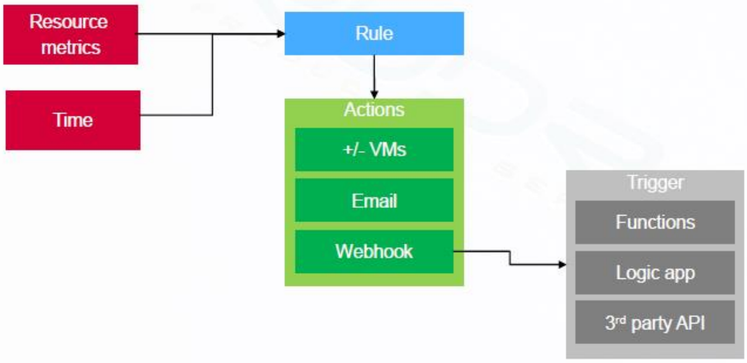
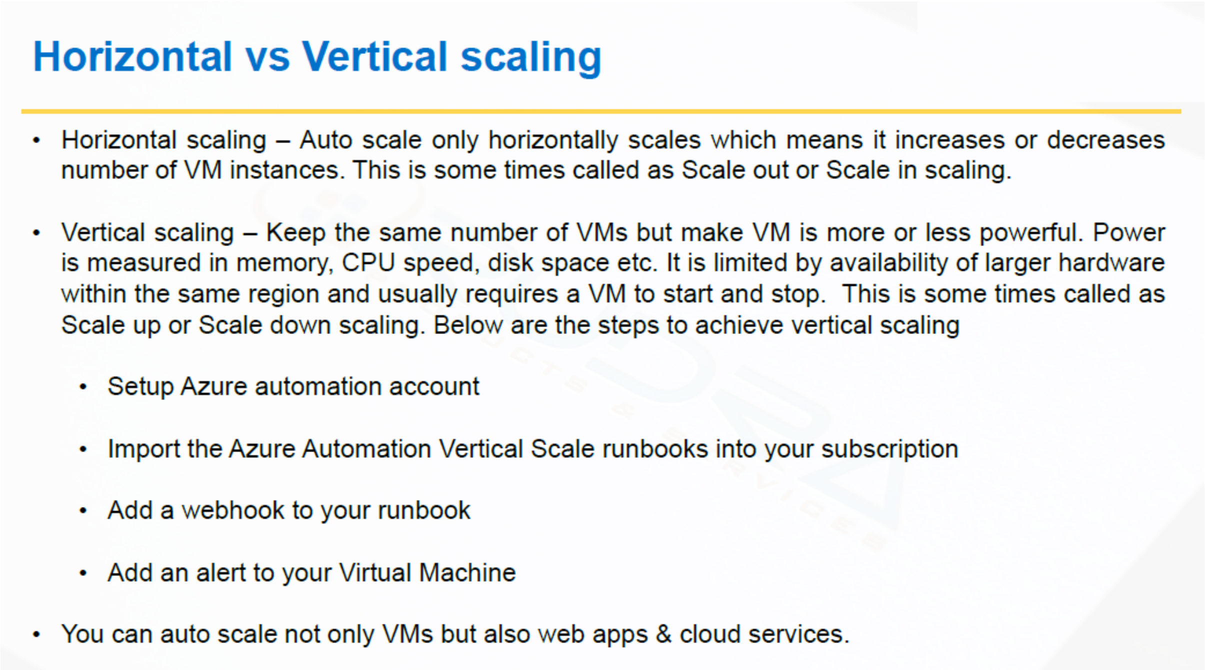
Virtual Machine scale sets

- Virtual Machine scale sets are an Azure compute resource that can be used to deploy and manage identical VMs. They are designed to support virtual machine auto scaling.
- VM Scale sets can be created using Azure portal, JSON templates and REST APIs.
- To increase or decrease number of VMs in the scale set, change the capacity property and redeploy the template.
- A VM scale set is created inside VNET and individual VMs in the scale set are not allocated with public IP addresses.

Auto scale

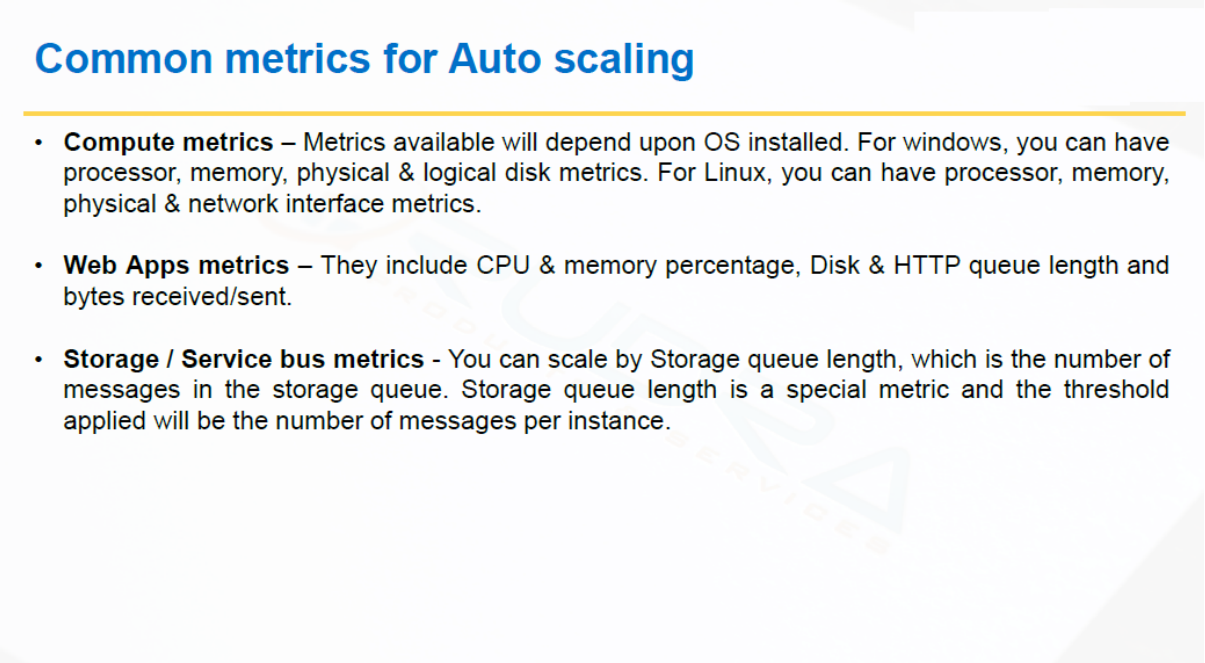
- Auto scale enables you to dynamically allocate or remove resources based on the load on the services. You can specify the maximum and minimum number of instances to run and add or remove VMs based on a set of rules within the range.





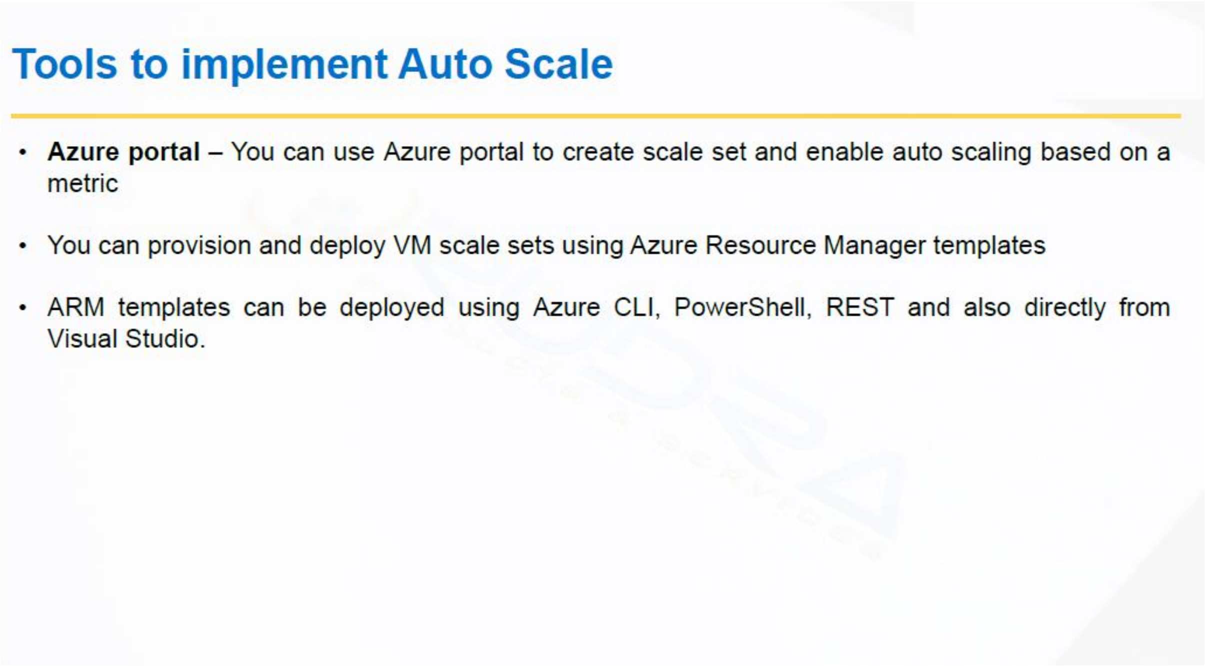
Horizontal vs Vertical scaling

- Horizontal scaling – Auto scale only horizontally scales which means it increases or decreases number of VM instances. This is some times called as Scale out or Scale in scaling.
- Vertical scaling – Keep the same number of VMs but make VM is more or less powerful. Power is measured in memory, CPU speed, disk space etc. It is limited by availability of larger hardware within the same region and usually requires a VM to start and stop. This is some times called as Scale up or Scale down scaling. Below are the steps to achieve vertical scaling
  - Setup Azure automation account
  - Import the Azure Automation Vertical Scale runbooks into your subscription
  - Add a webhook to your runbook
  - Add an alert to your Virtual Machine
- You can auto scale not only VMs but also web apps & cloud services.



Common metrics for Auto scaling

- **Compute metrics** – Metrics available will depend upon OS installed. For windows, you can have processor, memory, physical & logical disk metrics. For Linux, you can have processor, memory, physical & network interface metrics.
- **Web Apps metrics** – They include CPU & memory percentage, Disk & HTTP queue length and bytes received/sent.
- **Storage / Service bus metrics** - You can scale by Storage queue length, which is the number of messages in the storage queue. Storage queue length is a special metric and the threshold applied will be the number of messages per instance.



Tools to implement Auto Scale

- **Azure portal –** You can use Azure portal to create scale set and enable auto scaling based on a metric
- You can provision and deploy VM scale sets using Azure Resource Manager templates
- ARM templates can be deployed using Azure CLI, PowerShell, REST and also directly from Visual Studio.