# 1- high availability and scalability

When building or deploying a cloud application, two of the biggest considerations are uptime (or availability) and the ability to handle demand (or scale).

## High Availability

- it’s important the resources are available when needed.
- High availability focuses on ensuring maximum availability, regardless of disruptions or events that may occur
- SLA (Service-Level Agreements) is a percentage of a service that is available.
- difference between 99% and 99.9% of SLA is big.
- *99% SLA* - will not available for 1.68 hr.’s per week. 7.2 hr.’s per month.
- *99.9% SLA* - will not available for 10 mins per week. 43.2 mins per month.

## Scalability

- ability to adjust resources to meet demand
- to scale means you can add more resources to better handle the increased demand.
    - vertical Scaling
    - horizontal Scaling

### Vertical Scaling

Vertical scaling is focused on increasing or decreasing the capabilities of resources.

- CPU,
- RAM

### Vertical Scaling

Horizontal scaling is adding or subtracting the number of **resources**.

- *Scaling out* - you could add additional virtual machines or containers.
- *Scaling in* - removing VM’s or containers.