100 XP

Introduction

2 minutes

One of the chief advantages of moving IT resources to the cloud is *elasticity*. The term refers to the fact that resources can be dynamically brought online to meet increased demand and taken offline when no longer needed to reduce cost. Think of a balloon that expands in size when more capacity is needed, and contracts to its original size when demand diminishes. If you are charged for the volume of air stored in the balloon, you want it to be large enough but never larger than it needs to be.

A classic example of the need for elasticity occurs when an organization's web site experiences unusually high loads. If the site can't scale to meet demand, requests take longer to process because they're queued waiting for processor time. To the customer, the site seems slow and unresponsive. In extreme cases, the site might even appear to be down.

Some loads are predictable. For example, Domino's pizza sees peak demand for its web site on holidays such as Thanksgiving and New Year's Eve, and during major events such as the Super Bowl[1]. Other loads are not as predictable. They may occur because a tweet went viral or our favorite team won the World Cup, or due to other factors that an organization can't anticipate.

In this module, we examine the mechanics of elasticity. The enabling principle is that virtual machines and other cloud resources can be brought online quickly and deprovisioned when no longer needed. We first examine common load patterns that justify the need for elasticity. We then explore two scaling techniques -- scaling up and scaling out -- as well as *auto-scaling*, which enables resources to be scaled automatically in accordance with rules established by IT administrators. We will discuss load balancing and its role in ensuring that increased capacity is utilized. Finally, we discuss a recent innovation in cloud computing that makes auto-scaling truly automatic and is ideal for certain scenarios in which loads are highly variable: *serverless computing*.

Learning objectives

- Describe common load patterns and how they drive the need to scale
- Enumerate the strategies and considerations in scaling cloud applications
- Discuss the advantages of auto-scaling and the mechanisms used to achieve it
- Describe the importance of load balancing in cloud applications and enumerate various methods to achieve it

- List the primary benefits of serverless computing and explain the concept of serverless functions

Prerequisites

- Understand what cloud computing is, including cloud service models, and common cloud providers
- Recognize cloud service models such as IaaS, PaaS, and SaaS and differentiate between them
- Understand how cloud resource provisioning works
- Be familiar with different approaches to organizing and managing cloud resources

References

1. *Domino's. Domino's 101: Basic Facts. https://biz.dominos.com/web/public/about-dominos/fun-facts.*

Next unit: Compute load patterns

Continue >