Scaling up vs. scaling out

An intro to database scalability in cloud computing.

Data, data everywhere—what we talk about when we talk about scalability

Scalability in cloud computing is the ability to quickly and easily increase or decrease the size or power of an IT solution or resource. While the term scalability can refer to the capability of any system to handle a growing amount of work, when we talk about whether to scale up vs. scale out, we are frequently referring to databases and data—and lots of it.

Database scalability is top of mind for modern app developers. Let's say a new app takes off—and demand for it grows from a handful of users to millions of users worldwide. One of the most important abilities to help the app developers keep pace with demand and minimize downtime is the ability to scale efficiently.

This conversation on scaling out vs. scaling up focuses on the ways scalability helps us to adapt and handle the sheer volume and vast array of data, changing data volumes, and shifting workload patterns—all generated from the cloud, mobile, social media, and big data.

Learn more about databases >

Scaling out vs. scaling up

At its most basic level, database scalability can be divided into two types:

Vertical scaling, or scaling up or down, where you increase or decrease computing power or databases as needed—either by changing performance levels or by using elastic database pools to automatically adjust to your workload demands.

Horizontal scaling, or scaling out or in, where you add more databases or divide your large database into smaller nodes, using a data partitioning approach called sharding, which can be managed faster and more easily across servers.

Learn more about database sharding

Scaling up vertically

Vertical scaling is used when you need to react quickly to fix a performance issue that you can't resolve using classic database optimization techniques—such as query changes or indexing. Scaling up is useful to handle spikes in your workloads where the current performance level cannot satisfy all the demands. Scaling up lets you add more resources to easily handle peak workloads. Then, when the resources are not needed anymore, scaling down lets you go back to the original state and save on cloud costs.

Scale up when:

- You see that your workloads are hitting some performance limit such as CPU or I/O limits.
- You need to quickly react to fix performance issues that can't be solved with classic database optimization.
- You need a solution that allows you to change service tiers to adapt to changing latency requirements.

Scaling out horizontally

App developers start to consider scaling out or horizontal scaling when they can't get enough resources for their workloads, even operating on the highest performance levels. With horizontal scaling, data is split into several databases, or shards, across servers, and each shard can be scaled up or down independently.

How does partitioning data improve scalability? When you scale up a single database by adding resources such as virtual machines (VMs), it will eventually reach a physical hardware limit. Because data partitions are each hosted on a separate server, if you divide data across multiple shards, you can scale out a system almost limitlessly.

Some types of database technologies, most notably nonrelational or NoSQL databases, are developed with unique capabilities to scale out data horizontally by Database sharding—enabling them to handle large, unrelated, indeterminate, or rapidly changing data.

And, some relational (SQL) database services that originally offered services to scale up or down are beginning to offer exciting options to match the scalability advantages of nonrelational databases. Hyperscale services like Microsoft Azure SQL Database Hyperscale and Azure Database for PostgreSQL Hyperscale enable users to rapidly scale storage up to 100 TB, provide flexible, cloud-native architecture allowing storage to grow as needed, and enable nearly instantaneous backups and fast database restores in only minutes.

Scale out when:

- You have geo-distributed applications where every app should access part of the data in the region. Each app will access only the shard associated to that region without affecting other shards.
- You have a global sharding scenario—such as load balancing—where you have a large number of geo-distributed clients that insert data in their own dedicated shards.
- You've maxed out your performance requirements, even in the highest performance tiers of your service, or if your data cannot fit into a single database.

Autoscaling

Autoscaling is the process of automatically and dynamically matching resources to meet the performance requirements of a system. As the volume of work grows, apps may need additional resources to maintain the necessary performance levels or meet growing demand. If demand slows down and the additional resources are no longer needed, you can save on cloud spend by having an automatic service in place to de-allocate unused resources.

Autoscaling takes advantage of the elasticity of cloud-hosted environments. It eases management overhead by reducing the need for system operators to constantly make decisions about adding or removing resources or checking the system's performance.

While there are two main ways that apps can scale—vertically or horizontally— it's less common to automate vertical scaling, because scaling up often requires making the system temporarily unavailable while it is being redeployed.

Autoscaling is more common when scaling horizontally because scaling out or in means just adding or removing instances of a resource and your app continues running without interruption as new resources are provisioned. If demand drops, resources can be shut down seamlessly without downtime and de-allocated.

Many providers of cloud-based systems, such as Microsoft Azure, support automatic horizontal scaling.

- Learn more about fast NoSQL database autoscaling with Azure Cosmos DB
- Learn more about scalability with Azure SQL Database

Frequently asked questions

Learn more cloud computing terms >

Expand all | Collapse all

- What are databases?
- What are NoSQL databases?
- What is PostgreSQL
- What is caching?
- What is database sharding?
- What is a platform as a service (PaaS)?

Resources

Quickstarts and learning modules

- Explore Azure database and analytics services
- Dynamically scale database resources with minimal downtime with Azure SQL
- Alternative methods to scale your database
- Create a Hyperscale (Citus) server group in the Azure portal

Database migration

- Accelerate, guide, and automate your database migration
- Start your cloud journey at the Azure migration center
- Discover, assess, and migrate on-prem apps, infrastructure, and data with Azure Migrate

Explore cloud scalability with Azure

Discover a comprehensive approach to scaling up vs. scaling out—one that fits your own scenario across on-premises, multicloud, and edge environments. The Azure family of database services offers a choice of fully managed relational, NoSQL, and in-memory databases, spanning proprietary and open-source engines, to fit the needs of modern app developers.

Save time and money with automated infrastructure management—including automation solutions for scalability, availability, and security.

Find the right database products for your needs with Azure >

Related products and services

Azure SQL
Family of SQL cloud databases providing flexible options for app migration, modernization, and development.

Azure Cosmos DB
Fast NoSQL database with open APIs for any scale.

Azure PostgreSQL
Fully managed, intelligent, and scalable PostgreSQL.

Azure SQL Database
Managed intelligent SQL in the cloud.

Azure SQL Managed Instance
Managed, always up-to-date SQL instance in the cloud.

SQL Server on Virtual Machines
Migrate SQL Server workloads to the cloud at the lowest TCO.

Azure Database for MySQL
Fully managed, scalable MySQL database.

Azure Maria DB
Managed MariaDB database service for app developers.

Azure Cache for Redis
Accelerate applications with high-throughput, low-latency data caching.

Scale without limits with managed databases

Focus on building apps and make your job simpler with your databases managed by Microsoft Azure.

Get started

Ready when you are—let's set up your free account

Start free