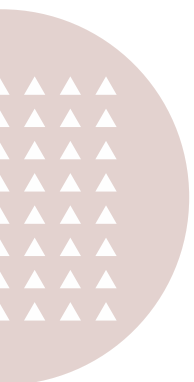




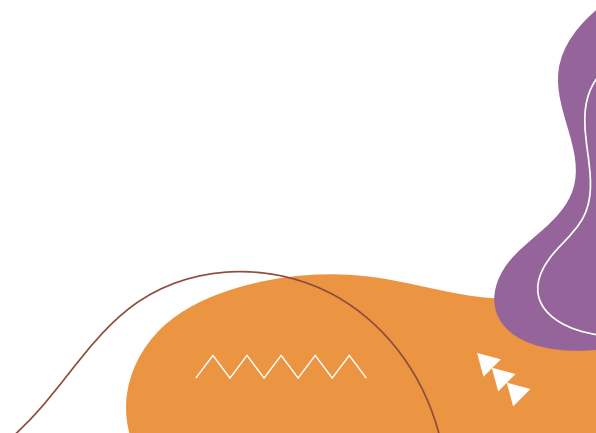
AUTO SCALING

O que é, como e quando usar uma estratégia de escalonamento automático.



SUMÁRIO

O que é Auto Scaling e por que essa técnica é importante.....	3
As terminologias importantes no universo Auto Scaling.....	5
Por que o AWS Auto Scaling é líder de mercado.....	8
5 estratégias Auto Scaling da AWS.....	10
Quando usar as estratégias de Auto Scaling da AWS.....	14
Como a Flexa Cloud pode ajudar.....	16
Sobre a Flexa Cloud.....	18





O que é Auto Scaling e por que essa técnica é importante

Auto Scaling, ou escalonamento automático, é um recurso que permite que as organizações aumentem ou diminuam automaticamente serviços em nuvem, como capacidades de servidor ou máquinas virtuais, com base em situações definidas, como níveis de utilização de tráfego.

Os principais recursos de escalonamento automático também permitem menor custo e desempenho confiável, aumentando e diminuindo continuamente novas instâncias conforme a demanda aumenta e diminui. Dessa forma, o escalonamento automático fornece consistência, apesar da demanda dinâmica e, às vezes, imprevisível de aplicativos.

O benefício geral do escalonamento automático é que ele elimina a necessidade de responder manualmente em tempo real aos picos de tráfego que merecem novos recursos e instâncias, alterando automaticamente o número ativo de servidores. Cada um desses servidores requer configuração, monitoramento e desativação, que é o núcleo do escalonamento automático.

Por exemplo, quando tal pico é causado por um ataque distribuído de negação de serviço (DDoS), pode ser difícil reconhecê-lo.





O monitoramento mais eficiente de métricas de escalonamento automático e melhores políticas de escalonamento automático podem, às vezes, ajudar um sistema a responder rapidamente a esse problema.

Da mesma forma, um banco de dados com escalonamento automático aumenta ou diminui automaticamente a capacidade, inicializa ou desliga com base nas necessidades de um aplicativo.





As terminologias importantes no universo Auto Scaling

Confira, a seguir, alguns termos que ajudam a compreender melhor o Auto Scaling!

- Uma **instância** é um único servidor ou máquina que está sujeito a regras de escalonamento automático criadas para um grupo de máquinas. O próprio grupo é um grupo de escalonamento automático, com cada instância do grupo sujeita a essas políticas de escalonamento automático.

Por exemplo, o Elastic Compute Cloud (EC2) é a plataforma de computação do ecossistema AWS. As instâncias do EC2 oferecem opções de servidor escalonáveis e personalizáveis na nuvem AWS. As instâncias do Amazon EC2 são virtuais, escalonadas elasticamente sob demanda e perfeitas para o usuário final.

- Um **grupo de escalonamento automático** é uma coleção lógica de instâncias do Amazon EC2 para fins de escalonamento automático. Cada instância do Amazon EC2 no grupo estará sujeita às mesmas políticas de escalonamento automático.

O **tamanho do grupo de dimensionamento automático** refere-se ao número de instâncias no grupo de dimensionamento automático. A capacidade ou tamanho desejado é o número ideal de instâncias





nesse grupo de escalonamento automático. Se houver uma diferença entre esses dois números, o grupo de escalonamento automático pode instanciar (provisionar e anexar) novas instâncias ou remover (desanexar e encerrar) instâncias.

- As **políticas de escalonamento automático** geralmente têm períodos de resfriamento associados para garantir que todo o sistema continue a gerenciar o tráfego.

Os períodos de resfriamento do escalonamento automático aumentam o tempo após as ações de escalonamento específicas para permitir que as instâncias recém instanciadas tenham tempo para começar a gerenciar o tráfego.

As alterações na capacidade desejada de um grupo de dimensionamento automático podem ser fixas ou incrementais. Mudanças fixas simplesmente fornecem um valor de capacidade desejado.

As alterações incrementais diminuem ou aumentam em um número específico, em vez de definir um valor final. As políticas de expansão, também chamadas de políticas de expansão, aumentam a capacidade desejada. Reduzir as políticas, também chamado de escalonamento nas políticas, diminui a capacidade desejada.

Para determinar se as instâncias anexadas estão funcionando corretamente, um grupo de escalonamento automático conduz uma verificação de integridade. Instâncias não íntegras devem ser marcadas para substituição.





- O **software de balanceamento de carga elástica** pode realizar verificações de integridade.

As verificações de status do Amazon EC2 também são uma opção, assim como as verificações de saúde personalizadas. Uma verificação de integridade positiva pode ser baseada em se a instância permanece registrada e em serviço com seu balanceador de carga associado ou se a instância ainda pode ser acessada e existe.

A configuração de inicialização descreve os scripts e parâmetros necessários para iniciar uma nova instância. Isso inclui o tipo de instância, imagem da máquina, zonas de disponibilidade potencial para lançamento, opções de compra (como spot vs. sob demanda) e scripts a serem executados no lançamento.





Por que o AWS Auto Scaling é líder de mercado

O **AWS Auto Scaling** usa automação para dimensionar recursos instantaneamente para atender à demanda e à carga do servidor.

Ao usar as ferramentas do AWS Auto Scaling, você pode ter certeza de que sempre terá instâncias suficientes para lidar com a carga do aplicativo, independentemente da intensidade ou do aumento repentino do tráfego. E, além de regular a capacidade de manter um desempenho constante, o faz pelo menor preço possível.

Se o AWS Auto Scaling parece uma opção poderosa para controlar custos e automatizar o escalonamento de recursos e serviços, é. Mas se você é novo na ferramenta, deve trabalhar com um parceiro experiente da AWS. Eles podem explicar suas várias opções de escalonamento automático e desenvolver e implementar um plano de escalonamento automático otimizado para as necessidades de sua empresa.

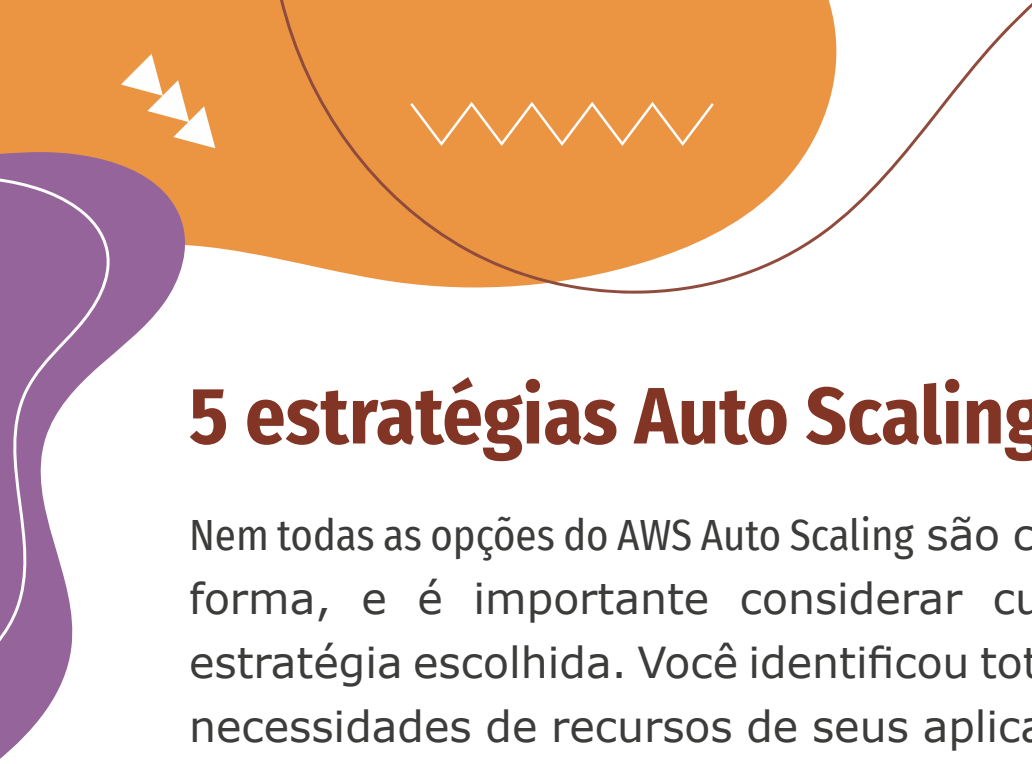
O AWS Auto Scaling oferece uma variedade de recursos e benefícios para garantir que seus aplicativos sempre tenham os recursos certos quando são mais necessários:

- **O AWS Auto Scaling fornece uma única interface de usuário** que torna o uso de recursos de dimensionamento para vários serviços da AWS um processo fácil e organizado.





- **O escalonamento automático não apenas adiciona poder de computação** para lidar com o aumento da carga de aplicativos, mas também o reduz para conservar recursos quando a demanda diminui.
- **O escalonamento automático funciona para instâncias EC2, frotas pontuais, [Amazon Elastic Container Service \(ECS\)](#) , [Amazon DynamoDB](#) ou [AWS Aurora](#) .**
- **O escalonamento de recursos é configurado e gerenciado de acordo com seu plano de escalonamento específico.**
- **As estratégias de escalonamento personalizadas são preditivas** e podem ajudá-lo com a previsão de carga e podem ajudá-lo a antecipar o comportamento da capacidade máxima.
- **Um consultor da AWS pode ajudá-lo a** personalizar sua estratégia de escalonamento automático para favorecer a disponibilidade ou o custo, ou um equilíbrio de ambos.



5 estratégias Auto Scaling da AWS

Nem todas as opções do AWS Auto Scaling são criadas da mesma forma, e é importante considerar cuidadosamente a estratégia escolhida. Você identificou totalmente todas as necessidades de recursos de seus aplicativos? Você sabe quais restrições e métricas são críticas para o seu sucesso com o escalonamento automático?

Se você estiver tendo problemas para responder a essas e outras perguntas, consulte um parceiro da AWS, que pode ajudá-lo a escolher qual dessas estratégias principais é a melhor para sua empresa.

1. Perpetuar os níveis de instância existentes indefinidamente

A primeira estratégia de escalonamento automático é simplesmente configurar o escalonamento automático para manter um número definido de instâncias indefinidamente.

O [Amazon EC2 Auto Scaling](#) verifica rotineiramente as instâncias para determinar sua integridade. Se ele detectar uma instância incorreta, ele irá encerrá-la e iniciar uma substituição. Isso dá a você um número predeterminado de instâncias, em execução o tempo todo.





2. Implementar escalonamento manual

Você sempre pode voltar para o escalonamento manual, que é a principal forma de escalonar os recursos. O Amazon EC2 Auto Scaling pode gerenciar a criação e o encerramento de instâncias para manter uma capacidade estável, que é um valor que você especificou.

Isso permite que você mantenha uma capacidade máxima, mínima ou outra capacidade desejada de sua escolha para o grupo de escalonamento automático.

3. Escala de acordo com um cronograma

Os eventos de escalonamento podem ser configurados para ocorrerem automaticamente em uma determinada data e hora. Isso é especialmente útil em situações em que você pode prever a demanda com precisão.

O que é diferente nessa estratégia é que seguir um cronograma prevê o número de recursos disponíveis em um determinado momento com antecedência, em vez de usar a automação para determinar as montagens apropriadas de momento a momento.

4. Escale junto com a demanda

Embora o AWS Auto Scaling *possa* executar todos os métodos de dimensionamento mais tradicionais mencionados nas estratégias de um a três, o dimensionamento junto com a demanda é onde os recursos exclusivos da AWS começam a brilhar.





A capacidade de alternar perfeitamente entre as estratégias mais tradicionais e aquelas discutidas nos números quatro e cinco é outro recurso interessante do AWS Auto Scaling por si só.

O escalonamento baseado na demanda é altamente responsivo ao tráfego flutuante e ajuda a acomodar picos de tráfego que você não pode prever. Isso a torna uma boa abordagem geral, “cubra todas as suas bases”. E também tem algumas configurações diferentes.

Por exemplo, você pode definir a utilização da CPU para permanecer em 50 por cento, caso a carga do aplicativo mude. É assim que a AWS responde à demanda de tráfego para que você não precise fazer isso.

5. Use escalonamento preditivo

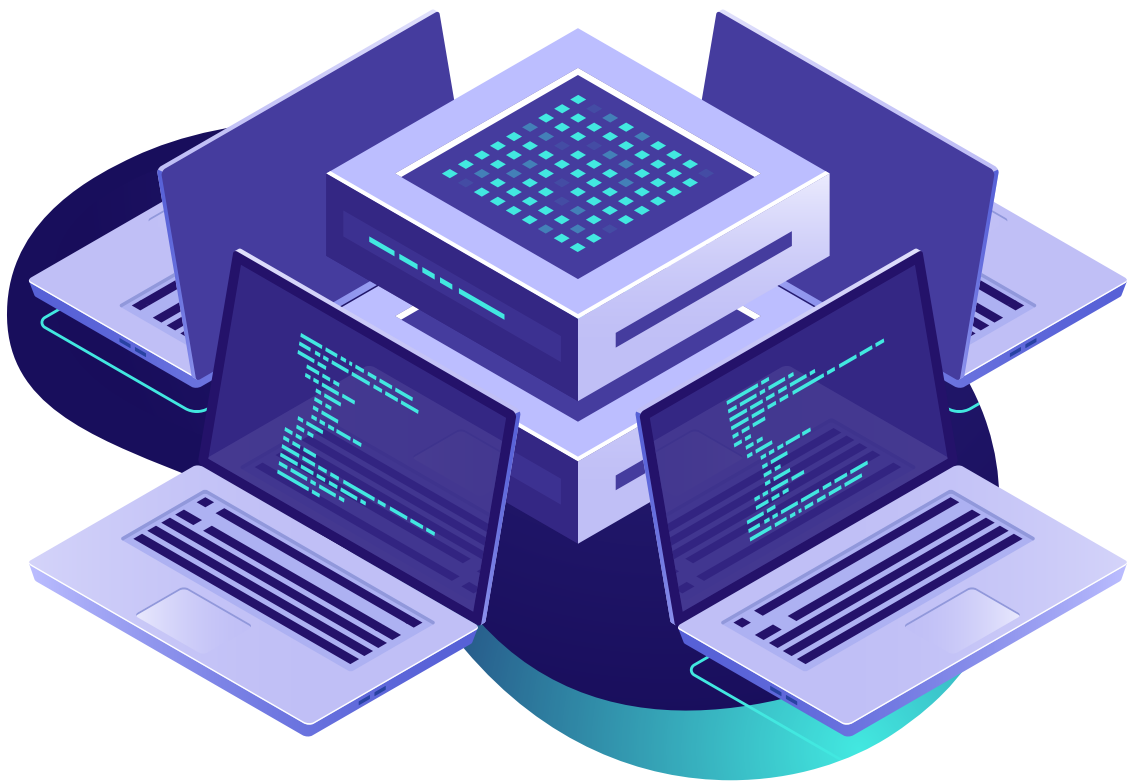
Por fim, você sempre pode combinar AWS Auto Scaling com Amazon EC2 Auto Scaling para dimensionar recursos em muitos aplicativos com dimensionamento preditivo. Isso inclui três sub opções:

1. **Previsão de carga:** este método preditivo analisa o histórico de até 14 dias para prever a demanda dos dois dias seguintes. Atualizado todos os dias, os dados são criados para refletir intervalos de uma hora.
2. **Ações de escalonamento programado:** esta opção adiciona ou remove recursos de acordo com uma previsão de carga. Isso mantém o uso de recursos estável e definido em seu valor predefinido.





3. **Comportamento da capacidade máxima:** Designe um valor mínimo e máximo de capacidade para cada recurso, e o AWS Auto Scaling manterá cada recurso dentro desse intervalo. Isso dá à AWS alguma flexibilidade dentro dos parâmetros definidos. E você pode controlar se os aplicativos podem adicionar mais recursos quando a demanda está prevista para ficar acima da capacidade máxima.





Quando usar as estratégias de Auto Scaling da AWS

Existem momentos ideais para usar essas diferentes estratégias de escalonamento automático. Basicamente, eles se resumem a se você está usando escalonamento dinâmico ou preditivo.

Enquanto o escalonamento preditivo prevê o tráfego futuro com base em tendências históricas, o escalonamento dinâmico usa um algoritmo para provisionamento automatizado de recursos.

→ Se você está tentando decidir qual usar ou quando, comece usando métricas para determinar o tráfego e os padrões de uso.

Primeiro, determine a consistência dos padrões de uso, bem como a frequência e intensidade dos picos de tráfego. Em seguida, defina suas prioridades; você deseja garantir que os clientes nunca experimentem tempos de resposta lentos ou pode pagar um pequeno nível de lentidão e, ao mesmo tempo, manter os custos no mínimo?

Esses fatores devem ser levados em consideração ao determinar os mínimos, máximos e limites para o dimensionamento. Agora você está pronto para determinar quando usar cada tipo.





Escalonamento dinâmico

O escalonamento dinâmico é a solução mais prática na maioria dos casos em que o tráfego da web ou a utilização de recursos variam uniformemente ao longo do tempo. Mas pode não ser capaz de responder rapidamente a picos acentuados, a menos que sua configuração AWS esteja configurada para limites de escalonamento agressivos.

Escalonamento preditivo

O escalonamento preditivo deve ser usado quando você sabe que pode esperar um nível elevado de uso.

Por exemplo, se um cliente tiver padrões de tráfego consistentes ao longo de uma semana, ele pode utilizar o dimensionamento preditivo para dimensionar a infraestrutura de forma proativa em preparação para o aumento do tráfego.

Também é muito útil ao planejar um evento de alto tráfego, como uma venda ou streaming de conteúdo programado.

Como a Flexa Cloud pode ajudar

A [Flexa Cloud](#) é parceira de negócios da Amazon Web Services (AWS) e fornecedora de serviços de TI. Dessa forma, estamos preparados para estruturar a estratégia de Auto Scaling da sua empresa.

Nossos profissionais são certificados em soluções e serviços AWS e têm vasta experiência em projetos de empresas de todos os portes e nos mais variados segmentos do mercado.

Confira, em nosso site, duas das principais frentes nas quais podemos ajudar o seu negócio a escalar suas capacidades:

e-learning

FLEXA CLOUD SCALE

Escale a capacidade da sua plataforma visual de ensino com o FLEXA CLOUD SCALE e-learning

[Clique aqui e saiba como.](#)



e-commerce

FLEXA CLOUD SCALE

Escale a capacidade da sua loja visual com o FLEXA CLOUD SCALE e-commerce

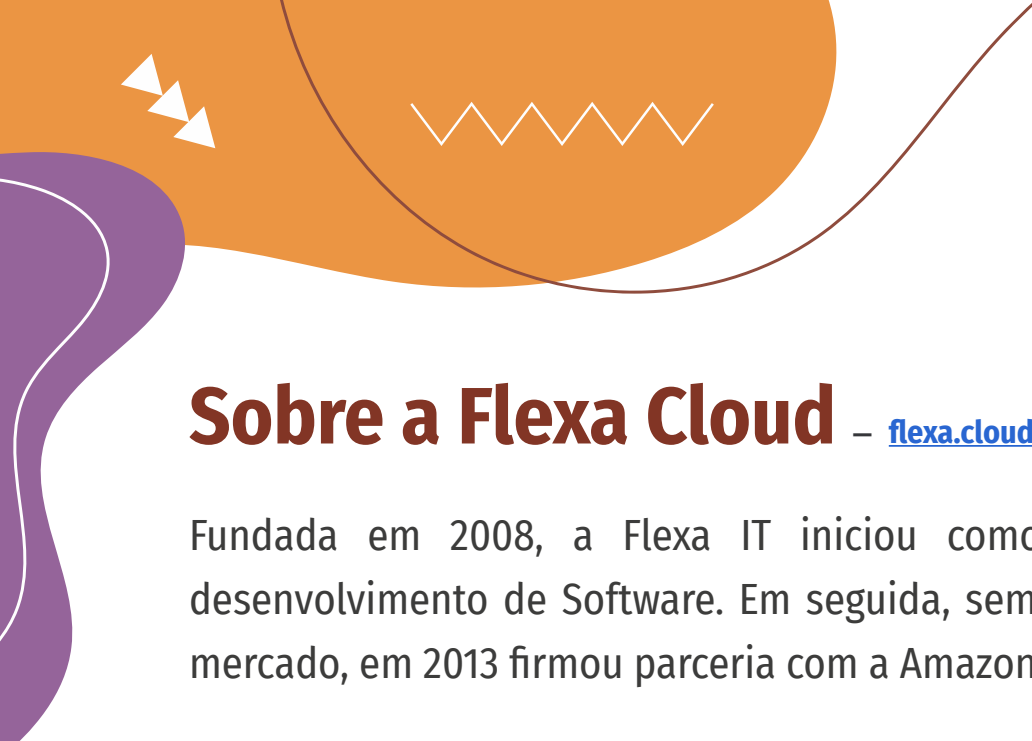
[Clique aqui e saiba como.](#)





→ **Conheça alguns dos nossos cases de sucesso:**

- [Grupo ASOEC escala EAD para 30 mil alunos com tecnologia AWS implementada pela Flexa Cloud!](#)
- [Artmark melhora infra de TI com soluções AWS implementadas pela Flexa Cloud!](#)
- [Saipos reduz custos com bancos de dados em 30% migrando do Oracle para PostgreSQL!](#)



Sobre a Flexa Cloud – flexa.cloud

Fundada em 2008, a Flexa IT iniciou como uma empresa de desenvolvimento de Software. Em seguida, sempre atenta ao mercado, em 2013 firmou parceria com a Amazon Web Services (AWS).

Com uma bandeira fincada no futuro, a Flexa IT transformou-se em uma empresa referência nacional em nuvem.

Em 2017 mudou seu nome para Flexa Cloud, consolidando definitivamente sua vocação em ajudar as empresas em sua jornada para a nuvem.

Com uma visão ampla, os serviços da Flexa consideram sempre os seguintes pilares:

- Excelência operacional;
- Segurança;
- Confiabilidade;
- Eficiência de performance;
- Otimização de custos.

Hoje, após centenas de cases bem-sucedidos, a Flexa Cloud se posicionou como uma empresa com expertise para conduzir com segurança as organizações em sua jornada para a nuvem.





Na esteira da busca em dar a melhor experiência aos nossos clientes, a Flexa Cloud faz ao longo do tempo parcerias estratégicas que visam complementar seu portfólio de serviços , a saber:

- New Relic;
- Bit Defender;
- Matillion;
- Dataguise.

A experiência acumulada nas várias implementações realizadas desde de 2013, encurta caminhos e traz uma robustez consistente na entrega dos serviços.

A etapa de implementação é estruturada com base nas melhores práticas de gerenciamento de projetos e tem no ITIL a base para os serviços na etapa de sustentação.

Se não bastasse isso, nossos colaboradores são treinados e certificados o que proporciona a tranquilidade necessária nessa jornada.

Aqui na Flexa Cloud, nossa premissa básica é o cliente. Por isso, nossos serviços buscam encantá-lo, fazendo com que ele seja o maior promotor da nossa marca.

[Faça contato](#), teremos o maior prazer em levar sua empresa às nuvens!

Faça contato conosco agora!



