UNIVERSIDADE DE BRASÍLIA
Faculdade do Gama

Sistemas de Banco de Dados 2

Trabalho Final (TF)

Big Data

Lucas Lopes Rocha – 202023903

Brasília, DF
2023

1. Big Data

a. Definição

O Big Data é um termo amplamente utilizado para descrever conjuntos de dados volumosos, complexos e de alta velocidade que ultrapassam a capacidade de processamento e armazenamento dos sistemas tradicionais de gerenciamento de dados (MAYER-SCHÖNBERGER; CUKIER, 2013). Esses dados são caracterizados por três principais aspectos: volume, variedade e velocidade.

O **volume** refere-se à imensa quantidade de dados que são gerados diariamente por diversas fontes, como redes sociais, sensores, transações comerciais, registros de dispositivos móveis, entre outros (MAYER-SCHÖNBERGER; CUKIER, 2013). Esses dados podem atingir proporções exorbitantes, exigindo soluções de armazenamento e processamento escaláveis.

A **variedade** diz respeito à diversidade de formatos e tipos de dados que são capturados e armazenados (MAYER-SCHÖNBERGER; CUKIER, 2013). Além dos dados estruturados tradicionais, como tabelas e bancos de dados relacionais, o Big Data abrange dados não estruturados, como texto, imagens, vídeos, áudio, feeds de mídia social e informações geoespaciais. A capacidade de lidar com essa diversidade de formatos é um dos desafios fundamentais do Big Data.

A **velocidade** refere-se à velocidade em que os dados são gerados, transmitidos e processados em tempo real (MAYER-SCHÖNBERGER; CUKIER, 2013). Em muitos casos, as organizações precisam analisar e tomar decisões com base em dados em tempo quase real, exigindo sistemas capazes de lidar com altas taxas de fluxo de dados e fornecer resultados em tempo hábil.

O Big Data tem sido aplicado em diversos contextos e setores. Na área de saúde, por exemplo, grandes volumes de dados podem ser analisados para identificar padrões e tendências em doenças, ajudando na prevenção e tratamento de doenças (ZIKOPOULOS et al., 2013). No setor financeiro, a análise de dados em tempo real pode fornecer insights valiosos para a detecção

de fraudes e previsão de mercado (ZIKOPOULOS et al., 2013). No campo da ciência, a análise de grandes volumes de dados pode ajudar a acelerar descobertas e avanços em várias áreas (LANEY, 2001).

Os fundamentos do Big Data envolvem técnicas e tecnologias específicas para lidar com os desafios associados aos dados em grande escala. Isso inclui tecnologias de armazenamento distribuído, como o Hadoop e sistemas de bancos de dados NoSQL, que são projetados para escalabilidade e tolerância a falhas (ZIKOPOULOS et al., 2013). Além disso, técnicas de processamento de dados em tempo real, como processamento de fluxo e análise de streaming, são essenciais para lidar com a velocidade dos dados (LANEY, 2001).

Os principais objetivos do Big Data podem ser resumidos em três aspectos principais: obter insights significativos, melhorar a eficiência e desempenho, e possibilitar a personalização e segmentação (MAYER-SCHÖNBERGER; CUKIER, 2013).

Em suma, o Big Data é uma área em constante crescimento que lida com a análise de grandes volumes de dados com alto volume, variedade e velocidade. A compreensão dos conceitos e fundamentos do Big Data, assim como seus contextos de utilização e objetivos, é fundamental para aproveitar todo o potencial desses conjuntos de dados em diferentes setores e contextos.

b. Objetivos principais

O Big Data, como campo em constante crescimento, apresenta uma ampla gama de objetivos que impulsionam sua adoção e aplicação em diferentes setores. Três principais objetivos do Big Data podem ser identificados: obter insights significativos, melhorar a eficiência e desempenho, e possibilitar a personalização e segmentação.

O primeiro objetivo do Big Data é obter insights significativos a partir da análise de grandes volumes de dados (MAYER-SCHÖNBERGER; CUKIER, 2013). Esses dados volumosos e diversos podem fornecer informações valiosas que antes eram difíceis de obter, permitindo a identificação de padrões,

correlações e tendências ocultas. Ao explorar e compreender esses insights, as organizações podem tomar decisões estratégicas embasadas em dados concretos, identificar oportunidades de negócios e antecipar mudanças no mercado.

O segundo objetivo do Big Data é melhorar a eficiência e o desempenho das organizações (MAYER-SCHÖNBERGER; CUKIER, 2013). A análise de grandes volumes de dados permite a identificação de ineficiências operacionais, gargalos e áreas de melhoria nos processos internos. Ao eliminar atividades desnecessárias, otimizar fluxos de trabalho e alocar recursos de forma mais eficiente, as organizações podem alcançar uma maior produtividade e reduzir custos operacionais.

Por fim, o Big Data busca possibilitar a personalização e segmentação de forma mais precisa (MAYER-SCHÖNBERGER; CUKIER, 2013). Ao coletar e analisar dados individuais de clientes, é possível oferecer experiências personalizadas, recomendações direcionadas e campanhas de marketing segmentadas. Essa abordagem baseada em dados permite uma maior compreensão dos clientes, suas preferências e necessidades, resultando em um maior engajamento e satisfação do público-alvo.

Os objetivos do Big Data são alcançados por meio do uso de tecnologias e técnicas específicas, como armazenamento distribuído, processamento em tempo real e análise avançada de dados (ZIKOPOULOS et al., 2013). Essas tecnologias permitem lidar com os desafios de volume, variedade e velocidade dos dados, garantindo a escalabilidade, a eficiência e a precisão necessárias para a obtenção dos objetivos propostos.

c. Vantagens

O Big Data oferece diversas vantagens que impulsionam sua adoção e aplicação em diferentes setores. A seguir, serão detalhadas as principais vantagens do Big Data, juntamente com um exemplo real que ilustra cada uma delas.

- **Tomada de decisões baseadas em dados:** Uma das principais vantagens do Big Data é a capacidade de fornecer insights e informações embasados em dados concretos, auxiliando na tomada de decisões estratégicas (MAYER-SCHÖNBERGER; CUKIER, 2013). Por exemplo, uma empresa de comércio eletrônico pode utilizar técnicas de análise de Big Data para examinar o comportamento de compra dos clientes, identificar padrões de consumo e antecipar tendências de mercado. Com base nesses insights, a empresa pode ajustar sua estratégia de marketing, adaptar seu catálogo de produtos e tomar decisões informadas para impulsionar as vendas.
- **Identificação de padrões e previsão de eventos:** O Big Data permite a identificação de padrões ocultos e a previsão de eventos futuros com base em análises avançadas de dados históricos (ZIKOPOULOS et al., 2013). Por exemplo, uma companhia de seguros pode analisar grandes volumes de dados relacionados a sinistros, informações do cliente e fatores externos para identificar padrões de comportamento que levam a reclamações fraudulentas. Essa análise preditiva ajuda a empresa a implementar medidas proativas de detecção e prevenção de fraudes, reduzindo os riscos e os custos associados.
- **Melhoria da eficiência operacional:** O Big Data permite identificar ineficiências operacionais e otimizar processos internos, resultando em maior eficiência e redução de custos (MAYER-SCHÖNBERGER; CUKIER, 2013). Por exemplo, uma empresa de logística pode analisar grandes volumes de dados de rotas de transporte, dados de sensores e informações meteorológicas para otimizar o planejamento de rotas, reduzir o consumo de combustível e minimizar o tempo de entrega. Essa abordagem baseada em dados ajuda a empresa a melhorar sua eficiência operacional, aumentar a satisfação do cliente e reduzir seus impactos ambientais.
- **Personalização e experiências do cliente:** O Big Data permite a personalização de produtos e serviços com base nas preferências e comportamentos individuais dos clientes (MAYER-SCHÖNBERGER; CUKIER, 2013). Por exemplo, empresas de streaming de música utilizam

algoritmos de recomendação baseados em Big Data para analisar o histórico de reprodução, preferências musicais e comportamento dos usuários, fornecendo recomendações personalizadas de novas músicas e artistas. Isso cria uma experiência mais relevante e envolvente para o cliente, aumentando a fidelidade e a satisfação.

d. Desvantagens

Embora o Big Data ofereça diversas vantagens, também apresenta algumas desvantagens que precisam ser consideradas. A seguir, serão abordadas as principais desvantagens do Big Data, juntamente com um exemplo real que ilustra cada uma delas.

- **Privacidade e segurança dos dados:** Uma das principais desvantagens do Big Data está relacionada à privacidade e segurança dos dados (MAYER-SCHÖNBERGER; CUKIER, 2013). À medida que grandes volumes de dados são coletados, armazenados e analisados, aumentam as preocupações com a proteção dos dados pessoais e a possibilidade de violações de segurança. Por exemplo, em 2019, a empresa de cartões de crédito Capital One sofreu um ataque cibernético que expôs dados pessoais de mais de 100 milhões de clientes. Esse incidente ressalta a importância de implementar medidas robustas de segurança e privacidade no contexto do Big Data.
- **Custo e infraestrutura requerida:** Outra desvantagem do Big Data é o custo associado à infraestrutura necessária para armazenar, processar e analisar grandes volumes de dados (ZIKOPOULOS et al., 2013). O investimento em hardware, software e recursos humanos qualificados pode ser significativo. Além disso, os custos de armazenamento de dados em larga escala podem se tornar um desafio. Por exemplo, uma organização que deseja implementar uma solução de Big Data precisa adquirir servidores de alto desempenho, implementar sistemas de armazenamento distribuído e contratar especialistas em análise de dados, o que demanda recursos financeiros consideráveis.

- **Viés e qualidade dos dados:** O Big Data pode estar sujeito a viés e qualidade inferior dos dados coletados (MAYER-SCHÖNBERGER; CUKIER, 2013). É fundamental garantir que os dados sejam representativos e confiáveis para obter insights precisos. Por exemplo, em 2020, um estudo revelou que os algoritmos usados em um sistema de reconhecimento facial apresentavam viés racial, resultando em taxas de erros mais altas para determinados grupos étnicos. Isso ressalta a importância de monitorar e avaliar constantemente a qualidade dos dados utilizados em análises de Big Data, a fim de evitar resultados tendenciosos e discriminatórios.
- **Complexidade e habilidades necessárias:** O Big Data envolve uma grande complexidade, requerendo habilidades especializadas para lidar com a coleta, o processamento e a análise dos dados (ZIKOPOULOS et al., 2013). A falta de profissionais qualificados pode representar um desafio para as organizações. Por exemplo, a demanda por cientistas de dados tem aumentado significativamente nos últimos anos, mas a oferta de profissionais com as habilidades necessárias ainda é limitada. Isso pode dificultar a implementação bem-sucedida de projetos de Big Data e a obtenção dos resultados desejados.

e. Exemplos de uso

i. Caso de sucesso

Uma história de sucesso que ilustra uma das vantagens do Big Data é a experiência da Netflix na personalização e recomendação de conteúdo para seus usuários.

A Netflix é uma plataforma de streaming de vídeo que utiliza o Big Data para entender as preferências de seus usuários e oferecer recomendações personalizadas de filmes e séries. A empresa coleta uma vasta quantidade de dados sobre o comportamento de visualização de seus assinantes, como os títulos assistidos, as avaliações atribuídas, a duração de visualização, entre outros dados demográficos e de perfil.

Por meio da análise desses dados, a Netflix utiliza algoritmos avançados de recomendação baseados em Big Data para identificar padrões de interesse e recomendar conteúdos relevantes para cada usuário individualmente. Isso proporciona uma experiência personalizada, na qual os usuários recebem sugestões de filmes e séries que se alinham com seus gostos e preferências.

Essa personalização baseada em Big Data resulta em um aumento significativo do engajamento dos usuários e na satisfação do público-alvo. Estima-se que cerca de 80% do conteúdo assistido na Netflix seja proveniente das recomendações feitas por meio de sua abordagem de Big Data. A capacidade da empresa em entender os interesses individuais de seus usuários e oferecer uma experiência personalizada é um dos principais fatores que contribuíram para o sucesso e a expansão da Netflix como líder no mercado de streaming de vídeo.

ii. Caso de insucesso

O Google Flu Trends era um projeto que utilizava o Big Data para tentar prever e rastrear surtos de gripe com base nas pesquisas realizadas pelos usuários no mecanismo de busca do Google. O objetivo era fornecer informações em tempo real sobre a propagação da gripe, permitindo uma resposta mais ágil por parte das autoridades de saúde.

No entanto, o Google Flu Trends enfrentou problemas relacionados ao viés e qualidade dos dados. O algoritmo utilizado para interpretar os termos de pesquisa associados à gripe começou a superestimar os casos de gripe, levando a resultados imprecisos.

Em 2013, o Google Flu Trends previu um número significativamente maior de casos de gripe do que o relatado pelos órgãos oficiais de saúde. Esse incidente revelou que a correlação entre os termos de pesquisa e os casos reais de gripe não era tão precisa quanto inicialmente esperado. O modelo utilizado pelo Google Flu Trends não conseguiu lidar adequadamente com a evolução dos padrões de busca e com a variabilidade das palavras-chave relacionadas à gripe.

Esse caso do Google Flu Trends ilustra como a qualidade dos dados utilizados em análises de Big Data pode afetar os resultados e levar a previsões imprecisas. A falta de controle sobre a qualidade e representatividade dos dados pode comprometer a confiabilidade das análises e das conclusões obtidas por meio do Big Data.

2. Referências Bibliográficas

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. Big Data: Como extrair volume, variedade, velocidade e valor da avalanche de informações cotidianas. Rio de Janeiro: Campus, 2013.

LANEY, Doug. 3D Data Management: Controlling Data Volume, Velocity and Variety. In: GARTNER SYMPOSIUM/ITXPO, 2001. Proceedings... [Local da conferência não especificado]: [Editora não especificada], 2001. p. 1-6.

ZIKOPOULOS, Paul et al. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. IBM Journal of Research and Development, Armonk, v. 57, n. 3/4, p. 1-12, 2013. DOI: 10.1147/JRD.2013.2247147.

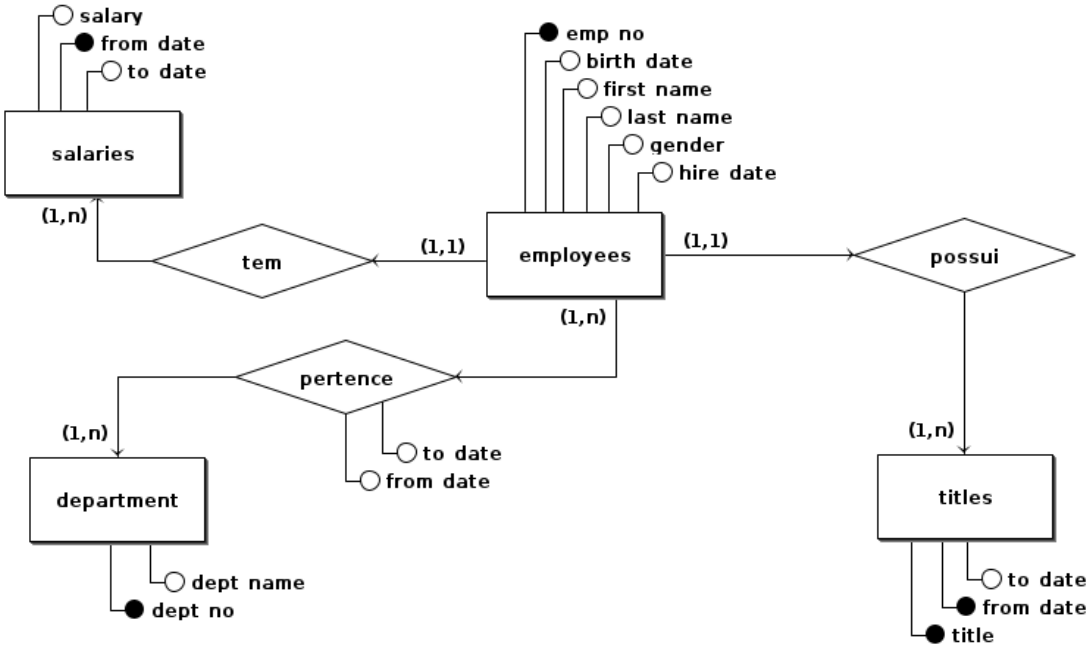
3. Base de Dados (Documentação)

Essa base de dados se trata das informações de funcionários de uma empresa, bem como os departamentos que trabalham, seus salários e títulos. Esta é uma base de dados em MySQL e conta com 6 tabelas e milhões de registros. A seguir alguns links importantes para esclarecimentos sobre a base utilizada no desenvolvimento desse projeto:

Documentação: https://dev.mysql.com/doc/employee/en/

Base de Dados no GitHub: https://github.com/datacharmer/test_db

a. Diagrama Entidade-Relacionamento (DE-R)



b. Diagrama Lógico de Dados (DLD)