Documento de Desenho Arquitetural

Entrega Desafio Final

Arquitetura de Soluções

Data: 20/09/2025

Aluno: Fabiano Correa da Silveira

Descrição da Solução Proposta

A arquitetura desenhada implementa uma infraestrutura de alta disponibilidade e escalável na nuvem AWS para uma aplicação de vendas on-line. A solução foi projetada para garantir que a aplicação permaneça operacional 24/7, seja resiliente a falhas e consiga se adaptar dinamicamente às flutuações de demanda dos usuários.

O fluxo de tráfego se inicia com o usuário acessando a aplicação pela internet. A requisição passa por um Application Load Balancer que a distribui de forma inteligente para um grupo de servidores de aplicação (instâncias EC2). Esses servidores estão configurados em um Auto Scaling Group, permitindo que a quantidade de instâncias aumente ou diminua conforme a necessidade. As instâncias EC2, por sua vez, se comunicam com um banco de dados relacional gerenciado (RDS) para persistir e consultar os dados da aplicação.

Toda a infraestrutura reside dentro de uma Virtual Private Cloud (VPC), que funciona como uma rede privada e isolada na nuvem. Para garantir a resiliência, os componentes são distribuídos em duas Zonas de Disponibilidade (AZs) distintas, que são datacenters fisicamente separados.

Objetivos da Arquitetura:

Esta arquitetura foi projetada para atingir os seguintes objetivos de negócio e técnicos:

- Alta Disponibilidade e Resiliência: O objetivo principal é garantir que a aplicação esteja sempre disponível. Ao distribuir as instâncias EC2 e o banco de dados RDS em múltiplas Zonas de Disponibilidade, a arquitetura é capaz de suportar a falha completa de um datacenter sem que o serviço seja interrompido.
- Escalabilidade e Elasticidade: A solução deve lidar com picos de tráfego (como em promoções ou datas comemorativas) sem degradação da performance. O Auto Scaling Group ajusta automaticamente a quantidade de servidores (entre 3 e 6), garantindo performance quando a demanda é alta e reduzindo custos quando a demanda é baixa.
- Distribuição de Carga: Evitar que um único servidor fique sobrecarregado. O Application Load Balancer distribui o tráfego de entrada de forma equilibrada entre todas as instâncias EC2 saudáveis, melhorando a performance e a confiabilidade.
- Segurança e Controle de Acesso: Proteger os dados e os recursos da aplicação. O uso de sub-redes privadas isola o banco de dados e os servidores da internet. O acesso ao banco de dados é concedido de forma segura às instâncias EC2

através de uma IAM Role, eliminando a necessidade de armazenar credenciais de acesso (chaves ou senhas) no código da aplicação.

- Recuperação de Desastres: A configuração do banco de dados RDS em modo Multi-AZ cria uma réplica síncrona (slave) em uma AZ diferente. Em caso de falha da instância primária (master), o sistema realiza o failover automático para a réplica, minimizando o tempo de inatividade e a perda de dados.
- Eficiência Operacional: Utilizar serviços gerenciados pela AWS (PaaS), como o RDS e o Application Load Balancer, para reduzir a carga de trabalho operacional da equipe de TI, que não precisa se preocupar com tarefas como provisionamento de hardware, patching de sistema operacional do banco de dados e backups.

Plataforma de Nuvem:

A solução proposta considera a plataforma de nuvem AWS, selecionada conforme familiaridade e preferências técnicas do cenário simulado.

Componentes Principais da Arquitetura:

A seguir, a descrição de cada componente presente no diagrama e sua função na arquitetura:

- VPC (Virtual Private Cloud): É a rede virtual privada e isolada na AWS onde todos os recursos da arquitetura são provisionados. Ela permite o controle total sobre o ambiente de rede.
- Internet Gateway (IGW): Componente que permite a comunicação entre a VPC e a internet. É o ponto de entrada e saída para todo o tráfego público.
- Zonas de Disponibilidade (Availability Zones - AZs): São datacenters físicos distintos dentro de uma mesma região da AWS. A utilização de duas AZs é o pilar da estratégia de alta disponibilidade.
- Sub-redes (Subnets):
    - Pública: Esta sub-rede tem uma rota direta para o Internet Gateway, permitindo que recursos nela (como o Load Balancer) sejam acessíveis a partir da internet.
    - Privada: Esta sub-rede não tem acesso direto à internet. Recursos nela (EC2, RDS) ficam protegidos e só podem ser acessados por outros recursos dentro da VPC.
- Application Load Balancer (ALB): Atua como o ponto de entrada para as requisições dos usuários. Ele distribui o tráfego HTTP/HTTPS de forma inteligente entre as instâncias EC2 nas diferentes Zonas de Disponibilidade, garantindo que apenas instâncias saudáveis recebam tráfego.
- Auto Scaling Group (ASG): Gerencia o ciclo de vida das instâncias EC2. Ele garante que um número mínimo de 3 instâncias esteja sempre em execução e

pode escalar horizontalmente até um máximo de 6, com base em métricas de demanda (como uso de CPU). Se uma instância falhar, o ASG a substitui automaticamente, promovendo a autorrecuperação (self-healing).

- Instâncias EC2 (Elastic Compute Cloud): São as máquinas virtuais (VMs) onde a aplicação web/backend é executada. No diagrama, elas rodam um sistema operacional Linux e são gerenciadas pelo Auto Scaling Group.
- RDS (Relational Database Service) Multi-AZ: É o serviço de banco de dados gerenciado (PaaS). A configuração Multi-AZ provisiona e mantém uma réplica síncrona (slave) em uma AZ diferente da instância primária (master). Em caso de falha, o RDS realiza o failover automaticamente para a réplica, garantindo a alta disponibilidade do banco de dados.
- IAM Role para EC2: É um mecanismo de segurança que concede permissões às instâncias EC2 para acessarem outros serviços da AWS (neste caso, o banco de dados RDS). As instâncias "assumem" essa role para obter credenciais temporárias, o que é muito mais seguro do que armazenar chaves de acesso permanentes.

Diagrama da Solução Proposta

O diagrama da solução proposta foi elaborado utilizando o PlantUML, com os ícones oficiais da nuvem escolhida (AWS), representando visualmente os componentes, suas interações e a distribuição entre zonas de disponibilidade: