
CLOUD
COMPUTING

Texto base

10

Elasticidade e Escalabilidade

Rodolfo Riyoei Goya

Resumo

Na "Computação em Nuvem", a agilidade com que recursos podem ser adquiridos e liberados permite o ajuste do dimensionamento de componentes como computação e armazenamento de um serviço à medida que demanda exigida deles muda: aumento em caso de maior demanda e redução em momento de menor demanda. Estas possibilidades permitem o uso mais eficiente dos recursos de modo a se pagar apenas pelo tempo e capacidade que estão sendo usados. Em alguns casos, aumento de capacidade pode ser proporcionado pelo uso de réplicas de leitura, ou seja, cópias usadas para operações que não envolvem alterações, apenas consultas. Abordam-se aqui, como os conceitos de redimensionamento dinâmico de elasticidade e escalabilidade são realizados nas implementações reais de nuvens com exemplos na nuvem da AWS.

10.1. Introdução

Quando se trata de computação na nuvem, como não errar por superdimensionar ou subdimensionar a quantidade de recursos necessários para um serviço? Como se adaptar rapidamente a demandas que crescem e diminuem? Como atender um grande pico de demanda de serviço que retorna posteriormente ao normal? E como fazer isso tudo a baixo custo e sem desperdícios?

10.1.1. Elasticidade

Elasticidade é a capacidade da infraestrutura de crescer ou decrescer quando necessário (como um elástico que estica quando se puxa e encolhe quando se solta). Deste modo, pode se adequar às mudanças de demanda conforme a hora do dia ou durante o dia da semana (ou para dias especiais como Dia das Mães ou Black Friday) mantendo a capacidade sempre adequada à demanda de cada momento. Com isso, evita-se o desperdício de recursos nas horas de menor uso. Na nuvem, a capacidade pode ser controlada pelo tamanho e pela quantidade dos equipamentos configurados.

10.1.2. Escalabilidade

Escalabilidade é a capacidade da infraestrutura para crescer (às vezes indefinidamente) se adequando a demandas crescentes nos negócios (crescimento no número de clientes, no volume de vendas, no leque de serviços, na área geográfica atendida, a infraestrutura precisa crescer para dar conta) com o passar do tempo. Isso é essencial para se evitar que alguma limitação da infraestrutura impeça o desenvolvimento do negócio.

10.1.3. Fail over

Quando um serviço é atendido por um grupo de equipamentos idênticos (por exemplo, instâncias EC2 lançadas de uma mesma imagem) a capacidade de se redirecionar o atendimento de um equipamento falho para outro idêntico, sem que isso afete o usuário final, é chamado “Fail over”.

Não basta a infraestrutura ser “Fail over”: as aplicações devem ser construídas de modo a serem independentes da instância onde estão sendo executadas (sem “afinidade”). Para isso, elas não devem ter estados internos (“stateless”) com toda informação persistente relevante armazenada fora da instância (a prática comum é colocar em bancos de dados).

10.1.4. Escalabilidade Vertical e Horizontal

Escalabilidade vertical é a capacidade de um componente crescer pelo aumento de seu tamanho (por exemplo, substituir uma CPU por outra capaz de executar mais instruções por segundo). Escalabilidade horizontal é crescer um serviço pelo aumento na quantidade de componentes usados para realizá-lo (por exemplo, aumentar de uma para duas CPU).

Escalabilidade horizontal é mais complexa de realizar que vertical pois, para ser eficiente, envolve a coordenação de vários componentes e exigir a distribuição uniforme de carga entre eles (caso contrário, em um conjunto de equipamentos, haverá alguns componentes ociosos e outros sobrecarregados), mas ela permite maior crescimento (por exemplo, é possível colocar centenas de CPUs para executar um serviço, mas é difícil obter uma única CPU com poder de processamento centenas de vezes maior).

10.1.5. Réplicas de leitura

Muitos serviços (por exemplo, bancos de dados relacionais) são apenas escaláveis verticalmente. Os requisitos para muitos serviços atuais, contudo, podem exigir capacidade de crescimento mais acelerada.

Assim, uma alternativa que permite parcialmente escalabilidade horizontal são as chamadas réplicas de leitura. Com elas, servidores contendo réplicas de bancos de dados podem ser acrescentadas na infraestrutura. Tais servidores não são capazes de processar atualizações no conteúdo de bancos de dados, mas permitem que aplicações sejam adaptadas para que sejam atendidas pelos servidores de réplica para os serviços que apenas consultam o banco sem alterá-lo (o que é uma ação muito comum em comércio eletrônico ou banco pela Internet, por exemplo, para consulta de estoque ou saldo).

Réplicas de conteúdo também são usadas para dar escalabilidade, reduzindo o tempo de resposta e incrementando o desempenho para acesso a páginas de Internet.

10.2. Monitoração

Para gerenciar o correto dimensionamento de serviços, evitando excesso/falta de capacidade e monitorar para substituir componentes com falhas é necessário medir o uso dos serviços e mantê-lo dentro de parâmetros adequados. Métricas usuais no monitoramento de servidores, por exemplo, incluem a porcentagem do tempo de uso de CPU (de 0% para completamente ociosa até 100% para totalmente congestionada) e o tráfego de dados pelas interfaces de rede. Com isso, os armazenamentos que ficam lotados podem ter seu tamanho aumentado.

Para executar monitoração, os serviços de nuvem incorporam "agentes" que medem o uso de recursos e transmitem os resultados desta monitoração para serviços dedicados especializados para o seu registro.

10.2.1. AWS CloudWatch

Na AWS, o serviço que centraliza coleta de medidas de desempenho geradas periodicamente e faz os registros de eventos de todos os serviços é o CloudWatch. Através dele, é possível produzir e transformar estes dados em relatórios de utilização/custos, gráficos/tabelas, alarmes indicando que limites programados foram atingidos ou ultrapassados e criar eventos periódicos. Além de proporcionar informações para gerenciamento, estes dados podem ser usados para automatizar o redimensionamento da infraestrutura.

10.2.2. AWS CloudWatch Agent

Na AWS, o serviço CloudWatch Agent é um programa (agente) que pode ser instalado no serviço a ser monitorados para enviar dados. Através dele, é possível produzir dados de monitoramento customizados com indicadores específicos. Ele pode ser usado tanto por serviço da AWS como para monitorar equipamentos on premises. Há, inclusive, versões de código aberto distribuídas publicamente, por exemplo no github, para uso e adaptação livre:
https://github.com/aws/amazon-cloudwatch-agent/

10.3. Vamos praticar?

10.3.1. Configure servidores EC2 com agentes Cloudwatch

O Cloudwatch Agent pode ser usado para gerar dados de monitoração (como uso de CPU, armazenamento e tráfego de rede) do um servidor executado em uma instância EC2. Veja mais detalhes sobre como criar um agente, configurá-lo e usá-lo para a monitoração no link:
https://docs.aws.amazon.com/pt_br/AmazonCloudWatch/latest/monitoring/Install-CloudWatch-Agent.html

10.3.2. Configure o monitoramento dos servidores do seu site com Cloudwatch

O Cloudwatch pode ser usado para coletar e exibir dados de monitoração produzidos pelo seu servidor Apache executado em uma instância EC2. Veja mais detalhes sobre como criar, configurar e usar a monitoração no link:
https://docs.aws.amazon.com/pt_br/AWSEC2/latest/UserGuide/using-cloudwatch-new.html

10.4. Você quer ler?

10.4.1. O que mais é possível monitorar com CloudWatch?

O serviço CloudWatch da AWS monitora bem mais que apenas uso de CPU. Ele inclui um grande número de métricas que permitem um controle fino do desempenho, capacidade e segurança dos serviços em execução. Veja os exemplos no link abaixo:
https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/WhatIsCloudWatch.html

10.4.2. Como configurar agentes para métricas adicionais no CloudWatch?

O serviço CloudWatch da AWS baseia sua monitoração em agentes instalados nos serviços a serem monitorados. Métricas, como a medida do uso de disco e memória RAM, podem ser obtidas de instâncias EC2. Veja os exemplos no link abaixo:
https://docs.aws.amazon.com/pt_br/AmazonCloudWatch/latest/logs/QuickStartEC2Instance.html

Referências

TAURION, Cezar. **Cloud Computing:** computação em nuvem: transformando o mundo da tecnologia da informação. Rio de Janeiro: Brasport, 2009.

VELTE, Anthony T.; VELTE, Toby J.; ELSENPRETER, Robert. **Cloud Computing:** a practical approach. EUA: McGraw-Hill, 2010.

MARSHALL, Nick; BROWN, Mike; BLAIR FRITZ, G.; JOHNSON, Ryan. **Mastering VMware vSphere 6.7.** New Jersey: Sybex, 2019. 848p.

SANTOS, Tiago. **Fundamentos da computação em nuvem**. São Paulo: Editora Senac, 2018. 211p. (Série Universitária).

ANDREWS, Joshua; HALL, Jon. **VMware Certified Professional Data Center Virtualization on vSphere 6.7 Study Guide:** Exam 2V0-21.19. New Jersey: Sybex, 2020. 640p.

Official Amazon Web Services (AWS) Documentation. **Amazon EC2 Auto Scaling User Guide.** Amazon. 368p. Disponível em: <https://docs.aws.amazon.com/autoscaling/ec2/userguide/as-dg.pdf>. Acesso em: 17 mar. 2022.

Official Amazon Web Services (AWS) Documentation. **Amazon CloudWatch User Guide.** Amazon. 985p. Disponível em: <https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/acw-ug.pdf>. Acesso em: 17 mar. 2022.

Official Amazon Web Services (AWS) Documentation. **Amazon Relational Database Service Guia do usuário.** Amazon. 2.459p. Disponível em: <https://docs.aws.amazon.com/pt_br/AmazonRDS/latest/UserGuide/rds-ug.pdf#USER_ReadRepl>. Acesso em: 17 mar. 2022.