# End-to-End Machine Learning Pipeline with Amazon SageMaker

Table of Contents

[End-to-End Machine Learning Pipeline with Amazon SageMaker 1](#_Toc202114928)

[Environment Setup 1](#_Toc202114929)

[Data Acquisition 1](#_Toc202114930)

[AWS CLI Installation (Windows) 2](#_Toc202114931)

[Creating IAM User(admin access) 3](#_Toc202114932)

[Setup security credentials 4](#_Toc202114933)

[S3bucket creation 6](#_Toc202114934)

[Role creation 7](#_Toc202114935)

In this project, we demonstrate how to build a complete machine learning pipeline using Amazon SageMaker, covering every stage from data acquisition and understanding to model development, deployment, and endpoint creation. SageMaker is a powerful AWS service that simplifies building, training, and deploying machine learning models at scale. By leveraging SageMaker, we can not only train and evaluate our models in a managed environment, but also deploy them as real-time endpoints to integrate seamlessly with other applications. This project serves as a practical, hands-on demonstration of these capabilities.

## Environment Setup

Before starting, create and activate a dedicated conda environment to manage dependencies cleanly:

**# Create a new conda environment with Python 3.11 in a specified folder (e.g., env)**

conda create --prefix ./env python=3.11

**# Activate the environment**

conda activate ./env

After activation, execute below line

(env) D:\aws\_sagemaker>

This ensures a clean and isolated environment for installing and running the necessary packages.

## Data Acquisition

The dataset for this project — **Mobile Price Classification** — was sourced from Kaggle. The Kaggle API was used to programmatically download and extract the dataset, ensuring that the data acquisition step is reproducible and automated.

**📌 Kaggle API Setup Instructions**

Before running the data download code, the Kaggle API must be configured. Follow these steps:

1️⃣ **Create your Kaggle API token**

* Visit [Kaggle account settings](https://www.kaggle.com/settings).
* Click on **“Create New API Token”**.
* This will download a kaggle.json file containing your API credentials.

2️⃣ **Place the kaggle.json file in the appropriate folder**

* Recommended location (Windows example):

C:\Users\<your\_username>\.kaggle\kaggle.json

or place it in the working directory where you will run your script.

3️⃣ **Verify you have kaggle installed**

pip install kaggle

4️⃣ **Then run the Python code( you can find in Jupyter code)**

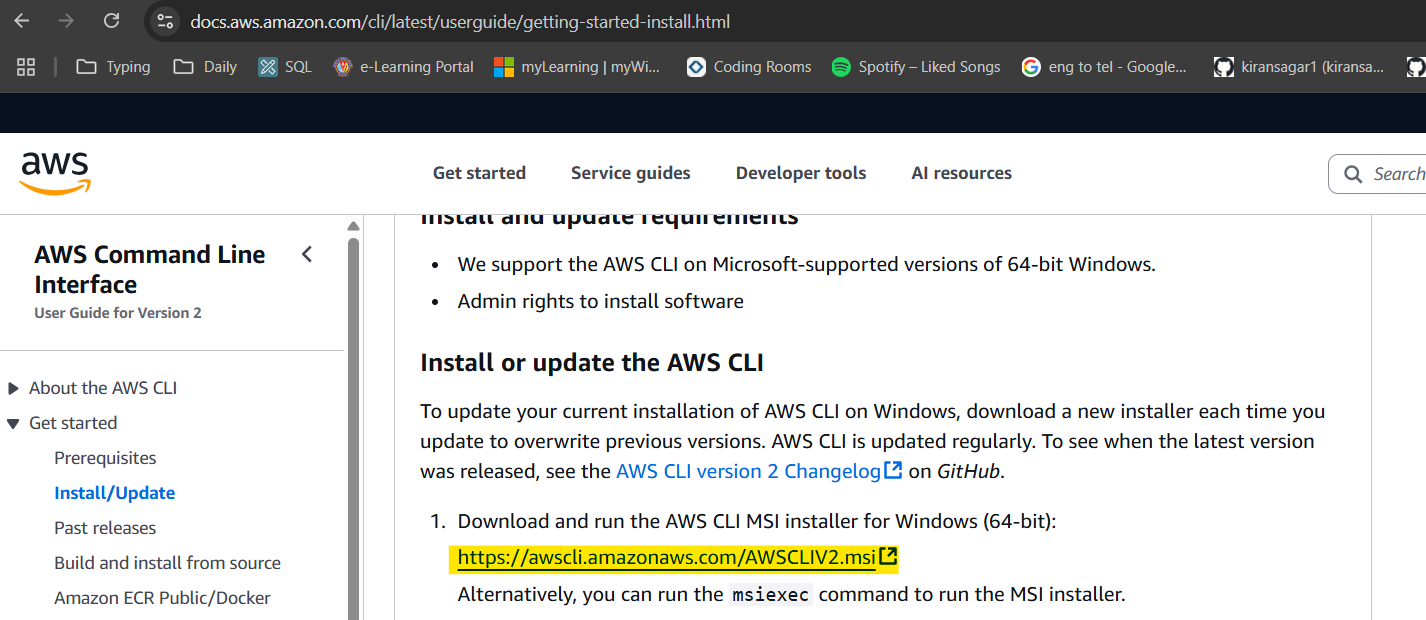
Alternatively, you can manually download the dataset from this link:  
[Mobile Price Classification Dataset on Kaggle](https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification?select=train.csv)

## AWS CLI Installation (Windows)

You can install the AWS CLI on Windows in two ways:

👉 **Option 1: Download directly**  
<https://awscli.amazonaws.com/AWSCLIV2.msi>

* Download the .msi installer from the above link
* Double-click to run the installer
* Follow the on-screen instructions



👉 **Option 2: Install via terminal**

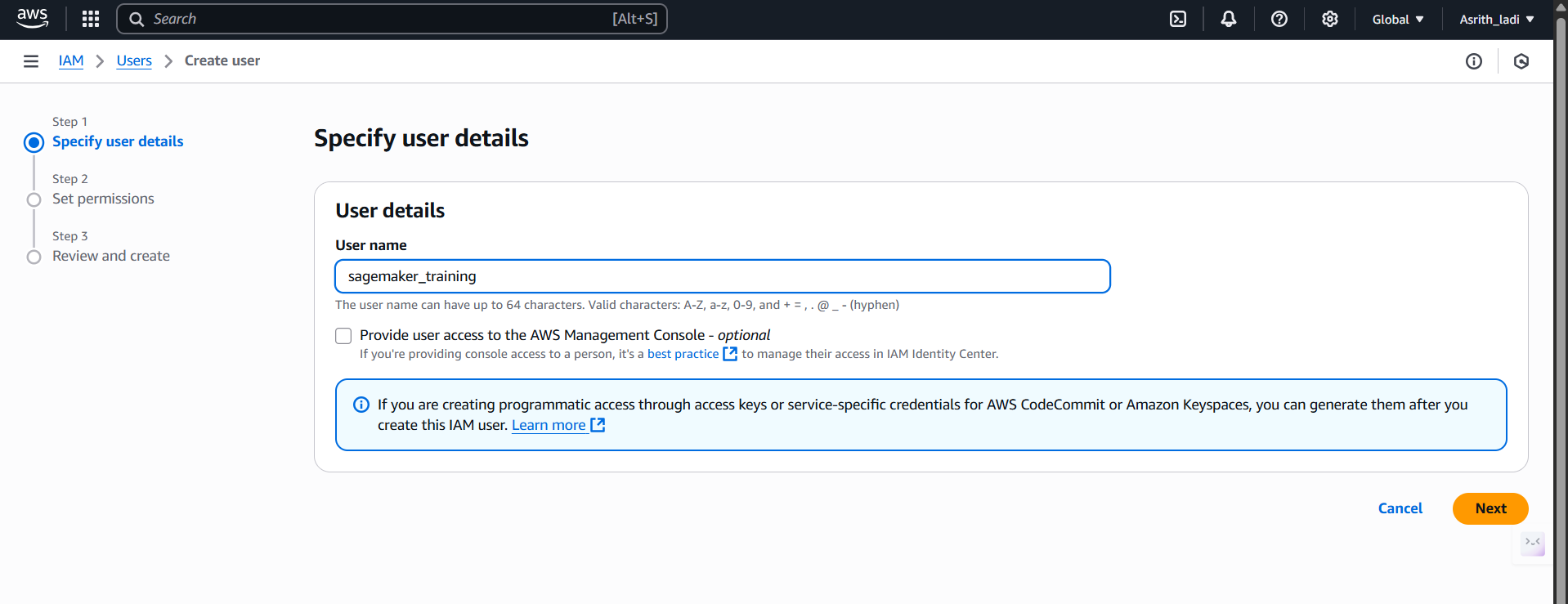
You can also install directly from the terminal with:

msiexec.exe /i https://awscli.amazonaws.com/AWSCLIV2.msi

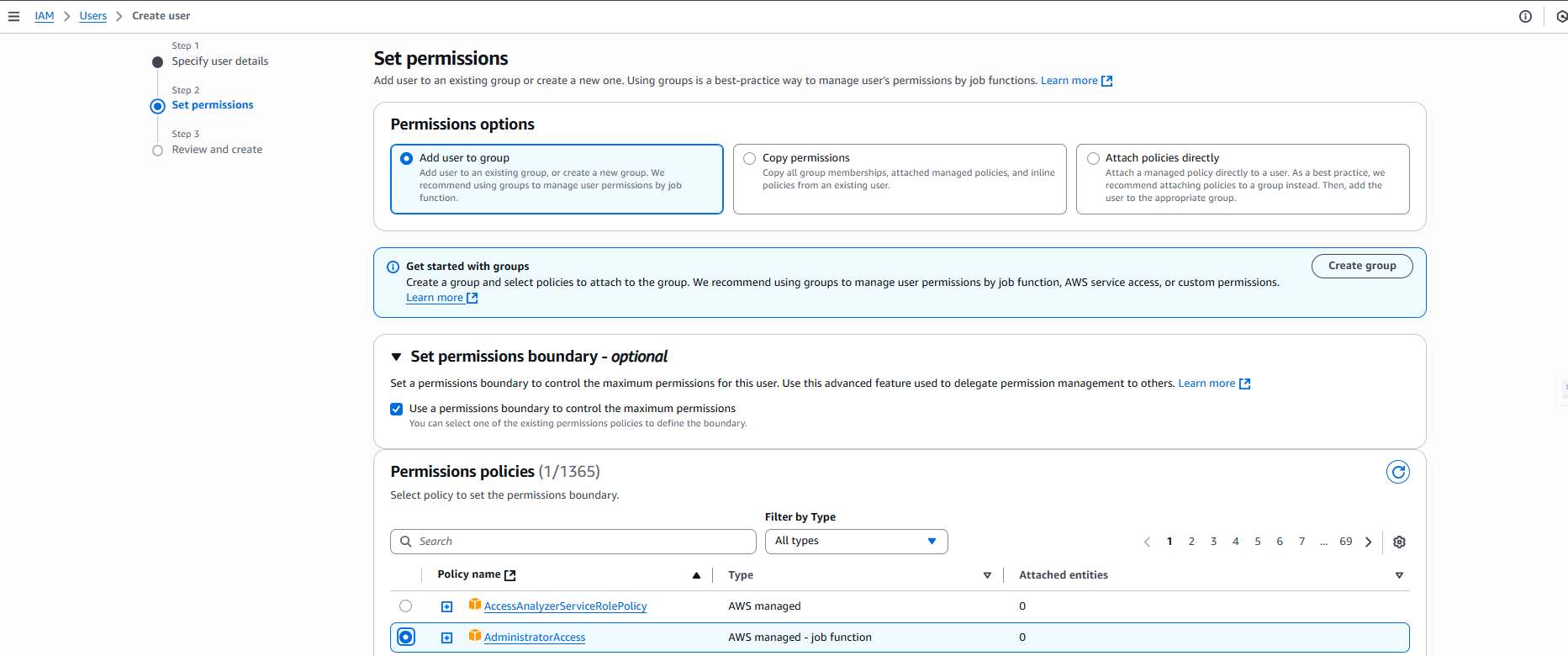
✅ After installation, confirm it is working:

aws --version

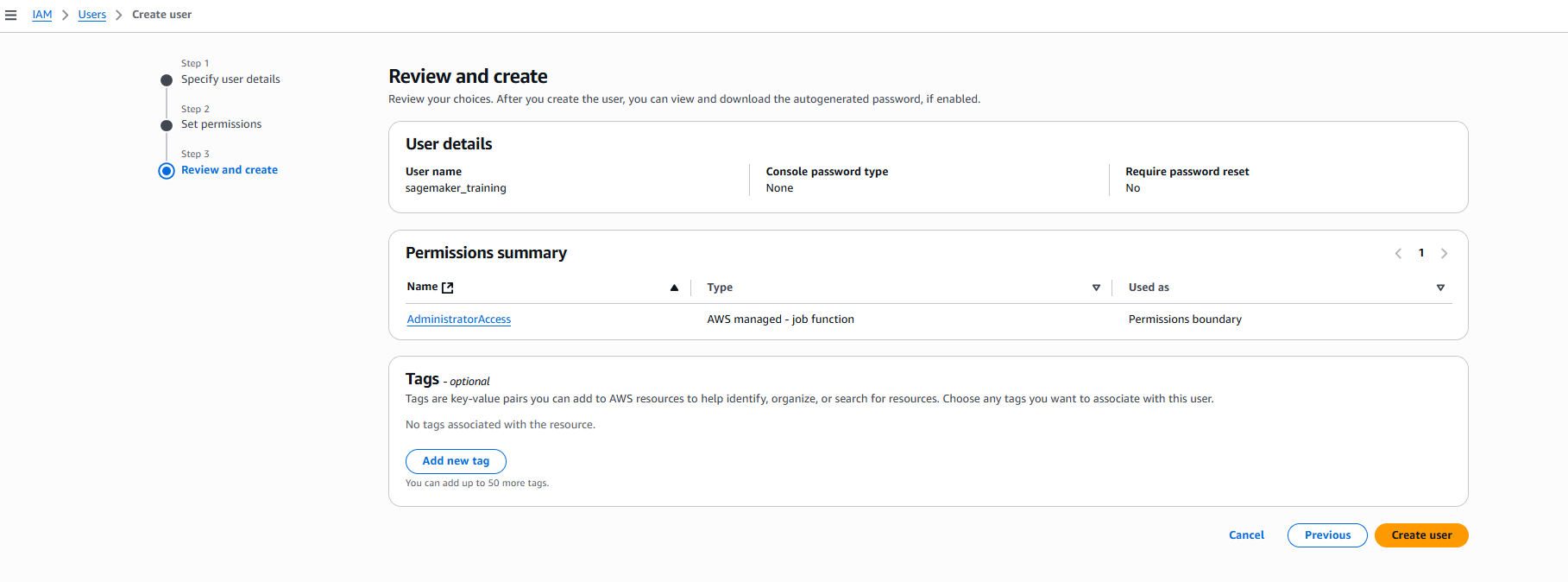
# Creating IAM User(admin access)



Click on Next

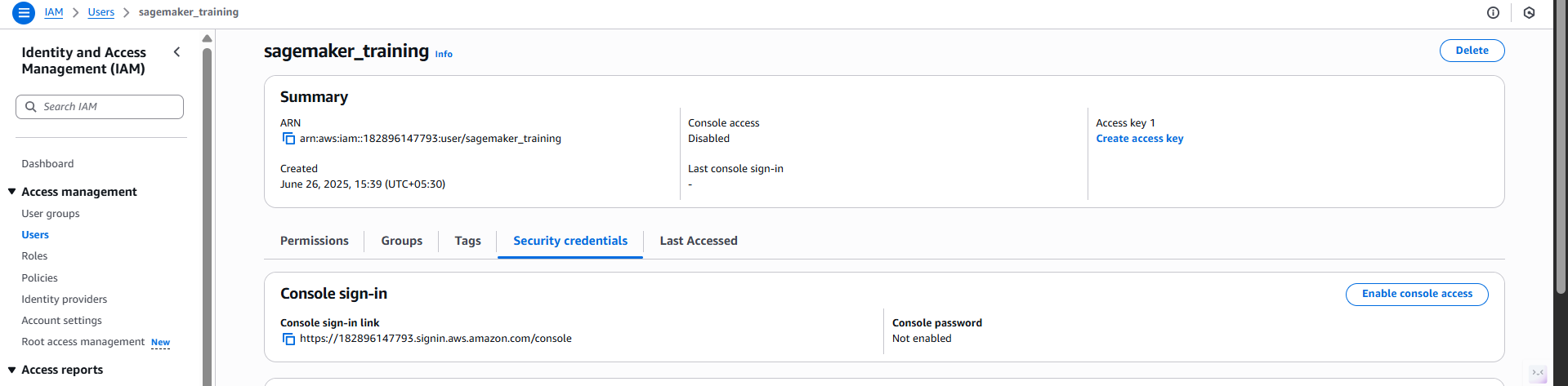


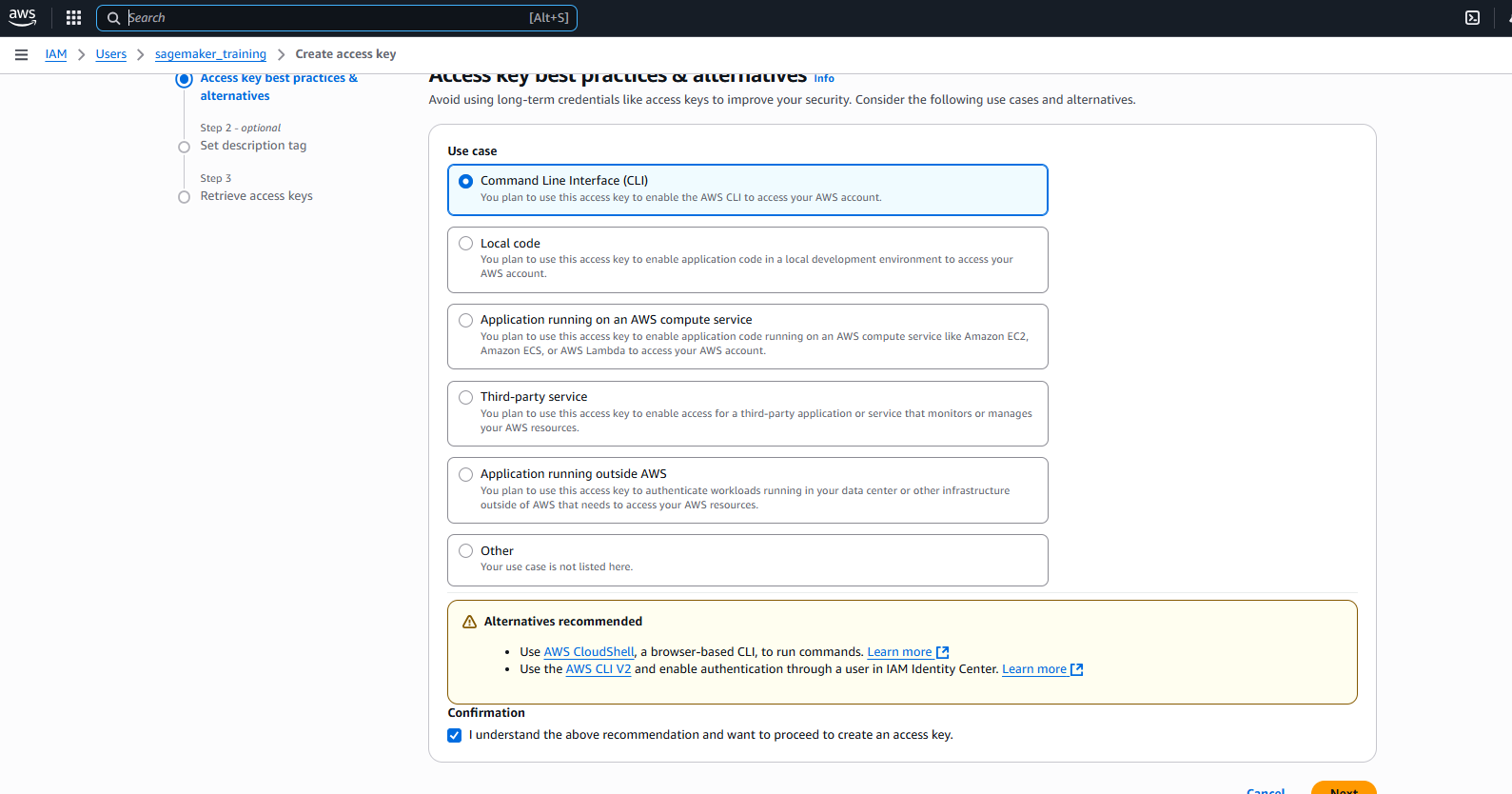
Click on create user



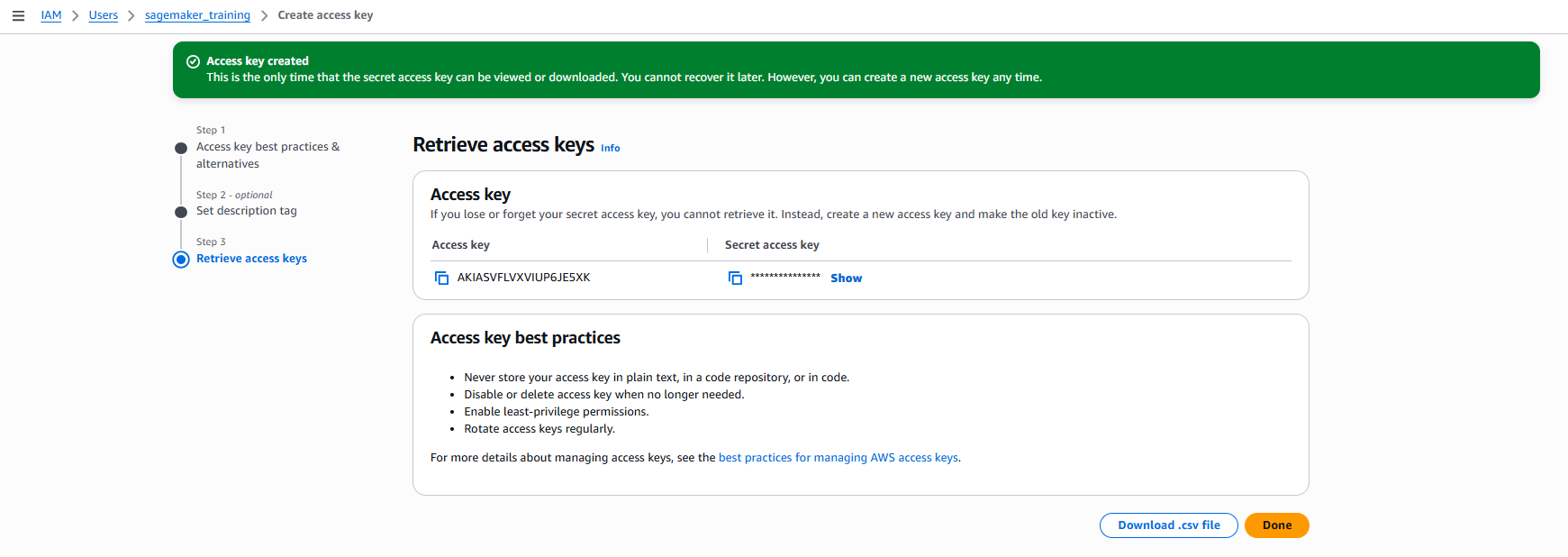
## Setup security credentials

Choose security credentials

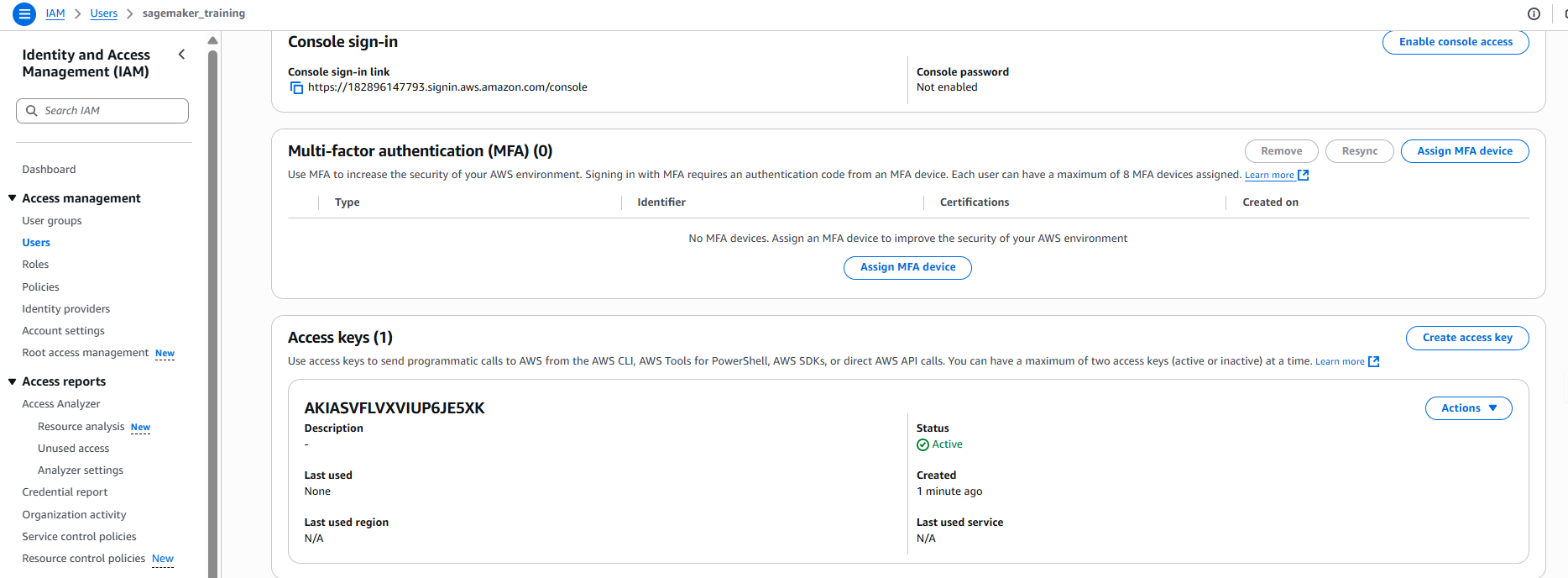




Copy the access credentials

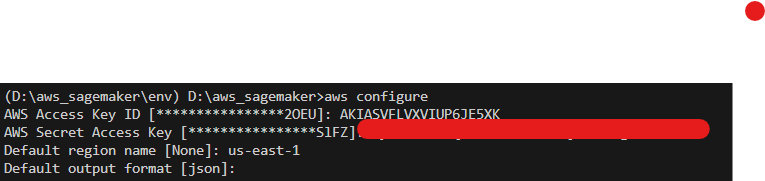


Download the .csv file locally or copy and store access credentials.



Open terminal

aws configure



>pip install -r requirements.txt

Requirement already satisfied: kagglehub in d:\aws\_sagemaker\env\lib\site-packages (from -r requirements.txt (line 1)) (0.3.12)

Requirement already satisfied: kaggle in d:\aws\_sagemaker\env\lib\site-packages (from -r requirements.txt (line 2)) (1.7.4.5)

Collecting sagemaker (from -r requirements.txt (line 3))

  Downloading sagemaker-2.247.1-py3-none-any.whl.metadata (17 kB)

Collecting scikit-learn (from -r requirements.txt (line 4))

  Downloading scikit\_learn-1.7.0-cp313-cp313-win\_amd64.whl.metadata (14 kB)

Requirement already satisfied: pandas in d:\aws\_sagemaker\env\lib\site-packages (from -r requirements.txt (line 5)) (2.2.3)

Requirement already satisfied: numpy in d:\aws\_sagemaker\env\lib\site-packages (from -r requirements.txt (line 6)) (2.2.5)

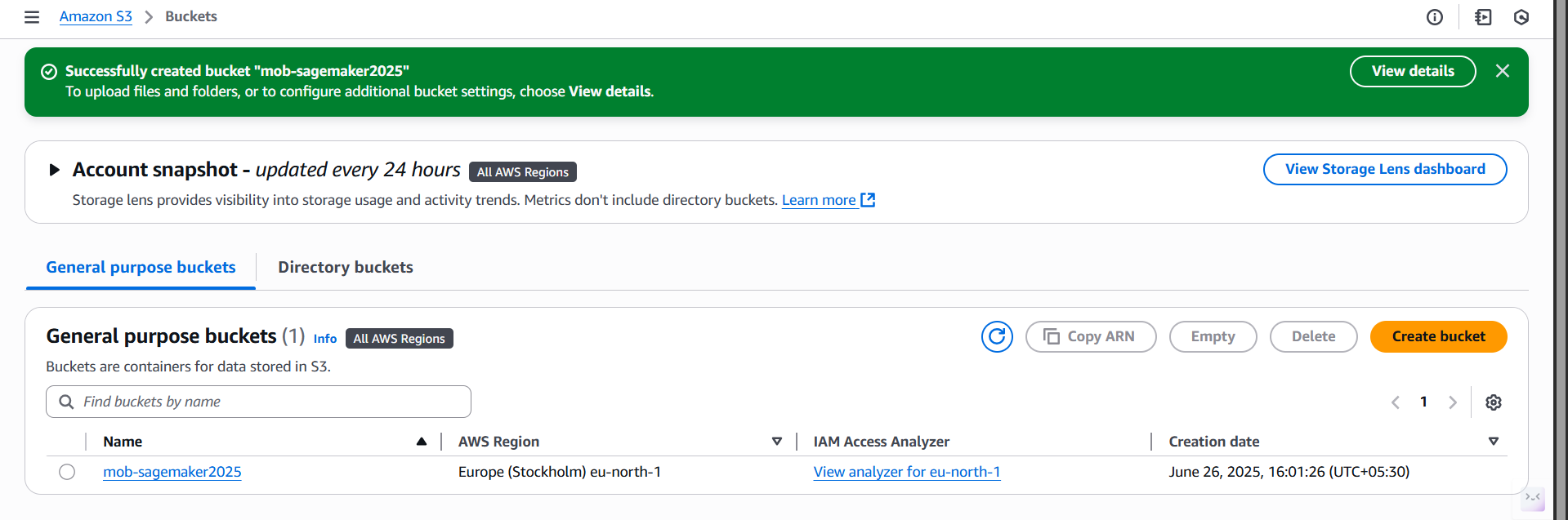
ERROR: Could not find a version that satisfies the requirement ipykernal (from versions: none)

ERROR: No matching distribution found for ipykernal

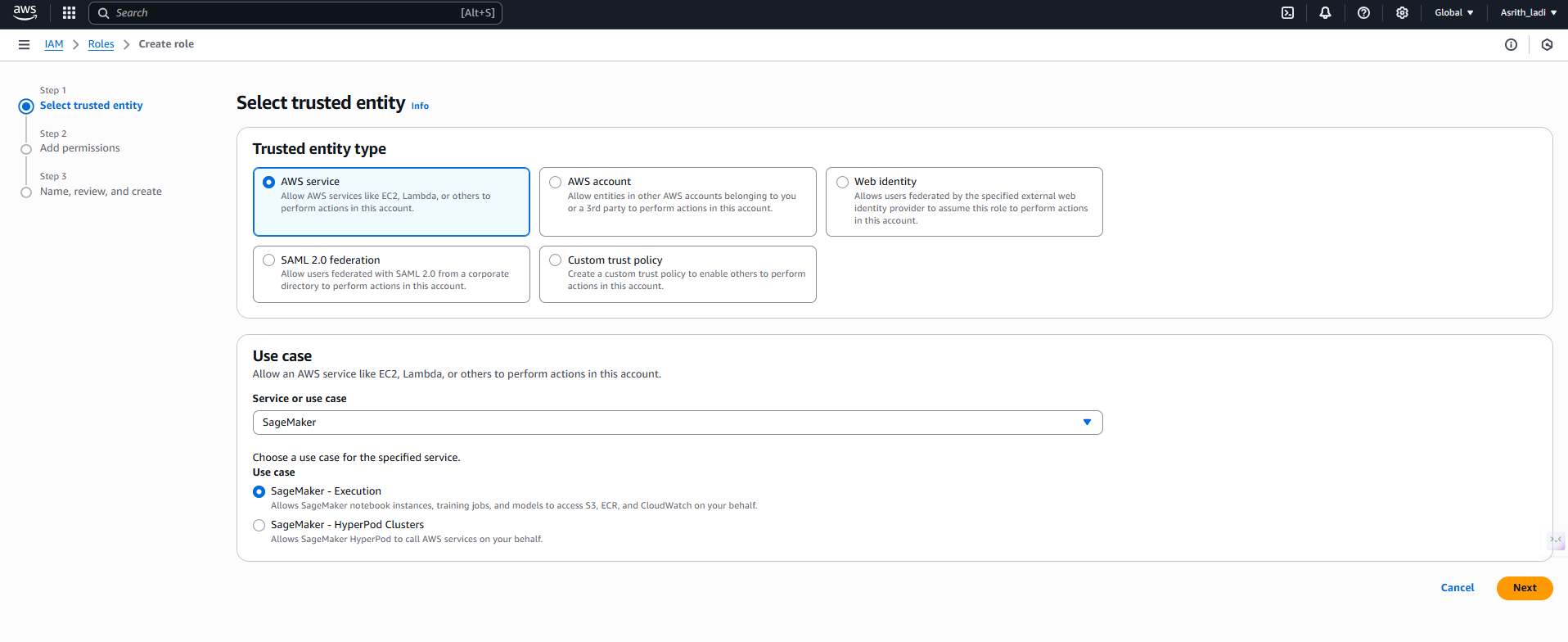
Got some issue while installation, so installed individually.

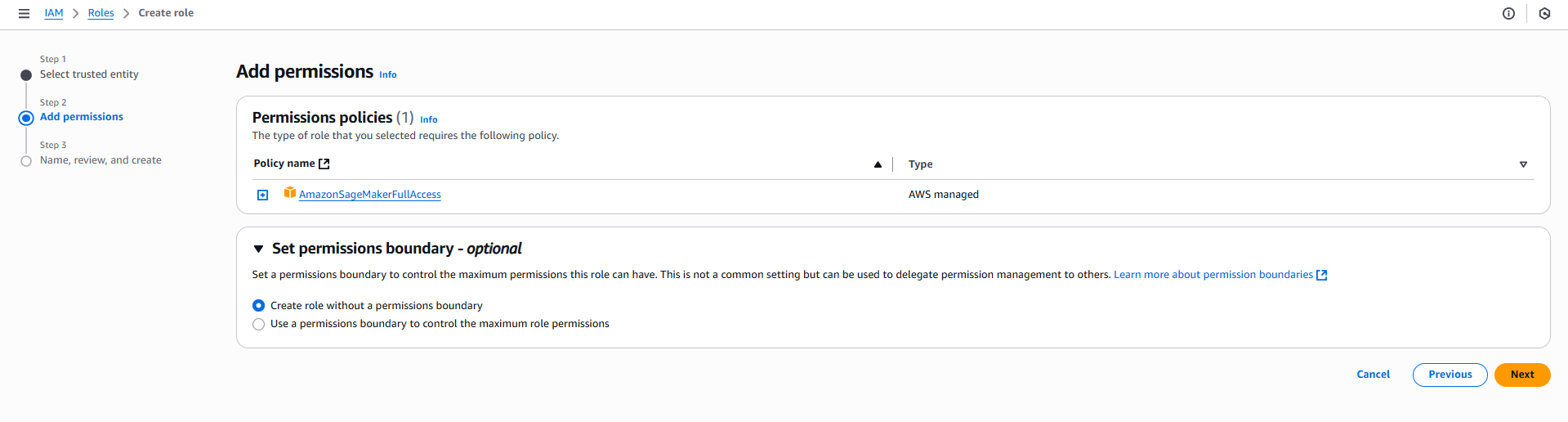
# S3bucket creation

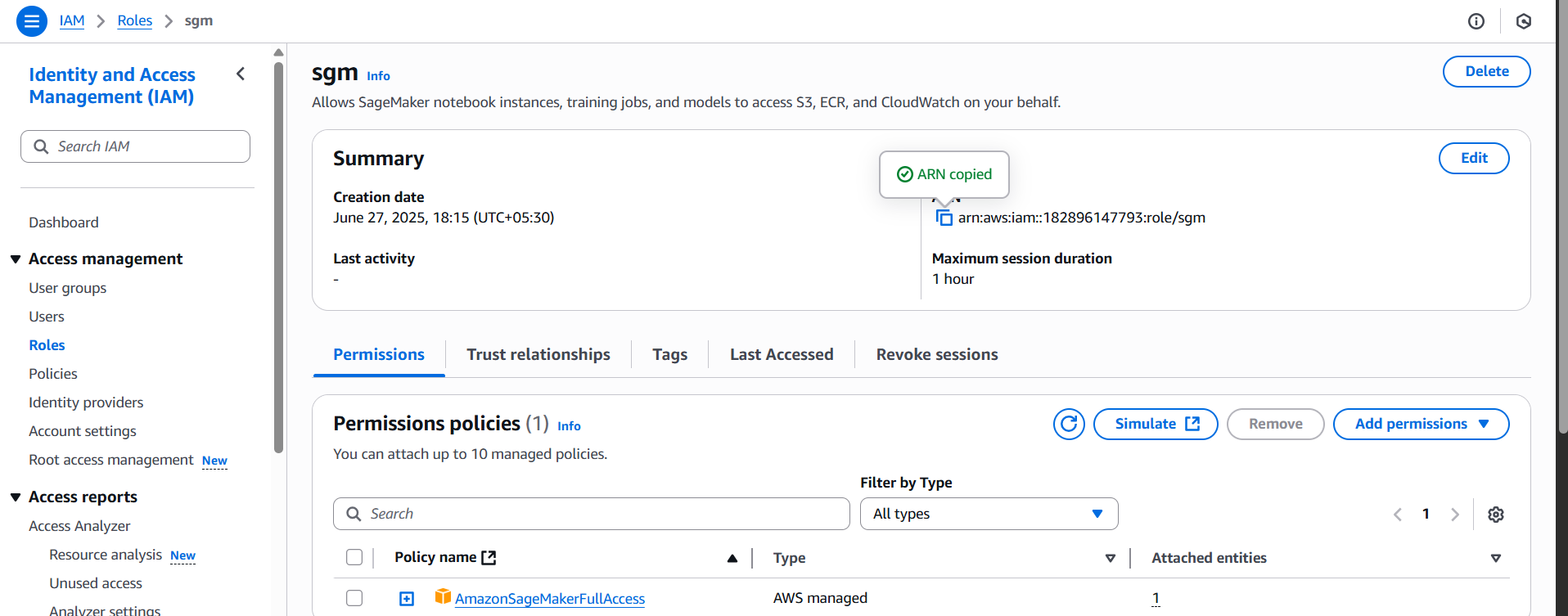
Open Aws console and navigate to S3.



# Role creation







Open your Jupyter Notebook and execute each code cell in sequence to run the entire workflow from data download through model deployment and endpoint testing.