### 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。

### 从GitHub下载源码包本地打开

### 

### 源码包主要的模块描述：

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。

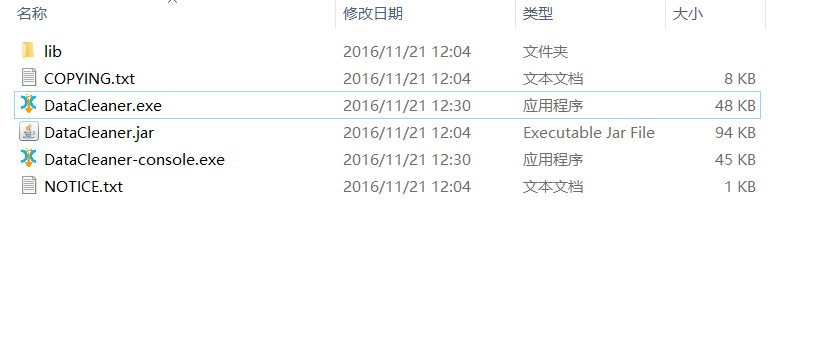
monitor

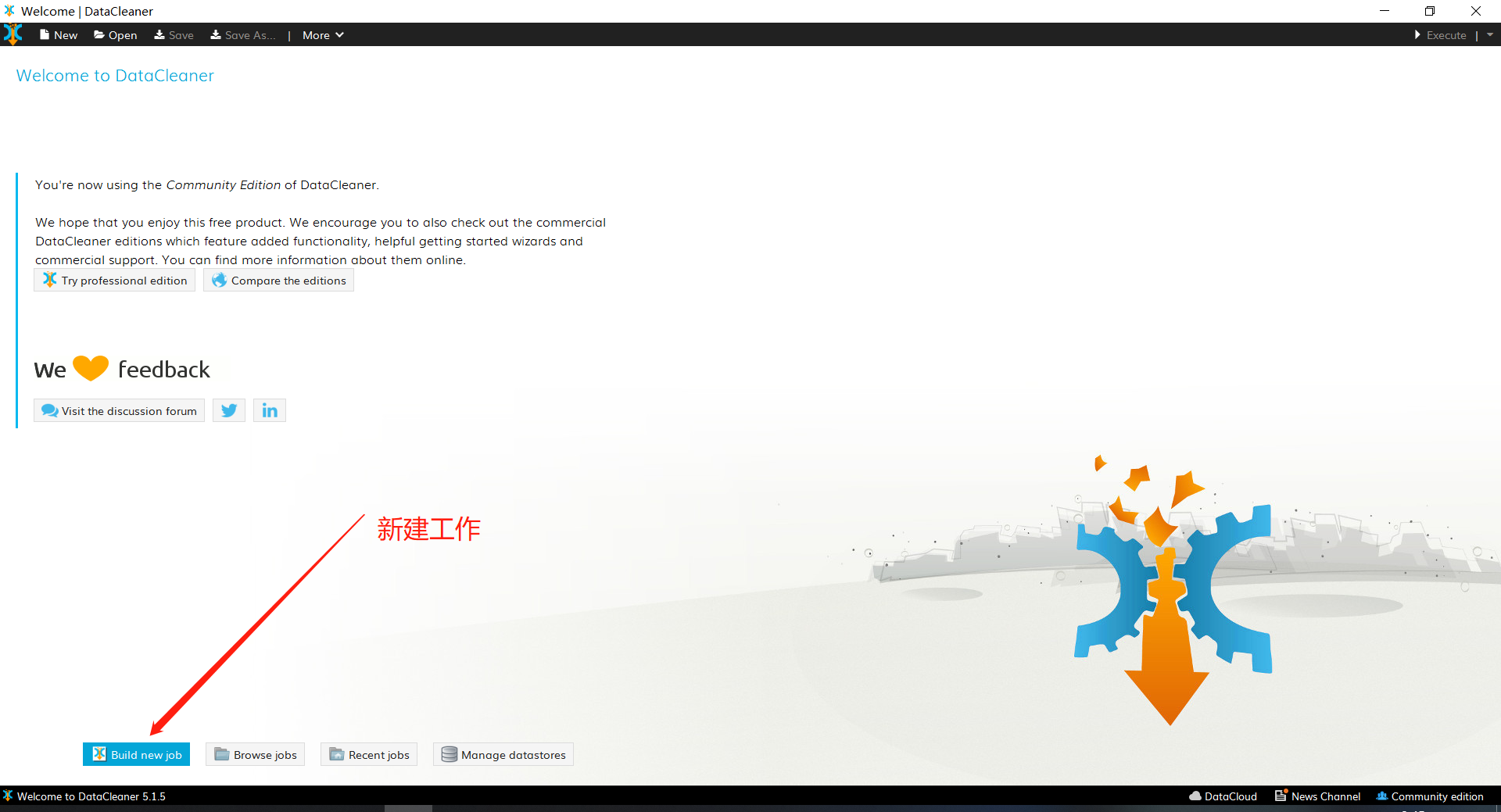
　- api DataCleaner监控模块的API类和接口

#### 安装环境要求

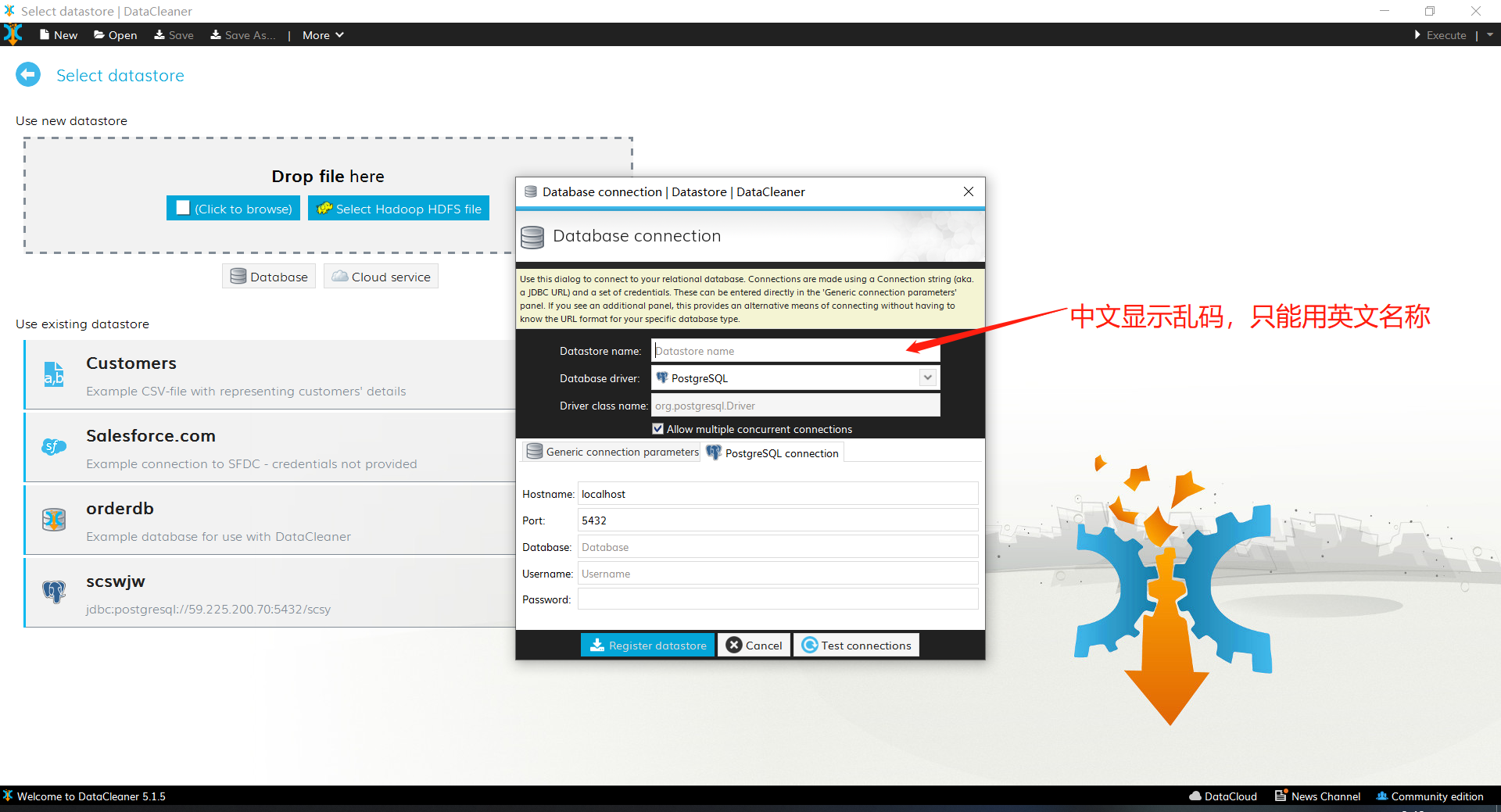
1. 一台有图形界面的计算机（命令行模式除外）
2. java7 或以上
3. DataCleaner 的软件许可文件

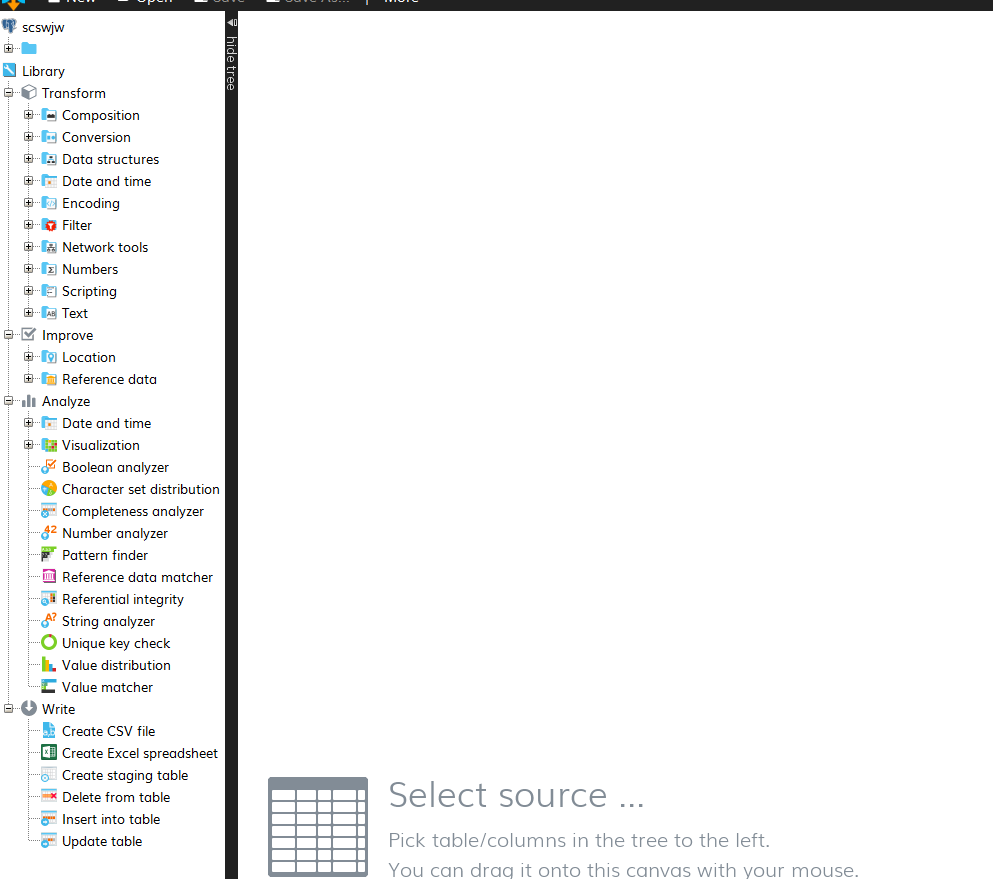
安装完成后，点击DataCleaner.exe,打开客户端











## 组件对应名称翻译

## 转换

组成

* [联盟](https://datacleaner.github.io/docs/5.4.0/components/union.html)
* [保险丝/合并领域](https://datacleaner.github.io/docs/5.4.0/components/fuse___coalesce_fields.html)
* [石斑鱼](https://datacleaner.github.io/docs/5.4.0/components/grouper.html)
* [调用子Analysis工作](https://datacleaner.github.io/docs/5.4.0/components/invoke_child_analysis_job.html)

### 转变

* [转换为布尔值](https://datacleaner.github.io/docs/5.4.0/components/convert_to_boolean.html)
* [转换为字符串](https://datacleaner.github.io/docs/5.4.0/components/convert_to_string.html)
* [转换为日期](https://datacleaner.github.io/docs/5.4.0/components/convert_to_date.html)
* [转换为数字](https://datacleaner.github.io/docs/5.4.0/components/convert_to_number.html)

### 数据结构

* [从地图中读取键和值](https://datacleaner.github.io/docs/5.4.0/components/read_keys_and_values_from_map.html)
* [从列表中选择值](https://datacleaner.github.io/docs/5.4.0/components/select_values_from_list.html)
* [构建列表](https://datacleaner.github.io/docs/5.4.0/components/build_list.html)
* [从键/值映射中选择值](https://datacleaner.github.io/docs/5.4.0/components/select_values_from_key_value_map.html)
* [构建键/值映射](https://datacleaner.github.io/docs/5.4.0/components/build_key_value_map.html)
* [编写和编写JSON文档](https://datacleaner.github.io/docs/5.4.0/components/compose_&_write_json_document.html)
* [从列表中读取元素](https://datacleaner.github.io/docs/5.4.0/components/read_elements_from_list.html)
* [阅读并解析JSON文档](https://datacleaner.github.io/docs/5.4.0/components/read_&_parse_json_document.html)

### 日期和时间

* [提取日期部分](https://datacleaner.github.io/docs/5.4.0/components/extract_date_part.html)
* [生成时间戳](https://datacleaner.github.io/docs/5.4.0/components/generate_timestamp.html)
* [时间戳转换器](https://datacleaner.github.io/docs/5.4.0/components/timestamp_converter.html)
* [日期差异/期间长度](https://datacleaner.github.io/docs/5.4.0/components/date_difference___period_length.html)
* [年龄日期](https://datacleaner.github.io/docs/5.4.0/components/date_to_age.html)
* [格式化日期](https://datacleaner.github.io/docs/5.4.0/components/format_date.html)
* [日期掩码匹配器](https://datacleaner.github.io/docs/5.4.0/components/date_mask_matcher.html)
* [捕获更改的记录](https://datacleaner.github.io/docs/5.4.0/components/capture_changed_records.html)

### 编码

* [XML编码器](https://datacleaner.github.io/docs/5.4.0/components/xml_encoder.html)
* [哈希值](https://datacleaner.github.io/docs/5.4.0/components/hash_value.html)
* [URL编码器](https://datacleaner.github.io/docs/5.4.0/components/url_encoder.html)
* [直译](https://datacleaner.github.io/docs/5.4.0/components/transliterate.html)
* [XML解码器](https://datacleaner.github.io/docs/5.4.0/components/xml_decoder.html)
* [HTML编码器](https://datacleaner.github.io/docs/5.4.0/components/html_encoder.html)

### 过滤

* [等于](https://datacleaner.github.io/docs/5.4.0/components/equals.html)
* [在字典中验证](https://datacleaner.github.io/docs/5.4.0/components/validate_in_dictionary.html)
* [相比](https://datacleaner.github.io/docs/5.4.0/components/compare.html)
* [字符串值范围](https://datacleaner.github.io/docs/5.4.0/components/string_value_range.html)
* [使用字符串模式进行验证](https://datacleaner.github.io/docs/5.4.0/components/validate_with_string_pattern.html)
* [空检查](https://datacleaner.github.io/docs/5.4.0/components/null_check.html)
* [最大行数](https://datacleaner.github.io/docs/5.4.0/components/max_rows.html)
* [数字范围](https://datacleaner.github.io/docs/5.4.0/components/number_range.html)
* [捕获更改的记录](https://datacleaner.github.io/docs/5.4.0/components/capture_changed_records.html)
* [字符串长度范围](https://datacleaner.github.io/docs/5.4.0/components/string_length_range.html)
* [日期范围](https://datacleaner.github.io/docs/5.4.0/components/date_range.html)

### 网络工具

* [将数字转换为IP](https://datacleaner.github.io/docs/5.4.0/components/convert_number_to_ip.html)
* [解析主机名](https://datacleaner.github.io/docs/5.4.0/components/resolve_hostname.html)
* [将IP转换为数字](https://datacleaner.github.io/docs/5.4.0/components/convert_ip_to_number.html)

### 数字

* [轮数](https://datacleaner.github.io/docs/5.4.0/components/round_number.html)
* [数学公式](https://datacleaner.github.io/docs/5.4.0/components/math_formula.html)
* [生成UUID](https://datacleaner.github.io/docs/5.4.0/components/generate_uuid.html)
* [生成ID](https://datacleaner.github.io/docs/5.4.0/components/generate_id.html)
* [增加数量](https://datacleaner.github.io/docs/5.4.0/components/increment_number.html)

### 脚本

* [Groovy变压器（高级）](https://datacleaner.github.io/docs/5.4.0/components/groovy_transformer_(advanced).html)
* [表达式语言（EL）变换器](https://datacleaner.github.io/docs/5.4.0/components/expression_language_(el)_transformer.html)
* [Groovy变压器（简单）](https://datacleaner.github.io/docs/5.4.0/components/groovy_transformer_(simple).html)
* [JavaScript变换器（简单）](https://datacleaner.github.io/docs/5.4.0/components/javascript_transformer_(simple).html)
* [JavaScript变换器（高级）](https://datacleaner.github.io/docs/5.4.0/components/javascript_transformer_(advanced).html)
* [JavaScript过滤器](https://datacleaner.github.io/docs/5.4.0/components/javascript_filter.html)

### 文本

* [URL解析器](https://datacleaner.github.io/docs/5.4.0/components/url_parser.html)
* [串接](https://datacleaner.github.io/docs/5.4.0/components/concatenator.html)
* [空白修剪器](https://datacleaner.github.io/docs/5.4.0/components/whitespace_trimmer.html)
* [简单搜索/替换](https://datacleaner.github.io/docs/5.4.0/components/plain_search_replace.html)
* [正则表达式解析器](https://datacleaner.github.io/docs/5.4.0/components/regex_parser.html)
* [删除子字符串](https://datacleaner.github.io/docs/5.4.0/components/remove_substring.html)
* [文本案例变换器](https://datacleaner.github.io/docs/5.4.0/components/text_case_transformer.html)
* [删除不需要的字符](https://datacleaner.github.io/docs/5.4.0/components/remove_unwanted_characters.html)
* [正则表达式搜索/替换](https://datacleaner.github.io/docs/5.4.0/components/regex_search_replace.html)
* [标记生成器](https://datacleaner.github.io/docs/5.4.0/components/tokenizer.html)

## 改善

### 本地

* [国家标准化者](https://datacleaner.github.io/docs/5.4.0/components/country_standardizer.html)

### 参考数据

* [字符串模式匹配器](https://datacleaner.github.io/docs/5.4.0/components/string_pattern_matcher.html)
* [删除字典匹配](https://datacleaner.github.io/docs/5.4.0/components/remove_dictionary_matches.html)
* [HTTP请求](https://datacleaner.github.io/docs/5.4.0/components/http_request.html)
* [字典匹配器](https://datacleaner.github.io/docs/5.4.0/components/dictionary_matcher.html)
* [表查找](https://datacleaner.github.io/docs/5.4.0/components/table_lookup.html)
* [同义词查找](https://datacleaner.github.io/docs/5.4.0/components/synonym_lookup.html)

## [分析](https://datacleaner.github.io/docs/5.4.0/components/#superCategoryPanel_Analyze)

### 日期和时间

* [月份分布](https://datacleaner.github.io/docs/5.4.0/components/month_distribution.html)
* [年份分布](https://datacleaner.github.io/docs/5.4.0/components/year_distribution.html)
* [平日分配](https://datacleaner.github.io/docs/5.4.0/components/weekday_distribution.html)
* [周数分布](https://datacleaner.github.io/docs/5.4.0/components/week_number_distribution.html)
* [日期/时间分析器](https://datacleaner.github.io/docs/5.4.0/components/date_time_analyzer.html)
* [日期差距分析仪](https://datacleaner.github.io/docs/5.4.0/components/date_gap_analyzer.html)

### 可视化

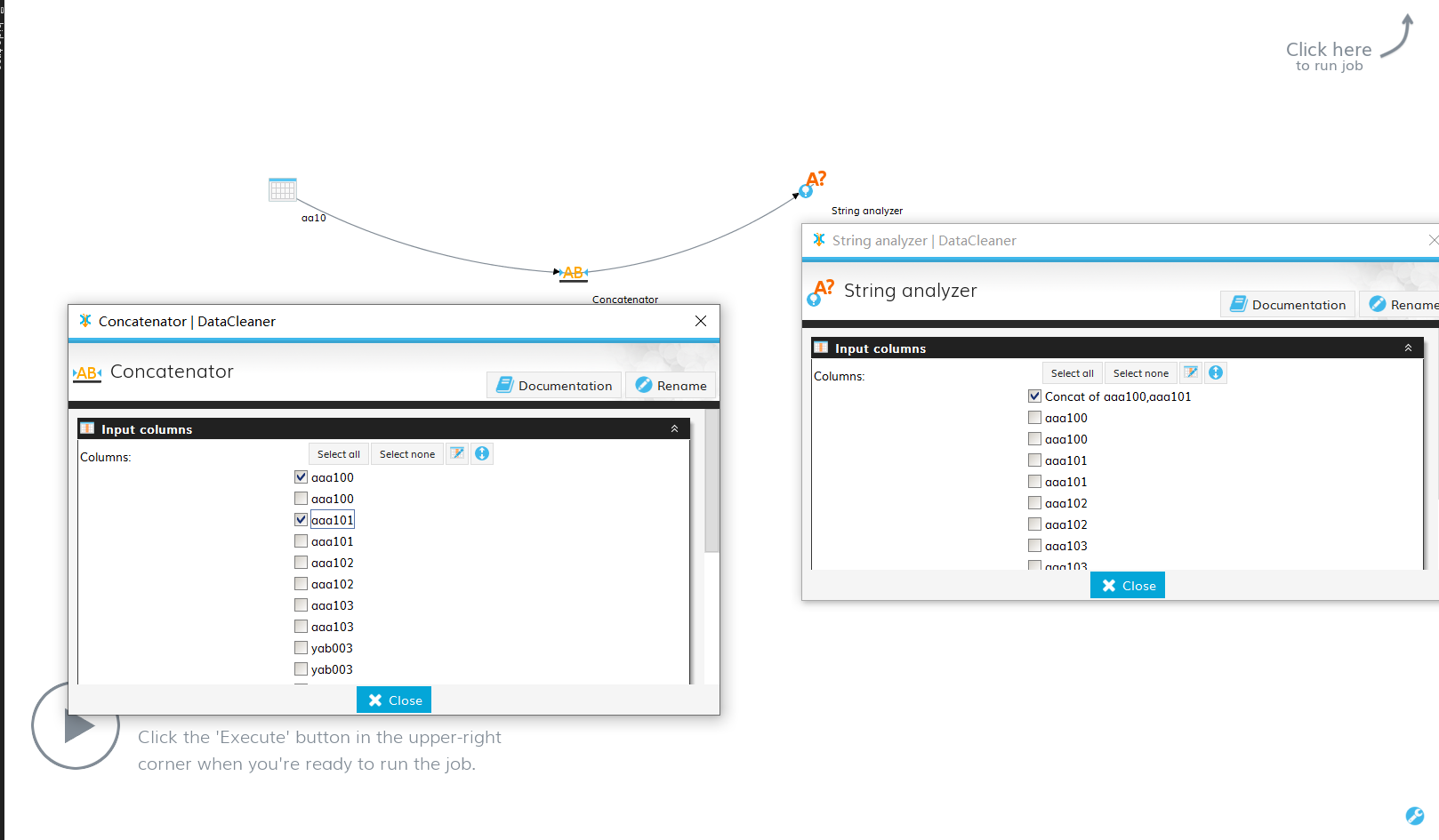
* [密度图](https://datacleaner.github.io/docs/5.4.0/components/density_plot.html)
* [堆积区域图](https://datacleaner.github.io/docs/5.4.0/components/stacked_area_plot.html)
* [散点图](https://datacleaner.github.io/docs/5.4.0/components/scatter_plot.html)
* [数字分析器](https://datacleaner.github.io/docs/5.4.0/components/number_analyzer.html)
* [填充图案](https://datacleaner.github.io/docs/5.4.0/components/fill_pattern.html)
* [独特的钥匙检查](https://datacleaner.github.io/docs/5.4.0/components/unique_key_check.html)
* [完整性分析仪](https://datacleaner.github.io/docs/5.4.0/components/completeness_analyzer.html)
* [参考数据匹配器](https://datacleaner.github.io/docs/5.4.0/components/reference_data_matcher.html)
* [价值匹配](https://datacleaner.github.io/docs/5.4.0/components/value_matcher.html)
* [价值分配](https://datacleaner.github.io/docs/5.4.0/components/value_distribution.html)
* [参照完整性](https://datacleaner.github.io/docs/5.4.0/components/referential_integrity.html)
* [图案查找器](https://datacleaner.github.io/docs/5.4.0/components/pattern_finder.html)
* [布尔分析器](https://datacleaner.github.io/docs/5.4.0/components/boolean_analyzer.html)
* [标记行](https://datacleaner.github.io/docs/5.4.0/components/mark_rows.html)
* [字符集分布](https://datacleaner.github.io/docs/5.4.0/components/character_set_distribution.html)
* [字符串分析器](https://datacleaner.github.io/docs/5.4.0/components/string_analyzer.html)

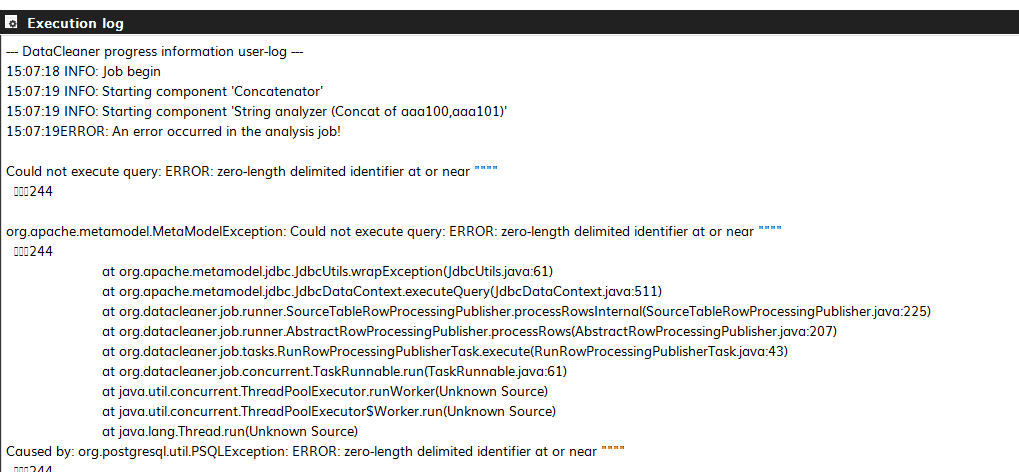
## [写](https://datacleaner.github.io/docs/5.4.0/components/#superCategoryPanel_Write)

* [创建CSV文件](https://datacleaner.github.io/docs/5.4.0/components/create_csv_file.html)
* [插入表格](https://datacleaner.github.io/docs/5.4.0/components/insert_into_table.html)
* [创建临时表](https://datacleaner.github.io/docs/5.4.0/components/create_staging_table.html)
* [从表中删除](https://datacleaner.github.io/docs/5.4.0/components/delete_from_table.html)
* [创建Excel电子表格](https://datacleaner.github.io/docs/5.4.0/components/create_excel_spreadsheet.html)
* [更新表](https://datacleaner.github.io/docs/5.4.0/components/update_table.html)

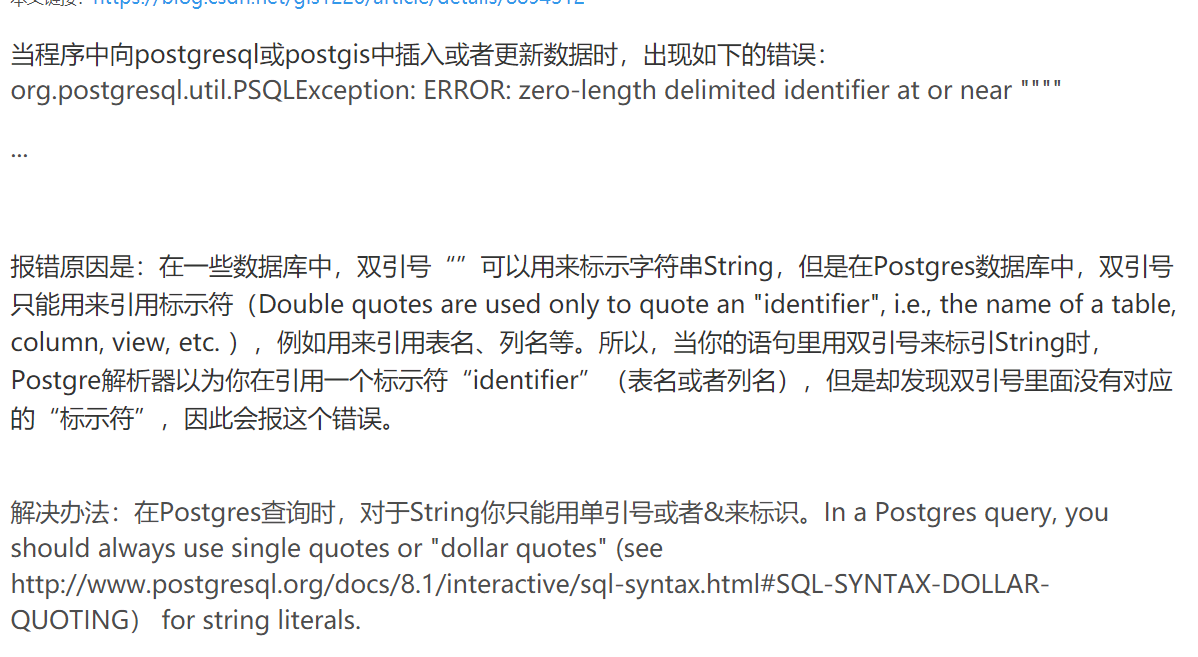
参考官方文档，配置好数据源，连接上省三医监控数据库

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

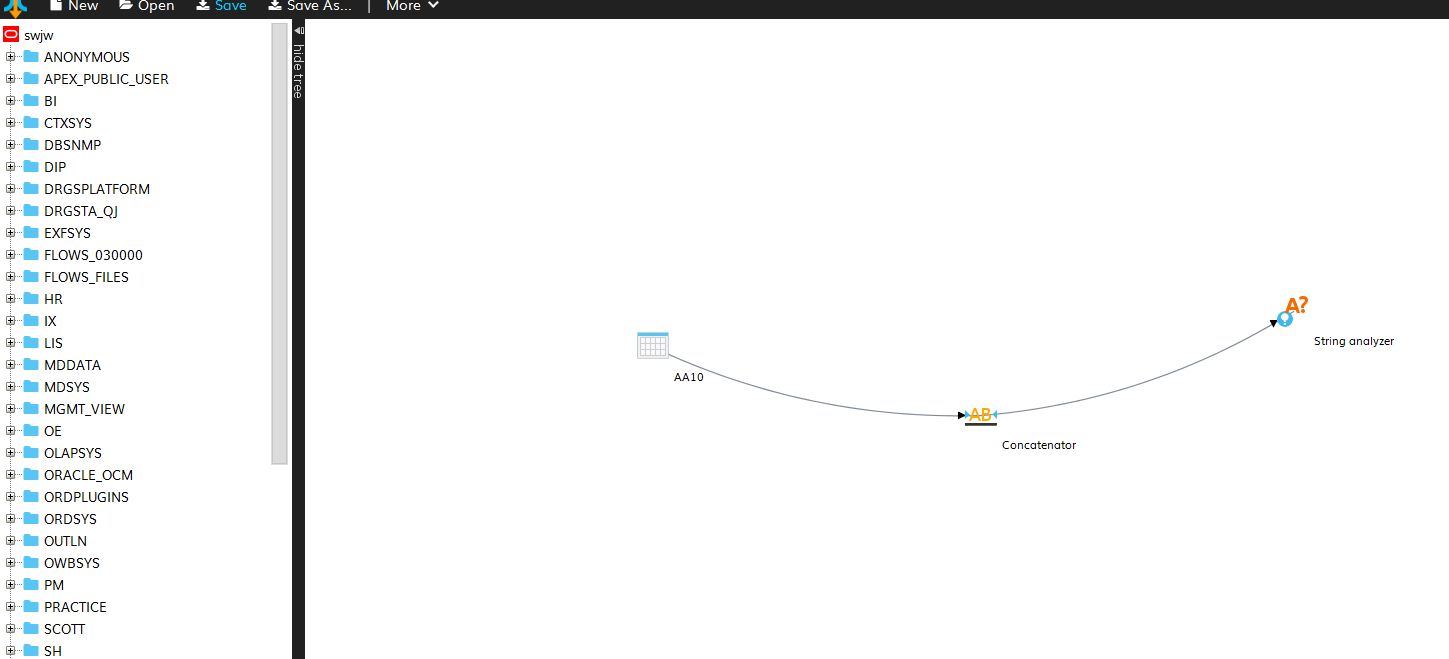


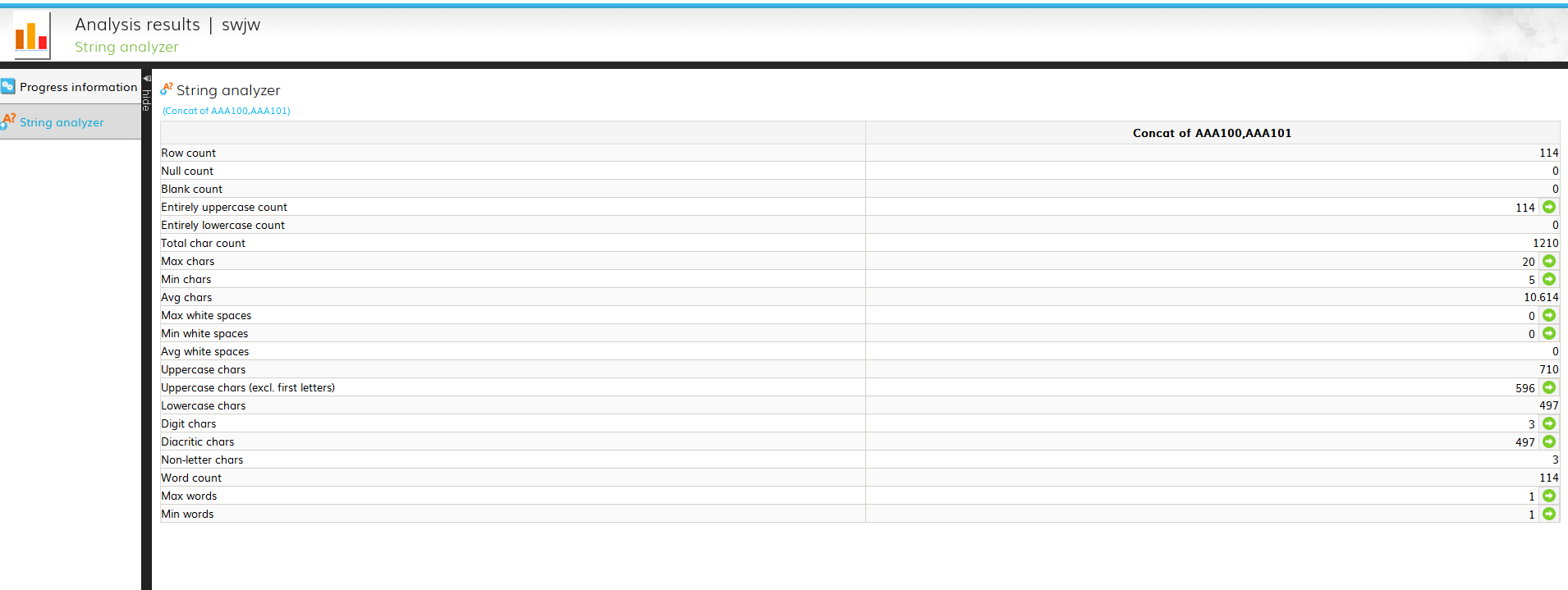


后经过多次不容分析任务测试，run job 时均会报这个错误，无法看到数据分析报告。解决思路如下



**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

则需要首先解决数据库的问题，才能做正常分析

学习过程中遇到的问题：

因网上相关资料较少，熟悉过程较为吃力。参考官方文档操作时，遇到报错问题解决较为困难。