

# Alta Disponibilidade: Aplicabilidade em Cloud Computing

## Resumo

---

### Referência bibliográfica

- AWS. **Amazon Relational Database Service**. 2022. Disponível em: <[https://docs.aws.amazon.com/pt\\_br/AmazonRDS/latest/UserGuide/rds-ug.pdf#Concepts.MultiAZ](https://docs.aws.amazon.com/pt_br/AmazonRDS/latest/UserGuide/rds-ug.pdf#Concepts.MultiAZ)> Acesso em: 09/05/2022.
- AWS. **Multi-Region Application Architecture**. 2021. Disponível em: <<https://aws.amazon.com/pt/solutions/implementations/multi-region-application-architecture/>> Acesso em: 09/05/2022.
- MOREIRA, Gabriel Araujo Siccardi. **VertElastic**: Um Módulo de decisão para explorando elasticidade vertical no Autoelastic.- São Leopoldo, 2018.
- SOARES, Juliane Adélia... [et al.]. **Redes de alta disponibilidade**. – Porto Alegre: SAGAH, 2020.
- SILVA, Fernanda Rosa da... [et al.]. **Cloud Computing**. – Porto Alegre: SAGAH, 2020.

### Escalabilidade e Elasticidade

Assim como para as nuvens de computadores, os conceitos de elasticidade e escalabilidade são fortes elementos para a Alta Disponibilidade, pois são capazes de sustentar as aplicações mantendo os níveis adequados de performance enquanto distribuem bem os recursos disponíveis. De acordo com Sharma et al. (2011 apud Moreira 2018, p.24, podemos definir elasticidade como sendo algo “[...]intimamente ligado à computação em nuvem, uma vez que se configura como a capacidade de contrair recursos computacionais de maneira dinâmica e em tempo real.

Agora, complementando este conceito, e desta vez com as palavras de Herbst, Kounev e Reussner (2013 apud Moreira 2018, p.24) temos que o conceito de elasticidade se refere ao nível de adaptabilidade de um sistema com relação a sua carga de trabalho, “provisionando e desprovisionando recursos de forma automática, de tal modo que em cada ponto no tempo, os recursos disponíveis correspondam à demanda atual, tanto quanto possível.”

Desta forma existe uma ligação direta entre o desempenho da aplicação hospedada em provedores da nuvem e a elasticidade destinada à sua carga de trabalho. Este recurso faz com que sejam modificados, dinamicamente, os parâmetros da aplicação para que seu desempenho nominal, seus requisitos de acesso aos recursos necessários para sua operação sejam garantidos.

Esta necessidade por elasticidade ocorre pois durante o funcionamento de uma aplicação existirão picos de requisições que podem superar de forma agressiva o volume normal da carga de trabalho definida para o aplicativo. Uma vez superado o número normal de requisições, os tempos de resposta e outros indicadores de desempenho devem apresentar diminuição de performance e a aplicação pode até mesmo deixar de responder a novas requisições.

Este cenário vai completamente contra o que preza o uso da alta disponibilidade, pois nele as aplicações devem funcionar continuamente em sua performance ideal. Agora a elasticidade não caminha solitária na alta disponibilidade, conta com outro importante recurso, a escalabilidade, que pode ser definido como um recurso que permite,

[...] escalar além dos limites dos recursos internos de TI existentes (processo conhecido como overprovision ou overcommitment of resources), sejam eles serviços de infraestrutura, como computação e armazenamento, ou mesmo aplicações. Para implementar esses recursos, o data center de nuvem deve ser projetado com o princípio da escalabilidade em diferentes dimensões, por exemplo, desempenho, tamanho e carga. Ele gerencia muitos recursos e usuários, que dependem da nuvem para obter o poder computacional que não podem obter nas instalações sem suportar custos administrativos e de manutenção consideráveis. (SILVA ET AL. 2020, p. 201).

Portanto, assim como a elasticidade, a escalabilidade habilita um sistema a suportar cargas de trabalho variáveis e crescentes de forma que respeite as especificações de desempenho da aplicação. Mas diferente da elasticidade, a escalabilidade tem relação direta com a adição de recursos como forma de se conquistar o atendimento a uma maior carga de trabalho. Esta diferenciação fica bem definida com as palavras de Moreira (2018, p. 25)

Uma relação equivocada entre os conceitos de escalabilidade e elasticidade no que tange à computação em nuvem se baseia no pensamento comumente difundido de que esta última, por suas características, oferece uma infraestrutura que permite a construção de sistemas escaláveis. No entanto, essa não é uma condição *sine qua non* da computação em nuvem. Para que um sistema seja escalável, ele precisa ser construído de forma a

utilizar a capacidade que o ambiente de nuvem oferece de alocar e liberar recursos dinamicamente.

Para tanto, pode haver a necessidade de um processo de adaptação específico — o que pode incluir uma etapa manual. Sem um processo de adaptação definido, um sistema escalável não pode se comportar de maneira elástica.

Podemos compreender que tanto a elasticidade quanto a escalabilidade são instrumentos poderosos na promoção da computação em nuvem e na oferta de serviços altamente disponíveis por permitir que as variações e incrementos na carga de trabalho sejam atendidas com mais recursos, mantendo o seu desempenho adequado e a satisfação dos clientes.

### **Casos de Aplicações de Alta Disponibilidade**

Para que a alta disponibilidade não seja um exagero, para que não se coloque o projeto num ciclo de duplicações e provisionamentos exagerados, promovendo um custo desnecessário, desproporcional, é preciso compreender a necessidade de cada sistema, cada aplicação. Mas a necessidade de sua existência já é algo bem difundido, devido ao nível de dependência que as empresas têm com a tecnologia nos dias atuais, conforme apresenta Soares et al. (2020, p, 199),

As aplicações são uma classe de software criada para ser utilizada por usuários. Atualmente, os negócios são completamente dependentes da tecnologia, o que torna as aplicações muito necessárias em qualquer ambiente, já que é por meio delas que os usuários executam as suas tarefas. Um exemplo de aplicação muito utilizada nas empresas e em acessos pessoais são os navegadores da web.

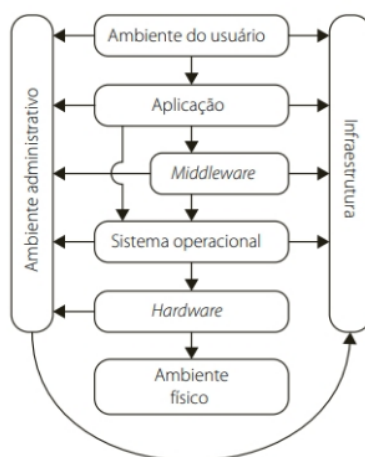
Desta forma, toda a dependência por tecnologia significa que a disponibilidade em aplicações representa também a disponibilidade do empreendimento, da empresa como um todo, pois a infraestrutura de funcionamento de um software é totalmente dependente destas tecnologias que também fazem funcionar outros aspectos do negócio, conforme complementa Soares et al. (2020, p, 199),

As aplicações são softwares projetados para a execução de um grupo de funções, tarefas ou atividades coordenadas para o benefício do usuário, como, por exemplo, navegador web e cliente de e-mail. A disponibilidade contínua dos aplicativos é uma necessidade, e os usuários estão cada vez mais exigindo um meio de garantir a alta disponibilidade de seus

aplicativos, de modo que possam realizar as suas tarefas e fornecer os mais altos níveis de serviço aos seus clientes.

Na figura a seguir temos a representação de uma pilha em um sistema, onde a camada de sistema tem sob sua estrutura a camada do ambiente do usuário, e desta forma podemos visualizar como em geral as aplicações dependem da disponibilidade de diversos componentes, tal como a empresa depende de diversos departamentos para entregar ao cliente seus produtos ou serviços.

Figura 1. Pilha do sistema.



Fonte: Adaptada de Schmidt (2006 *apud* SOARES ET AL. 2020, p.200)

Devemos sempre manter uma reflexão: nem todo recurso que pode ser acessado está disponível, pois em sistemas computacionais, um recurso disponível representa aquele que possui capacidade de responder de forma adequada as requisições de seus usuários. Desta forma, os recursos que uma aplicação demanda para que possa responder adequadamente sua carga de trabalho são analisados na busca pela sua alta disponibilidade, desta forma,

Nesse sentido, aumentar a disponibilidade das aplicações é uma das principais preocupações das organizações, pois a interrupção do fluxo de trabalho devido a falhas causa prejuízos financeiros. O custo do tempo de inatividade, ainda que programado, pode ter um impacto negativo substancial na receita das organizações, pois pode acarretar perdas de negócios imediatos ou perdas de fidelidade e confiança do cliente quando essas indisponibilidades ocorrem com frequência. (SOARES ET AL. 2020, p.200).

A grande maioria das aplicações tem como objetivo de desenvolvimento a busca por lucro aos seus desenvolvedores. Desta forma sua disponibilidade é buscada como forma de garantir e até

aprimorar, elevar a lucratividade. Em uma outra perspectiva, a indisponibilidade das aplicações e sistemas empresariais fazem com que a empresa perca oportunidades de negócio, atrase sua produção, algo que impacta severamente em sua receita e lucros.

## Implementações Multi-AZ para Alta Disponibilidade na AWS

A AWS oferece diversos recursos em sua infraestrutura que permitem e aprimoram a Alta Disponibilidade, um dos mais efetivos está inserido na forma com que ela, a AWS, distribui seus recursos computacionais, as suas zonas de disponibilidade. A alta disponibilidade, quando implementada para o cliente AWS representa o uso de diversas zonas de disponibilidade, o recurso denominado pela empresa como Multi-AZ, onde AZ representa as Availability, e de acordo com AWS (2022, p. 68),

Implantações Multi-AZ podem ter uma ou duas instâncias de banco de dados em espera. Quando a implantação tem uma instância de banco de dados em espera, ela é chamada de implantação de instância de banco de dados Multi-AZ. Uma implantação de instância de banco de dados Multi-AZ tem uma instância de banco de dados em espera que fornece suporte para failover, mas não serve tráfego de leitura. Quando a implantação tem duas instâncias de banco de dados em espera, ela é chamada de implantação de cluster de banco de dados Multi-AZ.

Um exemplo desta implementação em favor da alta disponibilidade diz respeito a implementação de clusters de bancos de dados em estruturas Multi-AZ que utilizam instâncias destes bancos de dados em modo de espera e assim podem manter o fluxo de dados da aplicação durante seu processo de failover ou como tráfego de leitura. O recurso pode ser gerenciado pelo AWS Management Console através do sistema Databases e em seguida determinando um DB identifier (Identificador de banco de dados). Tal implantação apresenta as seguintes características:

1. Para instância de banco de dados:
  - Há apenas uma linha para a instância de banco de dados.
  - O valor de Role (Função) é Instance (Instância) ou Primary (Principal).
  - O valor de multi-AZ é Yes (Sim). p. 68).
2. Para cluster de banco de dados:

- Há uma linha em nível de cluster com três linhas de instâncias de banco de dados abaixo dela.
- Para a linha no nível do cluster, o valor de Role (Função) é Multi-AZ DB cluster (Cluster de banco de dados multi-AZ).
- Para cada linha no nível da instância, o valor de Role (Função) é Writer instance (Instância de gravador) ou Reader instance (Instância de leitor).
- Para cada linha no nível da instância, o valor de Multi-AZ é 3 Zones (Três zonas). (AWS 2022, p. 68).

Dentro das estruturas de segurança e redundância oferecidas pela AWS existe uma possibilidade remota de que mais de uma zona de disponibilidade em uma mesma região sofram alguma falha e se tornem simultaneamente indisponíveis, o que torna o recurso um grande incremento de qualidade e alta disponibilidade aos serviços implementados desta maneira.

## Implementação de Infraestrutura Multi-regiões

A infraestrutura global da AWS está distribuída em regiões, espalhadas em locais onde existe maior concentração das atividades da Amazon, com seu e-commerce, e de seus clientes, para a oferta de latências e tempos de resposta competitivos. Assim, em suas regiões existem as Zonas de Disponibilidade onde ela mantém seus servidores. De forma geral, cada região apresenta ao menos 3 zonas de disponibilidades o que permite a implantação de uma arquitetura capaz de ser replicada tanto ente zonas quanto entre regiões. De acordo com a AWS (2021, online), a modelagem do workload dentro desta arquitetura pode ocorrer da seguinte forma,

Essa orientação implanta uma arquitetura de referência que modela uma workload ativa/passiva sem servidor com replicação assíncrona de dados de aplicações e failover de uma região da AWS primária para uma secundária. Para verificar se o failover regional está funcionando, também é possível implantar uma aplicação Web de exemplo para compartilhamento de fotos, que serve como demonstração visual das camadas de backend. Essa orientação permite um objetivo de ponto de recuperação (RPO) de 15 minutos e um objetivo de tempo de recuperação (RTO) de alguns segundos. (AWS 2021, ONLINE)

Portanto, esta arquitetura da orientação da Multi-Region Application Architecture tem suporte dentro do CloudFormation com o uso dos StackSets e assim permitir a aplicação da camada de

roteamento criando uma zona primária e uma secundária, está para o processo de failover. Mas esta infraestrutura permite a implantação de um segundo modelo, ainda dentro do CloudFormation, onde,

O modelo secundário implanta uma aplicação Web de exemplo do AWS Amplify hospedada em um bucket do Simple Storage Service (Amazon S3), uma distribuição do Amazon CloudFront para entregar a aplicação Web aos usuários e um grupo de usuários e um grupo de identidades do Amazon Cognito para permitir que os usuários acessem a aplicação Web, a camada de roteamento e os recursos da infraestrutura de backend. (AWS 2021, ONLINE)

Na prática esta estrutura funciona com a aplicação web realizando a consulta à camada de roteamento pelo estado da aplicação, ou seja, para saber se a aplicação solicitada está ativa, em isolamento ou em processo de failover. Este recurso é incrementado pelo AWS Amplify para que possa direcionar as solicitações à instância correta.

## Exercícios

---

1. A computação em nuvem figura como um conceito, que embora recente, traz consigo grande responsabilidade sobre o atual cenário computacional. Sua construção que permite a oferta de recursos de TI altamente complexos, apresentam uma característica, dentre muitas outras não menos importante, essencial para a eficiência de seus serviços, que representa o grau de adaptabilidade a mudanças de carga de trabalho promovendo o dimensionando o provisionamento de recursos de forma automática. Assinale a alternativa que apresente a característica que apresente esta descrição:
  - a) Disponibilidade
  - b) Elasticidade.
  - c) Confiabilidade.
  - d) Escalabilidade.
  - e) Confidencialidade.
  
2. Através da revolução tecnológica, a era da informática foi consolidada. Neste contexto, novas formas de ofertas de serviços foram instituídas através das grandes redes de computadores, como o cloud Computing, que pode ser descrito sucintamente com a oferta de serviços computacionais pela internet. Os serviços oferecidos pela computação em nuvem, como a escalabilidade e a elasticidade, são bastante atrativos, tornando estes serviços uma eficiente ferramenta de promoção organizacional. Neste contexto, assinale a alternativa que apresente o conceito de escalabilidade na nuvem?
  - a) Capacidade de dimensionar o provisionamento de recursos automaticamente.
  - b) Capacidade do sistema de modificar a estrutura física de um sistema.
  - c) Capacidade do sistema em abstrair detalhes da configuração da infraestrutura.
  - d) Grau em que um sistema se adapta a mudanças de carga de trabalho.
  - e) A capacidade de um sistema em suportar cargas de trabalho crescentes.
  
3. As aplicações representam uma categoria de softwares que foram desenvolvidas direcionadas para os usuários. Estas aplicações são projetadas para desempenhar certo as atividades que tem como objetivo atender as demandas cada vez mais crescentes deste usuário, sendo de grande funcionalidade para as operações organizacionais. Quais são as principais implicações para as organizações na ocorrência de indisponibilidade de aplicações?

- a) Perda de capacidade computacional, gerando assim constantes falhas nos sistemas internos.
- b) Prejuízos financeiros por perda de negócios imediatos e da fidelidade e confiança do cliente, caso a indisponibilidade for recorrente.
- c) Possível perda de escalabilidade no sistema, resultado de tempo decorrente de indisponibilidade.
- d) Maior dificuldade na recuperação de dados pós desastres, mesmo com um efetivo DR (Plano de recuperação de desastres).
- e) Uma recorrência nos períodos de indisponibilidade das aplicações resulta em uma instabilidade da disponibilidade permanente.

4. Uma instância de banco de dados representa um ambiente de banco de dados que se encontra isolado na nuvem. Um importante recurso, o Amazon RDS, quando utilizado com a opção Multi-AZ, é capaz de oferecer maior disponibilidade e durabilidade para estas instâncias de banco de dados, tornando estas instâncias funcionais para ambientes de produção. No caso de Implementações Multi-AZ, elas podem ter duas ou mais instâncias de banco de dados em espera. Como pode ser chamada uma implantação que possui apenas uma instância de banco de dados em espera?

- a) Implantação de cluster de banco de dados Multi-AZ
- b) Implantação de painel de navegação Multi-AZ.
- c) Implantação de data base Multi-AZ.
- d) Implantação de instância de banco de dados Multi-AZ.
- e) Implantação de alta disponibilidade.

5. Dentro de um ambiente em nuvem e de um modo geral, as instâncias representam um servidor virtual, em que é possível realizar a definição e a configuração do sistema operacional e das aplicações que são configuradas nestas instâncias. Entretanto, instâncias de banco de dados possuem características específicas. Assinale a alternativa que apresente o conceito de instância de banco de dados.

- a) Ambiente físico de banco de dados que fornece suporte para failover.
- b) São aplicações de alta disponibilidade dentro de um banco de dados.
- c) São dispositivos físicos de controle que são legíveis.
- d) Cluster de banco de dados que serve de tráfego de leitura.
- e) Ambiente de banco de dados que se encontra isolado em execução na nuvem.

6. A AWS (Amazon Web Services) oferece inúmeras ferramentas computacionais na nuvem, que visam viabilizar e facilitar o acesso a serviços tecnológicos a usuários que estão limitados pelos altos custos da implantação de infraestruturas computacionais. A Amazon RDS (Amazon Relational Database Service) são grupos de serviços que podem ser gerenciados, que oferecem ferramentas de configuração, operação e escalabilidade de banco de dados, além de disponibilizar a implantação de cluster de banco de dados Multi-AZ, que está diretamente relacionada com a promoção da disponibilidade. Como pode ser definida uma implantação de cluster de banco de dados Multi-AZ?

- a) Método de implantação de alta disponibilidade do Amazon RDS, que apresenta duas instâncias de banco de dados legíveis.
- b) Modo de implantação que oferece suporte para failover em instâncias de banco de dados que não servem tráfego de leitura.
- c) Modelo de implantação do Amazon S3, que apresenta uma instância de banco de dados legível.
- d) Método de implantação de alta disponibilidade do AWS Auto Scaling, que oferece suporte para failover em instância de banco de dados relacionais.
- e) Modelo de implantação de alta disponibilidade do Amazon CloudFront que apresenta duas instâncias de banco de dados.

## Gabarito

---

**1. Letra B.**

A alternativa está correta, pois a elasticidade, que representa uma das mais atrativas características da computação em nuvem, é capacidade do sistema de se adaptar às mudanças de carga de trabalho, tem a capacidade de provisionar e deprovisionar os recursos de forma automática, de forma a garantir que os recursos estejam disponíveis de acordo com a demanda de um determinado ponto no tempo.

**2. Letra E.**

A alternativa está correta, pois a escalabilidade representa a capacidade de um sistema em suportar e se adequar a cargas crescentes de trabalho, sempre com o desempenho adequado, mesmo com a adição de novos recursos computacionais.

**3. Letra B.**

Para as organizações a disponibilidade das aplicações garante uma gama de possibilidades de negócios, além de garantir a fidelidade de clientes que dispõe de acesso aos sistemas da organização sempre que este acesso for requerido. Deste modo a indisponibilidade pode gerar perda de negócios imediatos, além de gerar insatisfação do cliente, e muitas vezes perdendo sua fidelidade, resultando em prejuízos financeiros.

**4. Letra D.**

A Alternativa está correta, pois as implementações Multi-AZ podem ter uma ou duas instância de banco de dados em espera, no caso de apenas uma instância, esta é chamada de implementação de instância de banco de dados Multi-AZ, em que esta instância fornece suporte para o faillover, mas não serve tráfego de leitura.

**5. Letra E.**

A alternativa está correta, pois as instâncias de banco de dados representam um ambiente de banco de dados que pode conter vários bancos de dados criados pelo usuário que podem ser modificados por este usuário, estando de forma isolada na nuvem.

**6. Letra A.**

A alternativa está correta, pois uma implantação de cluster de banco de dados Multi-AZ representa um método de implantação de alta disponibilidade que apresenta duas instâncias de banco de dados legíveis, que fornecem suporte para failover e que podem servir de tráfego de leitura.