Tencent Cloud TCHouse-P

Data Warehouse Development

Service Notice

This document provides an overview of the as-is details of Tencent Cloud's products and services in their entirety or part. The descriptions of certain products and services may be subject to adjustments from time to time.

The commercial contract concluded by you and Tencent Cloud will provide the specific types of Tencent Cloud products and services you purchase and the service standards. Unless otherwise agreed upon by both parties, Tencent Cloud does not make any explicit or implied commitments or warranties regarding the content of this document.

Contact Us

We are committed to providing personalized pre-sales consultation and technical after-sale support. Don't hesitate to contact us at 4009100100 or 95716 for any inquiries or concerns.

Contents

Data Warehouse Development
Creating Airflow in Cloud

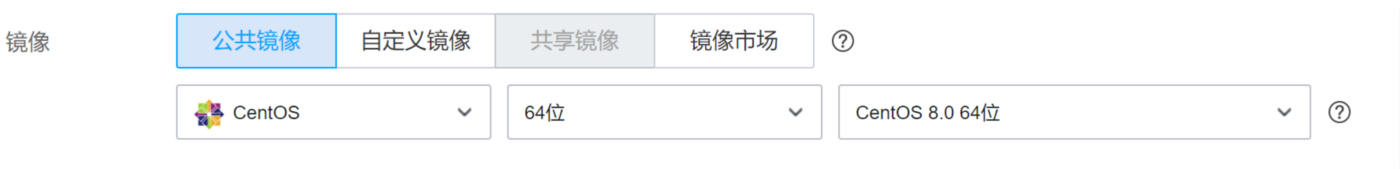
Last updated：2024-08-22 17:09:12

Apache Airflow is an open-source workflow management system that integrates orchestration, scheduling, monitoring, and graphical display features. In data warehouse scenarios, Airflow can be used for managing ETL tasks. This document mainly introduces how to set up Airflow on a cloud server.

Default Airflow Installation

1. Purchase CVM .

> **Note**
> This document uses CentOS 8.0 as an example.



镜像

公共镜像 | 自定义镜像 | 共享镜像 | 镜像市场

CentOS | 64位 | CentOS 8.0 64位

2. Install the dependent software.
Before installing Airflow, you need to install the following dependencies.

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

3. Create the Home Directory

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

The AIRFLOW_HOME variable can be configured in the `/etc/profile` file.

4. Install Airflow

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

5. Initialize the DB

```
airflow initdb
```

6. Configure Security Groups
Airflow's web UI starts on port 8080 by default. To access it via external network, you need to open port 8080 in the security groups.

7. Enable the web UI.
   Use the following command:

```
airflow webserver -D
```

   If you can access the UI at `url http://{ip}:8080/admin/`, the configuration is successful.

Processing Time Zone

Airflow uses the UTC time zone, which is eight hours behind Beijing time. As Airflow writes some fixed code, you need to modify the source code in addition to the configuration files in the following steps:

1. Modify `AIRFLOW_HOME` in the `airflow.cfg` file

```
Change default_timezone = utc to default_timezone = Asia/Shanghai
Change default_ui_timezone = UTC to default_ui_timezone = Asia/Shanghai
```

2. Modify the `/usr/local/lib/python3.6/site-packages/airflow/utils/timezone.py` file
   Add the following statement below the `utc = pendulum.timezone('UTC')` statement:

```
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
```

   Modify the `utcnow()` function:

```
Change d = dt.datetime.utcnow() to d = dt.datetime.now()
```

3. Modify the `/usr/local/lib/python3.6/site-packages/airflow/utils/sqlalchemy.py` file
   Add the following content below the `utc = pendulum.timezone('UTC')` statement:

```
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
```

   Comment the statement:

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

4. Modify the `/usr/local/lib/python3.6/site-packages/airflow/www/templates/admin/master.html` file

```
var UTCseconds = (x.getTime() + x.getTimezoneOffset()*60*1000);
Change to
var UTCseconds = x.getTime();
```

```
"timeFormat":"H:i:s %UTC%",
Change to
"timeFormat":"H:i:s",
```

5. Restart the webserver

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

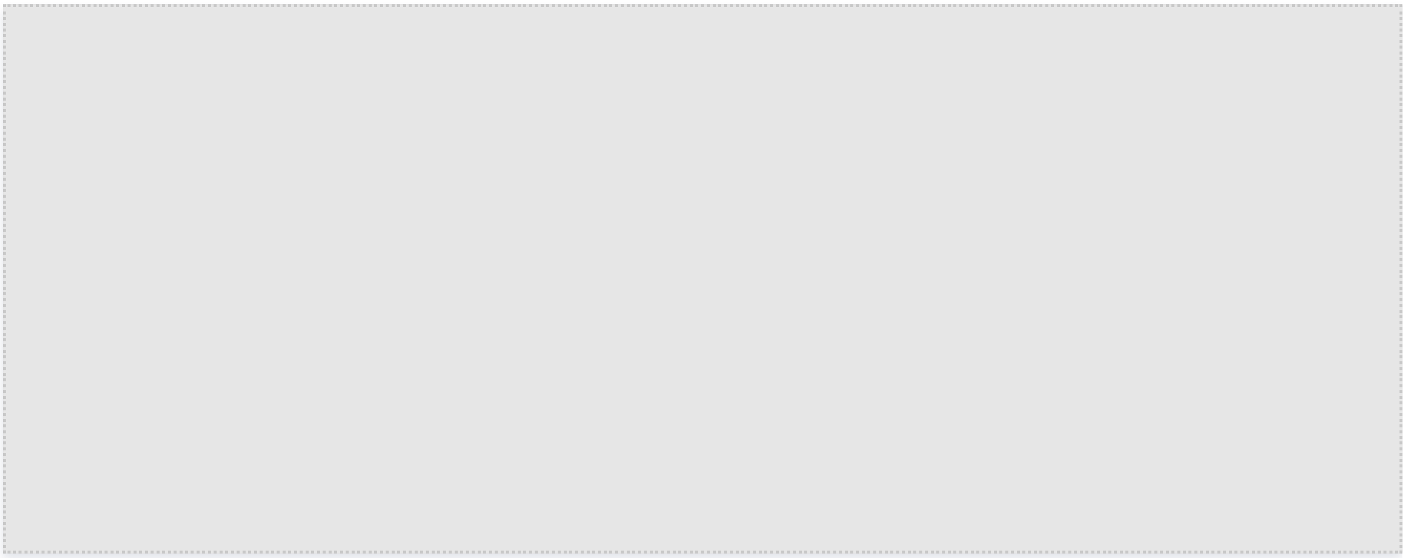
Using TencentDB for MySQL to Store Data

Airflow uses SQLite to store data by default. If you want to launch it in the production environment, you must ensure the high availability. In the following steps, TencentDB for MySQL is used as an example:

1. Purchase TencentDB for MySQL

> ⚠ **Note**
> Must be the High Availability Edition or Financial Edition. The Basic Edition cannot be used as storage for Airflow because it does not support the explicit_defaults_for_timestamp parameter.

2. Modify parameters
In the console, change the parameter explicit_defaults_for_timestamp to ON, after modification as follows:



3. Create a database and user.
Log in to MySQL and run the following statement, where you can change the username and password as needed.

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

4. Modify the configuration in `{AIRFLOW_HOME}/airflow.cfg`

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

5. Reinitialize the database

```
airflow initdb
```

If you want to retain the data from previous runs, run the following command:

```
airflow resetdb
```