St. JOSEPH'S
GROUP OF INSTITUTIONS
OMR, CHENNAI - 119

aws

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 : Harshana Perianayaki B Department : IT

St. JOSEPH'S COLLEGE OF ENGINEERING | St. JOSEPH'S INSTITUTE OF TECHNOLOGY
AUTONOMOUS INSTITUTIONS, AFFILIATED TO ANNA UNIVERSITY

INTRODUCTION

In today's cloud-driven environment, businesses require scalable and cost-effective solutions to manage fluctuating workloads efficiently. Auto-scaling is a crucial feature in cloud computing that enables applications to dynamically adjust resources based on demand. Cloud platforms, such as Microsoft Azure, AWS, and Google Cloud, provide robust auto-scaling capabilities to ensure optimal performance and cost efficiency.

OVERVIEW

Auto-scaling allows organizations to automatically increase or decrease the number of virtual machines (VMs) within an auto-scaling group based on predefined rules and real-time performance metrics. This ensures that applications can handle varying traffic loads while maintaining high availability and cost-effectiveness. Cloud-based auto-scaling solutions integrate with monitoring tools and load balancers to provide seamless scaling experiences.

OBJECTIVES

- **Efficiently Manage Workloads**: Automatically adjust resources based on real-time demand.
- **Improve Application Performance**: Maintain optimal response times by scaling resources up or down.

- **Reduce Operational Costs**: Avoid over-provisioning and only use necessary resources.
- **Enhance System Reliability**: Minimize downtime by ensuring the availability of sufficient resources during high traffic periods.
- **Automate Scaling Policies**: Utilize predefined metrics such as CPU utilization, memory usage, or custom application triggers for auto-scaling decisions.

IMPORTANCE

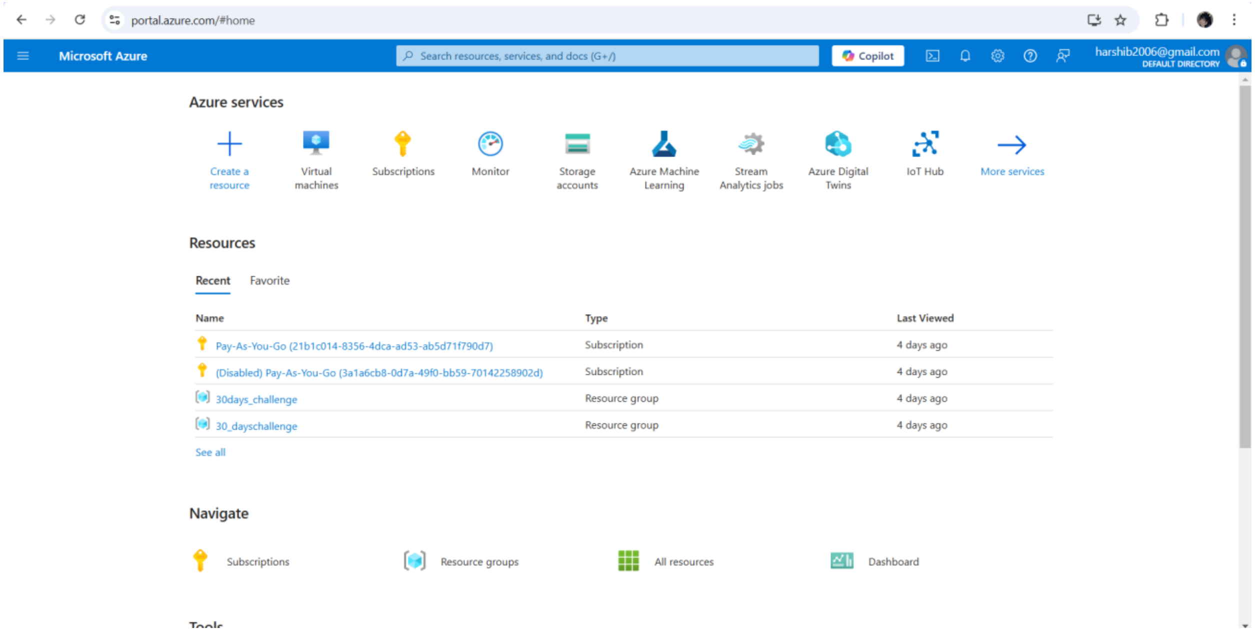
1. **Cost Efficiency** – Auto-scaling ensures that you only pay for the resources you use, reducing unnecessary infrastructure costs.
2. **High Availability** – Ensures applications remain available and responsive by scaling resources during peak loads.
3. **Performance Optimization** – Maintains optimal application performance by adding or removing instances based on demand.
4. **Resource Automation** – Reduces the need for manual intervention, enabling a more automated cloud infrastructure.
5. **Fault Tolerance** – Enhances system reliability by distributing workloads efficiently across multiple instances.

STEP BY STEP OVERVIEW

Setting up an **Auto-scaling Group** in **Microsoft Azure** is done through **Virtual Machine Scale Sets (VMSS)**. This allows you to automatically adjust the number of virtual machines based on demand. Below are the steps to implement auto-scaling in **Azure**:

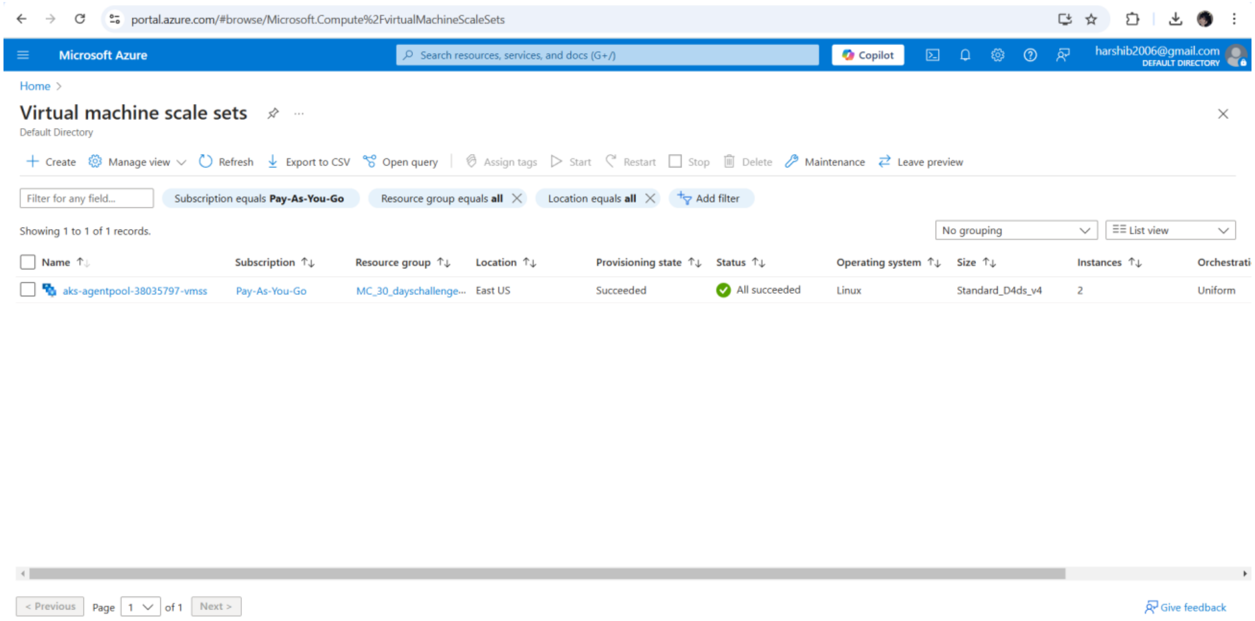
Step 1: Create a Virtual Machine Scale Set (VMSS)

1. **Log in to Azure Portal**:
   Go to **Azure Portal**



Name	Type	Last Viewed
Pay-As-You-Go (21b1c014-8356-4dca-ad53-ab5d71f790d7)	Subscription	4 days ago
(Disabled) Pay-As-You-Go (3a1a6cb8-0d7a-49f0-bb59-70142258902d)	Subscription	4 days ago
30days_challenge	Resource group	4 days ago
30_dayschallenge	Resource group	4 days ago

2. **Navigate to "Virtual Machine Scale Sets"**: Search for **Virtual Machine Scale Sets** in the top search bar and click **Create**.



Subscription equals **Pay-As-You-Go**   Resource group equals **all**   Location equals **all**

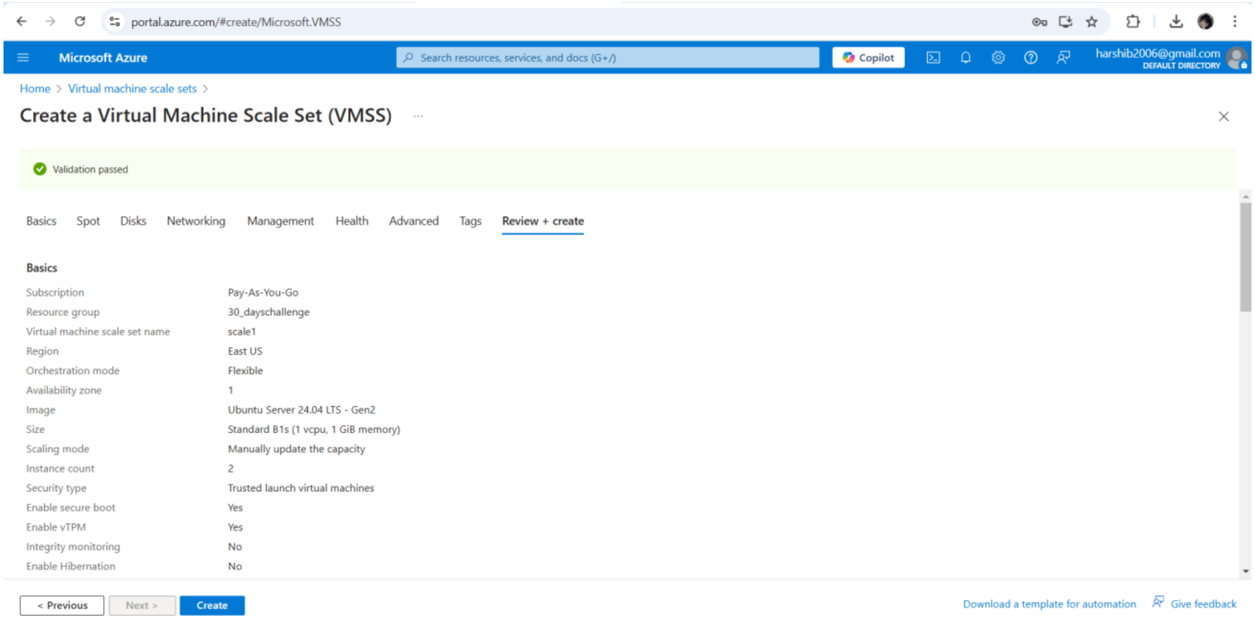
Showing 1 to 1 of 1 records.

Name	Subscription	Resource group	Location	Provisioning state	Status	Operating system	Size	Instances	Orchestrati
aks-agentpool-38035797-vmss	Pay-As-You-Go	MC_30_dayschallenge...	East US	Succeeded	✓ All succeeded	Linux	Standard_D4ds_v4	2	Uniform

3. **Configure Basic Settings**:
    a. Select your **Subscription** and **Resource Group**.
    b. Choose a **Region** where you want the scale set to be deployed.
    c. Set the **Instance details**:
        i. Virtual Machine Scale Set name
        ii. **Orchestration mode**: Choose **Uniform** for standard VM scaling or **Flexible** for greater instance independence.
        iii. Select an appropriate **Image** (e.g., Ubuntu, Windows Server).
        iv. Choose a **VM Size** (e.g., Standard_DS2_v2).
4. **Configure Scaling Policy**:
    a. Choose **Scaling Policy**: Manual or Automatic.
    b. Set the **Minimum**, **Maximum**, and **Default** number of instances.

    c. Select **Enable Autoscale**.

5. **Networking and Security**:
    a. Choose **Virtual Network (VNet)** and **Subnet**.
    b. Enable **Public IP** if needed.
    c. Configure **Load Balancer** (optional but recommended).
6. **Review and Create**:
   Click **Review + Create** and then **Create** to deploy the scale set.


portal.azure.com/#create/Microsoft.VMSS
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
harshib2006@gmail.com
DEFAULT DIRECTORY
Home > Virtual machine scale sets >

Create a Virtual Machine Scale Set (VMSS)

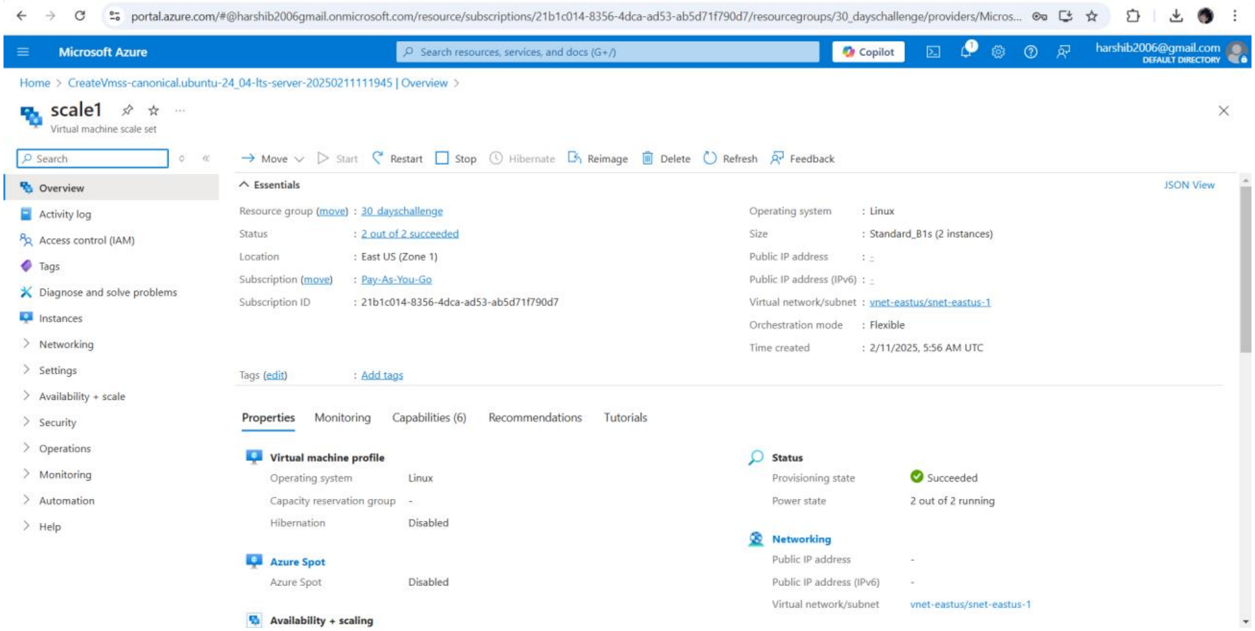
Validation passed

Basics

Subscription	Pay-As-You-Go
Resource group	30_dayschallenge
Virtual machine scale set name	scale1
Region	East US
Orchestration mode	Flexible
Availability zone	1
Image	Ubuntu Server 24.04 LTS - Gen2
Size	Standard B1s (1 vcpu, 1 GiB memory)
Scaling mode	Manually update the capacity
Instance count	2
Security type	Trusted launch virtual machines
Enable secure boot	Yes
Enable vTPM	Yes
Integrity monitoring	No
Enable Hibernation	No

Step 2: Configure Auto-scaling Rules

1. **Go to the Scale Set**:
    a. In the **Azure Portal**, navigate to your **Virtual Machine Scale Set**.



Resource group (move) : 30_dayschallenge
Status : 2 out of 2 succeeded
Location : East US (Zone 1)
Subscription (move) : Pay-As-You-Go
Subscription ID : 21b1c014-8356-4dca-ad53-ab5d71f790d7
Tags (edit) : Add tags

Operating system : Linux
Size : Standard_B1s (2 instances)
Public IP address : -
Public IP address (IPv6) : -
Virtual network/subnet : vnet-eastus/snet-eastus-1
Orchestration mode : Flexible
Time created : 2/11/2025, 5:56 AM UTC

Virtual machine profile

Operating system	Linux
Capacity reservation group	-
Hibernation	Disabled

Azure Spot

Azure Spot	Disabled

Availability + scaling

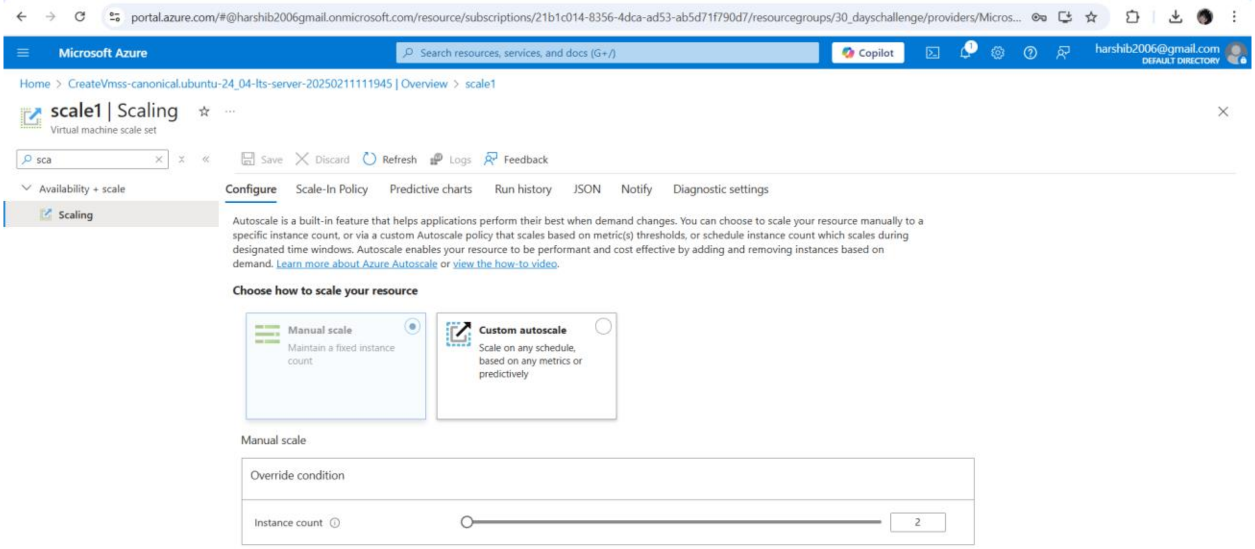
Status

Provisioning state	Succeeded
Power state	2 out of 2 running

Networking

Public IP address	-
Public IP address (IPv6)	-
Virtual network/subnet	vnet-eastus/snet-eastus-1

2. **Configure Autoscale Settings**:
   a. Click on **Scaling**.
   b. Select **Custom Autoscale**.
   c. Click **+ Add a rule** to create scaling conditions.


portal.azure.com/#@harshib2006gmail.onmicrosoft.com/resource/subscriptions/21b1c014-8356-4dca-ad53-ab5d71f790d7/resourcegroups/30_dayschallenge/providers/Micros...
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
harshib2006@gmail.com
DEFAULT DIRECTORY
Home > CreateVmss-canonical.ubuntu-24_04-lts-server-20250211111945 | Overview > scale1
scale1 | Scaling
Virtual machine scale set
sca
Availability + scale
Scaling
Save
Discard
Refresh
Logs
Feedback
Configure
Scale-In Policy
Predictive charts
Run history
JSON
Notify
Diagnostic settings

Autoscale is a built-in feature that helps applications perform their best when demand changes. You can choose to scale your resource manually to a specific instance count, or via a custom Autoscale policy that scales based on metric(s) thresholds, or schedule instance count which scales during designated time windows. Autoscale enables your resource to be performant and cost effective by adding and removing instances based on demand. Learn more about Azure Autoscale or view the how-to video.

Choose how to scale your resource

● Manual scale
Maintain a fixed instance count

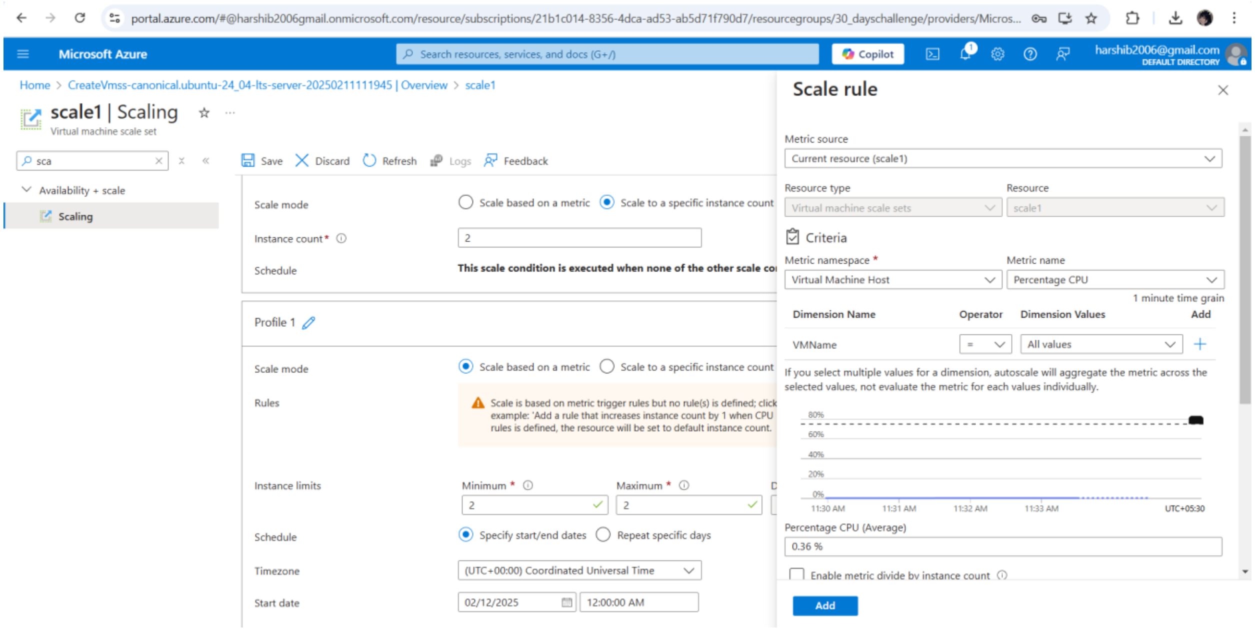
○ Custom autoscale
Scale on any schedule, based on any metrics or predictively

Manual scale

Override condition

Instance count: 2

3. **Define Scaling Triggers**:
    a. **Metric-based Scaling** (Recommended):
        i. Choose **CPU Percentage**, **Memory Usage**, or **Disk IOPS**.
        ii. Set **Thresholds** (e.g., scale out when CPU > 75% for 5 minutes).
        iii. Define **Action**: Increase instances by **1**.
        iv. Similarly, create a **Scale-in rule** (e.g., remove instances when CPU < 30%).
4. **Set Instance Limits**:
    a. **Minimum**: 1 (or as required)
    b. **Maximum**: Define based on expected workload
    c. **Default**: Initial number of instances
5. **Apply and Save**.



portal.azure.com/#@harshib2006gmail.onmicrosoft.com/resource/subscriptions/21b1c014-8356-4dca-ad53-ab5d71f790d7/resourcegroups/30_dayschallenge/providers/Micros...

Home > CreateVmss-canonical.ubuntu-24_04-lts-server-20250211111945 | Overview > scale1

scale1 | Scaling
Virtual machine scale set

Scale mode: ○ Scale based on a metric ● Scale to a specific instance count
Instance count *: 2
Schedule: This scale condition is executed when none of the other scale co

Profile 1

Scale mode: ● Scale based on a metric ○ Scale to a specific instance count
Rules: Scale is based on metric trigger rules but no rule(s) is defined; click example: 'Add a rule that increases instance count by 1 when CPU rules is defined, the resource will be set to default instance count.
Instance limits: Minimum *: 2 Maximum *: 2
Schedule: ● Specify start/end dates ○ Repeat specific days
Timezone: (UTC+00:00) Coordinated Universal Time
Start date: 02/12/2025 12:00:00 AM

Scale rule

Metric source: Current resource (scale1)
Resource type: Virtual machine scale sets
Resource: scale1

Criteria

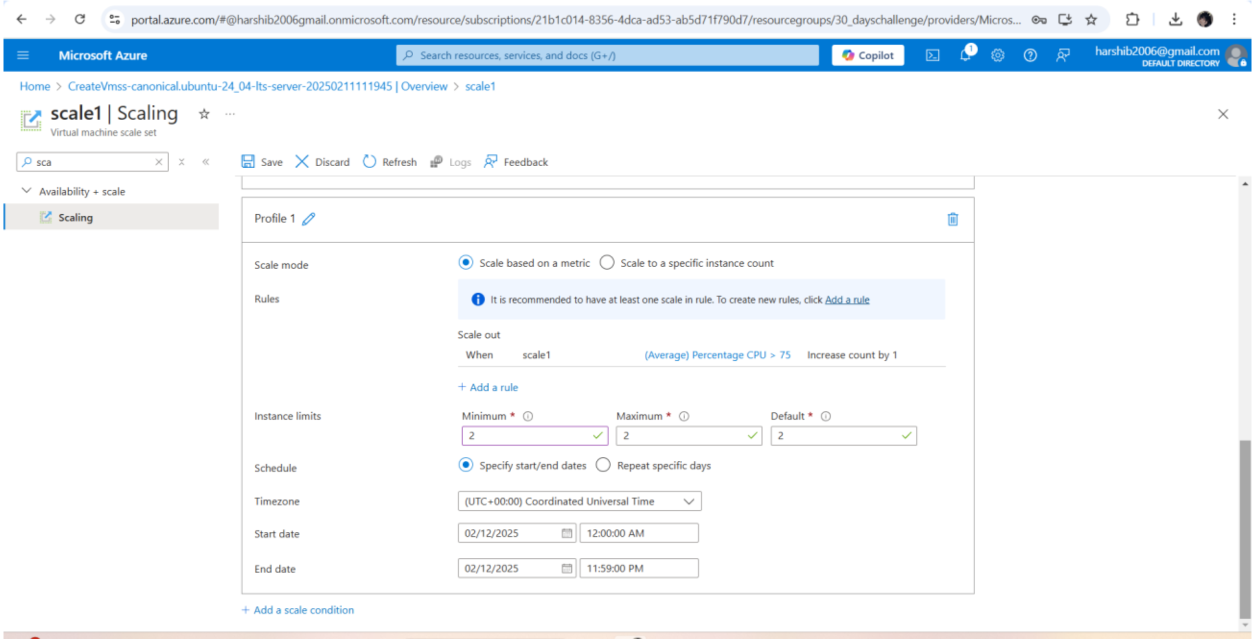
Metric namespace *: Virtual Machine Host
Metric name: Percentage CPU
1 minute time grain

Dimension Name	Operator	Dimension Values	Add
VMName	=	All values	+

If you select multiple values for a dimension, autoscale will aggregate the metric across the selected values, not evaluate the metric for each values individually.

Percentage CPU (Average): 0.36 %

□ Enable metric divide bv instance count


portal.azure.com/#@harshib2006gmail.onmicrosoft.com/resource/subscriptions/21b1c014-8356-4dca-ad53-ab5d71f790d7/resourcegroups/30_dayschallenge/providers/Micros...
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
harshib2006@gmail.com
DEFAULT DIRECTORY
Home > CreateVmss-canonical.ubuntu-24_04-lts-server-20250211111945 | Overview > scale1
sca
Availability + scale
Scaling
Save
Discard
Refresh
Logs
Feedback
+ Add a rule
+ Add a scale condition

scale1 | Scaling

Virtual machine scale set

Profile 1

Scale mode: ● Scale based on a metric ○ Scale to a specific instance count

Rules: It is recommended to have at least one scale in rule. To create new rules, click Add a rule

Scale out

When	scale1	(Average) Percentage CPU > 75	Increase count by 1

Instance limits: Minimum * 2 | Maximum * 2 | Default * 2

Schedule: ● Specify start/end dates ○ Repeat specific days

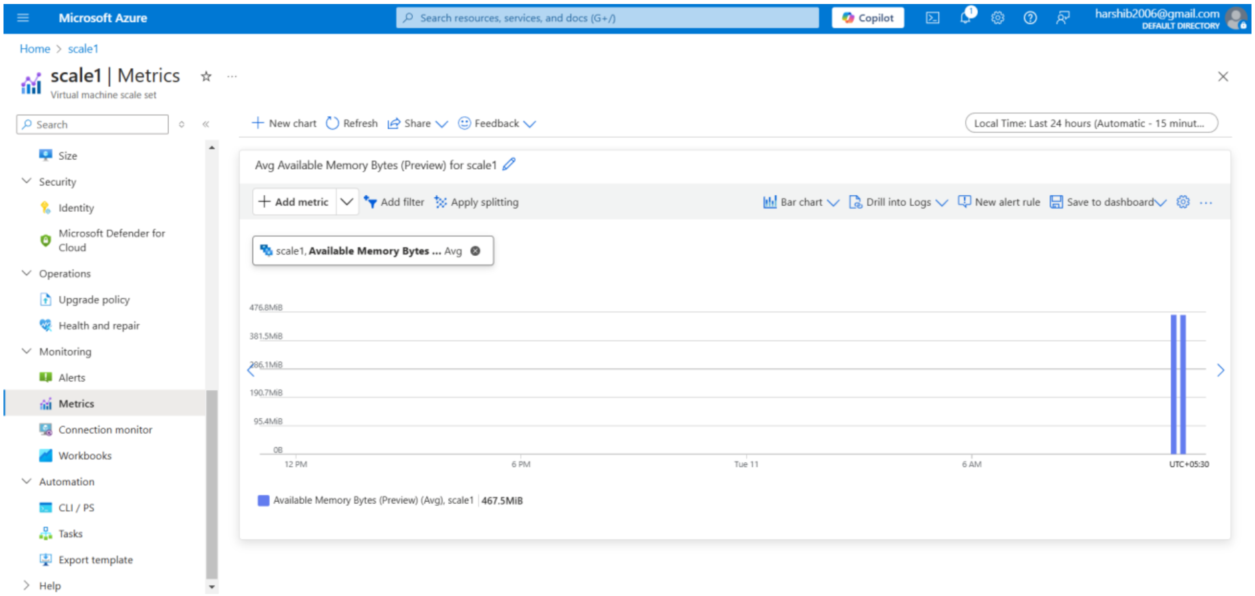
Timezone: (UTC+00:00) Coordinated Universal Time

Start date: 02/12/2025 12:00:00 AM

End date: 02/12/2025 11:59:00 PM

Step 3: Test the Auto-scaling Setup

1. **Generate Load**: Use a load-testing tool like **Apache JMeter**, **Azure Load Testing**, or **locust.io**.
2. **Monitor Scaling**:
   a. Go to **Azure Monitor** > **Metrics**.
   b. Check if new instances are created when demand increases.



3. **Verify Scale-in**:
    a. Reduce load and check if instances are removed.

OUTCOME

By implementing auto-scaling in the cloud, businesses can achieve increased efficiency, optimized resource utilization, and reduced costs. Applications remain highly available and responsive to fluctuating workloads without human intervention. This approach ensures sustainable infrastructure management, paving the way for enhanced business continuity and scalability in a cloud-native environment.