**Datacleaner**

**一．概要**

**1.描述：**

DataCleaner 是一个数据质量分析,比较,验证和监督的软件。DataCleaner包括一个独立的图形用户界面分析,比较和验证,并进行监测web应用。

**2.链接地址：**

官网地址：<https://datacleaner.org/>

源码包地址：<https://github.com/datacleaner/DataCleaner>

**3.源码包主要模块描述：**

api：DataCleaner的公共API。 主要是为了构建自己的扩展而提供的接口和注解。

resources：DataCleaner的静态资源

oss-branding：图标和颜色

test-ware：用于DataCleaner的单元测试的类和扩展代码

engine

core：核心引擎部分，它允许根据API执行作业和组件。

xml-config：包含读写任务文件和配置文件清理工具。

env：DataCleaner可以运行的不同/替代环境，例如Apache Spark或webapp-cluster

components

　- 许多子模块包含内置以及与DataCleaner一起使用的其他组件/扩展。

　- 标准组件：一个容器项目，它依赖于通常捆绑在DataCleaner社区版中的所有组件。

desktop

　- api DataCleaner桌面应用程序的公共API。

　- ui 适用于桌面用户的基于Swing的用户界面。

monitor

　- api DataCleaner监控模块的API类和接口

**二．安装运行桌面程序**

**1.安装环境要求：**

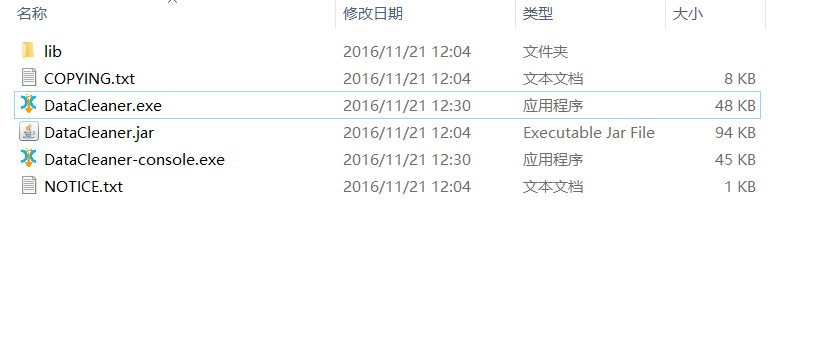
1. 一台有图形界面的计算机（命令行模式除外）
2. java7 或以上
3. DataCleaner 的软件许可文件

**2.运行桌面程序：**

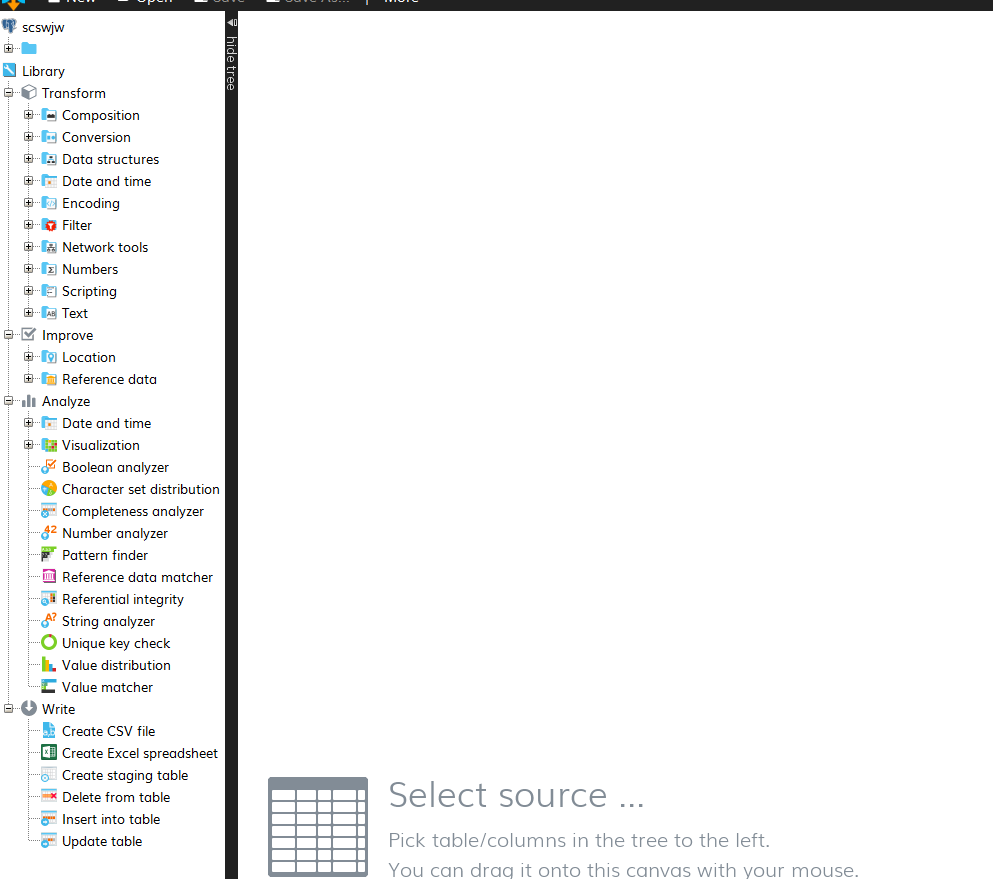
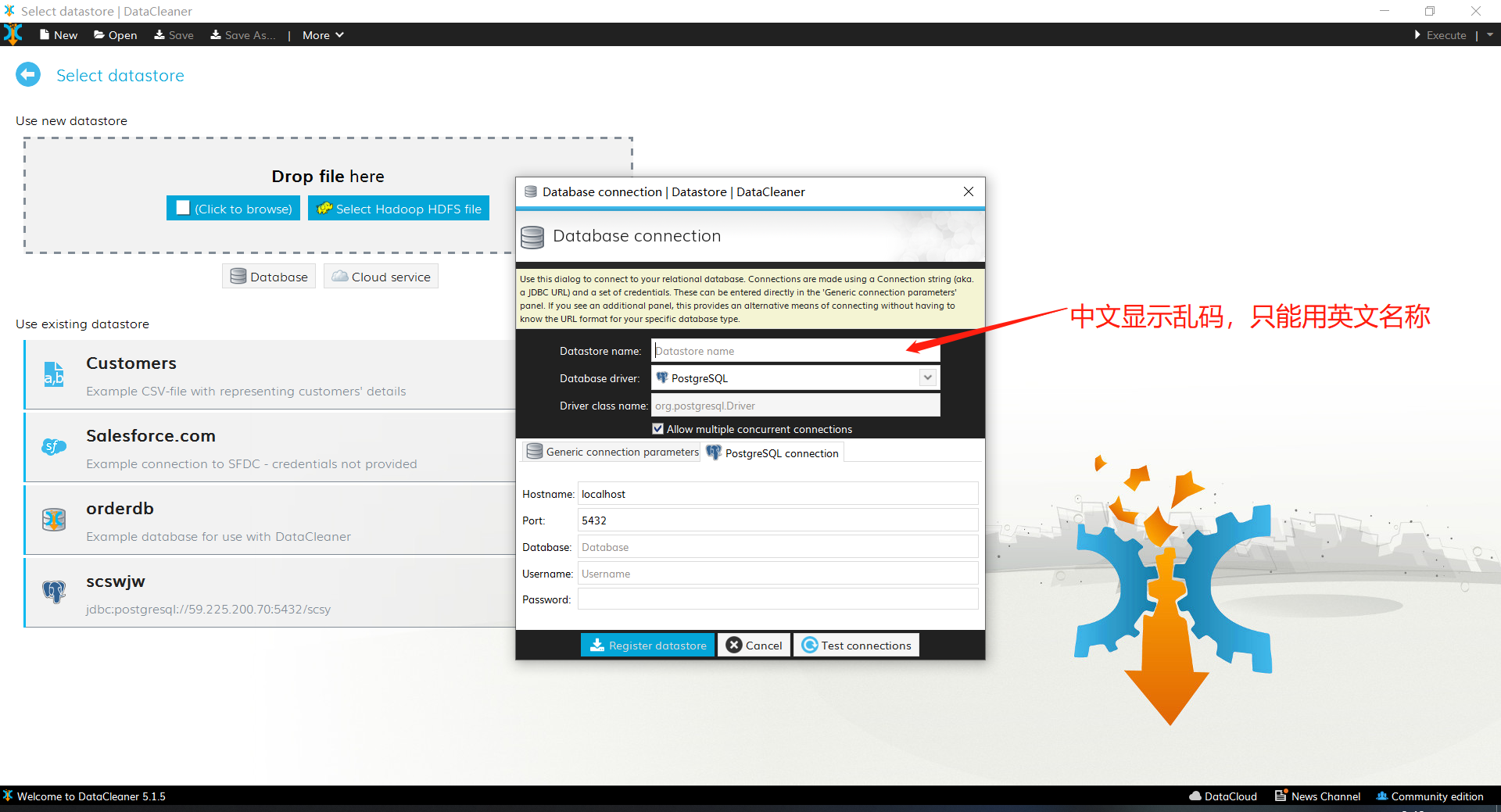
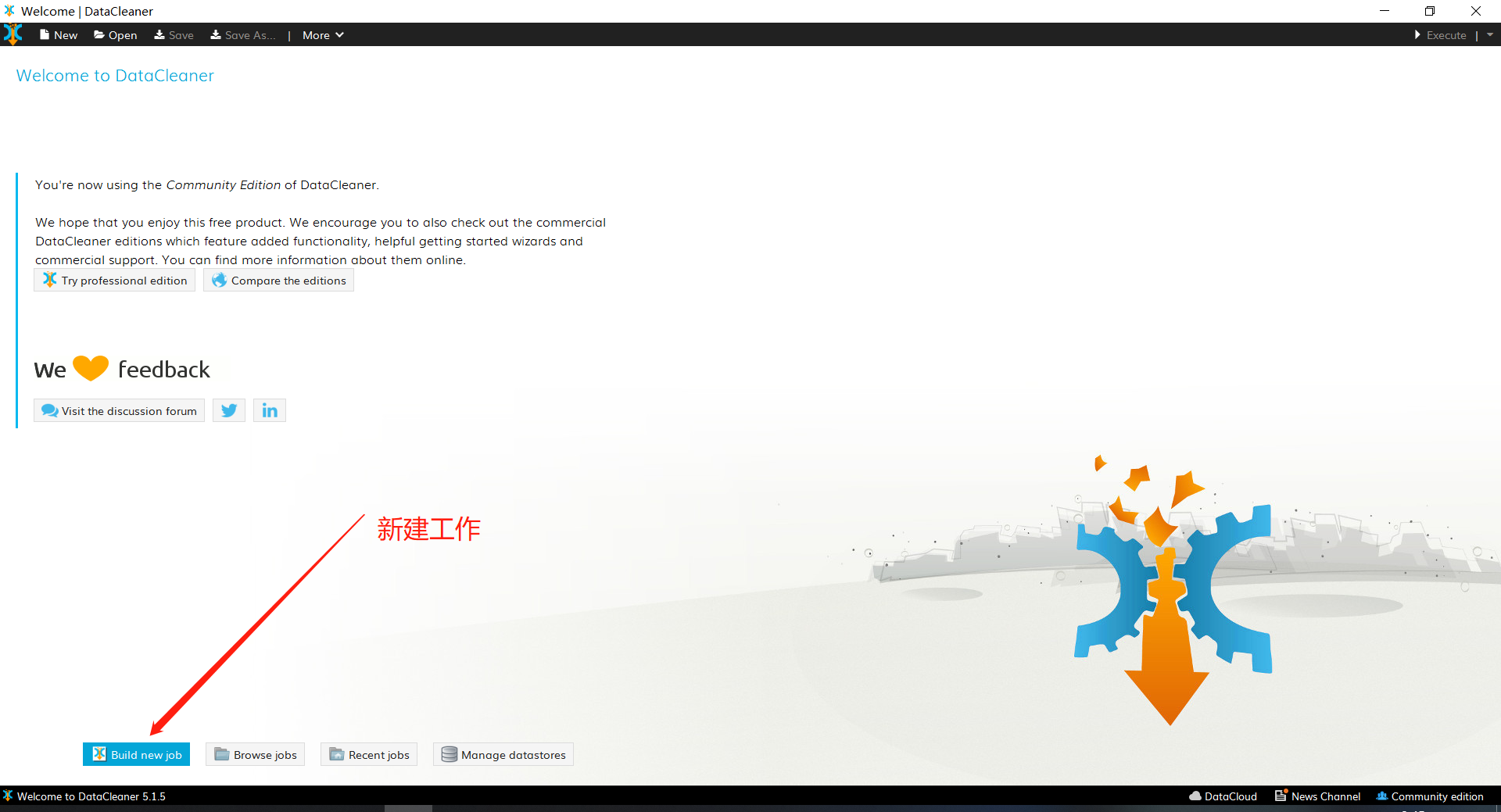
1.从官网下载Windows版本安装包，本地解压后得到可执行程序包

1）社区5.5.0版（及以上版本）zip下载后直接解压使用 [datacleaner.sh](http://datacleaner.sh/) 或 datacleaner.cmd启动

2) 本例使用社区5.2.0版zip下载后直接解压使用DataCleaner.exe,打开客户端



2.连接上数据库，可以快速对数据进行分析



3.组件菜单：

Library --- 组件库

…Transform --- 转换

……Composition --- 合并

…………[Union](https://datacleaner.github.io/docs/5.4.0/components/union.html) ---联合

…………Fuse/Coalesce fields --- 融合/合并字段

…………Grouper --- 聚合分组

…………Invoke child Analysis job --- 调用子级分析任务

……Conversion --- 转换

…………[Convert](https://datacleaner.github.io/docs/5.4.0/components/union.html) to boolean ---转换为布尔值

…………[Convert](https://datacleaner.github.io/docs/5.4.0/components/union.html) to string ---转换为字符串

…………[Convert](https://datacleaner.github.io/docs/5.4.0/components/union.html) to date ---转换为日期

…………[Convert](https://datacleaner.github.io/docs/5.4.0/components/union.html) to number ---转换为数字

……Data structures --- 数据结构

…………[Read](https://datacleaner.github.io/docs/5.4.0/components/union.html) keys and values from map ---从映射中读取键和值

…………[Select](https://datacleaner.github.io/docs/5.4.0/components/union.html) values from list ---从列表中选择值

…………[Bulid](https://datacleaner.github.io/docs/5.4.0/components/union.html) list ---构建列表

…………[Select](https://datacleaner.github.io/docs/5.4.0/components/union.html) values from key/value map ---从键/值映射中选择值

…………[Build](https://datacleaner.github.io/docs/5.4.0/components/union.html) key/value map ---构建键/值映射

…………[Compose](https://datacleaner.github.io/docs/5.4.0/components/union.html) & write JSON document ---组成并编写JSON文档

…………[Read](https://datacleaner.github.io/docs/5.4.0/components/union.html) elements form list ---从列表中读取元素

…………[Read](https://datacleaner.github.io/docs/5.4.0/components/union.html) & parse JSON documnet ---读取并解析JSON文档

……Date and time --- 日期和时间

…………[Extract](https://datacleaner.github.io/docs/5.4.0/components/union.html) date part ---日期截取

…………[Generate](https://datacleaner.github.io/docs/5.4.0/components/union.html) timestamp ---生成时间戳

…………[Timestamp](https://datacleaner.github.io/docs/5.4.0/components/union.html) converter ---时间戳转换

…………[Date](https://datacleaner.github.io/docs/5.4.0/components/union.html) difference / period length ---日期差异/周期长度

…………[Date](https://datacleaner.github.io/docs/5.4.0/components/union.html) to age ---日期转换年龄

…………[Format](https://datacleaner.github.io/docs/5.4.0/components/union.html) date ---日期格式化

…………[Date](https://datacleaner.github.io/docs/5.4.0/components/union.html) mask matcher ---日期掩码匹配器

…………[Capture](https://datacleaner.github.io/docs/5.4.0/components/union.html) changed records ---捕获更改记录

……Encoding --- 编码

…………[XML](https://datacleaner.github.io/docs/5.4.0/components/union.html) encoder ---XML编码器

…………[Hash](https://datacleaner.github.io/docs/5.4.0/components/union.html) value ---哈希值

…………[URL](https://datacleaner.github.io/docs/5.4.0/components/union.html) encoder ---URL编码器

…………[Transliterate](https://datacleaner.github.io/docs/5.4.0/components/union.html) ---直译

…………[XML](https://datacleaner.github.io/docs/5.4.0/components/union.html) decoder ---XML解码器

…………[HTML](https://datacleaner.github.io/docs/5.4.0/components/union.html) encoder ---HTML编码器

……Filter --- 过滤器

…………[Equals](https://datacleaner.github.io/docs/5.4.0/components/union.html) ---等于

…………[Validate](https://datacleaner.github.io/docs/5.4.0/components/union.html) in dictionary ---在字典中验证

…………[Compare](https://datacleaner.github.io/docs/5.4.0/components/union.html) ---比较

…………[String](https://datacleaner.github.io/docs/5.4.0/components/union.html) value range ---字符串值范围

…………[Validate](https://datacleaner.github.io/docs/5.4.0/components/union.html) with string pattern ---使用字符串模式验证

…………[Null](https://datacleaner.github.io/docs/5.4.0/components/union.html) check ---空检查

…………[Max](https://datacleaner.github.io/docs/5.4.0/components/union.html) rows ---最大行数

…………[Number](https://datacleaner.github.io/docs/5.4.0/components/union.html) range ---数字范围

…………[Capture](https://datacleaner.github.io/docs/5.4.0/components/union.html) changed records ---捕获更改的记录

…………[String](https://datacleaner.github.io/docs/5.4.0/components/union.html) length range ---字符串长度范围

…………[Date](https://datacleaner.github.io/docs/5.4.0/components/union.html) range ---日期范围

……Network tools --- 网络工具

…………[Convert](https://datacleaner.github.io/docs/5.4.0/components/union.html) number to IP ---将数字转换为IP

…………[Resolve](https://datacleaner.github.io/docs/5.4.0/components/union.html) hostname ---解析主机名

…………[Convert](https://datacleaner.github.io/docs/5.4.0/components/union.html) IP to number ---将IP转换为数字

……Numbers --- 数值

…………[Round](https://datacleaner.github.io/docs/5.4.0/components/union.html) number ---近似数

…………[Math](https://datacleaner.github.io/docs/5.4.0/components/union.html) formula ---数学公式

…………[Generate](https://datacleaner.github.io/docs/5.4.0/components/union.html) UUID ---生成UUID

…………[Generate](https://datacleaner.github.io/docs/5.4.0/components/union.html) ID ---生成ID

…………[Increment](https://datacleaner.github.io/docs/5.4.0/components/union.html) number ---增加数量

……Scripting --- 脚本

…………[Groovy](https://datacleaner.github.io/docs/5.4.0/components/union.html) transformer (advanced) ---Groovy高级转换器

…………[Expression](https://datacleaner.github.io/docs/5.4.0/components/union.html) language (EL) transformer ---EL表达式转换器

…………[Groovy](https://datacleaner.github.io/docs/5.4.0/components/union.html) transformer (simple) ---Groovy普通转换器

…………[JavaScript](https://datacleaner.github.io/docs/5.4.0/components/union.html) transformer (simple) ---JS普通转换器

…………[JavaScript](https://datacleaner.github.io/docs/5.4.0/components/union.html) transformer (advanced) ---JS高级转换器

…………[JavaScript](https://datacleaner.github.io/docs/5.4.0/components/union.html) filter ---JS过滤器

……Text --- 文本

…………[URL](https://datacleaner.github.io/docs/5.4.0/components/union.html) parser ---URL解析器

…………[Concatenator](https://datacleaner.github.io/docs/5.4.0/components/union.html) ---串接合并

…………[Whitespace](https://datacleaner.github.io/docs/5.4.0/components/union.html) trimmer ---空白微调

…………[Plain](https://datacleaner.github.io/docs/5.4.0/components/union.html) search/replace ---简单搜索/替换

…………[Regex](https://datacleaner.github.io/docs/5.4.0/components/union.html) parser ---正则表达式解析器

…………[Remove](https://datacleaner.github.io/docs/5.4.0/components/union.html) substring ---截取字符串

…………[Text](https://datacleaner.github.io/docs/5.4.0/components/union.html) case transformer ---文本案例转换器

…………[Remove](https://datacleaner.github.io/docs/5.4.0/components/union.html) unwanted characters ---删除不需要的字符

…………[Regex](https://datacleaner.github.io/docs/5.4.0/components/union.html) search/replace ---正则表达式搜索/替换

…………[Tokenizer](https://datacleaner.github.io/docs/5.4.0/components/union.html) ---分词器

…Improve --- 改善

……Location --- 本地

…………[Country](https://datacleaner.github.io/docs/5.4.0/components/union.html) standardizer ---国际标准化

……Reference data --- 参考数据

…………[String](https://datacleaner.github.io/docs/5.4.0/components/union.html) pattern matcher ---字符串模式匹配器

…………[Remove](https://datacleaner.github.io/docs/5.4.0/components/union.html) dictionary matches ---删除字典匹配

…………[HTTP](https://datacleaner.github.io/docs/5.4.0/components/union.html) request ---HTTP请求

…………[Dictionary](https://datacleaner.github.io/docs/5.4.0/components/union.html) matcher ---字典匹配器

…………[Table](https://datacleaner.github.io/docs/5.4.0/components/union.html) lookup ---表查找

…………[Synonym](https://datacleaner.github.io/docs/5.4.0/components/union.html) lookup ---同义词查找

…Analyze --- 分析

……Date and time --- 日期和时间

…………[Month](https://datacleaner.github.io/docs/5.4.0/components/union.html) distribution ---月份分布

…………[Year](https://datacleaner.github.io/docs/5.4.0/components/union.html) distribution ---年份分布

…………[Weekday](https://datacleaner.github.io/docs/5.4.0/components/union.html) distribution ---每周分布

…………[Week](https://datacleaner.github.io/docs/5.4.0/components/union.html) number distribution ---周数分布

…………[Date/time](https://datacleaner.github.io/docs/5.4.0/components/union.html) analyzer ---日期/时间分析器

…………[Date](https://datacleaner.github.io/docs/5.4.0/components/union.html) gap analyzer ---日期差距分析器

……Visualization --- 可视化

…………[Density](https://datacleaner.github.io/docs/5.4.0/components/union.html) plot ---密度图

…………[Stacked](https://datacleaner.github.io/docs/5.4.0/components/union.html) area plot ---堆叠面积图

…………[Scatter](https://datacleaner.github.io/docs/5.4.0/components/union.html) plot ---散点图

……Number analyzer ---数字分析仪

……Fill pattern ---填充模式

……Unique key check ---唯一键检查

……Completeness analyzer ---完整性分析仪

……Reference data matcher ---参考数据匹配器

……Value matcher ---值匹配器

……Value distribution ---值分布

……Referential integrity ---参照完整性

……Pattern finder ---模式查找器

……Boolean analyzer ---布尔分析器

……Mark rows ---标记行

……Character set distribution ---字符集分布

……String analyzer ---字符串分析器

…Write --- 写入

……Create CSV file ---创建CSV文件

……Insert into table ---向表中插入

……Create staging table ---创建临时表

……Delete from table ---从表中删除

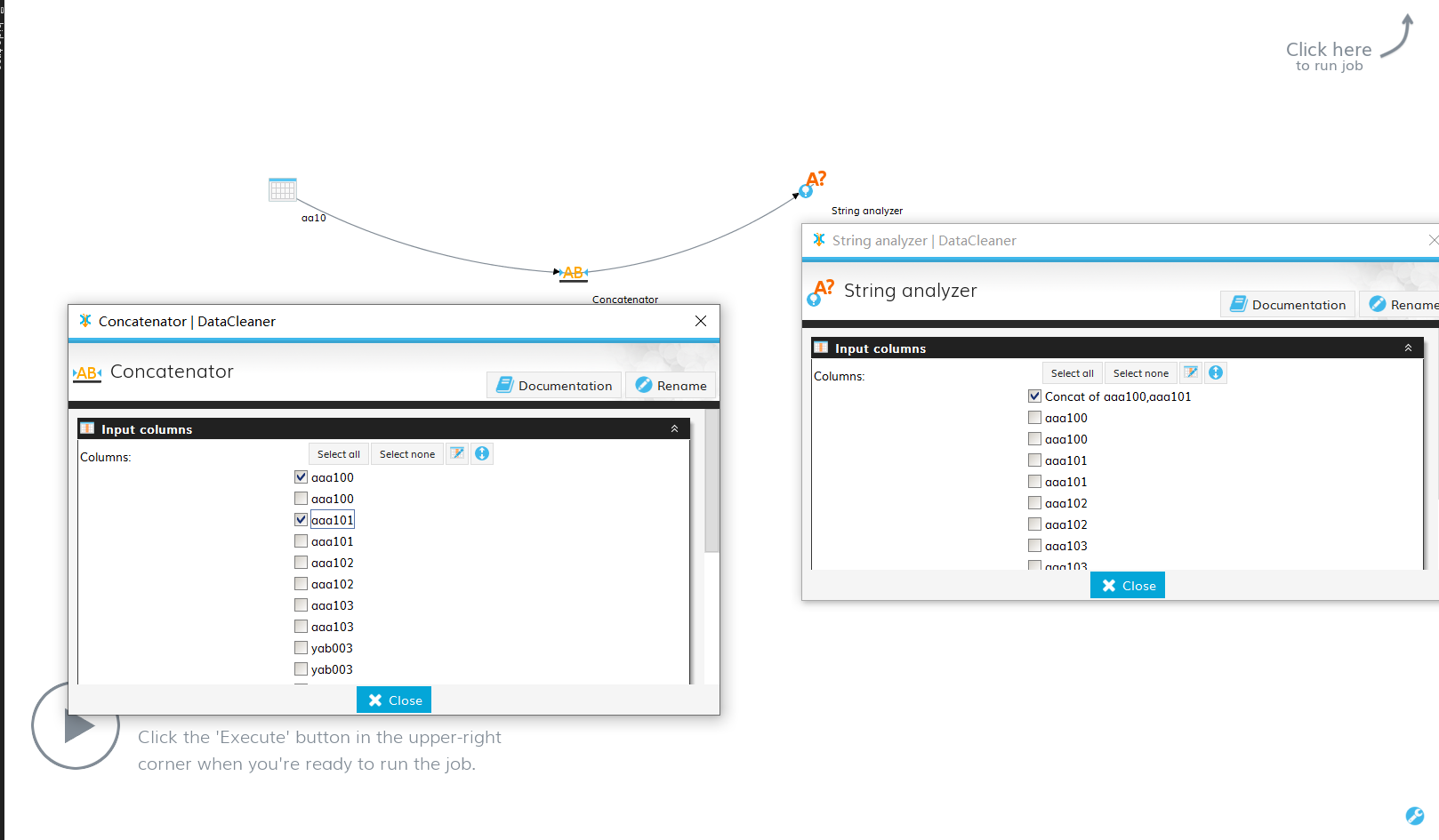
……Create Excel spreadsheet ---创建Excel电子表格

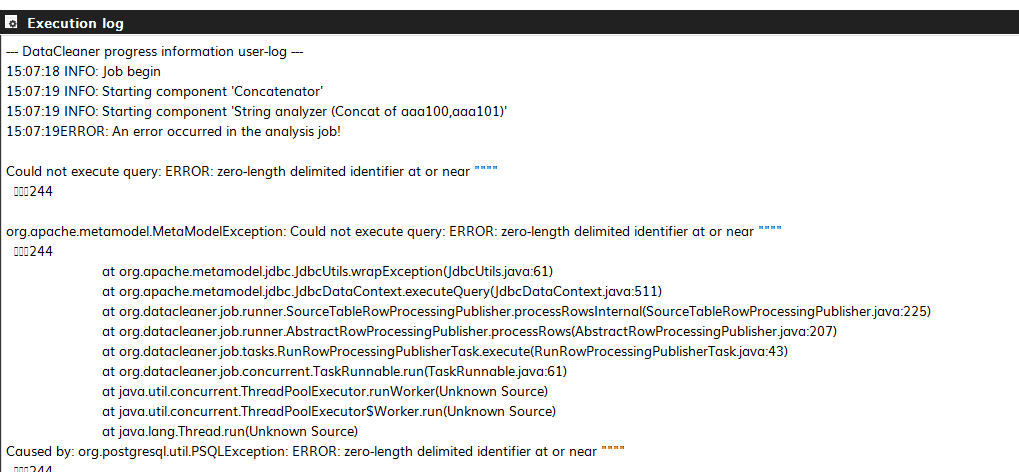
……Update table ---更新表

=============================分割线=========================

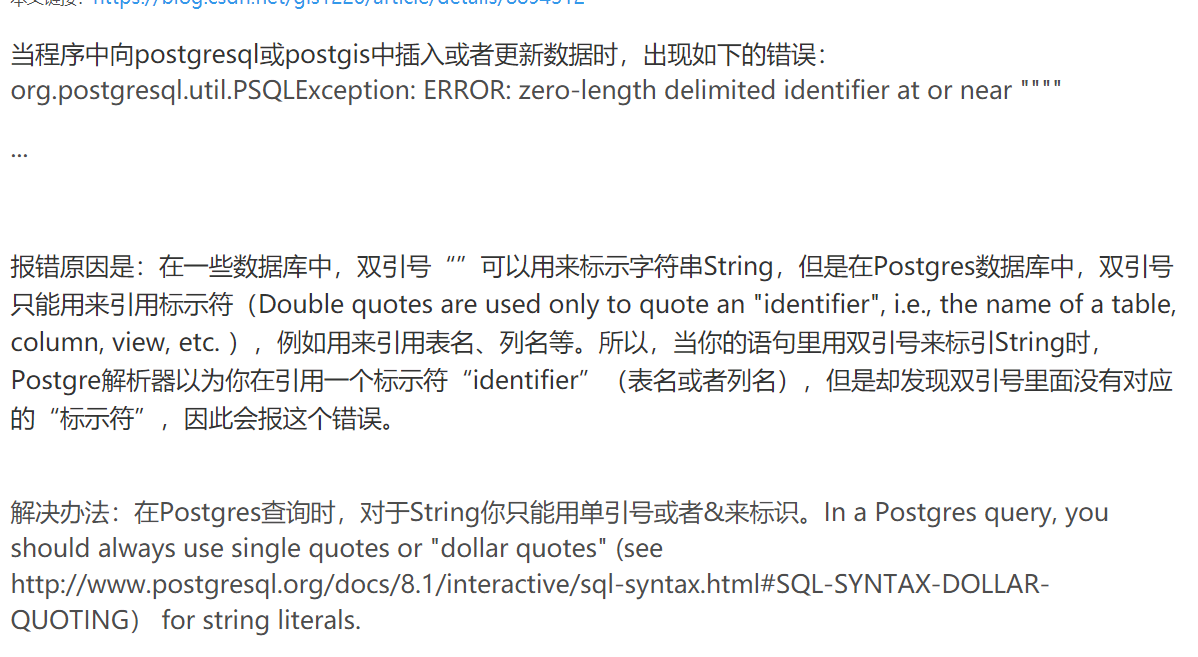
**记录遇到的迫切需要解决的问题**

对aa10表aaa100、aaa101字段数据合并，并对合并后的数据进行分析

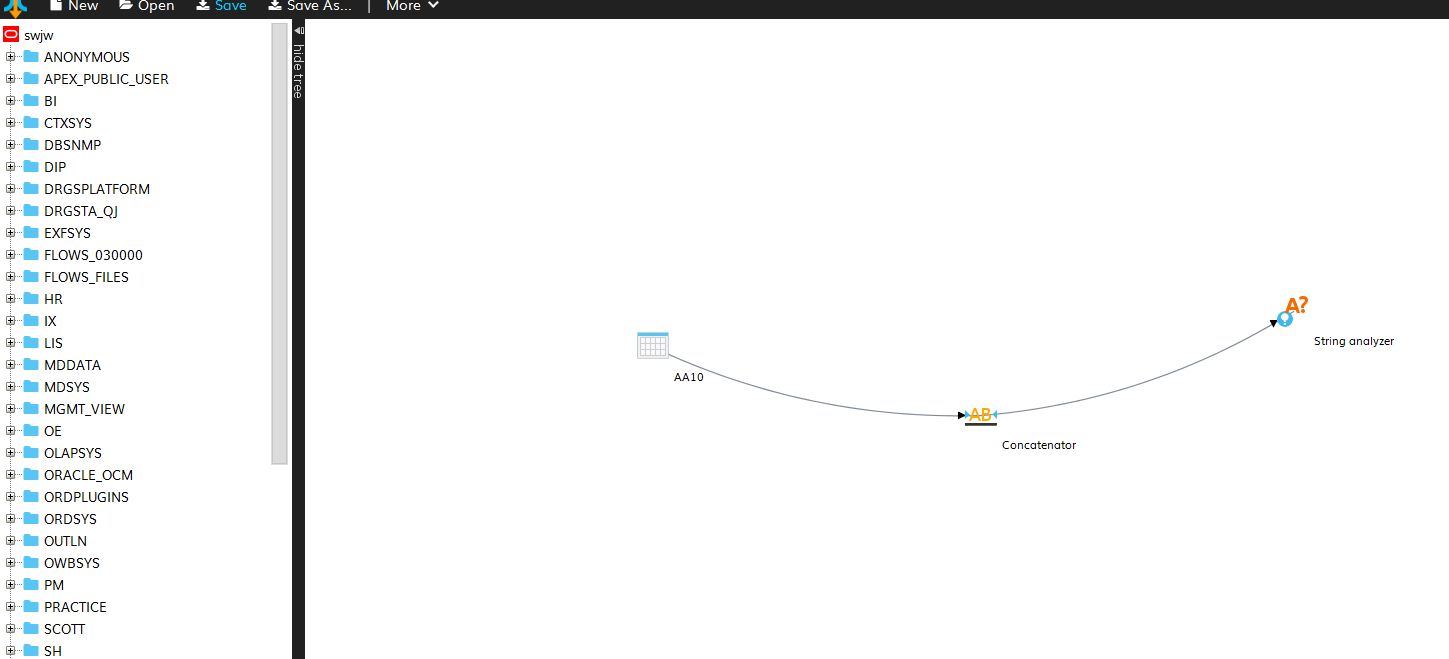


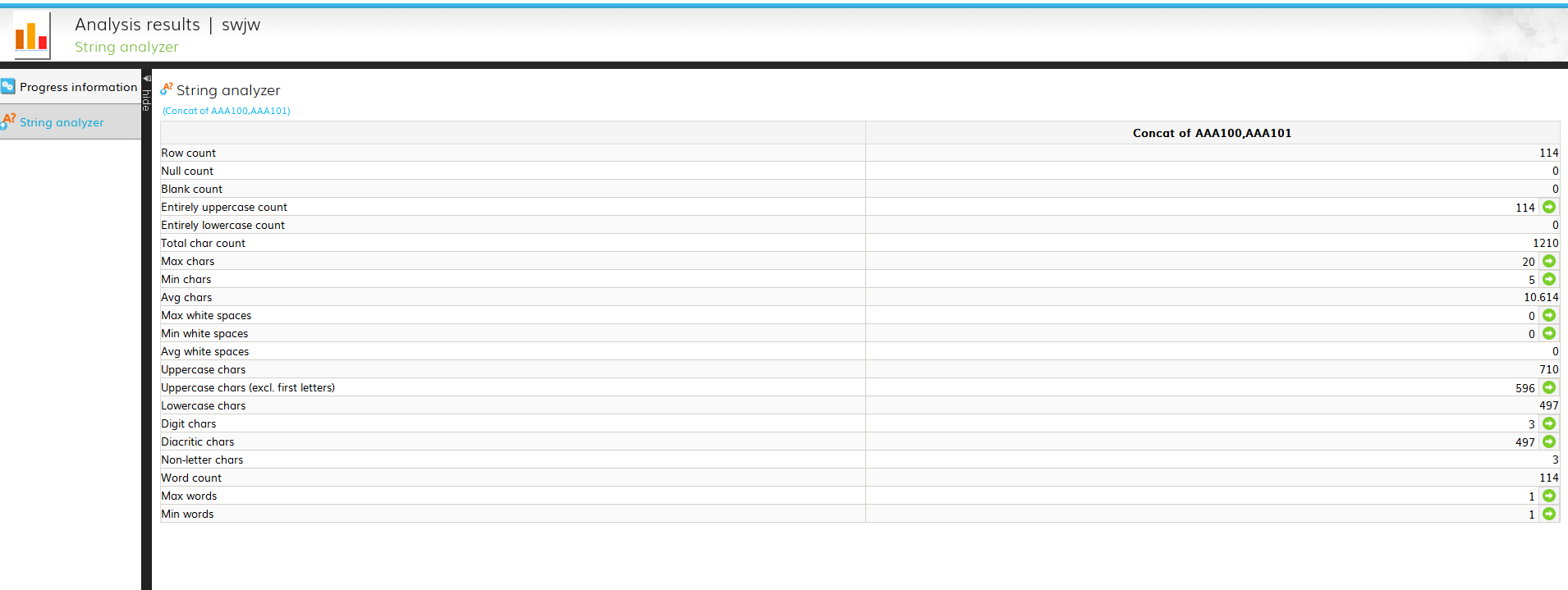


后经过多次不同分析任务测试，run job 时均会报这个错误，无法看到数据分析报告。解决思路如下



**用Oracle数据库测试一下，查看是否为数据库原因造成此问题**。



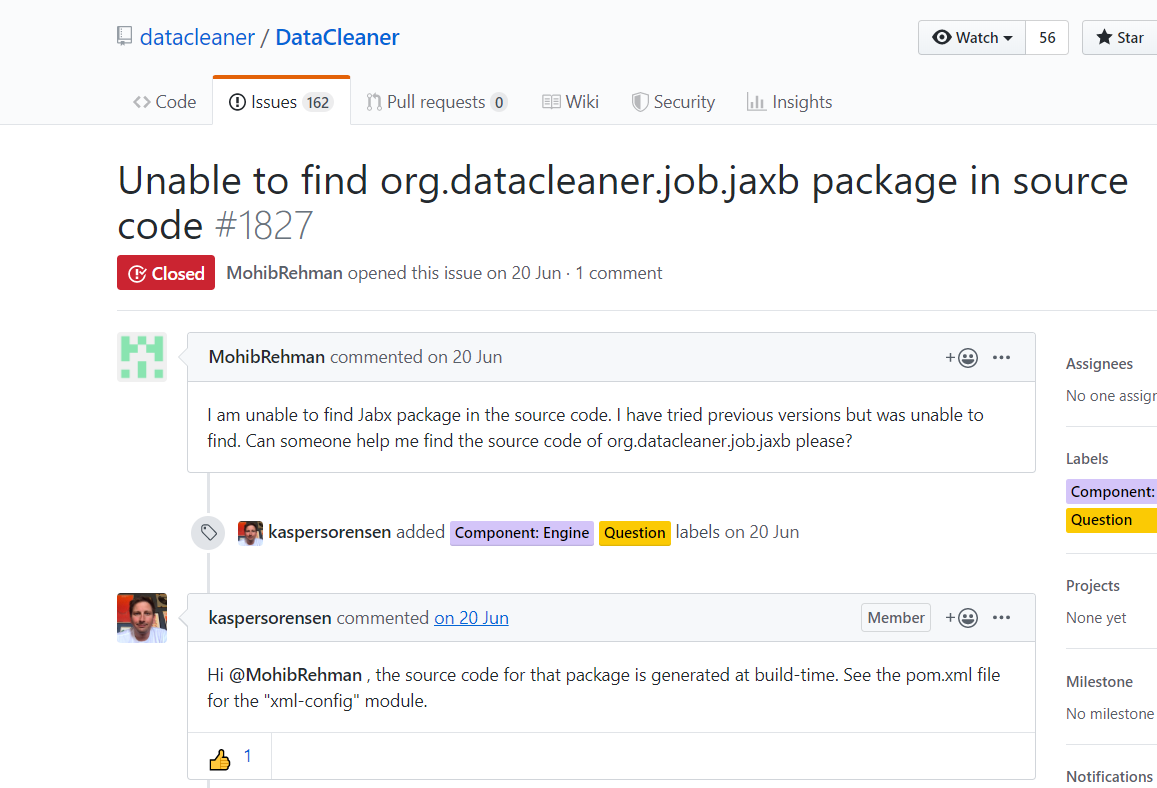


Oracle数据库经测试可以正常输出分析结果

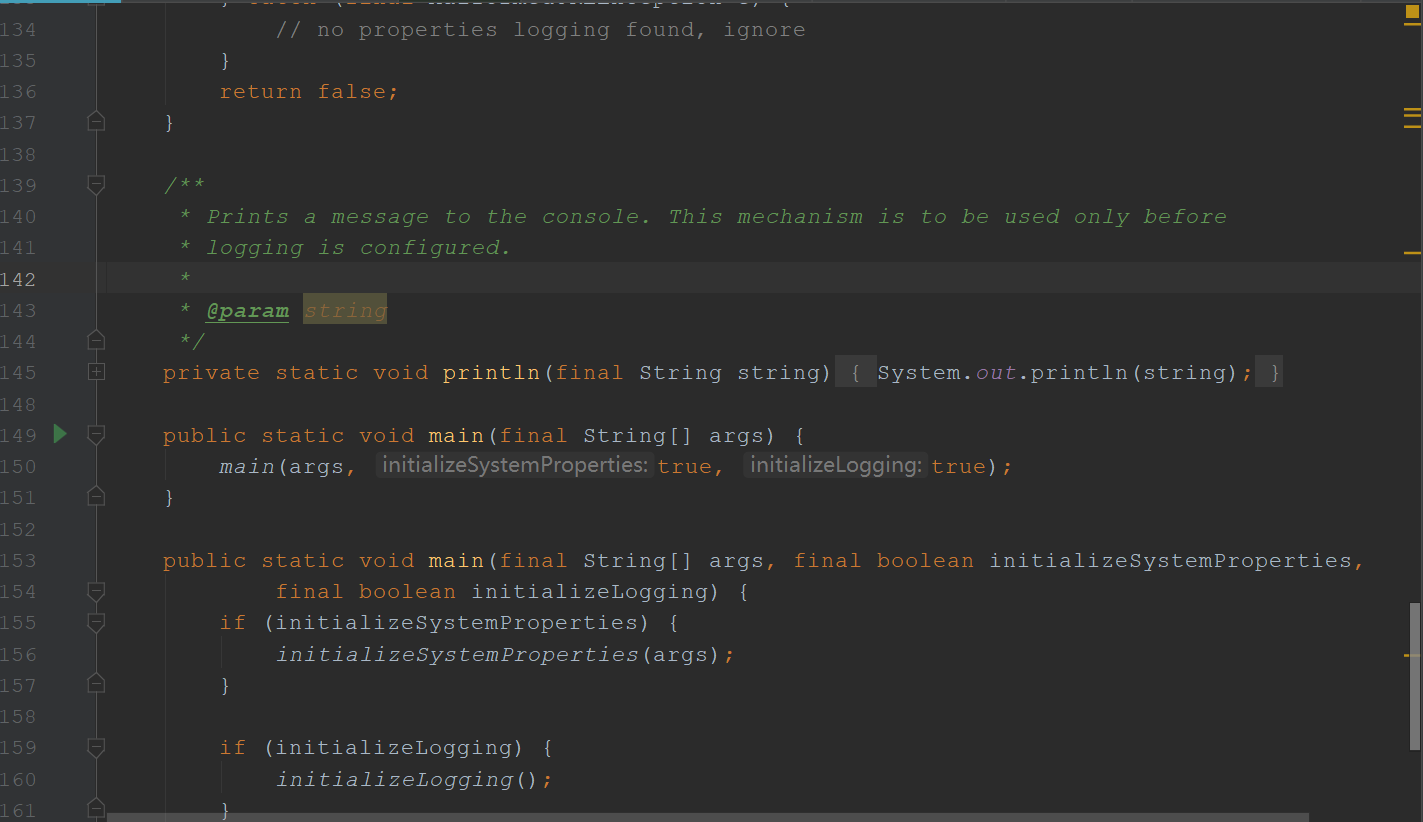
解决思路：

这里报错是直接在运行桌面程序时报的错，无法查看对应的一些错误日志信息。准备从源码部分着手，通过控制台、断点调试等方式来查看问题。

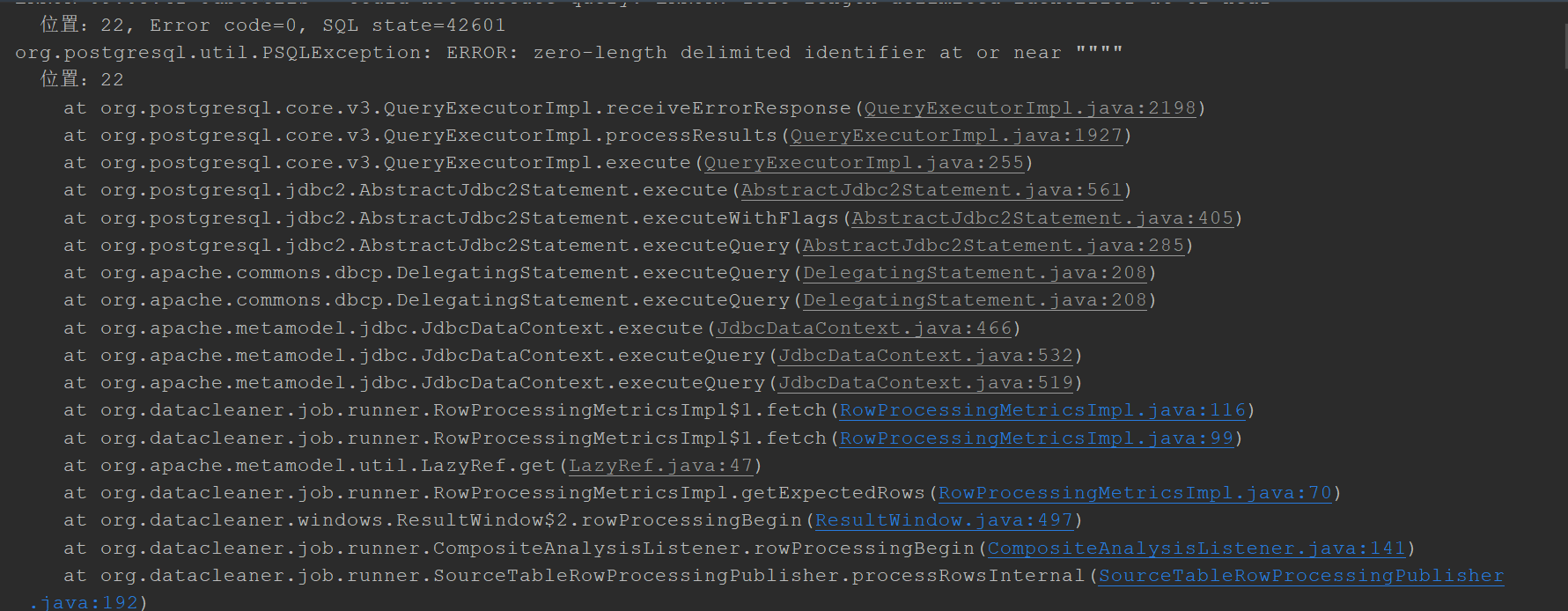
将datacleaner源码导入idea，根据pom.xml导入所有的依赖之后，找不到org.datacleaner.job.jaxb这个包，对应有几个类依然报错，一开始也找不到解决办法，几经周折最后在github上面找到相同问题，得到解决方法。



在代码里找到桌面程序的启动入口，运行main()方法，启动桌面程序



桌面程序启动后，配好任务，执行任务时依旧报同样的错，控制台输出:



接下来依次对蓝色标示的Java文件进行检查，以期查找到问题原因