#### **Launch an EC2 Instance for Airflow**

Provision an EC2 instance that will host Apache Airflow:

* **AMI**: Ubuntu Server (latest version)
* **Instance Type**: We choose t2.xlarge (16 GiB memory, 4 vCPUs) to handle the large dataset (~2.6 GB CSV from Redfin) and support in-memory transformations with Pandas.
* **Security Groups**:
  + Allow **HTTP (80)**, **HTTPS (443)**, and **SSH (22)** for web and remote access.
  + Add a custom **TCP rule for port 8080** to allow Airflow UI access.
* **Key Pair**: Create or use an existing .pem file for SSH access.

After provisioning, we SSH into the EC2 instance using either EC2 Instance Connect or **VS Code Remote SSH**.

#### **Connect to EC2 and Install Apache Airflow**

After the EC2 instance is running:

* Connect to it using **EC2 Instance Connect** from the AWS Console.
* Once in the terminal, install Python, libraries and Airflow with the following commands:

Bash:

sudo apt update

sudo apt install python3-pip -y

sudo apt install python3.12-venv -y

python3.12 -m venv airflow\_venv

source airflow\_venv/bin/activate

pip install pandas

pip install boto3

pip install awscli

pip install apache-airflow

airflow version

aws configure (*configure with your user Access Key, Secret Access Key, your region)*

airflow standalone

* This will initialize Airflow and provide you with a username and password for accessing the **Airflow UI** via http://<EC2-Public-IP>:8080.
* Make sure port **8080** is open in your EC2 Security Group.

#### **Connect EC2 to Visual Studio Code via SSH**

To manage code and configurations more easily, connect your EC2 instance to **Visual Studio Code**:

* Open the AWS Console → EC2 → Connect → SSH client.
* Copy the SSH connection command and make your .pem file read-only:

Bash:

chmod 400 airflow-key.pem

* In VS Code:
  1. Install the **Remote - SSH** extension.
  2. Open a Remote Window (bottom-left corner).
  3. Select **Configure SSH Hosts** and add your EC2 instance:

Host airflow-ec2

HostName <EC2-Public-IP>

User ubuntu

IdentityFile /path/to/your/airflow-key.pem

* Save and then choose **"Connect to Host"** from the Remote window to connect.

#### **Configure the Airflow DAG Folder**

Once connected to EC2 from VS Code:

* Navigate to /home/ubuntu/airflow/.
* Create a new directory named dags. This is where all your DAG (Directed Acyclic Graph) files will be placed.
* Create a new file inside the dags folder called redfin\_analytics.py. This is where your ETL pipeline will be defined.

#### **Configure airflow.cfg**

* Open airflow.cfg in /home/ubuntu/airflow/.
* Set the DAGs folder path:

dags\_folder = /home/ubuntu/airflow/dags

* Disable example DAGs from showing up in the Airflow UI:

load\_examples = False

#### **Restart Airflow to Apply Configuration Changes**

* Press Ctrl + C in the EC2 terminal to stop the running Airflow server.
* Restart it by running:

airflow standalone

* Once restarted, open your browser and navigate to http://<EC2-Public-IP>:8080.
* Log in using the credentials provided when you initialized Airflow.