**Placement Empowerment Program**

***Cloud Computing and DevOps Centre***

**Implement auto-scaling in the cloud :** Set up an auto-scaling group for your cloud VMs to handle variable workloads.

Name: Subiksha R Department: AML



**Introduction :**

Auto-scaling is a critical feature in cloud computing that ensures a system can dynamically adjust the number of resources (e.g., virtual machines) based on traffic demand. **Azure Virtual Machine Scale Sets (VMSS)** allow you to deploy and manage a group of load-balanced VMs that can automatically scale in or out depending on the defined rules. This approach enhances the availability, performance, and cost-effectiveness of applications, making it an essential component for workloads with variable resource requirements.

**Overview :**

In this task, we'll configure an **Auto Scaling Group** using **Azure Virtual Machine Scale Sets (VMSS)**. Azure provides a seamless and efficient way to automatically scale virtual machines to handle varying application loads. The scale set will automatically increase or decrease the number of virtual machines based on defined rules, ensuring that the right amount of computing power is always available to meet demand.

**Objective :**

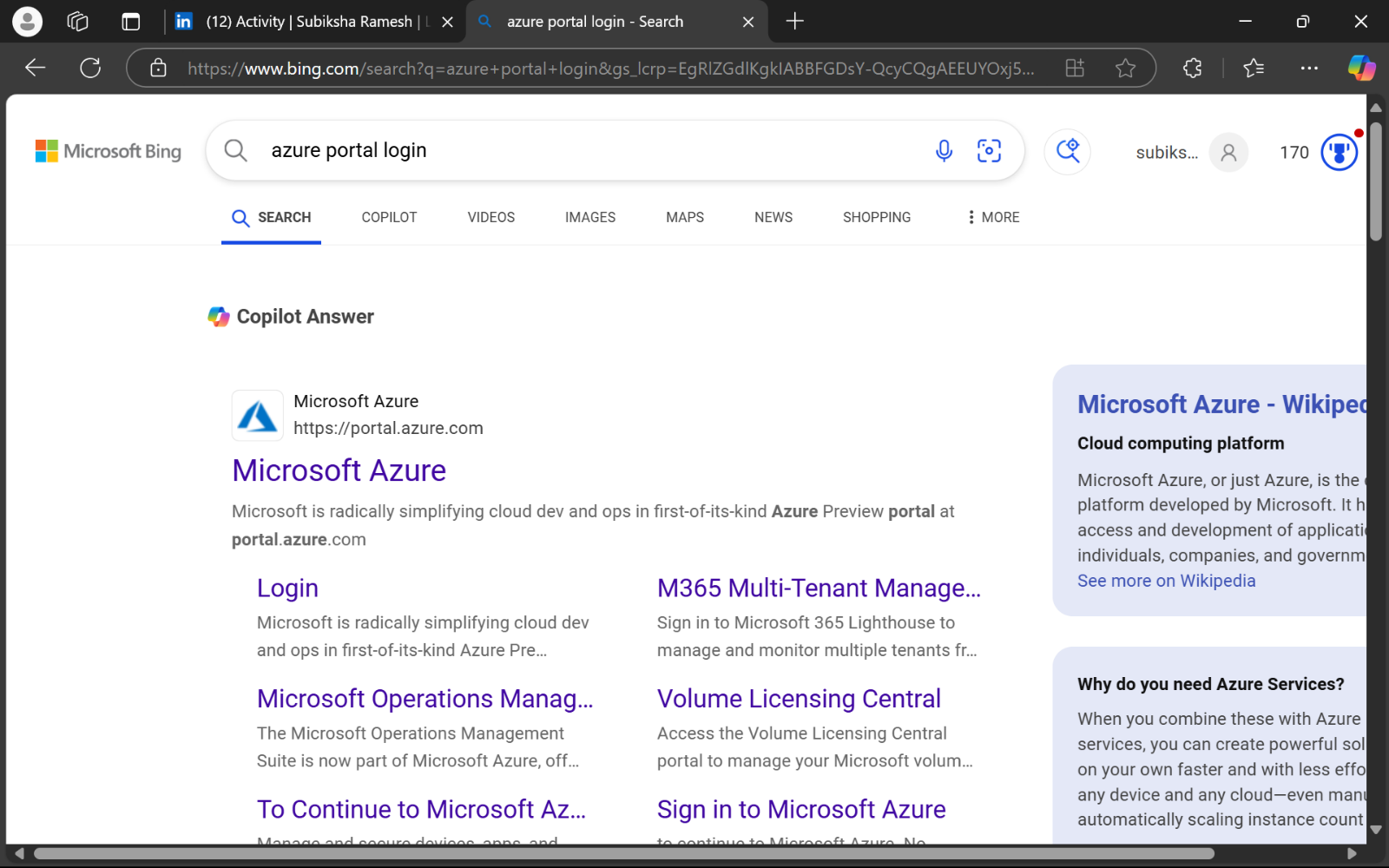
The objective of this task is to set up an Auto Scaling Group using Azure Virtual Machine Scale Sets (VMSS) to:

* Automatically scale virtual machines based on defined performance metrics (e.g., CPU usage).
* Ensure optimal resource allocation and cost efficiency.
* Maintain high availability and resilience for applications in the cloud.

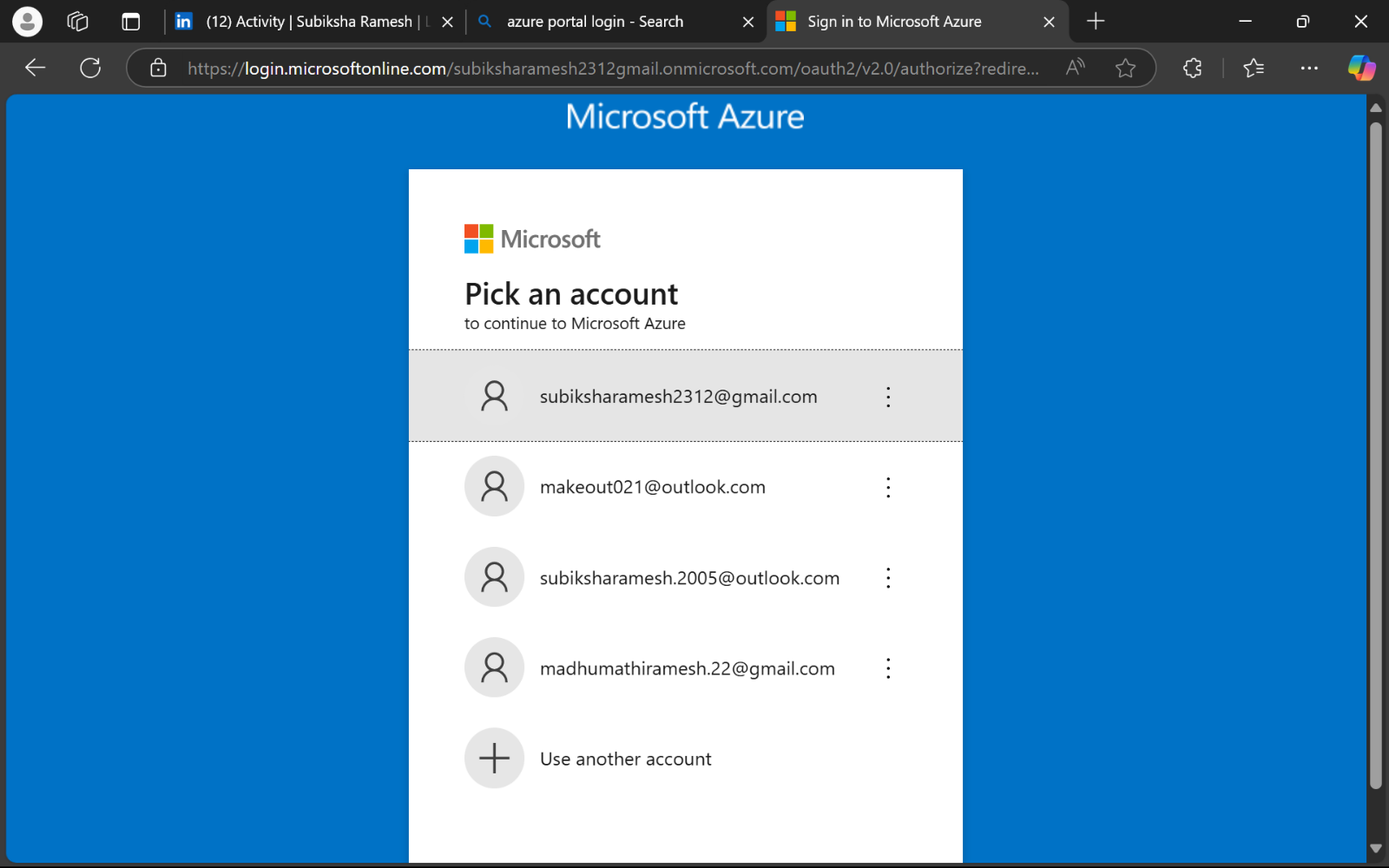
**Step-by-Step procedure :**

**STEP 1 :** Login to Azure Portal

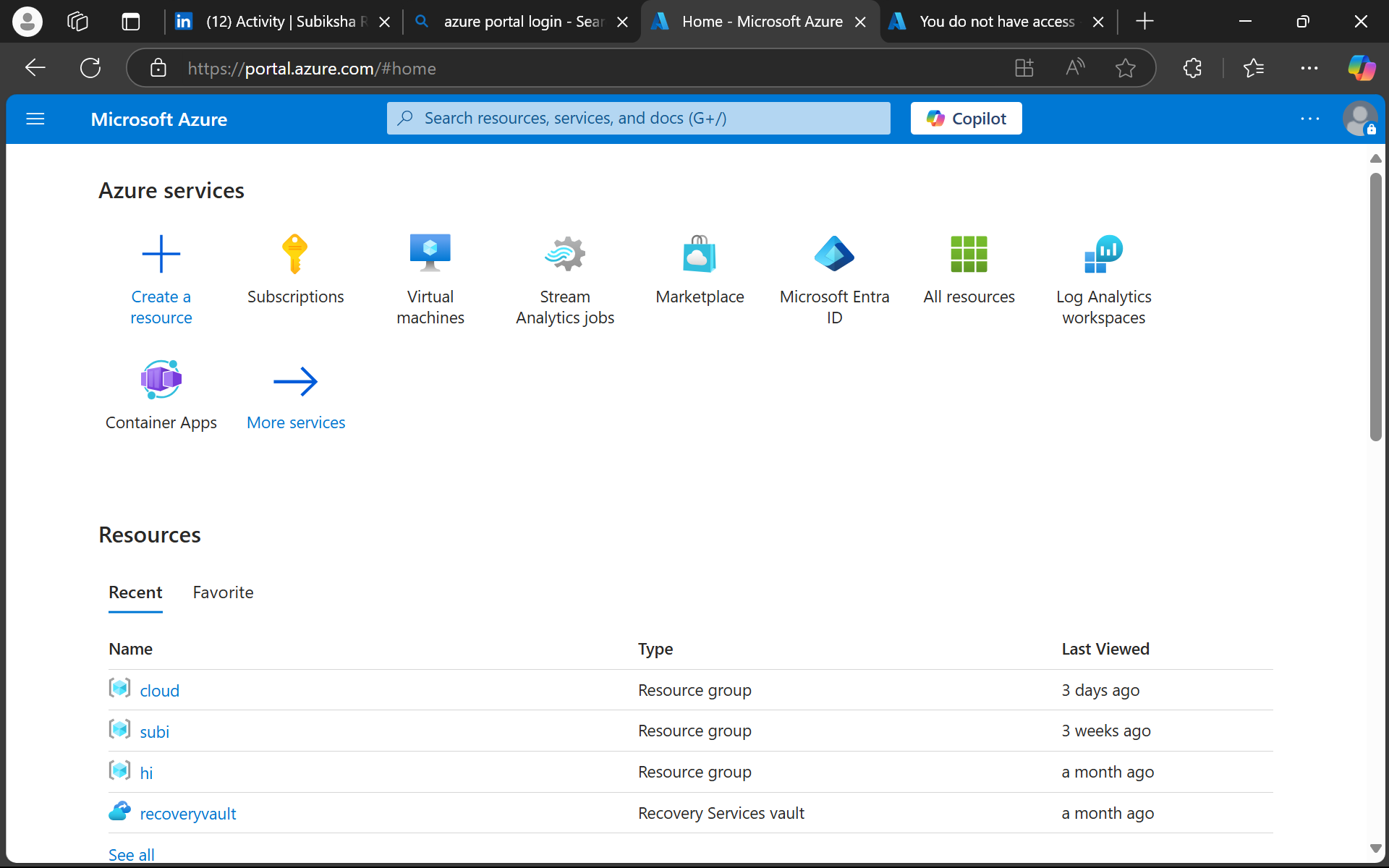
* Go to <https://portal.azure.com>.



**STEP 2 :** Choose the account in which you want to login.



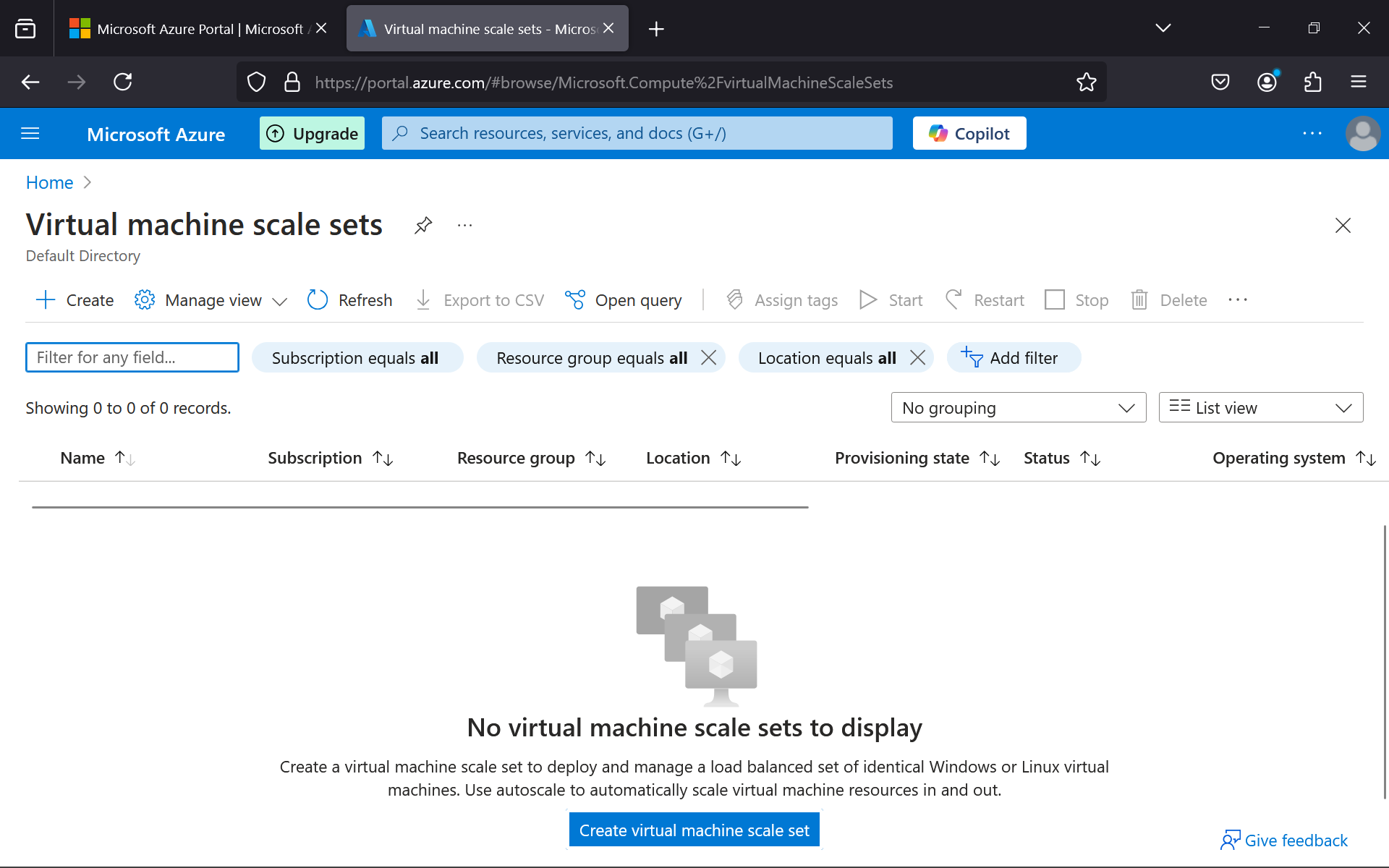
**STEP 3 :** Once logged in, you’ll see the Azure Portal dashboard.



**STEP 4 : Create a Virtual Machine Scale Set (VMSS)**

* From the Azure Portal, search for Virtual Machine Scale Sets in the search bar.
* Click + Create > Virtual Machine Scale Set.

.

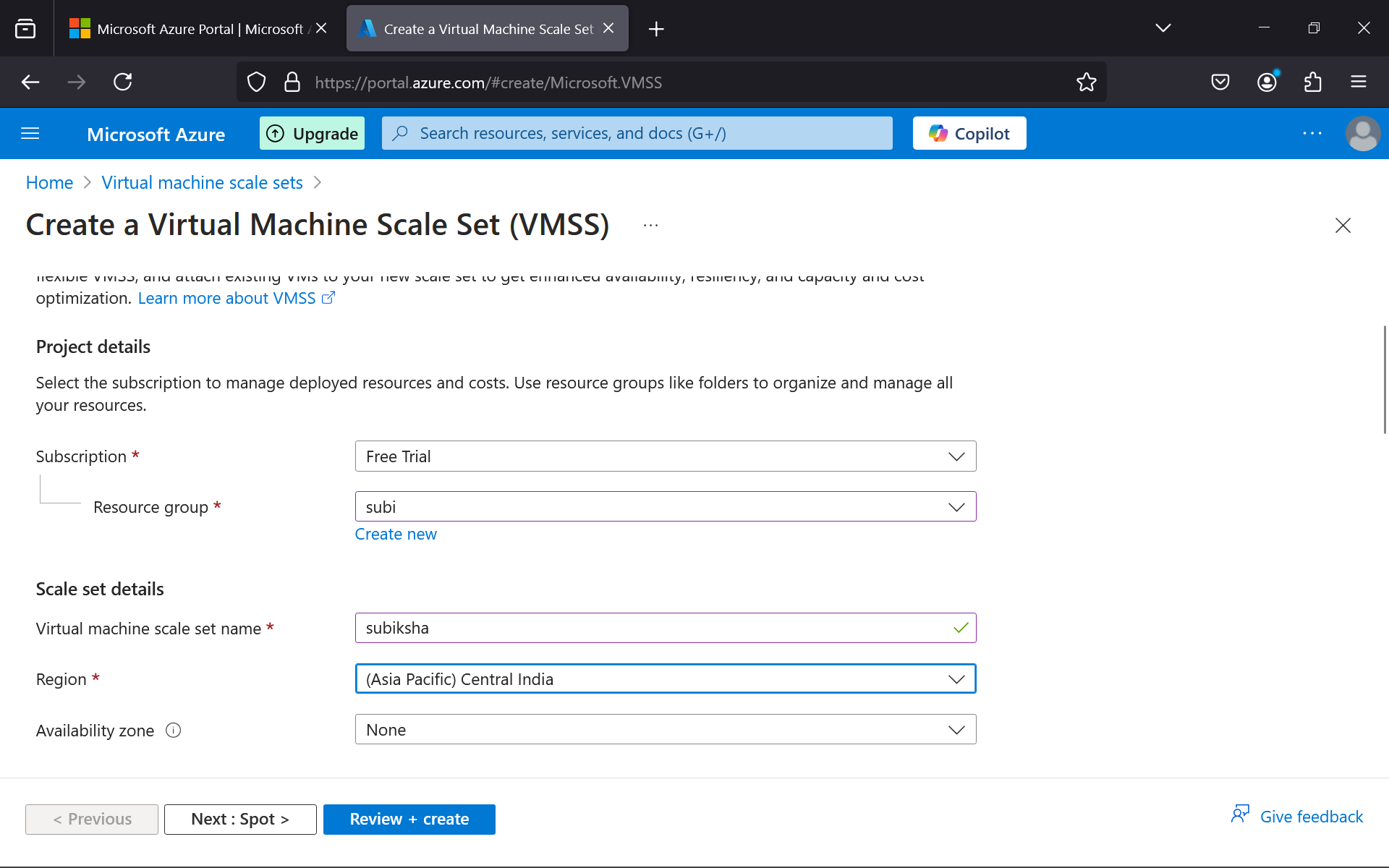
****

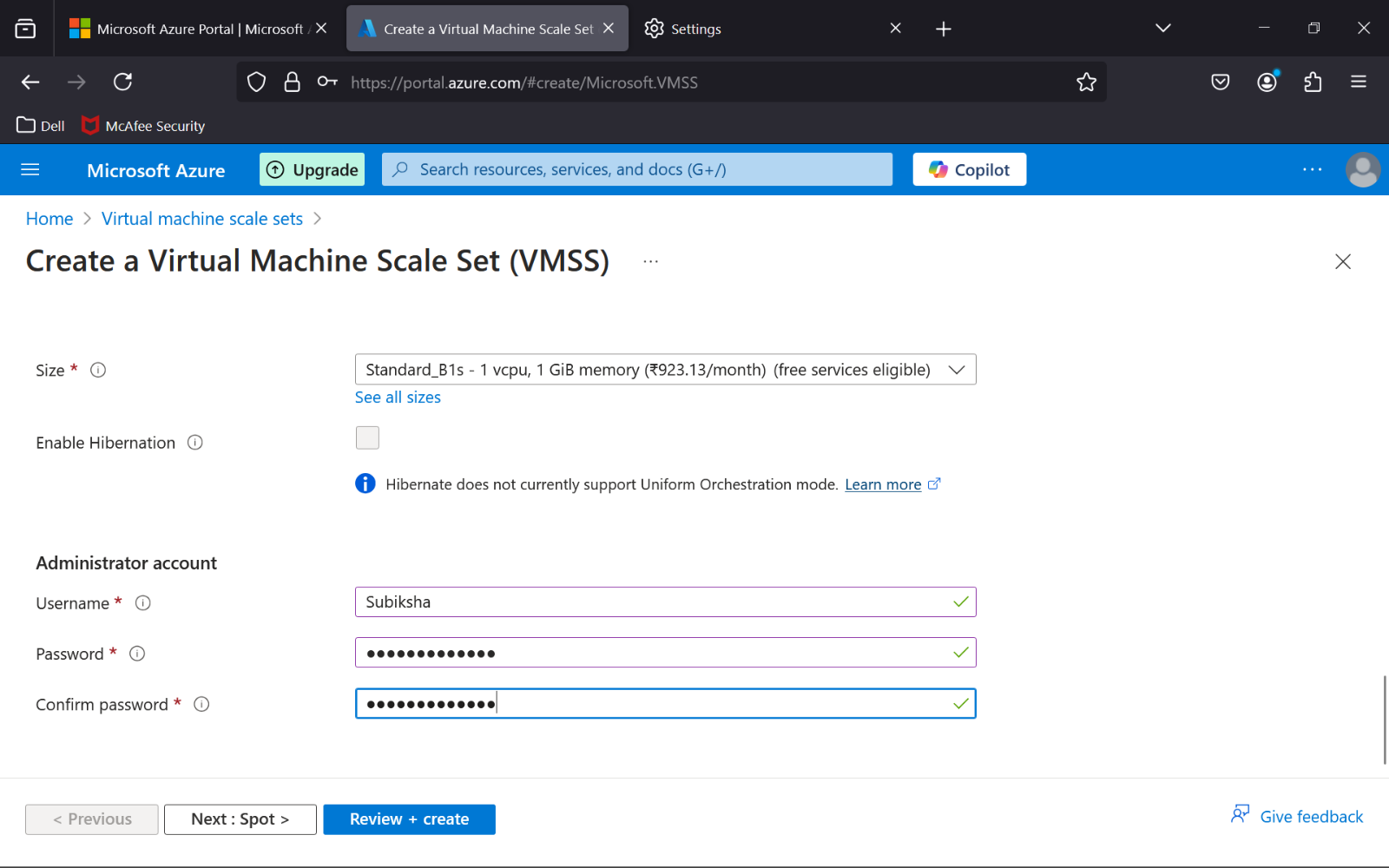
**STEP 5 : Configure Basic Settings**

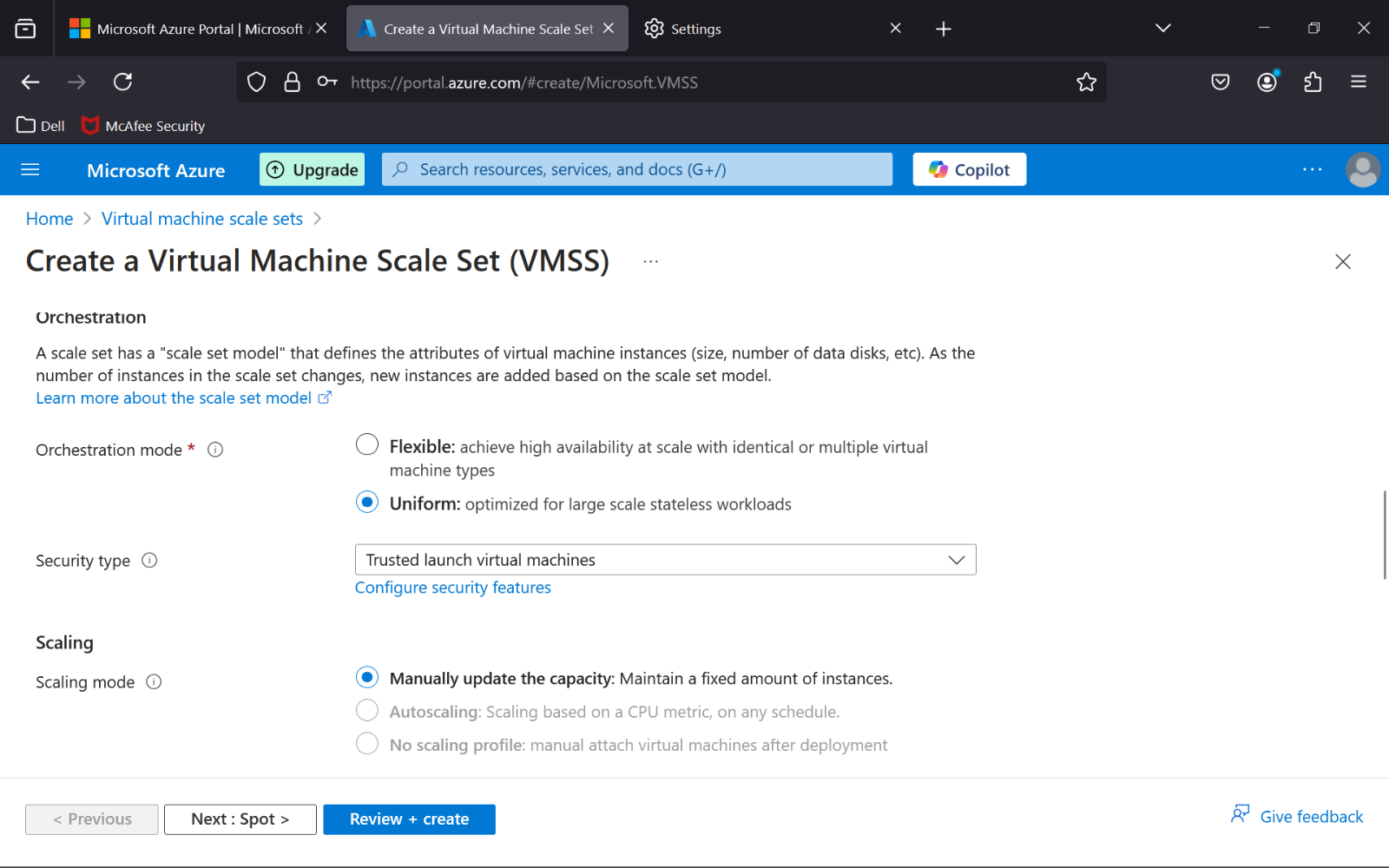
* Subscription: Select your Azure subscription.
* Resource Group: Choose an existing resource group or create a new one.
* Scale Set Name: Give a meaningful name to your scale set (e.g., "AutoScalingApp").
* Region: Choose the region nearest to your users.
* Orchestration Mode: Choose Uniform (for identical VMs).
* Image: Choose the desired operating system (e.g., Ubuntu Server 20.04 LTS or Windows Server).
* Instance Size: Choose an instance size (e.g., Standard\_B1ms for testing).

4. Configure Instance Settings

* Number of Instances: Start with a single instance for initial testing.
* Authentication: Use SSH for Linux VMs or create credentials for Windows VMs.

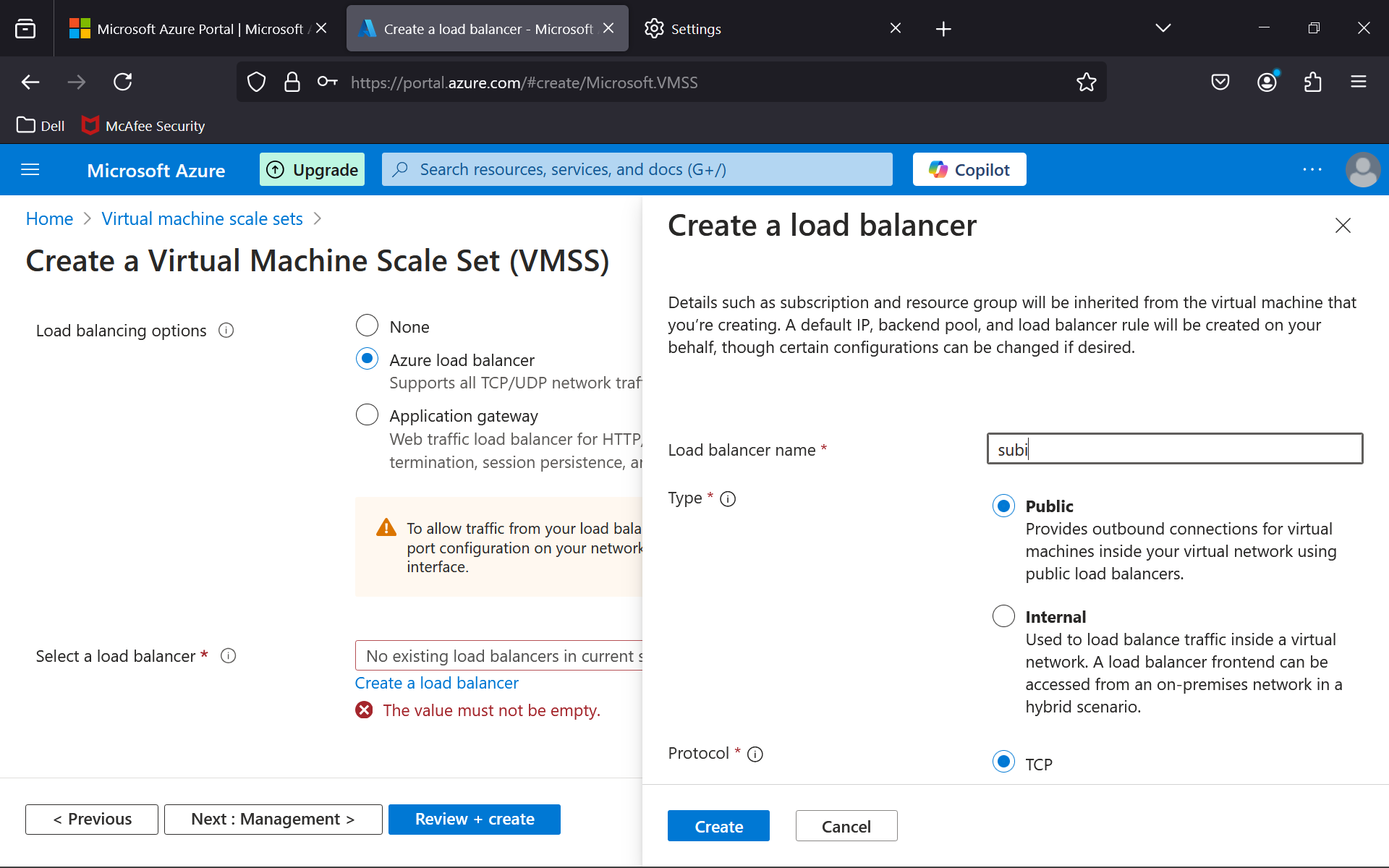






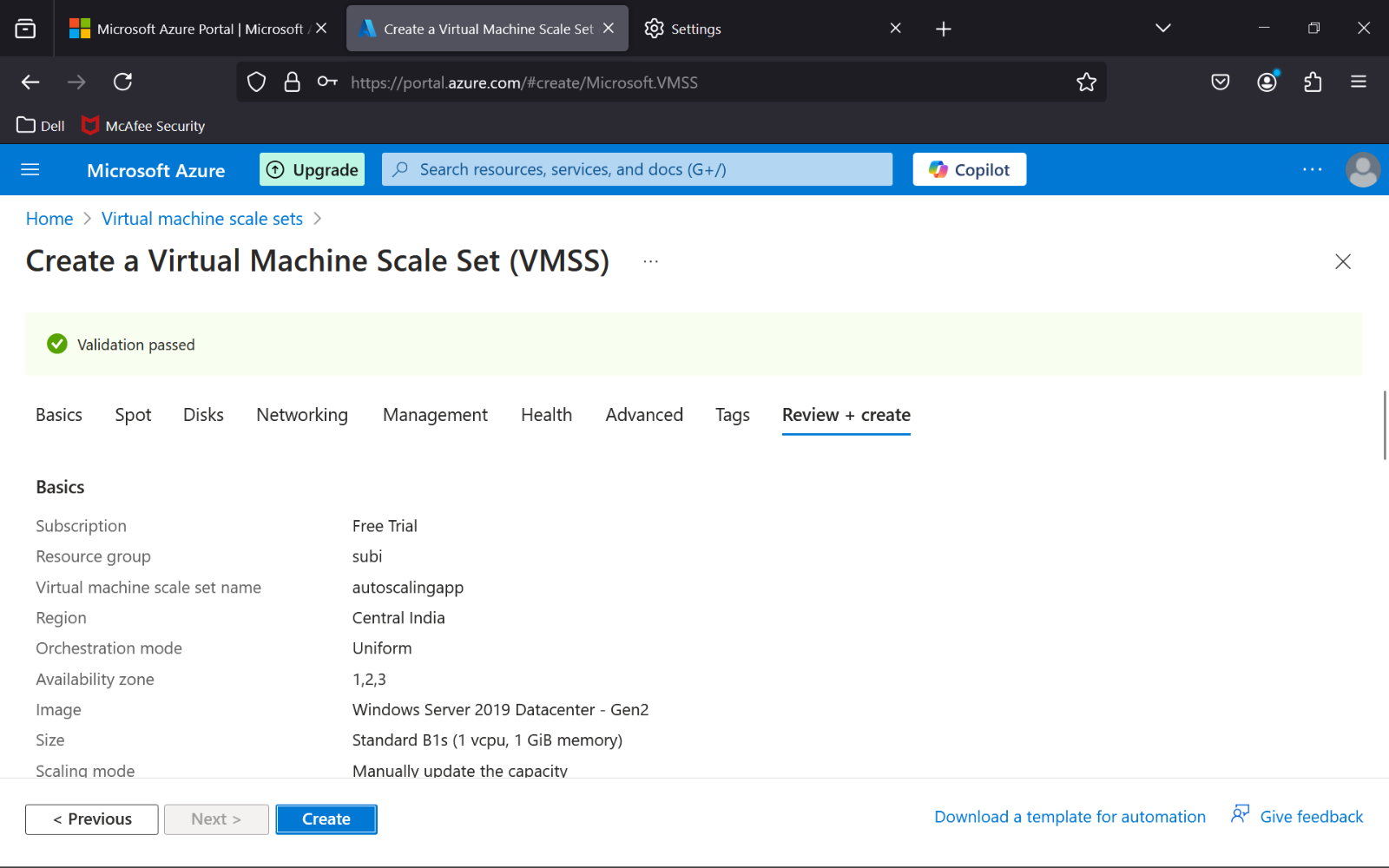
**STEP 6 : Set Load Balancer (Optional but Recommended)**

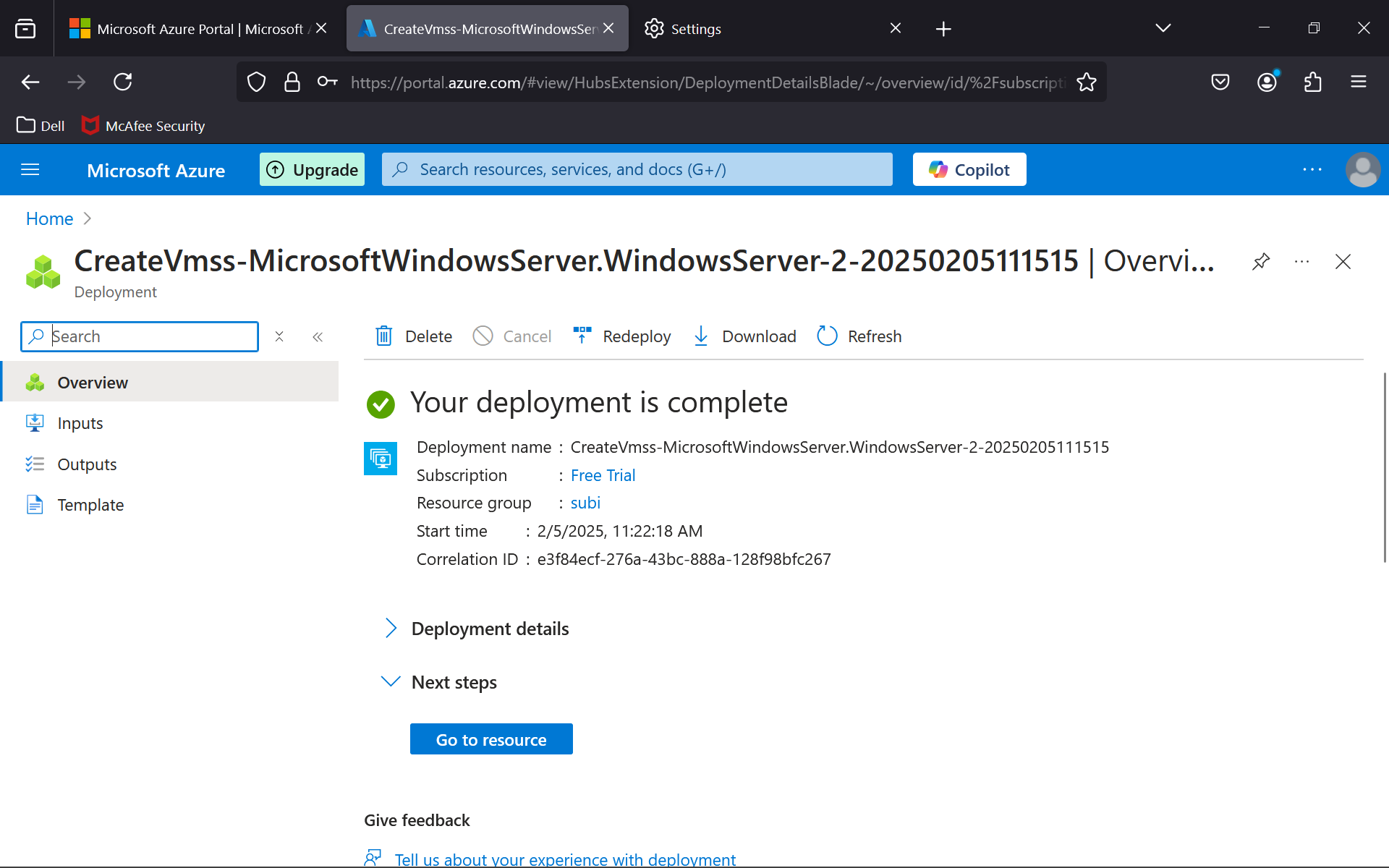
* Attach an **Azure Load Balancer** to your scale set to distribute incoming traffic across the VMs.
* Create a **Load Balancer** and configure it to distribute traffic for your VMSS.

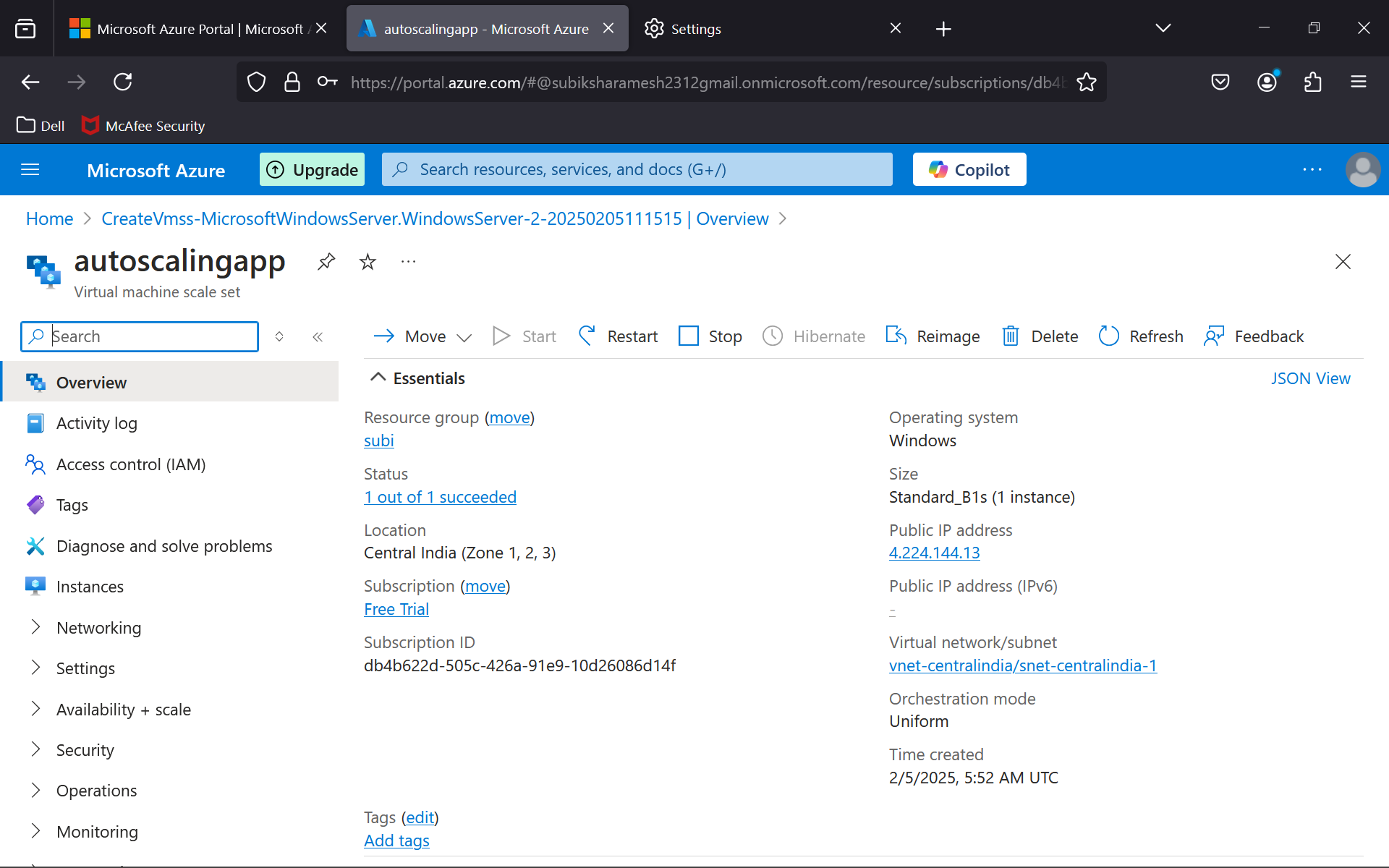


**STEP 7 : Deploy the Scale Set**

* Click **Review + Create** to review your settings.
* After verifying the configuration, click **Create** to deploy the VMSS.

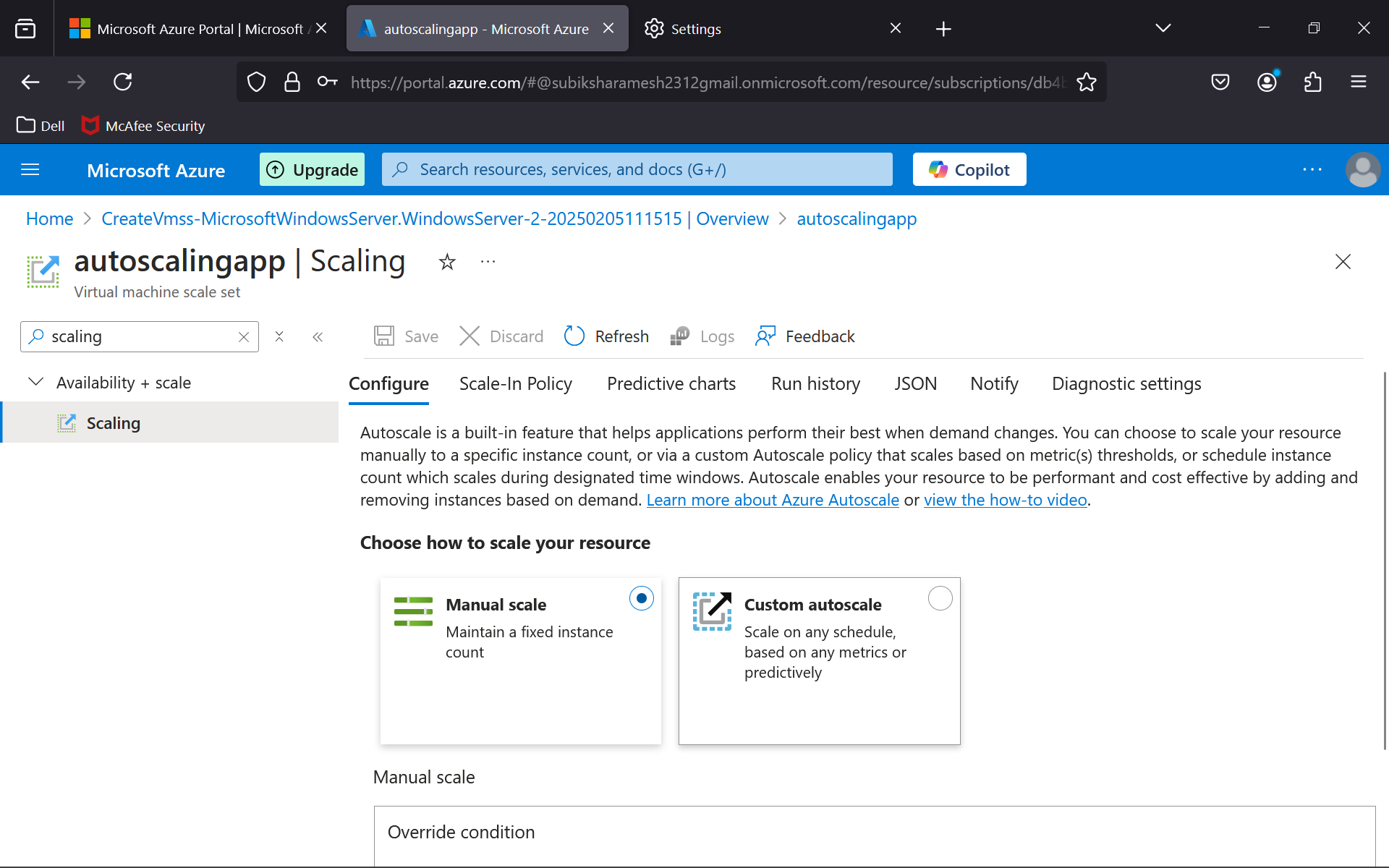


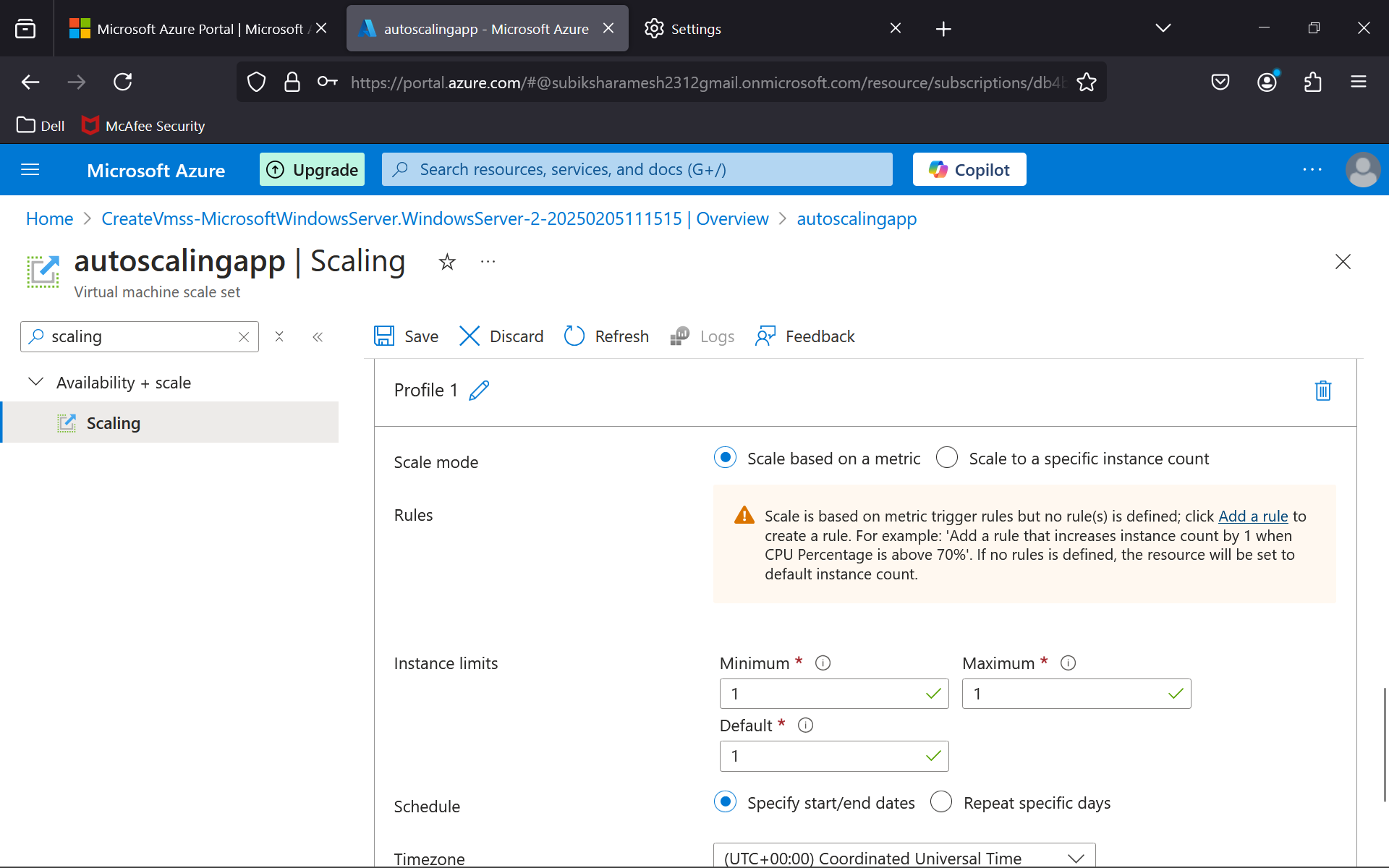
****

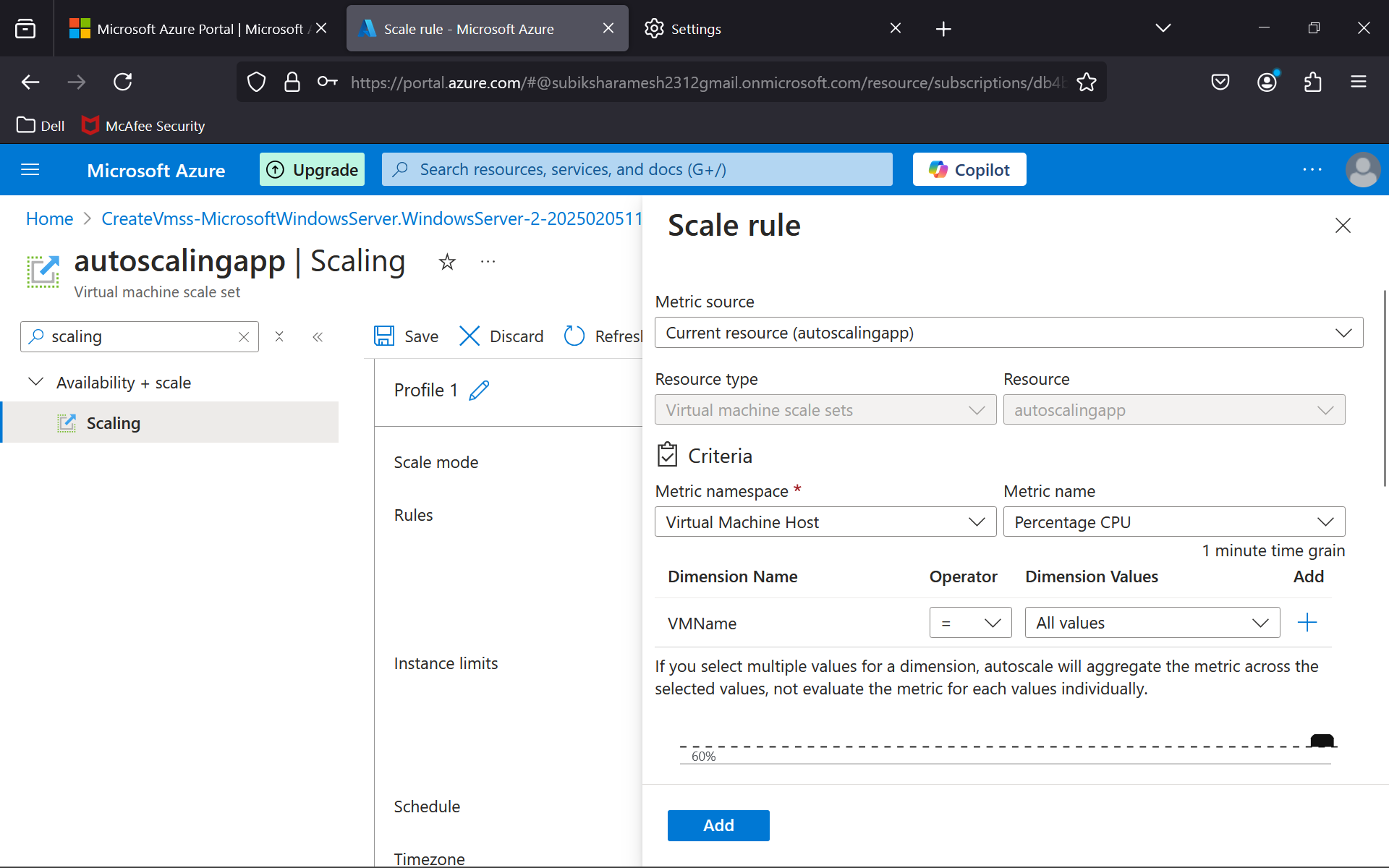
****

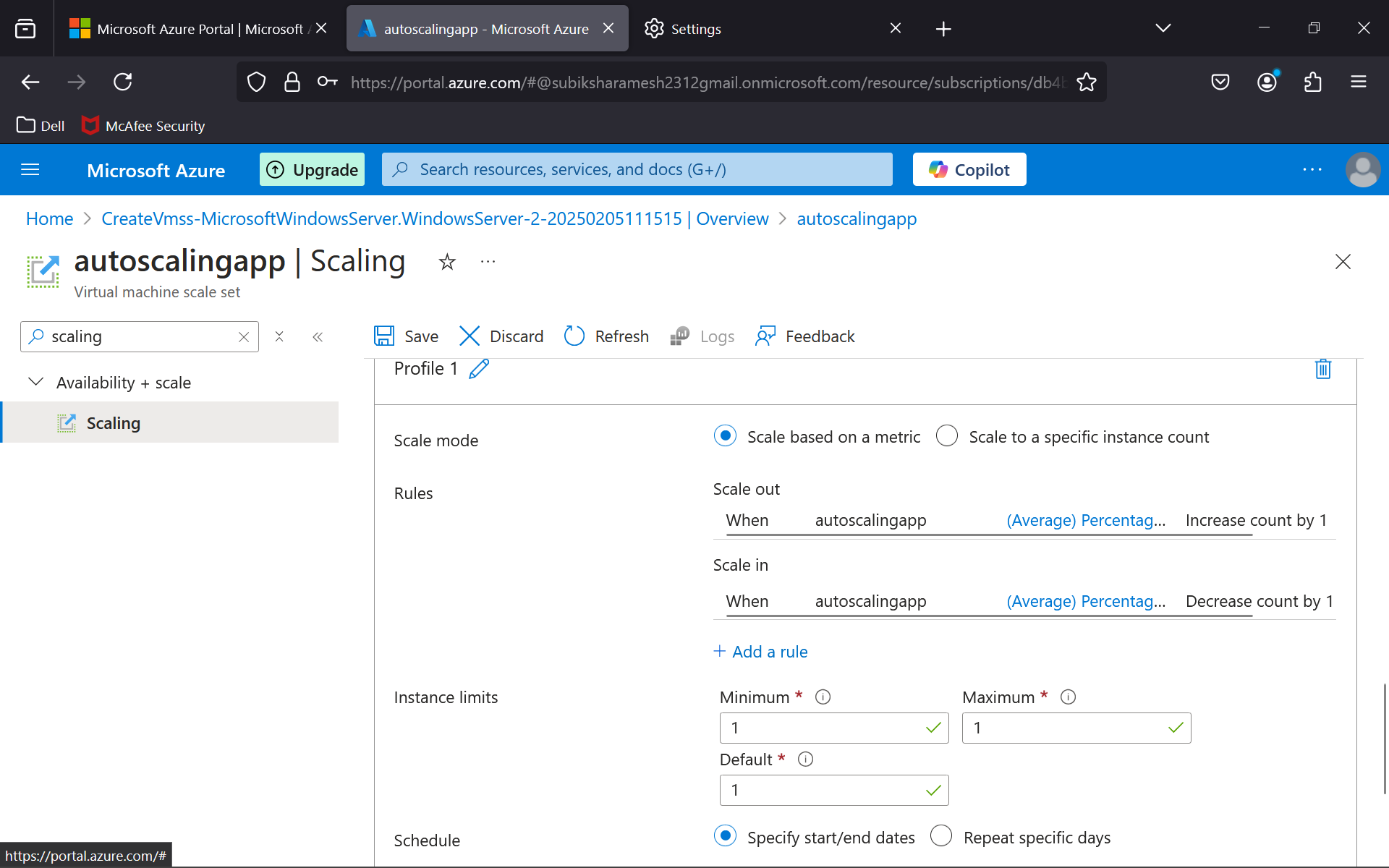
**STEP 8 : Set Up Scaling Policies**

* Choose **Manual Scaling** or configure **Autoscale** directly.
* Under **Scaling > Custom Autoscale**, click **Add Rule**.
* Configure autoscale rules based on a metric (e.g., CPU usage):
* **Metric:** CPU usage
* **Operator:** Greater than
* **Threshold:** 60%
* **Action:** Increase count by 1
* **Cool Down Period:** 5 minutes
* Add another rule to scale down when CPU usage is below 30%.





****

****

**Outcome :**

By following this procedure, you will achieve:

* **Auto-scaling:** The scale set will automatically adjust the number of VMs based on the defined metrics (such as CPU usage), ensuring your application can handle varying workloads.
* **Cost Optimization:** You will only pay for the resources you need at any given time, as the scale set will scale down when demand decreases.
* **High Availability:** The system ensures that enough VMs are running to handle the traffic, even during peak times, providing greater reliability and uptime.
* **Resilience:** If any VM fails, the scale set can quickly replace it, ensuring that there are no disruptions to your application’s availability.