

# Apache Airflow Installation on Ubuntu

Step 1 : update your local system using below command

```
sudo apt-get update
```

Step 2 : Apache Airflow requires pip for installation. Run below command.

```
sudo apt-get install python3-pip
```

```
hduser@bigdata-VirtualBox:~$ sudo apt-get install python3-pip
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  dh-python libpython3-dev libpython3.6 libpython3.6-dev libpython3.6-minimal
  libpython3.6-stdlib python3-dev python3-distutils python3-lib2to3
  python3-setuptools python3-wheel python3.6 python3.6-dev python3.6-minimal
Suggested packages:
  python-setuptools-doc python3.6-venv python3.6-doc binfmt-support
The following NEW packages will be installed:
  dh-python libpython3-dev libpython3.6-dev python3-dev python3-distutils
  python3-lib2to3 python3-pip python3-setuptools python3-wheel python3.6-dev
The following packages will be upgraded:
  libpython3.6 libpython3.6-minimal libpython3.6-stdlib python3.6
  python3.6-minimal
```

Step 3 : Install airflow using below command as follows

```
sudo pip3 install apache-airflow
```

```
hduser@bigdata-VirtualBox:~$ sudo pip3 install apache-airflow
The directory '/home/hduser/.cache/pip/http' or its parent directory is not owned by the current user and the cache has been disabled. Please check the permissions and owner of the directory. You may want to use the command 'pip --user install pip' to install and use pip with sudo, you may want sudo's -H flag.
The directory '/home/hduser/.cache/pip' or its parent directory is not owned by the current user and caching wheels has been disabled. Check the permissions and owner of the directory. You may want to use the command 'pip --user install pip' to install and use pip with sudo, you may want sudo's -H flag.
Collecting apache-airflow
  Downloading https://files.pythonhosted.org/packages/29/69/843d20697a04c262b3f3ceafc1ab7d2d0e5486db70435e587c69b265bd9b/apache-airflow-2.0.0-py3-none-any.whl (4.5MB)
100% |#####| 4.5MB 120kB/s
Collecting tzlocal<2.0.0,>=1.4 (from apache-airflow)
Collecting tenacity==6.2.0 (from apache-airflow)
  Downloading https://files.pythonhosted.org/packages/b5/05/ff089032442058bd3386f9cd991cd88ccac81dca1494d78751621ee35e62/tenacity-6.2.0-py3-none-any.whl (13kB)
100% |#####| 102kB 2.3MB/s
Requirement already satisfied: urllib3<1.26 in /usr/lib/python3/dist-packages (from apache-airflow)
Collecting apache-airflow-providers-http (from apache-airflow)
  Downloading https://files.pythonhosted.org/packages/65/ba/5a852e63b3e8f01b2c349bb554c418fb73d613d00b7d70ffcbcb37010c987/apache-airflow-providers-http-2.0.0-py3-none-any.whl (10kB)
100% |#####| 102kB 2.3MB/s
Collecting flask-login<0.5,>=0.3 (from apache-airflow)
Collecting python-daemon<=2.1.1 (from apache-airflow)
  Downloading https://files.pythonhosted.org/packages/5a/0c/57f15b1572661877ff1acbe66c2f5be9d999ae5fb128e22933d374f62aa1/python-daemon-2.2.4-py2.py3-none-any.whl (10kB)
100% |#####| 102kB 2.3MB/s
```

Step 4 : install mysql using below command

```
sudo apt install mysql-server
```

```
hduser@bigdata-VirtualBox:~$ sudo apt install mysql-server
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  libevent-core-2.1-6 libhtml-template-perl mysql-client-5.7 mysql-server-5.7
  mysql-server-core-5.7
Suggested packages:
  libipc-sharedcache-perl mailx tinycsa
The following NEW packages will be installed:
  libevent-core-2.1-6 libhtml-template-perl mysql-client-5.7 mysql-server
  mysql-server-5.7 mysql-server-core-5.7
0 upgraded, 6 newly installed, 0 to remove and 407 not upgraded.
Need to get 12.5 MB of archives.
After this operation, 124 MB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 mysql-client-5.7 amd64 5.7.32-0ubuntu0.18.04.1 [1,943
Get:2 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 mysql-server-core-5.7 amd64 5.7.32-0ubuntu0.18.04.1 [
```

Step 5 : Install the airflow mysql packages using below command

```
Sudo pip3 install "apache-airflow[mysql]"
```

```
hduser@bigdata-VirtualBox:~$ sudo pip3 install "apache-airflow[mysql]"
The directory '/home/hduser/.cache/pip/http' or its parent directory is not owned by the current user and the cache has been disabled.
g pip with sudo, you may want sudo's -H flag.
The directory '/home/hduser/.cache/pip' or its parent directory is not owned by the current user and caching wheels has been disabled.
ith sudo, you may want sudo's -H flag.
Requirement already satisfied: apache-airflow[mysql] in /usr/local/lib/python3.6/dist-packages
Requirement already satisfied: typing-extensions>=3.7.4; python_version < "3.8" in /usr/local/lib/python3.6/dist-packages (from ap
Requirement already satisfied: sqlalchemy<2,>=1.3.18 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[mysql])
Requirement already satisfied: termcolor>=1.1.0 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[mysql])
Requirement already satisfied: gunicorn<20.0,>=19.5.0 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[mysql])
Requirement already satisfied: flask<2.0,>=1.1.0 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[mysql])
Requirement already satisfied: flask-swagger==0.2.13 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[mysql])
Requirement already satisfied: jinja2<2.12.0,>=2.10.1 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[mysql])
Requirement already satisfied: pendulum==2.0 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[mysql])
Requirement already satisfied: tenacity==6.2.0 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[mysql])
Requirement already satisfied: colorlog==4.0.2 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[mysql])
Requirement already satisfied: thrift>=0.9.2 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[mysql])
```

Step 6 : Then create the DB, the user and grant the required privileges

```
mysql>CREATE DATABASE airflow CHARACTER SET utf8 COLLATE utf8_unicode_ci;
mysql>create user 'airflow'@'localhost' identified by 'airflow';
mysql>grant all privileges on * . * to 'airflow'@'localhost';
mysql>flush privileges;
mysql>quit
```

As follows below

```
hduser@bigdata-VirtualBox:~$ sudo mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 5
Server version: 5.7.32-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2020, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> CREATE DATABASE airflow CHARACTER SET utf8 COLLATE utf8_unicode_ci;
Query OK, 1 row affected (0.00 sec)

mysql> CREATE USER 'airflow' IDENTIFIED BY 'airflow';
Query OK, 0 rows affected (0.00 sec)

mysql> GRANT ALL PRIVILEGES ON airflow.* TO 'airflow';
Query OK, 0 rows affected (0.00 sec)
```

Step 7 : Install celery executor airflow package using below command

```
hduser@bigdata-VirtualBox:~/airflow$ sudo pip3 install 'apache-airflow[celery]'
The directory '/home/hduser/.cache/pip/http' or its parent directory is not owned by the current user and the cache has been disabled.
Upgrade pip with sudo, you may want sudo's -H flag.
The directory '/home/hduser/.cache/pip' or its parent directory is not owned by the current user and caching wheels has been disabled.
Upgrade pip with sudo, you may want sudo's -H flag.
Requirement already satisfied: apache-airflow[celery] in /usr/local/lib/python3.6/dist-packages
Requirement already satisfied: unicodedcsv>=0.14.1 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[celery])
Requirement already satisfied: python-daemon>=2.1.1 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[celery])
Requirement already satisfied: pandas<2.0,>=0.17.1 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[celery])
Requirement already satisfied: argcomplete~=1.10 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[celery])
Requirement already satisfied: apache-airflow-providers-sqlite in /usr/local/lib/python3.6/dist-packages (from apache-airflow[celery])
Requirement already satisfied: flask-wtf<0.15,>=0.14.3 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[celery])
Requirement already satisfied: python-dateutil<3,>=2.3 in /usr/lib/python3/dist-packages (from apache-airflow[celery])
Requirement already satisfied: cryptography>=0.9.3 in /usr/lib/python3/dist-packages (from apache-airflow[celery])
Requirement already satisfied: requests<2.24.0,>=2.20.0 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[celery])
Requirement already satisfied: pygments<3.0,>=2.0.1 in /usr/local/lib/python3.6/dist-packages (from apache-airflow[celery])
Requirement already satisfied: pep562~=1.0; python version < "3.7" in /usr/local/lib/python3.6/dist-packages (from apache-airflow[celery])
```

Step 8 : Update the airflow.cfg file (should be available in ~/airflow/ directory.)

sql connection as:

```
sql_alchemy_conn = mysql://airflow:airflow@localhost:3306/airflow
```

there should be a sql alchemy connection string and you can comment it and add the above

executor as:

```
executor = CeleryExecutor
```

Save the file and exit.



Step 9 : airflow initialization using below command

```
airflow db init
```

```
hduser@bigdata-VirtualBox:~/airflow$ airflow db init
DB: mysql://airflow:***@localhost:3306/airflow
[2021-01-07 16:08:57,117] {db.py:678} INFO - Creating tables
INFO [alembic.runtime.migration] Context impl MySQLImpl.
INFO [alembic.runtime.migration] Will assume non-transactional DDL.
INFO [alembic.runtime.migration] Running upgrade d2ae31099d61 -> 0e2a74e0fc9f, Add time zone awareness
INFO [alembic.runtime.migration] Running upgrade d2ae31099d61 -> 33ae817a1ff4, kubernetes_resource_checkpointing
INFO [alembic.runtime.migration] Running upgrade 33ae817a1ff4 -> 27c6a30d7c24, kubernetes_resource_checkpointing
INFO [alembic.runtime.migration] Running upgrade 27c6a30d7c24 -> 86770d1215c0, add kubernetes scheduler uniqueness
INFO [alembic.runtime.migration] Running upgrade 86770d1215c0, 0e2a74e0fc9f -> 05f30312d566, merge heads
INFO [alembic.runtime.migration] Running upgrade 05f30312d566 -> f23433877c24, fix mysql not null constraint
INFO [alembic.runtime.migration] Running upgrade f23433877c24 -> 856955da8476, fix sqlite foreign key
INFO [alembic.runtime.migration] Running upgrade 856955da8476 -> 9635ae0956e7, index-faskfail
INFO [alembic.runtime.migration] Running upgrade 9635ae0956e7 -> dd25f486b8ea, add idx_log_dag
INFO [alembic.runtime.migration] Running upgrade dd25f486b8ea -> bf00311e1990, add index to taskinstance
```

**Note:** when initializing the db if you get the below timestamp error

```
util.load_python_file(self.dir, "env.py")
File "/usr/local/lib/python3.6/dist-packages/alembic/util/pyfiles.py", line 98, in load_python_file
    module = load_module_py(module_id, path)
File "/usr/local/lib/python3.6/dist-packages/alembic/util/compat.py", line 184, in load_module_py
    spec.loader.exec_module(module)
File "<frozen importlib._bootstrap_external>", line 678, in exec_module
File "<frozen importlib._bootstrap>", line 219, in _call_with_frames_removed
File "/usr/local/lib/python3.6/dist-packages/airflow/migrations/env.py", line 108, in <module>
    run_migrations_online()
File "/usr/local/lib/python3.6/dist-packages/airflow/migrations/env.py", line 102, in run_migrations_online
    context.run_migrations()
File "<string>", line 8, in run_migrations
File "/usr/local/lib/python3.6/dist-packages/alembic/runtime/environment.py", line 846, in run_migrations
    self.get_context().run_migrations(**kw)
File "/usr/local/lib/python3.6/dist-packages/alembic/runtime/migration.py", line 522, in run_migrations
    step.migration_fn(**kw)
File "/usr/local/lib/python3.6/dist-packages/airflow/migrations/versions/0e2a74e0fc9f_add_time_zone_awareness.py", line 44, in upgrade
    raise Exception("Global variable explicit_defaults_for_timestamp needs to be on (1) for mysql")
Exception: Global variable explicit_defaults_for_timestamp needs to be on (1) for mysql
```

Open the mysql and update the timestamps as below image

```
hduser@bigdata-VirtualBox:~/airflow$ sudo mysql -u root
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 54
Server version: 5.7.32-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2020, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> SHOW GLOBAL VARIABLES LIKE '%timestamp%';
+-----+
| Variable_name | Value |
+-----+
| explicit_defaults_for_timestamp | OFF |
| log_timestamps | UTC |
+-----+
2 rows in set (0.01 sec)

mysql> SET GLOBAL explicit_defaults_for_timestamp = 1;
Query OK, 0 rows affected (0.00 sec)

mysql> SHOW GLOBAL VARIABLES LIKE '%timestamp%';
+-----+
| Variable_name | Value |
+-----+
| explicit_defaults_for_timestamp | ON |
| log_timestamps | UTC |
+-----+
2 rows in set (0.01 sec)

mysql> quit
Bye
```

Create admin user and give the password for login the web ui as follows

```
hduser@bigdata-VirtualBox:~/airflow$ airflow users create \
> --username admin \
> --firstname vanshikrishna \
> --lastname panwar \
> --role Admin \
> --email vanshikrishna.panwar@gmail.com

[2021-01-07 16:10:38,457] {manager.py:727} WARNING - No user yet created, use flask fab command to do it.
Password:
Repeat for confirmation:
Admin user admin created
```

Let's Start airflow web server and scheduler using this commands

```
admin@user-admin created
hduser@bigdata-VirtualBox:~/airflow$ airflow webserver --port 8080

      ____          _              _
     / __ \        | |             | |
    / /_/_         | |             | |
   / ___ \         | |             | |
  / /___\         _||_|            _||_|
 /_____\       _||_|            _||_|

[2021-01-07 16:12:06,483] {dagbag.py:440} INFO - Filling up the DagBag from /dev/null
Running the Gunicorn Server with:
Workers: 4 sync
Host: 0.0.0.0:8080
Timeout: 120
Logfiles: - -
Access Logformat:

=====
[2021-01-07 16:12:14 +0530] [15839] [INFO] Starting gunicorn 19.10.0
[2021-01-07 16:12:14 +0530] [15839] [INFO] Listening at: http://0.0.0.0:8080 (15839)
[2021-01-07 16:12:14 +0530] [15839] [INFO] Using worker: sync
[2021-01-07 16:12:14 +0530] [15843] [INFO] Booting worker with pid: 15843
[2021-01-07 16:12:14 +0530] [15844] [INFO] Booting worker with pid: 15844
[2021-01-07 16:12:14 +0530] [15845] [INFO] Booting worker with pid: 15845
```

Using below command :

```
airflow webserver --port 8080
```

Below command to start the scheduler

```
airflow scheduler
```

Open the browser and open the url : <http://localhost:8080/airflow>

The screenshot shows the Apache Airflow web interface. At the top, there is a navigation bar with the Airflow logo, the version number 0.0.0.0/0800/home, and a user profile icon. Below the navigation bar, the main content area is titled "DAGs". On the left side of the DAGs section, there are filters for "All" (29), "Active" (8), and "Paused" (26). A search bar labeled "Filter DAGs by tag" is also present. The main table lists the DAGs with the following columns: DAG, Owner, Runs, Schedule, Last Run, Recent Tasks, Actions, and Links. The table contains eight rows of example DAGs, each with a visual representation of its recent task history in the "Recent Tasks" column.

DAG	Owner	Runs	Schedule	Last Run	Recent Tasks	Actions	Links
example_bash_operator	airflow	3	@ 8 ***		3	▶ 🔍 🗑	...
example_branch_dop_operator_v3	airflow	3	@ 11 ***		3	▶ 🔍 🗑	...
example_branch_operator	airflow	3	@ daily		3	▶ 🔍 🗑	...
example_complex	airflow	3	None		3	▶ 🔍 🗑	...
example_dag_decorator	airflow	3	None		3	▶ 🔍 🗑	...
example_external_task_marker_child	airflow	3	None		3	▶ 🔍 🗑	...
example_external_task_marker_parent	airflow	3	None		3	▶ 🔍 🗑	...
example_kubernetes_executor	airflow	3	None		3	▶ 🔍 🗑	...

### Note : Disabling Examples

```
nano airflow/airflow.cfg...# Whether to load the examples that ship with Airflow. It's good to
# get started, but you probably want to set this to False in a production
# environment
```

```
load_examples = False
```