Cloud Architecture Terminologies

What is a Solutions Architect?
A role in a technical organization that architects a technical solution using multiple systems via researching, documentation, experimentation.

What is a Cloud Architect?
A solutions architect that is focused solely on architecting technical solutions using cloud services.

A cloud architect need to understand the following terms and factor them into their designed architecture based on the business requirements.

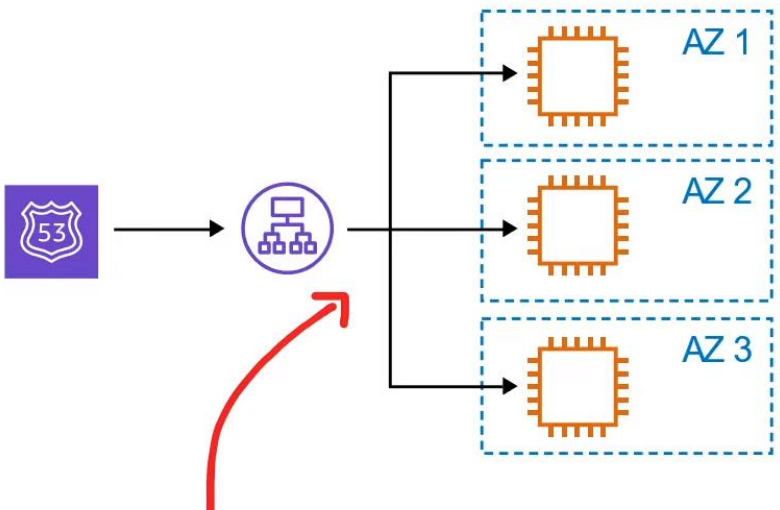
- **Availability** - Your ability to ensure a service remains available eg. **Highly Available (HA)**
- **Scalability** – Your ability to grow rapidly or unimpeded
- **Elasticity** – Your ability to shrink and grow to meet the demand
- **Fault Tolerance** – Your ability to prevent a failure
- **Disaster Recovery** - Your ability to recover from a failure eg. **Highly Durable (DR)**

A Solutions Architect needs to always consider the following business factors:

- (Security) How secure is this solution?
- (Cost) How much is this going to cost?

High Availability

Your ability for your service to **remain available** by ensuring there is ***no single point of failure** and/or ensure a certain level of performance



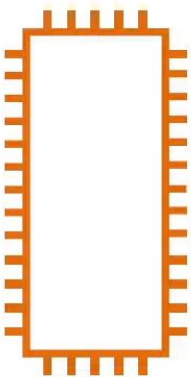
Elastic Load Balancer

A load balancer allows you to evenly distribute traffic to multiple servers in one or more datacenter. If a datacenter or server becomes unavailable (unhealthy) the load balancer will route the traffic to only available datacenters with servers.

Running your workload across multiple **Availability Zones** ensures that if 1 or 2 **AZs** become unavailable your service / applications remains available.

High Scalability

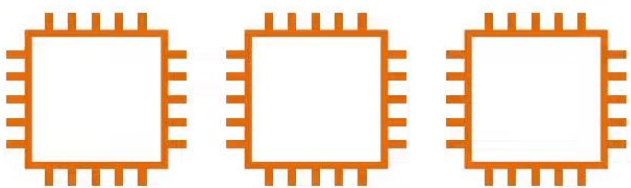
Your ability to **increase your capacity** based on the increasing demand of traffic, memory and computing power



Vertical Scaling

Scaling **Up**

Upgrade to a bigger server



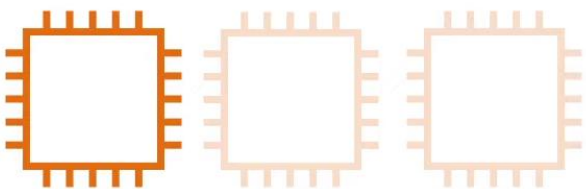
Horizonal Scaling

Scaling **Out**

Add more servers of the same size

High Elasticity

Your ability to **automatically** increase or decrease your capacity based on the current demand of traffic, memory and computing power



Auto Scaling Groups (ASG) is an AWS feature that will automatically add or remove servers based on scaling rules you define based on metrics

Horizonal Scaling

Scaling **Out** — Add more servers of the same size

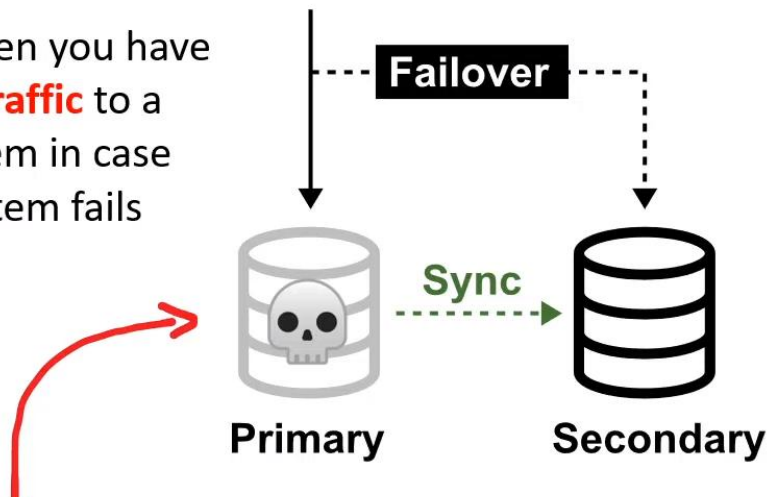
Scaling **In** — Removing underutilized servers of the same size

Vertical Scaling is generally hard for traditional architecture so you'll usually only see horizontal scaling described with Elasticity.

Highly Fault Tolerant

Your ability for your service to ensure there is no
no single point of failure. Preventing the chance of failure

Fail-overs is when you have a plan to **shift traffic** to a redundant system in case the primary system fails



RDS Multi-AZ is when you run a duplicate standby database in another Availability Zone in case your primary database fails.

A common example is having a copy (secondary) of your database where all ongoing changes are synced. The secondary system is not in-use until a fail over occurs and it becomes the primary database.

High Durability

Your ability to **recover** from a disaster and to prevent **the loss** of data
Solutions that recover from a disaster is known as **Disaster Recovery (DR)**

- Do you have a backup?
- How fast can you restore that backup?
- Does your backup still work?
- How do you ensure current live data is not corrupt?

CloudEndure Disaster Recovery continuously replicates your machines into a low-cost staging area in your target AWS account and preferred Region enabling fast and reliable recovery in case of IT data center failures.

Business Continuity Plan (BCP)

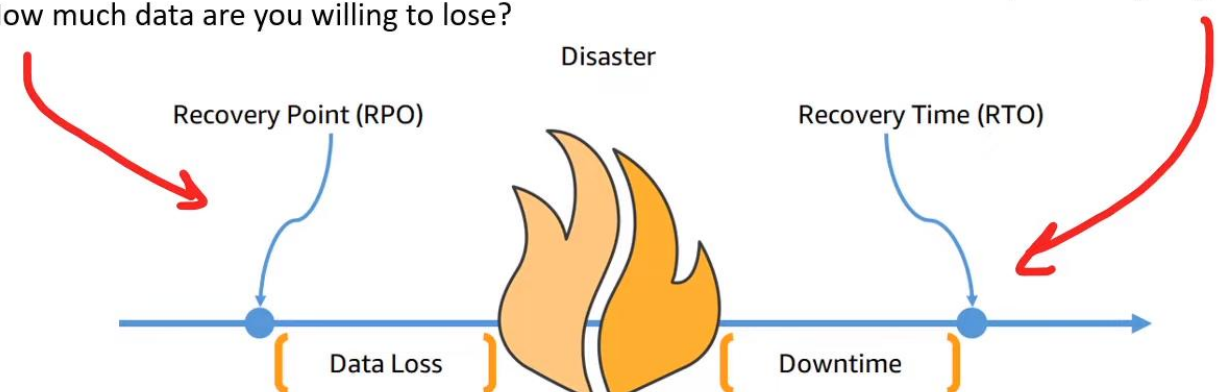
A **business continuity plan** (BCP) is a document that outlines how a business will continue operating **during an unplanned disruption in services**

Recovery Point Objective (RPO)
the maximum acceptable amount of data loss after an unplanned data-loss incident, expressed as an amount of time

How much data are you willing to lose?

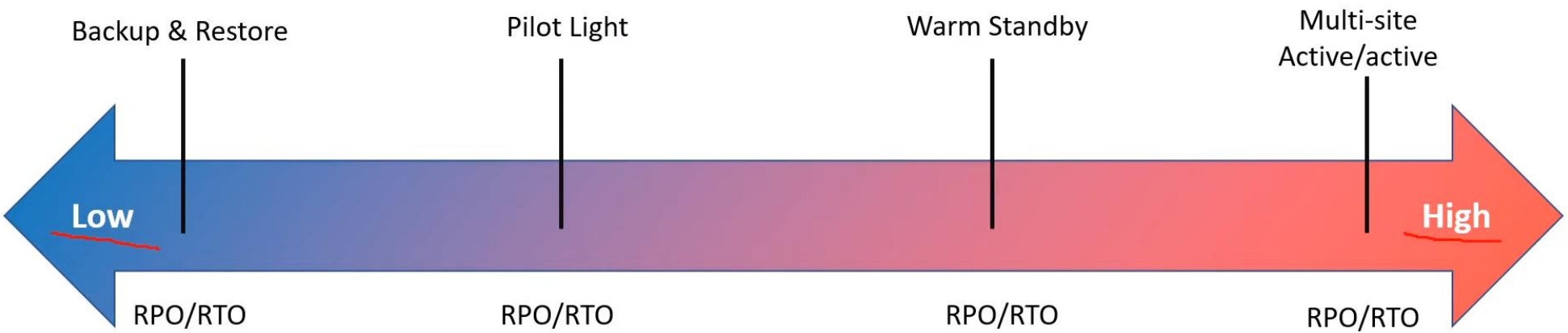
Recovery Time Objective (RTO)
the maximum amount of downtime your business can tolerate without incurring a significant financial loss

How much time are you willing to go down?



Disaster Recovery Options

There are multiple options for recovery that trade cost vs time to recover.



Backup & Restore

RPO/RTO
Hours

You backup your data and restore it to new infrastructure

- Lower priority use cases
- Restore data after event
- Deploy resources after event
- Cost $

Pilot Light

RPO/RTO
10mins

Data is replicated to another region with the minimal services running

- Less stringent RTO &RPO
- Core Services
- Start & and scale resources after event
- Cost $$

Warm Standby

RPO/RTO
Minutes

Scaled down copy of your infrastructure running ready to scale up

- Business Critical Services
- Scale resources after event
- Cost $$$

Multi-site Active/active

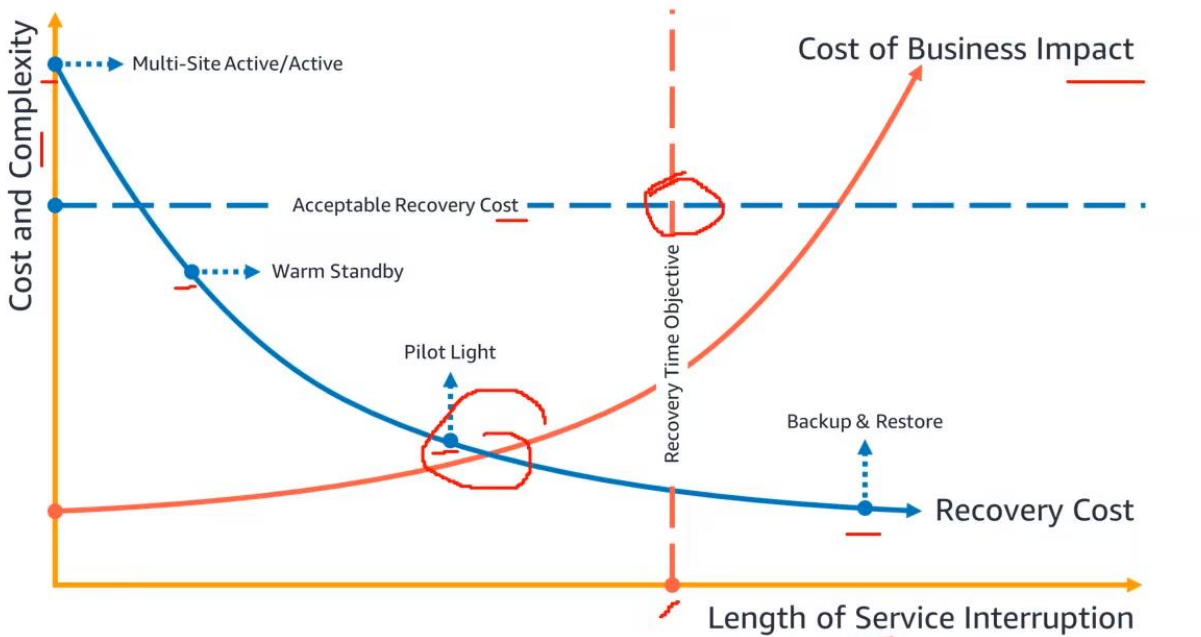
RPO/RTO
Real-time

Scaled up copy of your infrastructure in another region

- Zero downtime
- Nero zero loss
- Mission Critical Services
- Cost $$$$

RTO

Recovery Time Objective (RTO) is the maximum acceptable delay between the interruption of service and restoration of service. This objective determines what is considered an acceptable time window when service is unavailable and is defined by the organization.



RPO

Recovery Point Objective (RPO) is the maximum acceptable amount of time since the last data recovery point. This objective determines what is considered an acceptable loss of data between the last recovery point and the interruption of service and is defined by the organization.