Designing for Azure Autoscale

Jeff Landry
AUTHOR

Overview

Benefits of using autoscale:

- Positive financial impact
- Reduced manual operations

Key differences between:

- Vertical scaling
- Horizontal scaling

Demo: Scale up a virtual machine

Overview

Autoscaling rules and automation

Rules are based on metrics:

- CPU usage
- Memory usage

Creating a virtual machine scale set

- Provides high availability
- Improved performance

Enable autoscale on a VM scale set

Overview

Autoscaling with Azure App Service

- Similar to autoscaling a VM scale set

Scalability using Azure functions

- Serverless applications

Understanding Vertical Scaling

Older cluster technologies

- Could be complicated
- Time consuming
- Can lead to human mistakes

Virtual machines

- No more physical components
  - Drivers
  - Compatibility lists

Elastic computing

- Automation of your deployments

What is autoscale?

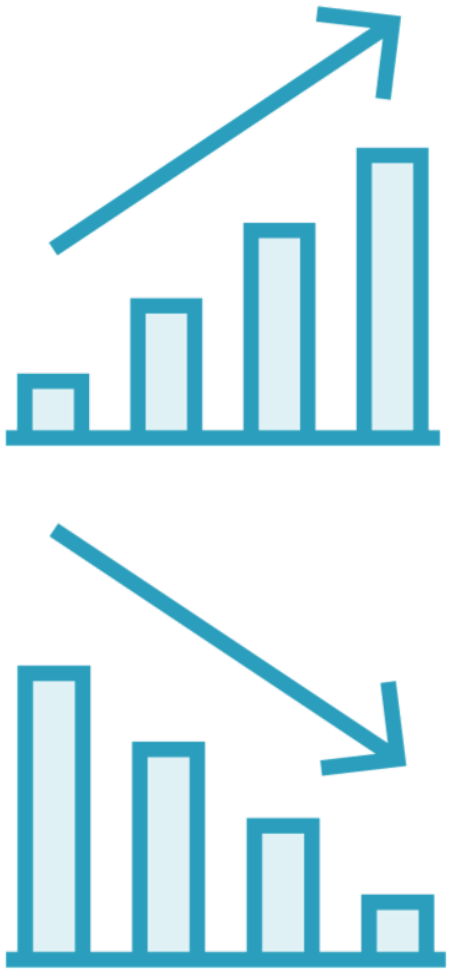
Even distribution of resources to your apps

New resources are added or removed

- When heavy workload occurs
- When resources are idle

Enhanced user experience

Can be modified anytime



Scaling works in two ways

- Vertical scaling
- Horizontal scaling

Vertical scaling

- Typically used with virtual machines
  - Scale up
  - Scale down

Examples of what can be scaled:

- Memory
- CPU
- Disk space

Limitations of Scaling Up or Down

Dependent on available hardware

May vary from one region to another

Must restart the virtual machine

Scaling Other Azure Services

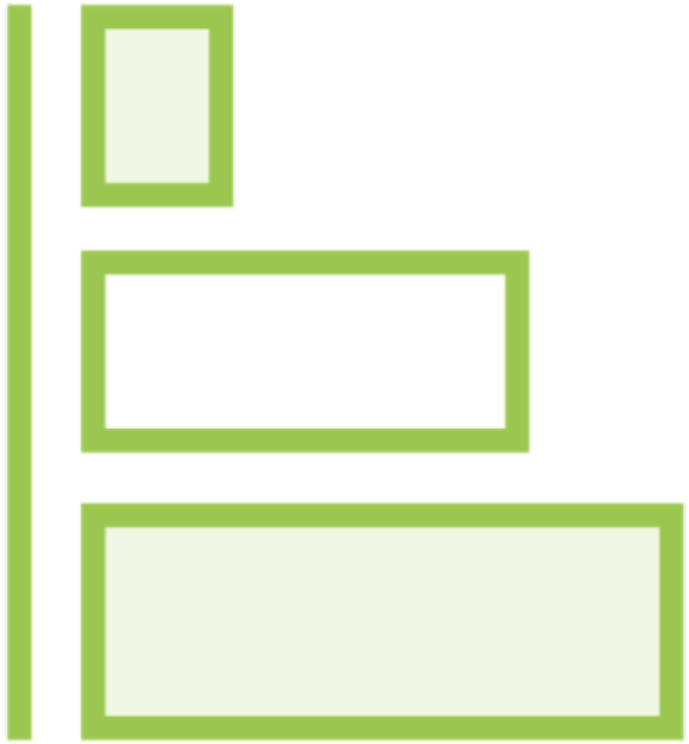
Azure SQL

Scale the plan for the DTUs – Data Transaction Units

Azure App Service

Scale the service plan to better fit the needs

Understanding Horizontal Scaling



Horizontal scaling is more flexible

- You do not scale up or down
- Scale out = Increase
- Scale in = Decrease

Scaling is done based on metrics

- CPU
- Memory

No service disruption during the operation

Scaling is done using autoscaling rules

Autoscale Rule Example

Scale out when average CPU > 80%

For a period of 10 minutes

Autoscale Rule Example

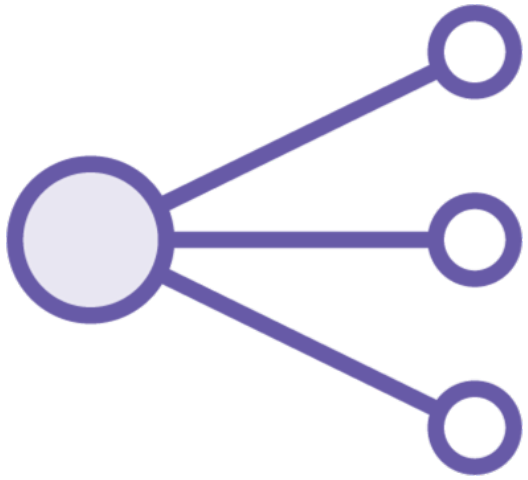
Scale in when average CPU < 40%

40%

For a period of 15 minutes

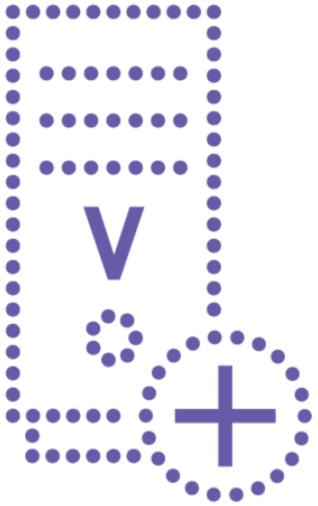
15 min

Load Balancer



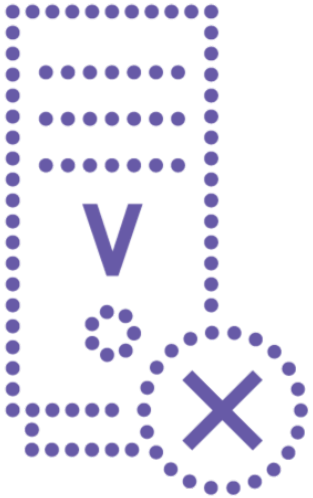
Load balancer

Even distribution of traffic to resources



Scale out

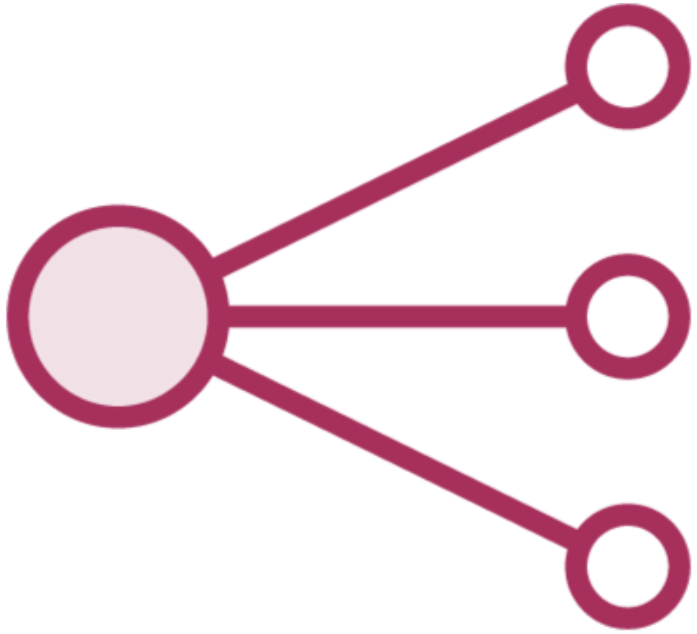
When a new virtual machine is added



Scale in

When a new virtual machine is removed

Designing Autoscaling Rules



Autoscaling rules

- Right amount of resources
- Supports load for the applications
- Flexibility

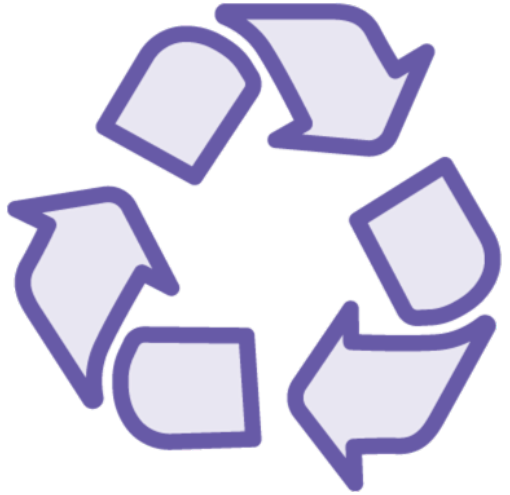
Configuring autoscaling rules

- Minimum virtual machine count
- Maximum virtual machine count
- Default virtual machine count

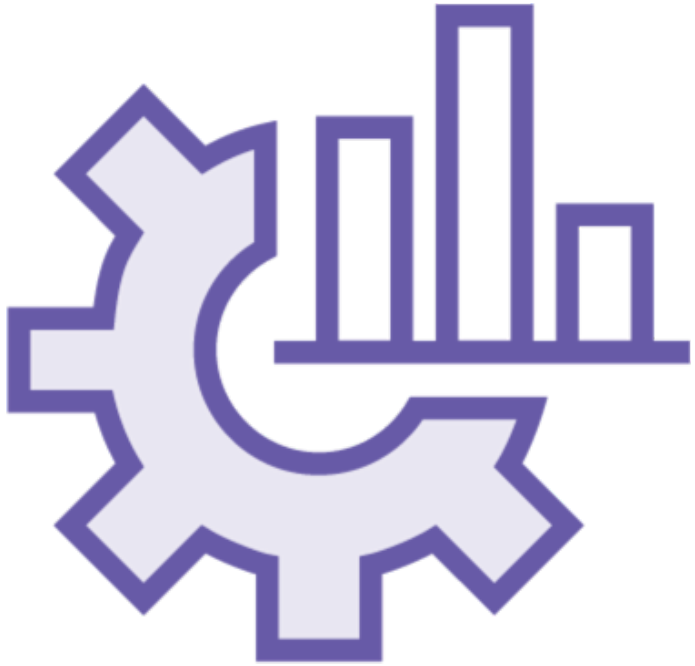
Autoscale Profiles

Regular profile

Fixed date profile



Recurrence profile



Available metrics

- Disk usage
- CPU
- Memory

Network metrics

- BytesReceived
- BytesSent

Many possibilities to choose from

Autoscaling Rules between Multiple Profiles

Fixed date profile

Recurrence profile

Regular profile

Notifications

Monitor autoscale actions in Azure

- Email sent to a distribution group
- Webhooks to send alerts

Monitor section in Azure portal

- Centralized information

Azure Functions Hosting Plans

Azure functions

- Serverless infrastructure

Develop and execute code

- Serverless applications
- No server management

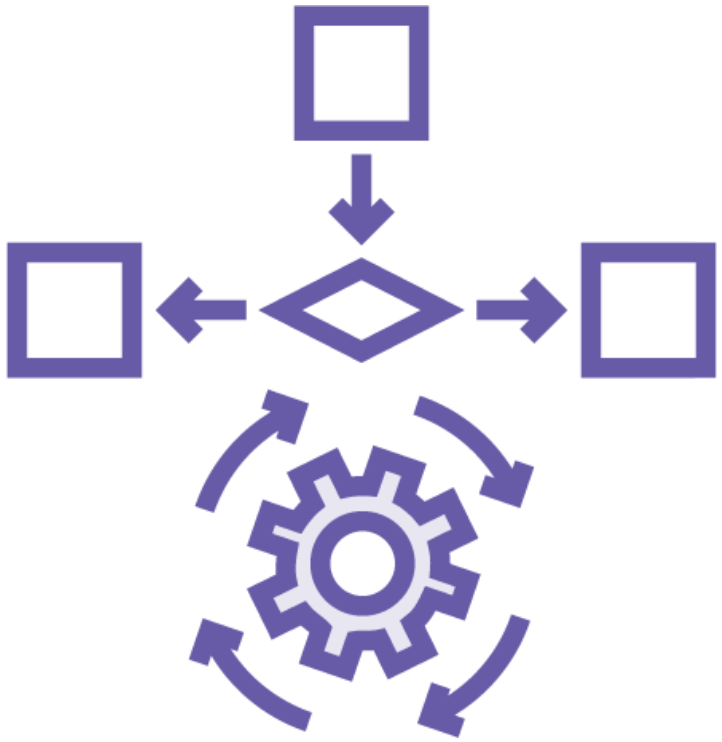
Functions are triggered by events

- Schedule
- Responding to a message
- Responding to an HTTP request

Examples of Azure Functions

Processing bulk data

Internet-of-things (IoT)



Dynamically scale applications

- Orchestrating tasks
- Automation

Building an Azure function app

- Similar to other resources
- Resource group
- Windows or Linux
- Hosting plan

Hosting Plans

Consumption plan

Default
Pay for what you use

Premium plan

“Pre-warmed” instances
Pay for all instances

App Service plan

Similar to web apps
Use existing plan

Summary

Vertical scaling

- Typically done with virtual machines
- Scale up CPU or memory
- App Service and Azure SQL plans

Manual process

- Service downtime is to be expected

Summary

Horizontal scaling

- No downtime during the operation
- Use only the resources you need
- Positive financial impact

Typical use is virtual machine scale sets

Provides for high availability

- Small or large-scale services

Summary

Autoscaling profiles

From the Azure portal:

- A VM scale set deployment
- Configure autoscaling rules

Azure functions

- Serverless applications