Horizontal Scaling vs. Vertical Scaling: How to Choose Which is Right for You

As your application gets more popular, you will reach a point where your servers will get to their maximum load. Before you reach that point, you must plan for how you will scale your database. Scalability is a complex topic, and you can achieve it in many different ways. In this article, we will cover what scalability is, what the different types of scaling are, and which one you should choose.

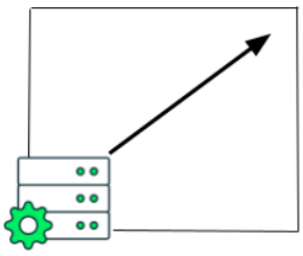
What is scalability for databases?

Database scalability is the ability for a database to adjust its resources to meet its demands constantly. As a project grows bigger or traffic increases, the original database server resources, such as RAM, CPU, and hard disks, might not suffice. This is when you'll need to start to scale your database. Scaling can occur temporarily if you expect a sudden burst of traffic due to some ad placements or more permanently when you see a constant increase in the popularity of your services. Either way, the scaling of the database can be achieved in two ways: horizontally or vertically.

You can find an in-depth explanation of both horizontal and vertical scaling in this article.

What is vertical scaling?

Probably the easiest way to scale a database is by using vertical scaling. Vertical scaling is the action of adding more resources to a server to handle an increasing load. Resources that you can add include CPUs, RAM, or hard disks.

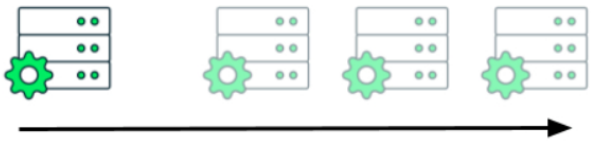


When the database server is deployed in a cloud environment, this is typically done through the management interface of the cloud provider or cloud platform.

What is horizontal scaling?

What is horizontal scaling?

Another way to scale a database server is to do so horizontally. Scaling horizontally means adding more servers so that the load is distributed across multiple nodes.



Scaling horizontally usually requires more effort than vertical scaling, but it is easier to scale indefinitely once set up. It is typically done through clustering and load-balancing. You might need some additional software if your architecture does not support horizontal scaling out of the box.

What are the differences between horizontal and vertical scaling?

	Horizontal scaling	Vertical scaling
Data location	Data is distributed across multiple nodes in a cluster. Parts of the data reside on different machines.	Data lives on a single server. If space is running out, larger hard drives are added to the server.
Downtime	There is no need to shut down the existing nodes when adding additional machines to an existing pool, meaning virtually no downtime.	When the server needs to be upgraded, the process might involve downtime.
Concurrency	Multiple jobs can be distributed across multiple machines over the network, reducing the load of a single node.	The software will need to rely heavily on multi-threading to optimize the incoming requests.
Pricing	Involves a higher investment up front to set up multiple servers, but the costs increase linearly as more machines are needed.	The initial investment is minimal and grows as the need for more hardware increases. However, there is a point at which the costs increase exponentially.

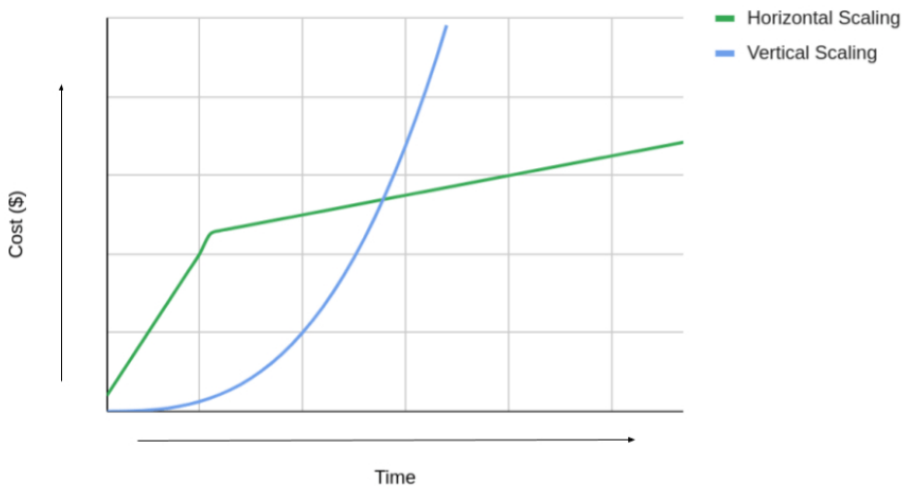
Vertical vs. horizontal scaling: pros and cons

The decision to scale horizontally or vertically can depend on a significant number of factors. Each approach has pros and cons, as described below.

Pricing

Vertical and horizontal scaling are two different approaches to database scaling. The first one involves adding more hardware resources, while the latter requires additional software considerations. Vertical scaling tends to be easy to do, but a point will be reached where the costs climb up very quickly.

Horizontal Scaling and Vertical Scaling



Horizontal scaling, on the other hand, is done by adding more servers to a cluster. The above diagram represents the costs over time of the different types of scaling. When planning for horizontal scaling, you can expect a significant upfront cost but a linear growth over time.

Downtime

Due to the nature of horizontal scaling, many servers are available to respond to incoming requests. If one of the servers goes down, the others can take over to offer almost zero downtime.

In vertical scaling, there is always a single point of failure. If the main database server were to fail, this would cause the entire system to fail.

Infrastructure management

When using vertical scaling, infrastructure management tends to be much simpler. There is a single server that is running at all times. If the server can't handle the current load, it can be upgraded with more hardware resources.

In horizontal scaling, multiple servers are added to a network and managing the server farm can offer some challenges. Ensuring that the internal network is always working correctly and that all servers are running can be difficult without the help of some automated processes and a monitoring system.

Horizontal vs. vertical scaling: which one to choose?

The quick answer to this question is that it depends. There are many considerations to take into account when deciding whether you should scale horizontally or vertically.

- Temporality: If you only need to scale temporarily, vertical scaling is most likely what you'll need. This way, you can quickly scale back down afterwards and save some money.
- Current architecture: If you use a solution that supports horizontal scaling, such as MongoDB, scaling could be an easy addition. However, if you use a more traditional database, you might need some significant investments before scaling.
- Pricing: As you do more vertical scaling, you will inevitably hit a limit where adding more resources becomes increasingly expensive. At that point, it might be cheaper to start scaling horizontally.

Horizontal scaling hybrids

Scaling isn't all black or white. As your application grows, you might want to consider a hybrid approach. If your servers can already handle horizontal scaling, increasing the resources available on some or all servers is possible before adding more machines.

To determine whether you should scale horizontally or vertically, you should calculate the costs to increase your infrastructure's servers' resources and compare it to the cost of adding additional servers.

This approach is called horizontal scaling hybrid.

What is scalability in MongoDB Atlas?

If you are already using MongoDB Atlas, the database-as-a-service offering by MongoDB, you are in luck. MongoDB Atlas offers easy ways to scale up, both horizontally and vertically. You can even set your clusters to scale automatically, based on the current server loads.

The MongoDB Atlas UI provides you with intuitive and visual metrics to determine whether you should consider scaling your database or not. From the "Metrics" tab in the database, you can visualize critical parts of your infrastructure such as the CPU load, disk storage capacity, operations per second, and much more.

Analyzing and understanding these charts will help you make a better decision about how you need to scale.

To scale up vertically, you can change the server tier by going to the configuration tab in the Atlas UI. From there, you will be able to pick a new server size that better suits your current needs. This upgrade is done with no downtime when upgrading the servers because Atlas uses a three-server replica set by default, enabling high availability to your data, even when one of the members is down.

To scale horizontally, MongoDB provides you with a built-in mechanism to distribute the data across multiple servers. This process is called sharding and can be done through a toggle button available in the configuration screen of the Atlas UI. The sharding process can also be done with no downtime whatsoever.

What are the benefits of cloud MongoDB Atlas

MongoDB Atlas provides you with an intuitive UI or administration API to efficiently perform tasks that would otherwise be very difficult. Upgrading your servers or setting up sharding without having to shut down your servers can be a challenge, but MongoDB Atlas removes that difficulty layer. Scaling your databases with MongoDB can be done with a couple of clicks.

Conclusion

As your application gets more traffic, you will eventually need to scale up your database servers. This can be done in different ways, horizontally or vertically. MongoDB Atlas provides you with an intuitive interface to do so. When thinking about scaling, you might also want to consider other alternatives such as online archiving or data lakes.

FAQ

Is horizontal or vertical scaling better?

The way you will scale your database depends on your current needs. Horizontal scaling can potentially scale indefinitely, but it requires more setup up front as well as more hardware. Vertical scaling can usually be done more quickly but might cause some downtime on your servers and might hit a limit at some point.

What is more expensive, horizontal or vertical scaling?

Vertical scaling will cost less when the servers are still small. As the servers grow bigger, vertical scaling costs will grow exponentially. At that point, horizontal scaling becomes a cheaper alternative.

What are vertical and horizontal scaling?

Vertical scaling is adding more resources such as RAM, CPUs, and hard disks to your servers. Horizontal scaling is adding more servers to distribute the data across multiple servers.

What is the advantage of vertical scaling?

Vertical scaling can be cheaper to do on small servers. It also has the advantage of temporarily increasing your server capacity for limited periods, if needed. Once the extra resources are no longer required, you can scale down.

Trust Center

Twitch

Facebook