Elastic Load Balancing e Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling automatiza o processo de inicialização (redução) e encerramento (redução) de instâncias do Amazon EC2 com base na demanda de tráfego para seu aplicativo.

O Auto Scaling ajuda a garantir que você tenha o número correto de instâncias do EC2 disponíveis para lidar com a carga do aplicativo.

O Amazon EC2 Auto Scaling fornece elasticidade e escalabilidade.

Você cria coleções de instâncias do EC2, chamadas de grupo de Auto Scaling (ASG).

Você pode especificar o número mínimo de instâncias em cada ASG, e o AWS Auto Scaling garantirá que o grupo nunca fique abaixo desse tamanho.

Você também pode especificar o número máximo de instâncias em cada ASG e o grupo nunca ultrapassará esse tamanho.

Uma capacidade desejada pode ser configurada e o AWS Auto Scaling garantirá que o grupo tenha esse número de instâncias.

Você também pode especificar políticas de escalabilidade que controlam quando o Auto Scaling inicia ou encerra instâncias.

As políticas de dimensionamento determinam quando, se e como o ASG é dimensionado e reduzido (escalonamento sob demanda/dinâmico, dimensionamento cíclico/agendado).

Os planos de dimensionamento definem os gatilhos e quando as instâncias devem ser provisionadas/desprovisionadas.

Uma configuração de execução é o modelo usado para criar novas instâncias do EC2 e inclui parâmetros como família de instâncias, tipo de instância, AMI, par de chaves e grupos de segurança.

Amazon Elastic Load Balancing (ELB)

O ELB distribui automaticamente o tráfego de aplicativos de entrada em vários destinos, como instâncias, contêineres e endereços IP do Amazon EC2.

O ELB pode lidar com a carga variável do tráfego do seu aplicativo em uma única zona de disponibilidade ou em várias zonas de disponibilidade.

O ELB apresenta alta disponibilidade, dimensionamento automático e segurança robusta necessária para tornar seus aplicativos tolerantes a falhas.

Existem quatro tipos de Elastic Load Balancer (ELB) na AWS:

- Application Load Balancer (ALB) – balanceador de carga de camada 7 que roteia conexões com base no conteúdo da solicitação.
- Network Load Balancer (NLB) – balanceador de carga de camada 4 que roteia conexões com base em dados de protocolo IP.
- Classic Load Balancer (CLB) – este é o mais antigo dos três e fornece balanceamento de carga básico na camada 4 e na camada 7 (não está mais no exame).
- Gateway Load Balancer (GLB) – distribui conexões para dispositivos virtuais e as dimensiona para cima ou para baixo (não no exame).

Balanceador de carga de aplicativos (ALB)

O ALB é mais adequado para balanceamento de carga de tráfego HTTP e HTTPS e fornece roteamento de solicitação avançado direcionado à entrega de arquiteturas de aplicativos modernas, incluindo microsserviços e contêineres.

Operando no nível de solicitação individual (camada 7), o Application Load Balancer roteia o tráfego para destinos na Amazon Virtual Private Cloud (Amazon VPC) com base no conteúdo da solicitação.

Balanceador de carga de rede (NLB)

O NLB é mais adequado para balanceamento de carga de tráfego TCP onde é necessário um desempenho extremo.

Operando no nível de conexão (camada 4), o Network Load Balancer roteia o tráfego para destinos na Amazon Virtual Private Cloud (Amazon VPC) e é capaz de lidar com milhões de solicitações por segundo, mantendo latências ultrabaixas.

O Network Load Balancer também é otimizado para lidar com padrões de tráfego repentinos e voláteis.