江苏国泰新点软件有限公司

地址：江苏张家港市经济开发区**(http://www.epoint.com.cn)**

电话：0512-58188000传真：0512-58132373



**大数据清洗服务V3.0**

**部署手册**

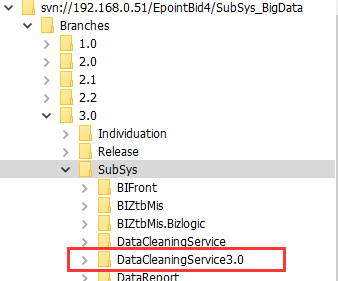
**版 本 历 史**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本/状态 | 作者 | 参与者 | 起止日期 | 备注 |
| 0.1 | 黄佳亮 |  | 2019.02.25 |  |
|  |  |  |  |  |
|  |  |  |  |  |

1. **代码更新**
2. **数据清洗服务工程代码**

位于大数据3.0SVN的SubSys中。

svn://192.168.0.51/EpointBid4/SubSys\_BigData/Branches/3.0/SubSys/DataCleaningService3.0



1. **大数据平台——后台及页面**

BIZtbMis.Bizlogic/com.epoint.ztbbi.tempdataview30.service.CleanStatus

BIZtbMis.Bizlogic/com.epoint.ztbbi.tempdataview30.service.TempDataViewListService

BIZtbMis/com.epoint.ztbbi.tempdataview30.action.TempDataViewListAction

BIZtbMis/src/main/webapp/datacleaning/tempdataview30/tempdataview\_list.html

1. **配置项说明**
2. **清洗服务——datacleaning.properties**
   1. RuleXMLDic（个性化必配）

规则XML的文件夹名，对应开发路径为

src/main/resources/RuleXML/Standard

用于项目个性化。

默认配置：Standard

* 1. AllowIPs（必配）

规定可以访问清洗服务的ip地址。多个以;隔开。

判断规则为contains

推荐配置：127.0.0.1，0:0:0:0:0:0:0:1，本机ip，大数据平台的服务器ip，维护工作人员使用的ip

* 1. CleaningJobCron（必配）

清洗作业的cron表达式。用于定时作业。

推荐配置：0 0 1 \* \* ?

* 1. ThreadCount

作业中的批量处理开启的线程数，在cleanimpl验证、批量设置异常状态位时使用。

默认配置：10

* 1. LimitSizeEveryTime

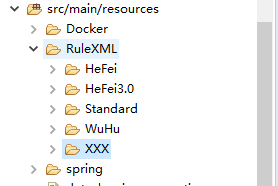
作业中的批量处理每天线程读取的记录数，在cleanimpl验证时使用。

默认配置：5000

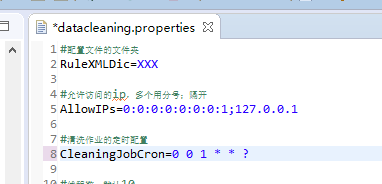
1. **大数据平台——ztb.properties**
   1. CleanRESTfulAPI

数据清洗服务的地址。用于数据治理页面。

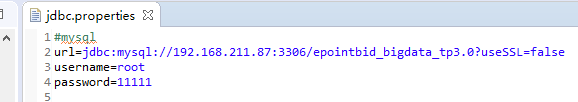
1. **部署步骤**
2. **在resources/RuleXML/下复制Standard为个性化项目，例如XXX**



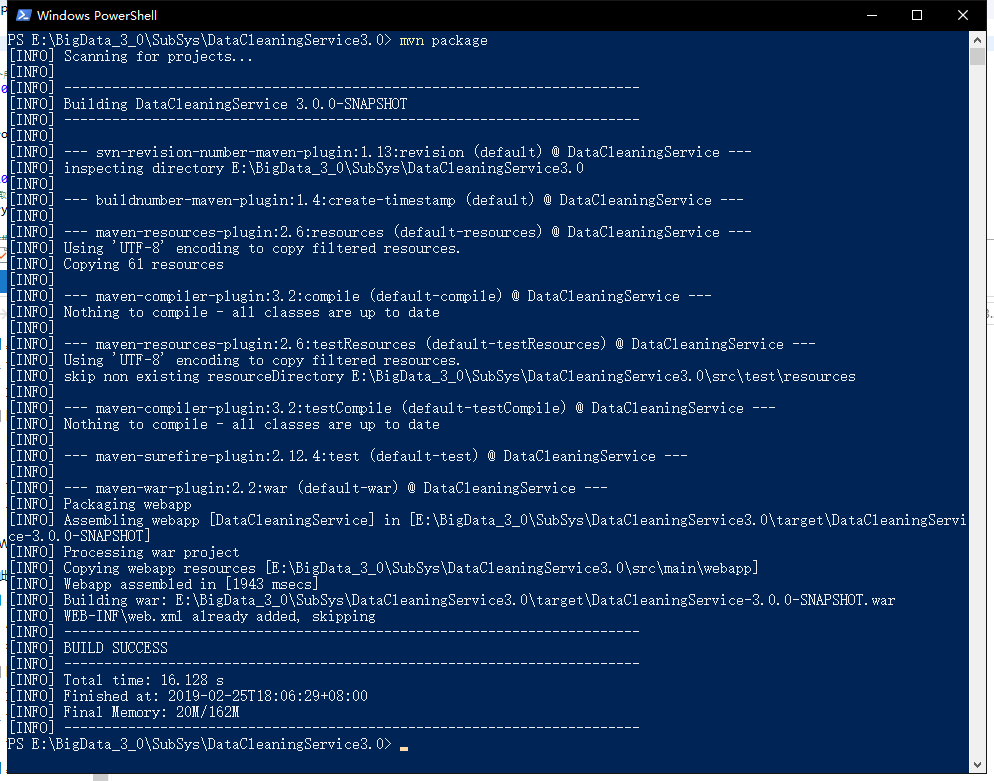
1. **设置必配项**



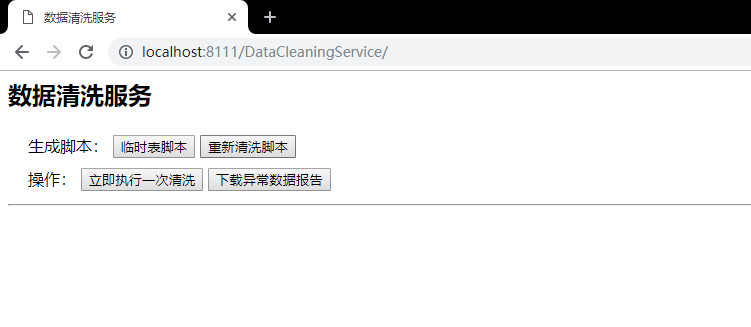
1. **修改数据库连接**



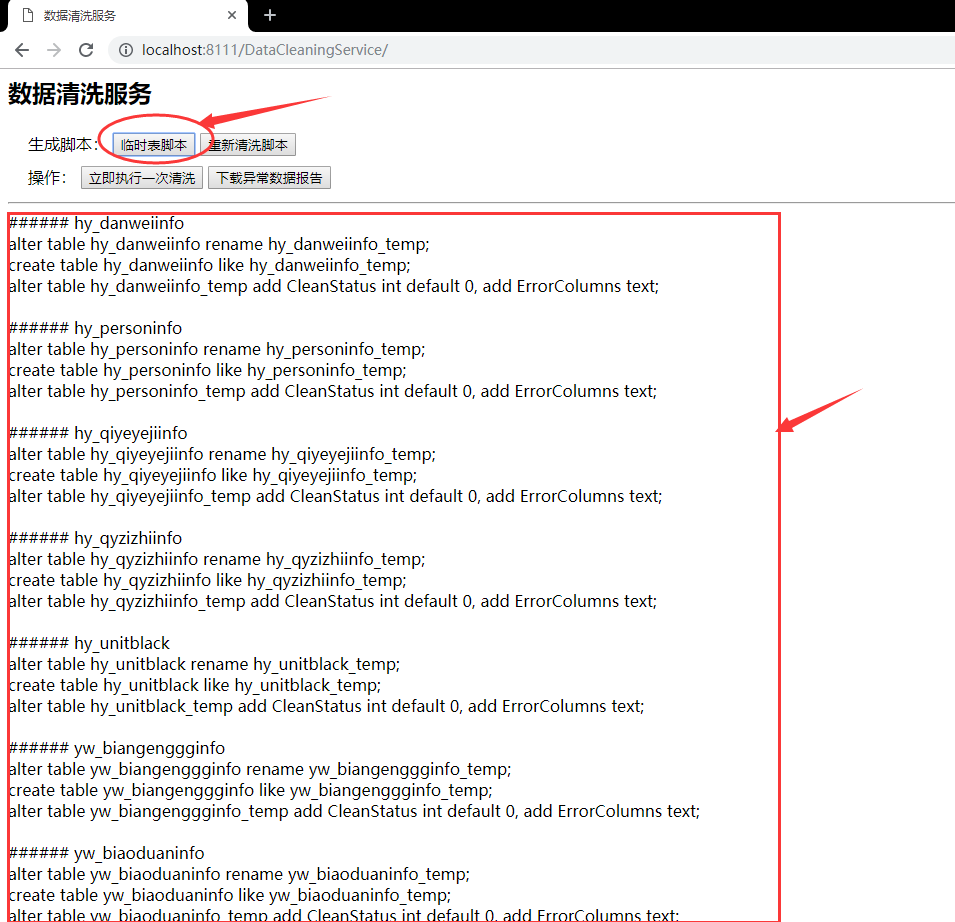
1. **maven打包**



1. **部署并启动服务**



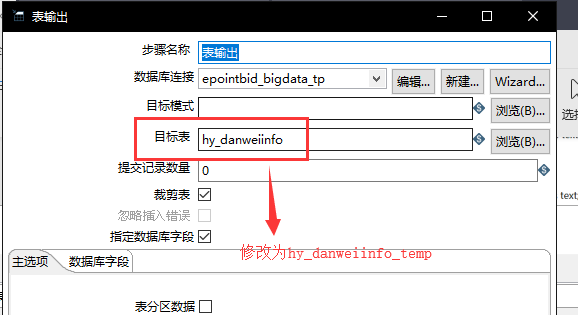
1. **点击临时表脚本按钮，复制SQL并放到数据库中执行**



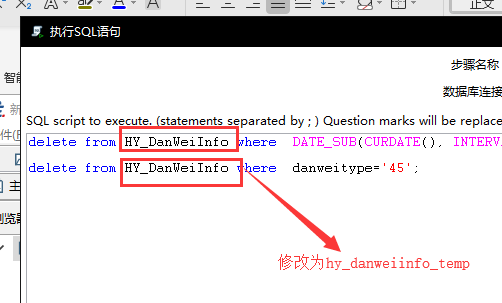
1. **修改所有需要清洗基础表的全量及增量抽取**

（以hy\_danweiinfo表为例）

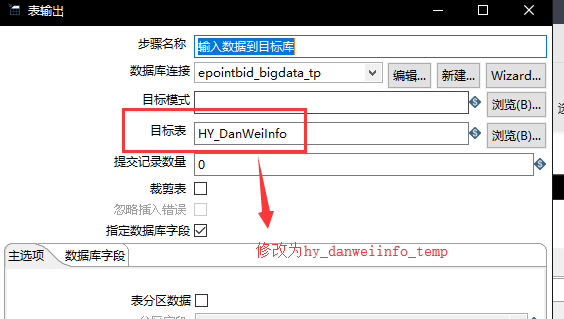
* 1. 修改全量中的“表输出”目标表，加上\_temp



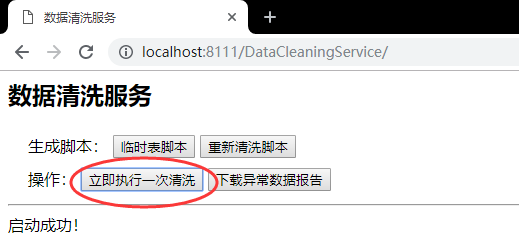
* 1. 修改增量中的“删除目标库指定时间内的数据”SQL

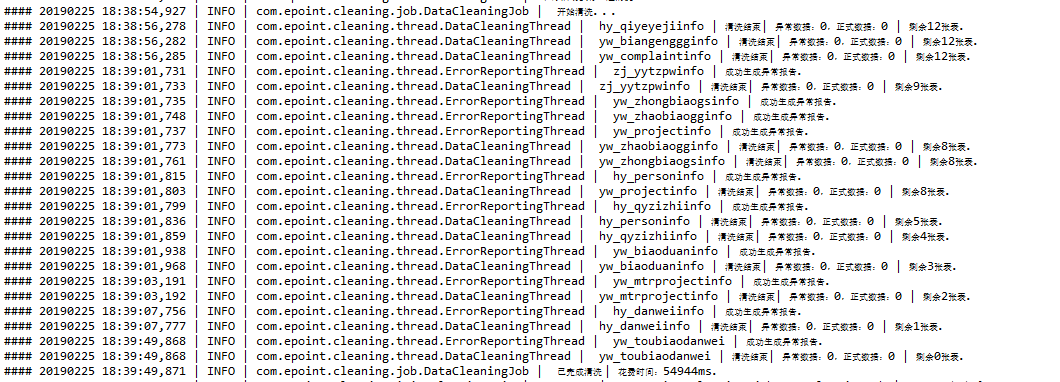


* 1. 修改增量中的“输入数据到目标表”目标表，加上\_temp



1. **点击立即执行一次清洗，并查看日志**



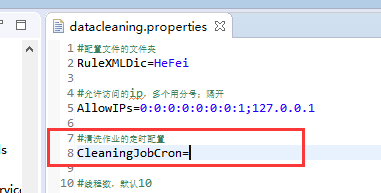


1. **下载异常报告**

http://localhost:8091/EpointBI/datacleaning/tempdataview30/tempdataview\_list

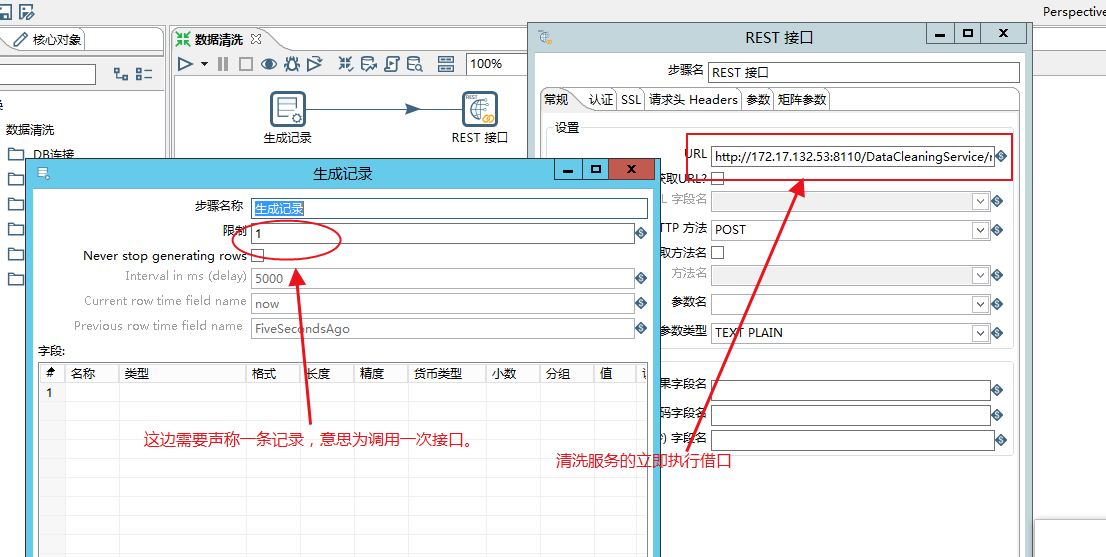


1. **加入EDI总作业**
2. **去除清洗服务的作业配置**



1. **EDI中新建转换并设置REST接口**

URL：http://localhost:8111/DataCleaningService/rest/cleanimmediately



1. **加入总作业**

放到临时表抽取后面，统计表之前。

