Capítulo 3 - Tudo gera dados

# **O que é Big Data?**

Dados são informações provenientes de uma variedade de fontes, como pessoas, fotos, texto, sensores e sites da Web. Os dados também são fornecidos por dispositivos tecnológicos, como telefones celulares, computadores, quiosques, tablets e caixas registradoras. Recentemente, ocorreu um pico no volume de dados gerados por sensores. Agora os sensores são instalados em um número cada vez maior de locais e objetos. Eles incluem câmeras de segurança, semáforos, carros inteligentes, termômetros e até mesmo videiras!

Big Data é um grande volume de dados, mas o que é um grande volume? Ninguém sabe o número exato que diz quando os dados de uma empresa são considerados "Big Data". Estas são as três características que indicam que uma empresa pode estar lidando com Big Data:

* Tem um grande volume de dados que requer cada vez mais espaço de armazenamento (volume).
* Tem um volume de dados que está crescendo exponencialmente rápido (velocidade).
* Tem dados que são gerados em diferentes formatos (variedade).

Os sensores coletam quantos dados? Veja aqui alguns exemplos estimados:

* Os sensores de um carro autônomo podem gerar 4.000 gigabits (Gb) de dados por dia.
* Um Airbus A380 Engine gera 1 petabyte (PB) de dados em um voo de Londres para Cingapura.
* Os sensores de segurança em operações de extração podem gerar até 2,4 terabits (TB) de dados a cada minuto.
* Os sensores em uma smart connected home podem produzir 1 gigabyte (GB) de informações por semana.

Embora o Big Data realmente crie desafios para as empresas em termos de armazenamento e análise, ele também pode fornecer informações imprescindíveis para ajustar as operações e melhorar a satisfação do cliente.

Clique [aqui](http://cs.co/9004DiluE) para assistir a um vídeo sobre como o sistema de trânsito metropolitano de San Diego usa Big Data.

# **Conjuntos de dados grandes**

As empresas não devem, necessariamente, gerar seus próprios Big Data. As empresas menores podem não ter os sensores, o volume de clientes ou a capacidade de gerar a variedade de informações que podem se beneficiar suas empresas. Existem fontes de conjuntos de dados gratuitos disponível, prontas para serem usadas e analisadas por qualquer pessoa disposta a procurá-las.

Muitas empresas de vários portes acham que precisam coletar seus próprios dados para ver os benefícios dos dados analíticos de Big Data, mas isso não é verdade.

Clique [aqui](https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/) para visualizar um site que identifica os bancos de dados disponíveis.

# **Quais são os desafios do Big Data?**

Estimativas de Big Data da IBM concluem que "cada dia criamos 2,5 quintilhões de bytes de dados". Para colocar isso no contexto, a cada minuto todos os dias:

* Carregamos mais de 300 horas de vídeo do YouTube.
* Enviamos mais de 3,5 milhões de mensagens de texto.
* Fazemos streaming de mais de 86 mil horas de vídeo da Netflix.
* Curtimos mais de 4 milhões de postagens do Facebook.
* Solicitamos mais de 14 milhões de previsões do The Weather Channel.

Para ver mais estatísticas da Internet em tempo real, clique [aqui](http://www.internetlivestats.com/).

O rápido crescimento de dados pode ser uma vantagem ou um obstáculo, quando se trata de atingir as metas empresariais. Para obter sucesso, as empresas devem ser capazes de acessar e gerenciar seus ativos de dados com facilidade.

Com este enorme volume de dados sendo criado constantemente, as tecnologias e depósitos de dados tradicionais não podem acompanhar as necessidades de armazenamento. Mesmo com as instalações de armazenamento em nuvem, que são disponibilizadas por empresas como a Amazon, Google, Microsoft e muitas outras, a segurança dos dados armazenados se torna um grande problema. As soluções de Big Data devem ser seguras, ter uma grande tolerância a falhas e usar a replicação para assegurar que os dados não sejam perdidos. O armazenamento de Big Data não se trata apenas de armazenar os dados, mas também de gerenciá-los protegê-los.

Existem cinco problemas de armazenamento de dados principais relacionados ao Big Data:

* Gerenciamento;
* Segurança;
* Redundância;
* Dados analítiocs;
* Acesso.

# 

# **Onde podemos armazenar Big Data?**

Em geral, o big data é armazenado em vários servidores, normalmente dentro de data centers. Para segurança, acessibilidade e redundância, normalmente os dados são distribuídos e/ou replicados em vários servidores diferentes, em diversos data centers.

**Fog Computing (computação em neblina)**

A computação em neblina é uma arquitetura que usa dispositivos clientes ou “edge” dos usuários finais para fazer uma quantia considerável do pré=processamento e do armazenamento necessários para uma organização. A computação em neblina foi criada para manter os dados mais próximas da origem para pré processamento.

Os dados de sensores, especificamente, podem ser pré-processados mais próximos de onde foram coletados. As informações obtidas dessa análise processada previamente podem ser devolvidas aos sistemas das empresas, para modificar os processos, se necessário. Como os dados de sensores são pré-processados pelos dispositivos finais no sistema da empresa, as comunicações entre os servidores e dispositivos seriam mais rápidas. Isso requer menos largura de banda do que uma saída constante para a nuvem.

Após o pré-processamento de dados, muitas vezes os dados são enviados para armazenamento de longo prazo, backup ou uma análise mais detalhada dentro da nuvem.

# **A nuvem e a computação em nuvem**

Como mencionado anteriormente, a nuvem é um conjunto de data centers ou grupos de servidores conectados. O acesso ao software, armazenamento e serviços disponíveis nos servidores é obtido por meio da Internet através de uma interface de navegador. Os serviços em nuvem são fornecidos por muitas empresas de grande porte, como a Google, Microsoft e Apple. Os serviços de armazenamento em nuvem são fornecidos por diferentes provedores, como: Google Drive, Apple iCloud, Microsoft OneDrive e Dropbox.

Do ponto de vista individual, o uso dos serviços em nuvem permite:

* Armazenar todos os dados, como fotos, músicas, filmes e e-mails, liberando espaço em disco rígido local
* Acessar vários aplicativos, em vez de baixá-los no dispositivo local
* Acessar os dados e aplicativos em qualquer lugar, a qualquer momento e em qualquer dispositivo

Uma das desvantagens do uso de nuvem é que os dados poderiam cair em mãos erradas. Os dados ficam à mercê da robustez de segurança do provedor de nuvem escolhido.

Do ponto de vista de uma empresa, os serviços e computação em nuvem dão suporte a vários problemas de gerenciamento de dados:

* Permite acesso a dados organizacionais em qualquer lugar e a qualquer momento.
* Simplifica as operações de TI de uma empresa com a assinatura apenas dos serviços necessários.
* Elimina ou reduz a necessidade de equipamentos, manutenção e gerenciamento de TI no local
* Reduz o custo de equipamentos, de energia, de requisitos de plantas físicas e de necessidades de treinamento de funcionários.
* Permite respostas rápidas aos requisitos cada vez maiores de volume de dados.

# **Processamento distribuído**

Do ponto de vista de gerenciamento de dados, a análise era simples quando só os seres humanos criavam dados. O volume de dados era gerenciável e relativamente fácil de examinar. No entanto, com a explosão dos sistemas de automação empresarial e o crescimento exponencial de aplicativos da Web e dados gerados por máquina, a análise está se tornando cada vez mais difícil de gerenciar. Na verdade, 90% dos dados existentes atualmente foram gerados apenas nos últimos dois anos. Esse volume maior em um curto período é uma propriedade de crescimento exponencial. Esse grande volume de dados é difícil de processar e analisar em um prazo aceitável.

Em vez de grandes bancos de dados sendo processados por mainframes grandes e eficientes e armazenados em matrizes de disco gigantes (dimensionamento vertical), **o processamento de dados distribuídos** pega o grande volume de dados e o fragmenta em partes menores. Esses menores volumes de dados são distribuídos em vários locais para serem processadas por muitos computadores com processadores menores. Cada computador da arquitetura distribuída analisa sua parte do total do Big Data (dimensionamento horizontal).

Sistemas de arquivo mais distribuídos são desenvolvidos para serem invisíveis para os programas de cliente. O sistema de arquivo distribuído localiza os arquivos e transfere os dados, mas os usuários não têm como saber que os arquivos são distribuídos entre os vários servidores ou nós diferentes. Os usuários acessam esses arquivos como se fossem locais em seus próprios computadores. Todos os usuários veem a mesma exibição do sistema de arquivo e são capazes de acessar os dados simultaneamente com outros usuários.

[Hadoop](http://hadoop.apache.org/) foi criado para lidar com esses volumes de Big Data. O projeto Hadoop foi iniciado com duas facetas: O Hadoop Distributed File System (HDFS) é um sistema de arquivo distribuído tolerante a falhas e o MapReduce, que é uma forma distribuída de processar os dados. Agora Hadoop evoluiu para um ecossistema de software muito abrangente para o gerenciamento de Big Data.

Hadoop é o software de código aberto que permite o processamento distribuído de grandes conjuntos de dados, que podem ser em terabytes e que estão armazenadas em clusters de computadores. O Hadoop foi criado para escalar de servidores simples para milhares de máquinas e cada uma oferece armazenamento e computação locais. Para maior eficiência, o Hadoop pode ser instalado e executado em muitos VMs. Esses VMs podem trabalham em conjunto, paralelamente, para processar e armazenar os dados.

O Hadoop tem duas funcionalidades principais que o tornaram o padrão do setor para controlar o Big Data:

* Escalabilidade - Tamanhos maiores de cluster melhoram o desempenho e oferecem recursos de processamento de dados mais altos. Com o Hadoop, o tamanho de cluster pode facilmente escalar de um cluster de cinco nós a um cluster de mil nós, sem aumentar excessivamente o ônus administrativo.
* Tolerância a falhas – O Hadoop replica automaticamente os dados em todos os clusters para assegurar que os dados não serão perdidos. Se um disco, um nó ou um rack inteiro falhar, os dados estarão seguros.

# **Por que as empresas analisam os dados?**

Toda empresa deve se tornar mais eficiente e inovadora para permanecer competitiva e relevante no mundo digitalizado. A IoT faz parte do cumprimento dessa eficiência e inovação.

O objetivo de muitas empresas é coletar e analisar as enormes quantidades de dados de uso de produtos novos para obter informações valiosas. A análise de dados permite que as empresas entendam melhor o impacto de seus produtos e serviços, ajustem seus objetivos e métodos e ofereçam aos clientes melhores produtos com mais rapidez. A capacidade de obter novas ideias a partir dos dados agrega valor às empresas.

Para as empresas, agora os dados são o novo petróleo. Como o petróleo bruto, é valioso, mas se for não refinado não pode ser usado facilmente. O petróleo bruto deve ser transformado em gasolina, plástico, produtos químicos e outras substâncias para criar um produto de valor. É a mesma coisa com os dados. Os dados devem ser fragmentados e analisados para que tenham valor.

O valor é obtido dos dois tipos principais de dados processados, transacionais e analíticos. Informações transacionais são coletadas e processadas quando ocorrerem eventos. Informações transacionais são usadas para analisar relatórios diários de vendas e agendamentos de produção para determinar quanto de inventário deve ser transportado. Informações analíticas oferecem suporte a tarefas de análise gerencial, como determinar se a empresa deve criar uma nova fábrica ou contratar equipe de vendas adicional.

# **Fontes de informações**

A fonte de dados nos grandes conjuntos de dados é variada. Além dos dados de sensores, outros dados são provenientes de tudo que foi digitalizado, inserido e lançado na Internet de fontes como:

* Sites de mídia social - Facebook, YouTube, eHarmony e Twitter
* HTTP, páginas da Web e mecanismos de pesquisa na Internet
* Dados históricos de arquivos públicos e privados
* Metadados anexados a e-mails, documentos transmitidos e fotos
* Formulários médicos, formulários de seguros e formulários de impostos
* Pesquisa de genômica usando DNA

Os dados coletados podem ser categorizados como estruturados ou não estruturados.

Dados estruturados são criados por aplicativos que usam entrada no formato "fixo", como planilhas ou formulários médicos. Mesmo se os dados forem considerados estruturados, diferentes aplicações criam arquivos em diferentes formatos, que não são necessariamente compatíveis entre si. Os dados estruturados podem precisar ser manipulados em um formato comum, como CSV.

Os arquivos de valores separados por vírgula (CSV) são um tipo de arquivo de texto simples, que usa vírgulas para separar colunas em uma tabela de dados e o caractere de quebra de linha para separar linhas. Cada linha é um registro. Embora sejam usados normalmente para importação e exportação nos bancos de dados e planilhas tradicionais, não existe um padrão específico. JSON e XML também são tipos de arquivo de texto simples, que usam uma forma padrão de representação de registros de dados. Esses formatos de arquivo são compatíveis com uma ampla gama de aplicativos. A conversão de dados em um formato comum é uma forma valiosa para combinar dados de fontes diferentes.

Dados não estruturados são gerados em um estilo "freeform", como áudio, vídeo, páginas da Web e tweets. Os dados não estruturados requerem diferentes ferramentas para preparar os dados para processamento ou análise. Veja a seguir dois exemplos:

* As páginas da Web são criadas para fornecer dados aos seres humanos e não às máquinas. Ferramentas de “reaproveitamento da Web” extraem, automaticamente, dados de páginas HTML. É semelhante a um Web Crawler ou spider de um mecanismo de pesquisa. Ele explora a Web para extrair os dados e criar o banco de dados para responder às consultas de pesquisa. O software de captura da Web pode usar o Hypertext Transfer Protocol ou um navegador da Web para acessar a World Wide Web. Normalmente, a captura da Web é um processo automático que usa um bot ou web crawler para fazer a extração de dados. Os dados específicos são reunidos e copiados da Web para um banco de dados ou uma planilha. Os dados podem ser analisados com facilidade.
* Muitos provedores de serviços da Web de grande porte, como Facebook, fornecem interfaces padronizadas para coletar os dados automaticamente, usando interfaces de programação de aplicação (APIs). A abordagem mais comum é usar RESTful APIs. Os RESTful APIs usam o HTTP como o protocolo de comunicação e a estrutura JSON para codificar os dados. Sites da Internet, como Google e Twitter, coletam grandes volumes de dados em série de tempo e estática. O conhecimento das APIs para esses sites permite que os analistas de dados e engenheiros acessem os grandes volume de dados que estão sendo gerados constantemente na Internet.

# **Visualização dos dados**

A mineração de dados é o processo de transformar dados brutos em informações significativas, detectando padrões e relacionamentos em grandes conjuntos de dados.

Para ter valor, os dados extraídos devem ser analisados e apresentados para os gerentes e os tomadores de decisões. Existem muitas visualizações diferentes que podem ser usadas para apresentar o valor dos dados. A determinação do melhor gráfico para usar irá variar, de acordo com o seguinte:

* Número de variáveis para mostrar
* Número de pontos de dados em cada variável
* Os dados estão representando uma linha do tempo?
* Itens que exigem comparações

Alguns dos tipos de gráfico mais populares são de linha, coluna, barra, pizza e dispersão.

**Tipos de gráficos**

* Gráfico de linhas;
* Gráfico de pontos de dispersão;
* Gráfico de pizza;
* Gráfico de colunas;
* Gráfico de barras.

# **Analisar o Big Data para uso eficaz nas empresas**

Big data é exatamente isso – GRANDE! É muito útil se você puder analisá-lo para obter valor. A análise de dados é o processo de inspeção, limpeza, transformação e modelagem de dados para descobrir informações úteis. A análise de Big Data normalmente exige ferramentas e aplicativos criados para essa finalidade. Essas ferramentas de análise foram desenvolvidas para fornecer às empresas informações detalhadas, padrões e insights valiosos.

Antes de iniciar uma análise, é fundamental saber qual é o problema que a empresa está tentando solucionar ou quais informações a empresa está procurando. Estão interessados no comportamento do cliente em estados específicos, padrões de consumo de energia em diferentes quadrantes da cidade ou no número de "curtidas" do Facebook com base na idade?

Ter uma estratégia ajuda a empresa a determinar o tipo de análise necessária e a melhor ferramenta para fazer a análise. Uma estratégia também ajuda a determinar a maneira mais eficiente para apresentar os resultados para a gerência.

As ferramentas e aplicativos variam, do uso de uma planilha Excel ou do Google Analytics para amostras de dados pequenas a médias até aplicativos dedicados à manipulação e análise de fichas de dados realmente grandes.

Existem muitas ferramentas de dados analíticos de Big Data que uma empresa pode selecionar, como: Knime, OpenRefine, Orange e RapidMiner. Clique nos logotipos corporativos para saber mais sobre suas ofertas.

# **Resumo**

Geralmente, o Big Data tem três características. É um grande volume de dados que requer cada vez mais espaço de armazenamento (volume), que está crescendo exponencialmente rápido (velocidade) e é gerado em diferentes formatos (variedade).

A computação em neblina é uma arquitetura que usa dispositivos clientes ou “edge” dos usuários finais para fazer uma quantia considerável do pré=processamento e do armazenamento necessários para uma organização. A computação em neblina foi criada para manter os dados mais próximas da origem para pré processamento.

A nuvem é um conjunto de centros de dados ou grupos de servidores conectados que recebem, em qualquer lugar, acesso a software, armazenamento e serviços usando uma interface de navegador. Os serviços em nuvem fornecem um maior armazenamento de dados, conforme necessário, e reduzem a necessidade de equipamentos, manutenção e gerenciamento de TI no local. Também reduzem o custo de equipamentos, de energia, de requisitos de plantas físicas e de necessidades de treinamento de funcionários.

O processamento de dados distribuído pega grandes volumes de dados de uma fonte e os fragmenta em partes menores. Esses menores volumes de dados são distribuídos em vários locais para serem processadas por muitos computadores com processadores menores. Cada computador da arquitetura distribuída analisa sua parte do total do Big Data.

As empresas agregam valor coletando e analisando enormes volumes de novos dados de uso de produtos para entender o impacto de seus produtos e serviços, ajustar seus objetivos e métodos e oferecer aos clientes melhores produtos com mais rapidez.

Os dados coletados podem ser categorizados como estruturados ou não estruturados. Dados estruturados são criados por aplicativos que usam entrada no formato "fixo", como planilhas ou formulários médicos. Dados não estruturados são gerados em um estilo "freeform", como áudio, vídeo, páginas da Web e tweets. Os dois formatos de dados precisam ser manipulados em um formato comum para serem analisados. CSV, JSON e XML são tipos de arquivo de texto simples, que usam uma forma padrão de representação de registros de dados. A conversão de dados em um formato comum é uma forma valiosa para combinar dados de fontes diferentes.

A mineração de dados é o processo de transformar dados brutos em informações significativas, detectando padrões e relacionamentos em grandes conjuntos de dados. Visualização de dados é o processo de pegar os dados analisados e usar gráficos como linha, coluna, barra, pizza ou dispersão para apresentar informações significativas. Uma estratégia ajuda a empresa a determinar o tipo de análise necessária e a melhor ferramenta para fazer a análise. Uma estratégia também ajuda a determinar a maneira mais eficiente para apresentar os resultados para a gerência.