O Big Data e os desafios de backup e armazenamento
Florence Perrin

Uma boa abordagem de proteção de dados para big data deve considerar o desenvolvimento de uma estratégia que contemple sistemas de backups e recuperação de dados

"Big Data" é um conceito que está conosco desde há muito tempo – nomeadamente no contexto de projetos científicos e de pesquisa nos quais eram criados grandes volumes de dados em um curto espaço de tempo.

O termo é usado para descrever dados – sejam estruturados ou não – tão maciços que a sua recolha, armazenamento, análise, partilha e replicação (esta última, necessária para requisitos de redundância) se torna um desafio. Os sistemas que manipulam Big Data incorporam centenas ou até milhares de processadores, infraestruturas de rede de alta velocidade e grandes sistemas de armazenamento com discos rígidos de classe empresarial de elevada capacidade, projetados para empresas convencionais mas também para ambientes de cloud e computação de alta escalabilidade.

No mundo de hoje, Big Data é gerado globalmente a partir de múltiplas fontes: grandes projetos científicos como o LHC criam e gerem cerca de 15 petabytes e redes sociais como o Facebook trabalham com bases de dados que têm mais de 50 bilhões de fotografias…

Apesar destes desafios serem significativos, as oportunidades que a Big Data apresenta também entusiasmam. Dependendo da ênfase de um projeto, os dados podem ser analisados de forma a deles extrairmos resultados conclusivos. Imagine-se um varejista online que tem acesso ao seu próprio Big Data – esta organização pode analisar os dados para encontrar tendências, identificar comportamentos dos clientes, analisar preços e até potencialmente criar publicidade que seja mais relevante.

A título de exemplo a gigante norte-americana Walmart tem trabalhado intensivamente com Big Data nos últimos tempos de forma a melhor compreender os seus clientes e oferecer- a eles produtos mais relevantes através de apps para smartphones.

Comparativamente com os dados tradicionais e estruturados que podem ser encontrados em uma base de dados relacional, a Big Data encontra-se muitas vezes armazenado de forma não-estruturada. A razão é que enquanto uma base de dados tradicional poderá apenas ser capaz de armazenar e analisar uma gama limitada de dados, tais como números ou datas, por exemplo, o Big Data compreende vários conjuntos de dados tais como texto, vídeo, áudio, dados provenientes de sensores, ficheiros de registo ("logs"), etc. Quando todos estes conjuntos de dados são analisados, oferecem o tipo de conhecimento pelo qual estas instituições anseiam

Big Data e a cloud privada

Existem diferentes tipos de modelos de implementação de clouds: pública, comunitária, híbrida e privada. Uma cloud privada é tipicamente criada para uma única entidade empresarial e pode ser gerida internamente ou por terceiros; o hardware físico e o software que constituem esta cloud podem estar nas instalações da empresa ou em uma localização externa.

Este modelo requer uma significativa alocação de recursos financeiros da empresa, tempo e gestão quando comparada com outros modelos de cloud. A sua maior vantagem é que a entidade pode assegurar a privacidade dos seus dados (enquanto os dados armazenados em clouds públicas estão acessíveis ao fornecedor do serviço de cloud).

Face aos desafios crescentes associados à gestão de dados provenientes de múltiplas fontes, bem como o enfoque em simulações complexas para obtenção dos resultados mais confiáveis, os problemas associados com Big Data podem ser resolvidos com uma nuvem privada. Contudo, para que este sistema seja eficaz, a nuvem privada tem de ser capaz também de ser configurada para lidar com agregação de dados, ser flexível em termos de capacidade,

oferecer instalações para alocação de recursos e ser capaz de enfrentar desafios de responsabilização.

Instituições interessadas em uma nuvem privada têm de fazer investimentos iniciais significativos em termos de hardware. A infraestrutura específica necessária para criar uma nuvem privada pode variar – instituições que já possuam um centro de dados próprio, por exemplo, poderão ter apenas de adicionar capacidade de forma a satisfazer os pré-requisitos da plataforma de cloud pela qual optaram. Tipicamente, sistemas de nuvem privada requerem uma infraestrutura de rede interna completa, bem como ligações externas de alta velocidade (para acessibilidade remota, etc.), para além de elementos como armazenamento de base de dados, servidores de aplicações, firewall, nós de controle e software de plataforma de cloud.

Ao optar por uma nuvem privada, uma instituição terá inicialmente um bom grau de privacidade de dados, uma vez que esta funciona tipicamente em uma infraestrutura localizada nas suas instalações ("on premise"). Contudo, se uma nuvem privada for gerida por um terceiro ou estiver instalada fora das instalações da empresa, é preciso criar processos que garantam a privacidade dos dados. Uma empresa poderá, a título de exemplo, criptografar os dados e dotar os seus funcionários com passwords rotativas que garantam acesso aos dados da empresa armazenados na nuvem privada.

A segurança no caso das nuvens privadas é extremamente importante devido às características muito específicas dos dados armazenados neste tipo de sistemas. Um hacker, por exemplo, estará muito mais interessado em tentar entrar em uma nuvem privada de um banco do que em uma nuvem pública, a qual poderá armazenar fotos e outros dados pertencentes a diversas empresas e que, na sua maioria, lhe serão inúteis.

Em termos de segurança, uma instituição deverá sempre realizar processos de avaliação de riscos, analisar a propriedade dos dados, ver de que forma os dados estão classificados, realizar auditorias e processos de monitorização e ter preparado um plano completo para lidar com potenciais brechas de segurança.

A cloud privada deverá estar sempre atrás de um firewall robusto e os clientes remotos deverão estar seguros antes de lhes ser permitido o acesso à nuvem.

Backup de dados na nova era

Há muitas ferramentas que podem ser usadas para realizar backups e recuperação de dados. No passado, as organizações que necessitavam fazer cópias de segurança dos seus dados recorriam sobretudo a sistemas de backup em fita magnética. Esta tecnologia é usada ainda hoje, mas a sua proeminência está declinando rapidamente (menos 30% de 2011 para 2012, segundo um estudo do Santa Clara Consulting Group), à medida a que as organizações mudam para sistemas de arquivos mais flexíveis que oferecem redundância sem necessidade sequer de considerar soluções de backup e recuperação.

Sistemas de arquivos como o XtreemFS, que é open source e foi projetado para as necessidades dos data centers e das nuvens, oferecem vários benefícios: podem ser instalados através de vários data centers e interligados via Internet, e serem capazes de lidar com falhas que ocorram através de uma área alargada. A ideia de sistemas de arquivos como este é que os dados armazenados sejam replicados através de todos os sistemas de armazenamento da rede. Como tal, caso um arquivo fique corrompido ou inacessível devido a um problema na rede ou falha de hardware (ou qualquer outra razão), o sistema de arquivos irá automaticamente recuperar esse arquivo a partir de outro ponto de armazenamento dentro da rede.

Uma vez que existam múltiplas cópias de dados, este sistema de arquivos oferece redundância em tempo real, o que significa que nunca será necessário recorrer a um processo de backup e recuperação. Estes sistemas de arquivos oferecem também outros benefícios, incluindo "lag" reduzido – uma vez que múltiplas cópias de dados pressupõem que existem múltiplos pontos de acesso e um caminho mais rápido para os dados para qualquer usuário.

(*) Florence Perrin é Senior Sales Manager Southern Europe da Western Digital

Fonte: CIO. [Portal]. Disponível em: <http://cio.uol.com.br/opiniao/2013/05/15/o-big-data-e-os-desafios-de-backup-e-armazenamento//>. Acesso em: 15 maio 2013.