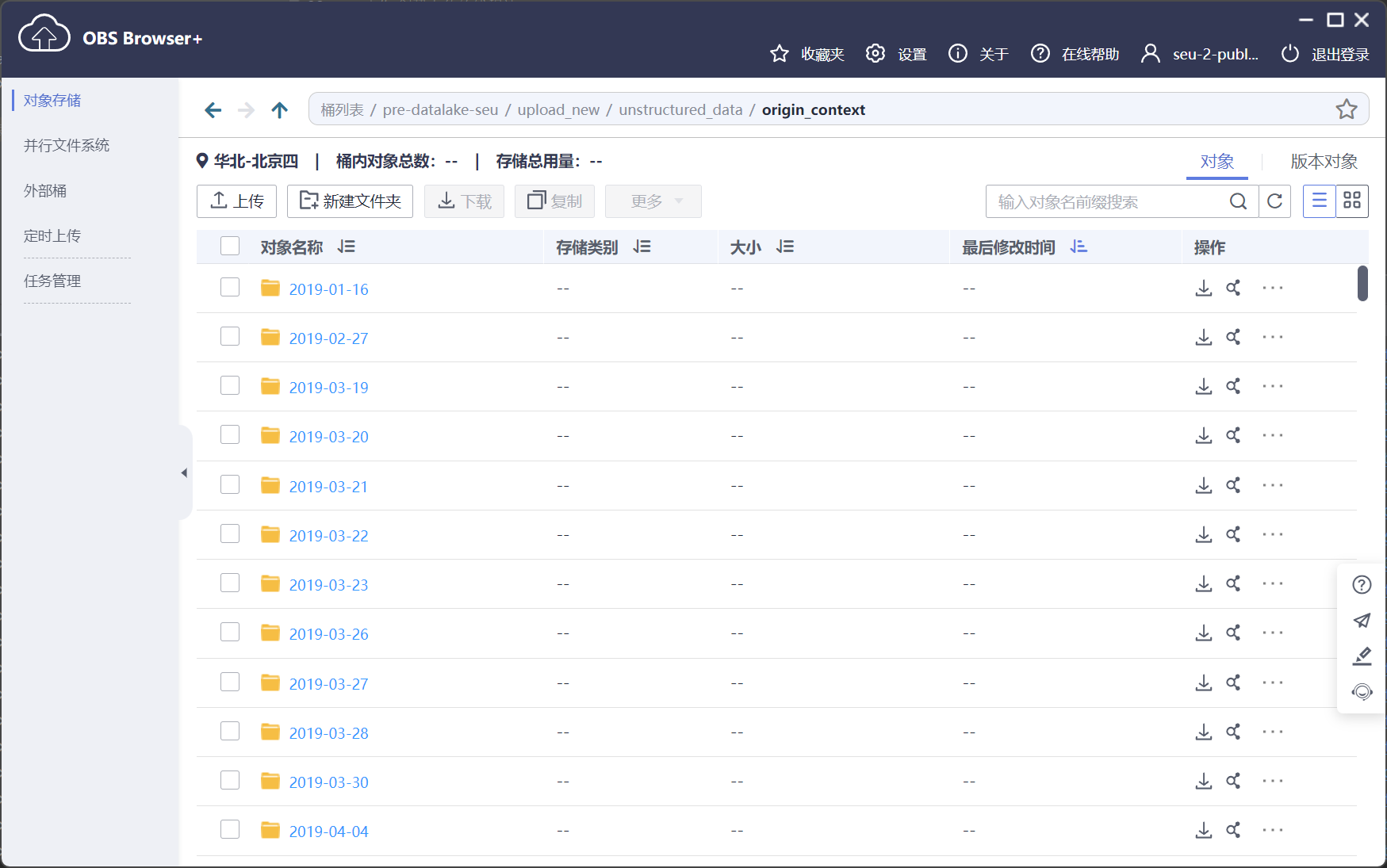
报错：页面由 JavaScript 渲染，使用模拟浏览器进行渲染

JS 页面加载异常：Message: timeout: Timed out receiving message from renderer: -0.002

(Session info: chrome=134.0.6998.88)

* 1. 微信公众号环境的配置

远程桌面的连接问题还没解决，在本地将更新的公众号文章爬取，上传到obs中。



* 1. Ocr识别

Ocr识别：目前存储在/Unstructured\_Data/origin\_contents/日期/源名称/url hash/ocr目录下，测试了ocr运行时的内存占用情况，在处理361张图片，在其它程序暂停的情况下，初始内存占比12%，ocr程序占用65.8%的内存后，进程被杀死，。且ocr运行时会与其他线程冲突。

