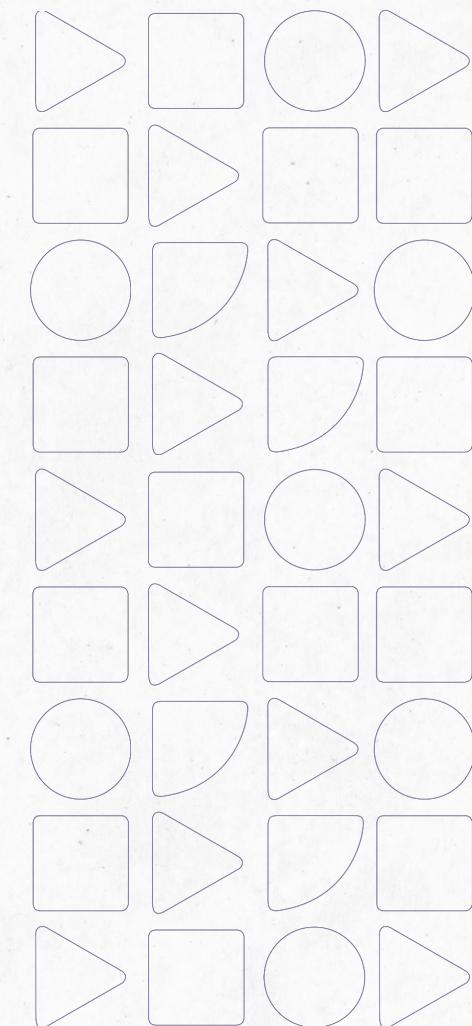


Big data

Disciplina: Sistemas Inteligentes



Conteúdos:

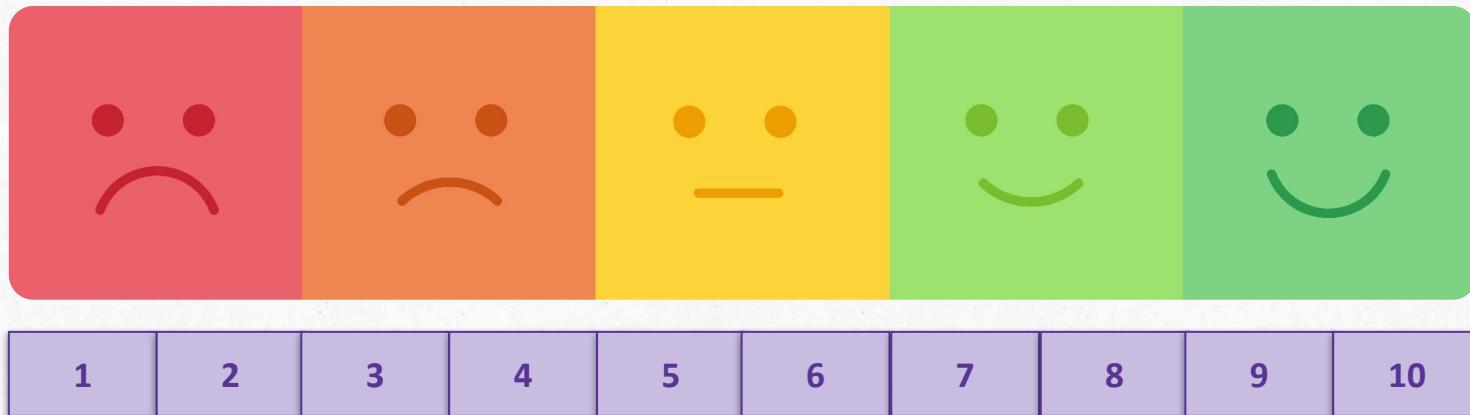
Big data.

Habilidade(s):

Lidar com conjuntos de dados extensos e complexos.

Bloco 1

Animômetro!



Big data

Na última década, a nossa sociedade passou por uma série de revoluções culturais e tecnológicas. Um dos destaques dessa transformação foi o crescimento exponencial do conceito de *big data*.

Trata-se de um termo que surgiu por volta de 2010 e que se refere a grandes volumes de dados de diversas origens e formatos, tornando a sua armazenagem, a sua análise e o seu processamento um desafio.

O conceito de *big data* ainda gera incertezas em relação à sua definição, às suas características, às suas aplicações e aos seus desafios.



Big data

Possível definição

Big data refere-se aos conjuntos de dados com enormes volumes, frequentemente provenientes de fontes variadas.

Os problemas do *big data*

A dificuldade de lidar com esses dados é amplificada quando se trata de dados não estruturados, desafiando os sistemas de banco de dados tradicionais.

Importante diferenciar

É importante distinguir entre a ciência de dados, que se concentra em criar modelos e extrair *insights*, e o *big data*, que lida com a infraestrutura necessária para lidar com volumes excepcionais de dados.

Era dos dados e fontes de dados

Era dos dados

Vivemos em uma era de geração contínua de dados por indivíduos e empresas. Os dados agora são considerados ativos valiosos que, quando gerenciados e analisados adequadamente, podem trazer benefícios e lucros significativos para as empresas.

Fonte de dados

Esses dados provêm de várias fontes, incluindo redes sociais, motores de busca na internet, *e-commerce* e muito mais. Eles podem ser classificados como dados estruturados, não estruturados ou semiestruturados, dependendo da sua forma e organização.

Os 5 Vs do *big data*

Volume

Refere-se ao grande volume de dados envolvidos, que podem variar de *megabytes* a *terabytes*. Esses dados provêm de diversas fontes, como redes sociais e motores de busca, e desafiam os sistemas de gerenciamento de banco de dados tradicionais.

Velocidade

A velocidade com que os dados são capturados e disponibilizados para análise é um aspecto crítico do *big data*. O processamento eficiente desses dados é essencial para evitar atrasos e gargalos.

Veracidade

Garantir a veracidade dos dados coletados é um desafio significativo no contexto do *big data*. É necessário verificar as fontes, detectar viés nos dados e considerar as datas de publicação para garantir que as informações sejam confiáveis.

Variedade

A variedade de tipos de dados é outro aspecto do *big data*. Os dados podem ser textuais, numéricos, imagens, *tags* e muito mais. É essencial que sistemas de *big data* possam lidar com essa heterogeneidade.

Valor

Isso se refere a atribuir valores aos dados. Transformar dados em informações relevantes é uma característica importante, ajudando na tomada de decisões e diferenciação no mercado.

***Big data* em diferentes setores**

O uso do *big data* não está restrito a nenhum setor específico e possui aplicabilidade em diversas áreas. Um exemplo de aplicação é na área financeira, em que a análise de grandes volumes de dados pode ajudar em decisões de investimento, previsão de mercado e muito mais.

Vamos praticar?

Façam grupos para realizar a atividade.

Primeiro momento

15 minutos

Pesquisem uma aplicação do *big data* e criem um *folder* sobre ela.

Segundo momento

10 minutos

Mostre para a turma o que vocês produziram.



Bloco 2

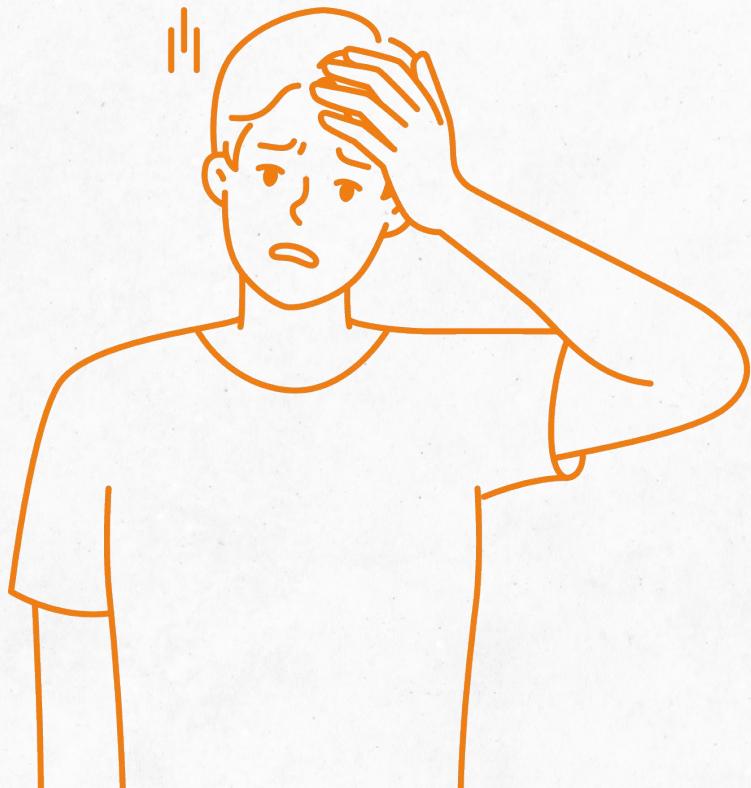
Fofoca do bem

Contem para a turma, em forma de fofoca, um pouco sobre *big data*.



Mito ou verdade?

1. *Big data* significa "muitos" dados;
2. Dados precisam ser "limpos";
3. Esperar para aperfeiçoar seus dados;
4. Possuir um lago de dados;
5. Análise de dados é cara;
6. Algoritmos substituirão analistas humanos.



1. Mito

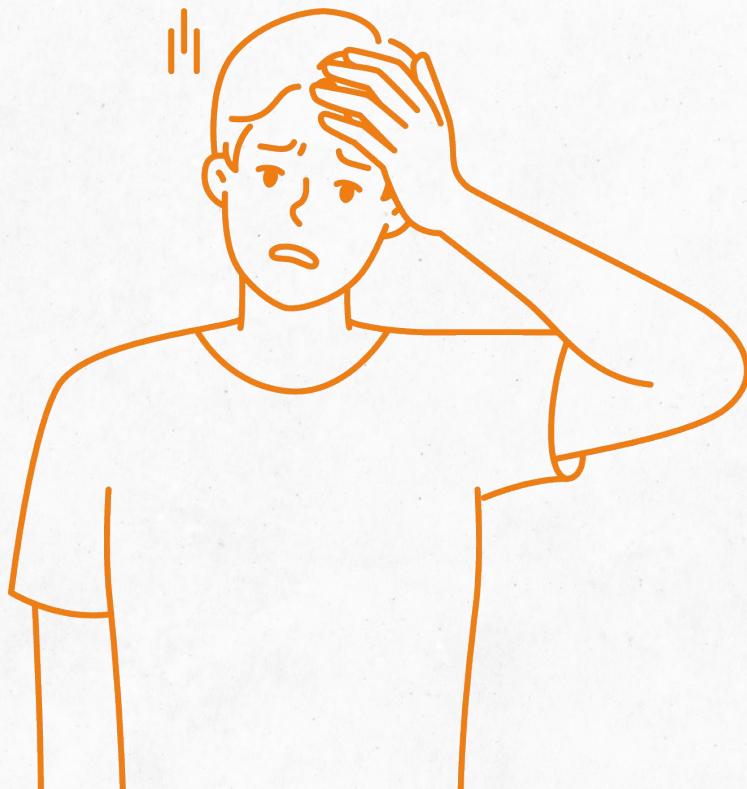
Big data significa "muitos" dados.

O *Big Data* não se limita apenas a grandes volumes de dados. Ele é mais complexo do que parece. Os dados são organizados e analisados para fornecer *insights* valiosos.

2. Mito

Dados precisam ser "limpos".

Às vezes, os dados capturados podem estar incompletos ou incorretos, o que pode levar a decisões equivocadas. É importante identificar e corrigir dados "sujos" para obter uma visão mais clara da situação.



3. Mito

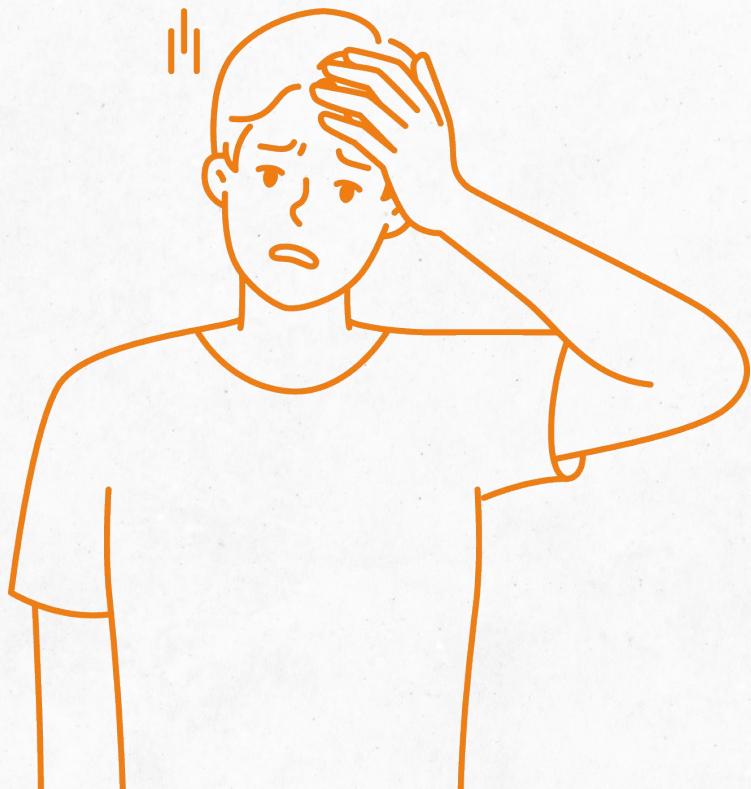
Esperar para aperfeiçoar seus dados.

Após a limpeza inicial dos dados, o processo de refinamento e aperfeiçoamento pode continuar. Refinamentos subsequentes podem levar a análises mais precisas.

4. Mito

Possuir um lago de dados.

O conceito de "*lake data*" não envolve o despejo indiscriminado de dados em um repositório. Os dados são curados e distribuídos em silos para conformidade e governança.



5. Mito

Análise de dados é cara.

A análise de dados não precisa ser um investimento exorbitante, pois existem ferramentas acessíveis disponíveis, além de abordagens econômicas para análise de dados.

6. Mito

Algoritmos substituirão analistas humanos.

Embora os algoritmos sejam eficazes na análise de grandes volumes de dados, eles não substituem completamente os analistas humanos.

Benefícios do *big data*

O *big data* oferece possibilidades incontáveis em várias áreas e setores. Empresas de diferentes indústrias podem se beneficiar do uso estratégico dele para obter *insights* valiosos.

Tomando decisões baseadas em dados

Tomada de decisão informada

A tomada de decisão baseada em dados é fundamental para o sucesso dos negócios. Os dados podem ser coletados, analisados e transformados em informações valiosas.

Coleta de dados abrangente

A coleta de dados abrange várias fontes, incluindo internas, externas, redes sociais, transações financeiras e muito mais. Conhecer os hábitos e as necessidades dos clientes é essencial para oferecer produtos e serviços adequados.

Bloco 3

Tipos de dados

Os tipos de dados são os blocos fundamentais para a organização e a compreensão das informações em qualquer contexto. Os dois principais tipos de dados são os **dados estruturados** e os **não estruturados**.

Dados estruturados

Dados estruturados são informações organizadas em formatos predefinidos e, geralmente, representam números, datas e texto. São **amplamente utilizados em bancos de dados relacionais**, em que a estrutura precisa estar rigidamente definida para estabelecer relações.

Qualitativos

Representam atributos categorizáveis, como gênero, estado civil e raça.

Quantitativos

Envolvem valores numéricos, como idade, altura e peso.

Dados estruturados

Nome	Idade	Altura	Data Nasc.	Sexo	Estado civil
ANA	22	1,72	25/01/1999	F	Solteira
MARCIO	19	1,78	31/07/2002	M	Solteiro
JOÃO	76	1,69	14/06/1945	M	Viúvo
MARIA	43	1,67	06/02/1978	F	Casada

Dados não estruturados

Dados não estruturados são informações desprovidas de uma estrutura predefinida e são normalmente encontrados em mídias visuais, como imagens, fotografias, vídeos e mídias sociais.



Preparação dos dados

A preparação dos dados é fundamental e inclui tarefas como **transformação, engenharia de atributos e verificação de consistência**. Também é crucial identificar atributos redundantes e tratar dados faltantes, que são desafios comuns em problemas reais.



Análise exploratória dos dados

A análise exploratória, por meio de estatísticas descritivas e visualizações, **revela padrões, tendências e valores atípicos nos dados**. O tipo de dado influencia o tipo de análise aplicada, seja qualitativa ou quantitativa.

Medidas de frequência, como contagem e porcentagem, são aplicadas a dados categóricos. Histogramas são úteis para representar a frequência de valores em dados contínuos.

Escolha e avaliação dos modelos de análise

A seleção do método de análise depende da natureza do problema em questão. Modelos preditivos, prescritivos, descritivos e diagnósticos são usados para resolver diferentes tipos de problemas. A eficácia do modelo escolhido está diretamente relacionada ao tipo de problema abordado. A avaliação inclui a **divisão dos dados em conjuntos de treinamento e avaliação para verificar o desempenho do modelo.**

Vamos praticar?

Primeiro momento

15 minutos

Escolha um dado não estruturado e o converta em um dado estruturado.

Segundo momento

10 minutos

Mostre para a turma o que você produziu.



Bloco 4

Data warehouse

A *data warehouse* é um tipo especial de banco de dados, que foi criado com o propósito de fornecer **suporte à tomada de decisões**. O *data warehouse* se tornou uma solução essencial para atender à necessidade de dados limpos e consistentes para análise.

Data warehouse

Motivação para os *data warehouses*

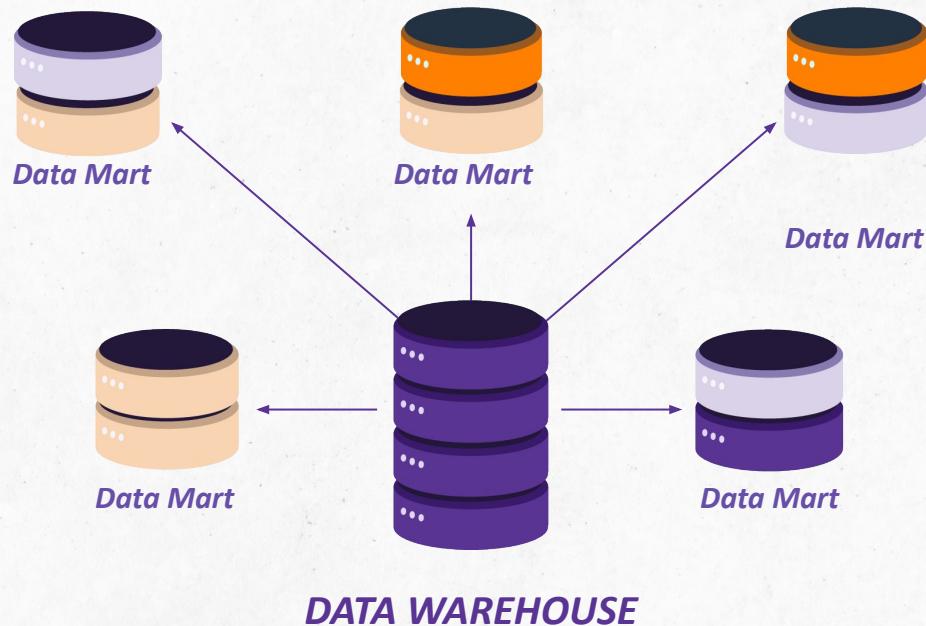
Data warehouses surgiram devido à necessidade de fornecer uma única fonte de dados limpa e consistente para tomada de decisões. Eles foram projetados para não impactar os sistemas operacionais enquanto atendem às demandas de consultas intensivas dos usuários.

Desafios dos *data warehouses*

Data warehouses enfrentam desafios significativos, incluindo consultas intensivas, volumes crescentes de dados e questões de escalabilidade. Erros de projeto, uso ineficaz de operações relacionais e fraquezas na implementação do modelo relacional também podem complicar as coisas.

Data mart

Data marts são subconjuntos especializados de *data warehouses*, destinados a atender às necessidades específicas de análise. Eles podem solucionar a ineficiência de executar repetidamente operações sobre o mesmo conjunto de dados.



Tipos de *data marts*

Data marts podem ser criados de três maneiras principais: extraídos do *data warehouse*, construídos independentemente ou como parte de uma estratégia "*data mart* primeiro".

Extraídos do *data warehouse*

Os dados podem simplesmente ser extraídos do *data warehouse* com efeito, seguindo uma tática de “dividir e conquistar” para a carga de trabalho global de apoio à decisão, a fim de obter melhor desempenho e escalabilidade.

Independente

Essa técnica poderia ser apropriada se o *data warehouse* estivesse inacessível.

Parte do *data mart* primeiro

Data marts são criados conforme a necessidade, com o *data warehouse* global sendo criado finalmente como uma consolidação dos diversos *data marts*.

Granularidade dos dados

A granularidade refere-se ao **nível mais baixo de agregação de dados** mantidos no banco de dados.

Decidir a granularidade certa é crucial para o desempenho e a eficiência do *data mart*.

Data mart e ferramentas analíticas

O projeto físico de um *data mart* frequentemente é influenciado pelas ferramentas analíticas específicas a serem usadas. A escolha cuidadosa do projeto e da granularidade dos dados é essencial para o sucesso do *data mart*.

Discussões e reflexões



Bloco 5

Mineração de dados

A mineração de dados é uma técnica poderosa que nos permite desvendar informações valiosas a partir de conjuntos de dados complexos. Ela combina princípios da estatística, inteligência artificial, máquinas de estados e bancos de dados para construir modelos analíticos e revelar *insights* ocultos.

A chave para o sucesso na mineração de dados reside no processamento de grandes volumes de dados e na capacidade de dimensionamento conforme a necessidade.



Desafios da mineração de dados

- Com frequência, os dados estão incompletos ou contaminados por ruído, o que pode comprometer a identificação de padrões e a confiabilidade das análises;
- É fundamental tomar decisões criteriosas sobre quais algoritmos de mineração aplicar a conjuntos de dados específicos, sintetizar os resultados obtidos, empregar ferramentas de apoio à decisão e iterar no processo para refinar os resultados.

Etapas do processo de mineração de dados

Começando pela **escolha dos algoritmos** apropriados, passando pela **aplicação desses algoritmos** a conjuntos de dados e variáveis específicas, até a **síntese dos resultados** e o **uso de ferramentas de apoio à decisão**. Cada uma dessas etapas desempenha um papel crucial na extração de *insights* valiosos a partir dos dados.

Desde algoritmos de associação e identificação de itens frequentes até técnicas de agrupamento, árvores de decisão, classificação bayesiana e mineração com redes neurais, cada algoritmo possui um propósito específico na análise de dados.

Data lake e data swamp

Data lake

Um *data lake* é um **repositório de dados** que pode armazenar uma grande variedade de tipos e formatos de dados. É projetado para **armazenar dados brutos e não processados na sua forma original**. A sua principal vantagem é a **flexibilidade** para lidar com uma ampla gama de dados.

VS

Data swamp

Refere-se a um ambiente de **armazenamento de dados caótico, desorganizado e sem governança**.

Quando um *data lake* não é gerenciado adequadamente, ele pode se transformar em um *data swamp*, tornando os dados difíceis de encontrar, acessar e entender. Isso **compromete a utilidade dos dados**.

Governança de dados no *data lake*

É fundamental a governança de dados em um ambiente de *data lake*. A governança de dados desempenha um papel vital na organização e no gerenciamento das informações, tornando-as mais acessíveis e valiosas, ao mesmo tempo em que evita que o *data lake* se torne um *data swamp*.

Discussões e reflexões



Bloco 6

A evolução do *data warehousing* e análise de dados

- A área de *data warehousing* e de análise de dados está em constante crescimento, impulsionada pela demanda crescente por *insights* valiosos;
- Muitas empresas estão adaptando os seus produtos e serviços para atender às necessidades em constante mudança dos clientes e isso está moldando o campo;
- O conceito de *data warehousing* e análise de dados, que já foi considerado novo, tornou-se uma das ferramentas mais cruciais para as grandes empresas em todo o mundo;
- As empresas agora implementam processos avançados de transformação e unificação de dados por meio de ETL (*Extract, Transform & Load*) e ELT (*Extract, Load & Transform*) para aproveitar ao máximo seus dados.

Preparando os dados para análise

A preparação dos dados envolve etapas críticas, como a seleção cuidadosa dos dados a serem analisados, que podem ser originários de diversas fontes. É necessário realizar a limpeza dos dados para remover inconsistências e preencher lacunas.

A adequação do formato dos dados e a criação de novos atributos são passos essenciais que dependem da tarefa e da técnica de mineração a serem utilizadas.

ETL

O ETL (Extração, Transformação e Carga) é um processo meticulosamente planejado que desempenha um papel fundamental na jornada dos dados.

Começa com a extração de dados de várias fontes, incluindo arquivos de origem e bancos de dados OLTP (*On-line Transactional Processing*).

Um banco de dados OLTP é uma parte vital desse processo, em que transações em tempo real são registradas em sistemas como Oracle, SQL Server e IBM DB2.

A camada de processamento é diversificada, pois os requisitos variam conforme a necessidade de extrair diferentes tipos de *insights* dos dados.

ELT

ELT

Embora o ELT (Extração, Carga e Transformação) seja frequentemente associado aos processos de *data warehouse*, ele compartilha alguns conceitos essenciais com o ETL.

Operações

As operações típicas no processo ELT envolvem a modificação dos dados recebidos, a formatação, categorização, filtragem e validação para garantir a conformidade com os requisitos.

Staging Area

- A *Staging Area*, ou Área de Preparação dos Dados, desempenha um papel crucial na fase de pré-processamento dos dados antes da modelagem;
- É um **ambiente temporário** em que os dados das fontes originais são copiados e submetidos a processos de transformação e limpeza;
- Isso **evita a necessidade de acessar diretamente as fontes originais** e permite que os dados sejam processados de forma eficiente.

Desafios

O processamento repetitivo e iterativo é uma característica fundamental no cenário dos dados em grande escala. Os desafios não se limitam apenas ao tamanho dos dados, mas também incluem a velocidade, o processamento e as características únicas que os dados grandes apresentam.

Vamos praticar?

Façam grupos para realizar a atividade.

Primeiro momento

10 minutos

Pesquisem uma aplicação do ETL ou ELT e leiam um pouco sobre.

Segundo momento

15 minutos

Falem para a turma o que você encontrou.



Fechamento

Diga uma palavra sobre o que você achou da aula.



Referências Bibliográficas

PROZ EDUCAÇÃO. *Apostila de Sistemas Inteligentes*. 2023.