Resumo do Projeto: Escalabilidade e Alta Disponibilidade com AWS

Application Load Balancer (ALB)

O que é?

É o "porteiro" e "gerente de tráfego" da nossa aplicação. Ele recebe todas as visitas (requisições HTTP) da internet e as distribui de forma inteligente entre as nossas instâncias EC2. Sua principal função é garantir que nenhuma instância fique sobrecarregada e que, se uma delas falhar, o tráfego seja redirecionado para as que estão saudáveis.

O que aprendemos?

- O ALB é o **ponto de entrada público** do nosso site e vive nas sub-redes públicas.
- Ele usa **Listeners** para saber em qual porta "ouvir" (no nosso caso, a porta 80 para HTTP).
- O tráfego é encaminhado para um **Target Group**, que é o grupo de instâncias EC2 que estão rodando o WordPress.
- A configuração mais crítica é o **Health Check (Verificação de Saúde)**. O ALB constantemente "pergunta" às instâncias se elas estão bem. Se uma instância não responder corretamente, o ALB para de enviar tráfego para ela, garantindo a estabilidade do site.

Passo a passo simplificado:

1. **Criar o Target Group:** Definir o grupo de instâncias (VPC, protocolo `HTTP`, porta `80`).
2. **Configurar o Health Check:** Definir o caminho que o ALB deve testar (ex: `/` ou `/healthcheck.html`).
3. **Criar o Load Balancer:** Escolher o tipo "Application", selecionar a VPC e as **sub-redes públicas**.
4. **Criar o Listener:** Configurar a regra para "ouvir" na porta 80 e encaminhar o tráfego para o Target Group criado.

Auto Scaling Group (ASG)

O que é?

É o "gerente de equipe" das nossas instâncias EC2. Sua função é garantir que sempre tenhamos o número desejado de instâncias rodando e saudáveis. Se uma instância é considerada "doente" (pelo Health Check do ALB), o ASG a demite (termina) e contrata uma nova (lança) automaticamente.

O que aprendemos?

- O ASG usa um **Launch Template** como uma "receita de bolo" para criar novas instâncias. A receita define a AMI, o tipo de instância, o script `UserData`, etc.
- As configurações `DesiredCapacity`, `MinSize` e `MaxSize` controlam o número de instâncias que o ASG deve manter.
- O ASG confia no status do **Health Check do ALB** para saber quando uma instância precisa ser substituída. Foi por isso que, durante a depuração, nossas instâncias com problemas de health check eram constantemente terminadas.
- Uma instância gerenciada por um ASG não deve ser parada manualmente, pois o ASG a interpretará como uma falha e a substituirá.

Passo a passo simplificado:

1. **Criar um Launch Template:** Definir a "receita" da instância (AMI, tipo, `UserData`, etc.).
2. **Criar o Auto Scaling Group:** Dar um nome, associar o Launch Template e escolher a VPC e as **sub-redes privadas**.
3. **Definir a Capacidade:** Configurar o número de instâncias desejadas, mínimas e máximas.
4. **Integrar com o ALB:** Associar o ASG ao Target Group do ALB.

Regras de Scaling (Scaling Up e Down)

O que são?

São as "regras de negócio" que dizem ao Auto Scaling Group quando contratar mais "funcionários" (Scaling Up) ou dispensar alguns (Scaling Down). Elas tornam a nossa infraestrutura verdadeiramente elástica, adaptando-se automaticamente ao volume de visitas.

O que aprendemos?

- Elas são baseadas em métricas, geralmente o **uso médio de CPU** de todas as instâncias.
- Uma política de **Scaling Up** adiciona instâncias quando o tráfego aumenta (ex: "se a CPU média passar de 70%, adicione uma instância").
- Uma política de **Scaling Down** remove instâncias para economizar custos quando o tráfego diminui (ex: "se a CPU média cair abaixo de 30%, remova uma instância").
- Ter ambas as políticas é fundamental para criar uma aplicação que é ao mesmo tempo performática e custo-eficiente.

Passo a passo simplificado:

1. **Selecionar o Auto Scaling Group** no console do EC2.
2. Ir na aba de **"Escalabilidade Automática"** e criar uma **"Política de escalabilidade dinâmica"**.
3. **Criar a Política de Scale Up:**

- Definir a métrica (ex: `Average CPU Utilization`).
- Definir o valor alvo (ex: `70`).
- Definir a ação (ex: adicionar `1` instância).

4. **Criar a Política de Scale Down:**
    - Repetir o processo, mas com um valor alvo menor (ex: `30`) e a ação de remover instâncias.