Big Data e o apoio à decisão

Conceituação e importância do Big Data na tomada de decisão pela empresa, envolvendo os principais Vs do Big Data, como volume, velocidade e valor.

Prof.ª Daisy Albuquerque

Propósito

O conhecimento dos conceitos, importância e aplicabilidade do Big Data é essencial para os profissionais de Tecnologia da Informação (TI). O Big Data está relacionado à coleta, análise, transformação e interpretação de um grande volume de dados carregado de grande variedade e sendo gerado em alta velocidade. Para isso, são necessárias soluções específicas que permitam aos profissionais de TI manipular esses dados, estruturados ou não.

Objetivos

- Reconhecer os conceitos básicos de Big Data.
- Analisar a infraestrutura e tecnologia de projeto de Big Data.
- Analisar a estratégia de Big Data nas empresas.

Introdução

O Big Data é uma tecnologia que prevê os comportamentos dos consumidores e clientes capazes de influenciar as decisões de negócio com o objetivo de conquistar melhores resultados.

Por isso, neste conteúdo, entenderemos o conceito e a importância do uso de Big Data na tomada de decisão pelas empresas. Veremos, também, que Big Data é muito mais do que um grande volume de dados, pois igualmente está relacionado com a velocidade de produção desses dados, suas formas e diversas origens.

Além do conhecimento e da importância do Big Data, aprenderemos sobre sua infraestrutura e tecnologias envolvidas na sua implementação e aplicabilidade.

Por fim, identificaremos a estratégia a ser utilizada pelas empresas na implementação e aplicação das ferramentas de Big Data com o objetivo de agregar conhecimento estratégico para a tomada de decisão.

A história do Big Data

A história da humanidade é repleta de fatos que comprovam a importância dos dados. A seguir, conheceremos um pouco mais da história do Big Data no decorrer dos séculos. Então, acomode-se, pois lá vem história!

Era dos dados

7.000 AP

Contabilidade na mesopotâmia
Na Mesopotâmia, 7.000 anos atrás, a contabilidade era utilizada para registrar o crescimento de colheitas e rebanhos. Na História, consta este fato como um dos primeiros usos dos dados para agregar valor.

1662

Método de graunt
O demógrafo John Graunt criou o método para calcular as mortes provocadas pela peste bubônica que assolava a Europa na época.
O estudioso lançou as bases para a demografia e sua obra foi publicada no *Natural and Political Observations upon the Bills of Mortality*, sendo considerada uma das pioneiras no estudo atuarial de mortalidade. Basicamente, a obra continha uma rudimentar tábua de vida, obtida por meio de dados sobre enterros em Londres. Devido ao seu trabalho, Graunt é considerado o pai da Estatística.

1865

Inteligência de negócios
O termo **Inteligência de Negócios** é usado por Richard Millar Devens em *Cyclopaedia of Commercial and Business Anecdotes*; baseado em dados recuperados antes da concorrência, o documento descreve como o banqueiro Henry Furnese ganhou a concorrência se baseando em dados.

1880

Estatísticas nos EUA
O Departamento Americano de Estatística se preparava para calcular o censo no ano. Segundo estimativas da época, o trabalho levaria aproximadamente 8 anos para ser concluído. No mesmo ano, estimaram que com a tecnologia da época, os dados do censo de 1890 demorariam cerca de 10 anos para serem tabulados.

1890

Invenção de hollerith
O Dr. Herman Hollerith, estatístico norte-americano, utilizando-se da invenção do matemático inglês Charles Babbage (a máquina de perfurar cartões e de tabular e ordenar), conseguiu reduzir o processamento dos dados do censo de 1890 para dois anos e meio.

As décadas seguintes e a revolução dos dispositivos móveis formataram o que conhecemos atualmente como Big Data.

Era da informação

1927

Armazenamento em fita
Fritz Pfleumer, engenheiro austro-alemão, criou uma forma de armazenamento de dados em fitas magnéticas. O método colava faixas de metal em papéis de cigarro para evitar as manchas nos lábios dos fumantes pelas mortalhas disponíveis na época. O engenheiro resolveu usar essa técnica para criar uma tira magnética que poderia ser usada para substituir a tecnologia de gravação de fios, usada na época e muito rudimentar. Quando se gravava em fios, e o fio torcia, a gravação ficava do outro lado. A tecnologia de gravação de fios permite que o fio passe rapidamente através de uma cabeça de leitura e gravação, que magnetiza cada ponto ao longo do fio de acordo com a intensidade e a polaridade da energia elétrica do sinal de áudio que está sendo enviado para a cabeça.

1943

Máquina de Turing
Allan Turing, britânico, matemático e cientista da computação, desenvolveu a máquina de processamento de dados para decifrar os códigos nazistas durante a Segunda Guerra Mundial. Colossus, nome da máquina, procurava padrões em mensagens interceptadas com uma taxa de 5.000 caracteres por segundo.

1945

Primeiro artigo EDVAC
John von Neumann, brilhante matemático, publica o artigo *Computador Eletrônico Variável Discreto Eletrônico (EDVAC)* e lança a primeira discussão documentada sobre armazenamento de dados.

1965

Centro de dados
O governo dos Estados Unidos cria o Centro de Dados para armazenar em torno de 742 milhões de declarações físicas e 175 milhões de impressões digitais em fitas magnéticas de computador. Inicia-se a era do armazenamento eletrônico de dados.

1989

Criação da WWW
A World Wide Web (www) é criada pelo cientista da computação britânico, Tim Berners. O objetivo da www era facilitar o compartilhamento de informações por meio do sistema de hipertexto. A partir de então, os dados são criados à medida que mais e mais dispositivos são conectados à Internet.

Era do Big Data

2000

Estudo de Lyman e Varian
Em outubro desse ano, Peter Lyman e Hal R. Varian, professores da Universidade da Califórnia, Berkeley School of Information, publicaram o estudo *How much information?*, considerado a primeira publicação abrangente, em termos de armazenamento de computador, sobre a quantidade total de dados novos e originais criados no mundo anualmente e armazenados em mídias como papel, filme, CDs, DVDs e fita magnética. O estudo concluiu que, em 1990, produzimos 1,5 exabytes de dados.

2002

Artigo de Laney
Doug Laney, analista famoso do Grupo Meta, publicou o artigo: *Gestão de dados 3D: controlar o volume de dados, velocidade e variedade*. Ele articulou a definição de Big Data incluindo três Vs: volume, velocidade e variedade. Uma década depois, esses Vs foram considerados as três dimensões aceitas para a definição de Big Data. Nesse mesmo ano, segundo Peter Lyman e Hal R. Varian o mundo produziu cerca de 5 exabytes de dados em 2002 e 92% desses dados foram armazenados em mídias magnéticas, como os discos rígidos.

2005

Termo Web 2.0
Um ano após a criação do termo Web 2.0, Roger Magoulas, da O'Reilly Media, assinalou o termo Big Data pela primeira vez. O termo criado na época fazia referência a um grande conjunto de dados quase impossível de gerenciar e processar usando técnicas e ferramentas tradicionais de Business Intelligence. Nesse mesmo ano, a Yahoo! cria o Hadoop, baseado no MapReduce, da Google. Na verdade, essas ferramentas foram criadas com o objetivo inicial de indexar toda a World Wide Web. Atualmente, o Hadoop é um código aberto usado por muitas organizações para processar grandes volumes de dados, os Big Data.

A Web 2.0 evoluiu com o surgimento das redes sociais, e atualmente muito mais dados são criados diariamente.

2009

Armazenamento de dados
O governo indiano resolveu fazer uma varredura da íris (leitura/escaneamento da íris do olho, usado como identificação única da pessoa), uma impressão digital e uma fotografia de todos os seus 1,2 bilhões de habitantes. Nasce, assim, a maior base de dados biométrica do mundo.

2010

Conferência Techonomy
Na Conferência Techonomy, em Lake Tahoe, Califórnia, Eric Schmidt, doutor em Ciência da Computação e ex-CEO da Google, falou que, dos cinco exabytes de dados criados no mundo até 2003, agora essa mesma quantidade é criada a cada dois dias.

2011

Relatório da McKinsey
McKinsey, empresa de consultoria empresarial americana, publica o relatório: *Big Data: a próxima fronteira para inovação, competição e produtividade*, e afirma que em 2018 os EUA terão uma grande escassez de cientistas de dados (em torno de 1400.000 a 190.000 cientistas) e gerenciadores de dados.

2018

Relatório do Fórum Econômico
De acordo com o Fórum Econômico Mundial e a IBM, a matéria *Data Analyst, the most in-demand job of the coming years* indica que a demanda anual para cientistas de dados, desenvolvedores e engenheiros de dados pode chegar a 700.000 novos recrutamentos em 2020.

Além dos acontecimentos anteriormente relatados, outros fatores como a evolução dos computadores, Internet, equipamentos sensoriais (Internet das Coisas – IoT) e mídias sociais vêm modificando a natureza dos volumes de dados.

Diversas soluções de mercado, como cartão de crédito e **gateways de pagamento**, também possuem uma importância em fornecer cada vez mais dados; sendo assim, a evolução da tecnologia moderna está entrelaçada com a evolução do Big Data.

Gateway de pagamento

É um sistema de pagamento on-line que serve de comunicação direta entre o consumidor, o banco e a operadora do cartão de crédito.

O que é Big Data?

Big Data é um grande conjunto de dados com milhões de registros ou é um software?

Se termo Big, ao ser traduzido para o português, significa grande, e Data significa dados, então podemos conceituar Big Data como grandes dados ou megadados. Certo?

> Na verdade, Big Data é a extração de informações de um volume enorme de dados. Além de extrair informações, também atribui significado a elas, pautando estratégias e ações.

Outra conceituação possível de Big Data é um conjunto de metodologias utilizadas para capturar, armazenar e processar um volume imenso de dados de várias fontes, desde tabelas, fotos, textos e vídeos, com o objetivo de acelerar a tomada de decisão e trazer vantagem competitiva às organizações.

Dessa maneira, o Big Data deixa de ser apenas uma ferramenta de volume de dados e passa a ser um **mecanismo estratégico de análise**. Isso porque ao coletar, organizar e permitir a interpretação dos dados obtidos, é possível retirar **insights** importantes sobre questões variadas, como associações entre itens do supermercado para uma melhor organização com o foco do aumento das vendas, ou até padrões de clientes para uma produção de marketing direcionado.

insights

Palavra que vem do inglês e significa capacidade de obter uma compreensão intuitiva precisa e profunda de uma pessoa ou objeto, como um pensamento que surge e se encaixa com outro, trazendo solução ou entendimento para uma questão.

O grande diferencial do Big Data está na possibilidade de cruzamento dos dados oriundos de diversas fontes, permitindo insights rápidos e preciosos. Por isso, o Big Data é essencial nos dias atuais.

Ao manipulá-lo, é possível obter dados de mercado por meio dos consumidores extraindo registros de insatisfações, satisfações, desejos, necessidades, entre outros, além de captar das mídias sociais e efetuar o cruzamento com os dados internos das organizações.

Comentário

A essência do conceito de Big Data está em gerar valor para os negócios. Quanto mais dados coletamos, maior o esforço de processamento para gerar informações. E, assim, a velocidade para obter a informação faz parte do sucesso que o Big Data pode proporcionar às organizações.

Os Vs do Big Data

Inicialmente, o conceito de Big Data era composto por três Vs:

Volume

Velocidade

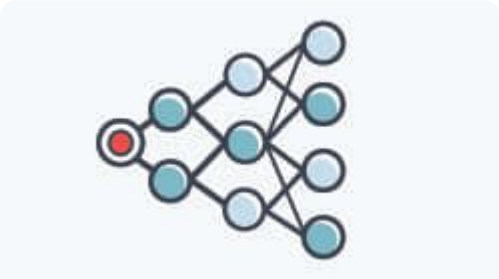
Variedade

Podemos dizer que os três primeiros Vs são os **principais pilares do Big Data**, porém, não são os únicos. Estudos mais recentes citam a existência de **sete Vs**, acrescentando os seguintes:

Veracidade

Valor

Visualização



Variabilidade

A seguir, conheceremos detalhes sobre cada um dos Vs do Big Data.

Volume

Volume é o primeiro grande desafio ao se manipular o Big Data. Atualmente, o volume de dados já é grande, mas a tendência é que continue a crescer ainda mais.

Segundo a IBM (International Business Machines Corporation), nós geramos mais de 2,5 exabytes de dados por dia. Imagine a situação, aproximadamente 90% dos dados gerados por nós foram criados nos últimos dois anos. E tudo indica que esse volume de dados dobre a cada ano, ou pelo menos nos próximos cinco anos. Uma prova disso é a quantidade de dados gerados por algumas das redes sociais mais utilizadas:

YouTube

Possui mais de 1 bilhão de usuários, que geram bilhões de visualizações.

Facebook

Possui bilhões de usuários gerando bilhões de curtidas por dia.

Twitter

Contabiliza em média 12 terabytes diariamente.

Em 2012, foram gerados cerca de 2.834 exabytes apenas no Twitter, cerca de milhões de Gigabytes e a previsão é de que, a partir de 2021, sejam gerados anualmente 40.026 exabytes de dados. Essa realidade não é apenas dessas grandes empresas; empresas do varejo também geram milhões de transações por dia.

> Atualmente, já passamos de gigabytes para terabytes, e agora já estamos nos referindo a petabytes ao manipular Big Data.

Mas, afinal, o que significa giga, tera e pera bytes? Veja a seguir:

1. **1 byte**
   É o suficiente para armazenar um caractere de texto em um computador.

2. **1 kilobyte**
   1.024 bytes são aproximadamente a informação contida em uma página de um livro.

3. **1 megabyte**
   1.048.576 bytes são aproximadamente o valor necessário para armazenar 1/5 da obra de William Shakespeare.

4. **1 gigabyte**
   1.073.741.824 bytes equivalem a uma hora de vídeo em baixa resolução.

5. **1 terabyte**
   1.099.511.627.776 bytes, na verdade, 385 terabytes, são necessários para guardar o catálogo da biblioteca do Congresso americano.

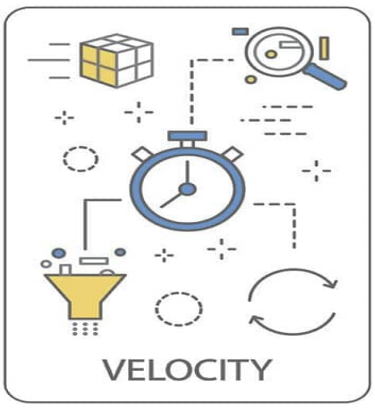
6. **1 petabyte**
   1.125.899.906.842.624 bytes equivalem ao armazenamento de mais de 4.000 fotos digitais por dia, durante toda a sua vida.

7. **1 exabyte**
   1.152.921.504.606.846.976 bytes, 3 exabytes são suficientes para guardar tudo que foi produzido pela humanidade em 1986; atualmente, produzimos quase o dobro em apenas dois dias.

8. **1 zettabyte**
   1.180.591.620.717.411.303.424 bytes, estima-se que todos os dados do mundo correspondem a 44 zettabytes no ano de 2021.

Ainda temos o yottabytes, com 1024 zettabytes, o brontobytes, com 1024 yottabytes, e, por último, o geobytes, com 1024 brontobytes.

Velocidade



A velocidade em que os dados são produzidos é impressionante, e eles se tornam desatualizados de forma vertiginosa. Em muitos casos, as respostas precisam ser rápidas ou em tempo real para o tratamento do grande volume de dados. Por exemplo, para um sistema antifraude, dois minutos pode ser muito tempo.

Quando nos referimos ao uso de Big Data de uma forma estratégica, a velocidade pode ser considerada até mais importante do que o volume em determinadas situações. Dessa forma, o segundo desafio do Big Data é a velocidade de processamento desses dados, que gera valor real e aplicabilidade. Assim, se faz necessária a utilização desses dados antes que se fiquem desatualizados. O objetivo é alcançar uma forma de processar os dados em tempo real.

Dica

Em se tratando de vantagem competitiva, pode ser mais interessante manipular uma quantidade um pouco menor de dados em tempo real do que uma enorme quantidade, que só poderá ser disponibilizada para uso depois de um tempo considerável.

Ao manipular um Big Data, a velocidade da coleta, organização e análise precisa ser proporcional ao alto volume de dados. Do contrário, seria um esforço incapaz de acompanhar os dados mais recentes.

O Big Data permite uma análise em tempo real de grandes volumes de dados sem necessariamente armazená-los em um banco de dados. Já existem plataformas totalmente automatizadas para capturar somente os dados estratégicos e relevantes em tempo real, de acordo com a natureza do negócio e os objetivos de cada organização. Vejamos, a seguir, exemplos de tomada de decisão para os casos de retorno em tempo real:

- Detecção de fraude em transação financeira;
- Oferta de produtos quando o cliente está navegando pela Internet ou quando está próximo ao ponto de venda;
- Determinação de doença grave no momento de realização de um check-up;
- Detecção de criminosos quando passam por uma câmera;
- Alerta de acidente aos condutores de veículos.

Variedade

Os dados que manipulamos atualmente são oriundos de diversas fontes, como: redes sociais, aplicativos, cookies, IoT, e-mails, dentre outros. Atualmente, cerca de 70% dos dados gerados são dados não estruturados, que não seguem uma modelagem padrão.

Variedade.

Isso significa que os dados não seguem um único padrão nem fornecem o mesmo tipo de informação. Dessa maneira, a tarefa de compilar esse grande volume em um banco de dados tradicional é impossível.

Atualmente, temos capacidade de capturar e analisar dados estruturados e não estruturados, texto, sensores, navegação Web, áudio, vídeo, arquivos de logs, catracas, centrais de ar-condicionado, entre outros. A complexidade em trabalhar com os dados é proporcional não só ao volume do dado, mas também em relação às diversas fontes dos dados, em formato muitas vezes distintos, e, assim, extrair informações valiosas deles.

Quanto mais dados e fontes eu tenho, maior é a complexidade em manipulá-los, porém também são maiores as possibilidades de gerar informação útil.

Comentário

Assim como o volume, a variedade de fontes de dados só tende a aumentar com o avanço tecnológico. Mas, assim como a velocidade, já existem ferramentas capazes de lidar com a heterogeneidade de dados, processá-los e agrupá-los de forma coerente.

Veracidade

A veracidade está relacionada à análise dos dados, se são verdadeiros ou não, pois precisam ser confiáveis. Lidar com os outros fatores, como volume, velocidade e variedade, pouco adianta se os dados não forem verdadeiros.

> É necessário que haja o máximo possível de consistência dos dados.

Diante do grande volume de dados que circula, é necessário estabelecer os dados verídicos e que ainda correspondem ao momento atual. Quando os dados se tornam desatualizados, podem ser considerados inverídicos; não porque tenham sido gerados com segundas intenções, mas porque já não correspondem à realidade e, dessa maneira, podem conduzir a organização a decisões equivocadas.

Comentário

Nosso desafio com a veracidade é determinar a relevância dos dados disponíveis de forma que esses dados possam servir como um tipo de guia para o seu planejamento com maior segurança. Nesse sentido, é importante a aplicação de filtros nos dados. É necessário destacar o que é risco no conteúdo para as organizações. Esse assunto tem sido pauta recorrente. O emaranhado de dados pode nos confundir, por isso todo cuidado é pouco para obtermos a veracidade dos dados.

Valor

Um dos mais importantes Vs do Big Data é o V de valor, valor que os dados geram para as organizações e para os usuários. A geração de valor é realizada a partir da aplicação dos aprendizados obtidos na análise e interpretação dos dados.

Já aprendemos que, para lidar com o Big Data, é necessário manipular um volume colossal de dados gerados a cada minuto de diversas fontes e formatos, dados estes que devem ser verídicos para gerar valor. O termo valor nada mais é do que o retorno relacionado ao investimento na tecnologia de Big Data, pois as organizações têm uma missão estratégica e o Big Data acaba surgindo para agregar valor e ajudar na execução desses objetivos.

> Na realidade, o valor é o resultado da combinação de todos os aspectos já estudados, pois o resultado não terá sentido algum se não trouxer benefícios significativos que compensem todo o investimento das organizações.

Visualização

O v de visualização está relacionado à maneira como os dados são apresentados, interpretados e consumidos.

Existem vários softwares desenvolvidos para a análise gráfica dos grandes volumes de dados. Alguns deles são o Tableau e o Power BI. E, dessa forma, é possível fazer uma análise das vendas de cada produto por ano, por trimestre ou por região, por exemplo.

O Tableau é uma plataforma de análise visual que está transformando a maneira como usamos dados para solucionar problemas, capacitando pessoas e organizações para que obtenham o máximo de seus dados. Já o Power BI é uma coleção de serviços de software, aplicativos e conectores que trabalham juntos para transformar suas fontes de dados não relacionadas em informações coerentes, visualmente envolventes e interativas.

Variabilidade

O v de variabilidade é diferente da variedade; trata-se do entendimento e interpretação dos dados com base no seu contexto.

Já sabemos que os dados são diferentes em relação aos dados estruturados e não estruturados, mas nem todos os dados se comportam da mesma maneira. Os dados podem fazer upload em velocidades diferentes, por exemplo.

> A compreensão da natureza e da extensão da variabilidade é essencial para o planejamento do processamento dos dados.

Para entendermos melhor, vamos imaginar uma cafeteria que oferece seis misturas diferentes de café: se você obtiver a mesma mistura todos os dias e tiver um sabor diferente a cada dia, isso é variabilidade. Da mesma forma, podemos fazer um paralelo com os dados; se o significado dos dados estiver mudando constantemente pode ocasionar um grande impacto na homogeneização da informação.

Variabilidade.

Viscosidade

Existe mais um v citado por alguns estudiosos, o v de **viscosidade**. A viscosidade está relacionada com a dificuldade de navegação entre os dados. Como os dados são variados, o algoritmo tem que ser capaz de lidar com diferentes fontes, e essa flexibilidade tem um custo.

Visão geral dos 8 Vs

Agora que conhecemos todos os Vs do Big Data, vamos entender melhor a sua importância. Na verdade, é fundamental conhecer esses fatores para podermos processar os dados de uma forma eficiente. Os oito Vs nos ajudarão a encontrar as melhores ferramentas para manipular os dados, para desenvolver os fluxos de trabalho baseados nos dados novos e nas diretrizes para manter a confiabilidade dos dados. E assim o Big Data fornecerá a análise e a inteligência de negócios de que precisamos para tomar decisões estratégicas e lucrativas.

Na imagem a seguir, é possível ter uma visualização de todos os 8 Vs.

Os 8 Vs do Big Data.

Estruturação dos dados

Dentre os Vs, a questão da variedade é marcante. Os dados são oriundos de diversas fontes, com alterações de acordo com o seu formato. Podemos classificar os dados em três formas de estrutura:

- Dados estruturados
- Dados não estruturados
- Dados semiestruturados

Veremos cada tipo a seguir.

Dados estruturados

Os dados estruturados são organizados e representados por uma estrutura rígida e previamente planejada para armazená-los. Algumas características dos dados estruturados são:

- Estrutura rígida e previamente planejada.
- Representação homogênea.
- Organização em blocos semânticos; no caso, as relações.
- Definições das mesmas descrições para dados de um mesmo grupo; no caso, os atributos.

Os dados estruturados possuem em cada campo um formato bem definido, considerado um padrão aceito pelo campo. Os dados que são do mesmo registro possuem uma relação entre eles. Os registros possuem valores diferentes, porém com os mesmos atributos. No caso, atributos ou campos são definidos por um determinado esquema. Para entendermos melhor, vamos analisar o exemplo. Imagine um formulário de cadastro com os seguintes campos:

Nome

O campo “nome” será do tipo texto, que corresponde a uma sequência de letras com ou sem a presença de espaços em branco, com um limite máximo de caracteres, e não será possível incluir números ou símbolos.

E-mail

O campo “e-mail” será do tipo textual, formado por uma sequência de caracteres e não só letras, pois admitirá números e alguns símbolos.

Idade

O campo “idade” será do tipo número, aceitando apenas um número inteiro positivo.

Uma pergunta com resposta sim ou não

O campo “pergunta” será do tipo binário, aceitando o valor 0 ou 1, sendo valor 0 para não e valor 1 para sim.

Dessa forma, cada campo possui um padrão bem definido que representa uma estrutura rígida e um formato previamente projetado para ele.

Na tabela a seguir, temos um exemplo dos dados do formulário.

Na tabela a seguir, temos um exemplo dos dados do formulário.

Nome	E-mail	Idade	Pergunta
Julia	julia@gmail.com	5 anos	1
Alana	alana@gmail.com	15 anos	1
Lucas	lucas@gmail.com	11 anos	0

Tabela: Exemplo de Formulário.
Daisy Albuquerque.

Os dados de um mesmo cadastro estão relacionados à mesma pessoa. Em outras palavras, os dados estruturados de um mesmo bloco ou registro possuem uma relação.

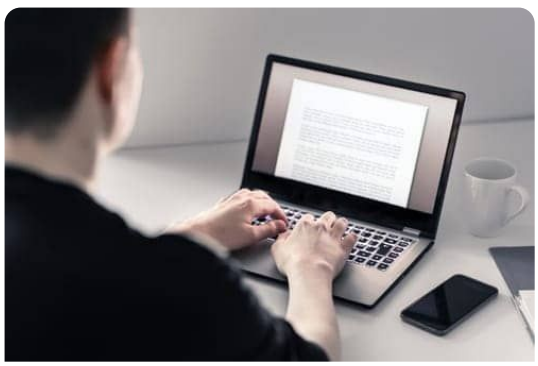
Registros ou grupos de dados diferentes (como de pessoas diferentes), possuem diferentes valores, mas utilizam a mesma representação estrutural homogênea para armazenar os dados. Ou seja, possuem os mesmos atributos (pense como sinônimo de campos no exemplo acima) e formatos, mas valores diferentes.

Comentário

O exemplo mais típico de dados estruturados é o banco de dados. Os dados são estruturados conforme a definição do esquema, que define as tabelas com seus campos ou atributos e seus formatos.

Dados não estruturados

Os dados não estruturados são o oposto dos dados estruturados. Nos dados não estruturados temos uma estrutura flexível e dinâmica, ou até mesmo sem estrutura.



Imagine um arquivo feito em um editor de texto. Nesse caso, é possível adicionar um texto sem se preocupar com campos, restrições e limites. No arquivo, também podemos incluir, além de texto, imagens, como gráficos e fotos misturados com textos.

> Atualmente, cerca de 70% do conteúdo digital gerado no mundo é do tipo não estruturado.

Algumas características de um dado não estruturado são:

- Sem estrutura predefinida;
- Constituem a maioria dos dados corporativos.

Alguns exemplos de dados não estruturados:

1 Arquivos de textos diversos
Páginas da Internet, relatórios, documentos, e-mails, mensagens em aplicativos como WhatsApp etc.

2 Arquivos de imagens
Fotos, gráficos, ilustrações, desenhos etc.

3 Arquivos de áudio
Música, streaming etc.

4 Arquivos de vídeo
Filmes, seriados, feitos por usuários etc.

5 Redes sociais
Blogs, Facebook, Twitter, Instagram, LinkedIn etc.

Dados semiestruturados

Os dados semiestruturados estão no meio termo entre os dados estruturados e os não estruturados. Um dado semiestruturado possui estrutura, porém ela é mais flexível. Sendo assim, ele agrega um pouco dos tipos de dados já estudados em termo de benefícios. Existe a facilidade de termos uma estrutura, mas também há uma certa flexibilidade.

O exemplo típico de um dado semiestruturado é um arquivo em XML (eXtensible Markup Language), que significa arquivo em linguagem de marcação estendida. Esse arquivo possui nós, que nada mais são do que rótulos de abertura e fechamento que aparecem precedidos com o símbolo "/", e cujos dados são inseridos entre eles. Imagine o seguinte texto:

Nome: Alana Silva

E-mail: alana@gmail.com

Endereço: Rua Siqueira Bueno, 1134, Liberdade, São Paulo

Agora, vamos transformar esses dados em um arquivo XML. O conteúdo ficará assim:

```
plain-text



 Alana Silva
 alana@gmail.com
 Rua Siqueira Bueno, 1134, Liberdade, São Paulo
```

A primeira linha identifica que o arquivo é uma estrutura XML e usa codificação de caracteres unicode.

Outros exemplos de dados semiestruturados são:

JSON

Acrônimo de JavaScript Object Notation, é um formato compacto, de padrão aberto independente, de troca de dados simples e rápida entre sistemas, especificado por Douglas Crockford em 2000, que utiliza texto legível a humanos, no formato atributo-valor.

RDF

Acrônimo de Resource Description Framework, é um modelo (de dados) abstrato para representação de informação na Web.

OWL

Significa Web Ontology Language, e é uma linguagem para definir e instanciar ontologias na World Wide Web.

Dados pequenos e dados em movimento ou em repouso

Já sabemos que os dados podem ter vários formatos, mas, além disso, os dados podem estar em movimento ou em repouso. Vamos conhecer os seguintes tipos de dados:

Dados em movimento

Quando os dados estão em movimento, são considerados dados de stream, ou melhor, dados em trânsito, que se movem na rede e de um lado para outro, ou de um nó para outro.

Para entendermos melhor, imagine uma live streaming (transmissão de vídeo via Internet e ao vivo). Os dados em movimento são mais difíceis de processar, principalmente devido ao custo; porém, se usados corretamente, são capazes de fornecer informações valiosas em tempo real para as organizações.

Esses tipos de dados são parte importante do Big Data, com o processamento e a análise em tempo real, à medida que estão sendo capturados.

Dados em repouso

Os dados em repouso, ou Data in Rest, são os dados armazenados em um determinado destino que não estão em uso nem viajando para outros destinos. Esses dados, ao chegarem no seu destino, ficam armazenados e recebem camadas de segurança como criptografia e proteção com senha. Os dados ficam em armazéns, com uma segurança envolvendo permissão de acesso aos usuários. A importância desses dados está em fornecer às organizações uma base para a operação e sua existência.

Dados pequenos

Os dados pequenos, ou Small Data, são um termo utilizado para se referenciar às pequenas quantidades de dados, ou uma quantia suficiente utilizada para a tomada de decisão. Esses dados possuem a característica de fornecer mais qualidade ao volume, pois o dado já está pronto, limpo e em condições de ser utilizado na análise do negócio.

Estamos falando dos dados provenientes de Sistemas de ERP ou CRM, que manipulam Small Data. Na verdade, o Small Data é uma solução complementar para o Big Data. Juntos podem ser utilizados para resolver problemas comuns, pois ambos possuem imenso significado nas organizações.

Em resumo: o Big Data é para as máquinas, enquanto o Small Data é para os seres humanos.

Big Data: definição e importância

No vídeo a seguir, abordaremos a definição de Big Data fazendo referência aos 8 Vs que compõem essa definição.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Vem que eu te explico!

Os vídeos a seguir abordam os assuntos mais relevantes do conteúdo que você acabou de estudar.

O que é Big Data?

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Os Vs do Big Data

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Estruturação dos dados

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

A utilização do Big Data, por sua importância, deve ser acompanhada de perto pelos profissionais da área da Tecnologia da Informação (TI), para que não ocorram problemas quando um sistema de armazenamento de

dados não suportar a massa de informação. Sobre os Vs que compõem o conceito de Big Data, avalie as afirmativas a seguir.

I - Velocidade se refere ao processamento ágil para gerar as informações necessárias.

II - Volume se refere ao valor obtido a partir dos dados, informação útil.

III - Valor se refere à grande quantidade de dados gerados.

IV - Variedade se refere às fontes de dados variadas, aumentando a complexidade.

Assinale a alternativa correta:

A

Somente a afirmativa I está correta.

B

Somente a afirmativa II está correta.

C

As alternativas I e IV estão corretas.

D

As alternativas II e III estão corretas.

E

As alternativas I e II estão corretas.

A alternativa C está correta.

A definição de Big Data é composta por vários termos, também conhecidos como os Vs do Big Data. Dentre eles, temos o volume, a velocidade, a variedade e o valor. Cada V tem um significado que agrega na definição final do Big Data, que é muito mais do que apenas uma grande base de dados.

No caso das alternativas I e IV, que estão corretas, velocidade se refere à agilidade de processamento do dado, e variedade corresponde ao grande número de tipos de dados.

Já nas alternativas II e III, incorretas, o volume está relacionado com quantidade, e não o valor; volume se refere à grande quantidade de dados gerados, e valor se refere ao valor obtido a partir dos dados.

Questão 2

Os dados podem ter vários formatos. Eles podem ser estruturados, não estruturados e semiestruturados. Assinale a alternativa que indica a qual tipo de dados, respectivamente, os exemplos abaixo se referem.

1. JSON e XML

2. Tabela de cadastros de CPF

3. E-mail e vídeo do YouTube

A

1- Dados semiestruturados; 2- Dados estruturados; 3- Dados não estruturados

B

1- Dados estruturados; 2- Dados não estruturados; 3- Dados semiestruturados

C

1- Dados semiestruturados; 2- Dados não estruturados; 3- Dados estruturados

D

1- Dados não estruturados; 2- Dados estruturados; 3- Dados semiestruturados

E

1- Dados não estruturados; 2- Dados semiestruturados; 3- Dados estruturados

A alternativa A está correta.

Os dados estruturados possuem como característica uma estrutura rígida como as tabelas de um banco de dados. Já os não estruturados se apresentam com uma estrutura mais flexível, como o e-mail. Por último, os dados semiestruturados possuem uma estrutura, porém com um formato mais flexível, como os arquivos XML.

Arquitetura para Big Data

A arquitetura de Big Data é projetada para lidar com coleta, armazenamento, processamento e análise de complexos e grandes volumes de dados estruturados e não estruturados.

O objetivo de um projeto de Big Data é tornar os dados corporativos disponíveis em um local centralizado, com redundância, alta disponibilidade e performance. Sendo assim, as características desejadas em uma arquitetura de Big Data são:

1

Reusabilidade

Facilidade de reutilização da arquitetura.

2

Manutenibilidade

Facilidade de modificar a Arquitetura a fim de se adequar a novos requisitos.

3

Modularidade

Arquitetura construída em módulos, ou seja, em ambientes separados.

4

Performance

Arquitetura de Big Data garante uma excelente performance na extração de conhecimento dos dados.

5

Escalabilidade

Capacidade de gerenciamento elevada e dados com uma previsão de aumento exorbitante.

Uma boa referência de estudos, padrões e arquiteturas está no diretório do NIST — National Institute of Standards and Technology — de Big Data Information NBD-PWG (NIST Big Data Public Working Group). No diretório, encontramos sete volumes de documentos.

O volume 2 fala sobre a taxonomia da arquitetura de Big Data, isto é, trata das determinações, descrições e classificações sistemáticas da arquitetura, descrevendo a hierarquia dos atores do processo em cinco principais componentes e duas camadas entrelaçadas.

Vamos conhecer alguns componentes e camadas, começando pelos componentes:

System Orchestrator

Responsável por definir e integrar as atividades de aplicativos de dados, necessárias em um Sistema Operacional Vertical.

Data Provider

Responsável por introduzir novos dados ou alimentações de informações no sistema Big Data.

Big Data Application Provider

Responsável por executar o ciclo de vida para atender aos requisitos de segurança e privacidade, bem como requisitos definidos pelo System Orchestrator.

Big Data Framework Provider

Responsável por estabelecer um framework de computação, no qual executa certas aplicações de transformação, protegendo simultaneamente a privacidade e a integridade dos dados.

Data Consumer

Responsável por incluir usuários finais ou outros sistemas que usam os resultados do Big Data Application Provider.

Agora, as camadas:

Management

Camada responsável pela gerência política de provisionamento, configuração, pacotes, software, backup, recurso e desempenho.

Security and Privacy

Camada responsável pela gerência das políticas de segurança.

Infraestrutura para armazenamento de Big Data

A Infraestrutura para armazenamento de Big Data é responsável por manter os dados coletados em um ambiente adequado. As principais opções de armazenamento são:

- Data Warehouse tradicional
- Data Lake
- Data Lakehouse

Vamos estudar cada uma dessas formas de armazenamento.

Data Warehouse

Por volta dos anos 1990, surge o conceito de **Data Warehouse (DW)**, ou armazéns de dados, depósitos que armazenam informações de uma forma mais consolidada, como dados históricos para classificação em blocos semânticos, chamadas relações.

Seu objetivo principal é integrar dados de diferentes sistemas com atualizações periódicas de longo prazo, possibilitando a visualização de relatórios gerenciais.

Data Warehouse.

> DW nada mais é do que um banco de dados relacional contendo principalmente dados estruturados. Os dados de um DW são divididos em subconjuntos chamados de Data Marts (mercado de dados).

Uma vez solicitados, os dados do Data Warehouse são disponibilizados em modo de leitura, conforme a demanda dos analistas de Big Data e **BI (Business Intelligence)**.

Para entendermos melhor, o DW é como um grande armazém da empresa onde todos os dados importantes estão armazenados. E, dentro do DW, os dados podem ser agrupados em conjuntos que fazem sentido para o negócio, como dados de RH, financeiro, vendas etc. Esses conjuntos são os **Data Marts**.

Os projetos de DW usam o processo de ETL, tradicionalmente, uma das etapas mais importantes e demoradas do Projeto de Dados. O processo de ETL se divide em três subprocessos:

Extrair (Extract)

Transformar (Transform)

Carregar (Load)

Unificados, livres de desvios e inconsistências, os dados do Data Warehouse rendem análises de alta precisão – que, por sua vez, geram informações e insights estratégicos.

Com o crescente e acelerado aumento da geração de dados, principalmente devido ao uso comercial e doméstico da Internet, manipular um grande volume de dados se tornou um desafio para os engenheiros de dados.

Para lidar com volume, velocidade e variedade, os três Vs principais do Big Data, os gestores de Tecnologia da Informação visualizaram o colapso das técnicas tradicionais de gestão da informação. Sendo assim, como tornar a gestão de dados mais eficiente, segura e economicamente sustentável? Para responder a essa pergunta, vamos estudar uma outra forma de armazenamento de dados, o **Data Lake**.

Data Lake

Nos anos 2000, surgiu uma tecnologia inovadora, o Data Lake (DL), ou lago de dados. A metáfora de um grande reservatório natural, cuja água pode ser filtrada para abastecer o seu entorno, foi criada por James Dixon, um dos fundadores do Pentaho.

> Data Lake é um repositório que centraliza e armazena todos os tipos de dados gerados pela e para a empresa.

As etapas de extração, carregamento e transformação em um Data Lake acontecem da seguinte forma:

1
Depósito dos dados brutos
Os dados são depositados no Data Lake ainda em estado bruto, sem processamento e análise, e até mesmo sem uma governança.

2
Extração dos dados
Após depositados, os dados são extraídos e carregados.

3
Carregamento dos dados
A etapa T do ETL (transformação) é pulada, isto é, ela ocorre após a etapa E e L. Dessa forma, após a extração é feito o carregamento dos dados.

4
Transformação dos dados
Por último, os dados são transformados.

Os projetos de DL usam o processo de ELT (extrair, carregar e transformar).

Sem tratamento, o repositório armazena volumes gigantescos de dados de qualquer tipo e em qualquer escala, podendo chegar às centenas de petabytes (1 PB é mais de mil terabytes!).

Uma grande vantagem de um Data Lake é o armazenamento dos dados na íntegra, com processamento sob demanda e de forma escalável. Por exemplo, voltando para o lago de dados, a água do lago pode ser filtrada para abastecer um caminhão-pipa ou garrafinhas de 500ml.

Como os dados do Data Lake, em sua grande parte, são não estruturados e, portanto, mais flexíveis, é possível poupar tempo e custo de armazenamento, facilitando a automação de processos e a inovação, impulsionando a transformação digital das empresas.

> Em resumo, o DL é uma solução que gerencia dados de forma econômica e dinâmica, alinhando a empresa com as tendências do mercado contemporâneo.

Na tabela a seguir, é possível visualizar uma comparação entre Data Warehouse e Data Lake.

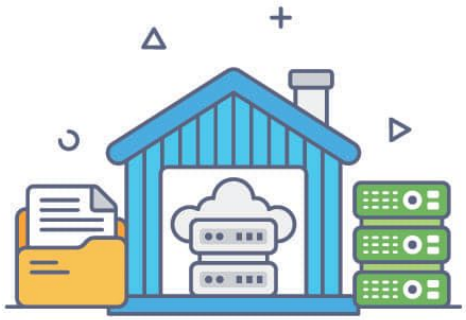
DATA WAREHOUSE	DATA LAKE
Dados estruturados.	Dados estruturados, não estruturados e semiestruturados.

DATA WAREHOUSE	DATA LAKE
ESQUEMA definido na escrita.	ESQUEMA definido na leitura.
ARMAZENAMENTO de dados frequentemente acessados, assim como dados agregados e sumarizados.	ARMAZENAMENTO de dados detalhados, brutos e processados.
CUSTO caro para grandes volumes de dados armazenados.	Projetado para baixo custo de armazenamento.
USUÁRIO: Business profissionais.	USUÁRIOS: cientistas de dados, analistas de dados.
UTILIZAÇÃO: relatórios business, análise de performance.	UTILIZAÇÃO IA (inteligência artificial) e modelos de *machine learning* .

Tabela: Data Warehouse vs Data Lake.
Daisy Albuquerque.

Data Lakehouse

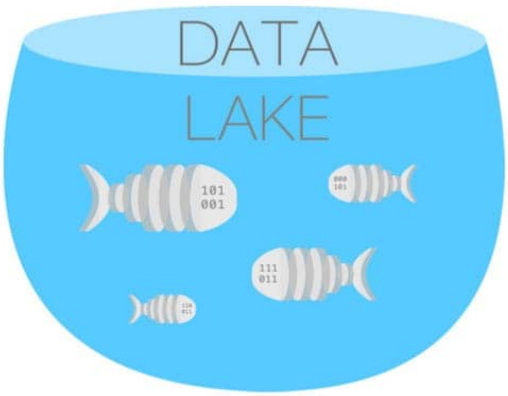
Os Data Warehouse são uma boa escolha quando se precisa de análise e processamento analítico de dados estruturados que se baseiam em dados históricos de várias fontes.



Data Warehouse.

Entretanto, atualmente, é cada vez mais necessário manipular dados não estruturados, semiestruturados e com alta variedade, velocidade e volume.

Sendo assim, as organizações começaram a coletar grandes quantidades de dados, em sua maioria, não estruturados e de muitas fontes diferentes. Os analistas sentiram necessidade de ter um único sistema para armazenar dados para muitos produtos analíticos e cargas de trabalho diferentes. E assim, cerca de uma década atrás, nasceram os Data Lake.



Data Lake.

Embora adequados para o armazenamento de dados, os lagos de dados precisam de alguns recursos essenciais.

Atualmente, as empresas usam sistemas para diversos aplicativos de dados, incluindo análise **SQL**, monitoramento em tempo real, ciência de dados e aprendizado de máquina.

SQL

Structured Query Language, ou Linguagem de Consulta Estruturada, é a linguagem de pesquisa declarativa padrão para banco de dados relacional.

Uma abordagem comum é usar vários sistemas – um Data Lake, vários Data Warehouses e outros sistemas especializados, como streaming, séries temporais, gráficos e bancos de dados de imagens. Nesse contexto, surgiram os **Data Lakehouse**.

> Os Data Lakehouse são repositórios que reúnem a implementação de estruturas de dados e recursos de gerenciamento de dados semelhantes aos de um Data Warehouse, com o tipo de armazenamento de baixo custo usado para Data Lakes.

O Data Lakehouse é um novo paradigma de armazenamento de dados com objetivo de simplificar radicalmente a infraestrutura de dados corporativos e acelerar a inovação em um período histórico em que o Machine Learning, ou aprendizado de máquina, está dominando diversos setores. As principais características de uma Data Lakehouse são:

1
Suporte a transações
Leitura e gravação de dados simultaneamente, normalmente usando SQL.

2
Aplicação e governança de esquema
Utilização de mecanismos robustos de governança e auditoria dos dados.

3
Suporte de BI
Utilização de ferramentas de BI diretamente nos dados de origem.

4
Armazenamento desacoplado da computação
O armazenamento e a computação usam clusters separados facilitando o escalonamento para vários usuários simultaneamente.

5
Abertura
Os formatos de armazenamento aberto e padronizados facilitam a utilização de ferramentas e mecanismos, como aprendizado de máquina, diretamente nos dados.

6
Suporte para diversos tipos de dados
Utilização do repositório para armazenar, refinar, analisar e acessar dados estruturados e não estruturados, incluindo imagens, vídeos, áudios, dados semiestruturados e textos.

7
Suporte para diversas cargas de trabalho
Utilização de ferramentas, incluindo ciência de dados e aprendizado de máquina, oferecendo suporte a toda carga de trabalho e dependente do mesmo repositório de dados.

8 Streaming de ponta a ponta
O suporte para streaming elimina a necessidade de sistemas separados dedicados a atender aplicativos de dados em tempo real.

Um Data Lakehouse combina os melhores elementos de um Data Lake e de um Data Warehouse.

Infraestrutura de computação e rede para Big Data

Normalmente, o Big Data é caracterizado por um volume extremo de dados, uma grande variedade de tipos e a velocidade com a qual devem ser processados.

Sendo assim, as características do Big Data impõem uma robusta infraestrutura de computação e de rede, que pode até sobrecarregar o servidor ou um cluster de servidores, gerando uma demanda de centenas ou milhares de servidores que distribuem o trabalho e operam de forma colaborativa.

> E como seria uma infraestrutura de computação e rede para Big Data?

Nesse contexto, surgiram as tecnologias de infraestrutura de computação e rede, que armazenam e processam os petabytes de dados e tecnologias analytics. Estamos nos referindo aos seguintes Bancos de Dados:

NoSQL

Para trabalhar processamento de muitos dados em tempo real, que permitem alto desempenho e recuperação baseada em índice.

MapReduce

Para processamento em lote, trata-se de um modelo computacional distribuído.

Vamos estudar, a seguir, o NoSQL, o MapReduce e o Hadoop.

NoSQL

Vamos conhecer um pouco da história do NoSQL:

1998

Origem do NoSQL
O termo NoSQL ou Not Only SQL foi usado pela primeira vez por Carlo Strozzi, como o nome de um banco de dados relacional de código aberto que tinha como objetivo ser uma implementação mais leve de um BD relacional; porém, sua principal característica era não expor a interface SQL.

06/2009

Movimento NoSQL
Finalmente, o movimento NoSQL ganhou força, em um encontro promovido por Johan Oskarsson e Eric Evans com o objetivo de buscar soluções open source de armazenamento de dados distribuídos não relacionais.

10/2009

Conferência NoSQL
Na conferência “no:sql(east)”, foi redefinido o uso do termo NoSQL para descrever soluções de armazenamento de dados não relacionais.

NoSQL nada mais é do que uma solução alternativa para os bancos de dados relacionais, com uma alta escalabilidade e desempenho. Seu surgimento é oriundo da necessidade de uma performance superior, uma maior flexibilidade e de uma alta escalabilidade para manipular Big Data.

Os bancos de dados relacionais atuais são muito restritos a isso, sendo necessária a distribuição vertical de servidores; ou seja, quanto mais dados, mais memória e mais disco um servidor precisa. Atualmente, no universo de ferramentas NoSQL, destacam-se quatro diferentes tipos de modelos de dados:

Modelo de dados orientado à Chave-Valor

Também conhecido como NoSQL Key-value, o Modelo de dados orientado à Chave-Valor utiliza-se de uma tabela hash com uma chave única e um indicador de um dado ou de um item em particular.

É recomendado para situações em que a integridade dos dados não é crítica, por exemplo, no caso de dados históricos e sessões de usuários, além de fóruns e websites de e-commerce.

Algumas soluções utilizadas atualmente no mercado são: DynamaDB, Redis, Voldemont.

Modelo de dados orientado à Família de Colunas (clone de BigTable)

Neste modelo, os dados são organizados em grupos de colunas, e tanto o armazenamento quanto as pesquisas de dados são baseadas em chaves.

Este modelo se tornou popular por meio do paper Big Table do Google, publicado em 2006, mostrando um sistema de armazenamento de dados distribuídos projetado para ter alta escalabilidade e suportar um grande volume de dados.

Este modelo é recomendado para cenários relacionados com a necessidade de lidar com grandes volumes de dados que precisam ser consultados com um tempo de resposta muito baixo, além da necessidade de armazenar uma estrutura de dados complexa.
Algumas soluções utilizadas atualmente no mercado são: Hbase, Cassandra e Accumulo.

Modelo de dados orientado a Documentos

Neste modelo, os dados têm o formato de documentos e cada documento pode ter campos similares ou não.

Este modelo permite armazenar qualquer documento, sem necessidade de definir previamente sua estrutura. Os documentos são armazenados em conjunto mesmo que não tenham nada em comum, esquema este chamado de schema-free.

Além disso, neste modelo existe a tendência de desnormalização dos dados, sendo possível deixar em um só documento ou registro todas as informações relacionadas.
Uma aplicação para este modelo são os sistemas de blog, os gerenciadores de conteúdo, pois em um único documento você pode agrupar o post e seus comentários.

Algumas soluções utilizadas atualmente no mercado são: MongoDB, Elasticsearch e CouchDB.

Modelo de dados orientado a Grafos

Este modelo é utilizado para extrair valor das relações entre os dados. Os bancos de grafos armazenam dados semiestruturados e suportam estruturas de grafo multirrelacionais, em que existem tipos diferentes de vértices (representando pessoas, lugares, itens) e diferentes tipos de arestas (como, por exemplo, amigo de, mora em, comprado por).

Usando a técnica Index freeadjacency, cada nó possui o endereço físico na memória RAM dos nodes adjacentes. Assim, os nodes apontam para objetos relacionados.

O exemplo mais comum é uma rede social, na qual pessoas, representadas por vértices, conhecem ou seguem outras pessoas, representadas por arestas do tipo Conhece ou Segue.

Algumas soluções utilizadas atualmente no mercado são: Neo4J, AllegroGraph, InfinitGraph.

MapReduce

MapReduce é um modelo projetado para usar computação paralela distribuída em Big Data e transformar os dados em pedaços menores, funcionando basicamente de duas formas: **mapeamento** e **redução**. O processo de redução dos dados acontece da seguinte forma:

1

Map

No processo de mapeamento (Map), os dados são separados em pares (chave-valor), transformados e filtrados.

2

Shuffle e sort

Entre o processo Map e o processo Reduce, existe uma fase chamada de shuffle e sort. Ela é responsável por agregar a lista de todos os nós e ordená-las, criando uma nova lista com elementos do tipo chave-valor. Esse resultado se torna a entrada do processo de redução (reduce).

3

Reduce

No processo de redução (Reduce), os dados são agregados em datasets menores.

Em resumo, no tratamento de Big Data é possível usar processos de mapeamento e redução para classificar os dados em pares key-value e reduzi-los em pares menores por meio de operações de agregação, combinando múltiplos valores de um dataset em um único valor.

Para entendermos melhor, vamos ao exemplo. Precisamos contar o número de vezes que as palavras ocorrem em determinados documentos da Internet. Para tanto, precisamos seguir os passos:

Passo 1

O primeiro passo seria tokenizar os documentos (dividir em palavras) e mapear cada palavra. O mapeador gerará uma relação com palavra e valor de 1.

Passo 2

A fase de agrupamento levantará todas as palavras e fará uma lista de 1s.

Passo 3

A fase de redução leva uma palavra e uma lista (uma lista de 1 para cada vez que a chave apareceu na Internet) e soma a lista. O redutor produz a palavra, junto com a contagem.
Ao final, teremos uma lista de todas as palavras dos documentos da Internet, juntamente com quantas vezes ela apareceu.

Hadoop

O Hadoop foi a primeira implementação open source para a solução MapReduce, desenvolvida pela Apache Software Foundation com o objetivo de processar grandes volumes de dados. O Hadoop utiliza clusters para armazenar os dados com elevada capacidade de computação quando combinados em distribuição paralela. Esse tipo de solução reduz dramaticamente os custos envolvidos no armazenamento de Big Data.

O Hadoop é uma combinação de dois projetos:

Hadoop.

Hadoop MapReduce (HMR) - Framework de processamento distribuído

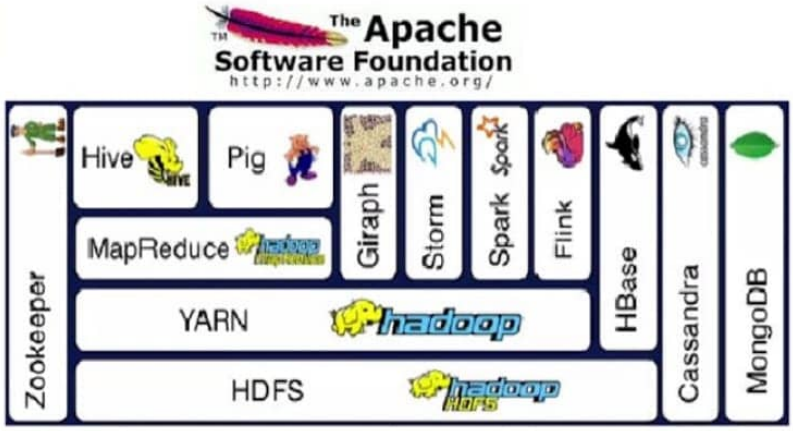
É um spin-off do MapReduce, software que o Google usa para acelerar as pesquisas endereçadas ao seu buscador.

Hadoop Distributed File System (HDFS)

É um sistema de arquivos distribuídos, otimizado para atuar em dados não estruturados, também baseado na tecnologia do Google (neste caso, o Google File System).

Em resumo, para que o HMR processe os dados, eles devem estar armazenados no HDFS. Embora os projetos do Hadoop mais conhecidos sejam o MapReduce (HMR) e seu sistema de arquivos distribuídos (HDFS), outros projetos também oferecem uma variedade de serviços complementares, adicionando abstrações de maior nível, facilitando o desenvolvimento.

Dentre os projetos, conforme a imagem a seguir, vale destacar:



Ecossistema Hadoop.

A seguir, temos a descrição dos elementos da imagem:

1

Yarn

Plataforma de gerenciamento de recursos e agendamento de serviços.

2

Hadoop Common

Repositório de bibliotecas e utilitários usados e compartilhados por outros módulos do Hadoop.

3

Avro

Sistema de serialização de dados de código aberto.

4

Pig

Plataforma paralelizada para trabalhar com grandes conjuntos de dados.

5

HBase, Cassandra e MongoDB

Base de dados distribuída NoSQL.

6

ZooKeeper

Serviço centralizado para coordenação de aplicações distribuídas.

7

Hive

Espécie de Data Warehouse distribuído, facilita a utilização de grandes conjuntos de dados (datasets) em ambientes de armazenamento paralelo.

8 Storm, Spark e Flink
Ferramentas para processamento em tempo real e em memória.

MapReduce: o que é e sua relação com Big Data

No vídeo a seguir, abordaremos a história, a definição e os detalhes do modelo de programação MapReduce e sua relação com o Big Data.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Vem que eu te explico!

Os vídeos a seguir abordam os assuntos mais relevantes do conteúdo que você acabou de estudar.

Arquitetura para Big Data

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Infraestrutura para armazenamento de Big Data

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Infraestrutura de computação e rede para Big Data

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

Big Data surgiu a partir da necessidade de manipular um grande volume de dados e, com isso, novos conceitos foram introduzidos, como o Data Lake, que

A

pode ser considerado um repositório de dados relacionados, sendo, portanto, um armazém de dados orientado por assunto.

B

pode ser considerado um conjunto de bancos de dados relacionais e com relacionamentos entre tabelas de diferentes esquemas de bancos de dados.

C

é o resultado de sucessivas operações de mineração de dados, sendo um ambiente no qual é possível ter relatórios e dashboards de maneira amigável para os analistas de negócio.

D

é projetado para armazenar dados de diversas fontes e formatos, não havendo a necessidade da definição de um esquema de dados para inserir novos itens.

E

é projetado para armazenar small data de diversas fontes e formatos, sendo, portanto, um lago de dados orientados por objeto.

A alternativa D está correta.

Data Lake é um repositório de armazenamento e processamento de Big Data que fornece armazenamento massivo para qualquer tipo de dados, enorme poder de processamento e capacidade de lidar com tarefas simultâneas, praticamente ilimitadas, além de possibilitar a criação de correlações e obtenção de insights para apoiar a tomada de decisão mais eficiente.

Essa arquitetura possibilita manter um grande repositório de dados "brutos", preservando o princípio de imutabilidade. Além disso, os cientistas de dados podem acessar e analisar dados com mais rapidez e precisão, e os analistas podem acessá-los para uma variedade de casos de uso, como análise de sentimento ou detecção de fraudes.

Questão 2

Sobre os bancos de dados NoSQL, assinale a afirmativa correta.

A

Bancos de dados NoSQL não podem ser indexados.

B

Bancos de dados NoSQL são considerados bancos de dados relacionais.

C

Nos bancos de dados NoSQL deve ser definido um esquema de dados fixo antes de qualquer operação.

D

São exemplos de bancos de dados NoSQL: MongoDB, Firebird, DynamoDB, SQLite, Microsoft Access e Azure Table Storage.

E

Os bancos de dados NoSQL usam diversos modelos para acessar e gerenciar dados, como documento, grafo, chave-valor e família de colunas.

A alternativa E está correta.

Bancos de dados NoSQL são criados para modelos de dados específicos e têm esquemas flexíveis para a criação de aplicativos modernos. São amplamente reconhecidos por sua facilidade de desenvolvimento, funcionalidade e performance em escala, e usam vários modelos de dados, incluindo documento e grafo.

Considerações iniciais

O Big Data chegou para ficar. O processamento em alta velocidade de um grande volume de dados oriundos de fontes estruturadas ou não vem sendo aplicado para gerar resultados de alto valor agregado para as empresas. O conceito de Big Data está associado às tarefas de **coletar, transformar, compartilhar, analisar** e **visualizar** dados que não são facilmente discerníveis em meio à grande quantidade de dados a que estamos expostos.

Em um mundo cada vez mais conectado, destacam-se as empresas que conseguem captar os dados e ler as informações de seus consumidores com maior agilidade e precisão. Nas redes sociais há um enorme e variado volume de dados sobre consumidores que estão disponíveis para análises das empresas dispostas a buscar oportunidades de negócios ou expansão de seu mercado.

> Mas como analisar bilhões de postagens feitas todos os dias por milhões e milhões de pessoas no Brasil, ou por bilhões de pessoas no mundo? Como estratificar esse grande universo de dados?

Esse é o tipo de desafio que o Big Data se propõe a solucionar, e que iremos analisar neste módulo, como uma estratégia para as empresas.

Vantagens do Big Data

Vamos conhecer as vantagens na utilização do Big Data.

Tomadas de decisão melhores

Dentre as maneiras mais eficazes na tomada de decisão, o Big Data se destaca. O motivo está na quantidade de dados confiáveis que ele é capaz de analisar e assim oferecer um perfil melhor do cliente, identificando as tendências do mercado, garantindo a retenção de um cliente atual e a conquista de um cliente novo.

Melhorar as estratégias de marketing

Quando se trabalha com marketing, as ações direcionadas para um público específico costumam ter muito mais sucesso do que aquelas sem nenhum direcionamento. Sendo assim, precisamos conhecer bem o comportamento do cliente.

Quando a empresa usa o Big Data, consegue analisar os dados de seu público-alvo e obter acesso a informações importantes, como comportamento, perfil de compra, rendimento financeiro e escolaridade. Dessa forma, é possível traçar estratégias de marketing bem estruturadas e decisivas.

Estar à frente da concorrência

Como já vimos, ao usar o Big Data, é possível identificar futuras tendências no mercado, fundamental para dominar a concorrência e se manter na frente como uma empresa pioneira, inovadora e capaz de conquistar cada vez mais clientes.

Por que implantar uma Estratégia de Big Data na empresa?

A popularidade do Big Data está relacionada a uma resposta de muitas empresas para períodos difíceis. Sendo assim, vamos conhecer alguns dos motivos para implantá-lo como estratégia durante uma crise:

1

Tomada de decisão

O auxílio na tomada de decisão é uma das grandes vantagens em utilizá-lo. Com tantos dados coletados e analisados será quase impossível tomar uma decisão errada. Ao utilizar o Big Data, as decisões serão tomadas com mais segurança e com a certeza de resultados positivos.

2

Obtenção de mais informações

A prática de análise de dados, como uma jornada do consumidor dentro do site, não é uma novidade; o diferencial do Big Data está em adquirir conhecimento profundo sobre a audiência, os concorrentes e o mercado.

3

Marketing digital

Ao unir Marketing digital (marketing realizado em ambiente digital, usando ferramentas como redes sociais, e-mail e sites) com Big Data, as milhares de informação disponíveis tornarão possível uma conexão individualizada com os internautas, permitindo conhecê-los melhor. O resultado dessa união está na viabilidade de projetos, como uma personalização de sites e ofertas. O marketing realizado em ambiente digital, usando ferramentas como redes sociais, e-mail e sites.

4

Reduzir custos

Reduzir custos

5

Reduzir custos

Ao fazer parte de um ambiente competitivo, ações como antecipar, entender e influenciar tendências são diferenciais somente possíveis para quem já tem uma cultura de análise de dados.

Em resumo, o que queremos com o Big Data pode ser visto a seguir:

Conhecer os hábitos e desejos do seu público-alvo

Encontrar os clientes certos, por meio de bancos de listas atualizadas de contato (mailing)

Reduzir tempo com prospecção de clientes

Desenvolver produtos e ofertas com assertividade

Tomar decisões mais inteligentes sobre seu negócio

Conhecer a causa raiz de falhas operacionais

Gerar preços e ofertas com base em hábitos reais de compra dos clientes

Detectar comportamentos fraudulentos

Traçar estratégias assertivas de marketing digital com base em dados

Boas práticas

A primeira a se destacar é o desenvolvimento de uma estratégia de Big Data, que nada mais é do que compreender profundamente as metas do negócio e os dados disponíveis para uso. Outra boa prática é analisar ou avaliar as necessidades de dados adicionais com o foco em cumprir os objetivos. Sendo assim, vamos conhecer as boas práticas no uso da estratégia de Big Data:

- Compreender as metas do negócio e seus dados de uso;
- Analisar e avaliar as necessidades de dados adicionais;
- Priorizar casos de uso planejados e aplicáveis;
- Identificar novos sistemas e ferramentas necessários;
- Criar um roteiro de implantação;
- Avaliar as habilidades internas para identificar a necessidade de retreinamento ou a contratação de especialistas;
- Implantar uma governança de dados e ter processos de gerenciamento de qualidade.

Lei Geral de Proteção de Dados (LGPD)

Agora, vamos conhecer a LGPD (ou Lei Geral de Proteção de Dados), Lei Nº 13.709/2018, que regula o tratamento de dados pessoais pelas organizações, sancionada em agosto de 2018 e que entrou em vigor em setembro de 2020. Essa lei busca proteger o usuário do uso indevido e excessivo dos dados, garantindo a segurança de suas informações e exigindo maior transparência das empresas.

Proteção de dados.

Em todo o mundo, temos em torno de 120 países com leis de proteção aos dados em vigor. São inquestionáveis sua importância e relevância, pois os dados sobre clientes e usuários permitem previsões, elaborar padrões de perfis de consumo, e segmentar opiniões. Porém, ela afeta áreas como prospecção e abordagem de clientes, sondagem ou identificação do perfil da demanda e o marketing digital.

> O não cumprimento da LGPD acarreta multas e sanções.

Na LGDP, a coleta e tratamento dos dados só são possíveis se justificados em bases legais. O uso dos dados só é possível nos seguintes casos:

- Mediante consentimento do titular;
- Para cumprimento de obrigação legal ou regulatória;
- Para execução de políticas públicas, como a vacinação;
- Para estudos por órgãos de pesquisa;
- Para a execução de contrato;
- Para exercício de direito em processo judicial, arbitral ou administrativo;
- Para proteção da vida ou da integridade física;
- Para tutela da saúde;
- Para legítimo interesse e para proteção do crédito.

⚠

Atenção

Os dados pessoais têm grande valor para as empresas, mas para manipulá-los é preciso que estejam suportados por uma base legal, com um objetivo que seja claro para o titular dos dados.

Passos para implantar uma estratégia de Big Data nas empresas

No meio da transformação digital que estamos presenciando, um aspecto que aflora nas organizações é a procura em acertar na identificação, captura, gerenciamento e análise de Big Data. Empresas dos mais variados setores estão cada vez mais interessadas em Big Data e insights estratégicos na condução da tomada de decisão.

O Big Data vem mudando a maneira como as organizações desenvolvem suas estratégias de mercado, avaliam como o público recebe seus produtos e serviços e identificam as principais tendências do seu setor.

> Mas como implementar o Big Data?

Elaboramos alguns passos importantes a serem seguidos na implementação de Big Data nas empresas. Vamos conhecê-los:

1º passo: identificar os desafios da empresa

Para se usar o Big Data, é necessário ter um desafio concreto a ser superado, ou um objetivo a ser conquistado. Fazendo uma analogia, sem se ter um desafio ou um objetivo, é como navegar sem uma bússola. Sendo assim, identifique os problemas da empresa em que o Big Data pode ser útil, como, por exemplo, melhorar processos existentes, reduzir custos operacionais, impulsionar a produtividade ou aumentar o seu market share.

Comece respondendo à pergunta: o que não podemos fazer sem o Big Data, e como isso está nos afetando? Use o Big Data para fazer Business Intelligence (ou Inteligência de Negócio), analisar e apresentar os dados por meio de um processo orientado pela tecnologia.

2º passo: priorizar os problemas de negócios

Nesse passo, já temos os problemas identificados; agora, vamos enumerá-los de acordo com sua prioridade usando uma lista do problema mais importante ao menos importante. Para isso, identificamos três ou quatro maiores desafios para levar ao Big Data.

3º passo: utilizar fontes de dados relevantes

Agora chegou o momento de descobrir os tipos de dados capazes de resolver os problemas levantados no passo 1. Sendo assim, vamos recorrer a fontes relevantes e confiáveis, como relatórios, pesquisas de mercado, tráfego em websites, engajamento nas redes sociais, pesquisas em sites de buscas, dente outras.

4º passo: recorrer a dados internos e externos

O conceito de Big Data está relacionado com grandes volumes de dados; sendo assim, recorra a dados além dos registros internos da empresa. Uma das formas é a utilização de ferramentas de marketing digital que monitoram redes sociais para saber o que as pessoas estão falando, bem como quem são essas pessoas.

5º passo: escolher a ferramenta

A escolha da ferramenta é um passo essencial. Dê preferência a ferramentas inteligentes para aplicar em processos extremamente importantes para a coleta, processamento, análise e armazenamento dos dados. Critérios como segurança, agilidade e precisão devem ser considerados ao escolher a ferramenta.

6º passo: saber quais dados usar e quais dados excluir

Em um grande conjunto de dados, saber diferenciar bons dados de dados inúteis é uma prática importante.

Que tipo de dados ajudará a identificar padrões de consumo ou preferências dos clientes?

A resposta da pergunta é uma boa maneira de filtrar os dados que realmente importam.

Estratégia de Big Data na empresa

No vídeo a seguir, abordaremos os passos a serem seguidos para implantar uma estratégia de Big Data.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Vem que eu te explico!

Os vídeos a seguir abordam os assuntos mais relevantes do conteúdo que você acabou de estudar.

Por que implantar uma estratégia de Big Data na empresa?

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Lei Geral de Proteção de Dados (LGPD)

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

A Lei Geral de Proteção de Dados Pessoais (Lei n.º 13.709/2018) se aplica às operações de tratamento de dados pessoais realizadas

A

no tratamento dos dados de clientes pelas empresas, exclusivamente.

B

na coleta de dados para fins exclusivamente jurídicos ou jornalísticos.

C

na extração de conhecimentos para fins exclusivos de segurança pública ou segurança do Estado.

D

na coleta e tratamento dos dados, desde que com o consentimento do titular.

E

para fins exclusivos de atividades de investigação e repressão de infrações penais.

A alternativa D está correta.

Na LGDP, a coleta e o tratamento dos dados só são possíveis se justificados em bases legais.

Questão 2

Sobre o Big Data e a implantação de uma estratégia de Big Data, julgue os seguintes itens e assinale a afirmativa correta.

I) Os fatores críticos de sucesso da análise de Big Data incluem uma sólida infraestrutura de dados, além de ferramentas analíticas e pessoal habilitado para lidar com elas.

II) A Big Data pode ser utilizada na EAD (Educação a Distância) para se entender as preferências e necessidades de aprendizagem dos alunos e, assim, contribuir para soluções mais eficientes de educação mediada por tecnologia.

III) O Big Data consiste em um grande depósito de dados estruturados, ao passo que os dados não estruturados são considerados importantes e agregam valores para as empresas.

A

Apenas a afirmativa I está correta.

B

Apenas a afirmativa II está correta.

C

Apenas a afirmativa III está correta.

D

As afirmativas I e III estão corretas.

E

As afirmativas I e II estão corretas.

A alternativa E está correta.

Dentre os fatores críticos de sucesso ao implantar uma Estratégia de Big Data na empresa, podemos citar: uma sólida infraestrutura (hardware), ferramentas (software) para facilitar a análise de grandes volumes de dados e uma equipe de profissionais habilitados.

Se a coleta de dados, no Big Data, for bem direcionada ao requerido nicho específico, a possibilidade de precisão é muito maior. Porém, o fator humano continua sendo o ponto crucial da análise dos números, visto que nela predomina a subjetividade.

O Big Data pode ser aplicado em diversas áreas: saúde, agricultura, gestão pública, transporte, educação etc. Na educação, é possível monitorar as ações dos alunos, o tempo que levam para responder a uma pergunta, quais fontes eles usam e quais perguntas deixam de responder.

Ao usar analytics, instituições de ensino podem criar programas personalizados, e o feedback é imediato para os professores.

Considerações finais

Como vimos, o Big Data é uma tecnologia inovadora com uma grande importância na agregação de valor na tomada de decisão pelas empresas.

Compreender seu conceito, suas características e sua importância é fundamental para o profissional de Tecnologia da Informação (TI) assessorar os gestores na tomada de decisão.

Conhecer a infraestrutura de Big Data é necessário para uma eficiente implementação. Por isso, abordamos metodologias e ferramentas usadas para manipular os grandes volumes de dados.

Um bom planejamento e uma boa estratégia na utilização dessa tecnologia se tornam diferenciais para a empresa, e uma forma de se manter na liderança no caso de uma concorrência.

Podcast

Ouça o podcast. Nele, falaremos sobre a importância da tecnologia de Big Data atualmente entre as organizações.

Conteúdo interativo

Acesse a versão digital para ouvir o áudio.

Explore +

Pesquise sobre a ferramenta **hadoop** em seu site e veja como ela aborda o processo de análise de Big Data.

Saiba mais sobre a **Lei Geral de Proteção aos Dados** no site da LGPD e avalie como o Big Data pode ser manipulado sem afetá-la.

Referências

NETO, J. A. R. **Big Data para executivos e profissionais de mercado.** 2. ed. 2019. *E-book*.

NIST. **Big Data interoperability framework:** volume 2. Big Data Taxonomies, v. 3, 2019. *E-book*.