链滴

Airflow 安装

作者：Mamba24L8

原文链接：https://ld246.com/article/1556442877206

来源网站：链滴

Airflow安装部署

新闻信息是通过爬虫获取，使用scrapy框架进行爬虫任务；使用airflow工作流监控平台对爬虫任务进管理、监控（可使用CeleryExecutor分布式，也可使用LocalExecutor多进程进行数据采集）。以下要是对airflow的安装和配置。

1.系统环境y

目前使用的系统环境为Centos Linux release 7.4.1708 (core),linux版本的内核Linux version 3.10.0-93.2.2e17.x86_64.

ip 地址：

- 外网： 47.104.191.52
- 内网： 172.31.178.92

2.准备python环境，安装Anaconda

2.1下载安装文件

下载地址1(官方网站)

下载地址2(清华开源镜像)

下载对应版本安装文件

2.2上传安装文件，开始安装

将下载的文件上传到Linux系统中 /opt

1、执行命令安装

cd /opt

sh Anaconda3-5.2.0-Linux-x86_64.sh (按回车键，直到出现>>> 输入yes)

/opt/anaconda3 (安装目录)

2、配置环境变量

echo "export PATH=/opt/anaconda3/bin:$PATH" >> /etc/profile

source /etc/profile

3.安装mysql （供airflow使用）、redis

mysql作为airflow数据库，主要是记录airflow信息；

redis作为celery的broker和backend（也可以用RabbitMQ），如果不使用CeleryExecutor则不需要redis配置。

4.安装配置airflow

1. 通过　anaconda安装虚拟环境news_push

```
/opt/anaconda3/bin/conda create -y --name news_push python=3.6.5
```

2. airflow安装、配置

- 激活虚拟环境　news_push

```
source activate news_push
```

- 通过pip安装airflow

```
pip install apache-airflow
```

- 配置airflow目录(先创建/opt/NewsPush项目目录)

```
echo "export AIRFLOW_HOME=/opt/NewsPush/airflow >> /etc/profile"
source /etc/profile
```

- 初始化数据库

```
airflow initdb
```

- 启动airflow

```
airflow webserver -p 5556
```

可到浏览器查看http://**ip**:5556/admin/

- 配置　airflow-更改数据库为mysql
  - 修改mysql配置文件参数（/etc/my.cnf），并重启mysql

```
explicit_defaults_for_timestamp=true
```

  - 登录mysql

```
mysql -uroot -p 回车后输入密码
```

  - 新建用户airflow

```
create user 'airflow'@'localhost' identified by 'airflow';
```

  - 创建数据库airflow

```
create database airflow;
```

  - 赋予权限

```
grant all privileges on airflow.* to 'airflow'@'%' identified by 'airflow';
flush privileges;
```

- 修改airflow配置文件

```
vim /opt/NewsPush/airflow/airflow.cfg
```

修改内容为：

```
executor = CeleryExecutor
sql_alchemy_conn=mysql://ariflow:airflow@localhost:3306/ariflow
load_examples = False
endpoint_url = http://localhost:5556
base_url = http://localhost:5556
web_server_port = 5556
```

原文链接：Airflow 安装

```
broker_url = redis://172.31.178.92:6379/3
celery_result_backend = redis://172.31.178.92:6379/4
flower_port = 5557
```

- 安装celery支持及celeryde redis组件

```
pip install airflow[celery]
pip install celery[redis]
```

- 安装MySQL-python

```
yum install MySQL-python
pip install PyMySQL==0.7.1
```

如果PyMySQL版本为0.8.0或以上则会有警告:

```
/opt/anaconda3/envs/news_push/lib/python3.6/site-packages/pymysql/cursors.py:170: Warn
ng: (1300, "Invalid utf8mb4 chara
  result = self._query(query)
```

- 再次初始化

```
airflow initdb
```

- 错误解决
  - 错误信息

```
Traceback (most recent call last):
  File "/opt/anaconda3/envs/news_push/bin/airflow", line 17, in <module>
    from airflow import configuration
  File "/opt/anaconda3/envs/news_push/lib/python3.6/site-packages/airflow/__init__.py", line
0, in <module>
    from airflow import settings
  File "/opt/anaconda3/envs/news_push/lib/python3.6/site-packages/airflow/settings.py", line
159, in <module>
    configure_orm()
  File "/opt/anaconda3/envs/news_push/lib/python3.6/site-packages/airflow/settings.py", line
147, in configure_orm
    engine = create_engine(SQL_ALCHEMY_CONN, **engine_args)
  File "/opt/anaconda3/envs/news_push/lib/python3.6/site-packages/sqlalchemy/engine/__ini
__.py", line 424, in create_engine
    return strategy.create(*args, **kwargs)
  File "/opt/anaconda3/envs/news_push/lib/python3.6/site-packages/sqlalchemy/engine/strat
gies.py", line 81, in create
    dbapi = dialect_cls.dbapi(**dbapi_args)
  File "/opt/anaconda3/envs/news_push/lib/python3.6/site-packages/sqlalchemy/dialects/my
ql/mysqldb.py", line 102, in dbapi
    return __import__('MySQLdb')
ModuleNotFoundError: No module named 'MySQLdb'
```

  - 解决(MySQLdb对python3.*支持)

```
vim /opt/anaconda3/envs/news_push/lib/python3.6/site-packages/sqlalchemy/dialects/mysql
mysqldb.py
```

（最后一行错误信息.py文件路径）

在代码开头增加

```
import pymysql
pymysql.install_as_MySQLdb()
```

- 再次初始化

```
airflow initdb
```

3. airflow启动及测试

- 创建一个dag(/opt/NewsPush/airflow/dags/hello_world.py)

```
from airflow import DAG
from airflow.utils.dates import days_ago
from airflow.operators.bash_operator import BashOperator
from airflow.operators.python_operator import PythonOperator

default_args = {
    'owner': 'airflow',
    'start_date': days_ago(1) #必须设置，尽量用固定时间，如果使用动态的当前时间会有意想不到问题。任务会先执行一次，再根据起始时间和schedule_interval设置开始执行
}

dag = DAG(
    'example_hello_world_dag',
    default_args=default_args,
    description='my first DAG',
    # schedule_interval=timedelta(days=1)
    schedule_interval='0 */1 * * *'    #每个小时执行一次
)

def print_hello():
    return 'Hello World!'

hello_operator = PythonOperator(
    task_id='hello_task',
    python_callable=print_hello,
    dag=dag
)
```

- airflow启动

以下命令都是单独开启一个窗口来启动，便于观察日志（也可以在后台启动）

注意：celery worker启动尽量不要用root用户启动，如果要用root用户启动则添加环境变量。

用其他用户启动则airflow启动命令也对应用用户启动，并更改项目目录权限属于此用户，否则日志录时没有权限会影响worker运行。

```
echo export C_FORCE_ROOT= true >> /etc/profile
source /etc/profile

airflow webserver #启动airflow web页面
airflow scheduler #启动调度器，执行任务调度，不过任务默认是关闭的，需要在页面手动开启
airflow worker #启动celery workd
airflow flower #启动flower监控页面
```

linux添加用户、用户组、密码

```
groupadd airflow #添加用户组airflow
useradd -g airflow airflow #添加用airflow到用户组airflow
passwd airflow #设置密码
```

更改项目目录权限为启动用户(airflow)权限

```
chowm -R airflow:airflow /opt/NewsPush/
```

airflow 浏览器访问地址：http://47.104.191.52/admin

flower 浏览器访问地址：http://47.104.191.52/

参考资料

Airflow使用

Airflow安装启动

Airflow框架下支持celery的问题