1. **作业**

DataCleaningJob: 数据清洗作业

**二、配置文件**

datacleaning.properties

RuleXMLDic:XML配置文件的文件夹名，默认为RuleXML

AllowIPs:允许访问此服务的IP地址

ThreadCount:线程数，默认为20

LimitSizeEveryTime:每次循环清洗的数据条数，默认为2000

**三、XML配置文件语法**

xml文件的文件名为正式表的表名。

标签：

<min>：最小值 —— sql清洗

included属性：值为false时不包含(>)，默认为true(>=)

dataformat属性：日期格式化

<max>：最大值 —— sql清洗

included属性：值为false时不包含(<)，默认为true(<=)

dateformat属性：日期格式化

round：保留小数位数

<required>：必填 —— sql清洗

<code>：代码项 —— sql清洗 多个用 逗号, 隔开

allowtext属性：允许文本内容，默认false

<in>：取值范围 —— sql清洗 多个用 逗号, 隔开

<notin>：<in>的反义 —— sql清洗 多个用 逗号, 隔开

<regex>：正则 —— sql清洗 符合正则的

<unregex>：正则的反义 —— sql清洗 不符合正则的

<where>：where语句判断 —— sql清洗 内容最好用 <![CDATA[ ]]> 包住

<cleanimpl>：[自定义逻辑，配置@Component的值，继承接口com.epoint.cleaning.iface.ICleaning，](mailto:自定义逻辑，配置@Component的值，继承接口com.epoint.cleaning.iface.ICleaning，)

params属性：额外的字段，多个参数以 , 隔开

函数：

config:参数名

取datacleaning.properties文件中的配置内容

**四、问题及说明**

1. 存在窗口期问题，增量抽取时可能会删除已经清理过的数据。
2. 因为大数据基础表中rowguid值一般都为空，所以该服务会给临时表填充rowguid值，用于区分数据。

**五、临时表清洗状态说明**

0：未清洗 —— 数据抽取过来后的初始值。

1：正在清洗 —— 清洗服务开始拿数据时，将所有未清洗的先置为预清洗。用于解决幻读风险。

2：异常数据 —— 被清洗服务判断为问题数据，需要修正。

3：正式数据 —— 通过清洗服务判定，并存入了正式表。

4：修正数据 —— 在修改页面存入正式表的。