**Naya College**

Data Engineering

**Apache Airflow**



**Version**: 2.0.0

**Date**: Aug 2024

**Content**

[Lab1 4](#_Toc175782226)

[Automating Remote Data Processing with Airflow: set up an SSH connection 4](#_Toc175782227)

[Objective: 4](#_Toc175782228)

[Step 1: Access the Airflow Web UI 4](#_Toc175782229)

[Step 2: Locate the ssh\_default Connection 5](#_Toc175782230)

[Step 3: Edit the Connection Details 5](#_Toc175782231)

[Step 4: Save the Connection 6](#_Toc175782232)

[Step 5: Test the Updated Connection 6](#_Toc175782233)

[Summary of Changes: 6](#_Toc175782234)

## Lab1

## Automating Remote Data Processing with Airflow: set up an SSH connection

## Objective:

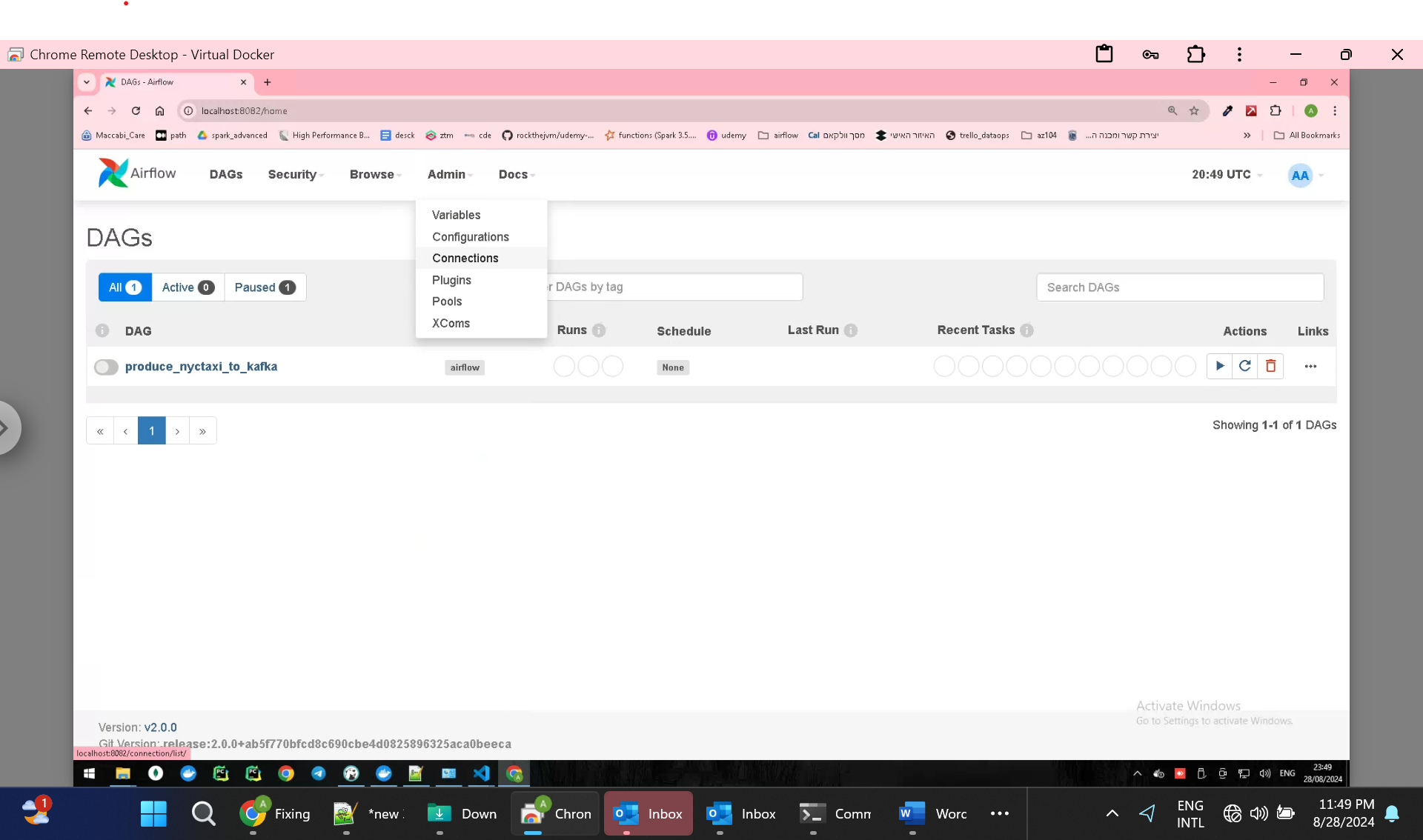
The objective of the task is to securely execute a Python script on a remote server (dev\_env) using SSH. This script, located in the specified directory on the remote server, is responsible for consuming data (e.g., NYC Taxi data) and processing it for further use in the data pipeline. By leveraging Airflow's SSHOperator, this task ensures that the script is executed in a consistent, automated manner, allowing for seamless integration into the broader ETL (Extract, Transform, Load) workflow.

This task is critical for initiating the data ingestion process from a remote environment, ensuring that data is reliably captured and made available for subsequent processing and analysis within the data pipeline.

To set up an SSH connection in Airflow with the hostname dev\_env, user developer, and password developer, follow these steps:

## Step 1: Access the Airflow Web UI

1. **Login** to your Airflow instance via a web browser.
2. Navigate to the **Admin** tab.
3. Select **Connections** from the dropdown menu.



## Step 2: Locate the ssh\_default Connection

1. In the **Connections** list, use the search bar or scroll through the list to find the connection named ssh\_default.
2. Click on the **Edit** button (usually represented by a pencil icon) next to the ssh\_default connection.

A screenshot of a computer

Description automatically generated

## Step 3: Edit the Connection Details

1. **Update the Hostname**:
   * **Host**: Change this to dev\_env if it's not already set.
2. **Update the Login Credentials**:
   * **Login**: Change this to developer.
   * **Password**: Update this to developer (if you’re using a password for SSH authentication).
3. **Review or Update the Port**:
   * **Port**: Ensure this is set to 22 (the default SSH port) unless your SSH server is configured to use a different port.

A screenshot of a computer

Description automatically generated

1. **Optional Fields**:
   * **Extra**: If you need to add or update any additional parameters (such as private key paths or connection timeout settings), this is where you would include them in JSON format. However, for basic setups, you can leave this as is.

## Step 4: Save the Connection

1. After making your changes, click the **Save** button at the bottom of the form to update the ssh\_default connection.

## Step 5: Test the Updated Connection

1. **Trigger a DAG**: Manually trigger a DAG that uses the ssh\_default connection to ensure that the updates are correct and the connection is functioning as expected.
2. **Check Logs**: Review the task logs in Airflow to confirm that the SSH connection was successful and that the remote command executed without issues.

## Summary of Changes:

* **Host**: dev\_env
* **Login**: developer
* **Password**: developer

By following these instructions, you'll successfully update the existing ssh\_default connection in Airflow to use the new credentials and hostname. This will ensure that any DAGs relying on this connection can connect to the correct server and execute their tasks.