BIG DATA

AULA 1

Prof. Luis Henrique Alves Lourenço

CONVERSA INICIAL

Nesta aula, será apresentado a você um panorama sobre o conceito de Big Data. Estudaremos o contexto do qual surge esse conceito, os fundamentos que o definem e que são importantes ao tema. E abordaremos as etapas necessárias para extrair informações valiosas dos dados e apresentá-las.

TEMA 1 – A ERA DOS DADOS

Os avanços tecnológicos das últimas décadas nos trouxeram cada vez mais capacidade para medir e avaliar os eventos que acontecem a cada instante à nossa volta. A necessidade de gerar informações valiosas com base nos dados obtidos por tantos mecanismos de medição tem sido discutida e estudada a fim de combinar o uso de tais tecnologias para capturar, administrar e processar a quantidade cada vez maior de dados gerados das mais distintas formas.

Desde a popularização dos computadores e demais equipamentos digitais, vivemos em uma época de transição entre um mundo em que todos os dados eram gerados e armazenados em mídias analógicas para outro, em que os dados são digitais. Com essa revolução digital, surge a possibilidade de um volume muito significativo de dados a explorar. Enquanto nos anos 1990 havia poucos setores digitalizáveis, limitados a alguns segmentos da música e mídia, no início dos anos 2000 setores como o comércio eletrônico e o internet banking iniciaram sua transição para o digital, e hoje quase todos os aspectos da vida cotidiana passam por formatos digitais.

Dados da web, mídias sociais, transações das mais diversas naturezas (entretenimento, financeiro, telecomunicações), dados biométricos, relatórios, logs, documentos e muitos outros tipos de dados gerados a cada instante permitem a construção de aplicações que antes pareciam impossíveis devido ao alto custo e complexidade. O volume e a variedade dos dados gerados continuam a crescer, tornando sua análise cada vez mais complexa. No entanto, é cada vez mais necessário que tais análises ocorram, para que se possa produzir informações valiosas em tempo real.

Para responder à demanda pelas informações valiosas que podem ser obtidas com os dados, deve-se prestar especial atenção a certos fatores para processá-los e analisá-los, como: a relevância dos dados; a velocidade

necessária para processá-los; quão variados precisam ser; qual seu nível de atualização; entre diversos outros fatores que podem ser cruciais para as informações extraídas dos dados e que podem refletir a realidade e gerar o valor esperado.

Dessa forma, começamos a definir o conceito de *Big Data* não só como uma solução empacotada que pode ser colocada em prática adquirindo certa tecnologia com um fornecedor, mas como o conjunto de práticas e técnicas que envolvem o processamento de um volume de dados confiáveis e variados com a velocidade necessária à geração de valor.

1.1 O crescimento do volume de dados

Um dos aspectos mais influentes no crescimento dos dados foi a conexão entre os equipamentos digitais pela internet. Em 1995, quando a internet estava nos primórdios, estima-se que menos de 1% dos dados eram armazenados em formato digital (Marquesone, 2016). A conexão entre diversos dispositivos eletrônicos permitiu a criação de uma diversidade antes inimaginável de serviços que hoje são amplamente utilizados, como a compra de passagens on-line, definição de trajeto por auxílio de GPS, reuniões por videoconferência, serviços de financiamento coletivo, busca e candidatura de vagas de trabalho on-line, serviços de streaming de vídeo e áudio, redes e mídias sociais, jogos on-line, compras via comércio eletrônico, compras coletivas, internet banking, entre tantos outros.

Outro fator importantíssimo foi a adoção em grande escala de dispositivos móveis, que só foram intensamente popularizados devido à redução do custo de produção de equipamentos com poder de armazenamento adequado. Mesmo que houvesse a intenção de explorar os dados gerados, enquanto não houvesse poder de processamento ou capacidade de armazenamento suficiente a custos acessíveis, a maioria dos dados seria simplesmente descartada. Portanto, o aumento no poder de processamento, combinado com a redução de custo de armazenamento, contribuiu com o aumento do volume de dados.

As empresas puderam explorar o potencial contido em diferentes tipos de dados. Dados obtidos por diversos tipos de sensores e equipamentos finalmente puderam ser analisados, passando a gerar valor, até o ponto atual, em que o recente desenvolvimento das novas tecnologias de redes móveis permite que

sensores cada vez menores possam gerar uma quantidade gigantesca de dados com a interação entre os próprios equipamentos ou seu ambiente, pelo conceito de *internet das coisas* (do inglês *internet of things* – IoT). Segundo Cezar Taurion (2013, p. 29),

> A internet das coisas vai criar uma rede de centenas de bilhões de objetos identificáveis e que poderão interoperar uns com os outros e com os *data centers* e suas nuvens computacionais. A internet das coisas vai aglutinar o mundo digital com o mundo físico, permitindo que os objetos façam parte dos sistemas de informação. Com a internet das coisas podemos adicionar inteligência à estrutura física que molda nossa sociedade.

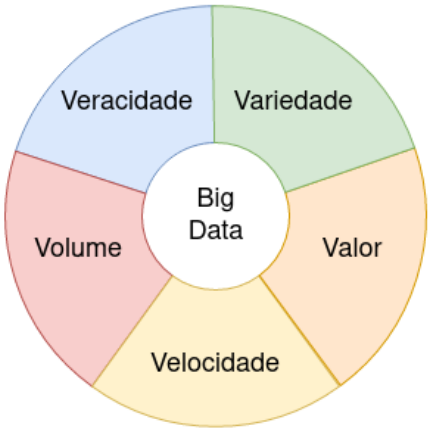
O crescimento da capacidade de processamento, segundo a lei de Moore, deve dobrar a cada 18 meses. Inicialmente esse crescimento ocorria devido à miniaturização dos processadores. Atualmente a paralelização do processamento permitiu que esse crescimento se mantivesse. Espera-se que no futuro o desenvolvimento de novas tecnologias, como a computação quântica, possa aumentar a capacidade de processamento. Além disso, as novas tecnologias de armazenamento, como os discos de estado sólido (SSDs), têm permitido seu barateamento e aumento de capacidade.

Podemos dizer que o aumento da geração de dados não deve ter seu ritmo reduzido tão cedo. Portanto, cada vez mais ferramentas poderosas são necessárias para analisar o imenso volume de dados gerados todos os dias no mundo.

TEMA 2 – OS Vs EM BIG DATA

Como acabamos de ver, o volume dos dados é um dos atributos mais relacionados ao conceito de Big Data, mas não é o único a defini-lo. Devemos considerar pelo menos outros dois aspectos: a velocidade em que os dados são processados e analisados, assim como a variedade dos dados, que podem ser obtidos de diversas fontes e se estruturar de diferentes formas.

Figura 1 – Os Vs em Big Data



2.1 Volume

O crescente volume de dados gerados a todo momento indica com relativa clareza que este é justamente o atributo mais significativo quando falamos em Big Data. No entanto, o que não fica muito claro é o ponto a partir do qual determinado conjunto de dados tem volume suficiente para ser considerado Big Data. Esse questionamento parte de uma premissa equivocada, pois o conceito de Big Data não pode ser definido única e exclusivamente pelo volume de dados processados. Um laboratório, por exemplo, pode necessitar de soluções de Big Data para visualizar imagens com 40 gigabytes, enquanto um observatório de astronomia pode necessitar de soluções de Big Data para analisar imagens e dados de sensores com terabytes de volume.

Portanto, o que define se um volume de dados necessita de soluções de Big Data não é seu tamanho, mas sua relação com a escalabilidade, eficiência, custo e complexidade. É o ponto em que a aplicação de soluções de Big Data supera os limites alcançados pelas tecnologias tradicionais que não foram projetadas para suportar esse volume.

2.2 Variedade

Um conjunto de dados de Big Data pode vir de diversas fontes e em diversos formatos. No entanto, as tecnologias tradicionais utilizam majoritariamente bancos de dados relacionais, ou seja, bancos que, embora

muito eficientes, são projetados para armazenar dados previamente estruturados, respeitando a propriedade Acid para que a integridade dos dados seja garantida da seguinte forma:

- **Atomicidade**: garante que transações não tenham atualizações parciais, ou seja, devem se comportar de forma indivisível e ser feitas por inteiro, ou então não são feitas;
- **Consistência**: garante que transações não afetem a consistência do banco. Dessa forma, as transações são completadas apenas se não ferirem nenhuma regra de integridade do banco, levando o banco de dados de um estado consistente a outro estado também consistente;
- **Isolamento**: garante que transações concorrentes não interfiram nos eventos umas das outras. As transações devem ter o mesmo resultado, como se fossem executadas uma após a outra;
- **Durabilidade**: garante que todos os efeitos de uma transação completada com sucesso persistam mesmo na ocorrência de falhas externas.

No entanto, estima-se que a maioria dos dados existentes seja de dados não estruturados ou semiestruturados (Marquesone, 2016). Os dados semiestruturados compreendem estruturas previamente definidas, mas que não exigem o mesmo rigor que os bancos de dados relacionais – é o caso de arquivos nos formatos JavaScript Object Notation (JSON) ou eXtensible Markup Language (XML). Já os dados não estruturados são todos aqueles excessivamente complexos para serem armazenados apenas com ferramentas tradicionais de armazenamento e gerenciamento de dados. É o caso de mídias como vídeos, imagens, áudios e até mesmo alguns formatos de texto.

Os dados não estruturados ou semiestruturados exigem que a tecnologia adotada forneça uma estrutura flexível o bastante para que dados tão diversos possam ser analisados. Além disso, a estrutura deve permitir a utilização de ambientes distribuídos, incluindo a variedade de soluções e tecnologias necessárias para atender a demanda específica que a solução requer.

Há uma variedade muito grande de dados em variados formatos, sendo utilizados por uma variedade também muito grande de soluções com necessidades muito específicas. A solução de Big Data deve ser capaz de integrar e interagir com toda essa diversidade de dados.

2.3 Velocidade

Devido à popularização da internet, das mídias e redes sociais, dos dispositivos móveis e da internet das coisas, dados são gerados de forma cada vez mais rápida e por cada vez mais agentes. Por exemplo, um único carro moderno pode ter cerca de 100 sensores, que geram dados a cada instante. Por isso a velocidade em que esses dados são coletados, analisados e utilizados se torna cada vez mais importante.

Dados perdem valor com o tempo; por exemplo, sites de comércio eletrônico que atualizam seus preços de acordo com a demanda de suprimentos podem maximizar as vendas. Outro exemplo são os serviços de tráfego, que podem oferecer melhores rotas ou sistemas críticos que dependem das informações geradas em tempo real. Portanto, para muitos serviços, de nada adianta ter a capacidade de processar um volume imenso de dados, de diferentes lugares e formas, se a solução não for capaz de responder no tempo necessário para que a informação seja útil e não perca seu valor.

2.4 Valor

Ao definir o conceito de Big Data, alguns autores fazem referência a apenas três fatores (Vs): **volume**, **variedade** e **velocidade**. Reunidos, definem Big Data como a coleta e análise de um volume imenso de dados que podem vir de diversas fontes e ter uma grande variedade de formatos, numa velocidade altíssima. No entanto, podemos avaliar também quão valioso e significativo um dado pode ser para uma solução. Dessa forma, temos como saber quais dados podem gerar maior valor para a solução e, por isso, devem ser priorizados. Ao priorizar ou escolher os dados corretos, é possível otimizar a solução para que o valor gerado seja o mais adequado.

2.5 Veracidade

Outro fator de grande importância é saber quão confiável é o conjunto de dados que estamos utilizando. Isso impacta diretamente na confiabilidade da informação extraída. Estima-se que dados de baixa qualidade custem à economia trilhões de dólares anualmente. Dados com uma precisão inadequada

ou falsos podem levar a informações incorretas e até inviabilizar soluções. Portanto, se a veracidade dos dados coletados não for avaliada e garantida, corremos o risco de afetar o valor e a validade das informações que geramos.

2.6 Temos uma definição para o conceito de Big Data?

Dado o que consideramos, a definição de Big Data pode variar bastante. Alguns autores podem resumir a definição nos três principais fatores (volume, variedade e velocidade), enquanto outros chegam a utilizar até dez fatores (ou mais). No entanto, não divergem muito dos 5 Vs que vimos até aqui. Muitas vezes acontece de alguns autores aglutinarem algumas ideias em um dos fatores ou as subdividem em novos fatores que podem ser mais adequados à solução específica que se está desenvolvendo.

Portanto, para nossos propósitos, vamos definir Big Data como o conjunto de práticas e técnicas que envolvem a coleta e análise de um grande volume de dados confiáveis e variados (tanto no formato quanto na origem), com a velocidade necessária para gerar o valor adequado à solução. O que realmente importa é o valor que podemos gerar quando nos livramos das limitações relativas ao volume, variedade, velocidade e veracidade dos dados processados.

TEMA 3 – OBTENÇÃO E ARMAZENAMENTO DE DADOS

No que diz respeito ao Big Data, tudo se inicia com a obtenção dos dados que serão processados. Como vimos, os dados podem vir de diferentes origens e ter diferentes formatos. Pode ser que os dados necessários ainda não existam e precisem ser gerados, que sejam internos à própria aplicação ou que devam ser buscados de fontes externas. Todas essas questões fazem parte da fase de obtenção de dados e, para isso, estratégias de como os dados serão coletados e armazenados devem ser definidas.

3.1 Obtenção de dados

A obtenção de dados pode ser compreendida com diferentes estratégias de captura e utilização de dados no projeto.

- **Dados internos**: são os dados que o proprietário do projeto (empresa) já tem e cujo controle já detém. Tais dados podem vir de sistemas de gerenciamento da própria empresa, como: sistemas de gerenciamento de projetos; automação de marketing; sistemas *customer relationship management* (CRM); sistemas *enterprise resource planning* (ERP); sistemas de gerenciamento de conteúdo; dados do departamento de recursos humanos; sistema do gerenciamento de talentos; procurações; dados da intranet e do portal da empresa; arquivos pertencentes à empresa, como documentos escaneados, formulários de seguros, correspondências, notas fiscais, entre outros; documentos gerados por colaboradores, como planilhas em XML, relatórios em PDF, dados em CSV e JSON, e-mails, documentos de texto em diversos formatos, apresentações e páginas web; e registros de log de eventos, de dados de servidores, logs de aplicações ou de auditoria, localização móvel, logs sobre o uso de aplicativos móveis e logs da web;
- **Dataficação**: é a transformação de ações sociais em dados quantificados de forma a permitir o monitoramento em tempo real e análises preditivas;
- **Dados de sensores**: são dados inseridos no contexto da internet das coisas, em que os objetos se comunicam com outros objetos e pessoas. Para esse tipo de solução, é necessário prover um meio de transmitir dados entre os sensores e um servidor capaz de armazenar os dados das interações. A obtenção de dados de sensores ocorre em tempo real, por isso os maiores desafios estão no volume e na velocidade com que os dados são gerados. Exemplos de dados de sensores são aqueles coletados de medidores inteligentes, sensores de carros, câmeras de vigilância, sensores do escritório, maquinários, aparelhos de ar-condicionado, caminhões e cargas;
- **Dados de fontes externas**: são dados obtidos de domínio público, como dados governamentais, dados econômicos, censo, finanças públicas, legislações, entre outros. Podemos considerar também qualquer tipo de dados obtidos por sites de terceiros, como mídias e textos de sites da web, além dos dados obtidos de mídias sociais. São basicamente todos os dados possíveis de obter por requisições na web ou uma *application programming interface* (API) dedicada. Muitos desses dados são

disponibilizados por APIs acessadas via *representational state transfer* (Rest), obtendo-se os dados requisitados em formato JSON.

3.2 Armazenamento

Como vimos, os bancos de dados relacionais foram por muito tempo o padrão mundial de armazenamento. Nesse modelo os dados são armazenados de forma previamente definida em estruturas de tabelas que podem ser relacionadas com outras tabelas da mesma base de dados. Uma das características mais importante desse tipo de armazenamento é o suporte a transações Acid.

Outra característica importante é o uso de *structured query language* (SQL) para operações de criação e manipulação de dados, o qual permitiu que os dados armazenados tivessem sua integridade garantida, e também permitiu gerar consultas mais complexas. No entanto, o crescimento constante na geração de dados mostrou os limites dos bancos de dados relacionais como única solução de armazenamento, principalmente no que se refere à **escalabilidade**, **disponibilidade** e **flexibilidade**.

3.2.1 Escalabilidade

Uma solução é considerada escalável quando mais carga é adicionada e mesmo assim o desempenho se mantém adequado. Com determinado volume de dados, os bancos de dados tradicionais conseguem manter esse desempenho apenas ao adicionar mais recursos computacionais à infraestrutura, o que é conhecido como *escalabilidade vertical*. No entanto, o volume de dados necessários aumentou tanto que esse tipo de solução não é mais viável em todos os casos, uma vez que os custos de tais recursos podem ser muito elevados.

3.2.2 Disponibilidade

Para que um serviço seja considerado de alta disponibilidade, o tempo em que ele se mantém operando deve ser priorizado em comparação às demais propriedades Acid. Portanto, deve-se garantir que o serviço se mantenha operando mesmo em casos de falha na infraestrutura.

3.2.3 Flexibilidade

Um serviço flexível é aquele capaz de comportar uma grande diversidade de dados. O grande problema desse requisito é que muitas vezes é inviável modelar um conjunto de dados de forma antecipada e que contemple características não estruturadas.

Concluímos que os bancos de dados tradicionais já não são a solução mais adequada para suprir os requisitos exigidos em soluções de Big Data. Dessa forma, soluções alternativas foram criadas para atender a esse tipo de demanda.

3.3 NoSQL

A noção de NoSQL incorpora uma ampla variedade de tecnologias de bancos de dados desenvolvidos como resposta à demanda de aplicações modernas. Quando comparadas com bancos de dados relacionais, bancos NoSQL são mais escaláveis, têm melhor desempenho, e seu modelo de dados resolve questões que os bancos de dados relacionais não foram projetados para resolver, como grandes volumes de dados de estruturados, semiestruturados e não estruturados que se modificam rapidamente, arquiteturas distribuídas geograficamente, entre outras. Os modelos de bancos de dados NoSQL podem ser classificados de acordo com a estrutura em que os dados são armazenados. Existem vários modelos, e os quatro principais são: **orientado a chave-valor**, **orientado a documentos**, **orientado a colunas** e **orientado a grafos**.

3.3.1 Bancos de dados orientados a chave-valor

Os bancos de dados orientados a chave-valor são os modelos mais simples de NoSQL. Cada item é armazenado como um atributo-chave normalmente composto de um campo tipo *string* associado a um valor que pode conter diferentes tipos de dados. Esse modelo não exige um esquema predefinido, como acontece nos bancos de dados relacionais. Esse tipo de banco de dados pode ser utilizado tanto para armazenar os dados quanto para mantê-los em *cache* para agilizar o acesso. Portanto, é um tipo de banco muito importante para aplicações que realizam muitos acessos aos dados. Apesar das

vantagens dos bancos de dados chave-valor, ele tem limitações. A única forma de realizar consultas é por meio da chave, uma vez que não é possível indexar utilizando o campo valor.

3.3.2 Bancos de dados orientados a documentos

Os bancos de dados orientados a documentos são uma extensão dos bancos de chave-valor, uma vez que também associam uma chave a um valor. Mas nesse caso o valor é necessariamente uma estrutura de dados chamada *documento*. A noção de documento é o conceito central desse tipo de banco de dados, e consiste em estruturas de um padrão definido, tal como XML, YAML, JSON, ou até formatos binários.

O documento pode se comportar de forma muito semelhante ao conceito de *objeto* em programação. Além disso, permite-se um conjunto de operações muito semelhantes ao padrão Crud: *creation* (inserção), *retrieval* (busca, leitura ou requisição), *update* (edição ou atualização), e *deletion* (remoção, deleção). Isso permite criar consultas e filtros sobre os valores armazenados, e não somente pelo campo-chave.

Outra característica desse banco é a alta disponibilidade, uma vez que permite trabalhar com a replicação de dados em *cluster*, garantindo que o dado ficará disponível mesmo em caso de falha no servidor.

3.3.3 Bancos de dados orientados a colunas

Os bancos de dados orientados a colunas são otimizados para buscas em grandes bancos de dados. Podem ser interpretados como bancos chave-valor bidimensionais; neles, o que seria uma tabela num banco de dados relacional seria um item identificado por uma chave associada a um valor que pode conter vários conjuntos de chave-valor.

Tais conjuntos são o equivalente ao campo de uma coluna de determinado item. Isso permite flexibilidade, tal que cada registro pode ter um número diferente de colunas. Os bancos orientados a colunas também podem ter o conceito de famílias de colunas. Cada família tem múltiplas colunas que são utilizadas em conjunto, de maneira semelhante às tabelas dos bancos de

dados relacionais. Dentro de uma família de colunas, os dados são armazenados linha por linha, de forma que as colunas de uma linha sejam armazenadas juntas.

Esse banco de dados é altamente adequado a soluções que necessitam trabalhar com volumes imensos de dados, alto desempenho, alta disponibilidade no acesso, armazenamento de dados e flexibilidade na inclusão de campos. Além disso, sua solução tolera eventuais inconsistências.

3.3.4 Bancos de dados orientados a grafos

Os bancos de dados orientados a grafos são muito úteis quando as relações entre os dados são mais importantes que os dados em si. Esse tipo de banco é utilizado para armazenar a informação sobre as redes de dados. Em vez de os dados serem formatados em linhas e colunas, são estruturados em vértices, arestas, propriedades para armazenar os dados coletados e os relacionamentos entre esses dados.

Os vértices representam entidades ou instâncias. Equivalem a um registro, uma linha dos bancos de dados relacionais, ou um documento num banco de dados orientado a documentos. As arestas (ou relações) são as linhas que conectam os vértices, representando a relação entre os dois vértices conectados. As arestas podem ser direcionais ou não direcionais, e propriedades são informações relacionadas com os vértices.

As soluções NoSQL não foram desenvolvidas para substituir os bancos de dados relacionais, mas para complementá-los. A tendência é adotar soluções híbridas, em que cada banco é utilizado onde possa ser mais adequado.

3.4 Governança de dados

Para qualquer empresa que adote estratégias de Big Data em suas soluções, é muito importante gerenciar dados de forma que seu uso seja o mais eficiente e confiável possível. A governança de dados inclui as pessoas, os processos e as tecnologias necessárias para proteger os ativos de dados da companhia, de forma a garantir que os dados da empresa sejam compreensíveis, corretos, completos, confiáveis, seguros e detectáveis. De acordo com Marquesone (2016), os principais tópicos na governança de dados são:

- **Arquitetura dos dados**: é o que define o modelo para gerenciar ativos de dados, alinhando-se à estratégia da empresa para estabelecer requisitos e projetos de dados estratégicos que atendam a esses requisitos. Todas as políticas que padronizam os elementos de conjuntos de dados, os protocolos e boas práticas são criadas para garantir a adoção dos padrões definidos;
- **Auditoria**: compreende que os dados devem permitir o próprio rastreio, e deve ser possível conhecer quando os dados foram criados, como estão sendo utilizados e quais os seus impactos;
- **Metadados**: são dados a respeito de outros dados. No contexto da governança de dados, é o que permite a visualização holística e acionável de uma cadeia de suprimentos de informação. É o que permite gerenciar as alterações nos dados, a auditoria e rastreabilidade do fluxo de dados, e também a melhora na acessibilidade dos dados por buscas e mapas visuais;
- **Gerenciamento de dados-mestre (*master data management* – MDM)**: dados-mestre são aqueles essenciais para o negócio de uma empresa. Podemos compreendê-los como o estabelecimento e gerenciamento de dados no nível organizacional, que fornecem dados-mestre precisos, consistentes e completos por toda a empresa e para parceiros de negócio;
- **Modelagem dos dados**: como vimos, é importante que toda a variedade de dados disponíveis seja modelada, de forma a permitir a utilização de padrões de dados, evitar redundâncias, definir como os dados serão utilizados, e encontrar as melhores formas de construir uma arquitetura de dados mais ágil e governável;
- **Qualidade dos dados**: compreende os processos incorporados com o objetivo de aperfeiçoar a qualidade dos dados de forma que contenham menos erros, estejam mais completos e possam otimizar a utilidade dos dados. Inclui a criação de *profiles* de dados, estratégias de limpeza, filtragem e agrupamento de dados;
- **Segurança**: relaciona-se à gestão de risco relacionado à coleta, armazenamento, processamento e análise dos dados. Isso implica que todos os dados importantes e sensíveis devem ser utilizados de maneira correta e segura, de forma a prevenir o mau uso em todos os níveis. Pode-

se utilizar estratégias de criptografia, definição e proteção a dados sensíveis, políticas de proteção de integridade, disponibilidade, confiabilidade e autenticidade dos dados.

A governança de dados tem se tornado ainda mais importante com a adoção cada vez maior de soluções de Big Data, uma vez que a veracidade e o valor dos dados são diretamente influenciados por seus processos, permitindo a criação de modelos de negócios mais inovadores, confiáveis e eficientes.

TEMA 4 – PROCESSAMENTO DE DADOS

Tão logo os dados são capturados e armazenados, inicia-se a fase de processamento dos dados. Para isso, devemos avaliar algumas questões relacionadas ao processamento, como alocação de recursos, escalabilidade, disponibilidade, desempenho e o tipo de processamento.

4.1 Escalabilidade

Uma das questões mais importantes quando tratamos de um volume de dados que pode crescer imensamente é a escalabilidade. Um sistema escalável é aquele em que o desempenho não se deteriora com o aumento significativo de dados sendo processados. A capacidade de processamento da plataforma deve escalar proporcionalmente à demanda. Para isso, é necessário monitorar a execução de forma a impedir o esgotamento de recursos.

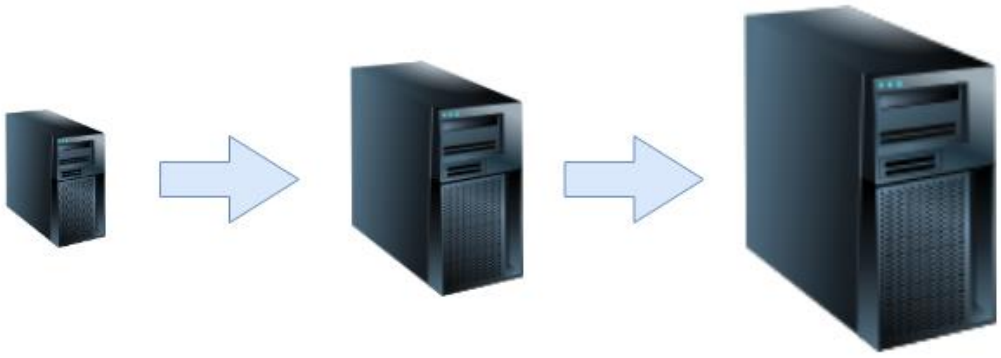
Outro aspecto importante é que não se deve sacrificar a disponibilidade da plataforma. Mesmo que ocorram falhas, o serviço deve manter-se ativo. No que se refere à escalabilidade, existem duas estratégias possíveis. Podemos adotar um modelo de **escalabilidade vertical** ou de **escalabilidade horizontal**.

4.1.1 Escalabilidade vertical

A escalabilidade vertical se refere à adição de capacidade de processamento de um único recurso com a atualização da infraestrutura. Aumenta-se a capacidade de processamento da plataforma pela atualização da infraestrutura.

Esse tipo de estratégia costuma comprometer a disponibilidade do serviço, a não ser que haja redundância na infraestrutura. Sua vantagem é não exigir modificações nos algoritmos, mas em geral é uma solução que não atende à demanda quando aplicada ao contexto em que o volume de dados cresce rapidamente, como é o caso de soluções de Big Data.

Figura 2 – Escalabilidade vertical

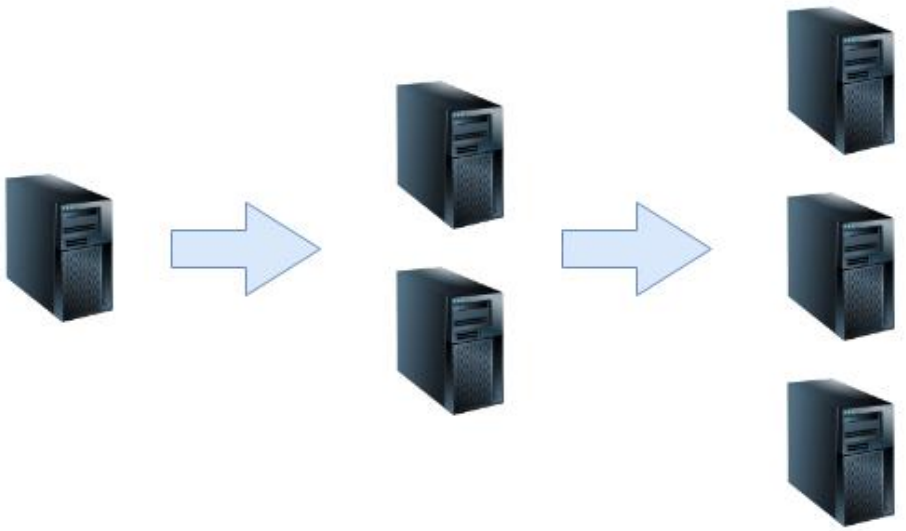


4.1.2 Escalabilidade horizontal

Uma estratégia muito mais adequada é a escalabilidade horizontal. Com ela o processamento é distribuído num conjunto de tarefas menores para serem processadas num *cluster* de recursos, de forma que o aumento da quantidade de recursos computacionais seja capaz de suprir o aumento na demanda de processamento. Além disso, os recursos do *cluster* se comportam de forma independente e redundante, colaborando com a disponibilidade da solução.

Dessa forma, o custo de melhorias na infraestrutura é reduzido, pois não é necessário interromper sua operação. A redistribuição de carga entre os recursos de um *cluster* é uma forma simples de regular a capacidade de processamento de acordo com a demanda; essa característica confere uma imensa capacidade de escalabilidade. Sua desvantagem é que a tecnologia tradicional não foi projetada para esse tipo de estratégia. Portanto, geralmente requer reimplementação de software para utilizar tecnologias de sistemas distribuídos.

Figura 3 – Escalabilidade horizontal



4.2 Processamento de dados com Hadoop

Originalmente, o Hadoop foi projetado para funcionar como um motor de busca de código aberto. O framework Hadoop se baseia em duas tecnologias que visam o suporte ao armazenamento e processamento distribuído de grandes volumes de dados:

- O sistema de arquivos distribuído Hadoop Distributed File System (HDFS);
- O modelo de programação distribuída MapReduce.

As principais características do Hadoop que envolvem o processamento de grandes volumes de dados são:

- **Baixo custo**: por ser projetado para utilizar em servidores tradicionais, não exige a implantação de hardware específico;
- **Escalabilidade**: devido à adoção de tecnologias distribuídas, sua capacidade de processamento escala linearmente. Isso significa que o aumento de recursos de computação se reflete diretamente na capacidade de processamento. E não é necessário alterar a base de código cada vez que a infraestrutura é atualizada;

- **Tolerância a falhas**: seu modelo de escalabilidade horizontal garante que, mesmo que algum dos recursos apresente falhas, os recursos restantes supram a demanda;
- **Balanceamento de carga**: a tecnologia de processamento distribuído evita gargalos que podem limitar o processamento de recursos. Dessa forma, todos os recursos operam de forma otimizada;
- **Comunicação entre máquinas e sua alocação**: ocorre de forma transparente para o usuário.

Todas essas características são implementadas pelo Hadoop e permitem que o desenvolvedor concentre seus esforços na lógica do problema. Dessa forma, a análise de um grande conjunto de dados – anteriormente ignorados devido a custos inviabilizantes – foi permitida com o surgimento das tecnologias distribuídas HDFS e Hadoop MapReduce.

O sistema de arquivos distribuídos HDFS foi criado para gerenciar o armazenamento das máquinas do *cluster*. Ele tem escalabilidade para armazenar grandes volumes de dados de forma tolerante a falhas, com recuperação automática. A disponibilidade é garantida pela replicação de dados, e o sistema se encarrega de quebrar o arquivo em blocos menores, replicando-os algumas vezes em diferentes servidores.

O Hadoop MapReduce foi projetado para gerenciar o processamento de dados distribuído com a divisão de uma aplicação em tarefas independentes executadas em paralelo nos servidores do *cluster*. O processamento é dividido nas seguintes etapas:

- **Map**: recebe uma entrada de dados e retorna um conjunto de pares no formato de pares chave-valor. As operações dessa etapa são definidas pelo desenvolvedor;
- **Shuffle**: os dados retornados pelo Map são organizados de forma a aglutinar todos os valores associados a uma única chave. Para cada chave teremos um par que a associa com uma lista contendo todos os valores relacionados a essa chave como valor. Essa etapa é feita automaticamente;
- **Reduce**: os dados organizados são recebidos, e operações definidas pelo desenvolvedor são realizadas, gerando o resultado da aplicação.

4.3 Processamento em tempo real

Apesar de todas as vantagens do Hadoop, ele não é adequado a todas as soluções de Big Data, uma vez que foi desenvolvido para processamento em lote. Isso significa que primeiramente são formados grupos de dados coletados num período de tempo, para só então os dados serem processados. Desde que os dados tenham sido gerados até seus resultados serem processados e, então, respondidos, temos uma quantidade de tempo significativa. Além disso, para muitos casos, o processamento dos dados deve se dar de forma contínua. No entanto, no modelo em lote, o processamento se encerra tão logo os resultados são retornados.

Diferentemente, muitas aplicações precisam que os dados sejam processados à medida que chegam à aplicação – ou seja, em tempo real –, e cada item de dados que chega à aplicação é processado imediatamente. Para isso, o processamento em tempo real tem alguns requisitos importantes:

- **Baixa latência**: o tempo de processamento de um item de dado deve ser no máximo igual ao tempo em que novos dados chegam ao fluxo;
- **Consistência**: a solução deve ser capaz de operar com imperfeições e manipular inconsistências;
- **Alta disponibilidade**: etapas de coleta, transmissão e processamento de dados podem causar grandes impactos se ficarem indisponíveis, resultando na perda de dados significativos para a aplicação.

O processamento em tempo real é importante para soluções web com o rastreamento de usuários e análises de preferências, detecção de fraudes, redes sociais, com a identificação de tendências, além da internet das coisas, com milhares de objetos e sensores que geram dados o tempo todo.

Se soluções de processamento em lote, como o Hadoop, tiverem dificuldades em atender às demandas de velocidade necessárias para o processamento em tempo real, precisamos de uma solução que faça o processamento de fluxos de dados.

4.4 Processamento de dados com Spark

Spark é uma tecnologia que tem se destacado no processamento em tempo real, devido ao seu desempenho. Trata-se de um framework que estende o modelo de programação MapReduce, otimizando o desempenho em programação distribuída. O Hadoop compreende tanto um componente de armazenamento – o HDFS – quanto um componente de processamento – o Hadoop MapReduce. No entanto, o Spark concentra seus esforços no processamento de dados, podendo muitas vezes ser utilizado com o Hadoop, uma vez que seu processamento costuma ter desempenho muito superior ao Hadoop MapReduce.

Os principais componentes do Spark são:

- **Spark Core**: disponibiliza as funções básicas para o processamento, como Map, Reduce, Filter, Collect, entre outras;
- **GraphX**: realiza o processamento sobre grafos;
- **SparkSQL**: para a utilização de SQL em consultas e processamento sobre dados;
- **MLlib**: disponibiliza a biblioteca de aprendizado de máquina.

O Spark não conta com um sistema próprio de gerenciamento de arquivos, portanto, precisa ser integrado a um, como o HDFS do Hadoop, como foi sugerido. Mas também é possível utilizá-lo com uma base de dados em *cloud computing*. A arquitetura Spark é definida no Spark Core e é composta principalmente de três componentes principais:

- **Driver Program**: aplicação principal que gerencia a criação e executa o processamento definido pelo programador;
- **Cluster Manager**: administra o *cluster* de máquinas quando a execução for distribuída;
- **Workers**: executam as tarefas enviadas pelo Driver Program.

Os conceitos mais importantes utilizados na programação e no desenvolvimento de soluções com Spark incluem:

- **Resilient Distributed Dataset (RDD)**: funciona como uma abstração do conjunto de objetos distribuídos pelo *cluster*. É o objeto principal do modelo de programação no Spark;
- **Operações**: são as transformações e ações realizadas num RDD;
- **Spark Context**: objeto que representa a conexão da aplicação com o *cluster*. Pode ser utilizado para criar RDDs, acumuladores e variáveis no *cluster*.

TEMA 5 – ANÁLISE E VISUALIZAÇÃO

Apenas recentemente a capacidade de armazenamento e processamento se tornaram suficientes para permitir que dados antes ignorados fossem analisados. Entretanto, além dos componentes tecnológicos, o analista de dados deve ser capaz de identificar quais dados se deve utilizar, como integrá-los, quais perguntas serão úteis para a tomada de decisão, e qual a melhor maneira de apresentar os resultados obtidos da análise.

5.1 Análise de dados

A extração de informações úteis pela análise de dados não é uma tarefa simples. Em muitos casos, os dados podem ter informações incompletas, inconsistências, caracteres indesejados, estarem corrompidos, duplicados, em formato inadequado, e outros tipos de problema. Segundo Marquesone (2016), cerca de 80% do tempo da análise de dados é utilizado apenas para limpar e preparar os dados.

Durante a análise, podemos constatar a importância da qualidade dos dados utilizados pois, sem um processo de inspeção, muitos dados incorretos podem ser descartados. No entanto, mesmo que a qualidade seja garantida, a busca por padrões nos dados ainda é um grande desafio. É muito fácil analisá-los de forma errada, ao não se identificar corretamente relações de correlação e causalidade, e propagar erros que invalidem toda a análise. Por fim, é necessário validar todos os resultados gerados pela análise dos dados antes de serem utilizados.

5.2 O processo de análise de dados

A análise de dados inclui como atividades a identificação de padrões nos dados, sua modelagem e classificação, detecção de grupos, entre muitas outras. Para isso, utilizamos técnicas matemáticas, estatísticas e de aprendizado de máquina. O aprendizado de máquina pode ser muito útil na automatização da construção de modelos analíticos. Pode-se extrair informações úteis e padrões ocultos em conjuntos massivos de dados.

Os processos de análise de dados podem ser definidos de diversas formas. Uma delas seria por um padrão aberto conhecido pela sigla CRISP-DM (*cross-industry standard process for data mining*), que define as seguintes fases e tarefas:

- **Entendimento de negócio**: determinar os objetivos de negócio, seu contexto e critérios de sucesso; avaliar recursos disponíveis, riscos e contingências, definir terminologias e calcular custos e benefícios; determinar os objetivos da mineração de dados e seus critérios de sucesso; e produzir um plano de projeto. Nessa fase são definidas as perguntas, os objetivos e os planos;
- **Compreensão dos dados**: fazer a coleta inicial e descrever os dados; fazer análise exploratória; e verificar a qualidade dos dados. O objetivo é entender a estrutura, atributos e contexto em que os dados estão inseridos;
- **Preparação dos dados**: descrever o conjunto, selecionar, filtrar e limpar os dados, e minimizar a geração de resultados incorretos; construir dados (atributos derivados, registros gerados); integrá-los (mesclagem ou redução); formatá-los e estruturá-los;
- **Modelagem dos dados**: selecionar técnicas de modelagem; projetar testes; definir e construir o modelo de dados, seus parâmetros e sua descrição; e validar o modelo e definir os parâmetros a revisar. Para construir o modelo, utilizamos tarefas de algoritmos de extração de padrões que podem ser agrupadas como atividades descritivas ou preditivas;

- **Avaliação do modelo**: avaliar os resultados do modelo; revisar processos; e determinar os passos seguintes. Avalia-se a precisão dos resultados gerados com os modelos de dados;
- **Utilização do modelo**: planejar a entrega; planejar o monitoramento e a manutenção; produzir relatório final; e documentar e revisar o projeto. Os modelos aprovados são então utilizados e monitorados.

5.3 Visualização de dados

Obtidos os resultados pela análise dos dados, ou até antes disso, é interessante poder comunicar as informações obtidas. Ao mesmo tempo, é importante entender quais informações são mais relevantes e qual a melhor forma de apresentá-las com clareza. Uma vez que nós, como humanos, somos dotados de grande percepção visual, a representação dos dados de forma gráfica é muito eficiente para expressar as informações que obtivemos dos dados. Assim, a visualização de dados é definida como **a comunicação da informação utilizando representações gráficas**.

5.3.1 Visualização exploratória

A análise de dados requer que eles sejam avaliados de forma detalhada. Para melhorar a compreensão deles nessa etapa, é possível usar a visualização exploratória, que auxilia na identificação de estruturas das variáveis, das tendências e de relações, permitindo até mesmo a detecção de anomalias nos dados.

Existem muitas formas de representar dados graficamente. Cada uma delas é capaz de exibir certo nível de detalhamento ou destacar características específicas. Inclusive é muito comum que o analista de dados use diferentes tipos de gráfico para estruturar os dados conforme necessário, para melhorar sua compreensão.

5.3.2 Visualização explanatória

Quando o analista já tiver resultados concretos, ele está pronto para comunicar suas percepções. Essa etapa é definida como *visualização explanatória*. Durante ela, o interesse do analista é destacar os detalhes

importantes, de forma a comunicar os resultados obtidos de um grande volume de dados em informações mais concisas e de fácil compreensão no formato de uma interface visual. Em muitos casos, essa informação permite revelar tendências e desvios que podem servir de apoio a tomadas de decisão.

Uma interface visual permite que o leitor veja características específicas e detalhadas dos dados. Para isso, existem muitos atributos que devem ser analisados durante a criação dos gráficos. Cada tipo de gráfico pode ser mais adequado para comunicar certa informação a respeito de seus dados. Por exemplo, gráficos de colunas, barras, áreas circulares, linhas e de dispersão são mais adequados para comparar valores, enquanto gráficos de dispersão, histogramas e gráficos de área são mais adequados para destacar a distribuição de um conjunto de dados. Para cada necessidade, existe algum tipo de gráfico apropriado.

5.4 A visualização de dados

A literatura mostra que existem sete etapas para a visualização de dados, incluindo algumas que fazem parte das etapas de coleta, armazenamento, processamento e análise de dados.

1. **Aquisição**: etapa em que ocorre a coleta de dados;
2. **Estruturação**: define-se a estrutura em que os dados são padronizados;
3. **Filtragem**: dados incorretos, incompletos ou desinteressantes para a análise são removidos;
4. **Mineração**: parte da etapa de análise de dados que extrai informações dos dados;
5. **Representação**: etapa da análise exploratória que gera um modelo visual básico de dados;
6. **Refinamento**: técnicas gráficas para tornar a visualização mais eficiente;
7. **Interação**: inclui funcionalidades que oferecem melhor experiência para o leitor.

Existem ferramentas que auxiliam a visualização de dados a ponto de automatizar muitas etapas. Como vimos, ela pode ser importante durante a análise dos dados, pois contribui para resultados com maior precisão, ou ainda atende soluções que necessitam de visualização em tempo real.

FINALIZANDO

Neste capítulo, vimos uma introdução aos principais fundamentos que compõem o conceito de Big Data. Iniciamos com uma breve contextualização histórica da evolução da geração de dados e dos avanços tecnológicos que aumentaram o volume de dados a uma escala imensa, permitindo que dados anteriormente ignorados ou descartados passassem a ser analisados de forma cada vez mais detalhada.

Vimos que uma solução Big Data pode ser definida pela coleta, processamento, análise e visualização de um volume muito grande de dados confiáveis (quanto à veracidade) e variados (tanto no formato quanto na origem), com a velocidade necessária para gerar o valor adequado à solução. Vimos um pouco sobre os processos de coleta, armazenamento, processamento, análise e visualização de dados, passando por noções básicas de bancos de dados não relacionais (NoSQL), Hadoop, Spark, entre outros.

REFERÊNCIAS

TAURION, C. **Big Data**. Rio de Janeiro: Brasport, 2013.

MARQUESONE, R. **Big Data**: técnicas e tecnologias para extração de valor dos dados. São Paulo: Casa do Código, 2016.