2-1-1: Examining Cloud Scaling Types

After completing this episode, you should be able to:

- Identify and explain cloud resource scaling types, given a scenario

Description: In this episode, the learner will examine cloud resource scaling and the characteristics of the cloud that allow for scaling, such as scalability, elasticity, and multi-regional support. We will explore common resource scaling types, such as horizontal, vertical, and diagonal.

- Describe cloud resource scaling
    - A technique that involves dynamically adjusting the capacity of cloud resources to meet changing demands.
- Describe the benefits of dynamic resource scaling in the cloud
    - Ensures that cloud infrastructure is both flexible and resilient, allowing it to adapt to workload variations and maintain optimal performance.
- Describe the common scaling types
    - Horizontal Scaling (demo - using VM Scale Sets in Azure)
        - Also known as Scale Out/In
        - Scaling involves adding or removing additional instances or resources to meet demand.
        - Applications
            - Common in microservices and containerized architectures
            - Ideal for systems that need to handle increased load by distributing it across multiple instances.
        - Benefits
            - Horizontal scaling offers flexibility, redundancy, and resilience by spreading workloads across multiple resources.
    - Vertical Scaling (demo - adjusting the performance tier on a VM hosting a database in Azure)
        - Also known as Scale Up/Down
        - Involves increasing or decreasing the capacity of existing resources, such as upgrading to a more powerful virtual machine or adding more memory.
        - Applications
            - Suitable for monolithic applications or database systems where scaling involves increasing the power of a single resource.
            - Benefits
                - Can be simpler to manage and doesn't require complex load balancing or orchestration.
    - Hybrid Scaling (Diagonal Scaling):
    - Combines elements of both horizontal and vertical scaling
    - Allows systems to scale in multiple directions based on the workload and architecture.
    - Applications
        - Can be useful for complex systems where some components require vertical scaling (like databases), while others benefit from horizontal scaling (like stateless microservices).
    - Benefits
        - Can offer flexibility, allowing for more efficient resource utilization by using the most appropriate scaling strategy for each part of the system.
- Describe common benefits of leveraging these cloud characteristics with scaling
    - Scalability
        - Manual or Automated
            - Scalability can be achieved through automation or manual adjustments, depending on the infrastructure's design.
        - Scale Up and Scale Out
            - See below
        - Support for growth

            - Ensures that cloud systems can accommodate expected or unforeseen growth in workloads over time.
    - Elasticity
        - Automatic adjustments
            - Elasticity relies on automation to increase or decrease resource capacity as needed.
        - Real-time response
            - It addresses immediate workload changes, providing resources when required and releasing them when demand decreases.
        - Cost-Efficiency
            - Elasticity helps reduce costs by using resources only when needed, minimizing overprovisioning.
    - Regional zone availability (demo - creating a VM with Availability Zones, possibly Availability Sets)
        - Redundancy and resilience
            - By distributing resources across regions or zones, cloud providers enhance resilience and reduce the impact of localized failures.
            - Compliance and data sovereignty (demo - Azure Compliance)
            - Regional zones help meet regulatory requirements and data sovereignty by allowing data to be stored in specific locations.
- Describe real-world scenarios using scaling in the cloud
    - Use the Azure Resource Manager to upgrade the performance tier on the VM hosting a database (vertical scaling)
    - Use the Azure Admin Console to create a VM Scale Set (horizontal scaling)
    - Use the Azure Admin Console to create a VM spanning multiple Availability Zones for service resiliency