工业企业关联信息挖掘与知识图谱构建技术说明文档

# 企业信息概念内涵

1. 企业概况：包括企业名称、所在地、行业类别、企业类型、成立时间、注册资本等基本信息。
2. 股东信息：包括股东名称、出资额、出资比例、联系方式等。
3. 组织架构：包括企业内部组织结构、管理层、员工人数、股东会、董事会等基本信息。
4. 财务信息：包括财务报表、销售收入、利润总额、净利润等信息。
5. 物资采购：包括主要供应商、采购渠道、物料类别等信息。
6. 生产管理：包括生产计划、实际生产量、生产效率、产品质量等方面的信息。
7. 销售与客户：包括销售渠道、重要客户、销售收入等信息。
8. 研发与知识产权：包括研发项目、专利申请及授权情况等信息。
9. 人力资源：包括员工招聘、培训、薪酬福利等基本信息。
10. 企业文化与社会责任：包括企业愿景、核心价值观、员工行为规范等信息以及企业在社会责任方面的表现。
11. 法律事务：包括企业涉及的法律法规、诉讼仲裁案件等信息。
12. 其他信息：包括企业其他重要事项，如投资情况、重要事件等。

# 企业信息—专利数据处理流程

### 数据样例展示

表 1 北京市两年内煤炭开采和洗选业专利数据样例（完整）

|  |  |  |
| --- | --- | --- |
| 类别 | 标识 | 样例1 |
| 行业大类编码 | big\_industry\_code | 06 |
| 行业大类名称 | big\_industry\_name | 煤炭开采和洗选业 |
| 省份编码 | province\_code | 110000 |
| 省份名称 | province\_name | 北京市 |
| 企业id | company\_id | CAICT\_COM\_100006\_C2A9FF4B9A889A4B168E99F8EB845B2E |
| 统一社会信用diamante | company\_credit\_code | 9111000010000085X6 |
| 公开号 | publication\_no | CN116687362A |
| 公开日期 | publication\_date | 2023-09-05 |
| 申请号 | apply\_no | CN202310711693.4 |
| 专利类型 | patent\_type | 发明专利 |
| 专利类型代码 | patent\_type\_code | 11 |
| 申请日期 | apply\_date | 2023-06-15 |
| 专利名称 | patent\_name | 一种岗前健康快速筛查系统及方法 |
| 主分类号 | main\_classification\_no | A61B5/0205 |
| 分类号 | classification\_no | A61B5/0205;A61B5/16;A61B5/00 |
| 申请人 | applicant | 中国中煤能源集团有限公司 |
| 发明人 | inventor\_ame | 周亚东;周应江;闫建华;王云泉;薛泽宇;赵学江;杨轲;赵玉宝;郭超荣;陈洪林;杜大文;葛庆岗;何华刚;史劢;刘波;杜祎康 |
| 代理机构 | agency\_org | 北京集佳知识产权代理有限公司 |
| 代理人 | agency\_name | 张雪娇 |
| 摘要 | abstract | 本申请提供一种岗前健康快速筛查系统及方法，方法包括：显示健康快速筛查界面；其中，健康快速筛查界面包含用于表征多项待测指标的检测信息；待测指标包括血压指标、肺功能指标、呼吸率、心率、精神状态、反应时间间和心理状态；根据待测指标，检测对应的检测数据；其中，检测数据包括血压数据、肺功能数据、心率数据、呼吸数据、精神状态数据、反应时间间和心理测试结果；根据检测数据，生成检测结果；显示检测结果。系统包括终端、肺功能检测设备、心率呼吸检测设备和血压检测设备，本方案可以通过系统同时检测人们身体的各项指标，提高了健康快速筛查的检测效率。本方案还可以打印检测结果，以便于人们更好地掌握自己的健康状况。 |



图 1 北京市两年内煤炭开采和洗选业专利数据样例（部分）

### 关键词抽取及数据处理算法部分

#### ：python环境安装

##### python安装（3.9版本，必须，其他版本未实验）

##### conda安装（可选）

##### pip安装包（需换国内源可参考）

C:\Users\用户名\AppData\Roaming\pip

[global]

index-url = http://mirrors.aliyun.com/pypi/simple

trusted-host = mirrors.aliyun.com

##### vscode安装（可选）

##### VS code下使用python环境图示

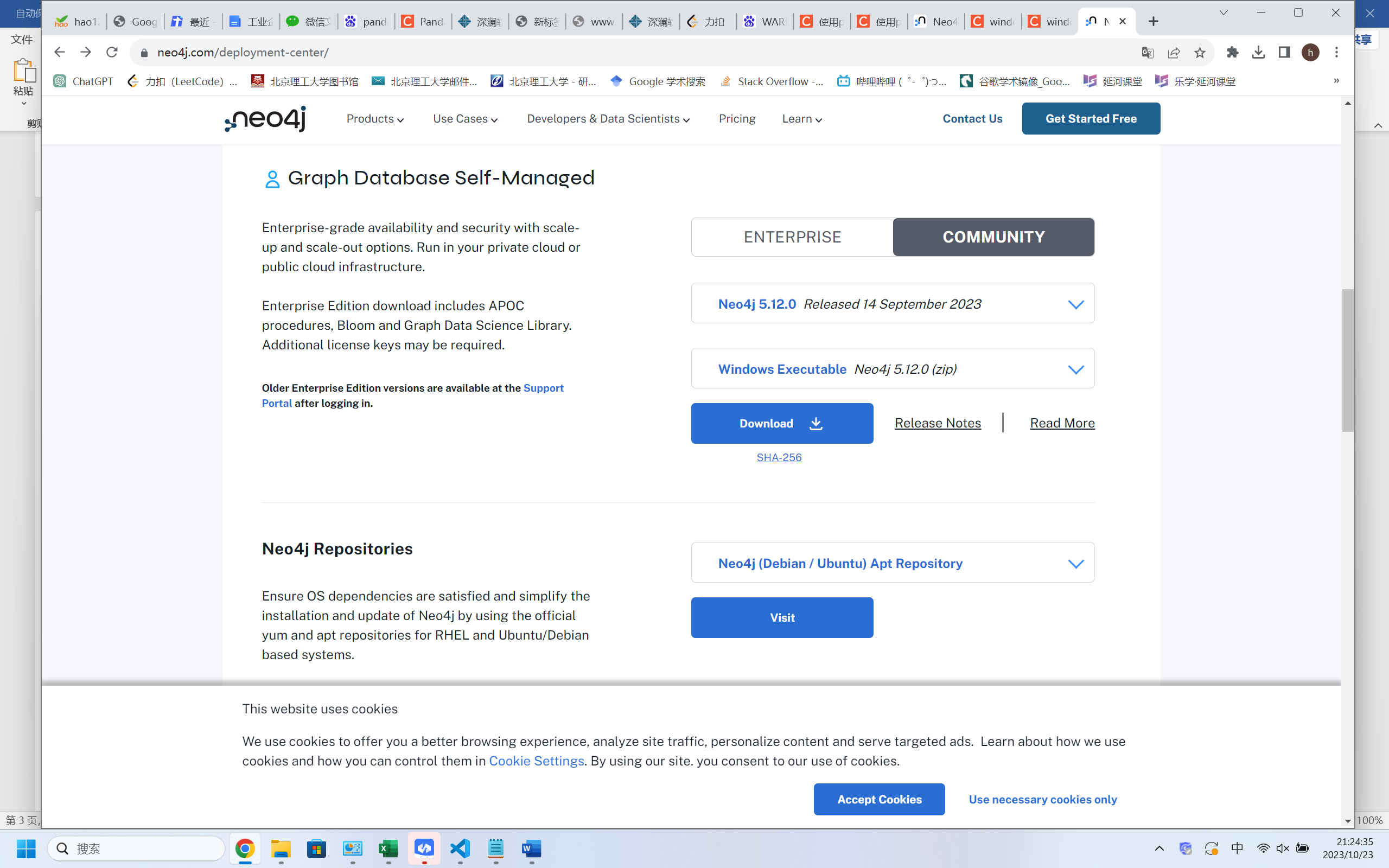
### 数据处理流程

CSV 文件的第一行包含了列名接下来的行包含了数据。脚本会逐行读取数据，并将其存入 Neo4j 数据库中。首先安装neo4j

#### Java环境及Neo4j安装

以windows下为例：

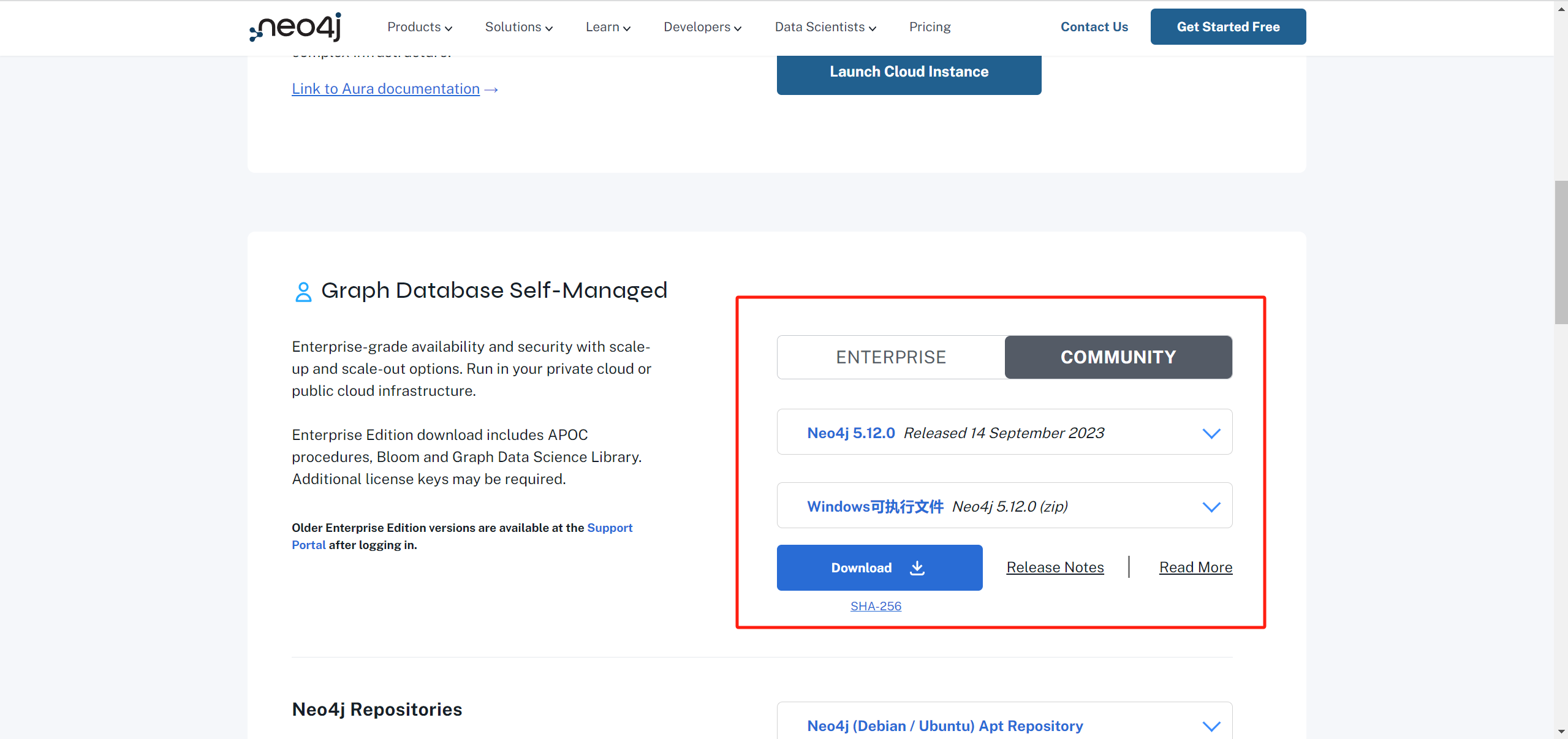
##### neo4j安装压缩文件下载



Neo4j下载页面：<https://neo4j.com/deployment-center/#community>

可参考：https://blog.csdn.net/happypipi123/article/details/124513215。

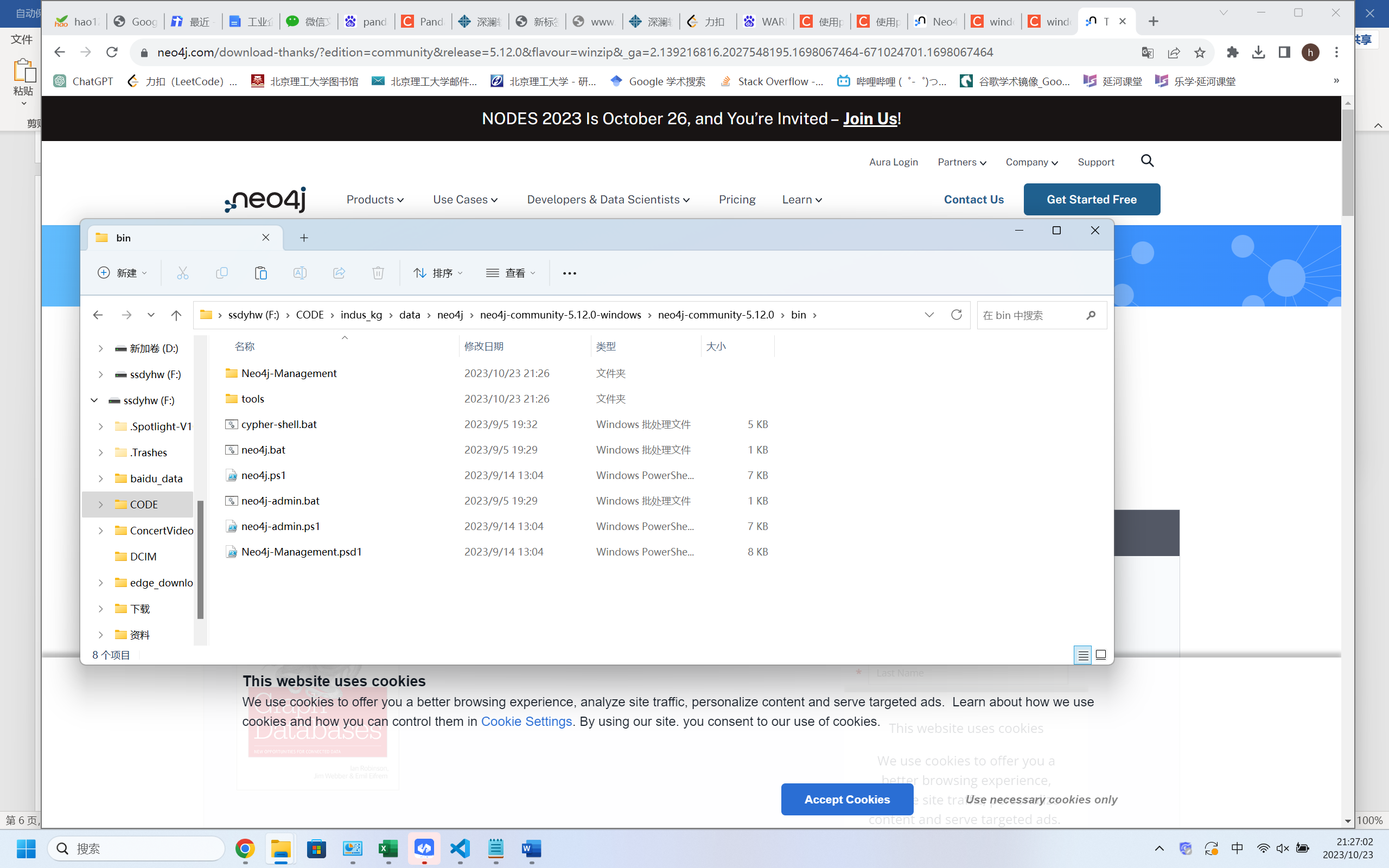
##### 配置Java 17的环境变量可参考https://blog.csdn.net/weixin\_44329942/article/details/125314276



如使用以上Neo4j下载链接，

Jdk版本注意Please use Oracle(R) Java(TM) 17, OpenJDK(TM) 17 to run Neo4j Server，下载链接：<https://jdk.java.net/java-se-ri/17>，

2.安装参考：https://blog.csdn.net/happypipi123/article/details/124513215

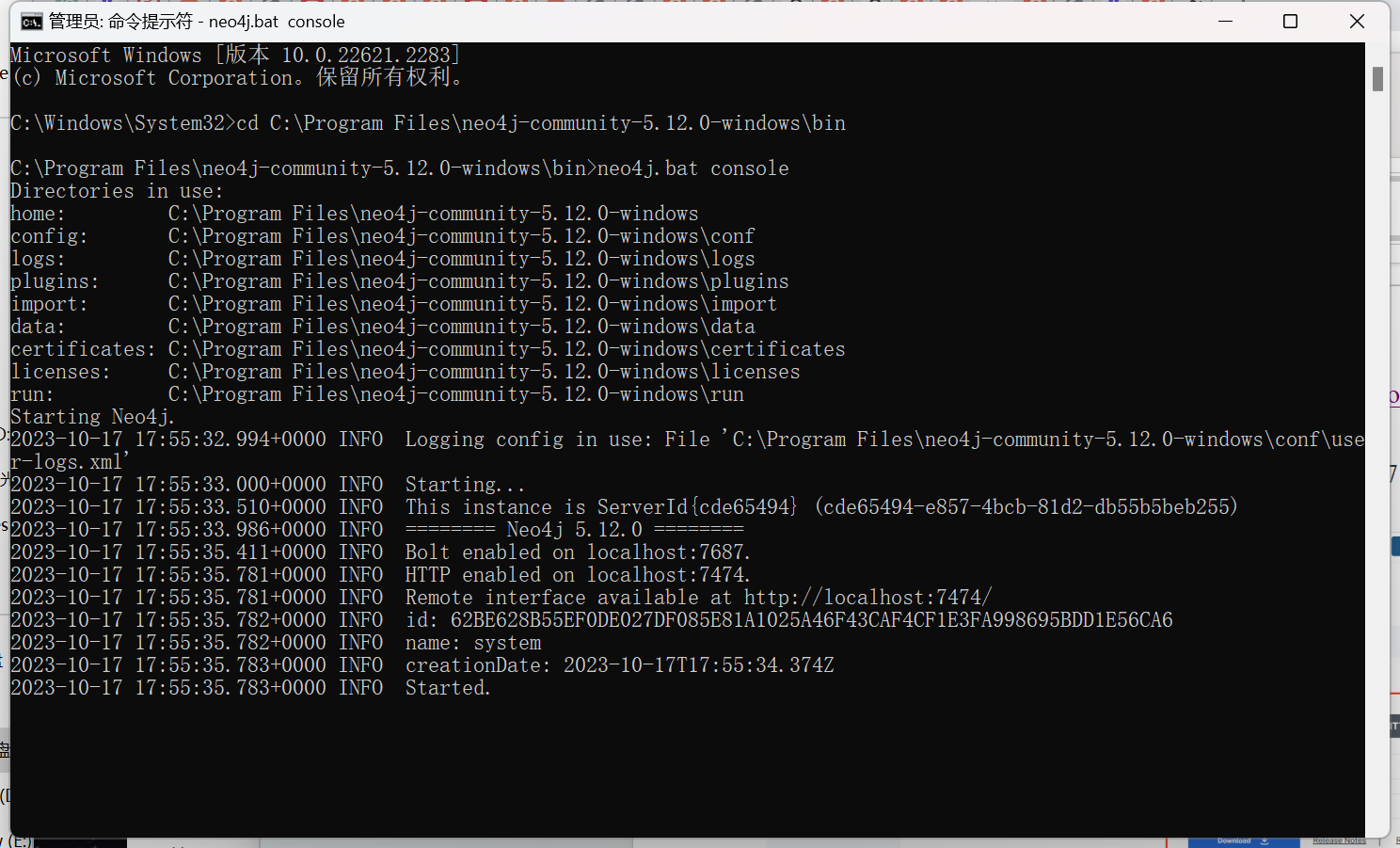
用**管理员身份**打开CMD，切换到neo4j的bin目录下，输入neo4j.bat console

如：

cd C:\Program Files\neo4j-community-5.12.0-windows\bin

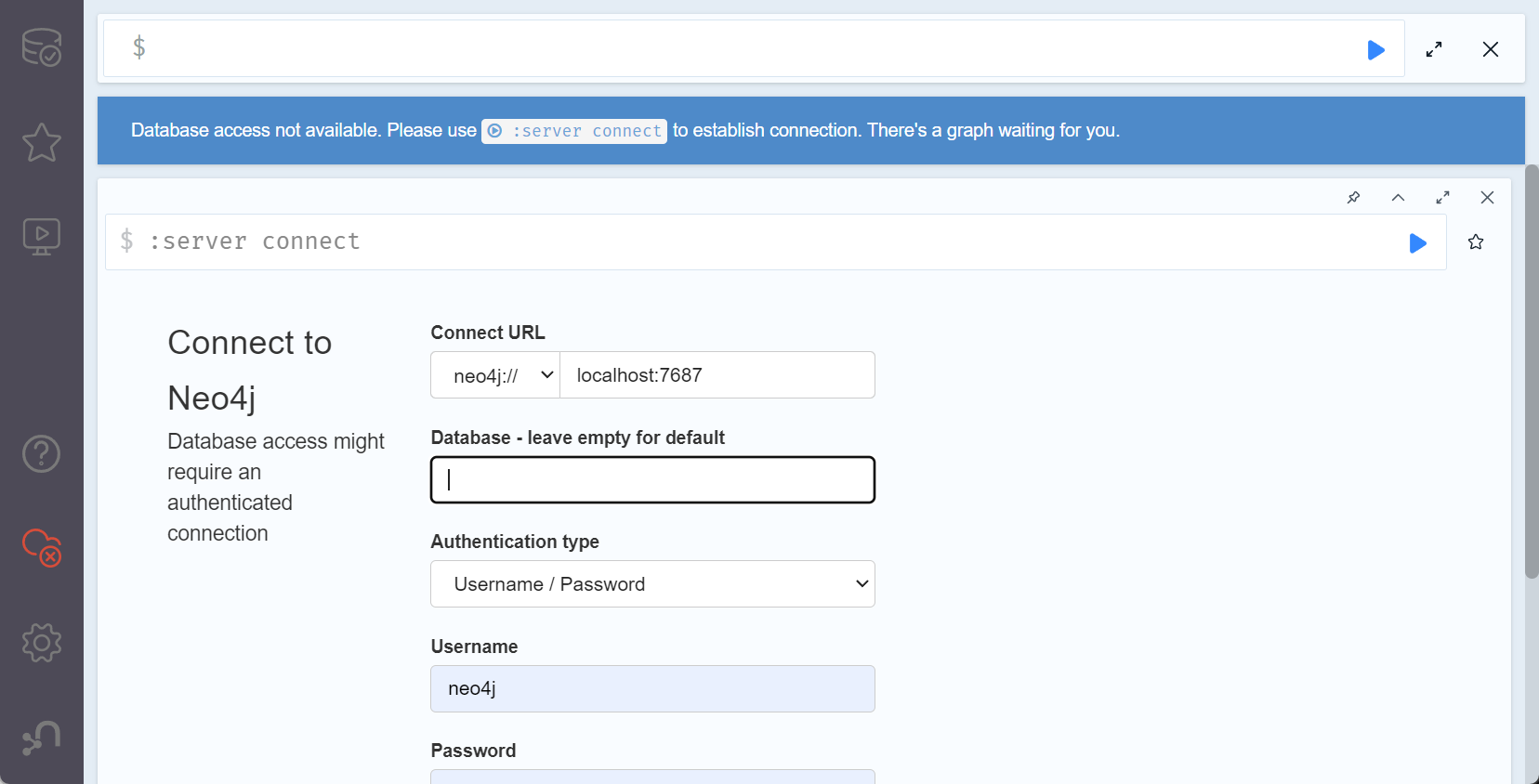
neo4j.bat console

出现下图即启动成功。

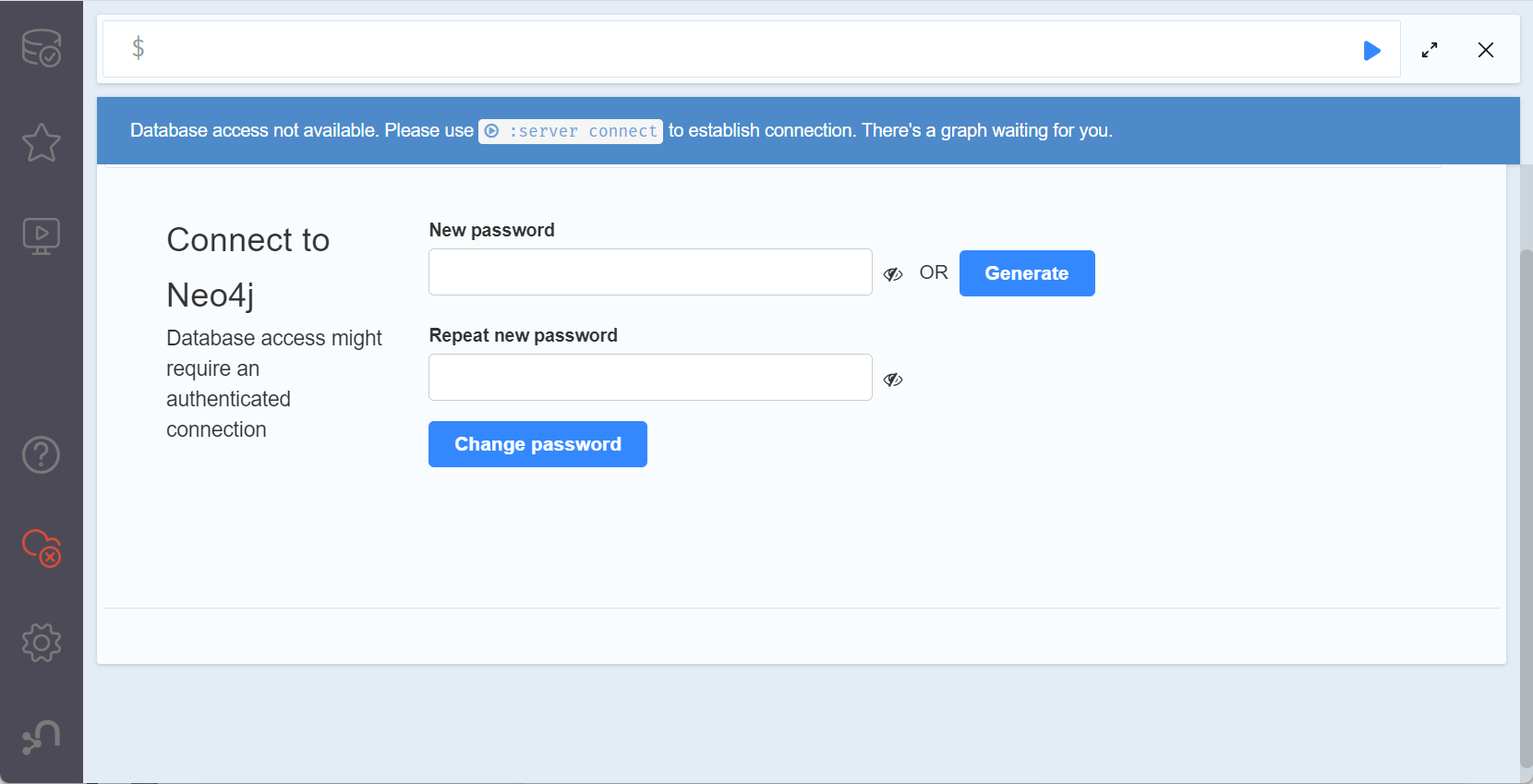


登录到：

<http://localhost:7474/browser/>



初始用户名和登录密码都是neo4j，进入更改密码，再登录即可使用。



#### 实体意义

##### 节点类型：

Industry（行业）, Province（省份）, Company（企业）, Patent（专利）, Inventor（发明人）, Agency（代理机构）

##### 关系类型：

BELONGS\_TO（行业与省份之间的关系）, LOCATED\_IN（企业所在省份的关系）, FILED\_BY（专利由企业申请的关系）, INVENTED\_BY（专利由发明人发明的关系）, HANDLED\_BY（专利由代理机构处理的关系）