O Que Você Realmente Precisa Saber Sobre

OPUS
SOFTWARE

Opus Software

O Que Você Realmente Precisa Saber Sobre

Computação em Nuvem

Primeira Edição

São Paulo - SP - Brasil

Edição realizada por Opus Software Com. e Repr. Ltda

2015

Produção Do Livro

COPRODUZIMOS SOLUÇÕES DE SOFTWARE QUE TRANSFORMAM CONHECIMENTO EM VALOR

Desenvolvemos soluções de software exclusivas que potencializam o talento das pessoas e o diferencial das organizações. Através de um processo de coprodução, unimos nossa competência tecnológica ao conhecimento de negócio de nossos clientes para a construção de aplicações que efetivamente geram resultados mensuráveis. A oferta da OPUS Software inclui a terceirização do desenvolvimento e manutenção evolutiva de sistemas, acompanhando todo o ciclo de vida das aplicações. Além disso, oferecemos serviços de migração de sistemas para a nuvem e gerenciamento, monitoramento e sustentação dos ambientes migrados. A OPUS Software é parceiro oficial da Amazon Web Services (AWS), da Microsoft e da IBM.

ou indiretamente pelo conteúdo deste livro.

A Opus Software procurou informar a origem dos dados de todas empresas e produtos mencionados neste livro, através de links de referência pesquisados no ano de 2014. No entanto, nem a Opus Software, nem seus distribuidores podem garantir a precisão dessas informações.

Primeira Publicação: Abril de 2015
Publicado por Opus Software Ltda
Rua Butantã, 500/518 – 2º Andar
Pinheiros - CEP 05424-000 - São Paulo – SP
(+55 11) 3816.2200
cloud@opus-software.com.br
www.opus-software.com.br
ISBN 978-85-69180-00-5

Conteúdo

- Comprando Infraestrutura como Serviço (IaaS)
  - Configuração de servidores
  - Armazenamento de dados
  - Banda Internet
  - Tráfego de E/S
  - Softwares e imagens binárias
  - Controle de acesso
  - Facilidade de gerenciamento
  - Custos
- Amazon Web Services (AWS)
  - Servidores
  - Armazenamento de dados
    - Amazon EBS (Elastic Block Store)
    - Amazon S3 (Simple Storage Service)
    - Amazon Glacier
  - Bancos de dados
    - Amazon RDS (Relational Database Service)
    - Amazon Aurora
    - NoSQL
  - Outros Serviços
- Google Compute Engine
  - Servidores
  - Armazenamento de dados
    - Discos persistentes
    - Google Cloud Storage
  - Bancos de dados
    - Cloud SQL
    - Cloud Datastore
    - BigQuery

Introdução

A Computação em Nuvem está sendo responsável por uma das maiores revoluções ocorridas nos últimos anos na área de Tecnologia da Informação. Os impactos dessa transformação têm crescido e se acelerado, na medida em que a nuvem oferece cada vez mais serviços, com mais segurança, com maiores recursos e com custos cada vez mais atraentes e competitivos. É uma indústria que definitivamente muda o modo de fazer as coisas na área de TI e que, embora ainda jovem, já proporciona resultados consolidados, fazendo com que sua adoção seja uma opção segura.

O fundamental a ser entendido na Computação em Nuvem é que não é uma revolução tecnológica encerrada em si mesma. Além de mudar o modo como se produzem os serviços de TI das empresas, ela potencializa a mudança do modo como a empresa oferece seus produtos e serviços, atinge seus novos clientes, acompanha os clientes existentes e pratica o seu marketing no dia a dia, tanto para a captação de novos clientes quanto para a manutenção dos existentes.

Além disso, a Computação em Nuvem permite que organizações de qualquer porte tenham acesso a recursos que antes só estavam disponíveis para grandes empresas, por exigirem elevados investimentos, e agora podem ser pagos sob demanda. Isso muda as condições de competitividade nos mercados, criando oportunidades ímpares de crescimento acelerado sem exigir a antecipação de grandes investimentos na área de infraestrutura de tecnologia. Nesse sentido estamos num novo mundo, onde TI é fundamental para qualquer negócio e, por outro lado, qualquer negócio pode ter sofisticados serviços de TI. Isso muda as relações de força entre as empresas, em qualquer setor.

É preciso entender bem o que é Computação em Nuvem, para não reduzi-la a uma simples oportunidade de diminuir alguns custos de TI ou investimentos em tecnologia. Seu impacto é muito mais profundo:

- Muda o perfil de qualificação de todo o pessoal especializado em TI. Mais do que nunca, o homem de TI se transforma num profissional que precisa entender os negócios de sua empresa, além de enxergar que tudo o que é feito na sua área deve, em última instância, se traduzir em serviços adequados, velozes, seguros e econômicos. É imperativo para essa classe profissional acompanhar estas mudanças na maneira de pensar e agir, sob o risco de perder sua função no mercado. O profissional de TI está se transformando num homem de negócios que entende de tecnologia e que sabe conversar com os líderes das empresas das áreas de marketing, finanças, produção, etc., com foco nos serviços essenciais para os negócios da empresa – e não mais em hardware e software.
- Muda o modo como a empresa pensa e pratica o seu Marketing. Com a nuvem a internet se transforma de fato, cada vez mais, na internet das coisas, onde tudo está interligado e qualquer dispositivo eletrônico pode estar permanentemente armazenando informações na rede e recebendo estímulos dela. Isso muda o marketing tradicional para o “Marketing das Coisas” ou o “Marketing pras Gentes“, no qual essa integração vai se dar com pessoas, através de redes sociais cada vez mais amplas, trocando e armazenando dados que antes estariam indisponíveis. O “Marketing pras Gentes” revoluciona o marketing tradicional: contato instantâneo e contínuo com clientes potenciais e efetivos, armazenagem de dados sem limites, possibilidades de análises mercadológicas sem precedentes e sem a necessidade de “pesquisas de mercado“. Os clientes potenciais e efetivos se transformam na rede de negócios da própria empresa, muito além de uma simples rede social para troca de fotos, filmes e opiniões. Clientes em rede, marketing em rede. Capacidade massiva de tratar individualmente cada pessoa da rede, de acordo com seus dados, seu perfil de demanda, seu histórico de compras, seu comportamento, seus dados sociodemográficos e por aí vai. A custos baixos como nunca antes imaginados ou possíveis.
- Muda a forma como a empresa produz e entrega seus produtos e serviços. Na verdade, o ciclo tradicional de produção, venda e transferência de posse de um serviço atômico ou de um bem, que

> caracterizava a missão de uma empresa, está rapidamente deixando de existir. Nos novos modelos de negócio possibilitados por essa transformação social e tecnológica, o consumidor “aluga” o produto ou serviço e fica com ele enquanto o mesmo é capaz de agregar valor à sua vida. A pós-venda passa a ter importância relativamente maior que a venda e a pré-venda, na medida em que a satisfação de uso é a única garantia de continuidade do negócio. Neste sentido, esta integração da empresa com sua rede de usuários é extremamente alavancada pelo uso dos serviços em nuvem.

Além disso, estamos assistindo a uma digitalização crescente da economia. Todos os negócios estão se transformando em negócios de Tecnologia da Informação, e a economia digital está se estabelecendo de forma rápida e irreversível. Muitos dos novos modelos de negócios só são viáveis através do uso da tecnologia, fazendo com que ela desempenhe um papel de relevância crescente nas organizações. À medida que as fronteiras entre TI e negócios se tornam nebulosas, a tecnologia se torna mais crítica do que nunca na execução das estratégias de negócio. Assim, o entendimento das novas possibilidades trazidas pelo avanço da tecnologia passa a ser uma função de todos dentro das organizações, não se restringindo mais apenas à área de TI.

Por isso tudo nós, da Opus Software, entendemos que seria importante tornar disponíveis conhecimentos fundamentais sobre Computação em Nuvem e seu potencial transformador, uma vez que temos acompanhado essa revolução desde seu início. Nosso conhecimento não é meramente técnico e vem da experiência de termos trabalhado em conjunto com nossos clientes, produzindo e implantando com sucesso várias soluções baseadas em tecnologia. Cabe destacar que o nosso negócio também foi profundamente alterado por tais transformações, e onde antes desenvolvíamos soluções de TI, hoje desenvolvemos soluções de negócios alavancadas por TI.

Temos certeza que a leitura deste livro vai dar uma excelente ideia do que é a propalada Computação em Nuvem e do que é possível revolucionar na sua empresa. Se você é um CEO, esperamos que os conceitos apresentados aqui ampliem suas possibilidades e ajudem a transformar o modo de você

pensar seu próprio negócio. Se você é um homem de marketing, esperamos proporcionar novas ferramentas para praticar o “Marketing pras Gentes” e transformar o relacionamento com seus mercados. Se você é um homem de TI, esperamos ajudá-lo a repensar seu próprio papel e a desenhar serviços e soluções cada vez mais inovadores. E se você atua em qualquer outra área, nossa expectativa é que o conhecimento sobre os fundamentos da Computação em Nuvem permita-lhe vislumbrar novos horizontes sobre como desenvolver novas soluções de negócios alavancadas por TI.

Boa leitura.

Por que Computação em Nuvem?

Desde o início da computação comercial, nos anos 1950, os fornecedores de tecnologia estiveram focados nos grandes clientes e seus grandes orçamentos. Mainframes, redes privadas de alta velocidade, arquiteturas de alta disponibilidade, “disaster recovery”, computação distribuída… Tudo muito eficiente e robusto, mas praticamente inacessível às pequenas e médias empresas dado seu alto custo e excessiva complexidade.

Esse fato acabou deixando pequenas e médias empresas sem condições de acesso à Tecnologia da Informação de ponta. Quantas empresas não pensaram em implementar alguma inovação de TI e acabaram desistindo por causa dos custos e da complexidade envolvidos?

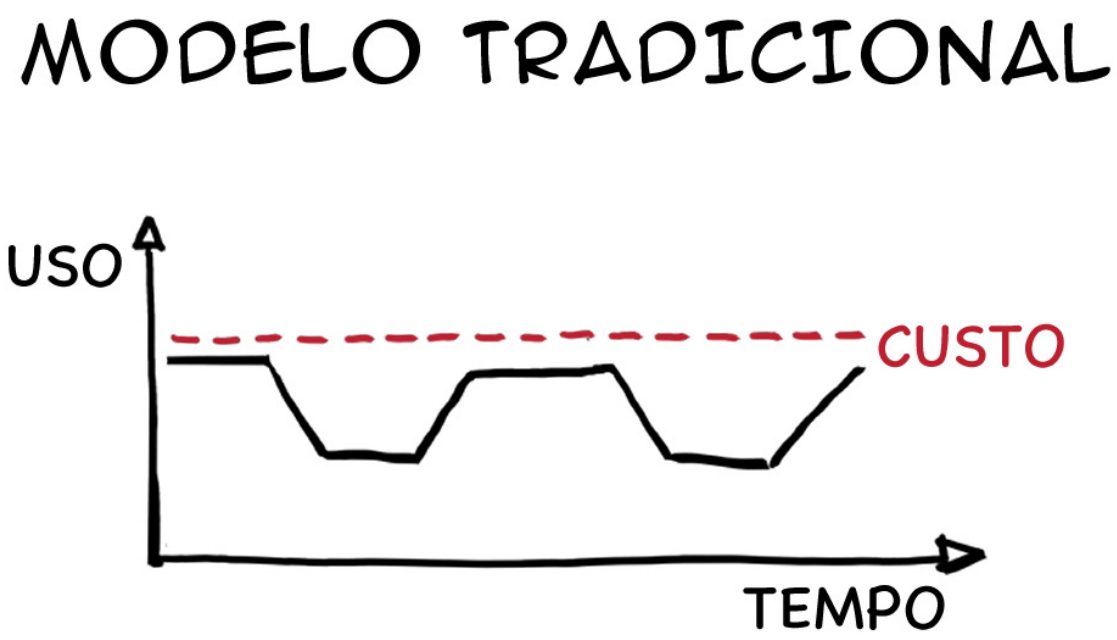
Com a Computação em Nuvem o jogo é diferente. Qualquer empresa, por menor que seja, pode ter acesso a todos os recursos disponíveis. O mercado do bairro pode usar a mesma tecnologia que dá suporte à *Netflix*. Sem investimentos nem custos fixos, tendo somente custos variáveis. Como a utilização de recursos da empresa pequena ou média é bem menor que a das grandes empresas, sua conta é sempre proporcional ao uso. É isso que viabiliza economicamente sua utilização.

Fala-se muito sobre a redução de custo e a agilidade que a Computação em Nuvem promove, até porque esses são atributos mais fáceis de entender. Mas, ao possibilitar que todas as empresas, independentemente de seu porte, façam uso dos mesmos recursos tecnológicos e da mesma infraestrutura que lhes dá suporte, a Computação em Nuvem possibilita que todas possam competir – o que não acontecia antigamente. Por isso a nuvem apresenta um poder transformador nos negócios muito maior a médio e longo prazo: uma startup hoje tem acesso exatamente aos mesmos recursos que o maior banco do país.

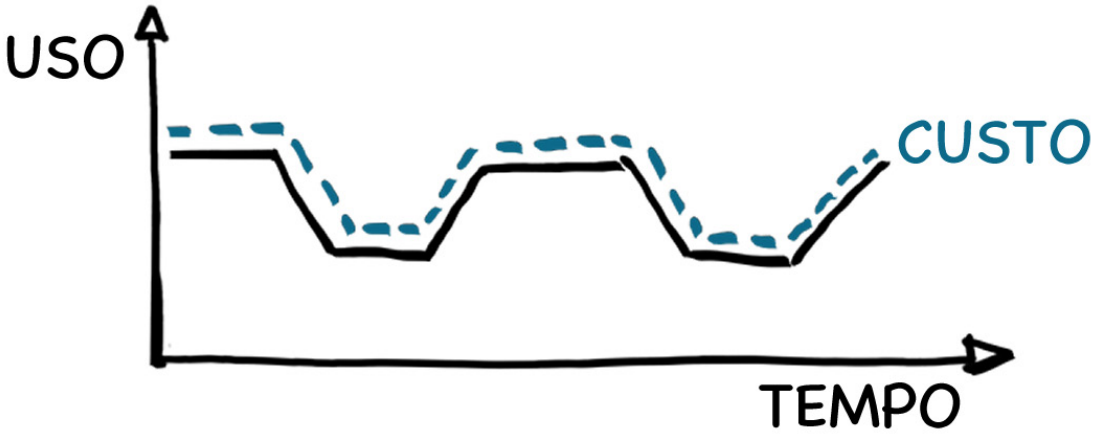
Bons motivos para adotar a Computação em Nuvem

Custo

Como você só paga pelo que realmente usa, o custo de uma infraestrutura na nuvem é na maioria das vezes menor do que o custo de uma infraestrutura convencional, em que você compra as máquinas para a sua empresa. O sistema não precisa funcionar de madrugada? Basta programar o desligamento das máquinas e deixar de pagar por elas nesse período. O volume de acessos está baixo? Troque a máquina que está usando por uma menor e pague menos. Simples assim. No jargão da Computação em Nuvem, esse modelo dinâmico de alocação e liberação de recursos é conhecido pelo termo *pagamento pelo uso* (“*pay-per-use*”), uma vez que só se paga pelo que é efetivamente utilizado.



MODELO DE NUVEM



Agilidade

Provedores de infraestrutura convencionais demoram dias, ou até semanas, para entregar novas máquinas, discos ou qualquer outro recurso para seus clientes. Observe que isso acontece independentemente de a infraestrutura ser própria, instalada em um *data center* local ao negócio, ou terceirizada, rodando em um provedor de hospedagem. Na Computação em Nuvem, o processo de alocação (ou diminuição) de novos recursos demora apenas alguns minutos. Basta escolher a máquina desejada, seu sistema operacional e parametrizar alguns dados: pronto, a nova máquina está no ar. Sem qualquer investimento.

Flexibilidade

Não precisa mais de uma máquina? É só desligar. Precisa de uma máquina maior? É só aumentar o tamanho. Mais disco? É só definir um novo tamanho para seu disco. Mais memória dinâmica? É só definir quanto é preciso. A qualquer momento. Instantaneamente.

Alta disponibilidade

Para que um ambiente computacional convencional garanta alta disponibilidade é necessário duplicá-lo, criando um ambiente de contingência que é acionado em caso de falha do principal. Claro, essa duplicação de recursos implica em desperdícios, uma vez que o ambiente de contingência fica ocioso a maior parte do tempo em condições normais. Mesmo não estando em operação, a infraestrutura de contingência demandou investimentos e está gerando despesas de depreciação, ocupando espaço, requerendo segurança e supervisão – ou seja, há geração de custos mesmo sem a infraestrutura estar no ar. Somente quando algo falha é que a infraestrutura adicional passa a ser utilizada.

Com a Computação em Nuvem, a garantia de alta disponibilidade passou a ser muito mais acessível[1]: o ambiente de contingência pode ser previamente configurado, mas mantido em formato reduzido e de baixo custo. Quando necessário, esse ambiente pode ser aumentado em poucos minutos, assumindo uma configuração semelhante ao do ambiente principal. A alta disponibilidade na nuvem tem custos expressivamente menores que quaisquer outras soluções, dentro da empresa ou em provedores onde se paga o aluguel de máquinas.

Por que a Computação em Nuvem é inevitável?

Pode-se falar muito sobre as vantagens tecnológicas da Computação em Nuvem em relação ao modelo tradicional. Mas o principal motivo pelo qual o movimento para a Computação em Nuvem é inevitável é baseado em sólidos fundamentos da Ciência Econômica, isto é, a principal vantagem dessa nova tecnologia é econômica. Vejamos a seguir.

A Computação em Nuvem favorece a Inovação

A Computação em Nuvem favorece a criação de novos negócios ou a

inovação em negócios tradicionais por eliminar a necessidade de grandes investimentos iniciais para a montagem de uma infraestrutura computacional para acomodar a nova iniciativa. Isso reduz os riscos, pois esses investimentos iniciais são evitados e não se tornam um “problema econômico“ caso o novo empreendimento não dê certo.

Se o novo negócio der certo, a agilidade da Computação em Nuvem para alocar ou liberar recursos de maneira quase instantânea permite que as empresas ajustem dinamicamente seus gastos com processamento de dados de acordo com a demanda, sem precisar provisionar recursos para uma necessidade que pode não se concretizar[2]. Dessa forma, o modelo de *pagamento pelo uso* da Computação em Nuvem, em que se paga apenas pelo que é efetivamente utilizado, permite a diminuição do chamado *custo enterrado* em novas iniciativas, além de garantir que o custo estará sempre ajustado à demanda.

Economia de escala[3]

A vantagem econômica advém da economia de escala resultante do uso mais eficiente dos recursos computacionais de propósito geral, que são compartilhados, o que não acontece quando se usa os mesmos recursos em *data centers* próprios ou terceirizados. Ora, nas Ciências Econômicas, o que caracteriza o conceito de *economia de escala* é justamente a organização do processo produtivo de forma a utilizar melhor os elementos de produção envolvidos nesse processo, o que resulta em diminuição de custos. No caso da Computação em Nuvem, essa organização dos processos é um problema do fornecedor do ambiente em nuvem – e não da sua empresa.

Na Computação em Nuvem, quanto mais o fornecedor organiza e otimiza a sua infraestrutura, mais seus custos diminuem, e a concorrência de mercado garante que essa diminuição seja normalmente repassada para os clientes de tempos em tempos. Ou seja, a economia de escala que acontece do lado dos fornecedores acaba sendo repassada para os clientes, gerando economia de escala também do lado da demanda.

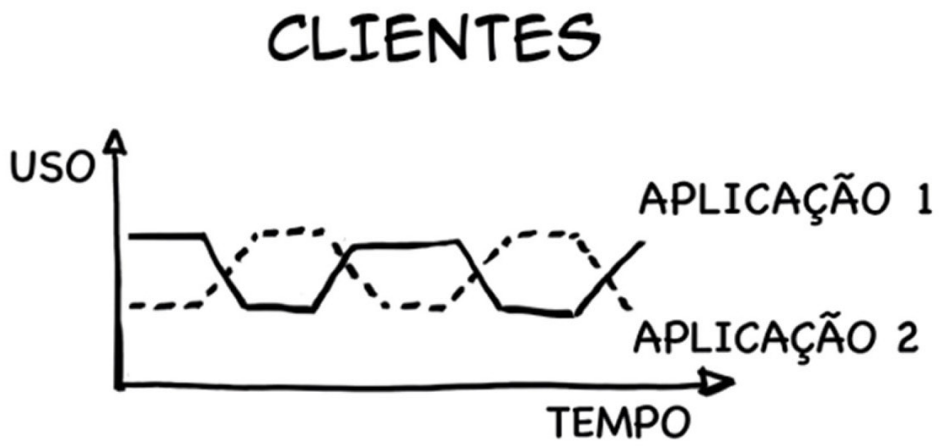
Economia de escala do lado dos fornecedores

Do lado dos fornecedores, quanto maior é o número de seus clientes, tanto maior é o volume de recursos necessários para atender a todos. Esse grande volume de concentração de recursos permite negociar melhores preços para adquirir novos servidores, comprar energia elétrica e outros insumos de produção. Além disso, o custo total do serviço de administração desses servidores, por ser dividido por um maior número de máquinas, também é minimizado. Por tudo isso, os custos da sua empresa com Computação em Nuvem são sensivelmente menores que os custos que você teria “in-house“ e mesmo o custo com provedores tradicionais de internet.

Economia de escala do lado da demanda

Do lado da demanda, a Computação em Nuvem permite agregar diferentes necessidades de processamento, suavizando os picos e vales de utilização dos equipamentos. Isso faz com que o uso médio dos servidores seja muito mais elevado do que o de servidores dedicados para fins específicos.

Por exemplo, existem determinadas aplicações que são mais usadas no horário comercial, enquanto outras são mais utilizadas à noite. No modelo tradicional, em que servidores são alocados para cada aplicação de maneira exclusiva, boa parte do tempo os servidores ficam ociosos. Com a Computação em Nuvem, é possível alocar os mesmos recursos computacionais para aplicações de perfis complementares de utilização.



O que incentiva e o que dificulta o uso da Computação em Nuvem

Vantagens

Estas são algumas das vantagens da Computação em Nuvem em relação ao modelo tradicional:

- Redução de investimentos iniciais e eliminação dos custos de manutenção, segurança, eletricidade, espaço e outros que seriam

necessários;

- Elasticidade e escalabilidade, isto é, a capacidade de se ajustar dinamicamente à demanda, esticando ou encolhendo a capacidade computacional em função do uso, inclusive dos recursos de Internet;
- Maior rapidez de implementação, incluindo tempo para aprovação de novas iniciativas, uma vez que não exigem grande investimento inicial, reduzindo os riscos empresarias;
- Agilidade para colocar novas aplicações no ar;
- Estímulo à colaboração entre departamentos da organização e também na cadeia de fornecimento.

Outros fatores indicados como incentivadores da adoção da nuvem já eram previstos na teoria, mas são menos óbvios e é interessante ressaltar que o mercado normalmente também os leva em consideração:

- Possibilidade de se testar e realizar pilotos de uma nova solução em um ambiente sem riscos antes de efetivar sua adoção em produção;
- Acesso a melhores ferramentas para rastreamento e auditoria dos sistemas e da integridade dos dados, sem ter que investir nessas ferramentas.

Dificuldades

Já em relação aos fatores que dificultam a adoção da Computação em Nuvem pelas organizações, alguns podem até ser surpreendentes:

- Necessidade de melhor integração entre os sistemas que rodam na nuvem e os sistemas que rodam internamente na organização;
- Necessidade de acesso estável à internet e com banda de comunicação adequada para o nível de uso, principalmente quando os sistemas produzem serviços para clientes internos;
- Resistência da equipe interna, que considera que esse tipo de serviço aumenta a complexidade do trabalho, seja o desenvolvimento de novos sistemas, seja a configuração da infraestrutura na nuvem;
- Resistência dos gestores de TI, que temem perda de controle sobre o ambiente operacional e também perda de sua importância dentro da organização, resultando em obsolescência de suas funções;
- Aspectos legais e de segurança – nesse caso, os gestores querem saber a localização física dos recursos computacionais e, especialmente, quais as práticas legais da jurisdição desse local;
- Reações negativas e céticas em relação ao termo *Computação em Nuvem*.

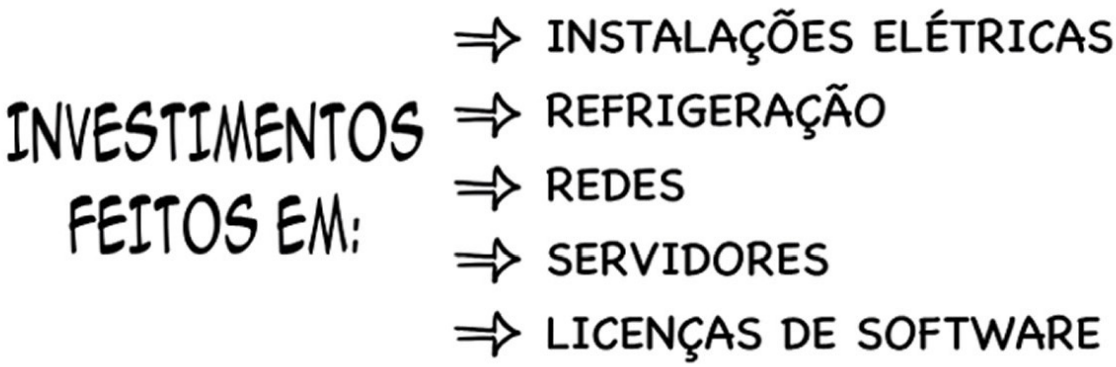
Em particular, vale a pena ressaltar o fator de *resistência dos gestores de TI*.

A Computação em Nuvem é evolução inevitável, e a principal prova disso é que os fornecedores tradicionais de tecnologia, como IBM, Microsoft, HP e Oracle têm concentrado investimentos massivos na criação de infraestruturas para oferta desse tipo de serviço no mundo inteiro. Esse movimento cria situações em que a inevitabilidade da mudança faz com que os mais ágeis aproveitem a onda para levá-los à frente antes dos outros, enquanto que os mais lentos perdem a oportunidade, construindo um futuro que confirma suas previsões mais pessimistas. Portanto, é uma questão de postura pessoal e profissional decidir de que grupo cada um quer fazer parte.

Custos enterrados: fator que deve ser considerado

Uma pesquisa da revista *Information Week* em 2013 buscou entender como os clientes corporativos enxergavam as soluções do pacote *Google in the Enterprise*, que é a oferta da empresa para sua suíte de aplicativos de automação para escritório completamente baseada na nuvem, que inclui editor de texto, e-mail e planilha eletrônica, entre outras aplicações.

Naturalmente, tendo como base a funcionalidade da solução alvo da pesquisa, os participantes focalizaram suas respostas na comparação entre a solução do Google e o pacote Office da Microsoft.



Independentemente do resultado – favorável para a Microsoft no curto prazo, mas apontando para uma competição mais acirrada nos próximos

anos – é importante ressaltar um aspecto evidenciado pela pesquisa: um número expressivo de usuários indicou que nem avalia ainda a solução do Google pelo fato de ter uma quantidade representativa de licenças da Microsoft cujos investimentos realizados precisam ser amortizados. E esse, em contrapartida, é um dos fatores menos analisados quando se fala em migração para a nuvem: os chamados “custos enterrados”, que são aqueles gastos já incorridos e irrecuperáveis.

Para as pequenas (e muitas médias) empresas, A Computação em Nuvem viabiliza soluções que antes não cabiam em seus orçamentos. Na maioria dos casos, a única alternativa seria a hospedagem em um *data center* tradicional, sem as vantagens da agilidade de alocação/desalocação de recursos e do *pagamento pelo uso*.

Entretanto, no caso das empresas grandes e médias que possuem *data centers* internos, quando se fala em migrar soluções para a nuvem, há que se considerar os tais custos já enterrados na construção da infraestrutura própria, incluindo sistemas de alimentação de energia, refrigeração e por aí vai. Além disso, há os equipamentos de rede, os servidores que já estão em operação e, principalmente, as licenças de software. Em certos casos, não é possível simplesmente transferir as licenças para os servidores que rodarão na nuvem - e essa avaliação deve ser feita com cuidado. Deve-se considerar ainda que, mesmo que se adote a nuvem apenas para as aplicações que mais se beneficiam das vantagens oferecidas por esse modelo, a estrutura interna provavelmente continuará sendo mantida para rodar outras aplicações que a empresa não tem interesse em migrar.

Em situações como essa, quando se fala em uma nova aplicação a ser implantada, mesmo que ela seja uma natural candidata à nuvem, pode ser muito tentador simplesmente comprar mais um servidor, aumentar um pouquinho a banda internet e tocar o barco assim mesmo, por ser a coisa mais fácil – e provavelmente mais barata – a fazer no curto prazo. Entretanto, essa forma de condução pode resultar em perda de competitividade ou menor retorno sobre os investimentos em TI logo ali na frente. O ideal é que as empresas que possuem infraestrutura própria definam uma estratégia clara em relação à Computação em Nuvem, baseada em uma classificação cuidadosa de seu portfólio de aplicações, dos

processos de negócio suportados por elas e por uma avaliação dos elementos que fundamentam a decisão de onde rodar cada tipo de aplicação.

---

1 Para uma discussão detalhada do tema, veja "Disaster Recovery as a Cloud Service: Economic Benefits & Deployment Challenges", de Wood et al.

2 Uma das situações em que a vantagem de custos da Computação em Nuvem é muito superior à de infraestruturas computacionais mantidas localmente é justamente quando a demanda pelos serviços é incerta. Veja "Above the Clouds: A Berkeley View of Cloud Computing", de Armbrust et al.

3 Uma abordagem detalhada sobre o tema é oferecida pelo artigo "The economics of the cloud", de Rolf Harms e Michael Yamartino.

Conceitos fundamentais

O que é Computação em Nuvem?

A final, o que é Computação em Nuvem? E o que não é? Desde o advento da internet, surgiram vários fornecedores que passaram a oferecer serviços de hospedagem e criaram *data centers* que absorveram parte expressiva do parque de equipamentos que antes ficava dentro das empresas. Mas isso *não é* Computação em Nuvem.

Quando de seu surgimento, e até que o conceito ficasse mais claro, a Computação em Nuvem foi caracterizada de maneira muito abrangente, incluindo toda e qualquer forma de virtualização de servidores e de terceirização de infraestrutura computacional. Dessa forma, durante algum tempo, o termo assumiu um caráter bastante genérico, e não caracterizava de maneira clara um modelo de funcionamento que permitisse identificar seus atributos e benefícios específicos. No entanto, à medida que as soluções oferecidas pelo mercado foram se consolidando, surgiram propostas de definição que caracterizam de maneira precisa o conceito de Computação em Nuvem.

Dentre as várias definições propostas, uma que vem tendo ampla aceitação pelo mercado e que é cada vez mais citada na literatura especializada é aquela proposta pelo NIST, o Instituto Nacional de Padrões e Tecnologia do Departamento de Comércio norte-americano, em 2011:

> A Computação em Nuvem é um modelo que permite acesso ubíquo[4], conveniente e sob demanda, via rede, para um conjunto compartilhado de recursos computacionais configuráveis (por exemplo, redes, servidores, armazenamento, aplicações e serviços) que podem ser alocados e liberados rapidamente com o mínimo esforço de gerenciamento ou interação com o provedor de serviços[5].

Por exemplo, um serviço de Computação em Nuvem que atenda à definição

proposta pelo NIST deve oferecer mecanismos automáticos para alocação de novos servidores através de regras que considerem a capacidade computacional em uso, algo como: “aloque um novo servidor sempre que o conjunto atual de servidores atingir 80% de sua capacidade máxima de processamento”.

Entretanto, o próprio NIST ressalta que o paradigma de Computação em Nuvem é um conceito que está em evolução. A definição que ele propõe não é definitiva e deverá evoluir ao longo do tempo.

Sob demanda

Em um serviço de nuvem, o sistema de aplicação consumidor deve ser capaz de alocar novos recursos automaticamente, sem interação humana com o provedor de serviços. Os recursos devem ser alocados e liberados de forma elástica, e de forma automática em alguns casos, permitindo a rápida adaptação ao aumento ou diminuição da demanda. Para a aplicação consumidora, os recursos disponíveis devem parecer ilimitados, sendo possível alocar a quantidade desejada desses recursos a qualquer momento.

Além disso, a empresa cliente pode reparametrizar as especificações do servidor, aumentando de modo fixo sua capacidade, se isso se mostrar realmente necessário e vantajoso do ponto de vista econômico, uma vez que o uso automático de recursos adicionais, se muito frequente, pode ser menos vantajoso que um aumento permanente de capacidade.

Assim, o serviço de Computação em Nuvem:

- Atende automaticamente a picos de demanda de processamento ou tráfego de dados;
- Não exige que se “encomendem” novos servidores;
- Não exige que o contrato de fornecimento de serviços seja alterado sempre que se deseje alterar os recursos computacionais disponíveis – inclusão ou remoção de servidores ou aumento de espaço em disco, por exemplo;
- Permite que os recursos sejam alocados, liberados ou reconfigurados sob demanda.

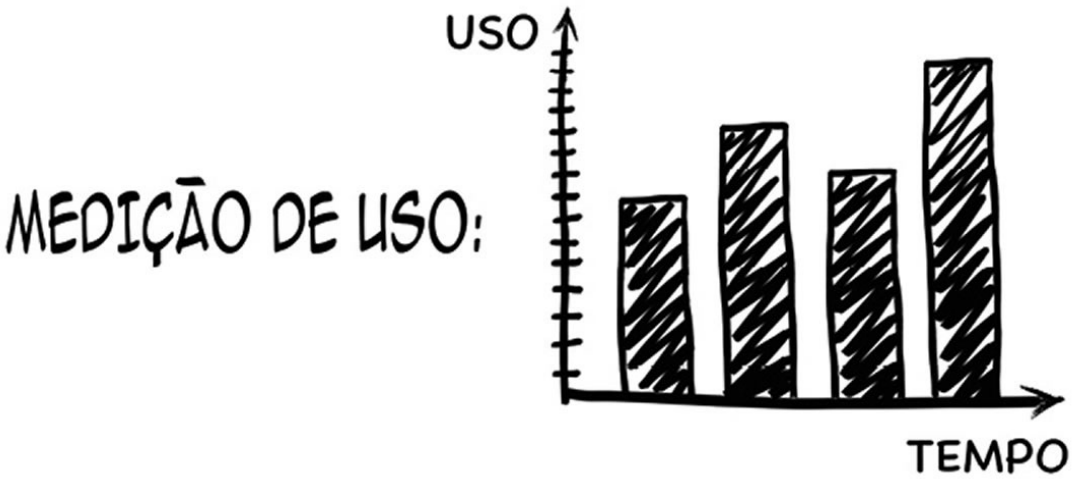
Acesso amplo

Os recursos devem estar disponíveis através da rede (internet) e devem ser acessíveis por mecanismos padrão, permitindo seu uso por diferentes dispositivos, tais como computadores pessoais, smartphones, tablets, etc.

Os recursos computacionais do provedor de serviços devem ser agrupados para servir a múltiplos clientes, com recursos físicos e virtuais sendo arranjados e rearranjados dinamicamente conforme a demanda desses clientes. Existe um senso de independência de localização, no qual o cliente

consumidor não tem um controle exato de onde os recursos utilizados estão localizados, mas deve ser possível especificar esse local em alto nível de abstração (país, unidade federativa ou *data center*).

Medição de uso



Os serviços de Computação em Nuvem devem controlar e otimizar os recursos de maneira automática, disponibilizando mecanismos de medição apropriados para o tipo de recurso utilizado (por exemplo, quantidade de espaço de armazenamento, velocidade de comunicação, capacidade de processamento, número de usuários ativos, etc.). Deve ser possível monitorar, controlar e consultar o uso dos recursos, oferecendo transparência tanto para o cliente quanto para o provedor dos serviços.

Provedores de Computação em Nuvem

Alguns dos principais provedores de Computação em Nuvem são:

- Amazon Web Services (AWS);
- Microsoft Azure;
- Google Cloud Platform;
- Softlayer (IBM);

É importante observar que todos esses provedores oferecem os recursos fundamentais que caracterizam a Computação em Nuvem. Infelizmente, alguns provedores de hospedagem e terceirização de *data centers* insistem em apresentar suas ofertas como “Computação em Nuvem”, o que dificulta ao mercado entender a diferença desse novo modelo para o tradicional. Portanto, vale lembrar novamente: se um provedor de serviços não oferece provisionamento e liberação de recursos sem intervenção humana, ou se a alocação de novos recursos (como servidores, espaço em disco) não é realizada em segundos ou poucos minutos, ele não está oferecendo Computação em Nuvem, mas ainda está usando o termo *nuvem* de forma genérica, sem proporcionar os benefícios do novo modelo.

Tipos de instâncias

Entender os modelos de compra de capacidade de processamento e a precificação de cada um deles pode parecer tarefa complicada para quem começa a percorrer o mundo da Computação em Nuvem. Porém essa sensação desaparece rapidamente assim que passamos a pensar de acordo com as regras da nuvem: não precisamos mais estocar recursos em forma de máquinas e discos para eventuais momentos de pico; podemos comprar capacidade sob demanda quando necessário.

Na Computação em Nuvem, os servidores podem ser alocados e desalocados dinamicamente, e rodam como máquinas virtuais no hardware do fornecedor de serviços. Quando em execução, as máquinas virtuais são chamadas de **instâncias**. As instâncias podem ter diferentes tipos, em função de seu modelo de cobrança e de sua forma de alocação. A AWS, por exemplo, oferece três tipos de instâncias: reservadas, sob demanda e *spot*. O Google, por sua vez, além da instância padrão também oferece instâncias com desconto para uso continuado. O Microsoft Azure tem as camadas Basic e Standard.

Vejamos como funcionam os tipos de instâncias da AWS.

Instâncias AWS sob demanda

São instâncias iniciadas quando necessário e o pagamento corresponde apenas à quantidade de horas utilizadas. Esse tipo de instância é a que reflete plenamente a flexibilidade da Computação em Nuvem. Em um momento de pico, em poucos minutos pode-se iniciar novas instâncias ou aumentar a capacidade das já existentes; além disso, num momento de baixo uso, as instâncias ociosas podem ser desligadas. A cobrança é feita por hora de uso, isto é, mesmo que uma instância seja utilizada por apenas um minuto, será cobrada a hora cheia.

Esse tipo de instância é interessante para aplicações que tenham um perfil de uso não constante ou imprevisível e que não possam ser interrompidas. Diante da necessidade de mais capacidade de processamento, instâncias sob demanda podem ser iniciadas, e se ficarem ociosas, podem ser desligadas. Tudo isso é possível sem ter de fazer um desembolso inicial para reserva de capacidade.

Instância AWS reservada

Ao comprar uma instância reservada, você paga um valor inicial para usá-la por um ou três anos. A taxa de utilização, paga por hora, é mais baixa do que a da instância sob demanda. A instância reservada estará sempre disponível. Embora esse tipo de instância seja mais econômica, em certo sentido ele contraria a flexibilidade do modelo de Computação em Nuvem, pois equivale a reservar recursos, como no caso da computação tradicional.

A desvantagem desse tipo de instância é que o usuário tem de desembolsar um bom dinheiro no pagamento inicial para reservá-la, além de perder a flexibilidade de aumentar a capacidade da instância ou de mudá-la de zona de disponibilidade (veja o conceito de zona de disponibilidade mais adiante) sempre que for conveniente.

A instância reservada pode ser usada por aplicações que exijam esse tipo de disponibilidade ou por aplicações que têm um uso constante e contínuo. Quanto menos variar a necessidade de processamento e quanto menos horas ela for subutilizada em um período, mais vantajosa é a compra de instâncias reservadas.

Instância spot

São as instâncias compradas da Amazon em uma espécie de leilão da capacidade ociosa. Você define qual o valor máximo por hora que está disposto a pagar por uma instância X e, caso a Amazon tenha capacidade de processamento ociosa, você leva a instância – geralmente por um valor bem abaixo do valor de uma instância sob demanda.

Como não existe almoço grátis, se a Amazon precisar de capacidade de processamento para iniciar instâncias sob demanda, ou se algum outro usuário oferecer um valor maior pela mesma instância (é um leilão!), você a perde na hora, sem aviso e sem dó. Por isso não é qualquer tipo de aplicação que está preparada para rodar em instâncias spot. As indicações de uso são:

- Aplicações que podem ser iniciadas ou terminadas a qualquer momento;
- Aplicações que só são viáveis com um custo bem baixo, e que podem esperar para realizar seu processamento quando houver disponibilidade de instâncias desse tipo;
- Para usuários com grande necessidade de processamento por um período muito curto.

Embora esses três modelos de compra de capacidade de processamento em nuvem da AWS tenham sido apresentados isoladamente, na prática a melhor relação custo/benefício para uma determinada aplicação encontra-se em uma combinação dos três modelos. Para chegar o mais próximo possível do ponto ótimo, não existe mágica. Um bom ponto de partida é a análise dos requisitos não funcionais da aplicação, como perfil de uso esperado, SLA, risco, etc. Com isto define-se uma combinação inicial que pode ser, então, monitorada e ajustada continuamente.

O modelo de instâncias do Google, por sua vez, apresenta algumas características diferentes que podem ser mais apropriadas, em termos de custo, para determinados tipos de aplicação.

Instância padrão do Google

Funciona como a instância sob demanda da AWS, porém difere na forma de cobrança. Ao invés de cobrar por hora de uso, a cobrança é por minuto, sendo que o mínimo é de 10 minutos. Por exemplo, se você usar 3 minutos, paga 10. Se usar 11 minutos e 25 segundos, paga 12 minutos.

De maneira geral, qualquer aplicação que precise realizar picos de processamento por curtos períodos de tempo podem se beneficiar desse modelo de cobrança. Ainda, desenvolvedores de software, que muitas vezes rodam instâncias para testar suas aplicações por períodos curtos, também se beneficiam dessa maior granularidade na contagem de tempo.

Instância com desconto para uso continuado do Google

Na verdade, esse não é um tipo diferente de instância, mas um modelo de cobrança diferenciado que favorece as instâncias sob demanda que rodam por períodos mais longos.

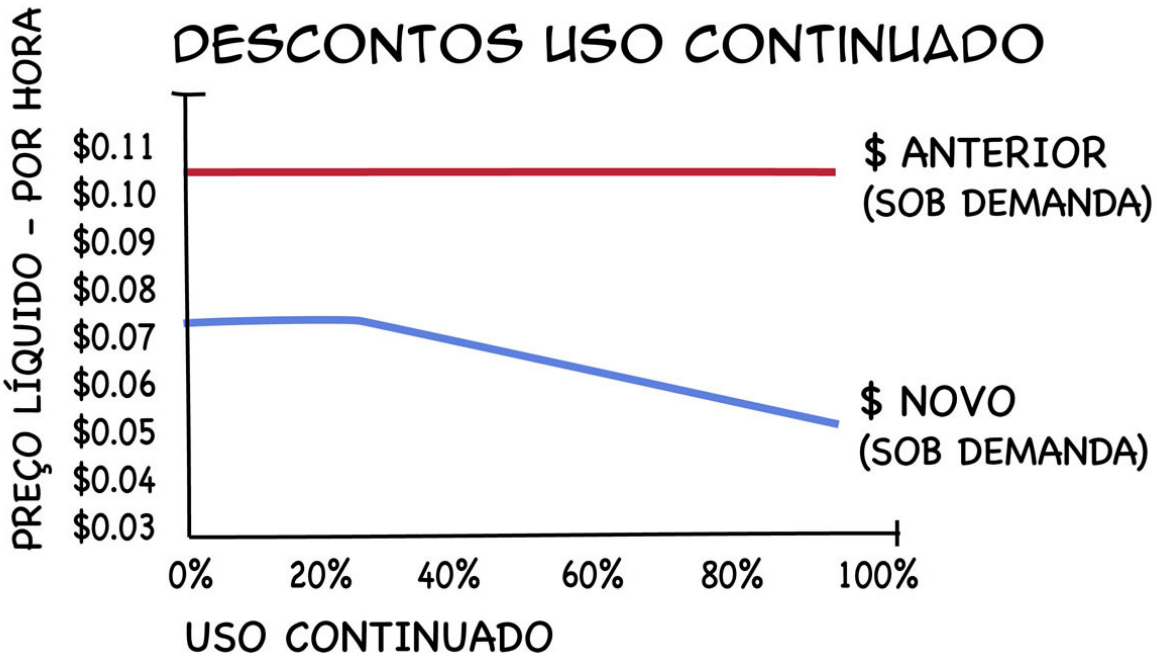
No modelo do Google, as instâncias sob demanda se beneficiam de descontos progressivos ao longo do tempo. O cálculo é feito considerando o mês de faturamento: uma instância sob demanda que rodar mais do que 25% do tempo de um mês começa a obter descontos. Por exemplo, considere uma instância que rodou o tempo equivalente 75% do mês:

- Os primeiros 25% de tempo são cobrados pelo preço normal da instância on demand;
- Os próximos 25% são cobrados com um desconto de 20%;
- Os próximos 25% são cobrados com um desconto de 40%.

Nesse exemplo, o desconto líquido total para a instância naquele mês é de 20% em relação ao preço cheio. O desconto máximo que pode ser obtido é de 30% do preço cheio.

Outro fato interessante é que o modelo de cobrança do Google agrupa as estatísticas de uso de suas instâncias não paralelas para fins de cálculo do desconto. Por exemplo, se você rodar uma instância por quinze dias, desligá-la e ligar outra com mesma configuração na mesma região, essa

segunda instância se beneficiará dos descontos progressivos, pois o Google considera que você está usando uma instância equivalente à primeira e aplica o desconto para uso continuado segundo as regras descritas acima.



Instâncias do Microsoft Azure

O Azure divide suas instâncias em duas famílias, que são chamadas de camadas: Basic e Standard.

A camada Basic oferece apenas instâncias para uso geral, e é indicada para aplicativos de produção de uma só instância, instâncias de desenvolvimento, servidores de teste e aplicativos de processamento em lote (batch). A camada Standard oferece instâncias para uso geral, para uso intenso de CPU ou para uso intenso de memória, e também possui mecanismos para redimensionamento automático de recursos computacionais (autoscaling) e balanceamento de carga. Por ter mais recursos, o preço desta camada é superior ao da Basic.

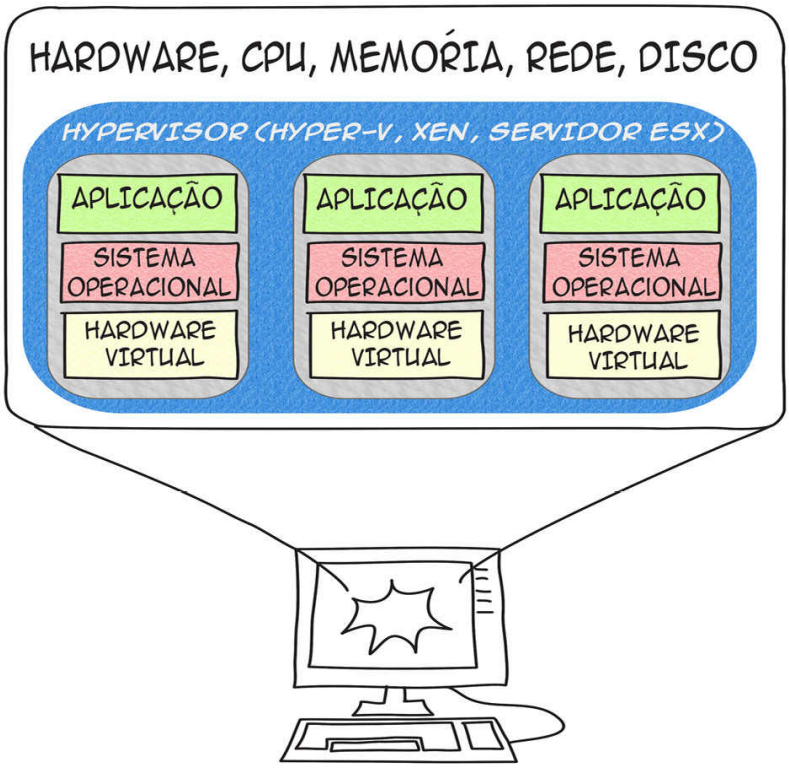
Embora os preços sejam apresentados por hora, o Azure cobra apenas os minutos efetivamente utilizados. A principal forma de cobrança é o pagamento pelo uso, ou seja, equivale a uma instância sob demanda.

Para empresas, existe a possibilidade de usar o Enterprise Agreement. O cliente estabelece um compromisso maior com plataforma Azure e passa a ter direito a descontos e outras condições adicionais. A empresa paga um valor antecipado para o consumo ao longo do ano. O que for excedido, será cobrado trimestralmente ou anualmente.

Virtualização X Computação em Nuvem

Virtualização de servidores é uma forma de se otimizar o uso de servidores físicos, fazendo com que vários servidores virtuais, sob o controle de um hipervisor (monitor das máquinas virtuais), possam rodar sobre o mesmo hardware[6]. A virtualização, ao simular ambientes autônomos em uma mesma máquina física, diminui a necessidade de hardware, de espaço físico e de energia. Além disso, os servidores virtuais, por poderem ser facilmente movidos entre diferentes máquinas físicas, favorecem a manutenção.

Já Computação em Nuvem abrange um conceito mais amplo: contém a ideia de *utility computing* (pagamento pelo uso e adaptação à demanda), em que usuários pagam pelo processamento, armazenamento e transferência de dados de acordo com o que efetivamente é utilizado, da mesma forma que se paga pela água, eletricidade e, de certa forma, telefonia.



A confusão entre os dois termos - virtualização e Computação em Nuvem - acontece porque a Computação em Nuvem usa a virtualização, ou seja, em cada máquina física do provedor de serviços de nuvem podem ser criadas várias máquinas virtuais, que são alocadas ou liberadas de acordo com a necessidade. Tipicamente, na Computação em Nuvem, a unidade básica de processamento é uma máquina virtual, e as máquinas virtuais são alocadas e liberadas pelo usuário de acordo com a demanda.

Computação em Nuvem e Virtualização não são, portanto, a mesma coisa. Há uma importante relação entre os dois conceitos, pelo fato de a virtualização ser uma das principais tecnologias de implementação do modelo de nuvem.

Elasticidade e Escalabilidade

O termo *elasticidade* teve origem nas áreas de Física e Economia, mas hoje também é bastante empregado na área de Computação.

Na Física, *elasticidade* é a propriedade que um material tem de retornar ao seu estado inicial depois de sofrer uma deformação. Um exemplo simples é dado por uma bola de futebol que, quando chutada, se deforma e, por possuir a virtude da elasticidade, retorna ao seu formato original – o que a impulsiona na direção dada pelo chute.

Na economia, *elasticidade* é o impacto que a alteração em uma variável causa em outra. Por exemplo, o quanto a alteração do preço de um produto se reflete no aumento ou diminuição de sua demanda.

Vemos que, nos dois casos, o conceito de elasticidade é praticamente intuitivo – mas pode ser descrito com precisão através de fórmulas puramente matemáticas.

Na área de Computação, *elasticidade* pode ser entendida como "a capacidade de um sistema se adaptar às alterações de carga de trabalho pela alocação e liberação de recursos, de maneira autônoma, de forma que, a qualquer momento, o conjunto de recursos utilizados é o mais compatível possível com a demanda instantânea"[7].

Observe-se, portanto, que na Computação a *elasticidade* diz respeito à adaptação automática de um sistema à variação da carga de trabalho de forma praticamente instantânea. Por exemplo, um site internet será elástico se for capaz de alocar e liberar recursos computacionais (servidores, discos, banda de comunicação) à medida que o número de usuários simultâneos aumenta ou diminui. Um ambiente computacional, por sua vez, será elástico se for capaz de proporcionar os recursos demandados pelos sistemas que rodam nele.

E é exatamente isso que a Computação em Nuvem proporciona: um ambiente elástico que permite a alocação e liberação dinâmica de recursos para os sistemas que rodam em sua infraestrutura. O ambiente de Computação em Nuvem gera a ilusão de que possui uma quantidade infinita de recursos para os sistemas que rodam nele. Sempre que um sistema demanda mais recursos, a nuvem deve ser capaz de provisioná-los.

O termo *escalabilidade*, por sua vez, pode ser conceituado como "a habilidade de um sistema de suportar com desempenho adequado cargas crescentes de trabalho, à medida que sejam adicionados novos recursos

computacionais”[8]. Ou seja, um sistema é dito *escalável* se for capaz de *fazer uso* de novos recursos computacionais disponibilizados para sua execução no sentido de acompanhar o aumento da demanda.

Portanto, a *elasticidade* diz respeito à capacidade de um sistema de “esticar” e “encolher” em termos de recursos utilizados em função da demanda. A *escalabilidade* é o atributo que indica que um sistema é capaz de fazer uso de recursos adicionais colocados à sua disposição. A *escalabilidade* é um pré-requisito para a *elasticidade*, uma vez que um sistema só pode ser elástico se for escalável.

Costuma-se dizer que a Computação em Nuvem oferece uma infraestrutura elástica que permite a construção de sistemas escaláveis. Porém, o fato de um sistema rodar na nuvem *não garante que ele seja escalável*. 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.

Na grande maioria dos casos, os sistemas desenvolvidos para rodar em ambientes computacionais convencionais não são escaláveis ou, quando possuem esse atributo, exigem intervenção manual de reconfiguração para poderem fazer uso de novos recursos. Dessa forma, sua execução no ambiente de nuvem não se beneficia completamente da elasticidade proporcionada. Quando um sistema é transferido de um ambiente convencional para a nuvem pode ser necessário adaptá-lo para que possa desfrutar das vantagens oferecidas pela Computação em Nuvem. No caso de sistemas que não possuem a característica de escalabilidade, essa adaptação pode ser bastante drástica, sendo necessário recriar o sistema quase completamente.

Em linhas gerais, um sistema a ser migrado de um ambiente computacional convencional para a nuvem pode ser classificado de acordo com seu grau de compatibilidade com a Computação em Nuvem[9]:

A. **Incompatível com a nuvem**: quando alguma de suas características o impede de ser executado no ambiente da nuvem;
B. **Compatível com a nuvem**: quando não há impeditivos para que seja

executado no ambiente da nuvem, embora algumas de suas características sejam incompatíveis com a nuvem;

C. **Pronto para a nuvem**: quando o sistema não possui nenhuma característica incompatível com a nuvem;

D. **Alinhado com a nuvem**: quando o sistema se beneficia da execução em nuvem – em termos de redução de custos ou uso de recursos específicos;

E. **Otimizado para a nuvem**: quando o sistema passa a explorar a elasticidade e os serviços exclusivos da nuvem, incluindo o paralelismo de operações proporcionado pelo ambiente, permitindo um uso ótimo dos recursos disponíveis.

Tipos de nuvem: pública, privada, híbrida

Quando se fala sobre as possíveis formas de implantação da Computação em Nuvem, os seguintes tipos são normalmente propostos[10]:

- Nuvem pública
- Nuvem privada
- Nuvem híbrida

Nuvem pública

É aquela oferecida pela internet por um provedor de serviços, em que os recursos computacionais são compartilhados pelos seus diversos clientes e o controle das instâncias, máquinas virtuais e recursos de processamento e armazenamento ficam completamente delegados ao provedor.

Nuvem privada

É aquela em que os recursos computacionais dedicados a uma determinada organização estão isolados dos utilizados por outras empresas. Não é necessariamente uma combinação de recursos computacionais em que os equipamentos pertencem e estão fisicamente alocados dentro de uma organização, como o nome pode dar a entender. Uma nuvem privada pode ser configurada em um provedor público.

Portanto, o termo “nuvem privada” pode significar duas coisas distintas:

- Uma nuvem criada na rede interna da empresa, em que a infraestrutura física é totalmente controlada e utilizada pela própria organização;

- Uma “nuvem privada virtual” (Virtual Private Cloud - VPC), em que a infraestrutura é controlada por um provedor de serviços, mas os recursos alocados para uma determinada organização são isolados dos recursos compartilhados pela nuvem pública.

Uma nuvem privada virtual oferecida por um provedor de serviços normalmente confere ao usuário os mesmos benefícios da nuvem pública, incluindo pagamento pelo uso e alocação e liberação de recursos computacionais sob demanda. Naturalmente, o maior controle oferecido por esse modelo acarreta também em maior esforço de gerenciamento quando comparado à nuvem pública.

Nuvem Híbrida

Já uma **nuvem híbrida** é constituída de uma junção de serviços de nuvem pública e privada. Por exemplo, a organização pode manter algumas aplicações na nuvem privada da empresa, e outras em serviços de nuvem pública ou privada virtual.

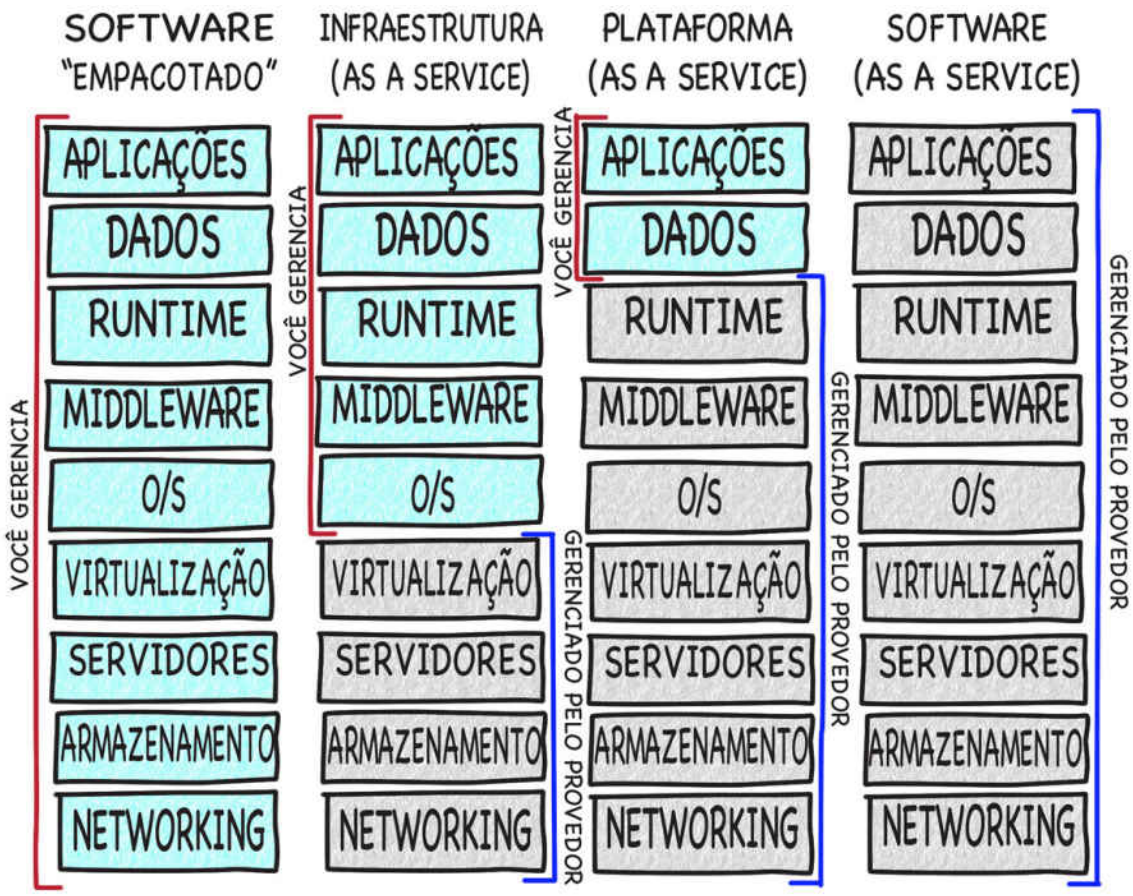
Em tese, uma nuvem privada oferece um grau maior de segurança do que uma nuvem pública, dado que o tráfego de informações e a migração de dados entre servidores virtuais e físicos é limitada aos recursos que estão sob controle direto da empresa cliente que a controla e administra. Entretanto, a questão de segurança dos sistemas está muito mais relacionada à sua arquitetura, mecanismos de proteção e utilização de técnicas de sigilo de dados do que ao tipo de nuvem no qual eles são executados. Ou seja, o uso de uma nuvem privada potencializa, mas não necessariamente oferece, de fato, maior segurança.

Tipos de serviços: IaaS, PaaS e SaaS

Os modelos de serviço de nuvem podem ser de três tipos:

- Infraestrutura como Serviço (IaaS – Infrastructure as a Service)
- Plataforma como Serviço (PaaS – Plataform as a Service)

> Software como Serviço (SaaS – Software as a Service)



Fonte: "IaaS, PaaS and Saas Terms Clearly Explained and Defined", Siverlight Hack.

Infraestrutura como Serviço (IaaS)

O serviço oferecido ao usuário é um conjunto de recursos computacionais básicos, tais como capacidade de processamento, armazenamento e redes, sobre os quais pode ser instalado e executado qualquer tipo de software, incluindo sistemas operacionais e aplicações. Neste caso, embora a infraestrutura de nuvem seja invisível para o usuário, ele pode controlar completamente os sistemas operacionais, espaço de armazenamento e aplicações alocados por ele. Exemplos desse tipo de serviço são o Amazon Web Services (AWS), o Google Compute Engine e o Microsoft Azure.

Plataforma como Serviço (PaaS)

O usuário pode instalar e gerenciar suas próprias aplicações, desenvolvidas por ele ou adquiridas de terceiros, utilizando as ferramentas e bibliotecas oferecidas pelo provedor. Ou seja, as aplicações que rodam numa plataforma como serviço são desenvolvidas especificamente para ela. Por exemplo, considere uma aplicação desenvolvida para a plataforma *Google App Engine* utilizando uma linguagem de programação padrão, digamos, Python. Para poder rodar em outra plataforma que suporte essa linguagem, como o *Heroku*, a aplicação precisaria ser adaptada. O uso de PaaS elimina a necessidade de comprar, configurar e gerenciar recursos de hardware e software. A infraestrutura é invisível para o desenvolvedor, mas ele pode configurar as aplicações e, eventualmente, aspectos referentes ao ambiente utilizado por elas. Além dos já citados Google App Engine e Heroku, outro exemplo de PaaS é o Microsoft Azure Cloud Services.

Software como Serviço (SaaS)

O usuário utiliza um software fornecido pelo provedor, e esse software roda em uma infraestrutura de Computação em Nuvem. A infraestrutura é invisível para o usuário, uma vez que o gerenciamento de recursos como espaço em disco, capacidade de rede, sistema operacional ou servidores fica a cargo do provedor de serviços. Um exemplo desse tipo de oferta é o Google Apps for Work, em que é possível criar e manter documentos, planilhas e apresentações nos servidores do provedor de serviços. Outros exemplos incluem o Microsoft Office 365 e o sistema de gestão de relacionamento com clientes (CRM – Customer Relationship Management) SalesForce.com.

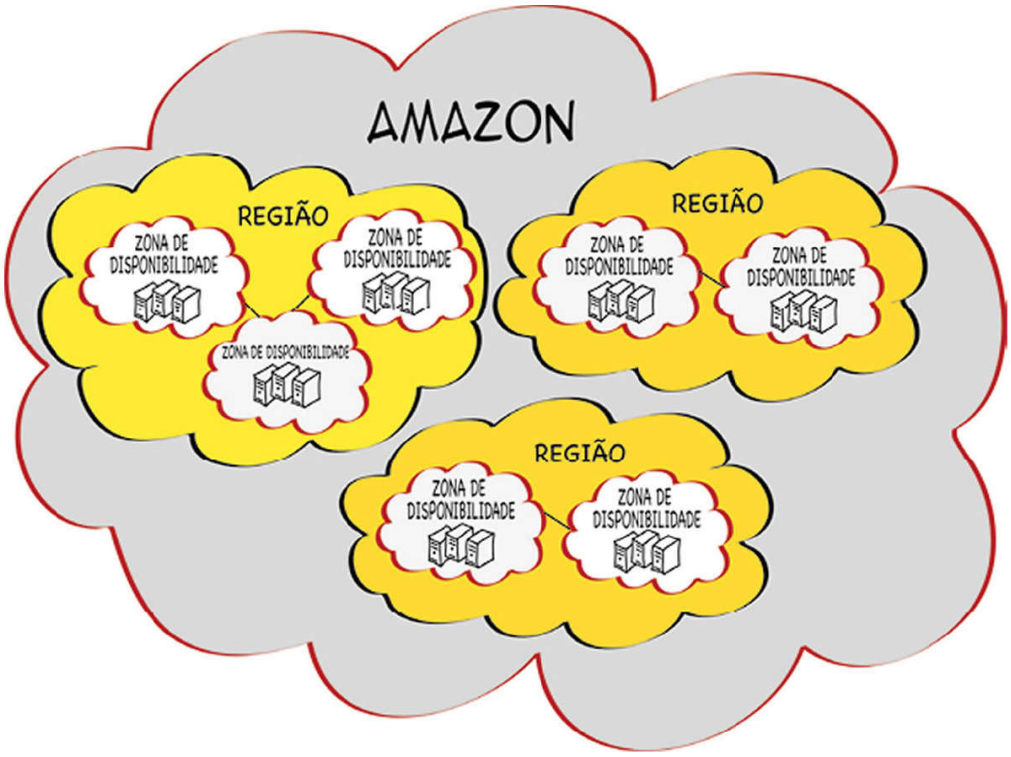
Regiões e zonas de disponibilidade

Os termos **região** e **zona de disponibilidade** são usados por um dos principais fornecedores de Computação em Nuvem, a Amazon Web Services (AWS). Embora sejam termos específicos desse fornecedor, podem ser úteis para se entender a estrutura global de um serviço em nuvem. Os outros fornecedores usam conceitos semelhantes.

Na AWS, as *regiões* estão distribuídas por vários locais do mundo, e cada uma delas é completamente independente e isolada das outras. Os recursos da AWS estão distribuídos em oito regiões[11]:

- Virgínia
- Califórnia
- Oregon
- São Paulo
- Irlanda
- Tóquio
- Singapura
- Sidney

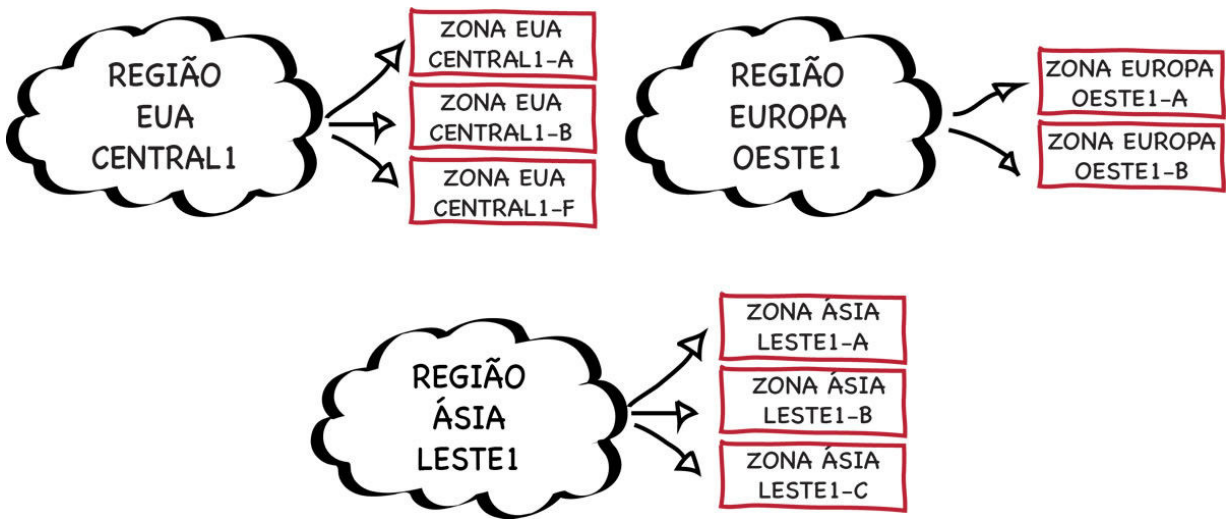
Cada região, por sua vez, oferece duas ou mais *zonas de disponibilidade*. Em São Paulo, por exemplo, há duas zonas de disponibilidade. Cada uma delas é um *data center* completo, com infraestrutura independente. Dentro de cada região, as zonas de disponibilidade são conectadas por links de baixa latência – isto é, velozes e com tempo de resposta baixo. A figura abaixo exemplifica a relação entre regiões e zonas de disponibilidade:



Essa estrutura, dividida em regiões e zonas de disponibilidade, possibilita a redundância dos recursos de maneira a garantir alta disponibilidade dos serviços e dados hospedados na nuvem. As múltiplas regiões permitem também que os serviços sejam espalhados geograficamente para atender de maneira mais veloz clientes de diferentes regiões do planeta.

O Google estrutura sua nuvem em três *regiões* – EUA, Europa e Ásia – e cada região possui pelo menos duas *zonas*, que equivalem às *zonas de disponibilidade* da AWS.

Diferentemente da AWS, o Google não é tão transparente em relação à sua estrutura. Ele possui três regiões[12], mas não dá maiores detalhes sobre onde seus *data centers* estão localizados.



Outro fornecedor é o Microsoft Azure, que possui *data centers* distribuídos em cinco áreas geográficas (*geos*) e 13 *regiões*[13]:

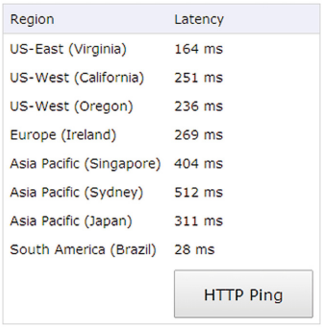
Estados Unidos:	**Europa:**	**Ásia - Pacífico:**	**Japão:**	**Brasil:**
Iowa	Ireland	Hong Kong	Saitama Prefecture	Sao Paulo State
Virginia	Netherlands	Singapore	Osaka Prefecture	
Virginia (2)				
Illinois				
Texas				
California				

GEOS E REGIÕES DO MICROSOFT AZURE

E faz diferença a região que você escolher para criar sua infraestrutura na nuvem? Sim, faz diferença. Um fator importante a ser considerado é a latência, isto é, o tempo de resposta a uma solicitação. Veja o resultado dos testes feitos com a AWS e o Microsoft Azure a partir do Brasil. Observe que a latência dos *data centers* localizados no Brasil é significativamente mais baixa:

Datacentre	Latency
South Brazil (Sao Paulo)	35 ms
East USA (Boydton, Virginia)	167 ms
North Central USA (Chicago, Illinois)	191 ms
West USA (California)	206 ms
West Europe (Amsterdam, Netherlands)	241 ms
North Europe (Dublin, Ireland)	253 ms
East Asia (Hong Kong, China)	421 ms
South East Asia (Singapore)	434 ms
East Japan (Saitama)	474 ms
West Japan (Osaka)	486 ms

Microsoft Azure: http://azureping.info/



Region	Latency
US-East (Virginia)	164 ms
US-West (California)	251 ms
US-West (Oregon)	236 ms
Europe (Ireland)	269 ms
Asia Pacific (Singapore)	404 ms
Asia Pacific (Sydney)	512 ms
Asia Pacific (Japan)	311 ms
South America (Brazil)	28 ms

HTTP Ping

AWS: http://www.cloudping.info/

Para otimizar a entrega de conteúdo, existem empresas que oferecem serviços de Content Delivery Network (CDN). Um CDN é um sistema de servidores distribuídos por diversos *data centers* ao redor do mundo, que tem como objetivo oferecer alta disponibilidade e alta velocidade do conteúdo a ser entregue ao usuário final. A ideia por trás desse serviço é distribuir cópias do conteúdo e entregá-lo a partir do *data center* que estiver mais próximo do usuário que faz a requisição. Uma das principais empresas especializadas nesse serviço é a Akamai. Entre os fornecedores de serviços de nuvem, alguns têm seu próprio serviço de CDN, como o Amazon CloudFront e o Microsoft Azure CDN.

Portanto, ao projetar a infraestrutura na nuvem é importante considerar dois elementos – latência e custo – de forma a escolher a região mais adequada para cada necessidade. Se a aplicação que vai rodar na nuvem exigir baixa latência, então é melhor escolher uma região mais próxima dos usuários. Caso a aplicação não tenha interatividade com o usuário e seja voltada para processamento de grandes lotes, vale a pena avaliar a diferença de preço entre as regiões, e eventualmente escolher uma região mais distante com preço menor. Além disso, atente para dois outros elementos importantes: a legislação sobre privacidade de dados do país que hospeda o *data center* e a cobrança de impostos que se aplicam às diferentes regiões.

Alta disponibilidade na nuvem

ma das primeiras coisas que pensamos quando ouvimos falar em nuvem é na alta disponibilidade das aplicações. Apesar de a associação ser válida, o simples fato de rodar uma aplicação na nuvem não garante que sua

U disponibilidade seja maior do que rodando em um *data center* tradicional. A nuvem entrega todas as ferramentas para tornar sua aplicação praticamente imune a falhas, mas cabe a você utilizar os recursos da nuvem de forma a tirar proveito das características do ambiente.

Ao planejar uma arquitetura para a nuvem, devemos primeiro levar em conta o *uptime* requerido, isto é, o tempo que sua aplicação deverá ficar rodando sem sofrer interrupções: 99,999%, 99,9% ou 99%? Essa é uma decisão importante, pois a receita para alta disponibilidade na nuvem é a mesma que a dos ambientes tradicionais: redundância. E redundância custa dinheiro.

Uma vez definido o *uptime* requerido, é preciso decidir em que nível haverá redundância. Em uma nuvem, as falhas podem ocorrer em cinco diferentes níveis[14]:

- Nível 1, recursos físicos
- Nível 2, recursos virtuais
- Nível 3, zonas de disponibilidade
- Nível 4, regiões
- Nível 5, provedor da nuvem

Nível 1 – Recursos físicos

Este nível envolve o conceito “n+1”: redundância de hardware, de *data center*, de rede, enfim, tudo que diz respeito à infraestrutura física de uma instalação. Esse tipo de redundância é usado nas instalações tradicionais; na nuvem, isso é de responsabilidade do fornecedor, e não está ao alcance do usuário.

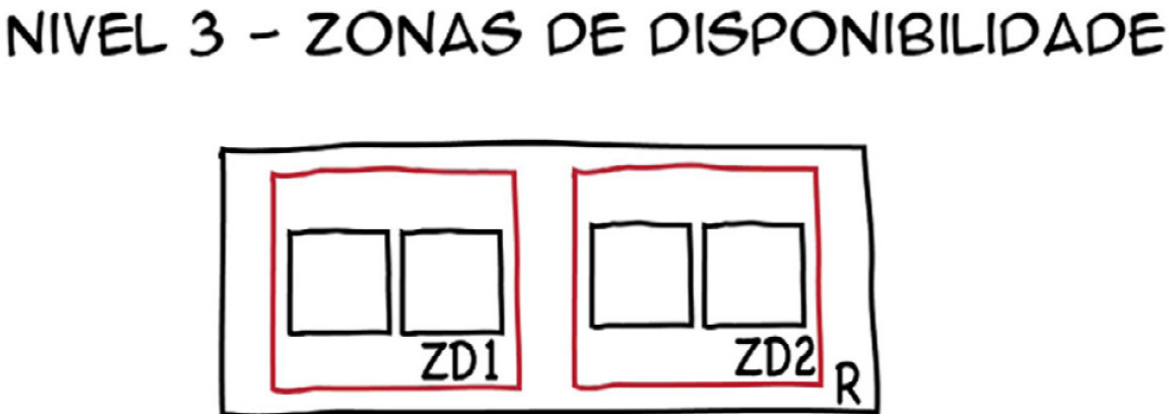
Nível 2 – Recursos virtuais

Para a maioria das aplicações, uma arquitetura distribuída com várias máquinas virtuais (instâncias) em uma mesma zona de disponibilidade

costuma ser suficiente. Nesses casos, para garantir tolerância a falhas devem ser utilizadas técnicas já tradicionais para o desenvolvimento de aplicações internet, como balanceamento de carga, arquitetura distribuída e replicação de dados.

Nível 3 – Zonas de disponibilidade

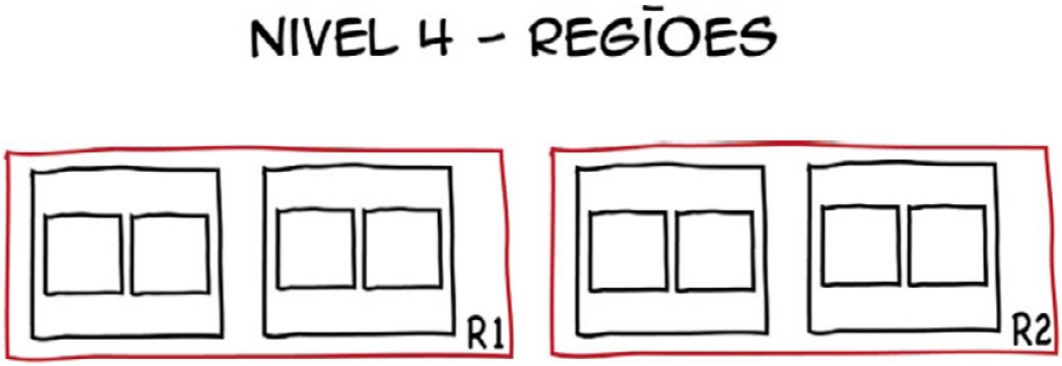
Para aplicações que exijam uma disponibilidade ainda maior, pode-se optar por uma arquitetura semelhante à do nível 2, mas com instâncias distribuídas em mais de uma zona de disponibilidade, dentro de uma mesma região. Isso acrescenta um grau de redundância, onde cada máquina virtual que tenha uma atribuição distinta em uma zona de disponibilidade precisa ser replicada na outra zona.



Nível 4 – Regiões

Para as aplicações realmente críticas, que devem continuar funcionando mesmo no caso de falhas de grandes proporções em regiões inteiras, podem-se construir aplicações distribuídas entre regiões distintas. Neste

caso, porém, o tráfego entre as regiões vai pela boa e velha internet, o que aumenta a complexidade do problema: construir aplicações distribuídas de grande porte pela internet não é tarefa trivial, mas é o preço a ser pago pela altíssima disponibilidade. Nesse caso, a aplicação precisa ser construída levando-se em consideração possíveis falhas de comunicação entre seus componentes, além da variação de latência nessa comunicação através da rede.



Nível 5 – Provedor de nuvem

Redundância de nuvem significa replicar a estrutura em mais de um fornecedor, o que permite sobreviver à perda completa de um provedor de nuvem. Naturalmente, além dos custos relacionados à completa redundância dos recursos de nuvem envolvidos, a complexidade desse tipo de solução inclui a construção de uma aplicação que seja capaz de rodar em dois ambientes completamente distintos, o que também aumenta o investimento.

Resumindo, a nuvem oferece ferramentas para a construção de aplicações de alta disponibilidade para todas as necessidades e bolsos. Entretanto, cabe ao desenvolvedor construir as aplicações considerando os recursos que expusemos, já que alta disponibilidade não vem gratuitamente só pelo fato de serem executadas na nuvem.

Aplicações que se beneficiam da Computação em Nuvem

Em princípio, a nuvem é um ambiente bastante democrático: qualquer aplicação pode ser colocada ali, independentemente de suas características e de como foi construída. Entretanto, há alguns tipos de aplicação que se beneficiam muito mais da Computação em Nuvem e da elasticidade de recursos oferecida por ela: são aquelas que possuem necessidades de processamento que variam significativamente de acordo com o tempo[15].

Aplicações com demanda variável

Exemplos comuns desta categoria são:

- Aplicações com uso intenso em horário comercial, mas que são pouco usadas fora desse período, como é o caso da maior parte das aplicações de uso interno das empresas;
- Portais de *e-commerce*, que apresentam aumento intenso de utilização às vésperas de datas comemorativas e depois voltam ao padrão normal de uso;
- Aplicações que automatizam tarefas atreladas a um calendário fixo, como portais de escolas (muito acessados quando da divulgação de notas ou no período de matrículas) ou de escritórios de contabilidade (em que é necessário calcular os impostos a pagar referentes ao mês anterior no início do mês seguinte).

Para tais aplicações, nos períodos de baixa demanda, poucos recursos da nuvem são utilizados. Conforme a demanda cresce, novos recursos podem ser adicionados à infraestrutura. Naturalmente, é possível automatizar através de scripts todo o trabalho de análise e alocação ou liberação de recursos.

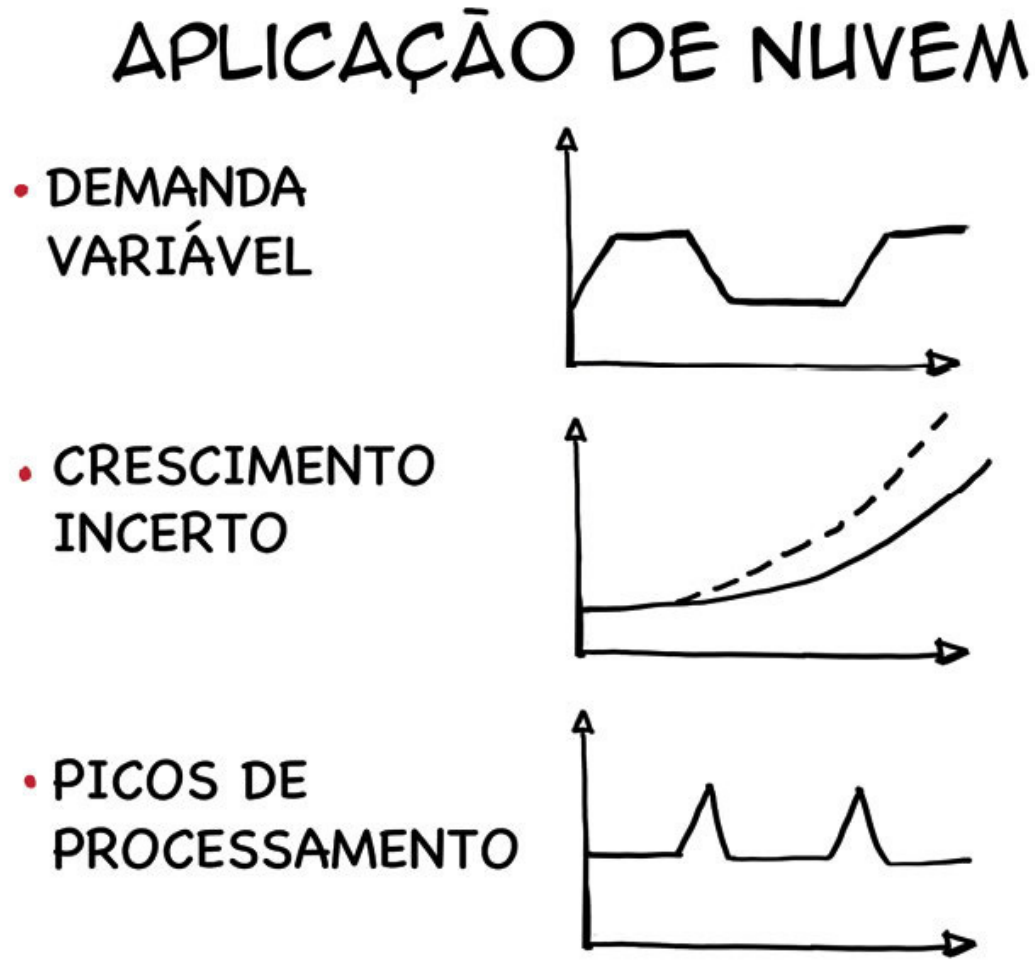
Aplicações com padrão de crescimento incerto

Esta é outra categoria que se beneficia bastante das características da Computação em Nuvem. Imagine uma empresa iniciante (*startup*) que lança um site com enorme potencial, mas que ainda é desconhecido. Deve-

se montar uma infraestrutura gigantesca que suporte o acesso de milhões de usuários? Além de custar caro, o site pode não ser tão bem sucedido quanto se espera. Por outro lado, se a infraestrutura for modesta e o site fizer muito sucesso, a empresa pode perder clientes por problemas no acesso. Na nuvem não é preciso se preocupar com o prévio dimensionamento. No início a estrutura pode ser pequena, com acréscimo de recursos de acordo com a necessidade. Em uma infraestrutura tradicional seriam necessárias várias semanas para a compra e preparo de novas máquinas. Na nuvem isso pode ser feito em minutos.

Aplicações com picos de processamento

Um exemplo deste tipo são as aplicações de processamento em lote (*batch applications*), como as usadas para cálculo de folha de pagamento. A demanda de processamento está fortemente concentrada nos dias anteriores ao fechamento da folha. A forma de resolver isso em uma infraestrutura tradicional é ter número suficiente de máquinas preparadas para suportar o pico de processamento naquele período, mesmo que elas fiquem ociosas a maior parte do tempo. Na nuvem, os recursos são alocados de acordo com a necessidade – e o melhor: 1 máquina por 24 horas custa o mesmo que 24 máquinas por 1 hora. Dessa forma, é possível alocar uma enorme quantidade de recursos para encurtar o tempo de processamento.



Esses três tipos de aplicação se beneficiam enormemente da elasticidade proporcionada pela Computação em Nuvem. O simples fato de executá-las em um ambiente que provê essa elasticidade já garante ganhos na forma de economia de recursos.

Porém, ganhos ainda mais significativos podem ser obtidos com o correto ajuste dessas aplicações e de seus ambientes para o modelo de Computação em Nuvem. Quando uma aplicação é desenvolvida para o modelo tradicional de computação centralizada, existe o pressuposto de que os recursos computacionais disponíveis são fixos. Dessa maneira, os programas que compõem a aplicação não consideram a possibilidade de se alocar e liberar recursos dinamicamente e, portanto, não consideram também que muitas das operações que realizam de maneira sequencial poderiam ser paralelizadas. Normalmente, depois que se realiza a migração de uma aplicação para a Computação em Nuvem, o passo seguinte ideal é

aperfeiçoá-la para que se beneficie ainda mais do novo modelo[16].

---

4 Que tem o dom da ubiquidade; que está ou pode estar em toda parte ao mesmo tempo; onipresente.

5 "The NIST definition of cloud computing", de Peter Mell e Timothy Grance, 2011.

6 Para conhecer mais profundamente o conceito, veja o artigo "Server Virtualization Architecture and Implementation", de Jeff Daniels.

7 Os conceitos de elasticidade e escalabilidade apresentados são aqueles apresentados no artigo "Elasticity in Cloud Computing: What It Is, and What It Is Not." de Herbst, Kounev e Roussner. No caso da escalabilidade, o conceito apresentado por esse artigo se limita à dimensão "desempenho". Para uma discussão mais aprofundada sobre o termo, veja "A framework for Modelling and Analysis of Software Systems Scalability." de Duboc, Rosenblum e Wicks.

8 Veja nota anterior.

9 A classificação apresentada vem do artigo "The CloudMIG Approach: Model-based Migration of Software Systems to Cloud-optimized Applications." de Sören Frey e Wilhelm Hasselbring.

10 Em "The NIST definition of cloud computing", de Peter Mell e Timothy Grance, é sugerida ainda a "nuvem comunitária", em que a infraestrutura de nuvem é compartilhada por uma comunidade específica de usuários de empresas que tenham interesses em comum.

11 Dados de 2014.

12 Dados de 2014.

13 Dados de 2014.

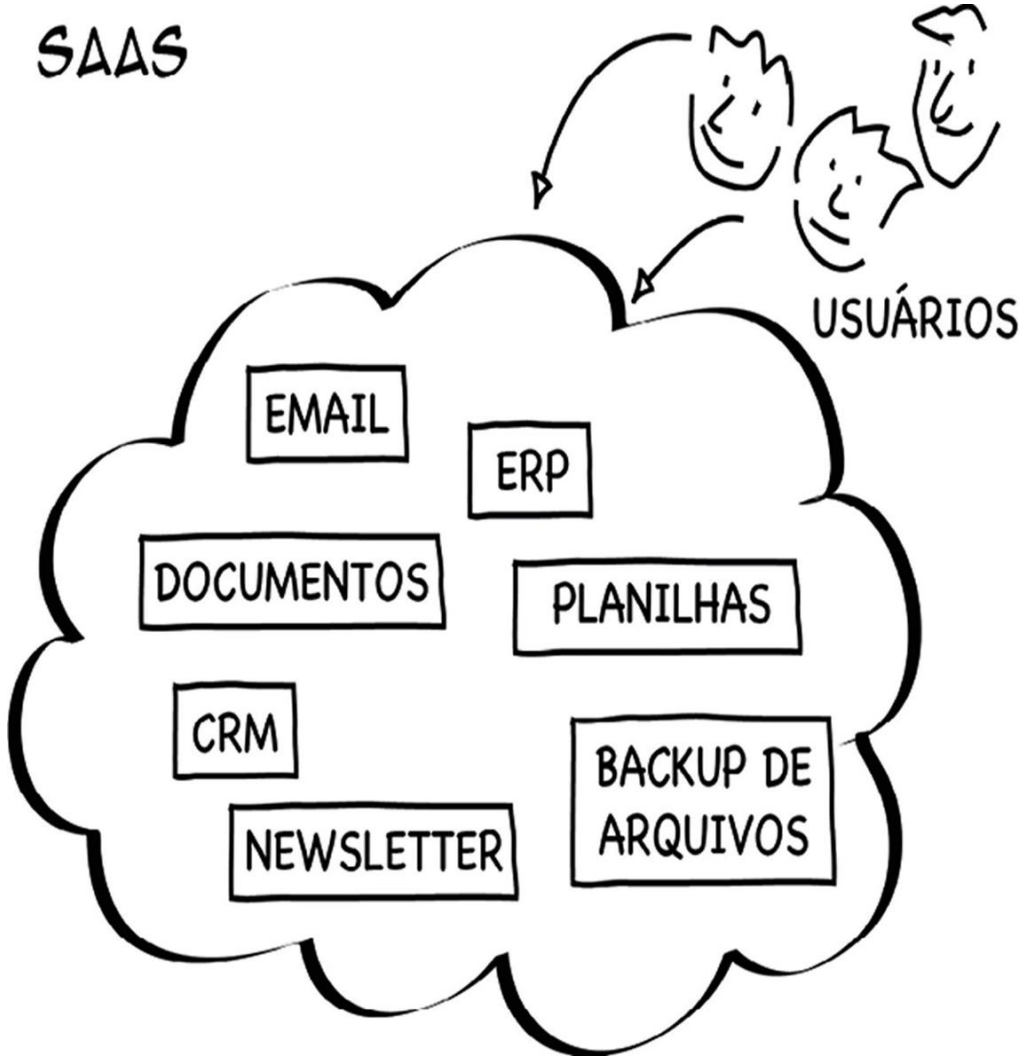
14 "Cloud application architectures: building applications and infrastructure in the cloud.", de George Reese.

15 "Above the Clouds: A Berkeley View of Cloud Computing", de Armbrust et al.

16 "The cloudmig approach: Model-based migration of software systems to cloud-optimized applications." de Sören Frey e Wilhelm Hasselbring.

Comprando Software como Serviço (SaaS)

Comprar Software como Serviço significa comprar apenas o acesso a uma aplicação, sem se preocupar com a infraestrutura que está por trás dela. Você será apenas o usuário do software, sem nenhuma tarefa de gerenciamento da infraestrutura necessária para executá-lo.



Qual o tamanho do disco? Qual sistema operacional é usado? Como os

servidores são configurados? Nada disso interessa ao usuário, que não precisa pensar em comprar equipamentos, instalar softwares, configurar servidores e redes; tudo isso é responsabilidade do fornecedor.

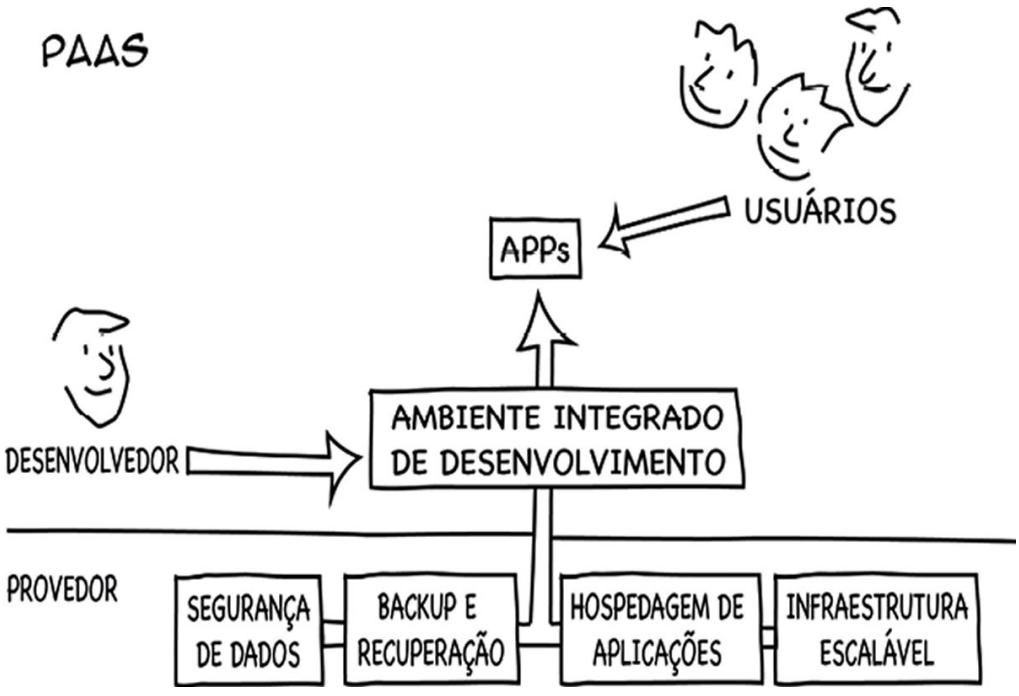
Um bom exemplo de Software como Serviço é o produto Google Apps, que oferece ferramentas como o Gmail, Agenda, editor de textos, planilha eletrônica e software para elaboração de apresentações, entre outros. Os usuários acessam as aplicações através de um navegador, de qualquer lugar e em qualquer dispositivo (computador, tablet, celular). Esses softwares podem ser usados gratuitamente por usuários que tenham uma conta Google, porém também é possível fazer uso corporativo através do Google Apps for Work. Os aplicativos podem ser personalizados para a empresa e o pagamento por usuário pode ser mensal ou anual.

Outro exemplo interessante de Software como Serviço é o SalesForce, um sistema de gestão do relacionamento com clientes (CRM – Customer Relationship Management) que roda inteiramente na nuvem. Dessa forma, a empresa que o utiliza pode reduzir investimentos, pois só paga um aluguel para usar o software, em vez de comprá-lo. A cobrança é feita por usuário/mês, faturado anualmente, e existem várias opções de planos.

De forma geral, a compra de Software como Serviço é um processo simples e uma vez realizada a assinatura, normalmente online, a aplicação já fica disponível para o uso. Alguns softwares permitem fazer configurações que personalizam a aplicação para a empresa e até para grupos de usuários, porém não é possível fazer customizações no sistema para adaptá-lo a necessidades muito específicas.

Comprando Plataforma como Serviço (PaaS)

Uma Plataforma como Serviço oferece um **ambiente de desenvolvimento de software. I**sso interessa diretamente àqueles que vão desenvolver novas aplicações para rodar na nuvem. Ao usar uma plataforma desse tipo, os desenvolvedores não precisam cuidar da administração do sistema operacional e de outras tarefas de gestão de infraestrutura[17]. Além disso, as plataformas oferecem muitos componentes já prontos para o desenvolvimento de aplicações. O modelo de PaaS também facilita muito o desenvolvimento de aplicações escaláveis e elásticas, uma vez que normalmente impõe aos programas a serem construídos uma arquitetura que garante esses atributos sem exigir do desenvolvedor um conhecimento detalhado dos mecanismos que a suportam[18].



Alguns fornecedores de PaaS oferecem serviços especializados, focalizados

em determinado tipo de desenvolvedor. Por exemplo, há fornecedores especializados em desenvolvimento de aplicações em linguagem Java. Outros têm como alvo os desenvolvedores que usam múltiplas linguagens, e oferecem suporte para PHP, Ruby, Python e outras ferramentas populares para desenvolvimento de aplicações web. Há também provedores de PaaS especializados na oferta de serviços voltados para o desenvolvimento de aplicações destinadas a interagir com dispositivos móveis.

Para nortear a escolha da plataforma a ser utilizada, é importante avaliar alguns elementos:

Linguagens de programação – A definição da linguagem que será usada no desenvolvimento das aplicações é fundamental, pois isso determina o modelo de programação que será usado e as ferramentas e componentes que devem estar disponíveis no ambiente.

Tecnologia usada no servidor – Por exemplo, se você pretende usar uma arquitetura baseada no ambiente .NET da Microsoft, faz sentido usar uma plataforma centrada nessa tecnologia. Por outro lado, se o projeto vai usar múltiplas linguagens e diferentes tecnologias do lado do servidor, o mais indicado é usar um PaaS que contemple esse cenário.

Armazenamento de dados – É importante dimensionar as necessidades de armazenamento e disponibilidade de dados da aplicação antes de escolher o fornecedor de PaaS. Se a aplicação exigir baixa latência, deve-se considerar uma solução com capacidade de prover grande número de operações por segundo. Por outro lado, se o fundamental para o tratamento de dados for a escalabilidade, então a escolha adequada pode ser um provedor que ofereça uma plataforma que suporte bancos de dados NoSQL[19].

Integração e suporte para ferramentas e aplicações – É interessante que a plataforma ofereça suporte e integração para ferramentas de desenvolvimento, tais como Visual Studio e Eclipse, além de ferramentas de gerenciamento de código, como Github. Além disso, é importante considerar como será a integração entre a aplicação que será desenvolvida na plataforma e outras aplicações. Os dados podem ser compartilhados com outras aplicações, ou será necessário exportá-los? É possível replicar os dados em outra base de dados automaticamente? Eventualmente será

preciso lidar com algumas restrições em função do PaaS escolhido.

Custo e orçamento – Naturalmente, é fundamental estimar o custo de executar as aplicações em um PaaS. Um serviço de PaaS provavelmente será mais caro do que um serviço de IaaS que ofereça uma capacidade de processamento equivalente. Deve-se avaliar se vale a pena gastar mais para não ter que fazer esse gerenciamento da infraestrutura, lembrando que a atividade de gerenciamento em si também acarreta em custos. Alguns ambientes de PaaS oferecem ferramentas de avaliação da eficiência da configuração de recursos para as aplicações, o que pode ser um instrumento importante para otimizar os custos.

Em seguida são apresentadas as principais características das plataformas de desenvolvimento oferecidas pelo Google e pela Microsoft. Embora nos concentremos nesses dois fornecedores, existem inúmeras plataformas no mercado. Algumas das mais populares são:

- AppFog
- Caspio
- Engine Yard
- Heroku
- Red Hat OpenShift
- Jelastic

Google App Engine

O Google App Engine é uma plataforma que permite construir e executar aplicativos utilizando a infraestrutura do Google. Para desenvolver aplicações no Google App Engine, o desenvolvedor utiliza o Kit de Desenvolvimento de Software (SDK– Software Development Kit) para as linguagens de programação Java, Phyton, PHP e Go[20], e pode acessar os recursos da plataforma através de interfaces de programação (API)[21] e bibliotecas[22].

Diversos serviços podem ser acessados pelo desenvolvedor através de uma interface web, a Admin Console, que permite gerenciar os aplicativos e ter acesso a detalhes sobre instâncias em execução, arquivos de log[23], atividades agendadas (Cron Jobs)[24], filas de tarefas, etc. Já a parte de infraestrutura fica por conta do Google, que também proporciona alocação automática de recursos e balanceamento de carga.

Uma aplicação desenvolvida no Google App Engine roda em uma *sandbox*, que é um ambiente isolado específico para cada aplicação. As aplicações não podem gravar dados no sistema de arquivos local, mas podem ser acessadas outras máquinas na internet para a gravação de arquivos desestruturados (de texto, por exemplo), através de funções de API oferecidas pela plataforma.

A aplicação pode ser executada de três formas: a partir de um pedido da web, de um evento na fila de tarefas ou de uma tarefa agendada. Dessa forma, as aplicações desenvolvidas para esse ambiente seguem um paradigma típico da Web, sendo estruturadas em torno de serviços implementados de maneira transacional, conforme preconiza a Arquitetura Orientada a Serviços (SOA – Software Oriented Architecture). Quando executada, cada transação deve retornar um resultado em até 60 segundos. Esse limite visa estimular o processamento modular em pequenas unidades, ou seja, sua aplicação pode ser projetada para quebrar o conjunto de dados em subunidades que possam ser processadas dentro do limite estabelecido. Esse tipo de processamento tira maior proveito da infraestrutura distribuída do Google. Por outro lado, se sua aplicação precisa processar grandes conjuntos de dados a partir de uma única requisição, pode ser que o Google App Engine não seja uma opção adequada, e que seja necessário recorrer a uma solução de IaaS, como o Google Compute Engine.

O modelo de arquitetura imposto pelo Google App Engine força que as aplicações sejam estruturadas com uma clara separação entre duas camadas, a de tratamento das requisições e a que lida com o armazenamento de dados. A primeira camada deverá necessariamente trabalhar sem salvar um estado interno entre as diferentes requisições, e a segunda camada manterá seus estados continuamente[25].

O Cloud Datastore é um banco de dados NoSQL que oferece o

armazenamento de dados semiestruturados e não estruturados e que não utiliza um “esquema de dados”. Permite o tratamento de grandes volumes de dados e gerencia o acesso aos dados inclusive em casos em que existam diversas instâncias trabalhando em paralelo. Oferece alta disponibilidade e transações atômicas.

Também é possível usar o Cloud SQL, que é um gerenciador de bancos de dados relacionais com os recursos e as funcionalidades do MySQL, com algumas características a mais e algumas restrições. É ideal para aplicações de pequeno e médio porte, mas também pode ser usado para aplicações de grande porte se for devidamente otimizado.

A cobrança dos serviços do Google App Engine é feita com base no uso. Cada aplicativo tem uma cota diária gratuita para instâncias, espaço de armazenamento e tráfego de entrada e saída. Quando a aplicação ultrapassa a cota gratuita diária, é feita a cobrança apenas pelo uso suplementar, até o montante máximo diário que for especificado.

Como no caso da maioria das plataformas oferecidas como serviços (PaaS), o Google App Engine é uma boa opção para os desenvolvedores que querem se concentrar essencialmente no desenvolvimento de aplicações e gastar o mínimo de tempo com a administração do sistema. O grande conjunto de funcionalidades disponíveis também acelera o desenvolvimento da aplicação, e o ambiente de desenvolvimento facilita a criação de aplicações escaláveis e elásticas para tratar grandes volumes de dados e um grande número de usuários simultâneos.

Microsoft Azure Cloud Services

A plataforma Microsoft Azure Cloud Services oferece um ambiente para execução de aplicações com base num modelo de programação bem definido, que visa proporcionar escalabilidade e alto grau de disponibilidade para as aplicações que rodam nela.

Embora suporte várias linguagens de programação populares como Java, Python, PHP, Ruby, JavaScript (através da plataforma node.js) e as linguagens do ambiente .NET (como C#), a plataforma é mais familiar para

os desenvolvedores acostumados ao modelo de desenvolvimento de aplicações e às bibliotecas da Microsoft. A empresa oferece um kit de desenvolvimento de software (SDK) específico para cada linguagem suportada, e a ferramenta de desenvolvimento Visual Studio também pode ser utilizada, inclusive com uma versão online.

Tipicamente, uma aplicação Azure Cloud Services deve ser estruturada em torno de dois papéis fundamentais: *web role* e *worker role*. São criadas máquinas virtuais distintas para cada papel, e a ideia é combinar múltiplas instâncias de cada uma para criar aplicações multicamada. Instâncias *web role* implementam o código responsável pelo tratamento de requisições vindas da web, como requisições de usuários, e normalmente as encaminha para instâncias *worker role*. Também é comum utilizar instâncias *web role* para implementar *Web Services*, que são serviços oferecidos para outras aplicações através da internet. Instâncias *worker role* rodam tarefas assíncronas que não dependem de interação com o usuário e que normalmente implementam as regras de negócio e o tratamento de dados da aplicação. A comunicação entre os dois tipos de instâncias é realizada através de filas de mensagens.

O modelo de programação do Azure Cloud Services garante que a aplicação possa suportar altas cargas de trabalho, permitindo aumentar ou diminuir automaticamente o número de instâncias de cada tipo de acordo com a demanda. Se alguma instância falhar, será automaticamente reinicializada, o que garante que a aplicação seja rapidamente recuperada. Por outro lado, o modelo impõem algumas restrições para garantir a escalabilidade e tolerância a falhas. Por exemplo, a aplicação não deve armazenar informações no sistema de arquivos de sua máquina virtual, porque em caso de falha ou queda da máquina virtual tais informações serão perdidas.

É possível flexibilizar a rigidez desse modelo de programação. Uma aplicação pode ser constituída, por exemplo, apenas de instâncias de um determinado tipo (*web role* ou *worker role*). Além disso, as aplicações podem interagir com outras que rodam fora do ambiente controlado pelo Azure Cloud Services, utilizando o serviço de barramento (Azure Service Bus) ou de filas de mensagens simples (objetos Queue do Azure Storage).

Para o armazenamento de dados através de recursos gerenciados pela

própria plataforma, existem quatro opções: SQL Database, **Table Storage**, DocumentDB e ***blobs***.

O **SQL Database** oferece a maioria dos recursos do consagrado gerenciador de bancos de dados relacionais Microsoft SQL Server, mas sem a sobrecarga de administração dos bancos de dados. O SQL Database também suporta grande parte das construções da linguagem de definição e manipulação de dados Transact-SQL, que já é bem conhecida pelos desenvolvedores acostumados com o Microsoft SQL Server.

O **Table Storage** é um banco de dados NoSQL do tipo chave-valor. Diferentemente de um banco de dados relacional, bancos de dados NoSQL típicos não controlam nem restringem os tipos de dados que podem ser armazenados nele e nem layout das tabelas de armazenamento, isto é, não utilizam "esquemas de dados". Bancos de dados do tipo chave-valor permitem o armazenamento de dados em qualquer formato, e atribuem um código (chave) a cada conjunto de dados gravado, permitindo depois a recuperação através desse código. A vantagem desse modelo é a flexibilidade que oferece, além da grande escalabilidade no acesso ou gravação dos dados. Além disso, o custo de armazenamento no Table Storage é menor do que no SQL Database. O Table Storage é um dos tipos específicos de objetos manipulados pelo serviço de armazenamento de objetos **Azure Storage**, e possui os mesmos atributos de redundância, durabilidade e baixo custo oferecidos pelo serviço.

O **DocumentDB** é um banco de dados NoSQL orientado a documentos que oferece escalabilidade, alta disponibilidade e excelente desempenho para tratamento de grandes volumes de dados. A principal diferença em relação ao Table Storage é que, em vez de manipular apenas registros do tipo chave-valor, o DocumentDB armazena dados semiestruturados (em formato *JSON*[26], familiar para quem desenvolve sites e interfaces de usuário utilizando a linguagem JavaScript). Cada elemento armazenado nas tabelas do banco de dados é identificado por um nome de propriedade. Cada valor armazenado na tabela é identificado por um nome de propriedade. O nome de propriedade pode ser usado posteriormente para especificar critérios de seleção sobre os dados. Uma coleção de propriedades e seus valores formam uma entidade. Pelo fato de também não utilizar "esquemas de

dados", o DocumentDB permite que duas entidades que contenham diferentes coleções de propriedades sejam armazenadas na mesma tabela.

O DocumentDB oferece recursos familiares para os desenvolvedores habituados aos bancos de dados relacionais tais como funções definidas pelo usuário (escritas usando a linguagem JavaScript), suporte a transações, gatilhos[27], *stored procedures*[28] e realização de consultas complexas utilizando um dialeto da linguagem SQL[29].

Blobs, ou "*binary large objects*", são objetos de armazenamento não estruturados destinados a armazenar dados binários arbitrários tais como documentos, imagens e vídeos. Assim como o Table Storage, o armazenamento de *blobs* também é proporcionado pelo serviço **Azure Storage**, que oferece redundância e durabilidade garantida das informações gravadas.

Finalmente, é importante observar que o modelo de programação do Azure Cloud Services também é suportado pelo sistema operacional Windows Server tradicional que roda localmente em data centers ou nuvens privadas. Isso permite que uma aplicação seja desenvolvida para rodar localmente dentro de uma organização e, mais tarde, seja levada para a nuvem sem precisar de nenhuma alteração. Essa característica faz parte da estratégia da Microsoft de apostar em nuvens híbridas, com as aplicações podendo rodar em nuvens privadas ou em sua nuvem pública de maneira indistinta e até colaborativa.

Os preços do Azure Cloud Services são baseados no tamanho das instâncias em execução, e também pode variar de acordo com a largura de banda e volume de dados trafegados e o tamanho dos bancos de dados.

---

17 "Developing Software Online with Platform-as-a-Service Technology.", de George Lawton.

18 "Above the Clouds: A Berkeley View of Cloud Computing", de Armbrust et al.

19 Convencionou-se chamar de NoSQL os bancos de dados que não usam a linguagem de consulta SQL e não utilizam esquemas de dados, isto é, que não exigem a pré-definição do formato das tabelas de armazenamento, flexibilizando o tipo de informações que podem armazenar (dados semiestruturados e desestruturados). Normalmente, oferecem alto desempenho atingido através de grande paralelismo de operações, e são capazes de armazenar e processar grandes volumes de dados.

20 Go é uma linguagem de programação de código aberto (open source) desenvolvida originalmente pelo Google.

21 Application Programming Interface: Conjunto de rotinas e convenções através das quais um software oferece sua funcionalidade para utilização por outros programas. Por exemplo, uma aplicação pode ativar a funcionalidade de tradução entre idiomas do Google Translate através da API oferecida por ele.

22 Uma biblioteca é um conjunto de funcionalidades prontas que podem ser adicionadas a uma aplicação. Por exemplo, uma biblioteca de funções matemáticas podem oferecer diversas rotinas de cálculo que podem ser integradas a uma aplicação, gerando reaproveitamento de funcionalidade e, portanto, diminuindo o esforço necessário para construir essa aplicação.

23 Arquivos de log são usados para registrar eventos ocorridos com um software. São úteis para entender o comportamento do software para detecção de falhas e otimização de desempenho, bem como para registrar informações de auditoria.

24 Cron Jobs são tarefas que rodam no ambiente do Google App Engine que podem ser agendadas para rodar em um horário específico ou periodicamente em intervalos regulares.

25 "Above the Clouds: A Berkeley View of Cloud Computing", de Armbrust et al.

26 JSON é o acrônimo de JavaScript Object Notation, que é um formato simples para descrição e intercâmbio de dados.

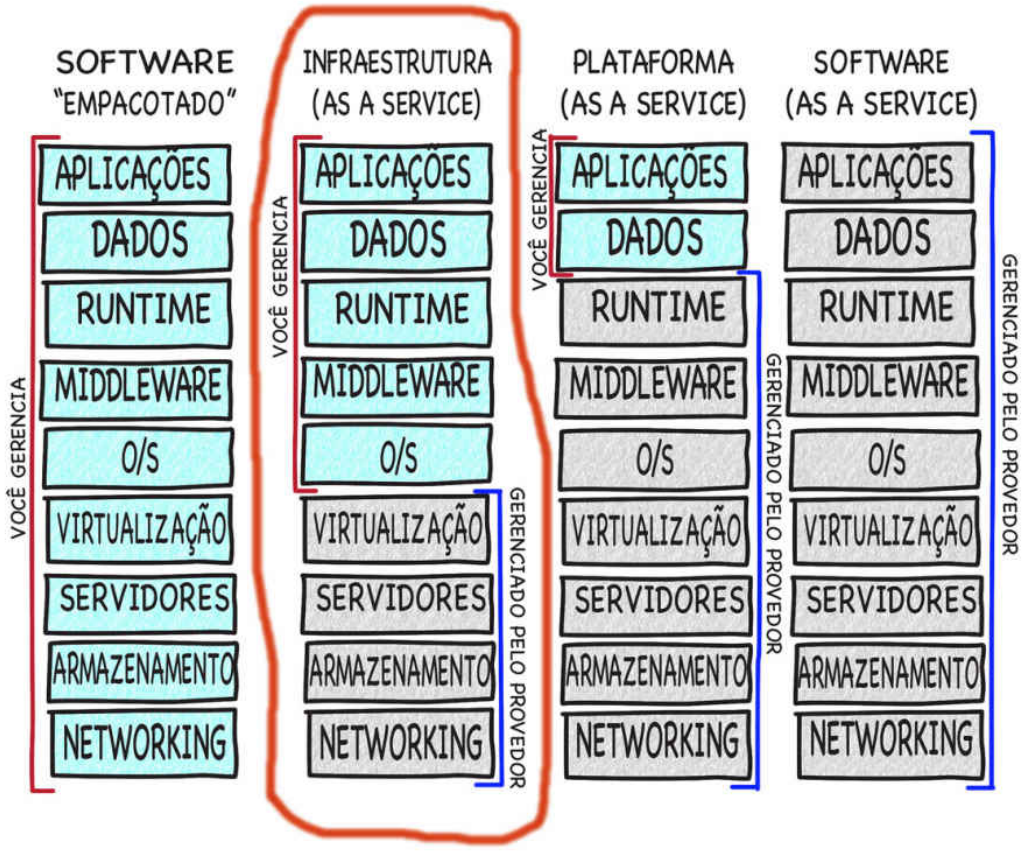
27 Gatilho ou trigger é um recurso para programação de bancos de dados que permite a execução de um trecho de código sempre que ocorre um determinado evento. São comumente utilizados para realizar verificações de consistência sempre que uma informação é gravada no banco de dados, ou para propagar para outras tabelas de armazenamento uma alteração realizada sobre uma determinada tabela.

28 Stored procedures ou Procedimentos armazenados são rotinas escritas em linguagem SQL e disponíveis para aplicações que acessam bancos de dados relacionais. Seu nome advém do fato de essas rotinas ficarem armazenadas diretamente no banco dados.

29 Structured Query Language (Linguagem de Consulta Estruturada) ou simplesmente SQL é uma linguagem padrão para definir e manipular dados tipicamente utilizada por gerenciadores de bancos de dados relacionais.

Comprando Infraestrutura como Serviço (IaaS)

A Computação em Nuvem baseia-se na alocação e desalocação de recursos de acordo com a demanda. No caso da Infraestrutura como Serviço, alocam-se principalmente recursos para processamento e armazenamento de dados. É preciso definir o tipo e o tamanho de cada máquina virtual, qual a sua capacidade de processamento, e como será feito o armazenamento de dados. Para tirar proveito de uma Infraestrutura como Serviço é fundamental uma boa integração entre esses diferentes tipos de recursos.



Os elementos que devem ser considerados na compra de infraestrutura na

nuvem são os seguintes:

- Configuração de servidores
- Armazenamento de dados
- Outros serviços:
  - Banda Internet
  - Tráfego de E/S
  - Softwares
  - Controle de acesso
  - Facilidade de gerenciamento
- Custos

Configuração de servidores

A compra de Infraestrutura como Serviço tem como ponto de partida a escolha da configuração de um servidor, que também é chamado de “instância” ou de “máquina virtual”. Do que você precisa: de uma máquina para uso geral, com mais demanda de memória ou de uso de CPU? Os fornecedores oferecem algumas opções para cada caso, que variam em tamanho e custo.

A escolha da região geográfica onde o servidor será criado faz toda diferença no tempo de latência, que impacta diretamente no desempenho de uso. Quanto mais distante dos usuários for a região, mais demorado será o tempo de resposta. Então, se a aplicação tiver muita interação, o melhor é que os servidores sejam criados numa região mais próxima dos usuários; caso os usuários estejam distribuídos geograficamente, podem-se criar instâncias em mais de uma região. Se a aplicação for apenas de processamento, sem interação, pode ser interessante escolher uma região onde os preços sejam mais baixos, mesmo que seja mais distante. Entretanto, para uma avaliação precisa dos custos envolvidos, considere a legislação e a incidência de impostos que são aplicáveis a cada região.

Armazenamento de dados

É preciso definir também quanto você vai precisar de espaço em disco. É importante observar que, normalmente, cada instância possui um disco temporário associado a ela, que pode ser usado durante o processamento dos dados, mas que não será mantido quando a instância for liberada – a ativação e desativação de instâncias são operações muito comuns na nuvem. Portanto, os discos temporários não podem ser usados para armazenar dados que devem permanecer após o processamento. A opção mais comum é utilizar um *disco persistente*, que será usado em conjunto com uma instância, e que será preservado ainda que a instância seja desativada. Essa opção é muito semelhante ao modelo de computação tradicional , em que um servidor físico possui um disco que é usado para armazenar os dados processados.

Além dos discos persistentes, que podem ser usados da maneira tradicional, os fornecedores de serviços de Computação em Nuvem normalmente oferecem uma opção de armazenamento bem mais barata, que tem menor desempenho em termos de velocidade de acesso, mas que pode ser usada como um repositório para qualquer tipo de dados, que é o **serviço de armazenamento de objetos**. Essa opção, combinada com o uso de discos persistentes, permite uma maior flexibilidade na estratégia de armazenamento de dados utilizada. O mais comum é manter os dados e arquivos de processamento mais frequente nos discos persistentes, e guardar no serviço de armazenamento de objetos informações que não se alteram ou que têm baixa frequência de atualização tais como dados históricos, imagens, vídeos, arquivos de áudio e cópias de segurança (backups). Os serviços de armazenamento de objetos costumam oferecer um alto grau de redundância, o que amplia a durabilidade e disponibilidade das informações salvas nele e o credencia como a solução ideal para dados históricos ou estáticos.

Banda Internet

Normalmente os fornecedores não cobram o tráfego de entrada, isto é, o volume de informações que são enviadas para a nuvem, mas cobram o

tráfego de saída. Determinados tipos de sistemas podem se beneficiar dessa política de cobrança. Por exemplo, sistemas que processam muitas informações mas geram resultados pouco volumosos, como é o caso de sistemas analíticos (Business Intelligence), que produzem relatórios sintéticos ou calculam índices a partir de grandes volumes de dados.

Uma decorrência importante desse modelo de cobrança é que, quando um sistema é hospedado em duas regiões diferentes para redundância, a eventual troca de dados entre eles para sincronização de bases de dados ou tarefas de controle será contabilizada e taxada nos mesmos moldes do tráfego da saída normal, uma vez que a internet será usada como meio para transporte das informações.

Tráfego de E/S

O tráfego de dados entre as máquinas virtuais e os discos persistentes também são cobrados. Isso significa que, mesmo que um site ou sistema envie poucos dados para o usuário (use pouca banda Internet), é importante considerar o volume de operações de entrada e saída dos sistemas a serem implantados na nuvem, isto é, o número de gravações e leituras no disco que ele realiza.

Softwares e imagens binárias

Quando uma instância é criada, é preciso especificar o sistema operacional que será usado. Os fornecedores oferecem instâncias já padronizadas e que vêm com algumas opções de configuração. Por exemplo, só sistema operacional ou sistema operacional mais banco de dados. Entretanto, pode ser que você precise de instâncias customizadas, diferentes daquelas oferecidas pelos fornecedores. Nesse caso, você pode criar suas próprias imagens binárias, instalando todos os softwares de que você precisa. Quando é criada uma nova instância, você pode especificar qual a imagem binária desejada.

Controle de acesso

É preciso considerar também o controle de acesso à sua estrutura de Computação em Nuvem. Avalie o grau de dificuldade de incorporar ou integrar à nuvem o sistema de identificação de usuários que já é usado pela sua empresa. Alguns fornecedores já oferecem integração com produtos de mercado. Considere também os recursos que o fornecedor oferece para o controle de acesso, tais como restrição de endereço IP ou o uso de autenticação de dois fatores.

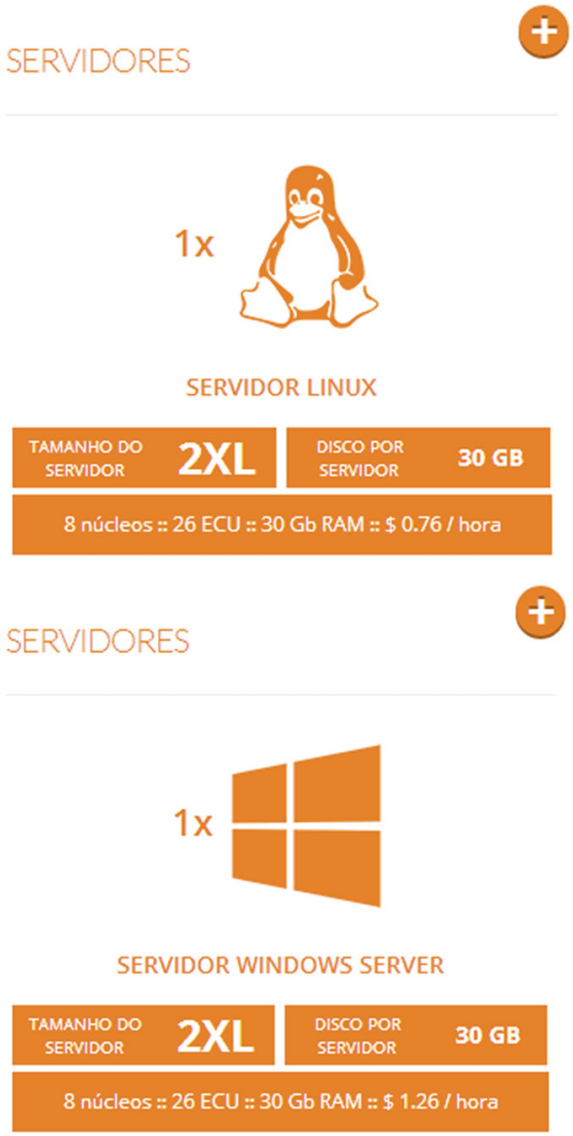
Facilidade de gerenciamento

Ao usar uma infraestrutura como serviço, muitas tarefas de gerenciamento dos recursos ficam sob sua responsabilidade. Avalie a facilidade de uso das ferramentas de administração oferecidas pelo provedor e o quanto essas ferramentas permitem que você acompanhe o desempenho do seu ambiente na nuvem.

Custos

O custo da infraestrutura varia de acordo com a forma planejada de uso, isto é, se as instâncias serão criadas sob demanda, se serão instâncias reservadas ou se terão uso continuado. Na seção "Tipos de Instâncias" apresentamos aquelas que são oferecidas por alguns fornecedores.

Na avaliação dos custos, um aspecto que também deve ser considerado é o licenciamento de software. Ao criar um servidor, o sistema operacional escolhido faz diferença no preço. Para exemplificar, usamos a calculadora NubExpress[30] para estimar o custo de servidores da AWS com a mesma configuração, alterando apenas o sistema operacional, mostrando a diferença no preço final advinda do licenciamento do software[31]:



A cotação exemplificada é de um servidor tamanho 2XL, com disco de 30Gb, 8 núcleos, 26 ECU[32] e 30Gb de RAM:

- Servidor com sistema operacional Linux: US$ 0,76 / hora
- Servidor com sistema operacional Windows Server: US$ 1,26 / hora

Além do sistema operacional, pode ser necessário licenciar um software de gerenciamento de bancos de dados (Microsoft SQL Server, Oracle, etc.). Em alguns casos o cliente pode aproveitar as licenças de software que já possui, em outros não pode. É preciso consultar o provedor de serviços na nuvem e o contrato de licenciamento do software para averiguar essa

possibilidade.

Portanto, é importante que fique claro que, quando falamos em compra de Infraestrutura como Serviço, não estamos tratando apenas de hardware, mas também de software. Além disso, é preciso considerar a incidência de impostos sobre serviços de processamento de dados e licenças de software.

A comparação entre fornecedores de Infraestrutura como Serviço não é simples. Embora os serviços sejam muito parecidos, as configurações oferecidas podem ser muito diferentes. Também é preciso saber qual o nome que cada fornecedor atribui para o mesmo tipo de serviço..

A seguir, apresentamos os principais serviços oferecidos por três dos maiores fornecedores de IaaS: Amazon, Google e Microsoft.

Amazon Web Services (AWS)

Servidores

Com o **Amazon EC2** (Elastic Compute Cloud) é possível criar instâncias de máquinas virtuais com os sistemas operacionais Windows Server, Linux ou FreeBSD, com a possibilidade de aumentar ou diminuir o número de instâncias de acordo com a demanda. A configuração dos servidores varia de acordo com a necessidade: instâncias de uso geral, instâncias com demanda de mais CPU ou instâncias com uso de mais memória.

A AWS oferece ainda o serviço **Amazon CloudWatch**, que monitora os recursos e emite relatórios detalhados sobre o a utilização da nuvem. Através do CloudWatch é possível ativar o recurso de ***Auto Scaling***, que usa regras pré-definidas para criar novas instâncias automaticamente em momentos de pico e remover as instâncias quando a demanda diminui.

Outro serviço bastante utilizado em combinação com o recurso de *Auto Scaling* é o **Elastic Load Balancing**, que permite distribuir a carga de trabalho entre várias instâncias, de forma a proporcionar um desempenho uniforme das aplicações à medida que a demanda aumenta ou diminui ou mesmo que alguma das instâncias falhe.

Claro, para que esse mecanismo de “esticar e encolher” funcione, a aplicação ou site que roda nesse ambiente deve estar preparada para suportar o aumento e diminuição automáticos do número de instâncias e da distribuição da carga de trabalho.

Armazenamento de dados

Estes são alguns dos serviços para armazenamento de dados oferecidos pela AWS:

Amazon EBS (Elastic Block Store)

É o serviço de disco persistente que pode ser usado junto com as instâncias EC2. Ou seja, como ele é persistente, se você desativar a instância EC2, os dados do EBS continuam lá. A unidade de alocação de espaço no EBS é um volume. Os volumes podem ter qualquer tamanho e o usuário pode redimensioná-los a qualquer momento.

Há opções de discos magnéticos (mais lentos e mais baratos) e discos SSD (Solid-State Disk), que são implementados com tecnologia de semicondutores, sem partes móveis, portanto são muito mais velozes. Há dois outros fatores que impactam diretamente no desempenho dos volumes EBS: o limite máximo de operações de entrada e saída por segundo e a quantidade máxima de transferência de dados por segundo. Por exemplo, os discos SSD suportam até 3.000 operações de entrada e saída por segundo, enquanto que os discos magnéticos são limitados a até duzentas operações[33]. Além disso, os discos SSD permitem a transferência de até 160 MBytes por segundo[34], enquanto os discos magnéticos são limitados a até 90.

Para aplicações que necessitam de altas frequências de operações de entrada e saída e/ou maiores quantidades de transferência de dados do que os oferecidos pelos volumes EBS padrão, a AWS oferece ainda uma terceira opção: volumes IOPS, que proporcionam armazenamento com desempenho consistente e baixa latência, tendo sido projetados para aplicações com uso intenso de operações de entrada e saída, como bancos de dados.

- Independentemente do tipo de volume EBS utilizado, todos são replicados automaticamente e de forma transparente para o usuário, para proteção contra falhas de componentes individuais, garantindo uma disponibilidade de 99,999%. Além disso, os volumes podem ser encriptados, garantindo que todos os dados armazenados sejam criptografados, bem como a transferência dos dados entre os discos e as instâncias de máquinas virtuais.

Amazon S3 (Simple Storage Service)

É o serviço de armazenamento de objetos da AWS. É possível salvar nele qualquer tipo de arquivo. Não é preciso pré-alocar espaço de armazenamento nele, e só se paga pelo que é efetivamente utilizado, não exigindo taxa mínima de utilização nem custo de configuração.

Sua principal característica é a altíssima taxa de durabilidade dos objetos ali armazenados. A AWS garante durabilidade de 99,999999999%, além de uma disponibilidade de 99,99% dos objetos ao longo do ano. A garantia de durabilidade e disponibilidade é suportada pela redundância automática dos dados em vários dispositivos e diferentes zonas de disponibilidade e regiões.

O acesso aos dados armazenados no S3 é bem mais lento do que o acesso aos volumes EBS, e deve ser usado como um repositório de arquivos e objetos. Para fins de comparação, enquanto o acesso a volumes EBS equivale ao de discos locais, na casa de milissegundos, o acesso ao S3 a partir de uma máquina virtual leva alguns segundos.

- Os dados armazenados no S3 podem ser automaticamente criptografados, garantindo segurança às informações. Como o serviço S3 é acessível pela internet através dos protocolos HTTP e HTTPS, ele pode ser usado para hospedar de maneira muito barata um site web que seja constituído apenas por páginas estáticas.

Amazon Glacier

- É um serviço que fornece armazenamento seguro e durável para backup e arquivamento de dados, geralmente dados históricos que você é

obrigado a guardar por anos e anos e que raramente precisa recuperar. O armazenamento tem um preço bem baixo, mas a recuperação dos dados pode levar algumas horas. Ou seja, este serviço não deve ser usado para armazenar dados que precisam ser recuperados com rapidez e que são usados frequentemente.

A AWS oferece ainda o **AWS Storage Gateway**. Trata-se de um software que pode ser instalado na infraestrutura local do usuário e que transfere automaticamente para a nuvem todos os dados gravados nos discos dos servidores da rede local. Dessa forma, os dados locais são replicados na nuvem de forma a poderem ser recuperados em caso de desastre.

Bancos de dados

Amazon RDS (Relational Database Service)

Este é um serviço de banco de dados relacionais e através dele é possível utilizar bancos de dados MySQL, Oracle, Microsoft SQL Server ou PostgreSQL.

Naturalmente, você não precisa do RDS para usar um banco de dados relacional na nuvem da AWS. Uma forma de ter este recurso seria instanciar uma máquina virtual e instalar nela por conta própria qualquer um desses softwares gerenciadores de bancos de dados. A diferença é que o serviço RDS garante a alocação de recursos computacionais otimizados para o melhor desempenho de seu banco de dados, além de realizar tarefas fundamentais de gerenciamento como atualização do software, alteração de espaço para o banco de dados e backup automático dos dados. Em outras palavras, o serviço RDS é o que se chama de DBaaS (Database as a Service ou Banco de Dados como Serviço), que libera o usuário das tarefas periódicas de gerenciamento e manutenção do banco de dados.

O RDS oferece ainda as chamadas instâncias Multi-AZ, em que o banco de dados é replicado e sincronizado automaticamente em outra zona de disponibilidade. Em caso de uma parada de manutenção programada, ou no caso de falha não programada, o banco de dados secundário – que está com os dados atualizados em função da replicação e sincronização online –

assume automaticamente, evitando parada no serviço.

Amazon Aurora

Esta é outra opção oferecida pelo RDS: um gerenciador de bancos de dados relacionais compatível com o MySQL, mas que oferece um desempenho muito superior, combinado com um altíssimo grau de disponibilidade. Algumas de suas características são:

- O serviço acrescenta automaticamente mais espaço de armazenamento quando necessário, em blocos de 10 GBytes, até o limite de 64 TBytes, sem necessidade de intervenção do usuário.
- Para garantir menor probabilidade de interrupção dos serviços, o banco de dados é replicado em três zonas de disponibilidade diferentes, com duas cópias dos dados em cada zona.
- A recuperação em caso de falhas de algum dispositivo de armazenamento ou de alguma instância do banco é automática. O processo de backup também se beneficia bastante dessa replicação, eliminando o aumento de carga de trabalho e degradação de desempenho durante sua realização, pois os dados a serem copiados podem ser lidos de qualquer das instâncias.

O Amazon Aurora permite também que existam até 15 réplicas de leitura. Tais réplicas compartilham os dispositivos de armazenamento com a instância principal, permitindo aumentar sobremaneira o volume de operações de leitura dos dados.

A inclusão do Amazon Aurora às opções suportadas pelo serviço RDS proporciona, segundo afirma a Amazon, um novo gerenciador de bancos de dados que alia a velocidade e disponibilidade dos bancos de dados comerciais de alto desempenho à simplicidade e baixo custo dos bancos de dados de código aberto (open source).

NoSQL

Além do serviço RDS, a AWS oferece dois outros mecanismos de banco de dados: o **SimpleDB** e o **DynamoDB**. Ambos são bancos de dados NoSQL

do tipo chave-atributo. A proposta do SimpleDB é oferecer simplicidade e flexibilidade em troca de limitações de escalabilidade e desempenho. Já o DynamoDB oferece desempenho e escalabilidade, acompanhando o crescimento do volume de dados e do número de operações das aplicações que o utilizam.

A AWS oferece também o **Amazon Redshift**, que é um serviço de data warehouse gerenciado e hospedado na nuvem. Utilizando uma arquitetura massivamente paralela, o serviço consegue realizar tarefas analíticas sobre grandes volumes de dados com excelente desempenho. Diversas ferramentas de Business Intelligence de mercado oferecem conectores para o Amazon Redshift, como Tableau, MicroStrategy e Jaspersoft.

Outros Serviços

Além dos recursos apresentados nas seções anteriores, a AWS oferece ainda um grande número de serviços adicionais que permitem a criação de uma infraestrutura completa de processamento na nuvem. Alguns dos serviços oferecidos pela AWS são:

- **Route53**, que é um sistema de nomes de domínios[35] (DNS) que oferece alto desempenho e diferentes mecanismos para balanceamento de carga.
- **CloudFront**, que é o serviço de entrega de conteúdo (CDN)[36] - Content Delivery Network da AWS. Com ele, é possível espalhar o conteúdo de um site pelas várias regiões da AWS pelo mundo, permitindo entregar o conteúdo para o usuário a partir do ponto mais próximo a ele, o que reduz o tempo de transferência das informações. O CloudFront suporta a entrega de conteúdo estático e dinâmico e também de serviços de streaming de vídeo, entre outros recursos.
- **VPC**, Virtual Private Cloud, que é um serviço que permite a criação de uma seção privada e isolada na nuvem da AWS para uso particular. O recurso de VPC permite a configuração de redes com diversas topologias para endereçar necessidades específicas. Por exemplo, com o VPC é possível criar uma sub-rede isolada da internet, onde podem ser colocados servidores com informações sensíveis que não devem ser

expostas na rede.

- **ElastiCache**, que é um serviço para implementar e operar um cachê de memória na nuvem. Usado para aumentar a velocidade de aplicações, criando uma camada intermediária de dados que reduz o acesso a bancos de dados em disco, o ElastiCache oferece dois mecanismos distintos baseados em código aberto (open source): Memcached e Redis. O serviço implementa recursos de tolerância a falhas, garantindo a substituição automática de nós que apresentem problemas de funcionamento.
- **EMR**, Elastic MapReduce, que é um serviço para o processamento de grandes quantidades de dados. O EMR utiliza o Hadoop, que é um software de código aberto que permite distribuir os dados e processá-los de maneira paralela. Para tanto, o EMR cria um grupo de instâncias de máquinas virtuais que pode ser redimensionado dinamicamente em função da necessidade de processamento.
- **SES**, Simple Email Service, que é um serviço de baixo custo de envio de e-mails em grande quantidade, tipicamente para mensagens de marketing ou qualquer outro tipo de conteúdo.
- **Elastic Transcoder**, que é um serviço de transcodificação de mídia na nuvem. Tipicamente usado para, a partir de um vídeo em formato digital, gerar várias alternativas de resolução e formato, normalmente necessárias para atender aos diversos tipos de dispositivos pelos quais os usuários acessam vídeos na internet. Sendo uma tarefa intensiva em termos de processamento, a tarefa de transcodificação oferecida pelo serviço permite o uso de recursos sob demanda e sem a necessidade de provisão de recursos dispendiosos e pouco utilizados.

Vários dos serviços adicionais oferecidos pela AWS estão além do que se convencionou chamar de Infraestrutura como Serviço (IaaS). Alguns desses serviços são, na verdade, grandes facilitadores para o desenvolvimento de novas aplicações na nuvem, reduzindo bastante a complexidade de construção dentro dos conceitos de elasticidade e escalabilidade inerentes ao ambiente.

Além dos serviços apresentados, a AWS oferece dezenas de outros serviços

voltados para o gerenciamento dos recursos na nuvem, além de mecanismos especializados ou voltados para a construção de aplicações, incluindo aplicações para dispositivos móveis.

Também há serviços voltados para gestão da segurança das informações e sistemas na nuvem, tais como o serviço de autenticação e gerenciamento de identidade e o serviço de criação e gerenciamento de chaves de criptografia, que armazena as chaves em módulos de hardware de segurança (HSM – Hardware Security Module). É possível também alocar módulos de hardware de segurança diretamente na nuvem.

Existem ainda diversas opções de softwares desenvolvidos por outras empresas para rodar na nuvem da AWS que estão catalogados no **AWS Marketplace**. Esses softwares estão disponíveis como imagens AMI (Amazon Machine Image), que são imagens binárias pré-configuradas de instâncias de máquinas virtuais com o software já instalado. Ao comprar o software é feita uma cópia da AMI para a sua máquina virtual no EC2 automaticamente, com o software já pronto para ser usado.

A AWS oferece também recursos específicos para aplicações que exigem computação de alta performance (HPC – High Performance Computing), voltados para a resolução de problemas científicos e de engenharia, tais como redes de baixa latência completamente compartimentadas, agrupamentos de unidades de processamento gráfico (GPU – Graphic Processing Unit) para construção e tratamento de imagens, além de instâncias específicas de alta capacidade de processamento ou alta capacidade para realizar tarefas de entrada e saída.

Para uma lista completa dos serviços oferecidos pela AWS, visite o catálogo de produtos da empresa.

Google Compute Engine

Servidores

Google Compute Engine é o componente da plataforma de nuvem do Google que permite criar servidores virtuais rodando os sistemas

operacionais Windows Server, Linux ou FreeBSD. O ambiente é implementado na mesma infraestrutura global utilizada pelo Google para rodar seu mecanismo de busca e outros serviços da empresa, como Gmail e YouTube. Assim como os outros fornecedores, também existem várias configurações de máquinas, que podem ser para uso geral, com mais demanda de memória ou alta utilização de CPU, desde micro máquinas virtuais até grandes instâncias.

O Google Compute Engine oferece um serviço de balanceamento de carga, de forma que sua aplicação possa distribuir o tráfego por múltiplas instâncias de máquinas virtuais. Também está disponível o recurso *Autoscaler*, que permite aumentar ou diminuir automaticamente o poder computacional utilizado em resposta à variação de demanda.

Conforme apresentado anteriormente (na seção *Tipos de Instâncias*), um diferencial interessante do Google Compute Engine em relação ao AWS EC2 é que, enquanto a Amazon cobra o uso do tempo de processamento por hora, as máquinas virtuais do Google são cobradas por minuto.

Outra característica diferencial da plataforma em termos de cobrança, também abordada em mais detalhes anteriormente, é o conceito de desconto para uso continuado. Diferentemente da AWS, que oferece descontos apenas para instâncias pré-alocadas por períodos longos de tempo (de um a três anos) – o que, em certo sentido, contraria o espírito da Computação em Nuvem de pagar apenas pelo uso – o Google oferece descontos progressivos de maneira automática para máquinas virtuais que rodam por mais tempo. Isso garante que a alocação e liberação de recursos possa acompanhar a variação de demanda por poder computacional sem a necessidade de um planejamento prévio ou um compromisso de uso mínimo.

Armazenamento de dados

Discos persistentes

O Cloud Compute Engine oferece discos padrão e discos SSD, que são mais velozes e mais caros. Os discos do Cloud Compute Engine podem ser

conectados e desconectados das máquinas virtuais criadas no ambiente, mantendo os dados de forma persistente. Como no caso da AWS, o desempenho dos discos está associado ao número máximo de operações de gravação e escrita por segundo (IOPS – Input/Output Operations Per Second) e ao volume máximo de transferência de dados por operação.

Todos os dados gravados nos discos do Cloud Compute Engine são criptografados já na máquina virtual, garantindo a inviolabilidade dos dados já no tráfego entre o servidor e os discos.

Google Cloud Storage

Este é o serviço de armazenamento de objetos do Cloud Compute Engine. Além de permitir o armazenamento de qualquer tipo de objeto e de alocar espaço dinamicamente, só cobrando pelo espaço efetivamente utilizado, o serviço oferece ainda opções de controle de versões de objetos e um mecanismo para gerenciar o tempo de vida dos objetos armazenados. Por exemplo, é possível estabelecer um tempo máximo de vida para um objeto, e o serviço automaticamente remove aquele objeto depois de vencido seu tempo de vida. Também é possível controlar automaticamente o número máximo de diferentes versões dos objetos armazenadas.

O Google Cloud Storage criptografa automaticamente os dados a serem armazenados, e essa criptografia é feita já no servidor, de forma que todos os dados que transitam entre as máquinas virtuais e o serviço de armazenamento de objetos são criptografados.

O serviço Google Cloud Storage oferece ainda uma opção mais barata de armazenamento de objetos, com a mesma garantia de segurança e durabilidade dos dados, mas com tempo de recuperação mais lento do que na opção padrão.

Bancos de dados

O Google oferece dois serviços de bancos de dados que podem ser usados tanto por aplicações tradicionais rodando no Google Compute Engine (IaaS) quanto por aplicações desenvolvidas para o Google Application Engine (PaaS do Google):

Cloud SQL

É um gerenciador de bancos de dados relacionais com recursos e funcionalidades do MySQL, com algumas características a mais e algumas restrições. Não requer instalação de software nem manutenção, e o serviço realiza cópias de backup autom aticamente. Oferece ainda mecanismos de redundância no armazenamento dos dados, garantindo um alto grau de tolerância a falhas. Além disso, o serviço realiza a alocação automática de espaço de armazenamento de acordo com a demanda, e apenas o espaço em disco efetivamente utilizado é cobrado. Esta solução é endereçada para aplicações de pequeno a médio porte, uma vez que existem restrições no tamanho máximo para as bases de dados armazenadas (250 GBytes)[37].

Cloud Datastore

Oferece armazenamento de dados semiestruturados ou não estruturados (NoSQL, embora o Google afirme que ele é um gerenciador de bancos de dados "*NoSQL-like*") e gerencia o acesso aos dados, inclusive em casos em que existam diversas instâncias trabalhando em paralelo. A grande vantagem deste serviço é a escalabilidade, podendo suportar grandes volumes de dados e múltiplos acessos simultâneos.

Na verdade, o Cloud DataStore é uma abstração construída sobre o mecanismo de armazenamento **Google BigTable**[38]. O serviço é oferecido de forma gerenciada, não exige dimensionamento prévio e garante redundância, disponibilidade e consistência de forma automática. Também oferece o recurso de gerenciamento de transações para garantia de integridade dos dados em operações de gravação complexas.

BigQuery

A plataforma de nuvem do Google oferece ainda este serviço de banco de dados analítico que permite realizar pesquisas sobre grandes volumes de dados utilizando um dialeto da linguagem SQL. As operações de consulta podem ser síncronas ou assíncronas e os resultados podem ser armazenados em tabelas temporárias ou persistentes. Atualmente, diversas ferramentas analíticas de Business Intelligence do mercado, como Tableau e QlikView, possuem conectores para o BigQuery, permitindo a manipulação de grandes

volumes de dados armazenados na nuvem.

Vale ainda lembrar que é possível rodar qualquer servidor de bancos de dados, relacional ou não, na infraestrutura do Google Compute Engine. Basta instanciar uma máquina virtual e carregar nela o software desejado. Porém, nesse caso, o gerenciamento do banco de dados fica todo por conta do usuário.

Outros Serviços

O Google Compute Engine oferece uma série de recursos e serviços adicionais para apoiar a construção de uma infraestrutura computacional abrangente, além de serviços específicos para a construção de aplicações. Alguns desses serviços adicionais são:

- **VPN**, que permite a construção de redes privadas virtuais entre a infraestrutura local do cliente e os recursos alocados na nuvem do Google através de um canal seguro.
- **Prediction API**, que é um serviço que permite integrar algoritmos de aprendizagem de máquina rodando na nuvem do Google a aplicações desenvolvidas pelos usuários. Algoritmos desse tipo permitem, por exemplo, a criação de mecanismos de recomendação de produtos para clientes baseados em seus históricos de preferências ou a previsão de tendências futuras baseadas em dados históricos.
- **Translate API**, que é um serviço que permite integrar aplicações desenvolvidas pelo usuário ao Google Tradutor. Dessa forma, é possível adicionar às funcionalidades de uma aplicação os recursos de tradução de documentos entre os vários idiomas suportados pelo serviço.

O Google Compute Engine oferece ainda diversos softwares de terceiros prontos para serem executados em sua plataforma, incluindo gerenciadores de bancos de dados e ferramentas de automação e gerenciamento de configuração, além de mecanismos de troca de mensagens entre aplicações.

Além das opções de infraestrutura como serviço (Google Compute Engine) e de plataforma como serviço (Google App Engine), o Google oferece um

grande número de ferramentas na modalidade SaaS (Software como Serviço): seu próprio mecanismo de busca que dá nome à empresa, Gmail, YouTube, Google Maps, Google Translate, Google Apps, Google Hangout e vários outros.

É bom lembrar que o Google é uma empresa 100% internet, e que a Computação em Nuvem sempre fez parte de sua estratégia, o que naturalmente o qualifica como um competidor forte nesse mercado agora e para o futuro.

Microsoft Azure

Servidores

O **Microsoft Azure** permite a criação de máquinas virtuais rodando sistemas operacionais Windows Server ou Linux. Da mesma forma que outros fornecedores, o ambiente oferece várias configurações possíveis de servidores para cenários de uso geral, maior demanda por capacidade de processamento ou com maior demanda por memória.

As diferentes configurações de instâncias são oferecidas em duas categorias, “Basic” e “Standard”. A diferença entre as duas categorias é que as instâncias “Basic”, mais baratas, não oferecem serviços de balanceamento de carga nem de redimensionamento dinâmico dos recursos computacionais (*Autoscaling*). Tipicamente, as instâncias do tipo “Basic” oferecem uma alternativa mais econômica para servidores de teste, aplicações não escaláveis – que rodam em apenas um servidor – e aplicações de processamento em lote (batch).

O serviço de balanceamento de carga do Microsoft Azure, o **Azure Traffic Manager**, permite a distribuição do tráfego entre múltiplas instâncias de uma aplicação baseada em diferentes políticas. É possível, por exemplo, rotear as requisições vindas da internet para as instâncias geograficamente mais próximas do usuário. Também é possível distribuir o tráfego igualmente entre os servidores, ou direcionar as requisições para um único conjunto de servidores, mantendo uma instalação de contingência para a

qual o tráfego só é direcionado em caso de falha da primeira. O mecanismo de *Autoscale* permite o redimensionamento dinâmico e automático dos recursos computacionais baseado em diversos contadores de desempenho.

Conforme citado anteriormente, embora a tabela de preços do Microsoft Azure indique os valores por hora de uso das instâncias, a cobrança é realizada por minutos de uso, como no Google Compute Engine. A principal forma de cobrança é o pagamento pelo uso, sem o conceito de instâncias reservadas da AWS. Entretanto, a Microsoft oferece um modelo diferenciado de cobrança para empresas: o *Enterprise Agreement*. Na verdade, trata-se do contrato de licenciamento de software que a empresa já mantém com seus clientes corporativos. É possível acrescentar a esse contrato um compromisso mínimo de consumo de serviços do Microsoft Azure: a empresa paga um valor antecipado para consumo ao longo do ano, e o que for excedido, será cobrado trimestralmente ou anualmente. Essa compra antecipada gera um desconto progressivo baseado no volume financeiro envolvido.

Armazenamento de dados

No Microsoft Azure, sempre que uma máquina virtual é instanciada ela já tem dois discos associados a ela: o de sistema operacional e o de dados temporários. O disco do sistema operacional é criado a partir de imagens binárias dos sistemas operacionais suportados. O custo do disco temporário já está incluído no custo da máquina virtual, e o disco de sistema operacional é taxado em função do espaço utilizado.

Os discos persistentes, chamados de *data disks*, podem ser discos padrão ou discos SSD, e também no caso da nuvem da Microsoft, o desempenho dos discos estará associado à tecnologia do disco e ao número máximo de operações de gravação e escrita por segundo (IOPS – Input/Output Operations Per Second), além do volume máximo de dados que podem ser transferidos em cada operação. É possível conectar vários discos persistentes a cada máquina virtual.

Azure Storage

No Microsoft Azure, os discos persistentes são apenas um caso especial do serviço de armazenamento de objetos, chamado de **Azure Storage**. Essa arquitetura proporciona uma vantagem interessante em termos de custo para o usuário: quando um disco persistente é criado, é necessário definir o seu tamanho lógico, mas apenas o espaço efetivamente utilizado será taxado, isto é, apenas o volume de dados realmente gravado no disco é que será cobrado. Portanto, se um volume de disco persistente tem tamanho lógico de 50 Gbytes, mas apenas 10 Gbytes são usados, o valor a ser pago mensalmente será o equivalente a 10 Gbytes.

O **Azure Storage** oferece alto grau de redundância das informações armazenadas nele, e a cobrança é feita apenas pelo volume de dados salvos, com descontos progressivos à medida que mais espaço é utilizado. Além disso, proporciona diferentes mecanismos de armazenamento para diferentes tipos de objetos:

- **Blob**: pode armazenar qualquer tipo de texto ou dados binários, como documentos, imagens ou arquivos de mídia.
- **Table**: para armazenar conjuntos de dados estruturados não relacionais. Na prática, é um banco de dados NoSQL do tipo chave-valor que permite armazenar grandes volumes de informações estruturadas, mas sem formato pré-definido, conforme citado anteriormente.
- **Queue**: implementa um mecanismo simplificado de filas persistentes, permitindo a comunicação entre componentes de aplicações.
- **Arquivos**: permite o compartilhamento de arquivos entre diferentes máquinas virtuais, como se estivessem acessando o mesmo servidor de arquivos em uma rede local. Oferece compatibilidade com o protocolo SMB, permitindo que mesmo aplicações legadas possam compartilhar arquivos na nuvem.

Azure Backup

Este é outro serviço importante oferecido pela Microsoft, um software que é instalado na infraestrutura local do usuário e que copia automaticamente os dados dos discos dos servidores locais para a nuvem. Os dados são encriptados antes de serem enviados, garantindo sua inviolabilidade tanto

no trânsito quanto no armazenamento dentro da infraestrutura do Azure. Há recursos integrados para compactação dos dados e limitação da largura de banda internet utilizada para sua transmissão. Os dados são armazenados de forma redundante, garantindo durabilidade e acessibilidade. Depois da cópia inicial, apenas os dados alterados são transferidos, garantindo um tempo mais rápido para a realização dos backups e minimizando o espaço utilizado na nuvem.

Bancos de dados

As opções de bancos de dados gerenciados do Microsoft Azure já foram apresentados na seção que fala sobre a solução PaaS da Microsoft, Azure Cloud Services:

- **SQL Database**: oferece a grande maioria dos recursos do gerenciador de bancos de dados relacionais Microsoft SQL Server, suportando também grande parte das construções da linguagem Transact-SQL.
- **Azure Storage Table**: banco de dados NoSQL do tipo chave-valor flexível que oferece grande escalabilidade e alto desempenho no acesso ou gravação de dados.
- **DocumentDB**: banco de dados NoSQL orientado a documentos que oferece escalabilidade, alta disponibilidade e excelente desempenho para tratamento de grandes volumes de dados. Oferece vários recursos não disponíveis em outros gerenciadores de bancos de dados desse tipo, como suporte a transações, gatilhos e realização de consultas complexas.

Além das opções de bancos de dados gerenciados, também é possível rodar na infraestrutura do Microsoft Azure qualquer sistema gerenciador de bancos de dados comercial que rode sobre as plataformas Windows Server ou Linux. Em particular, alguns produtos de mercado já possuem imagens binárias de máquinas virtuais já configuradas para rodá-los.

Um exemplo de gerenciador de bancos de dados comercial suportado pelo Microsoft Azure é o Oracle, sendo possível tanto instalar sua própria licença ou alugar uma a partir de uma imagem de máquina virtual pré-

configurada.

Outro exemplo é o **ClearDB**, que oferece uma versão do MySQL com recursos de redundância para suportar tolerância a falhas.

O Microsoft Azure oferece ainda um serviço para análise de dados baseada no Apache Hadoop: o **HDInsight**, que permite o processamento distribuído de grandes volumes de dados com escalabilidade e excelente desempenho. A partir da infraestrutura elástica do Microsoft Azure é possível aumentar ou diminuir o número de nós de processamento em função do volume de dados a serem analisados. O Hadoop é capaz de processar dados semiestruturados ou desestruturados, permitindo capturar e manipular informações advindas das mais diferentes fontes. Para permitir uma visualização amigável dos dados armazenados e processados pelo HDInsight, a Microsoft integrou ao Excel recursos para utilizar esse serviço como fonte de dados.

Outros Serviços

Dentro da estratégia de Computação em Nuvem da Microsoft, vários novos serviços vêm sendo integrados ao Azure. Alguns dos principais são:

- **Rede Virtual**, que é o serviço para criação de uma rede privada virtual (VPN) entre a infraestrutura local do usuário e a infraestrutura do Azure.
- **CDN**, que é o serviço de entrega de conteúdo do Azure, que espalha o conteúdo de um site pelos vários pontos de presença da nuvem da Microsoft, permitindo entregá-lo a partir do ponto geográfico mais próximo. O Azure CDN suporta a entrega de conteúdo estático, streaming de vídeo e também permite a interação dos usuários com as aplicações rodando na nuvem.
- **Serviços de Mídia**, que é um serviço para codificação, empacotamento e distribuição de vídeos e áudio em larga escala. É possível realizar o streaming de vídeos (inclusive ao vivo) e a criação de bibliotecas para entrega de vídeos sob demanda. Também possui recursos para cobrança e proteção de direitos autorais.

- **Batch**, que é um serviço para o agendamento de tarefas em lote em larga escala, podendo utilizar processamento paralelo e a realização de cargas de trabalho que exigem grande poder computacional.
- **Websites**, que é um serviço que facilita a disponibilização de sites Web, incluindo páginas estáticas e conteúdo dinâmico através de aplicações que podem ser desenvolvidas em diferentes linguagens de programação. Também permite criar sites baseados em ferramentas populares para construção de portais e gerenciamento de conteúdo como WordPress, Joomla e Drupal.
- **Cache**, que é um serviço para implementar e operar um mecanismo de cachê distribuído baseado no software de código aberto (open source) Redis. É usado para aumentar a velocidade de execução de aplicações, reduzindo o acesso a bancos de dados em disco.
- **Recuperação de Site**, que é um serviço que se propõe a replicar os dados da infraestrutura local do usuário e monitorá-la constantemente, sinalizando em caso de falhas. Além disso, permite uma rápida retomada em caso de desastres, orquestrando a recuperação dos serviços de forma ordenada.

Nuvens Híbridas

Um ponto importante da estratégia da Microsoft para Computação em Nuvem é a criação de nuvens híbridas baseadas em sua tecnologia. Considerando que o sistema operacional Windows Server está presente em uma imensa base de clientes corporativos, a empresa tem disponibilizado ferramentas para facilitar a construção de nuvens privadas, mantendo seu tradicional modelo de negócios baseado em venda de licenças de software, enquanto posiciona o Microsoft Azure como uma extensão natural do poder computacional das infraestruturas locais. Dessa forma, procura proporcionar uma transição suave para a tecnologia de nuvem através de uma adoção gradual pela sua base de clientes, além de garantir completa compatibilidade entre a infraestrutura privada e sua oferta de nuvem pública. É uma estratégia consistente que vem sendo executada de maneira muito competente.

Baseando-se nessa estratégia, o Microsoft Azure oferece alguns serviços que proporcionam integração entre aplicações que rodam na nuvem e as aplicações que rodam na infraestrutura privada de seus clientes, como é o caso do **Azure BizTalk** e o **Azure Active Directory**:

- **Azure Biztalk** – é a versão do Microsoft BizTalk rodando na nuvem, e oferece recursos de integração entre aplicações tais como troca de mensagens, conversão de diferentes formatos de dados e compatibilização entre diferentes protocolos de comunicação.
- **Azure Active Directory** – permite estabelecer as mesmas identidades de usuários e senhas para todas as aplicações na nuvem. Além disso, ele se integra ao Active Directory que roda nos servidores Windows Server da infraestrutura privada do cliente, permitindo o gerenciamento unificado e centralizado das identidades dos usuários locais e na nuvem, o que reduz custos e facilita o acesso aos recursos computacionais.

Outro produto da Microsoft para a integração entre infraestruturas locais e a nuvem pública é o **Azure StorSimple**. O produto é uma solução de armazenamento híbrida, que consiste em um hardware de SAN[39] instalado na infraestrutura local e integrado à nuvem. Nesse caso, os dados são armazenados localmente e, seguindo alguma política configurável, versões mais antigas das informações ou porções específicas dos dados são enviadas para nuvem, incluindo cópias de segurança (backup). A integração, nesse caso, fica transparente para os usuários locais, e a distribuição das informações entre a infraestrutura local e a nuvem é feita de maneira automática pelo equipamento.

Seguindo sua tradição de apoiar o desenvolvimento de aplicações para seus ambientes operacionais, a empresa oferece também serviços exclusivos na nuvem para facilitar a construção de aplicações e incentivar a adoção da Computação em Nuvem pelos desenvolvedores de software. Um dos serviços desse segmento oferece um engenho de análise preditiva e aprendizado de máquina, o **Azure Machine Learning**. Outro serviço oferece ferramentas para desenvolvimento de aplicativos móveis, o **Azure Mobile Services**.

O Microsoft Azure também oferece recursos para computação de alto

desempenho (HPC – High Performance Computing), tais como instâncias de alto poder de processamento, o serviço **Azure Batch** para processamento em lote já apresentado anteriormente, e uma opção de rede de baixa latência e alta taxa de transferência de dados. A empresa oferece ainda um pacote para a construção de uma infraestrutura de computação híbrida de alto desempenho baseada no Windows Server, de maneira a rodar localmente as tarefas que exigem alto desempenho até o limite de sua capacidade, e complementar o poder computacional de forma transparente utilizando recursos da nuvem pública.

Há ainda uma infinidade de soluções específicas desenvolvidas por terceiros, e que estão disponíveis como imagens binárias de máquinas virtuais para a rápida criação de instâncias. Essas soluções estão catalogadas no **Azure Marketplace**.

A Microsoft oferece ainda diversas opções de software como serviço (SaaS), especialmente versões de alguns de seus produtos mais consagrados. O carro-chefe é o Office 365, que é a versão na nuvem do tradicional conjunto de ferramentas de produtividade pessoal Microsoft Office. Outros produtos incluem o Microsoft Exchange, que permite a criação de um servidor de correio eletrônico corporativo diretamente na nuvem, e o Microsoft SharePoint, que é uma ferramenta para a construção de portais colaborativos.

A oferta de Computação em Nuvem da Microsoft é bastante abrangente e, considerando sua estratégia e a força de sua base instalada de produtos de software, a empresa está posicionada para manter sua relevância no mercado de tecnologia, mesmo com o avanço do uso da nuvem representando uma mudança importante em seu modelo de negócios tradicional, baseado na venda de licenças de software.

---

30 A calculadora NubeXpress oferece uma forma simplificada de estimar o custo de servidores rodando na AWS, sendo uma alternativa à complexa calculadora da própria AWS.

31 Preços de março de 2014.

32 ECU é o acrônimo de "EC2 Compute Unit", ou unidade computacional EC2 em tradução livre. Esta unidade foi criada para permitir a comparação do poder de processamento dos diversos tipos de servidores na AWS. Cada ECU é equivalente à capacidade de processamento de 1.0-1.2 GHz do processador Opteron 2007 ou Xeon 2007.

33 Dados de novembro de 2014.

34 idem.

35 Um sistema de nomes de domínios ou DNS (Domain Name System) é um serviço responsável pela tradução de nomes de domínios internet (como "exemplo.com.br") para endereços de rede IP (como 200.138.2.1) . Esse serviço facilita o uso da internet, pois sem ele, em vez de digitar, por exemplo, "google.com.br" no seu navegador, você precisaria conhecer o endereço IP (numérico) do servidor que desejasse acessar.

36 Content Delivery Network.

37 Informação de novembro de 2014.

38 BigTable é um mecanismo distribuído de armazenamento de dados de alto desempenho desenvolvido pelo Google para manipular grandes volumes de dados. Embora nunca tenha sido distribuído para uso externo pela empresa, sua arquitetura e funcionamento foram detalhados em um artigo publicado em 2006, e acredita-se que inspirou o projeto de diversos bancos de dados NoSQL, como Cassandra e Apache HBase (que é o banco de dados do Hadoop).

39 SAN - Storage Area Network, é um equipamento que implementa uma rede dedicada que funciona como um disco de rede de alto desempenho e com recursos de tolerância a falhas, centralizando o armazenamento de dados dos servidores em uma rede local. Normalmente, o SAN concentra os dispositivos de armazenamento da rede, gerando a ilusão de um dispositivo único e uniforme para os equipamentos cliente.

Disaster Recovery – recuperação de dados

Uma política de *Disaster Recovery* visa manter o negócio operando mesmo em caso de falhas de hardware, software, problemas na rede, falha humana, falta de energia elétrica, roubo de equipamentos, incêndio, catástrofes naturais, etc. Nos dias de hoje, em que as empresas dependem tanto das informações armazenadas em seus servidores de dados e em que os processos de negócio estão baseados em tarefas automatizadas pelos sistemas de software, é necessário garantir que as informações e os ambientes operacionais possam ser rapidamente restaurados em ocasiões de contingência. Normalmente, a política de *Disaster Recovery* se baseia na replicação de dados e ambientes de execução a fim de manter o negócio operando ou restaurá-lo rapidamente nessas situações de emergência.

Apesar de ser uma política comum em grandes empresas, a manutenção de uma estrutura de *Disaster Recovery* nunca esteve no radar das pequenas e médias, devido ao alto custo de replicar todos os ambiente operacionais envolvidos, incluindo hardware e software. Em um ambiente de computação tradicional, é preciso duplicar completamente a infraestrutura, além de manter tudo atualizado e pronto para entrar em operação. Isso significa duplicar os custos de implantação e de manutenção.

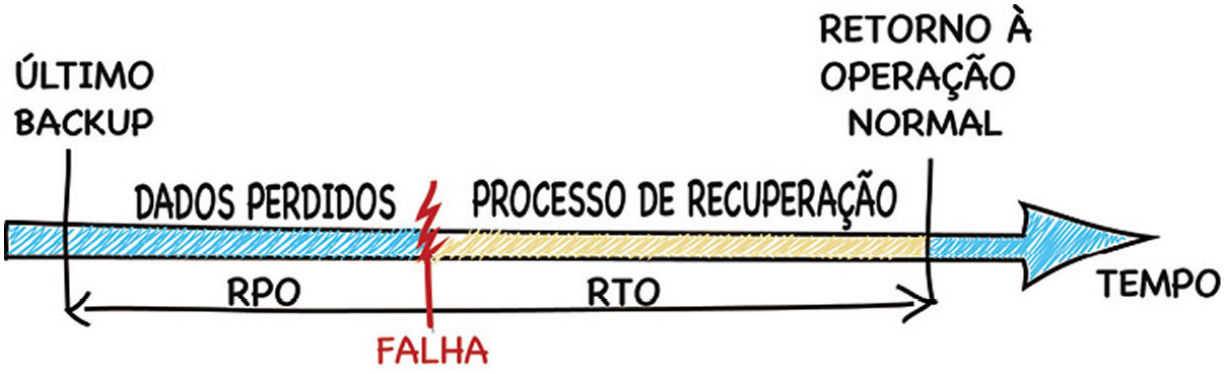
Com a popularização dos serviços de Computação em Nuvem, o cenário mudou. Manter uma estrutura paralela à principal passou a ser uma opção viável também para pequenas e médias empresas.

Na nuvem, o custo dos recursos é proporcional ao uso. Então, por exemplo, você pode criar na nuvem uma infraestrutura secundária que seja um espelho da infraestrutura primária, mas em menor escala, mantida com baixo custo. Em caso de necessidade, isto é, se a infraestrutura primária falhar, você pode ativar a infraestrutura secundária e ampliar em poucos

minutos os recursos utilizados por ela, sob demanda, para assumir o processamento.

Naturalmente, para que o processamento na estrutura secundária possa ser ativado quase imediatamente é necessário que seus dados estejam sincronizados com os da estrutura primária. Manter os dois ambientes completamente sincronizados pode ser custoso. Uma estratégia possível é a de realizar backups periódicos dos dados da infraestrutura primária para a secundária.

Nesse cenário, torna-se imperativo o planejamento e implantação de uma política de backup adequada para as necessidades de continuidade do negócio. Essa política deve estabelecer que as cópias de segurança sejam testadas periodicamente, para garantir que os mecanismos de backup estejam funcionando corretamente. Além disso, a eficiência de uma estrutura de recuperação de dados está ligada a dois fatores: **Recovery Time Objective (RTO)** e **Recovery Point Objective (RPO)**.



RPO E RTO: SEGUNDOS, MINUTOS, HORAS OU DIAS? DEPENDE DO GRAU DE CRITICIDADE DO SISTEMA.

Recovery Time Objective (RTO) é o período de tempo máximo em que cada aplicação deverá voltar a funcionar em caso de interrupção, ou seja, é o período máximo aceitável de interrupção do negócio.

Recovery Point Objective (RPO) define o tempo retroativo dos dados que serão recuperados em caso de falha. Deve-se considerar o que é uma perda aceitável que não cause grande prejuízo operacional. Por exemplo, considere um sistema com RPO de 4 horas, cujo último backup foi feito às 10 horas. Se o incidente ocorrer às 13 horas, tudo o que foi feito entre 10 e

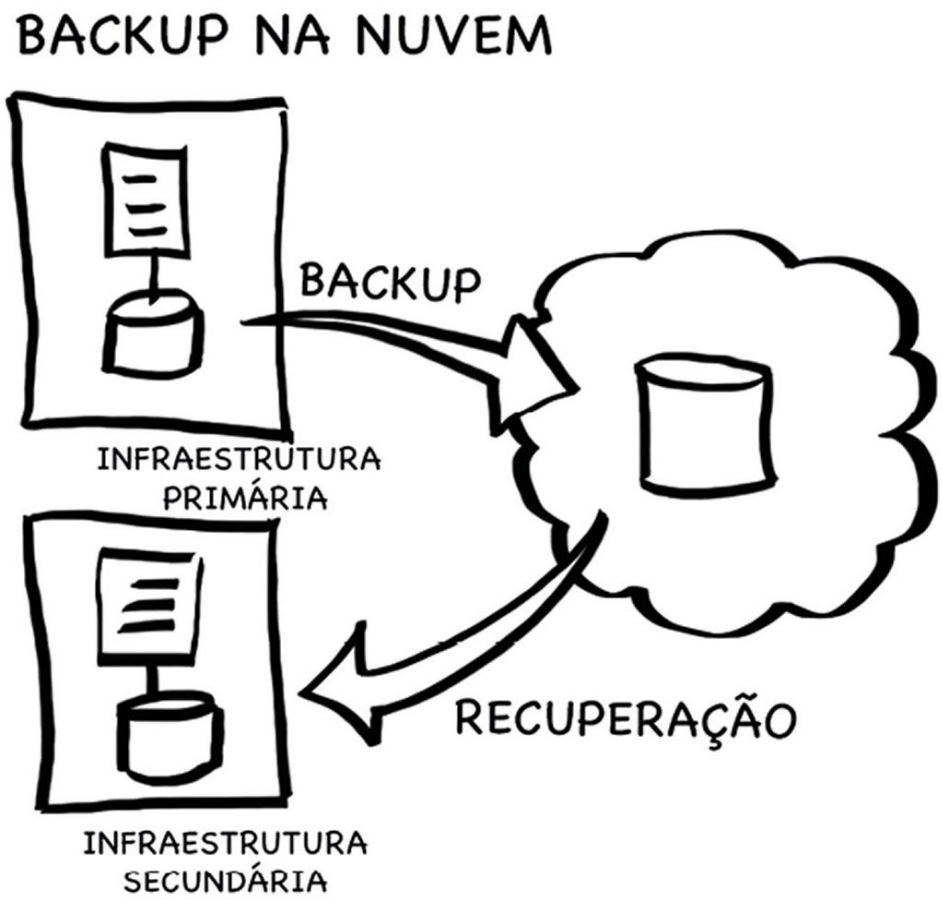
13 horas será perdido.

Ao analisar suas aplicações e escolher quais serão migradas para a nuvem, é preciso definir o RPO para cada uma delas de forma a estabelecer uma política de backup que seja adequada a essa necessidade.

O passo seguinte é definir o modelo de *Disaster Recovery* a ser implantado, considerando o RTO das aplicações. O modelo de *Disaster Recovery* indica quais os recursos devem ser replicados, e define o mecanismo a ser utilizado para garantir uma rápida recuperação em caso de necessidade. Apresentamos a seguir alguns modelos possíveis[40].

Modelo 1: Backup na nuvem

Também conhecido como "*cold start*", este é o modelo mais simples de ser implantado. O backup da estrutura primária é feito na nuvem, ou seja, a nuvem é usada somente como local externo de armazenamento de dados. Em caso de interrupção das atividades na estrutura primária, os dados são transferidos da nuvem para a infraestrutura secundária.



Modelo 2: Backup e infraestrutura secundária na nuvem

Este modelo também é conhecido como "*warm start*" ou "chama-piloto". A ideia de chama-piloto é análoga a de um aquecedor a gás. Nesse equipamento, uma pequena chama fica sempre acesa para que o acendimento total do aquecedor seja feito rapidamente, quando necessário. Neste cenário, a infraestrutura primária continua sob a tutela da empresa, porém tanto o backup quanto a infraestrutura secundária ficam na nuvem. Em caso de interrupção das atividades, a infraestrutura secundária entra em operação, utilizando os dados armazenados na nuvem.

Na infraestrutura secundária na nuvem devem ser instalados os sistemas mais críticos da empresa – a chama-piloto –, que devem ter seus dados atualizados em uma periodicidade compatível com seu RPO. Em caso de necessidade, esse núcleo crítico entra rapidamente em operação, eventualmente alocando mais recursos para que rode com um desempenho adequado. Em seguida, o ambiente de produção é ampliado para que os sistemas menos críticos também comecem a ser recuperados. A alocação dos recursos, em caso de falha, pode ser feita de forma automatizada, o que representa economia de tempo e maior segurança no processo de recuperação do ambiente de produção.

Modelo 3: Alta disponibilidade

Para aplicações que exigem um tempo de recuperação (RTO) muito curto, o modelo de alta disponibilidade, também conhecida como "*hot start*", é o mais indicado. A infraestrutura primária pode estar num *data center* controlado pela empresa ou na nuvem, e a secundária é criada na nuvem. Ambas ficam em operação, e o processamento é distribuído entre as duas. Se uma delas tiver problemas, a outra assume integralmente o processamento. A vantagem de essas infraestruturas estarem hospedadas em um ambiente de Computação em Nuvem é que, em caso de falha de uma delas, a outra pode ter sua capacidade de produção aumentada rapidamente para atender ao incremento de tráfego.

---

40 Para uma discussão detalhada do tema, veja "Disaster Recovery as a Cloud Service: Economic Benefits & Deployment Challenges", de Wood et al.

É hora de colher os benefícios da nuvem

A nuvem amplia consideravelmente as possibilidades para criação de uma infraestrutura de backup e recuperação de dados. Também diminui consideravelmente os custos de implantação de soluções modernas de processamento de dados e mecanismos de interação com seus clientes através de canais digitais, tornando acessível às pequenas e médias empresas recursos antes disponíveis somente para grandes organizações.

Entretanto, é importante deixar claro que os conceitos básicos para a implantação de uma infraestrutura sólida de TI não mudam com a adoção da nuvem: planejamento consistente, execução competente, testes frequentes e treinamento dos usuários continuam sendo fundamentais.

O baixo entendimento sobre o funcionamento de Computação em Nuvem muitas vezes leva algumas organizações a acreditarem em poderes mágicos associados a essa tecnologia. O simples fato de uma aplicação rodar na nuvem não garante a ela atributos como tolerância a falhas, replicação de dados, elasticidade e escalabilidade. Especialmente no caso de aplicações que foram originalmente desenvolvidas para rodar em infraestruturas locais estáticas, é necessário avaliar se suas arquiteturas são adequadas para a migração para a nuvem. Em muitos casos, pode ser necessária uma reengenharia em sua estrutura. Outro ponto importante diz respeito aos cuidados com políticas de backup e aspectos de segurança, da mesma forma que se faz com aplicações locais. Na Opus Software, em nossas consultorias e serviços, frequentemente temos nos deparado com situações em que a migração para a nuvem foi feita de maneira pouco planejada, o que acarreta em grandes riscos para o ambiente computacional das organizações e, consequentemente, para seus negócios.

Portanto, a Computação em Nuvem só vai gerar os benefícios que pode proporcionar se sua adoção for realizada de maneira ordenada e se seus

recursos forem utilizados da maneira adequada. Caso contrário, essa adoção poderá gerar uma falsa impressão de segurança, por puro desconhecimento. E, quando os riscos se materializarem, os prejuízos podem ser imensos.

O fato é que a tecnologia de Computação em Nuvem veio para ficar e já está transformando a forma como as empresas e organizações de todos os portes processam suas informações e oferecem seus produtos e serviços ao mercado. A adoção dessa tecnologia proporciona inúmeros benefícios, reduzindo custos e viabilizando soluções que eram inimagináveis antes. A questão não é mais perguntar **SE** uma organização vai adotar a nuvem, mas sim **QUANDO** ela irá dar esse passo inevitável. Como toda nova tecnologia transformadora que atinge o patamar da produtividade e viabilidade econômica, a Computação em Nuvem abre um mundo de novas possibilidades que permite ampliar o alcance e melhorar o desempenho das mais diversas atividades empresariais. É hora de usar esse imenso potencial a favor de seu negócio!

Referências

- Armbrust, Michael et al. “Above the Clouds: A Berkeley View of Cloud Computing” (2009).
- Chang, Fay, et al. “Bigtable: A distributed storage system for structured data.” ACM Transactions on Computer Systems (TOCS) 26.2 (2008): 4.
- Daniels, Jeff. “Server Virtualization Architecture and Implementation” Crossroads 16.1 (2009): 8-12.
- Duboc, Leticia, David S. Rosenblum, and Tony Wicks. “A framework for modelling and analysis of software systems scalability.” Proceedings of the 28th international conference on Software engineering. ACM (2006).
- Frey, Sören, and Wilhelm Hasselbring. “The cloudmig approach: Model-based migration of software systems to cloud-optimized applications.” International Journal on Advances in Software 4.3 and 4 (2011): 342-353.
- Harms, Rolf, and Michael Yamartino. “The economics of the cloud.”, Microsoft whitepaper, Microsoft Corporation (2010).
- Herbst, Nikolas Roman, Samuel Kounev and Ralf Reussner. “Elasticity in Cloud Computing: What It Is, and What It Is Not.” (2013).
- Lawton, George. “Developing software online with platform-as-a-service technology.” Computer 41.6 (2008): 13-15.
- Mell, Peter, and Timothy Grance, “The NIST definition of cloud computing.” (2011).
- Reese, George. Cloud application architectures: building applications and infrastructure in the cloud. “O’Reilly Media, Inc.”, 2009.
- Wood, Timothy, et al. “Disaster recovery as a cloud service: Economic

benefits & deployment challenges.” Proceedings of the 2nd USENIX conference on Hot topics in cloud computing. USENIX Association (2010).

- “Amazon Web Services (AWS) - Cloud Computing Services.” Amazon Web Services, Inc. Web. 23 Dec. 2014. <http://aws.amazon.com/>.
- “App Engine - Run Your Applications on a Fully-managed Platform-as-a-Service (PaaS) Using Built-in Services.” Google Developers. Web. 23 Dec. 2014. <https://cloud.google. /appengine/>.
- “Blog Opus Software – Dedicado ao Alinhamento entre TI e os Negócios.” Opus Software. Web. 23 Dec. 2014. <http://blog.opus-software.com.br/>.
- “Google Compute Engine - Cloud Computing & IaaS.” Google Developers. Web. 23 Dec. 2014. <https://cloud.google.com/compute/>.
- “Microsoft Azure: Plataforma em Nuvem da Microsoft” Microsoft Azure. Web. 23 Dec. 2014. <http://azure.microsoft.com/pt-br/>.

www.opus-software.com.br

OPUS
SOFTWARE

O Que Você Realmente Precisa Saber Sobre