

# 腾讯云数据仓库 TCHouse-P

## 数仓开发



腾讯云

---

**【 版权声明 】**

©2013–2024 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分內容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

**【 商标声明 】**

及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

**【 服务声明 】**

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。

您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

**【 联系我们 】**

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或95716。

---

## 文档目录

数仓开发

云上搭建 Airflow

# 数仓开发

## 云上搭建 Airflow

最近更新时间：2021-11-12 15:45:15

[Apache Airflow](#) 是一款开源的工作流管理系统，集成了编排、调度、监控以及图形化展示等功能。在数据仓库场景，Airflow 则可以应用于 ETL 任务的管理。本文主要介绍如何在云端服务器上搭建 Airflow。

### Airflow 默认安装


1. 购买 [云服务器](#)。

#### 注意

本文以 CentOS 8.0 为例。

镜像

[公共镜像](#) [自定义镜像](#) [共享镜像](#) [镜像市场](#) [?](#)

 CentOS [v](#) [64位](#) [v](#) [CentOS 8.0 64位](#) [?](#)

2. 安装依赖软件

安装 Airflow 前，需安装如下依赖。

```
yum install redhat-rpm-config -y
yum install mysql-devel -y
yum install python3-devel -y
dnf update gcc annobin -y
```

3. 创建 Home 目录

```
mkdir -p /usr/local/services/airflow
export AIRFLOW_HOME=/usr/local/services/airflow
```

AIRFLOW\_HOME 变量可以配置到 `/etc/profile` 文件中。

4. 安装 Airflow

```
pip install apache-airflow[mysql]
```

5. 初始化 DB

```
airflow initdb
```

6. 配置安全组

Airflow 的 webui 默认启动在8080端口，如果想要通过外网访问，需要打开安全组的8080。

<a href="#">添加规则</a>	<a href="#">导入规则</a>	<a href="#">排序</a>	<a href="#">删除</a>	<a href="#">一键放行</a>	<a href="#">教我设置</a>	
<input type="checkbox"/>	来源 <a href="#">?</a> <a href="#">v</a>	协议端口 <a href="#">?</a>	策略	备注	修改时间 <a href="#">?</a>	操作
<input type="checkbox"/>	0.0.0.0/0	TCP:8080	允许	-	2020-07-23 17:15:16	<a href="#">编辑</a> <a href="#">插入</a> <a href="#">删除</a>

7. 启动 webui

使用如下命令：

```
airflow webserver -D
```

如果通过 url `http://{ip}:8080/admin/` 可以正常访问页面，则代表配置成功。

## 处理时区

Airflow 使用 UTC 时间，与北京时间差8个小时，因此需要进行处理，由于 Airflow 写死部分代码，因此除了修改配置文件外，也需要修改源码，步骤如下：

1. 修改 AIRFLOW\_HOME 下的 airflow.cfg

```
default_timezone = utc 修改为 default_timezone = Asia/Shanghai
default_ui_timezone = UTC 修改为 default_ui_timezone = Asia/Shanghai
```

2. 修改文件 `/usr/local/lib/python3.6/site-packages/airflow/utils/timezone.py`

在语句 `utc = pendulum.timezone('UTC')` 下新增如下语句：

```
from airflow.configuration import conf
try:
    tz = conf.get("core", "default_timezone")
    if tz == "system":
        utc = pendulum.local_timezone()
    else:
        utc = pendulum.timezone(tz)
except Exception:
    pass
```

修改函数 `utcnow()` :

```
d = dt.datetime.utcnow() 修改为 d = dt.datetime.now()
```

3. 修改文件 `/usr/local/lib/python3.6/site-packages/airflow/utils/sqlalchemy.py`

在语句 `utc = pendulum.timezone('UTC')` 下添加如下内容：

```
from airflow.configuration import conf
try:
    tz = conf.get("core", "default_timezone")
    if tz == "system":
        utc = pendulum.local_timezone()
    else:
        utc = pendulum.timezone(tz)
except Exception:
    pass
```

注释语句：

```
cursor.execute("SET time_zone = '+00:00'")
```

4. 修改文件 `/usr/local/lib/python3.6/site-packages/airflow/www/templates/admin/master.html`

```
var UTCseconds = (x.getTime() + x.getTimezoneOffset()*60*1000);
修改为
var UTCseconds = x.getTime();
```

```
"timeFormat": "H:i:s %UTC%",
修改为
"timeFormat": "H:i:s",
```

## 5. 重启 webserver

```
cat {AIRFLOW_HOME}/airflow-webserver.pid
kill {pid}
airflow webserver -D
```

## 使用云数据库 MySQL 存储数据

Airflow 默认使用嵌入式的 Sqlite 存储数据，如果要上生产环境，必须满足高可用的要求，这里以云数据库 MySQL 为例，步骤如下：

### 1. 购买 [云数据库 MySQL](#)

#### ⚠ 注意

必须是高可用版或者金融版，基础版由于不支持 `explicit_defaults_for_timestamp` 参数，因此无法作为 Airflow 的存储。

### 2. 修改参数

在控制台修改参数 `explicit_defaults_for_timestamp` 为 ON，修改后如下：



The screenshot shows the 'Database Management' interface. The 'Parameters' tab is selected. A table lists various parameters, with 'explicit\_defaults\_for\_timestamp' highlighted in orange and its value 'ON' circled in red.

参数名	是否重启	参数默认值 <sup>①</sup>	参数运行值	参数可修改值
eq_range_index_dive_limit <sup>①</sup>	否	200	200	[1-4294967295]
event_scheduler <sup>①</sup>	否	OFF	OFF	[ON   OFF]
<b>explicit_defaults_for_timestamp<sup>①</sup></b>	是	OFF	<b>ON</b>	[ON   OFF]

### 3. 创建 DB 及用户

登录 MySQL，运行如下语句，其中用户名及密码可根据用户情况进行修改。

```
create database airflow;
create user 'airflowuser'@'%' identified by 'pwd123';
grant all on airflow.* to 'airflowuser'@'%';
flush privileges;
```

### 4. 修改 `{AIRFLOW_HOME}/airflow.cfg` 中的配置

```
sql_alchemy_conn = sqlite:///usr/local/services/airflow/airflow.db
修改为
sql_alchemy_conn = mysql://airflowuser:pwd123@{ip}/airflow
```

### 5. 重新初始化数据库

```
airflow initdb
```

如果想保留之前的运行数据，可运行如下命令：

```
airflow resetdb
```