应用系统负载分析与磁盘容量预测

1 项目背景

某大型企业为了信息化发展的需要，建设了办公自动化系统、人力资源管理系统、财务管理系统、企业信息门户系统等几大企业级应用系统。因应用系统在日常运行时，会对底层软硬件造成负荷。显著影响应用系统性能的因素包括：服务器、数据库、中间件、存储设备。任何一种资源负载过大，都可能会引起应用系统性能下降甚至瘫痪。因此需要关注服务器、数据库、中间件、存储设备的运行状态，及时了解当前应用系统的负载情况，以便提前预防，确保系统安全稳定运行。

应用系统的负载率可以通过对一段时间内软硬件性能的运行状况进行综合评分而获得。通过系统的当前负载率与历史平均负载率进行比较，获得负载率的当前趋势。通过负载率以及负载趋势可对系统进行负载分析，当出现应用系统的负载高或者负载趋势大的现象，代表系统目前处于高危工作环境中。如果系统管理员不及时进行相应的处理，系统很容易出现故障，从而导致用户无法访问系统，严重影响企业的利益。本章重点分析存储设备中磁盘容量预测，通过对磁盘容量进行预测，可预测磁盘未来的负载情况。避免应用系统出现存储容量耗尽的情况，从而导致应用系统负载率过高，最终引发系统故障。

2 项目目标

（1） 针对历史磁盘数据，采用时间序列分析方法，预测应用系统服务器磁盘已使用空间大小。

（2） 根据用户需求设置不同的预警等级，将预测值与容量值进行比较，对其结果进行预警判断，为系统管理员提供定制化的预警提示。

3 项目步骤

3.1 工程前期准备

3.1.1 导入数据

（1） 介绍数据

目前监控采集的性能数据主要包含 CPU 使用信息，内存使用信息，磁盘使用信息等，如表 3-1 所示。通过分析磁盘容量相关数据（见表 3-2），预测应用系统服务器磁盘空间是否满足系统健康运行的要求。

表 3-1 性能说明表

属性名称	属性说明	属性名称	属性说明
SYS_NAME	资产所在的系统名称	ENTITY	具体的属性
NAME	资产名称	VALUE	采集到的值
TARGET_ID	属性的标识号 183 表示磁盘容量大小 184 表示磁盘已使用大小	COLLECTTIME	采集的时间
DESCRIPTION	针对属性标识的说明		

表 3-2 磁盘原始数据集

SYS_NAME	NAME	TARGET_ID	DESCRIPTION	ENTITY	VALUE	COLLECTTIME
财务管理系统	CWXT_DB	184	磁盘已使用大小	C:\	34270787.33	2014/10/1
财务管理系统	CWXT_DB	184	磁盘已使用大小	D:\	80262592.65	2014/10/1
财务管理系统	CWXT_DB	183	磁盘容量	C:\	52323324	2014/10/1
财务管理系统	CWXT_DB	183	磁盘容量	D:\	157283328	2014/10/1
财务管理系统	CWXT_DB	184	磁盘已使用大小	C:\	34328899.02	2014/10/2
财务管理系统	CWXT_DB	184	磁盘已使用大小	D:\	83200151.65	2014/10/2
财务管理系统	CWXT_DB	183	磁盘容量	C:\	52323324	2014/10/2
财务管理系统	CWXT_DB	183	磁盘容量	D:\	157283328	2014/10/2
财务管理系统	CWXT_DB	184	磁盘已使用大小	C:\	34327553.5	2014/10/3
财务管理系统	CWXT_DB	184	磁盘已使用大小	D:\	83208320	2014/10/3
财务管理系统	CWXT_DB	183	磁盘容量	C:\	52323324	2014/10/3
财务管理系统	CWXT_DB	183	磁盘容量	D:\	157283328	2014/10/3
财务管理系统	CWXT_DB	184	磁盘已使用大小	C:\	34288672.21	2014/10/4
财务管理系统	CWXT_DB	184	磁盘已使用大小	D:\	83099271.65	2014/10/4
财务管理系统	CWXT_DB	183	磁盘容量	C:\	52323324	2014/10/4
财务管理系统	CWXT_DB	183	磁盘容量	D:\	157283328	2014/10/4
财务管理系统	CWXT_DB	184	磁盘已使用大小	C:\	34190978.41	2014/10/5
财务管理系统	CWXT_DB	184	磁盘已使用大小	D:\	82765171.65	2014/10/5
财务管理系统	CWXT_DB	183	磁盘容量	C:\	52323324	2014/10/5
财务管理系统	CWXT_DB	183	磁盘容量	D:\	157283328	2014/10/5
财务管理系统	CWXT_DB	184	磁盘已使用大小	C:\	34187614.43	2014/10/6
财务管理系统	CWXT_DB	184	磁盘已使用大小	D:\	82522895	2014/10/6
财务管理系统	CWXT_DB	183	磁盘容量	C:\	52323324	2014/10/6
财务管理系统	CWXT_DB	183	磁盘容量	D:\	157283328	2014/10/6

（2） 上传数据到 Python 数据挖掘建模平台

在新增数据源上，选择本地上传数据，如图 1 所示。

图 1 本地上传数据源

在本地路径上选择文件，填写在平台新建的目标表名，如图 2 所示。

图 2 本地选择文件上传

根据文件的数据，可以修改文件的字段名和类型，如图 3 所示。

图 3 字段设置

上传成功，可以在平台的数据源上查看数据，单击数据源操作的查看按钮如图 4 所示，数据预览如图 5 所示。

图 4 单击预览数据按钮

图 5 数据预览

3.1.2 新建空白工程

右击我的工程，新建一个空白的工程，如图 6 所示。

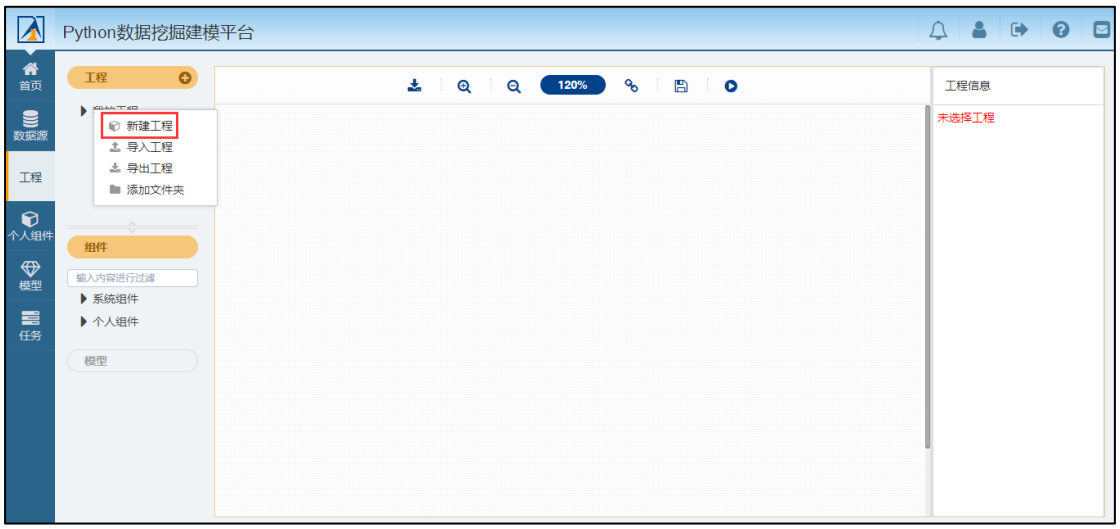

图 6 新建工程

填写工程的信息，包括工程名称和工程描述，如图 7 所示。

图 7 填写工程信息

3.2 数据预处理

读取 discdata 数据，步骤如图 8 所示。

(1) 选择工程。

(2) 选择输入源组件。

(3) 拖入输入源组件。

(4) 填写数据表名。

(5) 单击更新按钮，更新出数据。

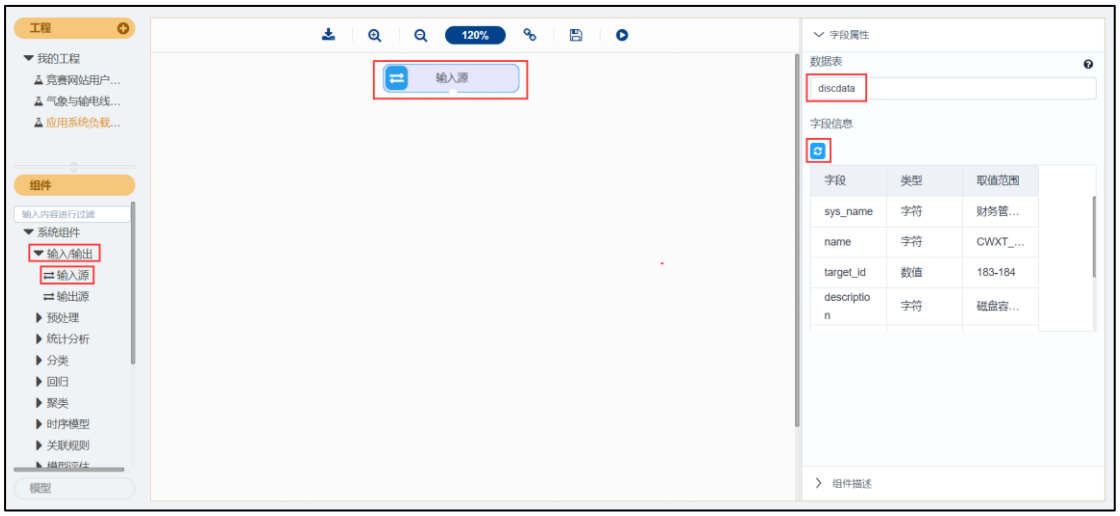

图 8 输入源组件

3.2.1 数据筛选

选择数据筛选，步骤如图 9、图 10 所示。

(1) 找到预处理→数据筛选。

(2) 拖入数据筛选组件，将输入源和数据筛选组件连接。

(3) 单击更新按钮，勾选全部字段作为输出字段。

(4) 选择参数设置，点击添加项，然后点击刷新，字段选择 target_id，函数选择=，值填入 184。

(5) 对数据筛选组件右键，选择运行该节点。

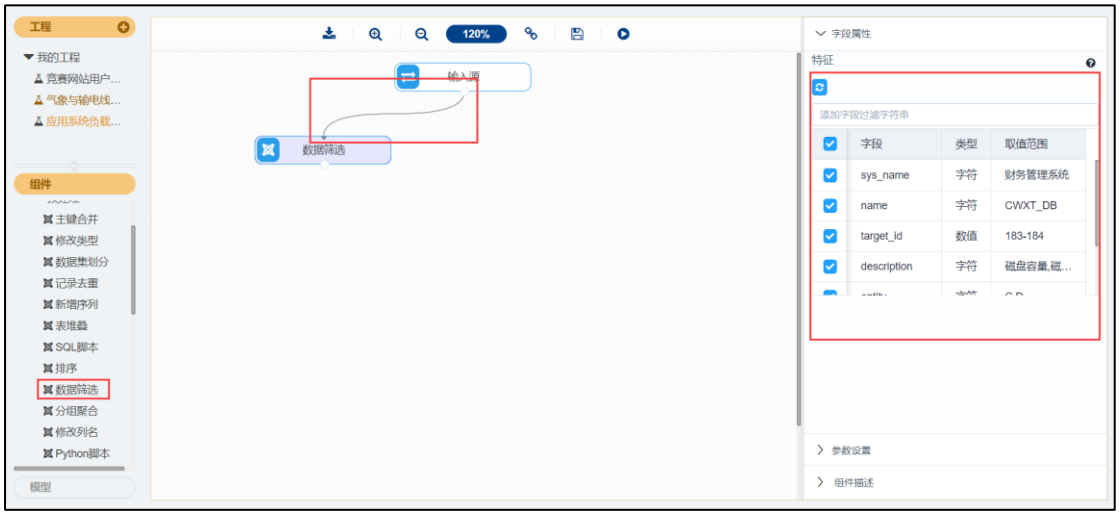

图 9 数据筛选组件_字段属性

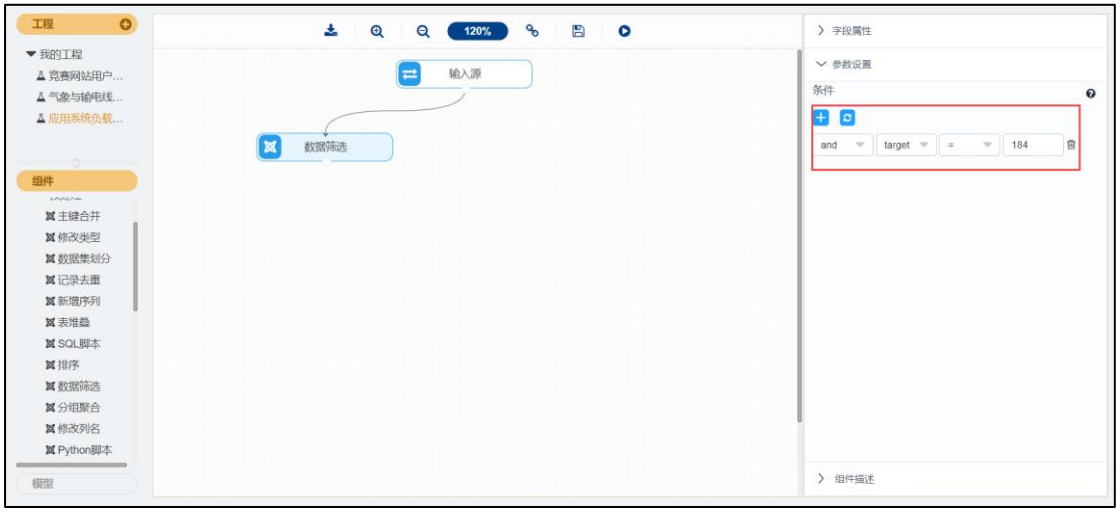

图 10 数据筛选组件_参数设置

(6) 运行完成后，对数据筛选组件右键，选择查看数据，数据筛选的输出表结果如图 11

所示。

sys_name	name	target_id	description	entity	value	collecttime
财务管理系统	CWXT_DB	184	磁盘已使用大小	C	34270787.33	2014-10-01
财务管理系统	CWXT_DB	184	磁盘已使用大小	D	80262592.65	2014-10-01
财务管理系统	CWXT_DB	184	磁盘已使用大小	C	34328899.02	2014-10-02
财务管理系统	CWXT_DB	184	磁盘已使用大小	D	83200151.65	2014-10-02
财务管理系统	CWXT_DB	184	磁盘已使用大小	C	34327553.5	2014-10-03
财务管理系统	CWXT_DB	184	磁盘已使用大小	D	83208320	2014-10-03
财务管理系统	CWXT_DB	184	磁盘已使用大小	C	34288672.21	2014-10-04
财务管理系统	CWXT_DB	184	磁盘已使用大小	D	83099271.65	2014-10-04

图 11 数据筛选结果

3.2.2 属性变换

接下来进行属性变换，步骤如图 12 所示。

(1) 找到预处理→Python 脚本组件。

(2) 拖入 Python 脚本组件，并将数据筛选和 Python 脚本组件连接。

(3) 选择字段属性，在脚本处填入数据变换代码，如表 3-3 所示。

(4) 对 Python 脚本组件右键，选择运行该节点。

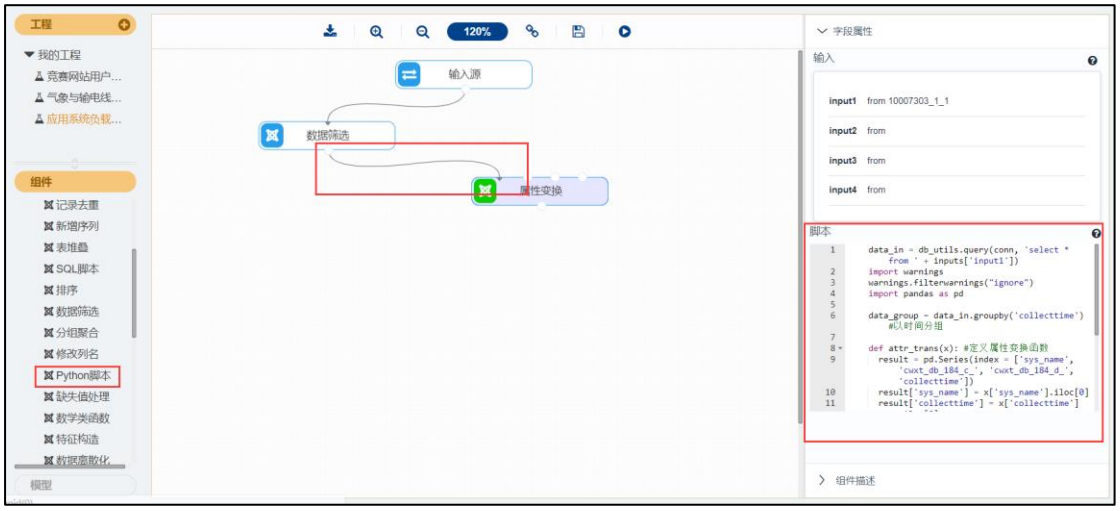

图 12 属性变换组件

表 3-3 属性变换代码

```
data_in = db_utils.query(conn, 'select * from ' + inputs['input1'])
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
```

```
data_group = data_in.groupby('collecttime') #以时间分组

def attr_trans(x): #定义属性变换函数
    result = pd.Series(index = ['sys_name', 'cwxt_db_184_c_', 'cwxt_db_184_d_',      'collecttime'])
    result['sys_name'] = x['sys_name'].iloc[0]
    result['collecttime'] = x['collecttime'].iloc[0]
    result['cwxt_db_184_c_'] = x['value'].iloc[1]
    result['cwxt_db_184_d_'] = x['value'].iloc[0]
    return result

data_processed = data_group.apply(attr_trans) #逐组处理

data_out = pd.DataFrame(data_processed)

return(data_out)
```

(5) 运行完成后，对 Python 脚本组件右键，选择查看数据，如图 13 所示。

预览数据

sys_name	cwxt_db_184_c_	cwxt_db_184_d_	collecttime
财务管理系统	80262592.65	34270787.33	2014-10-01
财务管理系统	83200151.65	34328899.02	2014-10-02
财务管理系统	83208320	34327553.5	2014-10-03
财务管理系统	83099271.65	34288672.21	2014-10-04
财务管理系统	82765171.65	34190978.41	2014-10-05
财务管理系统	82522895	34187614.43	2014-10-06
财务管理系统	82590885	34285280.22	2014-10-07
财务管理系统	82368173.3	34290578.41	2014-10-08

共 47 条 25 条/页 1 2 前往 1 页

图 13 属性变换结果

(6) 运行完成后，对 Python 脚本组件右键，重命名为属性变换。

3.2.3 生成训练数据

选择数据筛选，步骤如图 14、图 15 所示。

(1) 找到预处理→数据筛选。

(2) 拖入数据筛选组件，将属性变换和数据筛选组件连接。

(3) 单击更新按钮，勾选全部字段作为输出字段。

(4) 选择参数设置，点击添加项，然后点击刷新，字段选择 collecttime，函数选择<，值填入 2014-11-12。

(5) 对数据筛选组件右键，选择运行该节点。

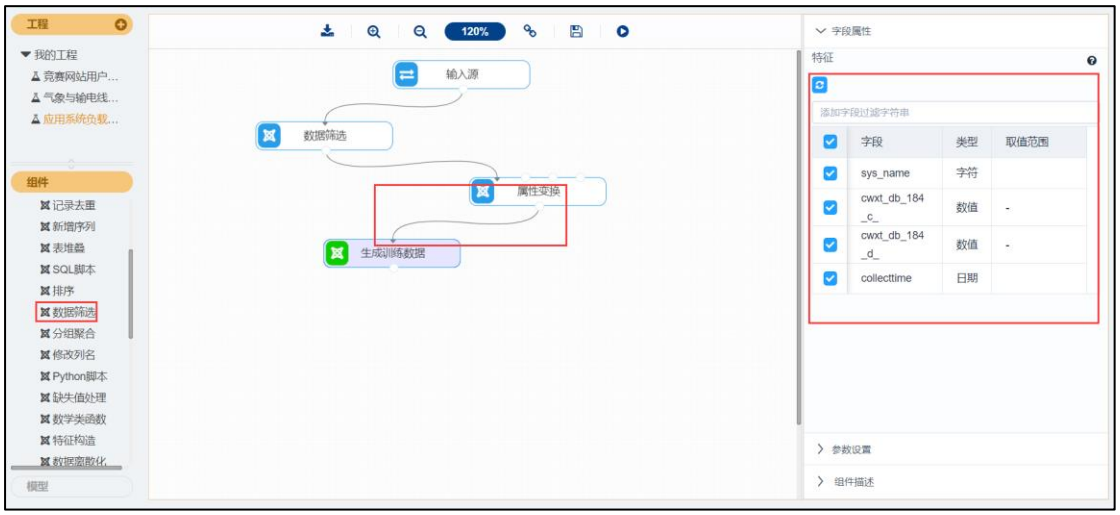

图 14 生成训练数据组件_字段属性

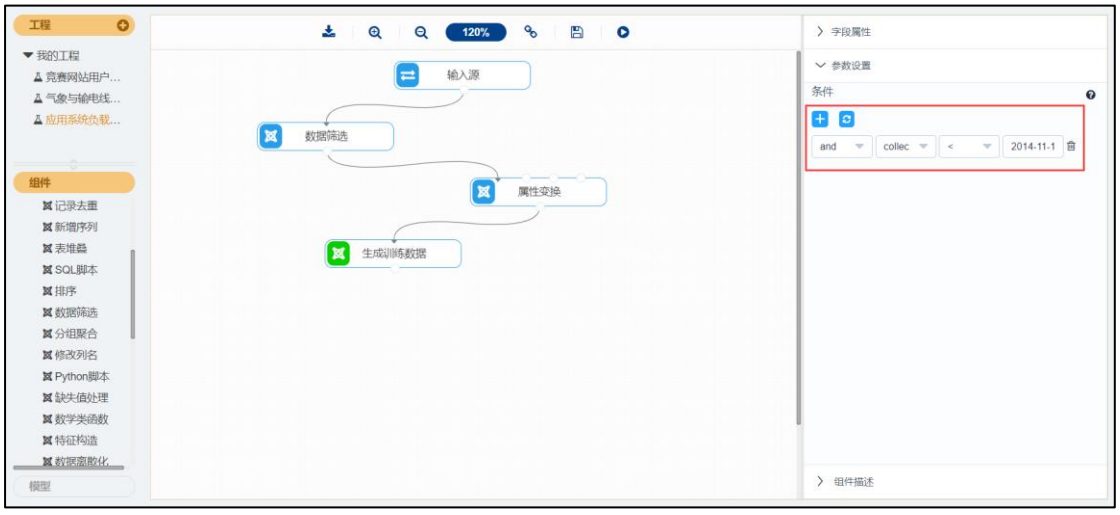

图 15 生成训练数据组件_参数设置

(6) 运行完成后，对数据筛选组件右键，选择查看数据，数据筛选的输出表结果如图 16 所示。

预览数据

sys_name	cwxt_db_184_c_	cwxt_db_184_d_	collecttime
财务管理系统	80262592.65	34270787.33	2014-10-01
财务管理系统	83200151.65	34328899.02	2014-10-02
财务管理系统	83208320	34327553.5	2014-10-03
财务管理系统	83099271.65	34288672.21	2014-10-04
财务管理系统	82765171.65	34190978.41	2014-10-05
财务管理系统	82522895	34187614.43	2014-10-06
财务管理系统	82590885	34285280.22	2014-10-07
财务管理系统	82368173.3	34290578.41	2014-10-08

共 42 条 25 条/页 1 2 前往 1 页

图 16 生成训练数据结果

(7) 运行完成后，对数据筛选组件右键，重命名为生成训练数据。

3.2.4 平稳性检验

选择平稳性检验，步骤如图 17 所示。

(1) 找到统计分析→平稳性检验。

(2) 拖入平稳性检验组件，将生成训练数据和平稳性检验组件连接。

(3) 单击更新按钮，时序特征勾选 cwxt_db_184_d_字段作为检验字段。

(4) 对平稳性检验组件右键，选择运行该节点。

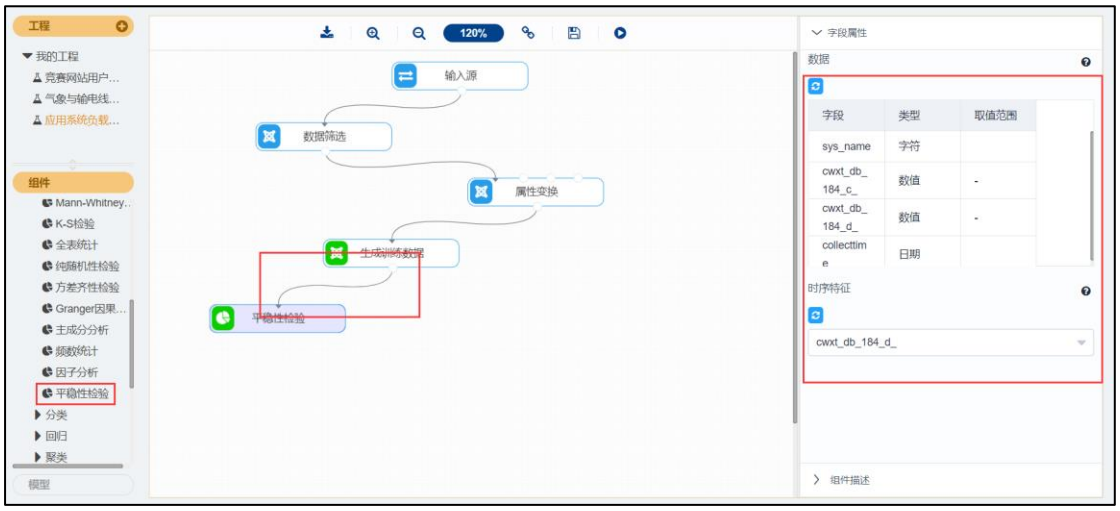

图 17 平稳性检验组件

(5) 运行完成后，对平稳性检验组件右键，选择查看报告，平稳性检验的报告结果如图 18 所示。

图 18 平稳性检验报告

3.2.5 纯随机性检验

选择平稳性检验，步骤如图 19 所示。

(1) 找到统计分析→纯随机性检验。

(2) 拖入平稳性检验组件，将生成训练数据和纯随机性检验组件连接。

(3) 单击更新按钮，特征勾选 cwxt_db_184_d_字段作为检验字段。

(4) 对纯随机性检验组件右键，选择运行该节点。

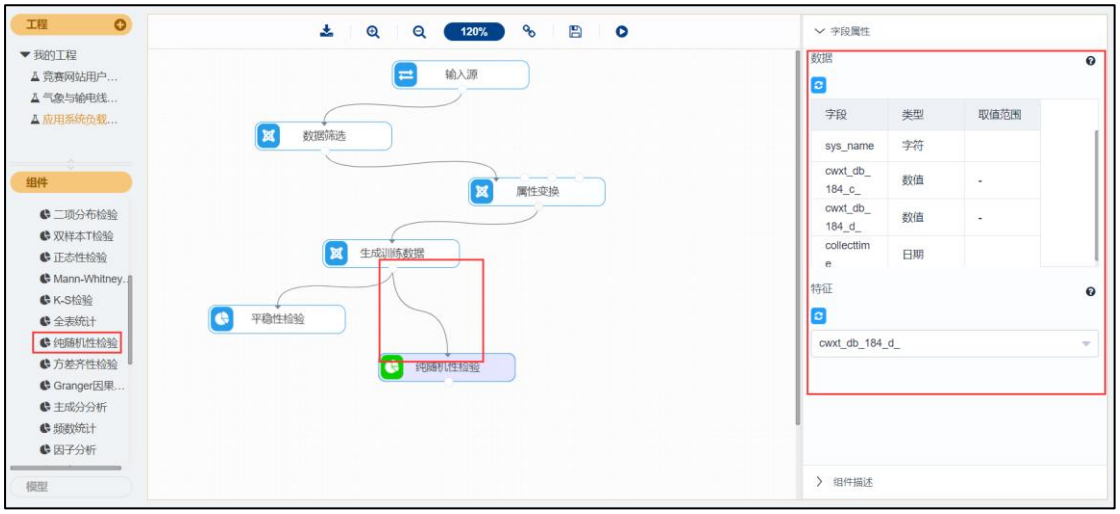

图 19 纯随机性检验组件

(5) 运行完成后，对纯随机性检验组件右键，选择查看报告，纯随机性检验的报告结果如图 20 所示。

图 20 纯随机性检验报告

3.3 模型构建

3.3.1 ARIMA 算法

选择 ARIMA 算法模型，步骤如图 21、图 22 所示。

(1) 找到时序模型→ARIMA 组件。

(2) 拖入 ARIMA 组件，将生成训练数据和 ARIMA 组件连接。

(3) 选择字段属性，单击更新数据，时序列勾选 cwxt_db_184_d_字段，时间列勾选 collecttime 字段。

(4) 选择参数设置，设置预测周期数的值为 5，设置自回归项数 p 的值为 0，设置差分次数 d 的值为 1，设置移动平均项数 q 的值为 3。

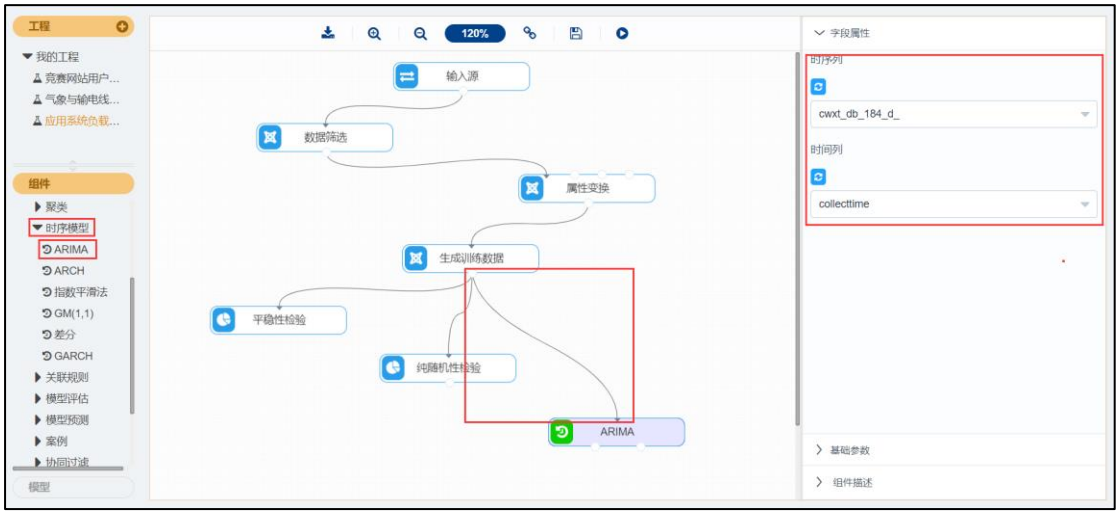

图 21 ARIMA 组件_字段属性

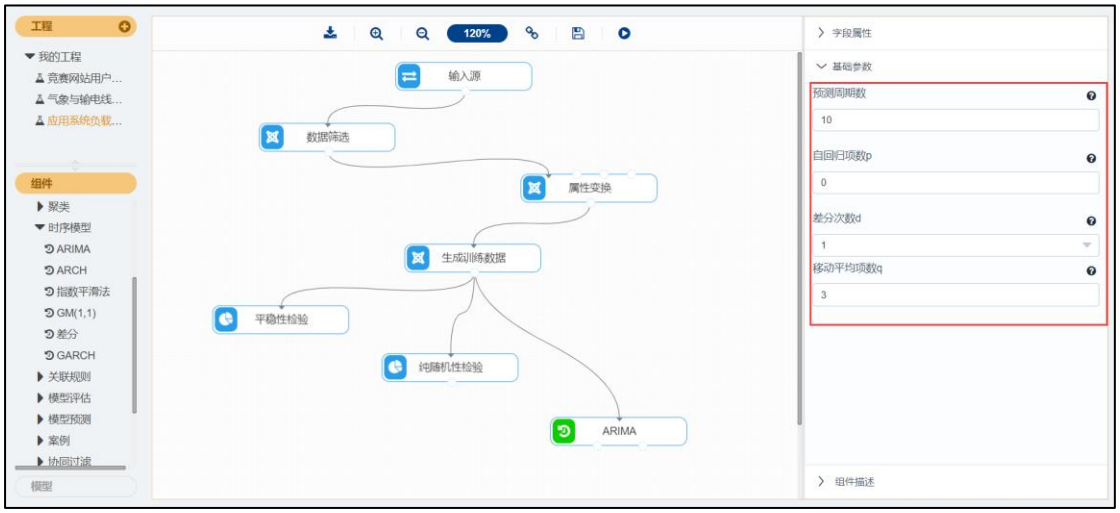

图 22 ARIMA 组件_参数设置

(5) 运行完成后，对 ARIMA 组件右键，选择查看报告，ARIMA 的报告如图 23 所示。

图 23 ARIMA 的报告