SIT226 Cloud Automation Technologies

Pass Task 8.1P

Vertical and Horizontal Scaling

Background

Early in this unit, we identified the problem of determining the ideal hardware for an application, both for current and future needs. This often resulted in over-estimating the resources required and unnecessary costs. In this task we explore how the scalability and elasticity are used to dynamically adjust the resources consumed by applications, eliminating this issue.

Get Prepared

Start by taking the time to ensure you understand (i) the challenges associated with estimating the hardware requirements for an application/network service (consider a web search for 'how to estimate hardware requirements' or similar), and (ii) the concepts of scalability and elasticity that relate to cloud computing and how applications benefit from these.

Finally, there are two types of scaling: vertical scaling (scaling up/down) and horizontal scaling (scaling out/in). Review these concepts, making sure that you focus on how these two approaches relate to applications. For example, what changes/modifications must be made to the application to support these scaling approaches, if any?

Complete the Task

Page Limit: 1 page of text formatted reasonably, e.g., 2cm margins, 11 or 12 point font, appropriate headings/spacing, etc.

Prepare a document according to the following requirements:

1. Reflect on the content for this week. In ½-¾ page, identify the most important lessons/topics this week relevant to your future studies/career and explain why they are the most important. Note: Do not present/explain topics, you are explaining why the things you learned are important!
2. Provide one or more screenshots for each activity, demonstrating that you have completed the lab session this week and briefly explain what is shown in each screenshot (one or two short sentences each).
3. In ¼-½ page, briefly explain why having scalable resources are beneficial for applications.
4. In ¼-½ page, briefly explain how applications may use vertical scaling and the advantages and disadvantages of this approach.
5. In ¼-½ page, briefly explain how applications may use horizontal scaling and the advantages and disadvantages of this approach.

Submit Your Task

Prepare your submission using the word processor of your choice and submit a PDF to OnTrack.

Taking it Further (Optional)

The topic of this task is the concept of scaling of an application. There are clearly two different aspects to this: the scaling of the hardware resources, but also the functionality of the application that allows it to scale to utilise additional hardware resources. There are a number of aspects worth investigating if you wish to pursue these ideas:

- It's worth reviewing the services that are provided by public cloud providers to gain some insight into the support that is provided for scaling applications. Some of the services you may wish to review include application scalers, container support, Kubernetes support (both managed and unmanaged), serverless architecture support, message queues, database scaling support, and so on.
- How applications scale is also a significant area. Understanding the difference between stateless and stateful deployments, why stateless deployments are preferable, algorithms for managing distributed and replicated data, how message queues are used, use of service meshes (such as Istio), and so on are all important areas.
- While the ability to scale is a clear benefit to applications, there are also costs associated with this approach. For example, a microservice scaled out to a dozen instances when there is excess demand is clearly beneficial, however, the application must have mechanisms in place to support the scaling of the same microservice when there is only one instance required. Reflect on what you learned in the above two points and consider how this would add overhead to an application.
- Discussions about scaling applications are usually based on assumptions that the application can scale infinitely, e.g., by consuming public cloud resources (vast but not infinite), and that costs are a secondary consideration, e.g., additional resources are only consumed when there is demand that would generate adequate revenue to cover those costs. In reality, neither assumption is necessarily true. Consider how an application might address both problems in different scenarios, e.g., if the application were deployed to a private cloud (limited resources) or if the demand was increasing due to a cybersecurity attack (no associated revenue).

Citations and Referencing

When completing any work, it is necessary to acknowledge the work of others that you have relied upon. For written assessment, we achieve this through the use of citations and references. Failing to correctly identify the work of others is known as plagiarism and is considered an issue of Academic Integrity.

If your submission to this task has involved the work of others, you must include citations and references where appropriate. Deakin provides a website that explains how to use citations and references, and includes explanations of various referencing styles:

https://www.deakin.edu.au/students/studying/study-support/referencing

You may select any style for your citations/references, however, you must be consistent in applying that style in this task (you can use other styles in other tasks if you wish).

Note that any bibliography/list of references is not included in page limits.