# Scale-up or scale-out: Is there one solution that fits all problem sizes?

Marcel Stolin
marcelstolin@gmail.com

**Abstract**

Since the beginning of Big Data, batch processing was the most popular choice for processing large amounts of generated data. These existing processing technologies are not suitable to process the large amount of data we face today. Research works developed a variety of technologies that focus on stream processing. Stream processing technologies bring significant performance improvements and new opportunities to handle Big Data. In this paper, we discuss the differences of batch and stream processing and we explore existing batch and stream processing technologies. We also explain the new possibilities that stream processing make possible.

## 1 Introduction

A slowdown in Moore‘s did not happen in recent years. With the increase of interest in Artificial Intelligence and the growing complexity of modern computing architectures, more powerful resources are needed to adapt to business needs. The cost of running computing environments grows with the capacity of its computing resources. The cloud computing architecture has contributed to the elasticity of modern computing environments. Cloud computing makes it easy to increase the computing capacity of a computing environment on demand [5]. Scaling computing environments can be achieved through horizontal and vertical scaling. For the horizontal scaling strategy, increasing the computing capacity is called scaling-out, in vertical scaling it is called scaling-up [3]. Although both strategies can be used to increase the computation capacity of an IT environment, each strategy follows a different approach.

The contribution of this essay is as follows. First, the two scaling strategies are introduced while mentioning examples. Second, the differences between both scaling strategies and their typical use-cases are described. Third, a summary of this essay is given with a logical conclusion.

# 2 State of the art

There are many research studies around scalability of IT environments and different use-cases for horizontal and vertical scaling. Rossi et al. [4] created a solution to control the horizontal and vertical elasticity of container-based applications using Reinforcement learning WHATS THE RESULT.

# 3 Background

This section provides information about the scale-up and scale-out strategies.

## 3.1 Scale-up

In vertical scaling, replacing an IT resource with another with higher capacity is called scaling-up [5]. By adding more powerful resources to a component, a component can take more throughput and perform more specialized tasks [3]. In a cloud environment, scaling-up is typically achieved by replacing a VM (Virtual Machine) with a more powerful VM [5].

## 3.2 Scale-out

Scale-out refers to horizontal scaling by cloning instances of a service or data in an IT environment. By cloning instances, the work can be distributed across instances to handle an increasing number of requests. A new clone is able to perform the same work and will respond as other clones. In cloud computing, instances are typically deployed using container technologies. Cloning containers is easy due to their lightweight architecture [3].

A typical example for using a scale-out strategy is, to create new instances of a database to serve read-operations for rapidly increasing requests.

# 4 Scale-up or scale-out

If an IT environment reaches the limitations of it‘s hardware capacity, options are to scale-up the environment by adding more powerful resources or scale-out to distribute the work across multiple instances. If an IT environment is not designed to scale-out, the only option remaining is to add more powerful hardware [2]. Scaling-up by adding more powerful hardware to the IT environment, is limited by maximum hardware capacity and causes downtimes due to IT resource replacements [5]. An economical threshold can be reached, where buying more powerful hardware is too expensive. Another limitation can be, that no more powerful hardware is available by the provider [2]. If the environment is designed to scale-out, buying more powerful resource is not needed. The computing capacity can be increased by adding more instances which are typically not limited [2]. In this case, the scaling-out strategy gives an IT environment the flexibility to scale without increasing the cost.

In a cloud environment, scaling-up is less considered by architects due to its significant drawbacks. A scaling-up operation is limited by maximum hardware capacity and downtimes caused by IT resource replacements [5]. While on the other hand, scaling-out in an IT environment is not limited and can be done on-demand with modern container technologies [2]. In general, IT environments should be designed to scale-out in the first place. While the cost of buying more powerful hardware grows proportionally, the cost of investing engineering power to design an environment to scale-out grows linear [2].

## 5 Conclusion

The ability to scale is an important characteristic of modern and complex IT environments. With the ability to scale, an IT environment is able to adapt to business needs. Scaling-up and scaling-out are two scaling approaches to increase the computing capacity of an IT environment. While scaling-up seems to be the logical solution to increase the computing capacity at the first place, it comes with significant disadvantages according to cost and flexibility. A modern IT environment should be designed to scale-out in the first place to handle performance spikes and keep the cost low.

# References

[1] Emiliano Casalicchio and Vanessa Perciballi. Auto-scaling of containers: The impact of relative and absolute metrics. 09 2017.

[2] Abbott Martin L. and Fisher Michael T. *Scalability Rules: 50 Principles for Scaling Web Sites*. Addison-Wesley Professional, May 2011.

[3] Abbott Martin L. and Fisher Michael T. *The Art of Scalability: Scalable Web Architecture, Processes, and Organizations for the Modern Enterprise, Second Edition*. Addison-Wesley Professional, June 2015.

[4] F. Rossi, M. Nardelli, and V. Cardellini. Horizontal and vertical scaling of container-based applications using reinforcement learning. In *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, pages 329–338, 2019.

[5] Mahmood Zaigham, Puttini Ricardo, and Erl Thomas. *Cloud Computing: Concepts, Technology & Architecture*. Pearson, May 2013.