Aula 00 (Profs. Felipe Mathias e Emannuelle Gouveia)

Correios (Analista de Correios - Analista de Sistema - Desenho de Sistemas) Ciência de Dados - 2024 (Pós-Edital)

Autor:
Emannuelle Gouveia Rolim, Felipe Mathias

12 de Outubro de 2024

Índice

APRESENTAÇÃO DA AULA

Olá, alunos! Bem-vindos a mais uma aula do curso de Tecnologia de Informação para concursos públicos, no Estratégia Concursos.

Me chamo Felipe Mathias e serei seu professor na aula de hoje. Sou um catarinense de 30 anos, programador *front end* (ex-programador, se preferirem haha) e atuo como professor de cursos de Tecnologia da Informação voltados a concursos há mais de um ano. Assim como você, também vivo a vida de concurseiro, aguardando minha nomeação como Auditor Fiscal da Secretaria de Fazenda de Minas Gerais (SEF-MG), onde figuro no cadastro de reserva. Atualmente, continuo, em paralelo, estudando para concursos aguardando o meu grande sonho – o cargo de Auditor Fiscal da SEF-SC, com especialidade em TI.

Minha aventura no mundo do ensino surgiu de uma vontade interna de atuar como professor – sempre amei explicar as coisas, além de ter certa facilidade em expressar conceitos mais complexos para pessoas que talvez não tenham tanta experiência na área.

Meu objetivo aqui é digerir assuntos, desde os mais simples aos mais complexos, para que qualquer aluno consiga os entender, seja um programador, operador de infraestrutura, ou simplesmente um leigo que resolveu adentrar no mundo dos concursos e se deparou com TI no seu edital.

Gostaria de pedir que **sempre** vejam as questões comentadas durante a aula. Elas trazem conteúdo essencial para o aprendizado, muitas vezes abordando alguns pontos que não foram abordados no conteúdo e são essenciais para a resolução de questões.

Caso tenha alguma dúvida, não tenha receio de entrar em contato comigo nas minhas redes sociais (especialmente no meu Instagram, que deixarei abaixo), ou no fórum de dúvidas que os responderei assim que possível.

Ah, posto bastante coisa interessante de TI direcionada para concursos lá, dá uma olhadinha que algumas coisas podem te interessar. Volta e meio acerto alguma questão de prova por lá ;)

@fe.fiscal

t.me/fefiscal

ESTRATÉGIA FLASHCARDS

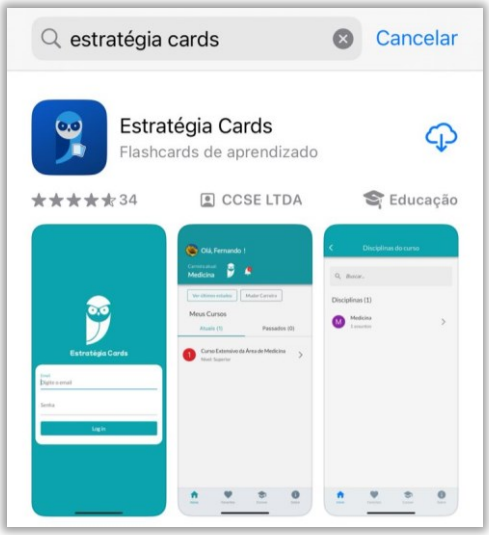
📚 Você tem dificuldade de estudar, memorizar e revisar os conteúdos que estuda em nossas aulas? Então nós temos a ferramenta perfeita para você!

Apresentamos o **Estratégia Cards**: app de flashcards que vai revolucionar sua forma de **estudar** e **revisar** conteúdos de provas de concurso público. Com nossa tecnologia inovadora e interface amigável, você dominará os tópicos mais complexos de maneira eficiente e divertida.

✨ Recursos do Estratégia Cards:

Curadoria de Flashcards	Flashcards criados e revisados por professores especializados em cada área, com qualidade e voltados para concursos públicos.
Flashcards Personalizados	Crie seus próprios flashcards, cobrindo os principais tópicos e matérias dos concursos públicos.
Repetição Espaçada	Técnica de aprendizagem que envolve revisar informações em intervalos crescentes para melhorar a retenção de longo prazo e combater o esquecimento.
Estatísticas Personalizadas	Visualize graficamente o percentual de acertos, erros ou dúvidas dos decks estudados.
Modo Offline	Estude em qualquer lugar, mesmo sem conexão à internet, fazendo o download dos decks.
Estudo por Áudio	*Está dirigindo ou fazendo esteira e quer continuar estudando?* Basta utilizar a opção "Escutar".
Decks Favoritos	Você pode escolher decks específicos como favoritos e visualizá-los em uma aba separada do app.
Opções de Estudo	Você poderá estudar todos os cards de um deck; ou apenas os que você errou; ou apenas os que você não estudou ainda; entre outras opções.

E como eu consigo baixar?



É muito fácil! Basta pesquisar por "Estratégia Cards" na loja oficial do seu smartphone.

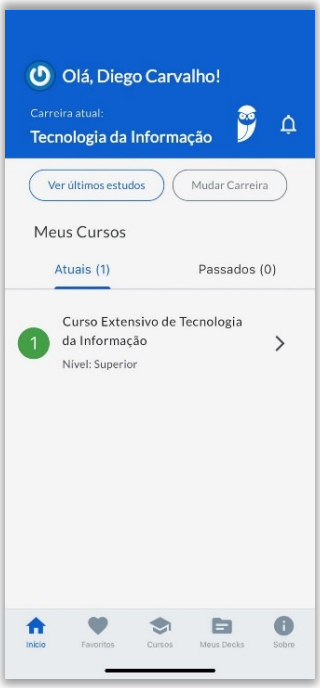
Se você tiver um Android, basta acessar a **Google Play**;

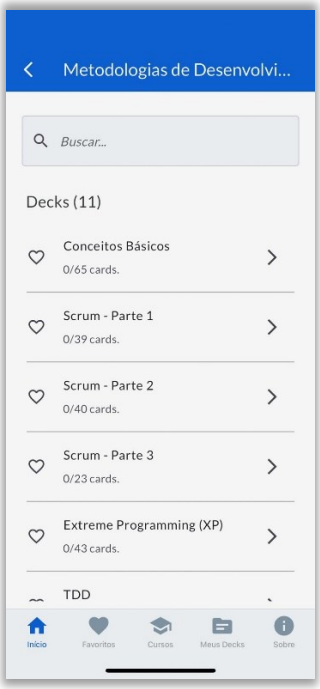
Se for tiver um iPhone, basta acessar a **App Store (iOS)**.

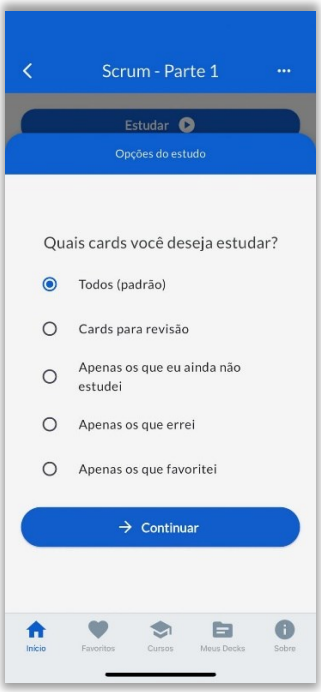
E para acessar?

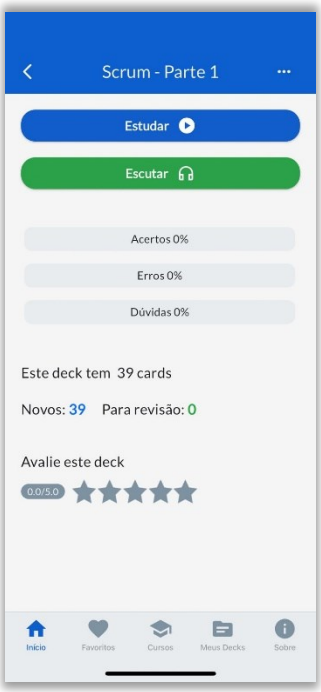
Para acessar, basta ter uma conta no Estratégia Concursos. Em seguida, utilize suas credenciais de login e senha para acessar o aplicativo. Por fim, acessa a carreira de Tecnologia da Informação.

Como utilizar o app:









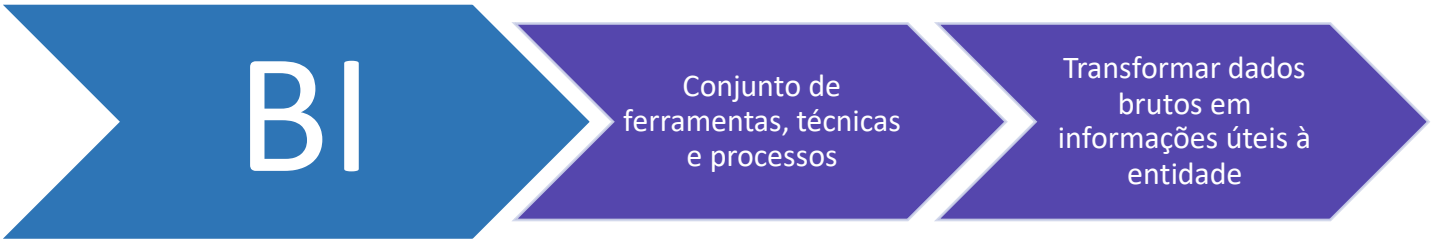
BUSINESS INTELLIGENCE

Conceitos Gerais

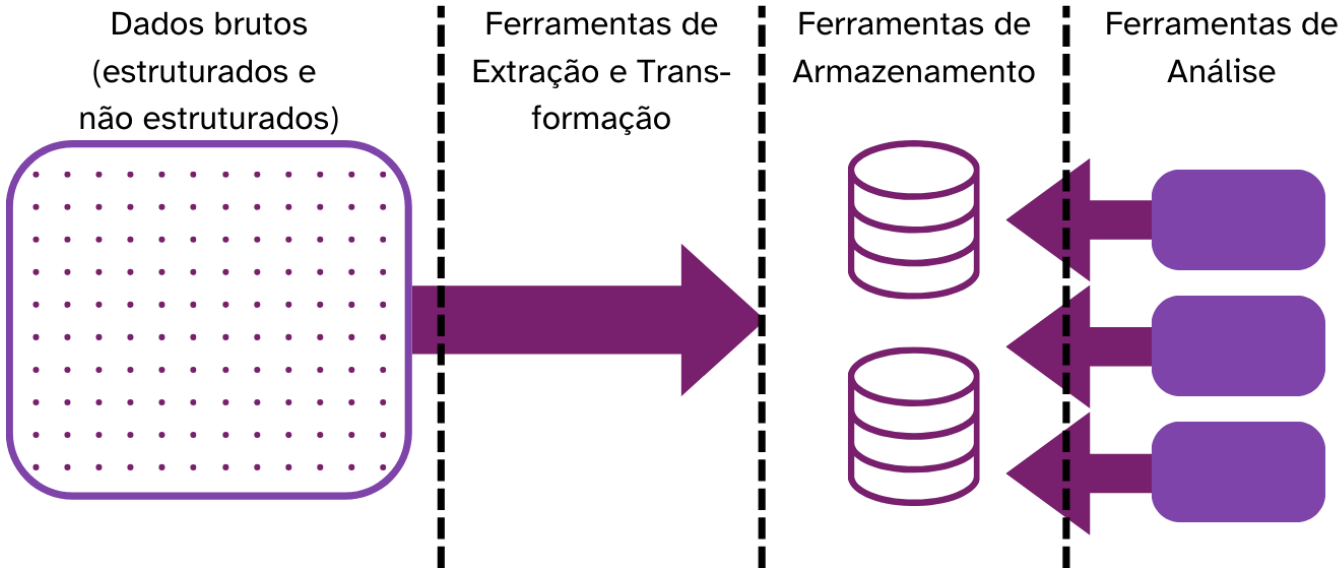
Antes de estudarmos o processamento e armazenamento de dados, precisamos entender o porquê da necessidade de movermos todos os dados de um lugar a outro e armazená-los. E esse motivo tem um nome: Business Intelligence.

A Inteligência de Negócios, Business Intelligence ou simplesmente BI, é um conjunto de técnicas, processos e ferramentas utilizadas para transformar dados brutos em informações significativas e úteis para a tomada de decisões estratégicas nas organizações.

O objetivo principal do BI é fornecer insights acionáveis que auxiliem na compreensão do desempenho atual do negócio, identificação de tendências, previsão de resultados futuros e suporte à formulação de estratégias, tudo a partir de uma análise de informações não triviais, isso é, informações que não estejam tão óbvias nos dados.



Para fazermos essa extração de informações, é necessário que sejam coletados os dados brutos, estruturados ou não, transformando-os para um padrão de uso requisitado pela entidade que irá os usar, armazená-los em conjuntos e utilizar ferramentas diversas de análise em cima desse conjunto de dados.

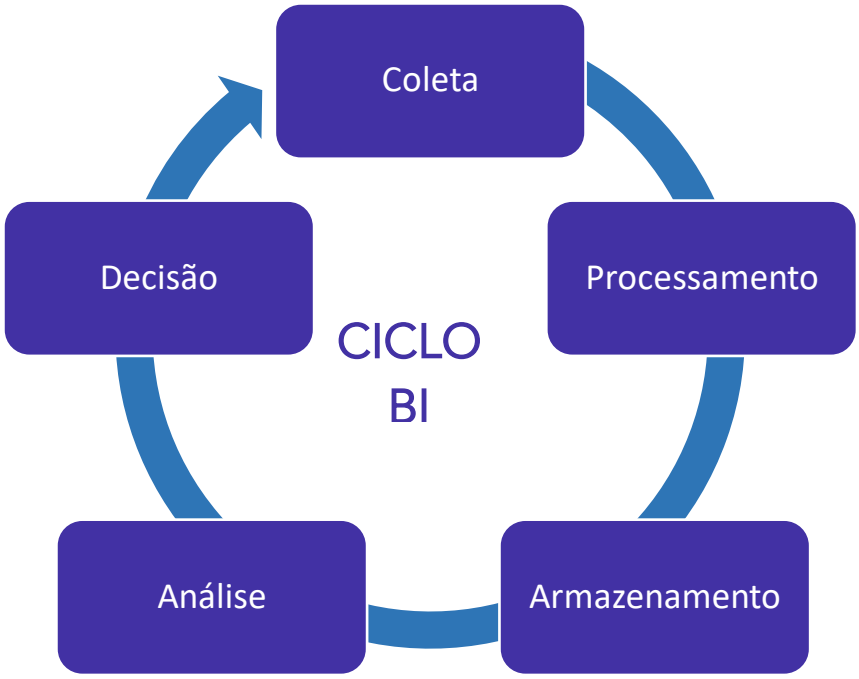


Cada parte do fluxograma possui ferramentas específicas. Veja:

- Ferramentas de extração e transformação:
  - ETL (Extract, Transform, Load)
  - ELT (Extract, Load, Transform)
- Ferramentas de Armazenamento
  - Data Warehouse
  - Data Lake
  - Data Mesh
- Ferramentas e Técnicas de Análise:
  - Visualização de dados
  - OLAP
  - Análises de tendência
  - Machine Learning
  - Programas de análise (Power BI, QLik View)

Podemos considerar que o processo do Business Intelligence envolve uma abordagem iterativa, onde cada ciclo de ações é repetida indeterminadas vezes, até que seja alcançado o resultado desejado: obter informações relevantes para a entidade. Podemos definir esse ciclo nas seguintes ações:

- Coleta de Dados: O primeiro passo do BI é justamente coletar os diferentes dados para serem usados nas decisões. Esses dados podem ser estruturados, oriundos de bancos de dados relacionais, por exemplo, ou não estruturados, oriundos das mais diversas fontes geradoras de informação;
- Processamento e Transformação: Muitas vezes é necessário que os dados estejam em algum padrão, tendo certo nível de estruturação, antes de serem armazenados. Por esse motivo, passam por um processo de análise, descarte de dados inutilizáveis e transformação para o padrão do repositório de destino;
- Armazenamento de dados: Precisamos de repositórios que suportem alto volume de dados, para permitir que as ferramentas de análise tenham um desempenho mais elevado;
- Análise: Utilizam-se diferentes técnicas, como visualização de dados (gráficos, tendências), mineração de dados, técnicas OLAP, para obter informações relevantes que auxiliarão a tomada de decisões da entidade;
- Tomada de decisão: Com base nas informações encontradas, busca-se tomar as decisões ótimas para a evolução da entidade.



QUESTÃO DE PROVA

(CEBRASPE/FUB/2022) A respeito de data warehouse, data mining e business intelligence, julgue o item subsequente.

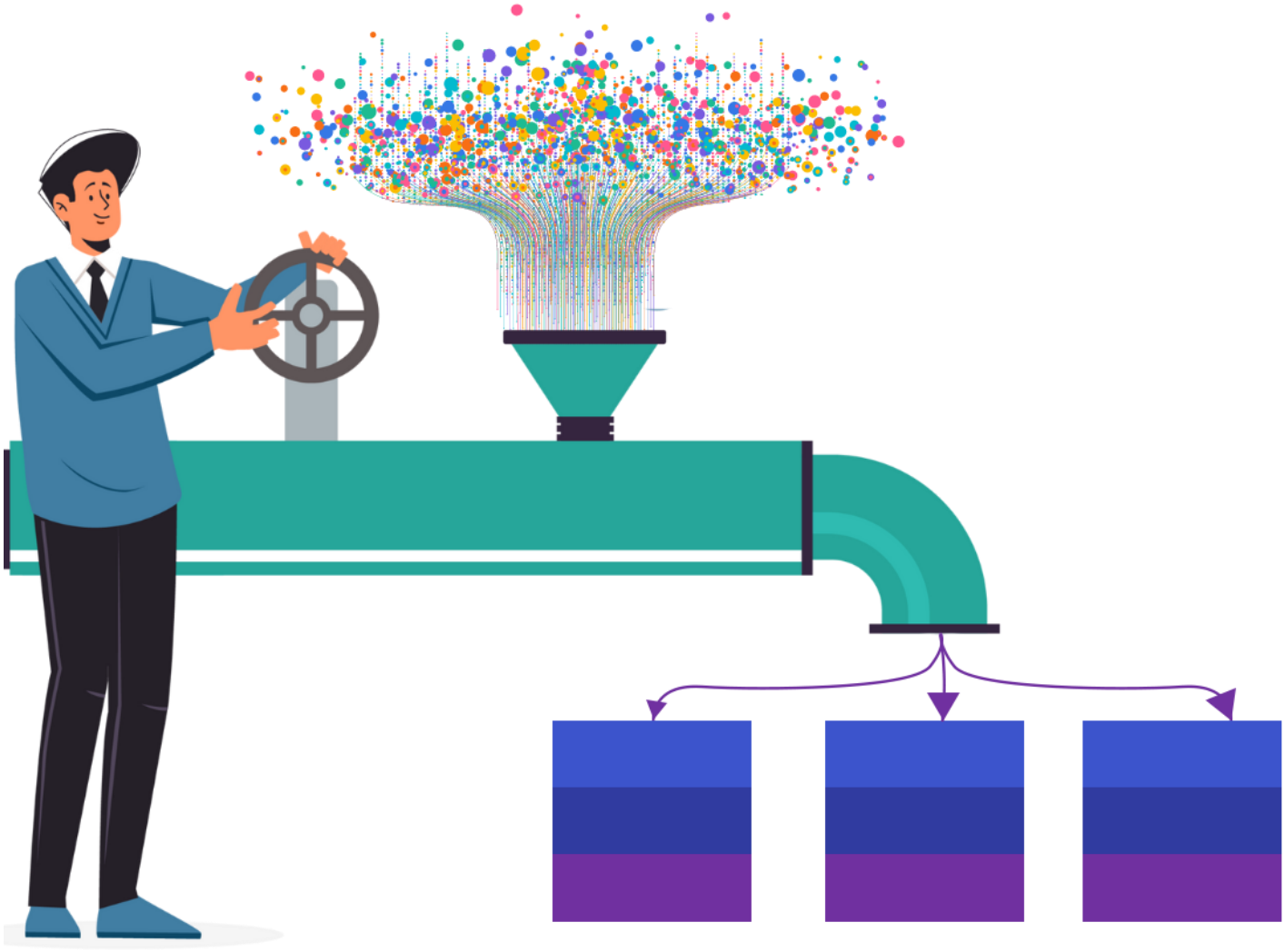
Coletar e transformar dados de várias fontes e descobrir tendências e inconsistências são etapas gerais dos processos de business intelligence.

Comentários:

Como acabamos de ver no nosso ciclo, essas são, de fato, etapas gerais do processo de descoberta de informações do BI. **(Gabarito: Certo)**

Processamento & Transformação

Nessa aula, iremos estudar as duas principais formas de transformar, processar e levar os dados da coleta até o repositório de destino: ETL e ELT. Também faremos relações diretas entre as duas formas.



Essas ferramentas formam o que chamamos de pipelines de dados. As *pipelines* são ferramentas que buscam levar um objeto de A a B. No nosso caso, visamos levar os dados dos pontos de coleta até o repositório de destino, fazendo as transformações necessárias nos dados no meio do caminho.

KDD

Muito do Business Intelligence se alinha ao conceito de Knowledge Discovery in Data, KDD - Descobrimento de Conhecimento em Dados. Ele é um processo interdisciplinar, que combina técnicas de banco de dados, estatística, aprendizado de máquina e visualização de dados. O objetivo é extrair conhecimento a partir de dados brutos - assim como o Business Intelligence.

De certa forma, podemos dizer que o KDD é o BI, numa abordagem um pouco mais "ampla", envolvendo o processo completo, desde a seleção de dados até a interpretação e avaliação das informações. Na verdade, o KDD traz um ciclo para a aquisição de informação bem definido, ao longo de 5 etapas:

- Seleção de Dados
- Pré-processamento
- Transformação
- Mineração de Dados
- Interpretação/Avaliação

Seleção de Dados

Na seleção de dados, consideramos quais dados serão relevantes para a análise, definindo as fontes de dados e os dados em si. Aqui, o objetivo é identificar fontes de dados apropriadas para os objetivos de negócio. Por exemplo, se a análise será de desempenho de vendas, não faz sentido selecionarmos dados de um banco de dados de Recursos Humanos - focaremos na parte de vendas.

Pré-processamento

Selecionados os dados, passamos por etapas de pré-processamento. O objetivo aqui é lidar com "problemas" nos dados: remoção de dados duplicados, remoção de dados ruidosos, tratamento de valores ausentes, remoção de *outliers*, entre outros. É aqui que tornamos o conjunto de dados de trabalho num conjunto mais "conciso". Os processos mais comuns aqui são:

- Deduplicação: remoção de valores duplicados ou aproximadamente duplicados;
- Remoção de ruídos: remoção de dados com pouca representatividade no conjunto;
- Tratamento de outliers: abordar valores aberrantes, escolhendo o melhor caminho a ser tomado, que pode envolver a remoção, tratamento de valores ou outros;
- Tratamento de valores ausentes: processo em que é decidido se valores ausentes devem ser imputados ou removidos;
- Correção de valores: medições podem conter erros, e o trabalho aqui é corrigi-los;
- Enriquecimento de dados: integrar fontes de dados externas aos dados para ter uma pluralidade de visões nos dados;

Caso sua prova exija um conhecimento mais aprofundado nesses temas de pré-processamento e de tratamento, eles serão vistos em um momento mais

adequado, quando você tiver um conhecimento mais aprofundado sobre tópicos subjacentes. Não se preocupe.

Transformação

Dados "corrigidos", vamos para a etapa de transformação dos dados. Aqui o objetivo é otimizar os dados para o melhor desempenho do modelo, alterando escalas de variações, diminuindo variáveis, discretizando ou agrupando dados. Assim como no pré-processamento, temos um conjunto extenso de atividades. Destaco as principais:

- Normalização: trazer um conjunto de variáveis para uma mesma escala de variação, facilitando a interpretação;
- Padronização: objetivo similar à normalização, trazendo as variáveis para uma variável normal padrão, com média igual a 0 e desvio padrão igual a 1
- Discretização: processo de transformar variáveis contínuas em variáveis discretas, isso é, que só podem assumir um ponto no espaço;
- Redução de dimensionalidade: redução da quantidade de variáveis trabalhadas em um modelo, buscando manter o máximo de variância nos dados;
- Codificação: Transformação de variáveis categóricas em numéricas;
- Agregação: junção de dados criando um conjunto resumido;

Mineração de Dados

Agora temos os dados prontos para serem trabalhados, passamos a utilizar diversas técnicas objetivando fazer a mineração dos dados. Assim como na mineração tradicional, onde tentamos extrair minérios valiosos de pedras, na mineração de dados procuramos extrair informações valiosas de dados.

Aqui o importante é conhecer as técnicas. Temos uma infinidade de técnicas, que serão aprofundadas ao longo do estudo específico de mineração de dados, mas podemos citar para você agora algumas delas:

- Clusterização/agrupamento: processo de divisão dos dados em diferentes grupos, conforme algum critério de separação - usualmente, através de similaridades;
- Classificação: atribuição de classes a conjuntos de dados;
- Regressão: processo de análise de variáveis, buscando entender correlações entre variáveis dependentes e independentes para prever valores futuros;
- Regras de associação: definição de regras que permitam inferir relações entre os dados;
- Detecção de anomalias: processo de detectar valores aberrantes e de difícil identificação manual, que possam identificar fraudes ou erros de medição;
- Mineração de texto: processo de extrair informações ou criar artefatos, como resumos, a partir de conjuntos de textos;

Avaliação e Interpretação dos Resultados

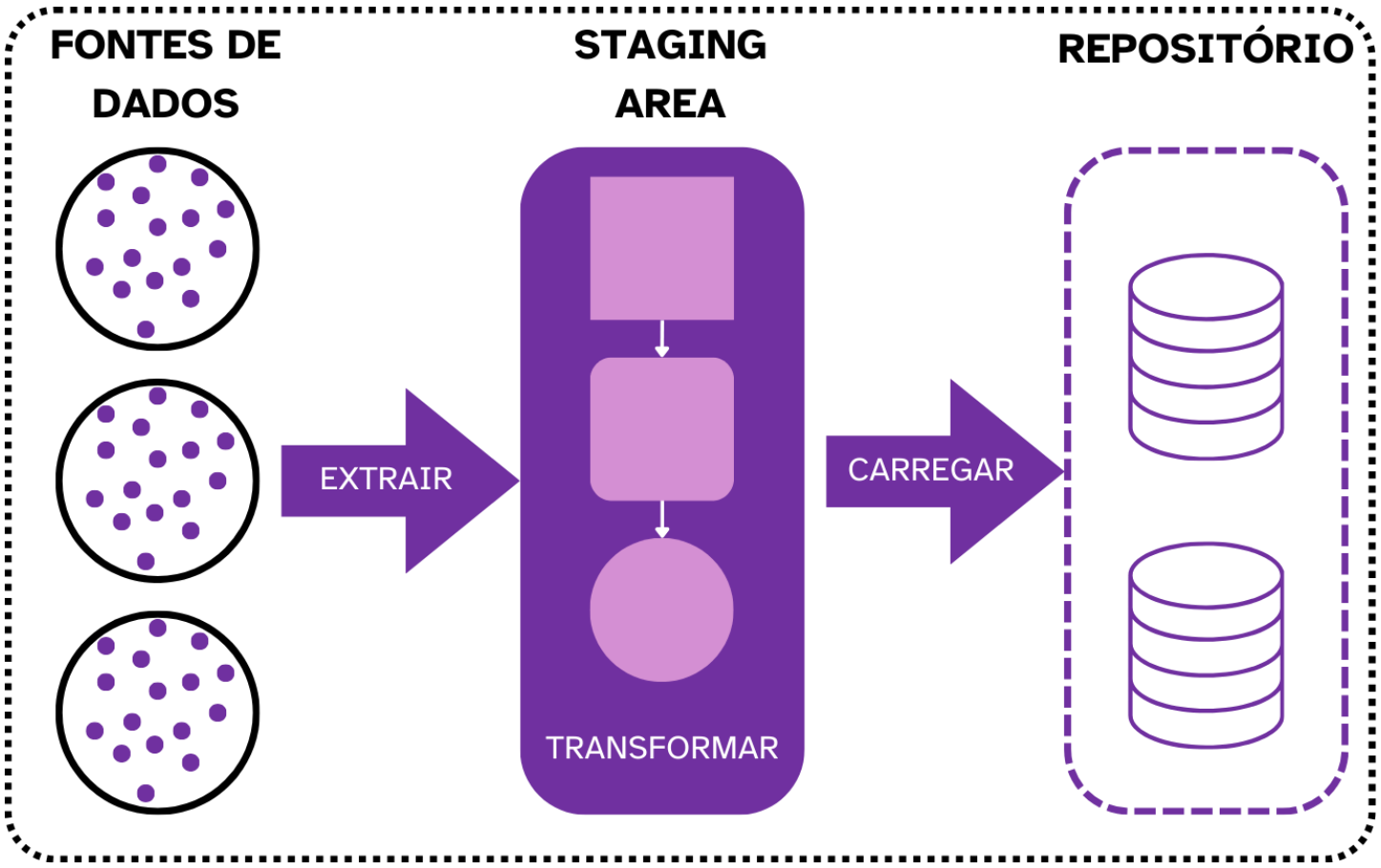
Na última etapa, temos todas as informações que precisamos em nossas mãos. Agora o objetivo é encontrar aplicações para elas, utilizar os padrões e *insights* descobertos para tomadas de decisões. Para isso, utilizamos padrões de visualização, como gráficos e *dashboards*, relatórios, e outras formas que permitam uma análise mais concisa.

ETL

Conceitos Gerais

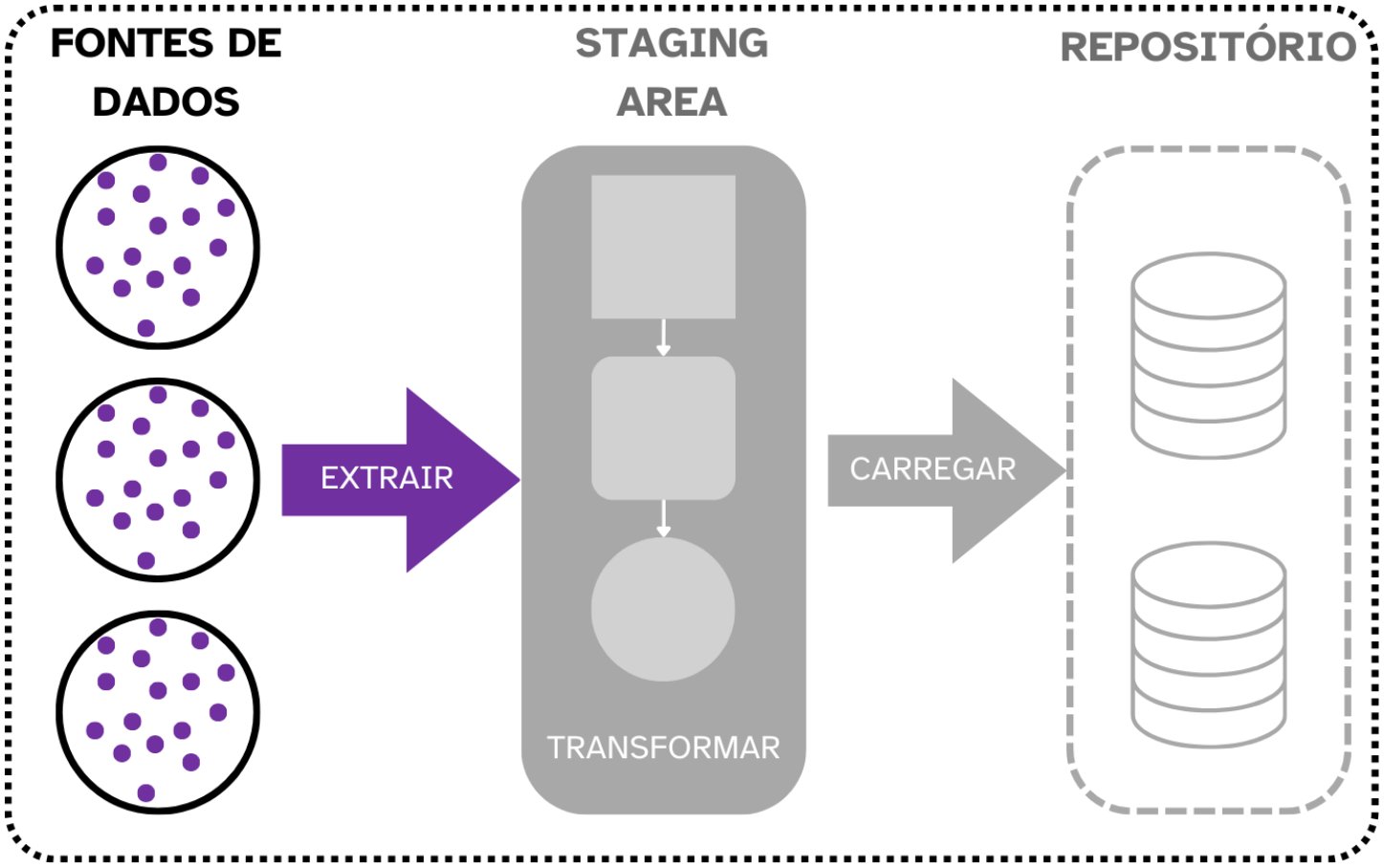
ETL é um acrônimo para "Extract, Transform and Load" – em tradução literal: Extrair, Transformar e Carregar. Nesse tipo de *pipeline*, temos três etapas distintas que ocorrem antes do dado ser armazenado no destino:

- Extração (Extraction)
- Transformação (Transformation)
- Carregamento (Load)



Vamos falar sobre cada uma dessas etapas do processo.

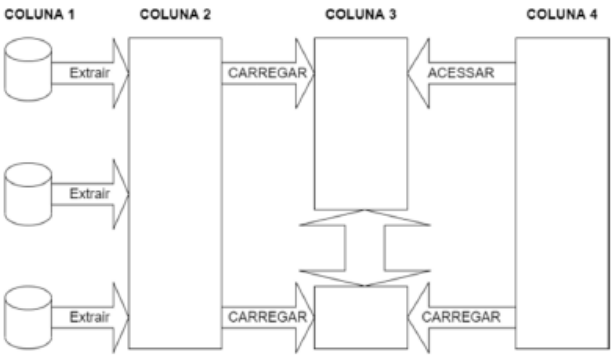
Extração



A etapa de extração é a primeira fase do processo, na qual os dados são coletados a partir de diversas fontes, como bancos de dados, arquivos CSV, APIs, sistemas legados, entre outros — perceba, então, que a extração poderá lidar com dados estruturados e não estruturados. Ela pode ser realizada de diferentes maneiras, dependendo da origem dos dados e das ferramentas utilizadas.

QUESTÃO DE PROVA

(CEBRASPE/FUNPRESP/2022) Uma empresa necessita estruturar, melhorar e utilizar cada vez mais recursos, a fim de gerar inteligência para o seu negócio. Nesse sentido, foi desenvolvido o esquema a seguir, a ser utilizado como uma visão dos elementos do seu respectivo data warehouse, que apoiará a inteligência do negócio.



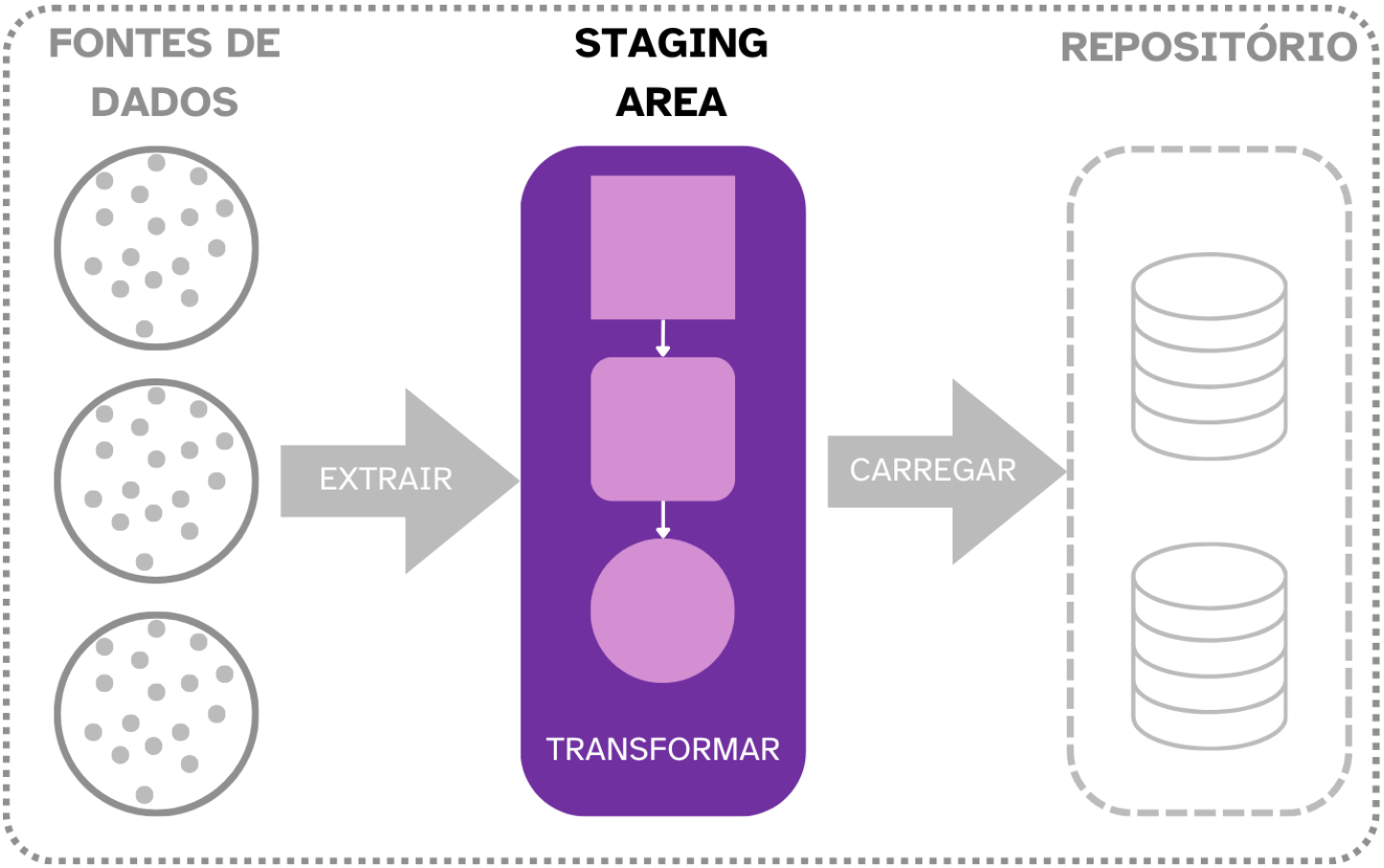
Com base nas informações apresentadas, julgue o item seguinte.

Ferramentas denominadas extract transformation load (ETL) são utilizadas para extrair dados da coluna 1 e disponibilizá-los na coluna 2.

Comentários:

Nossa coluna 1 representa aos repositórios de dados de onde extraímos nossos dados, que são disponibilizados para tratamento na coluna 2. (Gabarito: Certo)

Transformação



Após a etapa de extração, os dados passam pela fase de transformação em uma pipeline ETL. Nessa etapa, os dados extraídos são modificados, limpos, enriquecidos e reestruturados de acordo com as necessidades do destino final. A transformação de dados pode envolver uma série de atividades:

- Limpeza de Dados: Envolve a remoção de dados duplicados, tratamento de valores ausentes, correção de erros de formatação e padronização de dados.
- Integração de Dados: Quando os dados são provenientes de múltiplas fontes, é necessário integrá-los em um único conjunto de dados coerente. Isso pode envolver a resolução de discrepâncias de esquema, mapeamento de atributos e unificação de formatos.
- Enriquecimento de Dados: Às vezes, é útil enriquecer os dados com informações adicionais de fontes externas. Isso pode incluir a adição de dados demográficos, geoespaciais ou informações de terceiros para melhorar a análise.

Nesse assunto é importante que você conheça algumas técnicas empregadas para a transformação dos dados. Vamos vê-las.

OBS: Existem processos mais avançados de transformações, envolvendo o pré-processamento de dados, que necessitam de conhecimentos mais avançados. Caso seja necessária à sua prova, esse assunto será revisitado na profundidade adequada quando você tiver os conhecimentos necessários.

Discretização

A discretização é o processo de transformar variáveis contínuas em variáveis discretas, ou seja, agrupar valores em intervalos ou categorias discretas. Isso é útil quando se deseja simplificar a análise de dados contínuos ou quando os algoritmos de análise exigem que os dados estejam em formato discreto.

Variáveis contínuas e variáveis discretas são tipos de variáveis em estatística e análise de dados.

As variáveis contínuas podem assumir um número infinito de valores dentro de um intervalo específico e podem ser medidas em uma escala contínua. Por exemplo, altura, peso e temperatura são exemplos de variáveis contínuas, pois podem ter valores em qualquer ponto dentro de um intervalo, como 1,75 metros ou 27,3°C.

Já as variáveis discretas têm um conjunto finito ou enumerável de valores possíveis e são geralmente contadas em unidades inteiras. Por exemplo, o número de estudantes em uma sala de aula ou o número de carros em um estacionamento são exemplos de variáveis discretas, pois não podem ter valores fracionários e só podem assumir valores específicos, como 10 estudantes ou 20 carros.

(FGV/TJ SE/2023) Uma das etapas mais importantes do processo de Mineração de Dados é o pré-processamento dos dados das fontes que, normalmente, apresentam diversos tipos de heterogeneidade. A operação de pré-processamento que transforma dados quantitativos (contínuos) em dados qualitativos, ou seja, atributos numéricos em atributos discretos ou nominais com um número finito de intervalos, obtendo uma partição não sobreposta de um domínio contínuo, é a:

a) Redução;
b) Imputação;
c) Overfitting;
d) Discretização;
e) Undersampling.

Comentários:

A mineração de dados é uma das técnicas de análise de dados que iremos estudar durante as aulas ainda, mas ela envolve as etapas prévias realizadas no BI - como o tratamento dos dados. A transformação de dados contínuos em discretos recebe o nome de discretização. (Gabarito: Letra D)

Normalização e Padronização

A normalização e a padronização buscam colocar os dados em uma escala comum. Por exemplo, se estivermos lidando com duas variáveis distintas, como Idade e Salário, a escala absoluta de mudanças difere muito - 2 anos é bastante para idade, mas 2 reais é pouco para salário. Dessa forma, essas duas técnicas buscam trazer as variáveis para uma escala comum, de forma que uma variação de 1 unidade impacte o mesmo ambas delas.

A normalização, que não se confunde com as formas normais que vimos em aulas anteriores, é o processo de ajustar os valores de uma variável para que eles fiquem dentro de um intervalo específico, geralmente entre 0 e 1. Isso é feito calculando o valor relativo de cada ponto de dados em relação ao intervalo total dos dados. A normalização é útil quando os dados possuem intervalos muito diferentes e é importante que eles estejam na mesma escala para algoritmos de machine learning sensíveis à escala.

Já a padronização, também chamada de Z-Score, por outro lado, visa transformar os valores de uma variável para que eles tenham uma média zero e um desvio padrão de um — ou seja, transformar as variáveis em distribuições normais padrão. Isso é feito subtraindo a média dos dados e dividindo pelo desvio padrão. A padronização é útil quando os dados têm diferentes escalas e distribuições, e algoritmos como regressão linear e algoritmos baseados em distância, como *k-means*, podem se beneficiar de dados padronizados para melhor desempenho.

(IADES/UnDF/2022) O processo de normalização/padronização dos dados pode ser realizado de acordo com diferentes técnicas. O método cujo objetivo é transformar os dados de tal forma a serem reescalonados para uma distribuição com média 0 (zero) e desvio padrão 1 (um) denomina-se

a) Log.
b) Min-max.
c) Z-score (padronização).
d) Clipping.
e) KNN.

Comentários:

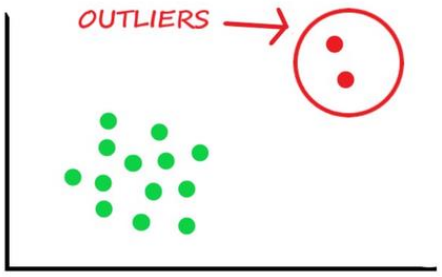
O processo de normalização responsável por "achatar" a variância e a variabilidade de um conjunto de dados, tornando sua média igual a 0 e variância igual a 1, é chamada de padronização, ou Z-score. (Gabarito: Letra C)

Tratamento de Dados Ausentes

O tratamento de dados ausentes visa lidar com valores ausentes ou faltantes em um conjunto determinado de dados. Valores ausentes podem surgir por diversas razões, como erros de coleta, falhas nos sistemas de armazenamento, ou simplesmente porque algumas informações não foram fornecidas. O tratamento adequado desses dados é fundamental para garantir a qualidade e a confiabilidade das análises e dos modelos construídos com esses dados

Esse tratamento é feito a partir da imputação de valores conforme várias técnicas, em sua maioria estatísticas. Esse é um ponto um pouco avançado para esse ponto na sua caminhada, mas podemos usar técnicas de aprendizado de máquina baseada em vizinhos (como o k-NN), médias, variâncias, entre outros.

Detecção e Remoção de Outliers



Outliers são pontos de dados que se diferenciam significativamente do restante do conjunto de dados. Eles podem representar observações incomuns, extremas ou até mesmo erros de medição. A presença de outliers pode distorcer a análise estatística e prejudicar a precisão dos modelos, tornando sua detecção e remoção uma etapa importante no pré-processamento de dados.

Redução de Dimensionalidade

A redução de dimensionalidade é uma técnica utilizada para reduzir o número de variáveis ou dimensões em um conjunto de dados. Em problemas com muitas variáveis, a remoção de dimensionalidade pode ser útil para simplificar a análise, reduzir a complexidade computacional e evitar o que é conhecido como a "maldição da dimensionalidade". Esta última se refere ao fenômeno em que a precisão dos modelos de análise diminui à medida que o número de variáveis aumenta, devido à dispersão dos dados em um espaço de alta dimensionalidade.

Aqui é importante encontrar o *middle ground*, um ponto intermediário na remoção de variáveis de forma a não termos variáveis excessivas, mas também não termos ausência de informações. Para isso, usamos diversas técnicas, mais notoriamente o Principal Component Analysis (PCA), que será analisado em momento oportuno.

Staging Area

A Staging Area não é uma técnica de tratamento de dados, e sim um "lugar" onde essas técnicas ocorrem. Ela serve como um espaço temporário onde os dados extraídos são armazenados e preparados antes de serem carregados no destino final, como um data warehouse.

Ela é uma área completamente isolada e efêmera, isso é, existe apenas com o único objetivo de transformações, não permitindo acesso externo, e tem sua existência apagada após esvaziada - isso é, após ter todos seus dados movidos ao destino final.

QUESTÃO DE PROVA

(CESGRANRIO/TRANSPETRO/2018) Os sistemas de data warehouse diferem de várias formas dos sistemas transacionais das empresas, como, por exemplo, em seu modelo de dados. Para transferir e transformar os dados dos sistemas transacionais para os sistemas de data warehousing, é comum utilizar, como estratégia, a existência de uma camada especial da arquitetura conhecida como

a) Data Marts
b) Data Staging Area

c) Dimensional Model Area
d) Presentation Area
e) Living Sample Area

Comentários:

A camada especial a que se refere nossa questão é justamente a staging área. (Gabarito: Letra B)

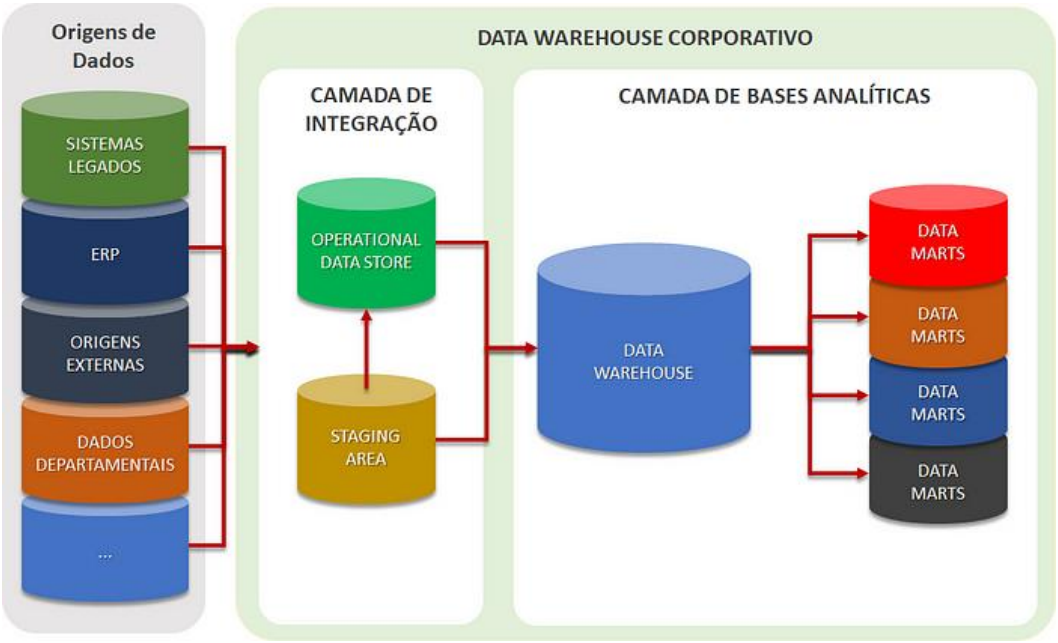
Operational DataBase (ODS)

O Operational Database (ODS), ou Banco de Dados Operacional, é um outro repositório intermediário. Porém, aqui temos a possibilidade de consultas detalhadas, tendo o espaço funcionando como uma forma de simplificar a criação de Data Warehouses (o destino final do ETL).

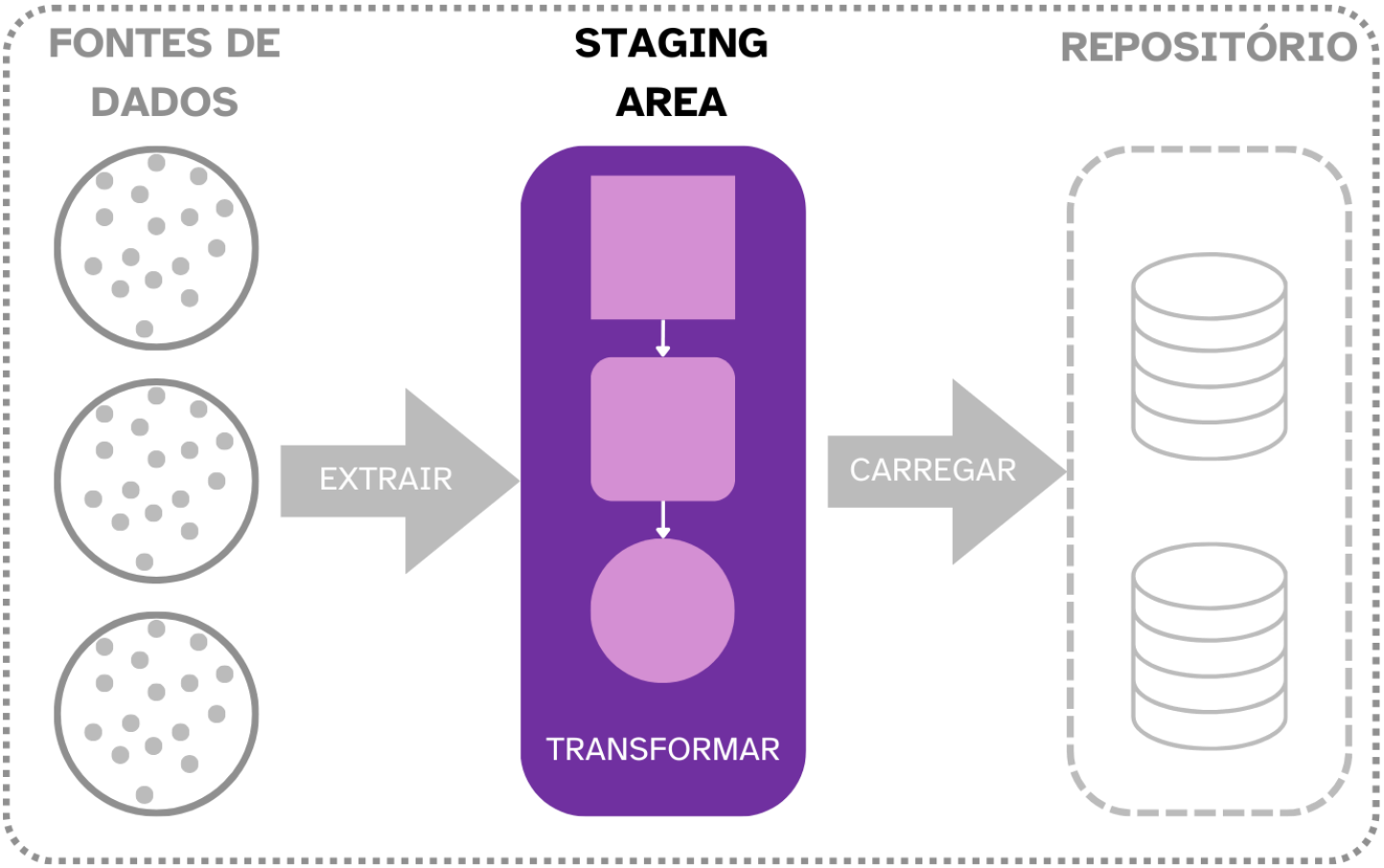
Ao contrário da Staging Area, o ODS é persistente entre as cargas, permitindo as consultas diretamente nele. Além disso, ele é recomendado:

- Quando há necessidade de detalhamento em relatórios ou consultas, não suportados por um modelo dimensional ou por acesso direto às origens
- Quando há necessidade de manter os dados para permitir atualizações e reconstrução do repositório de destino
- Quando são necessários processos complexos de uniformização, regularização ou pré-processamentos de múltiplas origens

Usualmente mantemos tanto uma *Staging Area* quanto um ODS no ecossistema - a *Staging Area* sendo alocada antes do ODS, alimentando-a. Esses elementos formam o que chamamos de camada de integração.


Origens de Dados
SISTEMAS LEGADOS
ERP
ORIGENS EXTERNAS
DADOS DEPARTAMENTAIS
...
DATA WAREHOUSE CORPORATIVO
CAMADA DE INTEGRAÇÃO
OPERATIONAL DATA STORE
STAGING AREA
CAMADA DE BASES ANALÍTICAS
DATA WAREHOUSE
DATA MARTS
DATA MARTS
DATA MARTS
DATA MARTS

Carregamento



O carregamento é a última etapa da nossa *pipeline* ETL. Ele é responsável por coletar o conjunto de dados transformados na *Staging Area* e carregá-los nos repositórios de dados, como o Data Warehouse e o Data Lake. Esse é um processo lento, visto que a quantidade de dados inserida, muitas vezes, é massiva, por isso recebe relevante atenção.

O carregamento é um mundo em si, com diferentes técnicas e formas de realizá-lo. Podemos adotar, basicamente, dois tipos de carga:

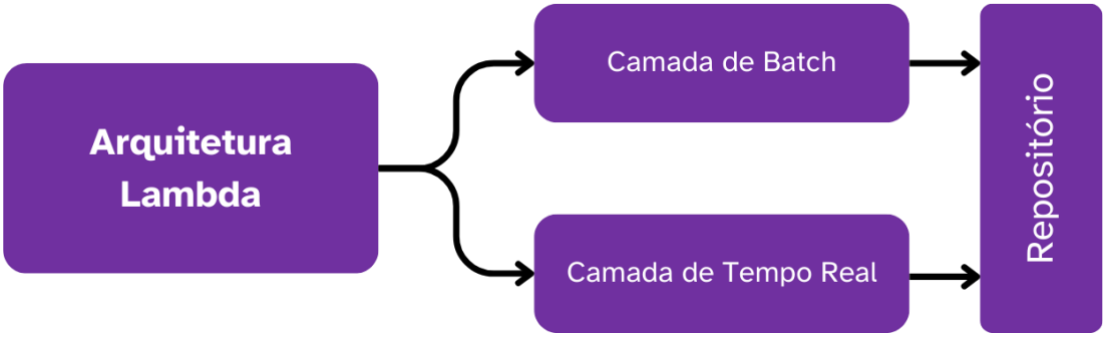
- Carga em lotes (ou *batches*), onde agregamos uma determinada quantidade de dados e o inserimos de uma só vez
- Utilizar o carregamento em tempo real. Aqui, os dados são inseridos conforme são gerados, através de um fluxo chamado de streaming de dados.

Nesse sentido, temos duas arquiteturas que ganham destaque. Vamos vê-las.

Arquitetura Lambda

A arquitetura Lambda é uma abordagem mais antiga ao processamento da dados, trabalhando com duas camadas distintas. A saber:

- Camada de Batch (Batch Layer): Nesta camada, os dados são processados em lotes, geralmente utilizando técnicas de ETL (Extract, Transform, Load), como discutido anteriormente. Os dados processados são então armazenados em um data warehouse, onde podem ser analisados por meio de consultas ad hoc e relatórios.
- Camada de Tempo Real (Speed Layer): Simultaneamente à camada de batch, os dados também são processados em tempo real à medida que chegam. Isso envolve o uso de sistemas de processamento de streaming, como Apache Kafka ou Apache Flink, para processar e analisar os dados em tempo real. Os resultados dessas análises em tempo real são então agregados aos dados armazenados na camada de batch.



Arquitetura Kappa

O modelo Lambda começou a apresentar um problema de latência. A latência nada mais é que o tempo de demora para ocorrer a carga dos dados - principalmente quando tratamos da camada de *batch*. Nesse sentido, surgiu o próximo modelo de arquitetura: a arquitetura Kappa. Esse modelo de arquitetura simplifica a estrutura do modelo Lambda, trazendo apenas uma camada de processamento, funcionando em um streaming contínuo.

- Camada de Processamento de Streaming: Nesta abordagem, os dados são processados em tempo real à medida que chegam, utilizando frameworks de processamento de streaming como Apache Kafka ou Apache Flink. Os dados são processados uma única vez e os resultados são agregados e armazenados diretamente no repositório.

(FGV/SEFAZM AM/2022) Com relação às arquiteturas de big data, analise as afirmativas a seguir.

I. As arquiteturas de big data suportam um ou mais tipos de carga de trabalho, por exemplo, processamento em lote de fontes de big data em repouso; processamento em tempo real de big data em movimento; exploração interativa de big data e análise preditiva e aprendizado de máquina.
II. A arquitetura kappa aborda o problema da baixa latência criando dois caminhos para o fluxo de dados. Todos os dados que entram no sistema passam por dois caminhos: a camada de lote (caminho frio) que armazena os dados de entrada em sua forma bruta e executa o processamento os dados em lote, e a camada de velocidade (hot path) que analisa os dados em tempo real. Essa camada é projetada para ter baixa latência, em detrimento da precisão.
III. A arquitetura lambda, posterior à kappa, foi proposta para ser uma alternativa para mitigar os problemas da baixa latência. Lambda tem os mesmos objetivos da kappa, mas com uma distinção importante: todos os dados fluem por um único caminho, usando um sistema de processamento de fluxo de dados. Semelhante à camada de velocidade da arquitetura lambda, todo o processamento de eventos é realizado através de um fluxo único de entrada.

Está correto o que se afirma em

a) I, apenas.
b) II, apenas.
c) III, apenas.
d) I e II, apenas.
e) II e III, apenas.

Comentários:

Vamos analisar cada uma das alternativas.

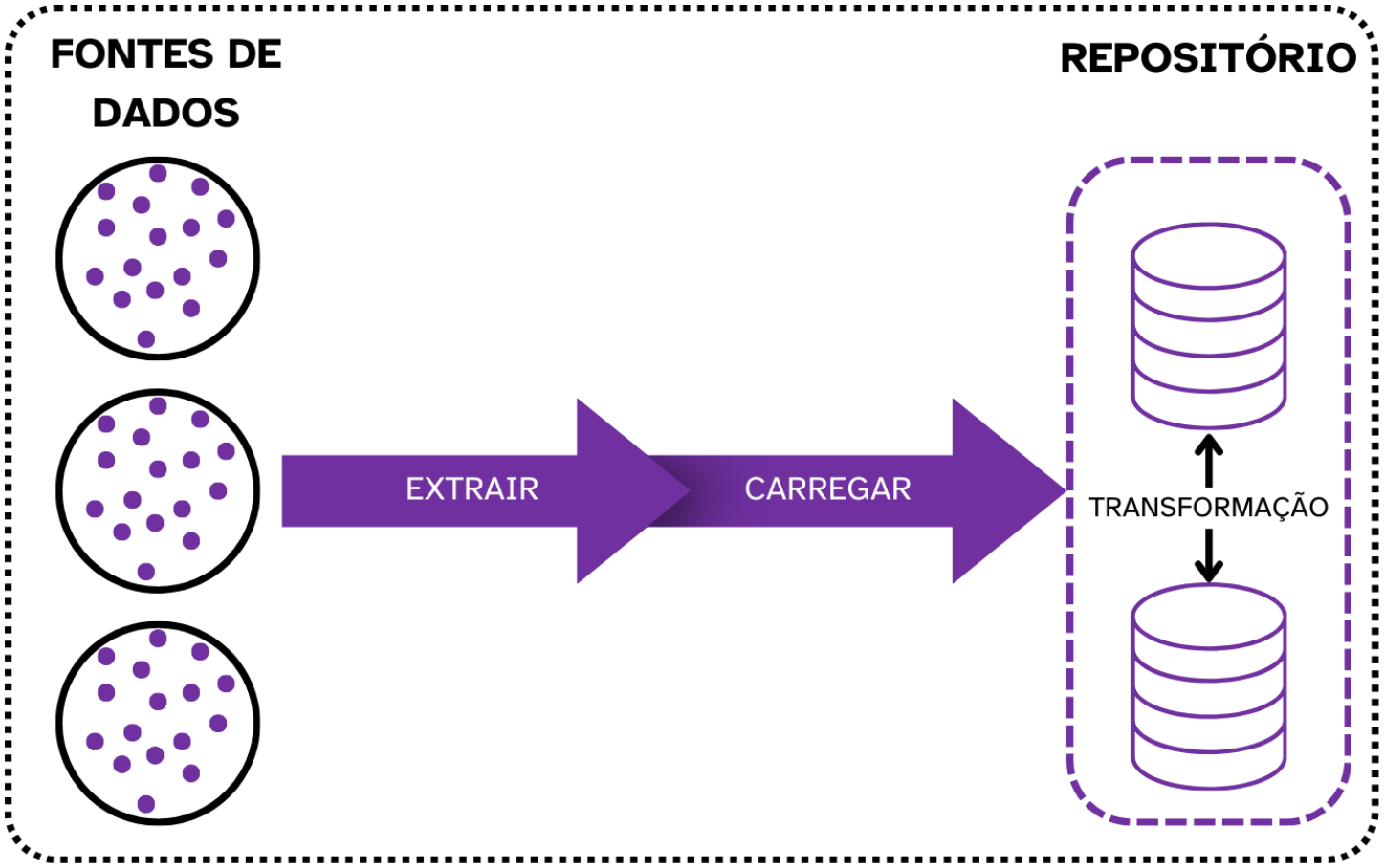
I - Ignore a primeira afirmativa, já que ainda não vimos o assunto – mas saiba que ela está correta. Vamos analisar as restantes.
II – Temos uma inversão do conceito. A afirmativa descreve a arquitetura lambda, e não a Kappa. Se aplicarmos todos os conceitos à arquitetura Lambda, a afirmativa estaria correta.
III – A mesma coisa acontece aqui – temos a inversão de conceitos, que descrevem, na verdade, a arquitetura Kappa.

Está correta, portanto, apenas a afirmativa I. (Gabarito: Letra A)

ELT

Conceitos Gerais

Bom, agora que você entendeu a estrutura geral do ETL, o ELT não deve ser um segredo para você. Novamente, temos um acrônimo para as três mesmas etapas - Extract (Extrair), Load (Carregar) e Transform (Transformar). Mas, como você deve ter percebido, temos uma inversão de etapas - aqui o carregamento ocorre antes da transformação.



"Mas Felipe, por que eu faria isso?"

Veja, quando você faz a transformação dos dados em uma etapa prévia, isso acaba aumentando a latência na carga ou, em outras palavras, aumentando o tempo que se leva para pegar o dado lá do ponto de origem e inserir no repositório de destino. Muitas vezes, é necessário que tenhamos o dado gerado rapidamente, principalmente quando estamos usando ferramentas que usam quantidades massivas de dados, como uma inteligência artificial.

ETL ou ELT?

A escolha entre ETL e ELT depende muito do contexto do sistema de BI da entidade, dos requisitos de negócio, de segurança, entre outros. Vou trazer uma tabela para fazermos comparações e depois trataremos melhor sobre elas.

Tópico	ETL	ELT
Ordem	Extração > Transformação > Carga	Extração > Carga > Transformação
Latência	Alta	Baixa
Segurança	Maior	Menor
Flexibilidade	Menor	Maior
Escalabiliade	Menor	Maior
Integração Principal	Data Warehouse	Data Lake
Espaço de Armazenamento	Menor	Maior

Em geral, operações de ETL são destinadas a interagir com Data Warehouses, repositórios de dados que possuem uma estrutura mais definida, mais padronizada. Esses tipos de repositórios são mais adequados a processos de Business Intelligence e em ambientes que necessitem de determinado nível de segurança, já que os dados são tratados antes do carregamento, de forma a ficar adequados a diversas diretrizes de segurança.

Em contrapartida, temos uma latência maior no carregamento, devido ao processo de alterações feitos antes da carga. De outro lado, como temos alterações feitas antes, envolvendo compactação e remoção de dados desnecessários, temos um espaço de armazenamento exigido muito menor.

De outro lado, as operações de ELT são direcionadas a interagir primariamente com Data Lakes, um tipo de repositório muito mais desestruturado, que veremos logo mais. Dessa forma, temos muito mais velocidade na carga e escalabilidade, funcionando idealmente para ferramentas que exigem essas quantidades massivas de dados, como ferramentas de *machine learning* e inteligência artificial.

O ponto negativo é a falta de organização e padronização, que acaba tornando esses processos mais suscetíveis a vulnerabilidades de segurança, e ocupando um espaço muito maior no sistema, já que temos muito "lixo" carregado conjuntamente.

QUESTÃO DE PROVA

(CEBRASPE/DATAPREV/2023) Julgue o item a seguir, relativos a ELT e ETL.

Para uma mesma base de dados, um ETL demanda menos espaço de armazenamento do que um ELT.

Comentários:

Perfeito. Como vimos, pelos dados estarem estruturados, sem repetições desnecessárias e dados sem utilidade, temos menos espaço de armazenamento demandado. (Gabarito: Certo)

(CEBRASPE/SEFAZ CE/2021) Em relação a big data e analytics, julgue o próximo item.

Comparado ao ETL, o ELT apresenta vantagens como tempos menores de carregamento e de transformação de dados, e, consequentemente, menor custo de manutenção.

Comentários:

Justamente pelo ELT realizar a transformação apenas após a inserção dos dados, ele tem uma maior velocidade. Além disso, por ter uma maior maleabilidade de sua estrutura, seu custo de manutenção acaba caindo drasticamente. (Gabarito: Certo)

REPOSITÓRIOS DE DADOS

Conceitos Gerais

Os repositórios de dados são onde armazenamos os dados carregados. Eles são destinados a suportar quantidades massivas de dados — e quando falo massivas, são massivas mesmo, chegando a petabytes, que corresponde a 1.024 TB (Terabytes), ou 1.000.000 GB (Gigabytes).

Por esses motivos, precisamos ter estruturas que, ao mesmo tempo, atendam a esses requisitos de massividade de conteúdo, e forneçam um suporte ao seu objetivo principal - a análise de dados visando o fornecimento de informações essenciais à entidade ou, em suma, o Business Intelligence.

Temos três tipos diferentes de repositórios principais:

- Data Warehouse (Armazém de Dados)
- Data Lake (Lago de Dados)
- Data Mesh (Malha de Dados)

ESQUEMATIZANDO

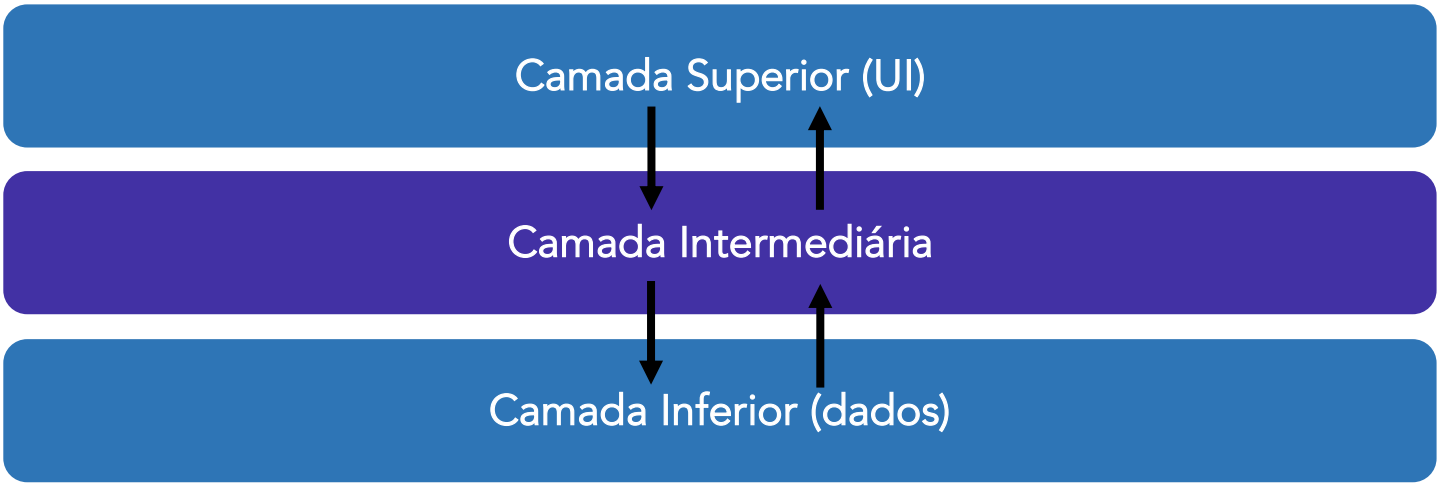
Tipos de Repositórios

Data Warehouse	Data Lake	Data Mesh

Data Warehouse

Um Data Warehouse, ou Armazém de Dados, é um repositório centralizado de dados, que trabalha com dados estruturados, e destinado a suportar análises de dados com foco no Business Intelligence. Usualmente, os dados dentro de Data Warehouses são armazenados em bancos de dados dimensionais, assuntos que veremos na próxima aula.

A arquitetura do Data Warehouse é feita em camadas. Temos uma camada superior, mais próxima ao usuário final, que fornece uma UI (User Interface, ou Interface de Usuário) para lidarmos com esses dados, aplicando, sobre ela, ferramentas de análise, relatórios e outra técnicas compatíveis. A camada intermediária é responsável por prever os mecanismos necessários para que as ferramentas de análise cheguem nos dados. Por fim, a camada inferior é o próprio servidor do banco de dados, onde o dado está armazenado.



O objetivo de um Data Warehouse é fornecer uma visão única e consistente dos dados de negócios da organização, independentemente de onde esses dados são originados ou armazenados. Isso permite que os usuários façam análises mais precisas e confiáveis, e tomem decisões melhores e mais informadas.

Os Data Warehouses possuem quatro pilares essenciais, que orientam todo o seu funcionamento. São as características mais preponderantes de sua estrutura: orientação a assunto, não volatilidade, variância no tempo, integração.

ORIENTADO A ASSUNTO

O Data Warehouse é orientado a assunto porque ele é projetado para lidar com informações específicas sobre um determinado assunto ou tópico de negócios, como vendas, finanças ou estoque. Isso significa que os dados são organizados em torno de assuntos de negócios, em vez de serem organizados em torno de transações ou atividades operacionais. Essa abordagem

permite que os usuários façam análises mais precisas e relevantes, com foco nas informações que são importantes para o seu trabalho.

NÃO VOLÁTIL

O Data Warehouse é não volátil porque ele é projetado para armazenar dados históricos de negócios, que não mudam após a inserção. Uma vez que os dados são carregados no Data Warehouse, eles são mantidos lá como um registro histórico, e não são atualizados, apenas recebendo novos dados. Isso garante que os dados históricos estejam sempre disponíveis para análises e relatórios, mesmo que ocorram alterações nos sistemas de origem.

VARIANTE NO TEMPO

O Data Warehouse é variante no tempo porque ele é projetado para manter informações históricas e permitir a análise de dados em diferentes períodos de tempo. Isso significa que os usuários podem analisar como as informações de negócios mudaram ao longo do tempo, identificar tendências e padrões, e fazer previsões com base nessas análises. Essa abordagem permite que as empresas respondam rapidamente às mudanças no mercado e tomem decisões mais informadas com base em informações históricas.

INTEGRADO

O Data Warehouse é integrado porque ele é projetado para integrar dados de diversas fontes, para fornecer uma visão única e consistente dos dados de negócios. Isso significa que os dados são consolidados em um único local, independentemente de onde eles são originados ou armazenados. A integração de dados garante que os usuários tenham acesso a informações precisas e confiáveis, e podem confiar nas análises e relatórios gerados pelo Data

Orientado a assuntos

Não volátil

Variante no tempo

Integrado

QUESTÃO DE PROVA

(CEBRASPE/SEFAZ CE/2021) Julgue o próximo item, relativo ao business intelligence (BI).

Um data warehouse (DW), ainda que seja não volátil — ou seja, após os dados serem inseridos nele os usuários não podem alterá-los — é variável no tempo, pois mantém um conjunto de dados históricos que oferecem suporte à tomada de decisões.

Comentários:

Perfeito! É justamente isso que quer dizer a não volatilidade e a variância no tempo de um DW. O dado em si não é atualizado, mas são feitas novas inserções com esse mesmo dado com valores diferentes, em momentos diferentes, criando um contexto histórico para o mesmo. *(Gabarito: Certo)*

(CEBRASPE/MEC/2015) No que se refere a data warehouse, julgue o item que se segue.

Data warehouse é um banco de dados projetado para obter melhor desempenho na consulta e análise de dados, em vez de processamento de transações.

Comentários:

Exatamente. O Data Warehouse tem foco total e completo na consulta e recuperação de dados, ao contrário dos bancos de dados transacionais (como o relacional), que são destinados a lidar com as transações do dia a dia e, portanto, focam nelas. Nesse sentido, correta a afirmativa. *(Gabarito: Certo)*

Construção

Os Data Warehouses são compostos de partes menores, mais especializadas e menos abstratas, chamadas de Data Marts. Cada Data Mart pode representar um conjunto específico de dados mais especializado, com um assunto mais definido.

Imagine que temos um Data Warehouse destinado a suportar decisões em um cursinho para concursos, que oferece cursos de diversos materiais. Pode ser interessante subdividir esse conjunto total em conjuntos menores, especificados — os Data Marts. Podemos ter um Data Mart para a área fiscal, outro para tribunais, ou até mesmo uma divisão por matérias, tendo um para TI, outro para português, e assim em diante.

Durante a construção de um Data Warehouse, temos duas diferentes abordagens, idealizadas por dois cientistas diferentes: William Inmon e Ralph Kimball.

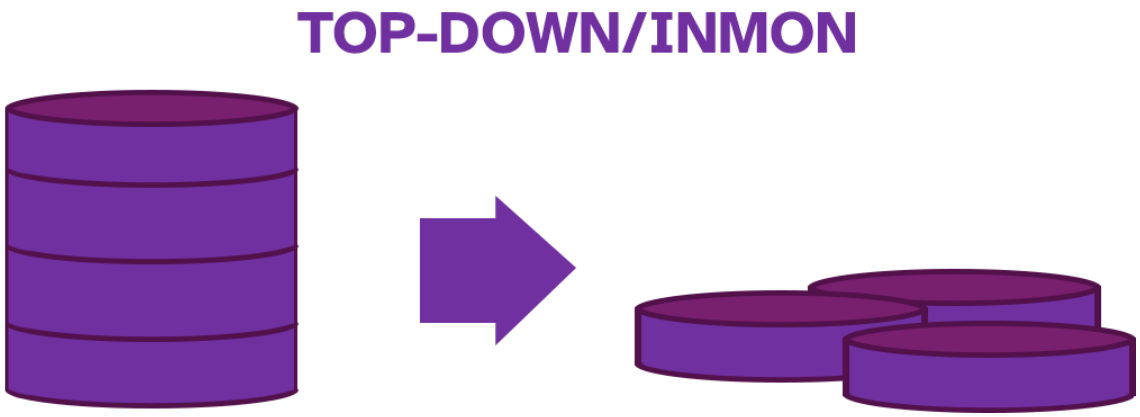
Abordagens na consturção do DW

Top-Down/Inmon	Bottom-Up/Kimballl

A abordagem de Inmon, também chamada de Top-Down, prega que primeiro deve-se construir o todo, para depois as subdivisões, Dessa forma, primeiro construímos o Data Warehouse como um todo, que recebe o nome de EDW - Enterprise Data Warehouse, ou Armazém de Dados Corporativo.

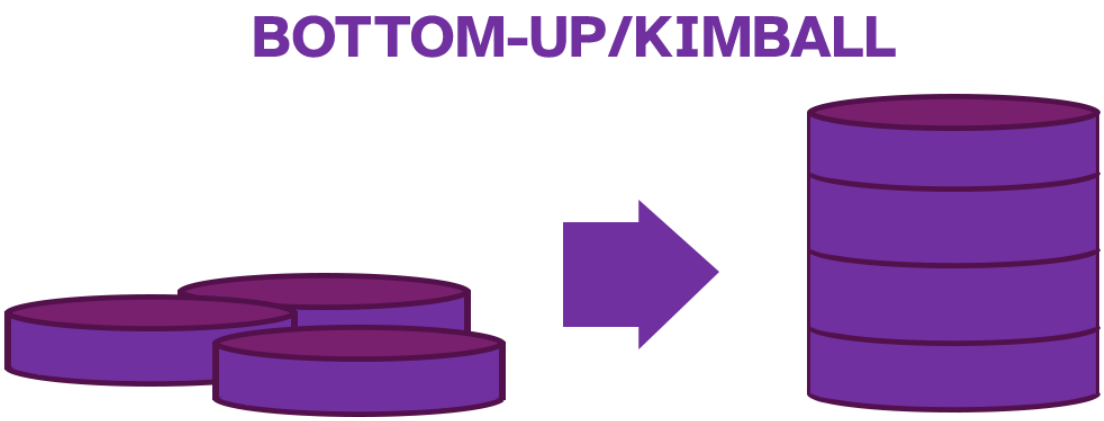
Nessa abordagem, temos o EDW funcinonando como *single source of trugh* (SST), ou ponto único de verdade, o que garante maior consistência nos dados. Aqui também é empregada uma modelagem normalizada, onde os dados são armazenados em tabelas altamente normalizadas - usualmente na 3FN.

Por fim, a partir do Data Warehouse, criamos os diferentes Data Marts setorizados, que irão receber as consultas das ferramentas de análise. Então perceba, nesse modelo a consulta é feita nos Data Marts, não no Data Warehouse em si — ele apenas funciona como um repositório central de onde os Data Marts retiram suas informações.



Já na abordagem de Kimball, temos a criação primária sendo baseada nos Data Marts, por isso recebe o nome de bottom-up. Assim, conseguimos atender necessidades de setores específicos antes de criarmos o "todo". Nesse sentido, Kimball introduz um conceito muito importante, que é o motivo pelo qual esse é o modelo mais abordado por questões: o banco de dados dimensional.

Veremos mais sobre ele na próxima aula, mas saiba que é um modelo de banco dados em que os dados são tratados em três dimensões, não apenas em tabelas relacionais. Com isso, conseguimos fazer uma entrega de valor mais para a entidade que estiver usando os Data Warehouses.



TOP-DOWN / INMON	BOTTOM-UP / KIMBALL
Data Warehouse → Data Marts	Data Marts → Data Warehouse
Bancos de Dados Relacionais (adaptados)	Bancos de Dados Dimensionais
Modelos normalizados	Opcionalmente normalizados
Voltado para profissionais da TI	Voltado para usuários finais
Consultas nos Data Marts	Consultas no Data Warehouse

(FCC/TRT 20/2016) Considere, por hipótese, que o Tribunal Regional do Trabalho da 20ª Região tenha optado pela implementação de um DW (Data Warehouse) que inicia com a extração, transformação e integração dos dados para vários DMs (Data Marts) antes que seja definida uma infraestrutura corporativa para o DW. Esta implementação

a) é conhecida como top down.
b) permite um retorno de investimento apenas em longo prazo, ou seja, um slower pay back.
c) tem como objetivo a construção de um sistema OLAP incremental a partir de DMs independentes.
d) não garante padronização dos metadados, podendo criar inconsistências de dados entre os DMs.
e) tem como vantagem a criação de legamarts ou DMs legados que facilitam e agilizam futuras integrações.

Comentários:

Vamos analisar cada alternativa, tendo em mente que a abordagem caracterizada por construir vários DMs antes do DW equivale à abordagem Bottom-Up.

a) Errado. Como estamos criando vários DMs antes do DW, estamos usando a abordagem bottom-up.
b) Errado. Nós já temos os DMs criados logo no início do processo, o que permite um retorno do investimento muito mais rápido, se comparado ao modelo top-down.
c) Errado. A alternativa descreve a abordagem top-down.
d) Certo. Como cada DM é trabalhado de forma independente num primeiro momento, podemos ter certas inconsistências.
e) Errado. As DMs legado são uma desvantagem. Sistemas legado são sistemas que se tornam obsoletos, mas ainda continuam em uso pela entidade, devido à importância da sua estrutura para as operações.

Nesse sentido, correta a letra D. *(Gabarito: Letra D)*

Matriz de Barramento

A matriz de barramento, ou bus matrix, é uma ferramenta crucial no contexto de projetos de data warehousing. Ela serve como um guia visual que mapeia os requisitos de negócios e as dimensões-chave do data warehouse.

A matriz de barramento é geralmente organizada em duas dimensões: os processos de negócios (ou áreas de negócio) e as dimensões que descrevem aspectos desses processos. Cada célula da matriz representa uma interseção entre um processo de negócio e uma dimensão. Nessas células, são listados os principais indicadores de desempenho (KPIs) ou métricas que são relevantes para entender e analisar o desempenho da empresa em relação a esse processo e dimensão específicos.

Business Process / Event	Time	Customer	Service	Rate Category	Local Svc Provider	Calling Party	Called Party	Long Dist Provider	Internal Organization	Employee	Location	Equipment Type	Supplier	Item Shipped	Account Status
Customer Billing	X	X	X	X	X			X			X				X
Service Orders	X	X	X		X			X	X	X	X	X			X
Trouble Reports	X	X	X		X	X		X	X	X	X	X	X	X	X
Yellow Page Ads	X	X		X		X			X	X	X				X
Customer Inquiries	X	X	X	X	X	X		X	X	X	X				X
Promotions & Communication	X	X	X	X	X	X		X	X	X	X	X	X	X	X
Billing Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X
Network Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X
Customer Inventory	X	X	X	X	X			X	X		X	X	X	X	X
Network Inventory	X		X						X	X	X	X	X	X	
Real Estate	X								X	X	X	X			
Labor & Payroll	X								X	X	X				
Computer Charges	X	X	X		X			X	X	X	X	X	X	X	
Purchase Orders	X								X	X	X	X	X	X	
Supplier Deliveries	X								X	X	X	X	X	X	

O propósito fundamental da matriz de barramento é alinhar as necessidades de análise de negócios com a estrutura do data warehouse. Ela ajuda a identificar quais dimensões são relevantes para cada processo de negócio e quais métricas são importantes para avaliar o desempenho desses processos. Isso é crucial para garantir que o data warehouse seja projetado de forma a fornecer informações significativas e úteis para os usuários finais.

Além disso, a matriz de barramento facilita a priorização de esforços de modelagem e desenvolvimento, pois ajuda a identificar quais áreas de negócio e dimensões são mais críticas e devem receber atenção prioritária durante o projeto do *data warehouse*.

Ao fornecer uma visão abrangente das inter-relações entre os processos de negócio e as dimensões, a matriz de barramento também promove a comunicação e o alinhamento entre as diferentes partes interessadas no projeto de *data warehousing*, incluindo analistas de negócios, especialistas em domínio, arquitetos de dados e desenvolvedores.

QUESTÃO DE PROVA

(FGV/TCE SP/2023) O TCE SP está implementando um Data Warehouse utilizando uma abordagem incremental, ou seja, constrói um Data Mart para um setor e depois para outro setor, compartilhando Dimensões.

A ferramenta de projeto, que representa as áreas do negócio e as dimensões associadas, utilizada para apoiar a implementação de modelos dimensionais de áreas de negócio distintos compartilhando dimensões padronizadas em um Data Warehouse Corporativo é o(a):

a) Data Lake;
b) Pipeline de dados;
c) Regras de Associação;
d) Matriz de Barramento;
e) Processamento distribuído.

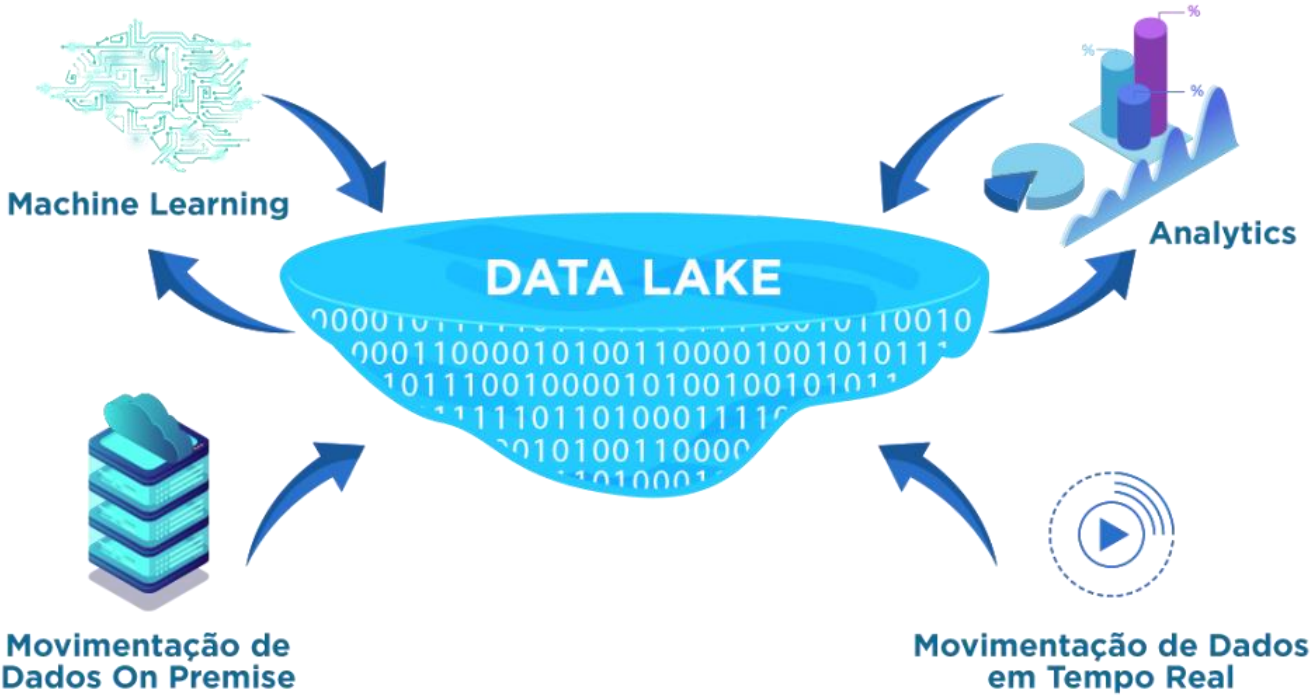
Comentários:

A ferramenta que apresenta as áreas de negócio e associa elas às dimensões dos dados, apoiando a implementação, é chamada de Matriz de Barramento. *(Gabarito: Letra D)*

Data Lake

Data Lake, ou lago de dados, são repositórios centralizados que permitem armazenar dados estruturados e não estruturados em qualquer escala. Ou seja, é possível o armazenamento dos dados sem a necessidade prévia de transformação e adaptação do dado ao padrão do repositório.

Graças a essas características, os Data Lakes oferecem uma alta escalabilidade, podendo receber um volume muito grande de dados em um curto período de tempo. Seu principal objetivo é fornecer subsídios para processamentos de Big Data e uma base de dados para ferramentas de Machine Learning.



Podemos operacionalizar o Data Lake de duas formas distintas:

- Como um intermediário, trabalhando como uma espécie de repositório para os dados da entidade. Nesse caso, as ferramentas de ETL ou ELT serão aplicadas sobre ele, carregando os dados no Data Warehouse
- Como um fim, recebendo uma camada de governança e metadados, que servirão para as ferramentas de análise de dados interagirem diretamente com esse repositório. Nesse caso, o Data Lake atua num formato parecido com o Warehouse - por isso recebe o nome de Data Lakehouse.

Na prática, ambas as formas são usadas simultaneamente, de forma que o Data Lake atua diretamente como um repositório final, fornecendo informações para ferramentas que necessitem

Para não virar uma "bagunça" completa, é indicado que o desenvolvimento de um Data Lake passe por quatro etapas diferentes, evitando transformar o lago em um pântano de dados (Data Swamp).

Estágio 1 – Inicial: captura de dados

No primeiro nível, o Data Lake é construído separado dos principais sistemas de TI e serve como um ambiente de "captura pura" de baixo custo e escalável. O Data Lake serve como uma fina camada de gerenciamento de dados dentro da pilha de tecnologia da empresa que permite que dados brutos sejam armazenados indefinidamente antes de serem preparados para uso em ambientes de computação. As organizações podem implantar o Data Lake com efeitos mínimos na arquitetura existente. Uma governança forte, incluindo classificação rigorosa de dados, é necessária durante essa fase inicial se as empresas desejarem evitar a criação de um pântano de dados (Data Swamp).

Estágio 2 – Estável: ambiente de Ciência de Dados

Neste próximo nível, as organizações podem começar a usar mais ativamente o Data Lake como uma plataforma para experimentação. Os Cientistas de Dados têm acesso fácil e rápido aos dados – e podem se concentrar mais na execução de experimentos com dados e na análise de dados, em vez de se concentrarem apenas na coleta e aquisição de dados. Neste "sandbox", eles podem trabalhar com dados inalterados para criar protótipos para programas analíticos. Eles podem implantar uma variedade de ferramentas comerciais e de código aberto ao lado do Data Lake para criar os ambientes de teste necessários.

Estágio 3 – Integrado: integração com Data Warehouses

No próximo nível, os Data Lakes estão começando a ser integrados aos EDWs existentes. Aproveitando os baixos custos de armazenamento associados a um Data Lake, as empresas podem armazenar dados "frios" (raramente usados ou inativos). Elas podem usar esses dados para gerar insights sem pressionar ou exceder as limitações de armazenamento, ou sem precisar aumentar drasticamente o tamanho dos Data Warehouses tradicionais. Enquanto isso, as empresas podem manter a extração de dados relacionais de alta intensidade em EDWs existentes, que têm o poder de lidar com eles. Elas podem migrar tarefas de extração e transformação de baixa intensidade para o Data Lake – por exemplo, uma pesquisa do tipo "agulha no palheiro", na qual os Cientistas de Dados precisam varrer bancos de dados para consultas não suportadas por estruturas de índices tradicionais.

Estágio 4 – Implementado: componente estruturante

Uma vez que as empresas cheguem a esse estágio de lançamento e desenvolvimento, é muito provável que grande parte das informações que circulam pela empresa estejam passando pelo

Data Lake. O Data Lake se torna uma parte essencial da infraestrutura de dados, substituindo Data Marts existentes ou armazenamentos de dados operacionais e permitindo o fornecimento de dados como um serviço. As empresas podem aproveitar ao máximo a natureza distribuída da tecnologia de Data Lake, bem como sua capacidade de lidar com tarefas de uso intensivo de computação, como aquelas exigidas para conduzir análises avançadas ou implantar programas de aprendizado de máquina (Machine Learning). Algumas empresas podem decidir criar aplicativos com uso intensivo de dados na parte superior do Data Lake – por exemplo, um painel de gerenciamento de desempenho. Ou elas podem implementar interfaces de programação de aplicativos para que possam combinar perfeitamente os insights obtidos dos recursos do Data Lake com insights obtidos de outros aplicativos.

ESQUEMATIZANDO

Fases do Data Lake

1. Inicial - DL age puramente funcionando como uma ferramenta de captura de dados

2. Estável - Dados consolidados, usados primariamente por cientistas de dados

3. Integrado - Chega a um nível de estruturação que permite a integração com os DWs

4. Implementado - Atua como elemento estruturante na entidade, substituindo repositórios padrão

(CEBRASPE/EMPREL/2023) No projeto de arquitetura de um data lake, a primeira etapa que deve estar prevista é a

a) criação de um ambiente virtual de captura de dados.
b) concessão de acesso ao banco de dados (somente leitura) aos cientistas de dados, para estes realizarem experimentos e testes.
c) integração dos dados do data lake aos data warehouses da empresa.
d) atualização dos dados dos repositórios de dados da empresa a partir dos dados já consolidados disponíveis no data lake.
e) visualização de dados e otimização das principais consultas.

Comentários:

Vamos analisar cada alternativa.

a) Certo. Na primeira etapa, o Data Lake funciona apenas como uma ferramenta de captura de dados da empresa, até "encher o lago".
b) Errado. O acesso aos cientistas de dados passa a se dar a partir da segunda etapa.
c) Errado. A integração aos Data Warehouses se dá a partir da terceira etapa.
d) Errado. A integração completa, com o Data Lake funcionando como um repositório puro de dados à empresa, só é possível na quarta etapa.
e) Errado. Não conseguimos ter visualização de dados na primeira etapa.

Dessa forma, correta a letra A. *(Gabarito: Letra A)*

Data Mesh

O Data Mesh, ou Malha de Dados, é um assunto relativamente novo, começando a aparecer em editais específicos de TI em 2023 — mas ainda sem cobranças. Ela é uma estrutura arquitetônica que resolve desafios avançados de segurança de dados por meio de propriedade distribuída e descentralizada.

As organizações têm várias fontes de dados de diferentes linhas de negócios que devem ser integradas para análise. Uma arquitetura de malha de dados une de forma efetiva as fontes de dados diferentes e as vincula por meio de diretrizes de governança e compartilhamento de dados gerenciados centralmente. As funções de negócios podem manter o controle sobre como os dados compartilhados são acessados, quem os acessa e em quais formatos são acessados. Uma malha de dados adiciona complexidades à arquitetura, mas também traz eficiência ao melhorar o acesso aos dados, a segurança e a escalabilidade.

A arquitetura do Data Mesh é orientada em 4 princípios primários:

Arquitetura orientada por domínio distribuído

Na abordagem de malha de dados, a ideia é dividir a responsabilidade de lidar com os dados de uma empresa em diferentes áreas ou equipes, cada uma focada em um aspecto específico do negócio. Essas equipes, chamadas de equipes de domínio, cuidam de coletar, arrumar e disponibilizar os dados relacionados à sua área. Em vez de todos os dados irem para um lugar central, cada equipe cuida dos seus próprios dados de uma forma que seja fácil de entender e usar. Por exemplo, em uma loja de roupas, pode haver uma equipe que cuida dos dados sobre os produtos vendidos e outra equipe que se dedica a entender o comportamento dos clientes no site da loja.

Dados como um produto

Para que uma implementação de malha de dados seja bem-sucedida, todas as equipes de domínio precisam aplicar o pensamento do produto aos conjuntos de dados que fornecem. Elas devem considerar seus ativos de dados como seus produtos e o restante das equipes de negócios e dados da organização como seus clientes.

Para a melhor experiência do usuário, os produtos de dados de domínio devem ter as seguintes qualidades básicas.

- Descobríveis: Cada conjunto de dados fica registrado em um catálogo central, para que todo mundo consiga achar facilmente.
- Endereçáveis: Cada conjunto de dados tem um endereço próprio, para que as pessoas consigam acessá-lo de forma programática. Esse endereço segue padrões de nomenclatura que são decididos pela empresa.
- Confiáveis: Os conjuntos de dados têm que ser atualizados de acordo com a realidade. Por exemplo, a equipe que cuida dos pedidos só publica os dados depois de verificar se o endereço e o telefone do cliente estão certos.
- Autodescritivos: Todos os conjuntos de dados têm instruções claras sobre como usá-los e o que significam, seguindo as regras de nomenclatura da empresa.

Infraestrutura de dados de autoatendimento

Uma arquitetura de dados distribuídos exige que cada domínio configure seu próprio pipeline de dados para limpar, filtrar e carregar seus próprios produtos de dados. Uma malha de dados introduz o conceito de uma plataforma de dados de autoatendimento para evitar a duplicação de esforços. Os engenheiros de dados configuram tecnologias para que todas as unidades de negócios possam processar e armazenar seus produtos de dados. A infraestrutura de autoatendimento permite, assim, uma divisão de responsabilidades. As equipes de engenharia de dados gerenciam a tecnologia enquanto as equipes de negócios gerenciam os dados.

Governança de dados federados

As arquiteturas de malha de dados implementam a segurança como uma responsabilidade compartilhada dentro da organização. A liderança determina padrões e políticas globais que você pode aplicar em todos os domínios. Ao mesmo tempo, a arquitetura de dados descentralizada permite um alto grau de autonomia sobre padrões e implementação de políticas dentro do domínio.

Qual escolher?

Novamente, assim como no ETL vs ELT, temos um dilema: qual abordagem deve ser escolhida? Vamos ver uma comparação primariamente, e depois discutiremos sobre.

Tópico	Data Warehouse	Data Lake
Tipo de Dados	Estruturado	Não estruturado, semiestruturado e estruturado
Processamento	Em lotes	Em lotes e em fluxo
Escalabilidade	Baixa	Alta
Uso Principal	Business Intelligence	Machine Learning
Pipeline	ETL	ELT
Latência	Alta	Baixa
Segurança	Alta	Baixa
Integração	Requer transformação prévia dos dados antes de serem integrados	Integra-se diretamente a diversas fontes, como dispositivos móveis e sistemas de gestão empresarial

Em geral, o Data Warehouse é estruturado focando na análise de negócios, no Business Intelligence em si. Ele alia uma arquitetura mais estruturada, usando dados estruturados e uma pipeline ETL. Dessa forma, apesar da maior latência e lentidão na sua criação, ele é uma alternativa mais segura, principalmente quando tratamos de dados mais delicados.

Já o Data Lake funciona tanto como um repositório geral de dados, coletando os mais diversos tipos de dados não estruturados e semiestruturados, além de se integrar com bancos de dados transacionais, de forma a receber dados estruturados. Seu objetivo é alimentar ferramentas de inteligência artificial e mineração de dados, e, paralelamente, outros Data Warehouses.

(FGV/TCU/2022) Uma organização deseja implementar um pipeline de dados e está avaliando a opção mais adequada para o seu contexto de operação. Em torno de 40% dos dados consumidos pela organização se encontram em planilhas eletrônicas que contêm dados sensíveis, produzidas semanalmente por suas unidades de negócio. Os outros 60% dos dados se encontram em alguns bancos de dados relacionais de sistemas de produção da organização. O tamanho da base é de moderado a pequeno, mas existe a necessidade de conformidade com normas de privacidade e confidencialidade dos dados. O objetivo do pipeline é fornecer insumos para um departamento que realiza análises de dados com métodos não supervisionados de aprendizagem de máquina para elaborar relatórios periódicos mensais. A organização está avaliando a construção de um Armazém de Dados (ETL) ou de um Lago de Dados (ELT).

A proposta de modelo adequada e corretamente justificada é:

a) Armazém de Dados. Ambos os modelos são adequados, mas Lago de Dados tem maior latência até a carga (L) e custo maior;
b) Armazém de Dados. Esse modelo possui menor latência até a carga (L) e, ao contrário do Lago de Dados, opera de forma eficiente com dados relacionais;
c) Armazém de Dados. O processo ETL é mais adequado para o tratamento dos dados sensíveis e os casos de uso são bem conhecidos;
d) Lago de Dados. Esse modelo possui menor latência até a carga (L) e permite a extração de dados semiestruturados e não estruturados;
e) Lago de Dados. Esse modelo não necessita de hardware especializado e, ao contrário do Armazém de Dados, possibilita tarefas de aprendizado de máquina.

Comentários:

Os dados tratados na questão são dados sensíveis, isso é, possuem uma política de segurança e tratamento exigíves deles - conforme o trecho "[...] existe a necessidade de conformidade com normas de privacidade e confidencialidade dos dados". Nesse sentido, precisaremos implementar um repositório estruturado, um Data Warehouse, que é alimentado por um processo de ETL. Com isso em mente, vamos analisar as alternativas:

a) Errado. O lago de dados tem latência de carga menor, e não maior.
b) Errado. O armazém de dados tem maior latência na carga de dados, e, além disso, o Lago de Dados opera de forma menos eficiente com dados relacionais, que são estruturados, se comparado com o Armazém.
c) Certo. Exatamente isso - devemos optar pelo Data Warehouse, que utiliza o ETL.
d) e e) Errado. Devemos optar pelo Armazém de Dados, e não pelo Lago.

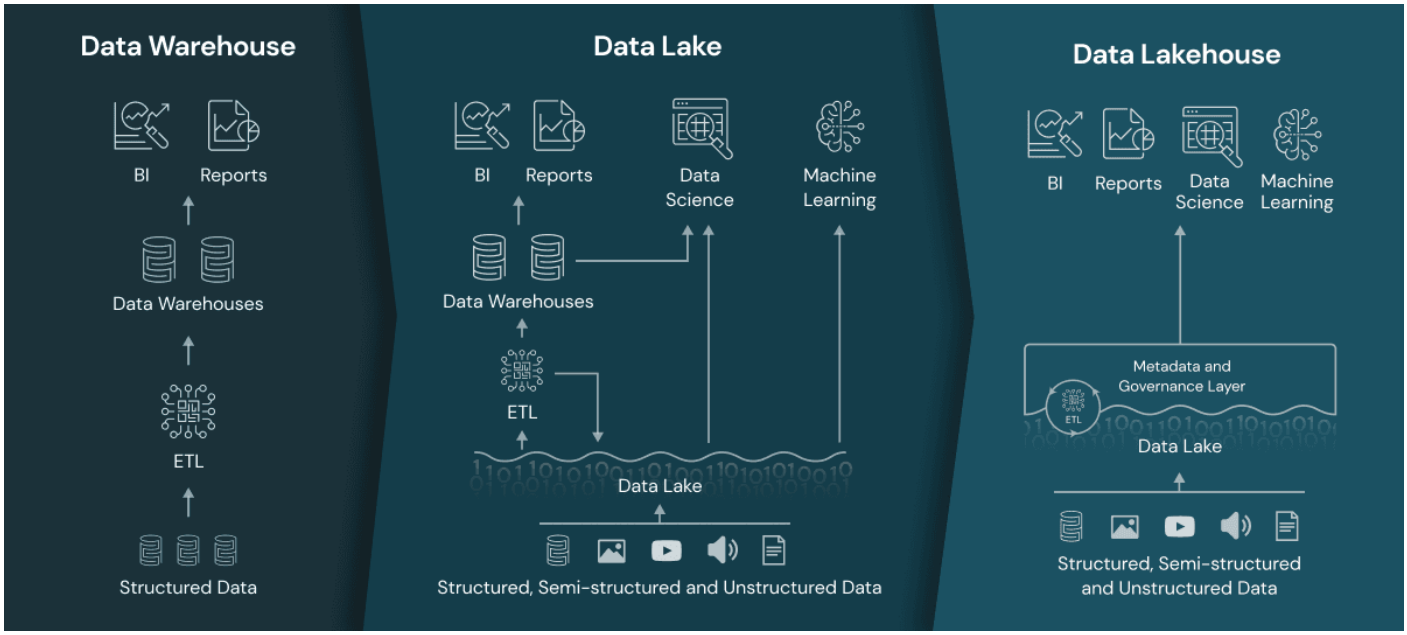
Dessa forma, correta a letra C. *(Gabarito: Letra C)*

Data LakeHouse

Tanto o Data Warehouse quanto o Data Lake possuem suas vantagens específicas, e casos de uso válidos. Recentemente, uma nova "versão" de armazenamento vem sendo adotada por diversas entidades, consistindo de uma mistura das duas abordagens, juntando os pontos positivos de ambos - criando o que é chamado de Data LakeHouse.

De forma geral, uma *Data LakeHouse* remove os silos entre um Data Lake e um Data Warehouse. Isso significa que os dados podem ser facilmente movidos entre o armazenamento flexível e de baixo custo de um Data Lake para um Data Warehouse e vice-versa, fornecendo acesso fácil às ferramentas de gerenciamento de um Data Warehouse para implementação de esquema e governança, geralmente alimentados por machine learning e inteligência artificial para limpeza de dados.

O resultado cria um repositório de dados que integra a coleção acessível e não estruturada de Data Lakes e a preparação robusta de um Data Warehouse. Ao fornecer o espaço para coletar de fontes de dados selecionadas enquanto usa ferramentas e recursos que preparam os dados para uso comercial, um Data Lakehouse acelera os processos. De certa forma, Data Lakehouses são Data Warehouses reformulados para nosso mundo moderno orientado por dados.



Essa estrutura é possibilitada graças a uma camada compartilhada de metadados e governança. Esta camada é responsável por gerenciar e organizar os dados dentro do sistema, assegurando que os dados sejam facilmente encontráveis, utilizáveis e seguros - além de permitir que os dados fluam entre o Data Lake e o Data Warehouse sem problemas.

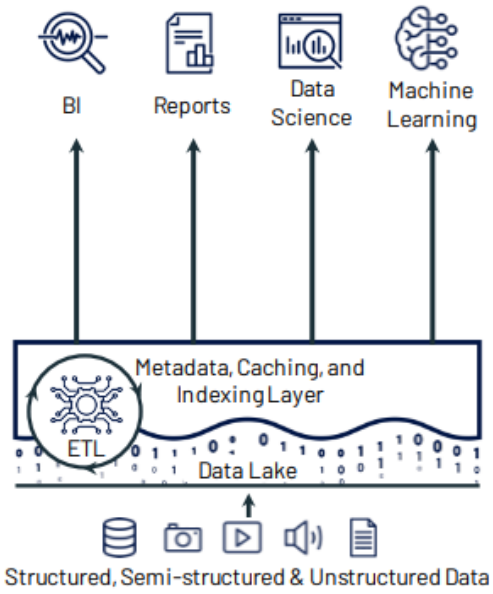
Uma das formas mais comuns de estruturação de um Data LakeHouse é uma arquitetura em três camadas lógicas. A primeira camada funciona para a ingestão de dados. Seja em lote ou em fluxo, os dados são alocados inicialmente no Data Lake, em um formato brutom, passando em seguida por uma verificação de integridade. Ferramentas como o Data Bricks (Azure) permitem a conversão desses dados em tabelas, habilitando processos de detecção de valores ausentes ou inesperados.

Depois de feita a verificação, começa o processo da segunda camada lógica, com a coleta e refinamento. Cientistas de dados e profissionais de Machine Learning usualmente usam essa camada para começar a combinar ou criar novos recursos e concluir a limpeza dos ados. Depois de concluída a etapa, os dados estarão prontos para serem integrados e reorganizados em dados estruturados, como tabelas, para atender às necessidades do negócio.

Por fim, a camada final fornece um serviço de dados, com dados limpos e enriquecidos para os usuários finais. As tabelas finais aqui deve ser projetadas para fornecer dados para todos os casos de uso previsíveis, utilizando um modelo de governança unificado, que permita acompanhar a linhagem dos dados de volta à origem.

Dessa forma, um LakeHouse oferece:

- Acesso aberto e direto aos dados armazenados em formatos de dados padrão.
- Protocolos de indexação otimizados para machine learning e ciência de dados.
- Baixa latência de consulta e alta confiabilidade para BI e análise avançada.


BI
Reports
Data Science
Machine Learning
Metadata, Caching, and Indexing Layer
ETL
Data Lake
Structured, Semi-structured & Unstructured Data

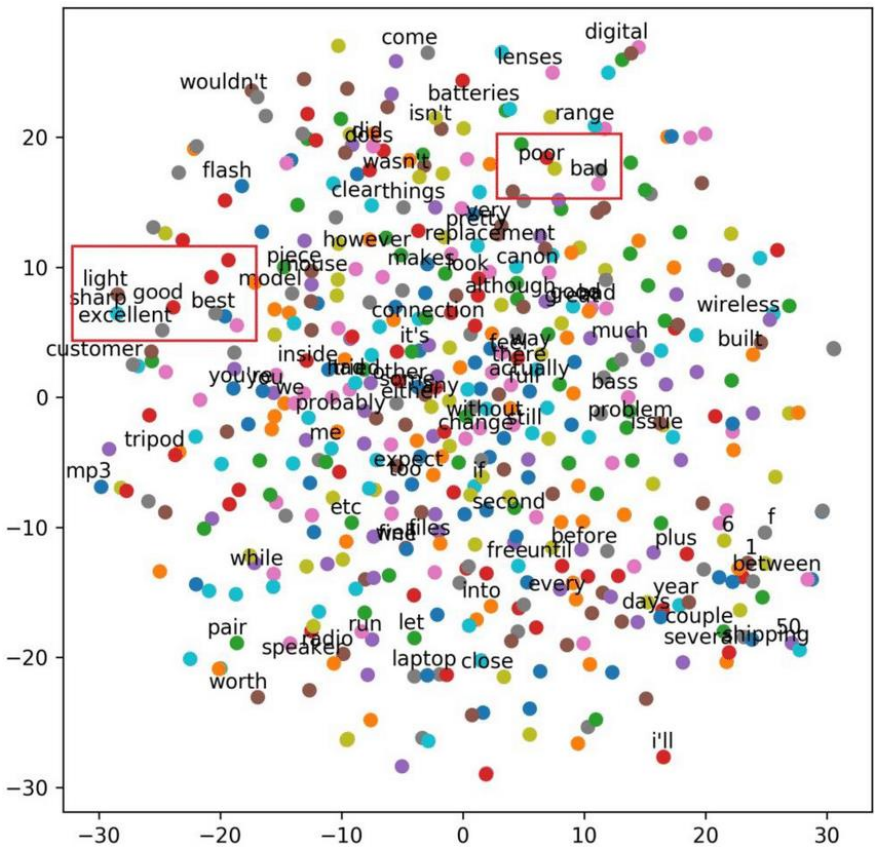
VECTOR STORAGES

Vetorização

Antes de falarmos especificamente desse repositório, precisamos falar sobre o que é vetorização. Esse é um tópico que é abordado mais profundamente quando se estuda Processamento de Linguagem Natural - mas vamos trazer uma explicação breve para que você possa entender do que se trata, e porque estamos vendo esse repositório de forma isolada.

Quando tratamos de dados textuais, nem sempre os computadores fazem um bom processamento dessas informações. Na verdade, é preferível que se trabalhe com números - os computadores, assim como eu, lidam melhor com número do que com textos e palavras. Para capturar os significados das palavras e, em alguns casos, do contexto que elas se inserem, usamos diversas técnicas para transformar letras, palavras e textos em números, ou vetores de números.

De forma geral, um vetor é uma representação matemática de palavras em um espaço vetorial, utilizado para capturar o significado e as relações semânticas entre essas palavras. Essa técnica permite que as palavras sejam manipuladas por algoritmos de aprendizado de máquina de maneira mais eficiente do que representações tradicionais, como bag-of-words.

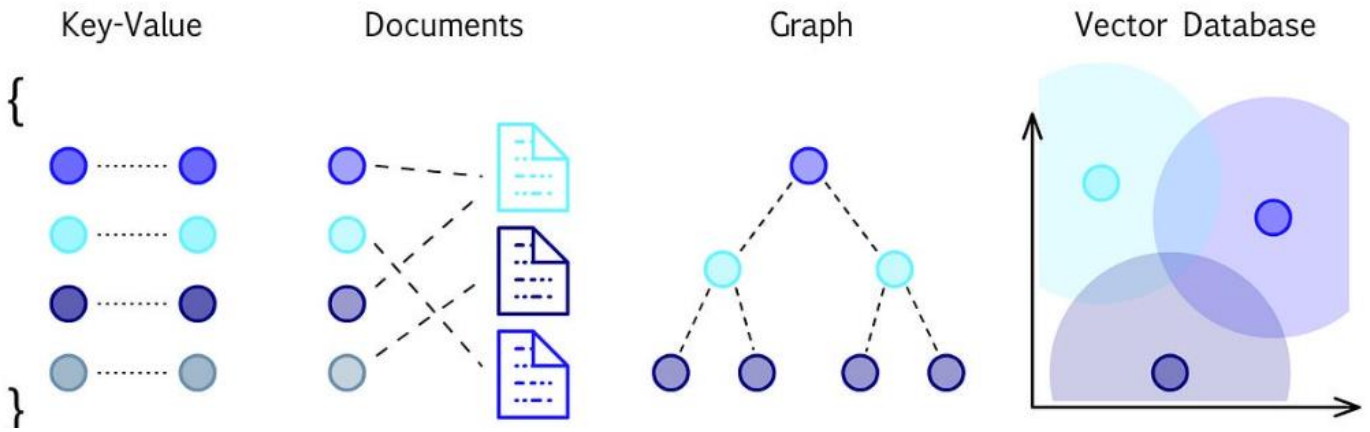


Armazenando Vetores

Vector Storages, também conhecidos como Vector Databases ou Armazenamentos de Vetores, são sistemas de banco de dados projetados especificamente para armazenar e gerenciar vetores de alta dimensão, isso é, com muitas variáveis. Esses vetores são frequentemente utilizados em aplicações de aprendizado de máquina e inteligência artificial, especialmente em tarefas que envolvem buscas e comparações baseadas em similaridade, como recomendação de produtos, recuperação de informação, reconhecimento de imagens, e processamento de linguagem natural.

A estrutura de um Vector Storage é otimizada para armazenar e manipular vetores. A estrutura típica de um Vector Storage inclui:

- Espaço para armazenamento de vetores, envolvendo indexação dos vetores, onde armazenamos os vetores em estruturas de dados otimizadas para busca, como as árvores de k-dimensões, e técnicas de compressão e redução de dimensionalidade.
- Suporte a consultas por similaridade, retornando palavras iguais ou aproximadamente iguais quando realizamos consultas - usualmente através de análises de distâncias no espaço vetorial, usando métricas como a distância euclidiana, de cosseno ou de Manhattan



A chave aqui está justamente nas consultas por similaridade, atuando como motor de pesquisas ou até mesmo em sistemas de recomendação. Por exemplo, você pesquisa "x-burguer" no *aiFudi* - mas não há nenhum disponível. O algoritmo pode entender que "x-salada" é uma aproximação válida, e sugerir esse produto com base nessa aproximação.

QUESTÕES COMENTADAS

01. (CEBRASPE/CNPq/2024) Acerca da análise de dados para tomada de decisão, da capacitação tecnológica e da competitividade, julgue o item a seguir.

Business intelligence (BI) pode ser definido corretamente como um conjunto de tecnologias que dão suporte a decisões gerenciais por meio de informações internas e externas às organizações, tendo grande impacto na estratégia corporativa, na performance e na competitividade.

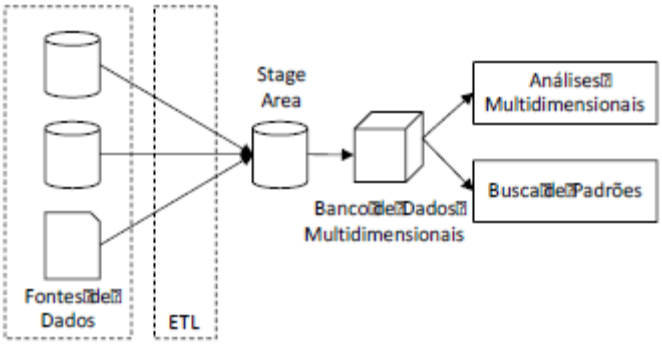
Comentários:

Certinho! O BI é o conjunto de técnicas, ferramentas e utilidades destinadas a extrair informações não triviais, isso é, não tão óbvias, para impactar na gestão e gerência de organizações.

Gabarito: Certo

---

02. (FGV/DPE RS/2023) Observe a seguinte arquitetura básica de uma solução de Business Intelligence implementada na empresa CleverBI.



O elemento arquitetural da solução de BI da CleverBI implementado por meio de operações OLAP, como slice, rotate, drill-down e drill-up, é o:

a) ETL;
b) Stage Area;
c) Banco de Dados Multidimensionais;
d) Análises Multidimensionais;
e) Busca de Padrões.

Comentários:

As ferramentas que implementam operações OLAP (que estudaremos mais profundamente em algumas aula) estão na parte de análise de dados, na última coluna da estrutura.

Gabarito: Letra D

---

03. (FGV/CGE SC/2023) Avalie se os componentes de um Data Warehouse incluem:

I. Sistemas de origem.
II. Infraestrutura de ETL (Extraction-transformation-load).
III. Data Warehouse.
IV. Aplicações de Front-end para o usuário final.

Estão corretos os itens

a) I e II, apenas.
b) III e IV, apenas.
c) I, II e III, apenas.
d) II, III e IV, apenas.
e) I, II, III e IV.

Comentários:

Numa infraestrutura de Data Warehouse temos todos os elementos apontados pela lista. Os sistemas de origem são os bancos de dados transacionais, usualmente relacionais, de onde são retirados os dados. A infraestrutura ETL leva esses dados até o Data Warehouse que, por fim, é acessado por ferramentas *front-end* pelo usuário final. *Front-end* é um termo que define aplicações de interface de usuário, aquelas com o qual o usuário final interage (digitando, apertando botões etc.).

Gabarito: Letra E

---

04. (FGV/ISS RJ/2023) No ambiente de Data Warehousing da SMFP/RJ, os dados são extraídos de diversas fontes e integrados para apoiar a fiscalização de rendas por meio do desenvolvimento de diversos artefatos de dados. Antes de começar as análises nesse ambiente, o fiscal de rendas Inácio fez consultas sobre o seu conteúdo, por exemplo:

- DW_Tributos, banco de dados analítico do tipo Data Warehouse que integra dados sobre os tributos arrecadados do Município do Rio de Janeiro.
- TP_EMPRESA, Caractere, 1, atributo que descreve o tipo da empresa contendo os seguintes valores: M - MEI ou S - Simples Nacional, e faz parte da tabela TB_EMPRESA.
- RL_Sit_Fiscal, relatório sobre a situação fiscal das empresas do Município do Rio de Janeiro.

O componente do ambiente de Data Warehousing, utilizado por Inácio, que foi desenvolvido para apoiar consultas sobre a descrição de cada artefato de dado, é:

a) ETL;
b) Métricas;
c) Dashboard;
d) Dimensões de Dados;
e) Repositório de Metadados.

Comentários:

Vimos mais sobre esse tipo de dado nas aulas passadas - mas ele não deixa de existir aqui nos Data Warehouses. Os metadados são os dados que descrevem toda a estrutura necessária para o Data Warehouse, tabelas, tipos de dado, relações etc. Como o Data Warehouse trabalha com grandes quantidades de dados, temos um repositório de metadados que lida com esses dados. É justamente esse componente descrito pela questão - um componente responsável por descrever os objetos do Data Warehouse.

Gabarito: Letra E

---

05. (CEBRASPE/MPE RO/2023) Em uma solução de BI (Business Intelligence), os dashboards são

a) fontes de dados.
b) insights.
c) usados no ETL.
d) armazéns de dados.
e) modelos semânticos de dados.

Comentários:

Dashboards são representações gráficas, visuais, que congregam as informações angariadas no processo do BI para auxiliar na decisão - fornecendo, portanto, ***insights*** para a entidade. Eles se parecem com isso:

Gabarito: Letra B

---

06. (FGV/PREF. SJC/2024) Relacione as características de um data warehouse listadas a seguir com suas descrições, conforme estabelecido por William Inmon.

1. Orientados a Assunto.
2. Integração.
3. Não Volátil.
4. Variante no Tempo.

( ) O foco de um data warehouse na mudança ao longo do tempo é essencial para descobrir tendências e identificar padrões e relacionamentos ocultos nos negócios, para isso os analistas precisam de grandes quantidades de dados. Isso contrasta muito com o processamento de transações on-line onde os requisitos de desempenho exigem que os dados históricos sejam movidos para arquivos.
( ) Os data warehouses devem colocar dados de fontes diferentes em um formato consistente. Eles devem resolver problemas como nomear conflitos e inconsistências entre unidades de medida.
( ) Significa que, uma vez inseridos no data warehouse, os dados não devem mudar. Essa característica é lógica porque o propósito de um data warehouse é permitir que um analista analise o que ocorreu no passado.
( ) Os data warehouses são projetados para ajudar os profissonais a analisar grandes volumes de dados. Por exemplo, para saber mais sobre os dados de vendas de uma empresa, o analista pode construir um data warehouse que concentre a venda. Usando esse data warehouse, ele poderá responder perguntas como "Quem foi nosso melhor cliente

para este item no ano passado?" ou "Quem provavelmente será nosso melhor cliente no próximo ano?"

A relação correta, na ordem dada, é:

a) 1 – 2 – 3 – 4.
b) 2 – 1 – 4 – 3.
c) 4 – 2 – 3 – 1.
d) 3 – 4 – 1 – 2.
e) 2 – 3 – 1 – 4.

Comentários:

Vamos fazer as relações.

( ) O foco de um data warehouse na mudança ao longo do tempo é essencial para descobrir tendências e identificar padrões e relacionamentos ocultos nos negócios, para isso os analistas precisam de grandes quantidades de dados. Isso contrasta muito com o processamento de transações on-line onde os requisitos de desempenho exigem que os dados históricos sejam movidos para arquivos.

Trata-se do princípio da variância no tempo, que garante uma série de dados históricos que auxiliarão na tomada de decisões. (4)

( ) Os data warehouses devem colocar dados de fontes diferentes em um formato consistente. Eles devem resolver problemas como nomear conflitos e inconsistências entre unidades de medida.

Esse item aborda o princípio da integração, que prega uma integração consistente entre diversas fontes de dados. (2)

( ) Significa que, uma vez inseridos no data warehouse, os dados não devem mudar. Essa característica é lógica porque o propósito de um data warehouse é permitir que um analista analise o que ocorreu no passado.

Aqui temos a não volatilidade, que define que os dados, uma vez que carregados, não podem ser alterados - somente excluídos. (3)

( ) Os data warehouses são projetados para ajudar os profissonais a analisar grandes volumes de dados. Por exemplo, para saber mais sobre os dados de vendas de uma empresa, o analista pode construir um data warehouse que concentre a venda. Usando esse data warehouse, ele poderá responder perguntas como "Quem foi nosso melhor cliente

para este item no ano passado?" ou "Quem provavelmente será nosso melhor cliente no próximo ano?"

Por fim, o último item aborda o tópico da orientação a assunto, que orienta a construção do DW com orientações afins específicas. (1)

Temos, portanto, 4-2-3-1.

Gabarito: Letra C

---

07. (FGV/PREF. SJC/2024) Com relação ao ETL, ELT e suas tecnologias, avalie as afirmativas a seguir e assinale V para a verdadeira e F para a falsa.

( ) ELT utiliza fluxos de trabalho de análise de dados e de aprendizado de máquina. O ELT é frequentemente usado por uma organização para: Extrair dados de sistemas legados, limpar os dados para melhorar sua qualidade e carregar dados em um banco de dados de destino. O ELT transforma dados no trânsito.
( ) ETL copia ou exporta os dados dos locais de origem, mas, em vez de carregá-los em uma área de preparação para transformação, ele carrega os dados em estado brutos diretamente no armazenamento de dados no destino para serem transformados conforme necessário. O ETL não transforma nenhum dado no trânsito.
( ) A ordem das etapas não é a única diferença entre ETL e ELT. No ELT, o armazenamento de dados de destino pode ser um armazém de dados, mas, mais frequentemente, é um data lake, que é um armazenamento central grande projetado para manter tanto dados estruturados quanto não estruturados em grande escala.

As afirmativas são, respectivamente,

a) F – V – F.
b) F – F – V.
c) F – V – V.
d) V – F – V.
e) V – V – F.

Comentários:

Vamos analisar cada uma das afirmativas.

(F) ELT utiliza fluxos de trabalho de análise de dados e de aprendizado de máquina. O ELT é frequentemente usado por uma organização para: Extrair dados de sistemas legados, limpar os dados para melhorar sua qualidade e carregar dados em um banco de dados de destino. O ELT transforma dados no trânsito.

Falso. A afirmativa descreve o ETL, que faz a limpeza e transformação no trânsito, e não o ELT.

> (F) ETL copia ou exporta os dados dos locais de origem, mas, em vez de carregá-los em uma área de preparação para transformação, ele carrega os dados em estado brutos diretamente no armazenamento de dados no destino para serem transformados conforme necessário. O ETL não transforma nenhum dado no trânsito.

Falso. Novamente, uma inversão de conceitos. A afirmativa descreve o ELT, não o ETL.

> (V) A ordem das etapas não é a única diferença entre ETL e ELT. No ELT, o armazenamento de dados de destino pode ser um armazém de dados, mas, mais frequentemente, é um data lake, que é um armazenamento central grande projetado para manter tanto dados estruturados quanto não estruturados em grande escala.

Verdadeiro. Exatamente o que trouxe a vocês, além da clara diferença na ordem das etapas, temos outras diferenças - como os repositórios de interação.

Temos, portanto, F-F-V.

Gabarito: Letra B

---

08. (FGV/PREF. SJC/2024) Com relação ao Data Warehousing e ao Business Intelligence, avalie as afirmativas a seguir e assinale V para a afirmativa verdadeira e F para a falsa.

> ( ) Eles têm como meta construir e manter o ambiente técnico e os processos técnicos e de negócios necessários para fornecer dados integrados em apoio às funções operacionais, requisitos de conformidade e atividades de inteligência de negócios.
> ( ) Ambos visam apoiar e permitir análises de negócios e tomadas de decisões mais eficazes por parte dos trabalhadores do conhecimento.
>
> ( ) O Data Warehousing concentra-se em permitir um contexto de negócios histórico e integrado em dados operacionais, aplicando regras de negócios e mantendo relacionamentos de dados de negócios apropriados. O armazenamento de dados também inclui processos que interagem com repositórios de metadados.

As afirmativas são, respectivamente,

a) F – F – F.
b) F – F – V.
c) F – V – V.

d) V – F – V.
e) V – V – V.

Comentários:

Vamos aos itens.

> (V) Eles têm como meta construir e manter o ambiente técnico e os processos técnicos e de negócios necessários para fornecer dados integrados em apoio às funções

Verdadeiro. Apesar de eu acreditar que tenham cortado um "em apoio às funções decisórios", a afirmativa está correta. O DW e o BI constroem e mantêm um ambiente técnico voltado ao apoio às funções decisórias da entidade.

> (V) Ambos visam apoiar e permitir análises de negócios e tomadas de decisões mais eficazes por parte dos trabalhadores do conhecimento.

Verdadeiro. Esse é o propósito da infraestrutura de BI.

> (V) O Data Warehousing concentra-se em permitir um contexto de negócios histórico e integrado em dados operacionais, aplicando regras de negócios e mantendo relacionamentos de dados de negócios apropriados. O armazenamento de dados também inclui processos que interagem com repositórios de metadados.

Verdadeiro. O objetivo do DW é criar um repositório de dados que acumulará uma evolução dos dados ao longo do tempo - o que chamamos de dados históricos.

Todas as afirmativas são verdadeiras.

Gabarito: Letra E

---

09. (CEBRASPE/SPELAN RR/2023) No que se refere às características de um banco de dados relacional, julgue o item que se segue.

Em contraste com os modelos relacionais, os modelos dimensionais, ou modelos de dados dimensionais kimball — modelos de dados baseados na técnica desenvolvida por Ralph Kimball —, são estruturas normalizadas projetadas para recuperar dados de um data warehouse.

Comentários:

Vou trazer a tabelinha pra você se lembrar:

TOP-DOWN / INMON	BOTTOM-UP / KIMBALL
Data Warehouse → Data Marts	Data Marts → Data Warehouse
Bancos de Dados Relacionais (adaptados)	Bancos de Dados Dimensionais
Modelos normalizados	Opcionalmente normalizados
Voltado para profissionais da TI	Voltado para usuários finais
Consultas nos Data Marts	Consultas no Data Warehouse

Então veja, o modelo normalizado é o modelo baseado na arquitetura de Inmon, não no de Kimball. Pense que, dos dois nomes (Kimball e Inmon), aquele que tem N é o que prega por normalização - fica mais fácil de decorar.

Gabarito: Errado

---

10. (FGV/CGE SC/2023) Em relação às diferenças de características técnicas entre um banco de dados planejado para lidar com informações transacionais (operações do dia a dia de uma empresa) e um Data Warehouse, é correto afirmar que

a) a normalização é essencial em um Data Warehouse, sobretudo no modelo dimensional estrela, de forma a evitar dados redundantes.
b) os processos analíticos normalmente usam uma pequena parcela de dados, reservando grandes porções de dados aos processos transacionais.
c) a questão de redundância de dados não é problema para o modelo dimensional (estrela), pois a normalização não é relevante entre fatos e dimensões.
d) os dados transacionais são acessados raramente, ao passo que os dados em um Data Warehouse são acessados frequentemente para o funcionamento operacional de uma empresa.
e) os dados salvos em um Data Warehouse são constantemente atualizados por meio de operações de UPDATE, ao passo que os dados transacionais recebem apenas novos registros (INSERT) e pedidos de leitura (SELECT).

Comentários:

Vamos analisar cada alternativa.

a) Errado. Vamos ver mais sobre na próxima aula, mas podemos ter dois modelos principais armazenados no Data Warehouse - o modelo estrela, não normalizado, e o floco-de-neve, normalizado. Percebe-se, portanto, que a normalização não é essencial.
b) Errado. Temos uma inversão, os processos transacionais, do dia a dia, que usam uma pequena parte dos dados - já os analíticos usam a maior quantidade possível de dados.

c) Certo. Novamente, é um tópico que veremos mais sobre na próxima aula, mas a redundância, que se refere à ocorrência de múltiplas instâncias de um mesmo dado, realmente não é um problema, já que no modelo estrela não temos normalizações.
d) Errado. Novamente, temos uma inversão de conceitos - os dados transacionais são acessados frequentemente.
e) Errado. Não podemos atualizar os dados de um Data Warehouse.

Gabarito: Letra C

---

11. (FGV/CGE SC/2023) Assinale a opção que apresenta uma diferença funcional entre um banco de dados planejado para lidar com informações transacionais (operações do dia a dia da empresa) e um Data Warehouse.

a) A finalidade de um banco de dados transacional é ser orientado para uma aplicação de negócio, e a de um Data Warehouse é ser orientado para um assunto de análise.
b) Um Data Warehouse é usado por todos os tipos de colaboradores em uma empresa, e um banco de dados transacional é usado apenas por gestores.
c) Um Data Warehouse deve ser orientado para uma aplicação de negócio, e um banco de dados transacional deve ser orientado para um assunto de análise.
d) A finalidade de um banco de dados transacional e de um Data Warehouse é a mesma: ser orientada para um assunto específico de análise.
e) Um Data Warehouse e um banco de dados transacional são igualmente utilizados por todos os colaboradores em uma empresa no nível operacional.

Comentários:

Em uma entidade, podemos ter dois tipos "gerais" de bancos de dados:

- Transacionais: são sistemas destinados a lidar com as operações do dia a dia das empresas, abrangendo atividades comuns como vendas, compras e ordens. Esses sistemas são amplamente utilizados pela maioria dos funcionários e registram as atividades operacionais da empresa. Geralmente, são compostos por bancos de dados relacionais (SQL), embora também possam incorporar tecnologias de bancos de dados NoSQL ou orientados a objetos.
- Orientados a assuntos: são bancos de dados projetados para fornecer informações cruciais para o processo decisório. Eles não são destinados a interações com o corpo geral de funcionários, mas sim a fornecer insights para analistas de dados e a alta gerência da organização. Dentro dessa categoria, destacam-se os Data Warehouses, que concentram dados de várias fontes para análise e consulta específicas.

Com isso em mente, vamos às alternativas.

a) Certo. Essa é uma das diferenças mais estruturantes - o objetivo do banco de dados. O transacional se orienta a aplicações do negócio, atividades realizadas frequentemente e que não envolvem um processo decisório, enquanto um Data Warehouse já tem uma orientação para assuntos específicos.
b) Errado. Na verdade, é o contrário. Os gestores e analistas de dado usam os DWs, enquanto a entidade como um todo usa o DB transacional
c) Errado. A afirmativa inverteu os conceitos – um DW é orientado por assunto, e um DB transacional é orientado por aplicação de negócio
d) Errado. Não faz sentido né? Ambos têm finalidades distintas: o DW tem fins analíticos, enquanto os DB transacionais têm fins mais operacionais
e) Errado. Não há como afirmar isso. Normalmente, o uso de um DB transacional é mais difundido pela empresa do que o uso do DW, reservado a analistas de dados e gestores

Gabarito: Letra A

---

12. (FGV/CGE SC/2023) Sobre a proposta geral do modelo dimensional em um Data Warehouse, não é correto afirmar que o modelo dimensional

a) cobre tanto dados detalhados quanto dados sumarizados.
b) cobre toda a empresa, e não apenas departamentos.
c) é escalável, podendo entregar relatórios com trilhões de registros.
d) é arquitetado apenas para um uso previsível, geralmente cobrindo os 10 relatórios mais acessados.
e) pode integrar diversas fontes de dados operacionais da empresa, inclusive fontes externas.

Comentários:

Da análise das afirmativas, a que apresenta um conceito errôneo é a letra D: o uso do DW é imprevisível, tende em vista que é usado primariamente por ferramentas de BI, que buscam detectar padrões e tendências não triviais. Além disso, não há essa restrição de cobrir os 10 relatórios mais acessados.

Gabarito: Letra D

---

13. (CEBRASPE/CNMP/2023) A respeito de data warehouse e data mining, julgue o próximo item.

Em data warehouse, o conceito de granularidade refere-se ao nível de detalhe ou resumo existente em uma unidade de dados, de forma que, quanto mais detalhes, mais alto o nível de granularidade.

Comentários:

Quando falamos de granularidade, estamos falando do nível de detalhamento das informações. Pensem que um grão grande é pouco detalhado e, conforme vamos detalhando o grão, vamos pegando partes cada vez menores dele – diminuindo-o.

Assim, portanto, ao aumentarmos o detalhamento de um dado, estamos diminuindo a granularidade de um dado. Errada a afirmativa.

Gabarito: Errado

---

14. (FCC/MPE PB/2023) Um Data Warehouse

a) de duas camadas divide as fontes de dados tangíveis em dois segmentos: os dados transacionais, utilizados unicamente por ferramentas analíticas, e os dados operacionais, utilizados exclusivamente por ferramentas de Data Mining.
b) trabalha com metadados, que definem níveis de acesso que permitem aos usuários transferir dados, sendo também usados para dividir as informações por departamento ou dentro de visões necessárias para os diferentes usuários.
c) cujo schema é baseado no modelo Star, permite a decomposição de uma ou mais dimensões que possuem hierarquias, sendo o mais utilizado por ser de baixa complexidade.
d) oferece consultas com nível de granularidade da tabela fato variável, de forma que quanto maior a granularidade, maior o nível de detalhamento e quanto menor a granularidade, menor o nível de detalhamento da informação.
e) possui como característica a não volatilidade, ou seja, os dados ficam disponíveis para consulta, não sendo estes alterados quando novas informações são carregadas.

Comentários:

Vamos analisar cada afirmativa:

a) Errado. Não há essa separação. Uma coisa são bancos de dados transacionais, que têm fins operacionais, outra são os Data Warehouses (DW).
b) Errado. Metadados são as descrições dos dados – usualmente encontrados em bancos de dado transacionais, não dimensionais.
c) Errado. A existência de hierarquia entre dimensões é presente no modelo snowflake, não no estrela.
d) Errado. Como vimos na questão anterior, maior granularidade representa um menor detalhamento, ao contrário do que a afirmativa induz.
e) Certo. Um dos princípios do DW é justamente e não volatilidade – isso é, os dados ficam armazenados no nosso repositório, criando dados históricos.

Gabarito: Letra E

---

15. (FGV/TJ RN/2023) Para integrar os dados de diversas fontes, Julia desenvolveu um ETL para executar ações sobre os dados como: extrair, limpar, agregar, transformar e carregar dados em um banco de dados destino visando apoiar análises históricas.

Para implementar as ações sobre os dados em um ETL, Julia utilizou:

a) steps e fluxos de dados;
b) repositório de metadados;
c) sequências temporais;
d) regras de associação;
e) data mining.

Comentários:

Julia utilizou "steps" (passos) e "fluxos de dados" para implementar as ações sobre os dados em um processo de ETL. Os "steps" representam as etapas individuais do processo, como extração, limpeza, transformação e carga dos dados. Os "fluxos de dados" referem-se à maneira como os dados são movidos e processados entre esses passos, seguindo uma sequência lógica definida para realizar as transformações necessárias antes de serem carregados no banco de dados destino.

Gabarito: Letra A

---

16. (FGV/CGE SC/2023) As informações analiticamente úteis das fontes de dados operacionais (das operações do dia a dia do negócio) são carregadas no Data Warehouse por meio do processo de ETL. Um dos recursos úteis em um DW é poder observar um mesmo item de dimensão em vários instantes de tempo (timestamps), como, por exemplo, observar o preço de venda de um produto ao longo dos anos. Assinale a opção que indica a técnica que torna possível a disposição desse recurso.

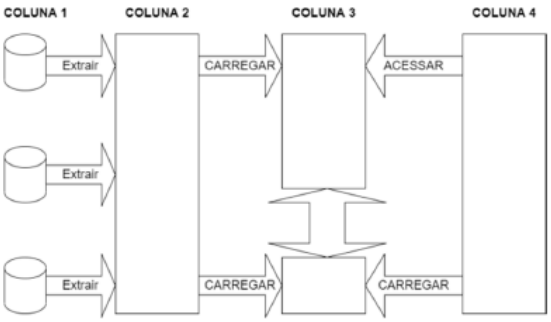
a) A supressão, no Data Warehouse, das chaves primárias do bando de dados operacional.
b) A criação de chaves primárias compostas por um atributo de chave substituta e um de chave primária do banco de dados operacional.
c) A substituição, e consequente supressão, das chaves primárias do banco de dados operacional por chaves substitutas no Data Warehouse.
d) A criação de chaves primárias substitutas no Data Warehouse, mantendo as chaves primárias do banco de dados operacional como atributos únicos no Data Warehouse.
e) A criação de chaves primárias substitutas no Data Warehouse, mantendo as chaves primárias do banco de dados operacional como atributos não chave no Data Warehouse.

Comentários:

A técnica utilizada é a substituição de chaves primárias. Essas chaves são consideradas chaves primárias na nova tabela, no nosso Data Warehouse, e coexistem com as chaves primárias antigas - que deixam de se chaves primárias, se tornando atributos não chave.

Gabarito: Letra E

Uma empresa necessita estruturar, melhorar e utilizar cada vez mais recursos, a fim de gerar inteligência para o seu negócio. Nesse sentido, foi desenvolvido o esquema a seguir, a ser utilizado como uma visão dos elementos do seu respectivo data warehouse, que apoiará a inteligência do negócio.



17. (CEBRASPE/FUNPRESP/2022) Dados que são armazenados na área representada pela coluna 2 não podem sofrer nenhum tipo de modificação, ou seja, devem permanecer sem nenhum tipo de alteração ou ajuste.

Comentários:

Errado, gente! Na coluna 2, a staging área, é justamente onde ocorre as transformações, limpeza e adaptação dos dados, para podermos carregar os dados na coluna 3.

Gabarito: Errado

---

18. (CEBRASPE/FUNPREP/2022) A área representada pela coluna 2 é alimentada com dados obtidos da coluna 1; ela é uma área acessível aos usuários para efetuarem suas consultas.

Comentários:

As consultas são feitas na coluna 3(DW), e não na coluna 2 (Staging Area)

Gabarito: Errado

---

19. (CEBRASPE/BNB/2022) Julgue o item a seguir, a respeito do conceito de data lake.
O termo data lake é usado para se referir a uma arquitetura em que os dados são armazenados em vários sistemas de armazenamento de dados e em diferentes formatos, inclusive em sistemas de arquivos, mas podem ser consultados em um único sistema.

Comentários:

Perfeito! Data Lakes são caracterizados por armazenar os mais variados tipos de dado, em sistemas diferentes, podendo ser consultados a partir de um sistema único de entrada.

Gabarito: Certo

---

20. (CEBRASPE/TCE RJ/2021) A respeito de bancos de dados relacionais e de modelagem dimensional, julgue o item subsequente.

A construção de um data mart antecede a criação de um data warehouse.

Comentários:

Como vimos na aula, há duas abordagens diferentes para a construção de um data mart – a top down, em que a construção do DW antecede a construção do DM, e a bottom up, onde construímos os DM antecipadamente.

Não há como afirmar que uma é padrão, e que deve ser seguida – pois ambas são abordagens válidas e diferentes. Por esse motivo, está errada a afirmativa.

Gabarito: Errado

---

21. (CEBRASPE/MPE AP/2021) Tecnologias que recuperam dados de muitas fontes, limpando-os e carregando-os em data warehouse, e que fazem parte de qualquer projeto centrado em dados denominam-se

a) Depósitos de Dados Operacionais (ODS).
b) ETL (Extract, Transform and Load).
c) BPM (Business Performance Management).
d) OLTP (online transaction processing).
e) KPI (Key Performance Indicators).

Comentários:

Questão tranquila e sem muito segredo, né? Quando estamos extraindo, limpado e carregando dados, estamos usando a pipeline ETL.

Gabarito: Letra B

---

22. (CEBRASPE/SEFAZ CE/2021) Julgue o próximo item, relativo ao business intelligence .

Um data warehouse (DW), ainda que seja não volátil — ou seja, após os dados serem inseridos nele os usuários não podem alterá-los — é variável no tempo, pois mantém um conjunto de dados históricos que oferecem suporte à tomada de decisões.

Comentários:

Perfeito! É justamente isso que quer dizer a não volatilidade e a variância no tempo de um DW. O dado em si não é atualizado, mas são feitas novas inserções com esse mesmo dado com valores diferentes, criando um contexto histórico para o mesmo.

Gabarito: Certo

---

23. (CEBRASPE/FUNPRESP/2022) A respeito de modelagem dimensional e data marts, julgue o item subsecutivo.

Na modelagem dimensional, data marts devem ser criados e utilizados para dados de resumo.

Comentários:

Data Marts usualmente são criados para separar os dados por áreas do negócio, como setores, fábricas etc. Não existe essa relação entre data marts e dados de resumo.

Gabarito: Errado

---

24. (CEBRASPE/PETROBRAS/2022) Quanto aos conceitos relativos à arquitetura de dados, julgue o item a seguir.

Data lakes são grandes armazenadores de informações, vindas de diversas fontes, na qual diversos usuários podem ter acesso para fazer a análise e coletar insights importantes para o negócio.

Resolução:

Perfeito! Essa é a caracterização correta dos Data Lakes. Perceba que, apesar de a coleta de insights importantes usualmente estar mais aliada ao Data Warehouse, nada impede que a mesma seja feita nos Data Lakes.

Gabarito: Certo

---

25. (CEBRASPE/TC DF/2023) No que se refere a Big Data, data lake, business intelligence e data warehousing, julgue o item seguinte.

Business intelligence é uma técnica utilizada para organizar dados em tabelas relacionadas a fatos e ocorrências, para otimizar as transações de inclusão.

Comentários:

O BI é puramente relacionado à descoberta de informações - não possui relação com a otimização de transações. Incorreta a afirmativa.

Gabarito: Errado

---

26. (CESGRANRIO/IPEA/2024) O processo de ingestão de dados é normalmente dividido em três etapas principais:

> 1 - Extração, ou coleta, de dados das fontes disponíveis;
> 2 - Transformação dos dados coletados para que atendam às necessidades específicas de processamento e análise; e
> 3 - Carga dos dados em algum repositório de destino, como um banco de dados relacional ou um data lake. Essas três etapas podem variar dependendo de os dados serem estruturados ou não.

Nesse contexto, verifica-se que, na etapa de

a) carga, os dados estruturados são sempre transferidos diretamente ao repositório de destino, sem necessidade de transformação.
b) carga, os dados não estruturados são sempre convertidos em formatos estruturados antes de serem armazenados.
c) extração, os dados estruturados são coletados exclusivamente através de APIs especializadas.
d) transformação, os dados estruturados podem requerer conversão para um formato não estruturado para facilitar a análise avançada.
e) transformação, os dados não estruturados podem necessitar de processamento de linguagem natural ou de técnicas de reconhecimento de imagens.

Comentários:

Vamos analisar cada alternativa.

a) Errado. Não é pelo dado ser estruturado que ele está já no padrão do repositório de destino. Por isso, nem sempre teremos a inclusão direta - só reparar que o ETL trabalha com dados estruturados e mesmo assim faz uma transformação deles.
b) Errado. Só há necessidade de conversão em modelos estruturados se estivermos inserindo em um DW. Se a inserção for em um Data Lake, não há essa necessidade.
c) Errado. Temos diversas formas de extração, não somente por APIs.
d) Errado. A transformação para dados estruturados que ajuda a fazer análises avançadas - e não o caminho oposto.
e) Certo. As técnicas apontadas são técnicas que visam enriquecer e estruturar os dados, empregadas na etapa de transformação.

Gabarito: Letra E

---

27. (CONSULPLAN/SEGER ES/2023) Os processos ELT (Extract, Load, Transform) e ETL (Extract, Transform, Load) lidam com tratamento de dados através da integração de dados de diversas fontes. Sobre os processos ELT e ETL, assinale a afirmativa correta.

a) Assim como no ETL, também para o ELT é na etapa "transformação" que há agregação de valor.
b) ETL é usado especificamente em Data Lake (DL), enquanto o ELT é usado em Data Warehouse (DW).
c) Na abordagem ELT, a transformação é a única etapa automática, ou seja, não é realizada sob demanda.
d) No ELT, o mesmo artefato responsável pela extração não se encarrega da carga dos dados no destino, já que as etapas de Extração e Carga (EL) são distintas.
e) Via de regra, sistemas de ETL são pequenos e simples e não demandam grande esforço, tempo e conhecimentos especializados, para construção e manutenção.

Comentários:

Vamos analisar cada uma das alternativas.

a) Certo. Apesar da ordem diferente, a etapa é a mesma - o Transform agrega valor aos dados, transformando-os e enriquecendo-os.
b) Errado. Não temos uma ligação obrigatória entre ambos - mas, usualmente, o ETL é usado com DWs, e o ELT com DLs.
c) Errado. Podemos ter todas as etapas do processo feitas de forma automatizada.
d) Errado. No ELT, a extração e carga são feitas conjuntamente, como se fosse uma operação só.
e) Errado. Muito pelo contrário, sistemas ETL são complexos pois lidam com grande quantidades de dados, de diferentes fontes.

Gabarito: Letra A

---

28. (CESGRANRIO/IPEA/2024) Existem várias abordagens para a ingestão de dados, sendo cada uma delas adequada para determinado tipo de necessidade e de cenário.

No caso da ingestão de dados em tempo real, streaming, os dados são

a) coletados e processados em intervalos regulares, por exemplo, diariamente ou semanalmente.
b) capturados e processados continuamente à medida que são gerados.

c) processados em pequenos lotes, com o processamento ocorrendo em intervalos curtos, mas não instantâneos.
d) processados apenas após um evento específico ser acionado, como, por exemplo, uma transação em banco de dados ou um clique de usuário.
e) armazenados em um data lake ou data warehouse, antes de qualquer forma de processamento ou de análise.

Comentários:

Podemos ter duas formas de ingestão de dados num Data Warehouse:

- Em lotes (batches): a ingestão ocorre periodicamente, em intervalos definidos, com uma quantidade alta de dados de cada vez.
- Em fluxo, ou tempo real: a ingestão é constante, e os dados são capturados, processados e inseridos assim que forem gerados.

A questão cobra conhecimentos acerca do processamento em tempo real e, a alternativa que melhor o descreve, é a letra B. Quanto às demais:

a) Errado. Descreve o carregamento em lotes.
b) Certo. É o nosso gabarito.
c) Errado. Descreve o que chamamos de *mini-batches*, ou mini lotes.
d) Errado. Esse é um tipo de ingestão baseado em eventos, pouco usado.
e) Errado. Até podemos ter um dado armazenado num Data Lake ou Data Warehouse antes da ingestão, mas não é esse o objetivo da ingestão em fluxo.

Gabarito: Letra D

---

29. (FGV/TJ RN/2023) A gestão do TJRN é apoiada por sistemas de informações digitais que estão em produção há mais de dez anos abrangendo diversos contextos, como gestão de pessoal, gestão orçamentária, pedidos de serviço, controle de viaturas etc. Para apoiar a tomada de decisão de alto nível do Tribunal, é necessário o desenvolvimento de um banco de dados analítico que seja orientado a assunto, não volátil e histórico, integrando dados estruturados dos diversos sistemas e contextos.

O banco de dados a ser desenvolvido é um Data:

a) Lake;
b) Mart;
c) Graph;
d) Mining;
e) Warehouse.

Comentários:

A descrição de características descreve perfeitamente o Data Warehouse - lembre-se:

Orientado a assuntos	Não volátil
Variante no tempo	Integrado

Gabarito: Letra E

---

30. (CEBRASPE/MCOM/2022) Julgue o item a seguir, a respeito de ETL, ELT e data lake.

O processo ETL (extrair, transformar e carregar) permite analisar grandes volumes de dados de forma rápida; para isso, é necessário duplicar o espaço em disco e triplicar o tempo no carregamento e na transformação de dados em relação ao ELT (extrair, carregar e transformar), que compacta os dados no carregamento.

Comentários:

A afirmação de que o processo ETL duplica o espaço em disco e triplica o tempo no carregamento e na transformação de dados em relação ao ELT, não é uma generalização precisa. Na verdade, a eficiência e o desempenho de ambos os processos dependem de vários fatores, incluindo a quantidade de dados, a complexidade das transformações, a infraestrutura de hardware e software utilizada, entre outros.

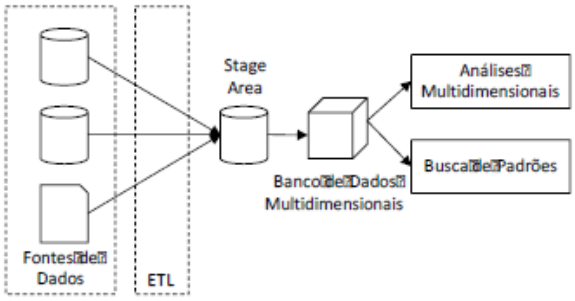
Gabarito: Errado

LISTA DE QUESTÕES

01. (CEBRASPE/CNPq/2024) Acerca da análise de dados para tomada de decisão, da capacitação tecnológica e da competitividade, julgue o item a seguir.

Business intelligence (BI) pode ser definido corretamente como um conjunto de tecnologias que dão suporte a decisões gerenciais por meio de informações internas e externas às organizações, tendo grande impacto na estratégia corporativa, na performance e na competitividade.

02. (FGV/DPE RS/2023) Observe a seguinte arquitetura básica de uma solução de Business Intelligence implementada na empresa CleverBI.



O elemento arquitetural da solução de BI da CleverBI implementado por meio de operações OLAP, como slice, rotate, drill-down e drill-up, é o:

a) ETL;
b) Stage Area;
c) Banco de Dados Multidimensionais;
d) Análises Multidimensionais;
e) Busca de Padrões.

03. (FGV/CGE SC/2023) Avalie se os componentes de um Data Warehouse incluem:

I. Sistemas de origem.
II. Infraestrutura de ETL (Extraction-transformation-load).
III. Data Warehouse.
IV. Aplicações de Front-end para o usuário final.

Estão corretos os itens

a) I e II, apenas.
b) III e IV, apenas.
c) I, II e III, apenas.
d) II, III e IV, apenas.

e) I, II, III e IV.

04. (FGV/ISS RJ/2023) No ambiente de Data Warehousing da SMFP/RJ, os dados são extraídos de diversas fontes e integrados para apoiar a fiscalização de rendas por meio do desenvolvimento de diversos artefatos de dados. Antes de começar as análises nesse ambiente, o fiscal de rendas Inácio fez consultas sobre o seu conteúdo, por exemplo:

- DW_Tributos, banco de dados analítico do tipo Data Warehouse que integra dados sobre os tributos arrecadados do Município do Rio de Janeiro.
- TP_EMPRESA, Caractere, 1, atributo que descreve o tipo da empresa contendo os seguintes valores: M - MEI ou S - Simples Nacional, e faz parte da tabela TB_EMPRESA.
- RL_Sit_Fiscal, relatório sobre a situação fiscal das empresas do Município do Rio de Janeiro.

O componente do ambiente de Data Warehousing, utilizado por Inácio, que foi desenvolvido para apoiar consultas sobre a descrição de cada artefato de dado, é:

a) ETL;
b) Métricas;
c) Dashboard;
d) Dimensões de Dados;
e) Repositório de Metadados.

05. (CEBRASPE/MPE RO/2023) Em uma solução de BI (Business Intelligence), os dashboards são

a) fontes de dados.
b) insights.
c) usados no ETL.
d) armazéns de dados.
e) modelos semânticos de dados.

06. (FGV/PREF. SJC/2024) Relacione as características de um data warehouse listadas a seguir com suas descrições, conforme estabelecido por William Inmon.

1. Orientados a Assunto.
2. Integração.
3. Não Volátil.
4. Variante no Tempo.

( ) O foco de um data warehouse na mudança ao longo do tempo é essencial para descobrir tendências e identificar padrões e relacionamentos ocultos nos negócios, para isso os analistas precisam de grandes quantidades de dados. Isso contrasta muito com o

processamento de transações on-line onde os requisitos de desempenho exigem que os dados históricos sejam movidos para arquivos.
( ) Os data warehouses devem colocar dados de fontes diferentes em um formato consistente. Eles devem resolver problemas como nomear conflitos e inconsistências entre unidades de medida.
( ) Significa que, uma vez inseridos no data warehouse, os dados não devem mudar. Essa característica é lógica porque o propósito de um data warehouse é permitir que um analista analise o que ocorreu no passado.
( ) Os data warehouses são projetados para ajudar os profissonais a analisar grandes volumes de dados. Por exemplo, para saber mais sobre os dados de vendas de uma empresa, o analista pode construir um data warehouse que concentre a venda. Usando esse data warehouse, ele poderá responder perguntas como "Quem foi nosso melhor cliente para este item no ano passado?" ou "Quem provavelmente será nosso melhor cliente no próximo ano?"

A relação correta, na ordem dada, é:

a) 1 – 2 – 3 – 4.
b) 2 – 1 – 4 – 3.
c) 4 – 2 – 3 – 1.
d) 3 – 4 – 1 – 2.
e) 2 – 3 – 1 – 4.

07. (FGV/PREF. SJC/2024) Com relação ao ETL, ELT e suas tecnologias, avalie as afirmativas a seguir e assinale V para a verdadeira e F para a falsa.

( ) ELT utiliza fluxos de trabalho de análise de dados e de aprendizado de máquina. O ELT é frequentemente usado por uma organização para: Extrair dados de sistemas legados, limpar os dados para melhorar sua qualidade e carregar dados em um banco de dados de destino. O ELT transforma dados no trânsito.
( ) ETL copia ou exporta os dados dos locais de origem, mas, em vez de carregá-los em uma área de preparação para transformação, ele carrega os dados em estado brutos diretamente no armazenamento de dados no destino para serem transformados conforme necessário. O ETL não transforma nenhum dado no trânsito.
( ) A ordem das etapas não é a única diferença entre ETL e ELT. No ELT, o armazenamento de dados de destino pode ser um armazém de dados, mas, mais frequentemente, é um data lake, que é um armazenamento central grande projetado para manter tanto dados estruturados quanto não estruturados em grande escala.

As afirmativas são, respectivamente,

a) F – V – F.

b) F – F – V.
c) F – V – V.
d) V – F – V.
e) V – V – F.

08. (FGV/PREF. SJC/2024) Com relação ao Data Warehousing e ao Business Intelligence, avalie as afirmativas a seguir e assinale V para a afirmativa verdadeira e F para a falsa.

( ) Eles têm como meta construir e manter o ambiente técnico e os processos técnicos e de negócios necessários para fornecer dados integrados em apoio às funções operacionais, requisitos de conformidade e atividades de inteligência de negócios.
( ) Ambos visam apoiar e permitir análises de negócios e tomadas de decisões mais eficazes por parte dos trabalhadores do conhecimento.
( ) O Data Warehousing concentra-se em permitir um contexto de negócios histórico e integrado em dados operacionais, aplicando regras de negócios e mantendo relacionamentos de dados de negócios apropriados. O armazenamento de dados também inclui processos que interagem com repositórios de metadados.

As afirmativas são, respectivamente,

a) F – F – F.
b) F – F – V.
c) F – V – V.
d) V – F – V.
e) V – V – V.

09. (CEBRASPE/SPELAN RR/2023) No que se refere às características de um banco de dados relacional, julgue o item que se segue.

Em contraste com os modelos relacionais, os modelos dimensionais, ou modelos de dados dimensionais kimball — modelos de dados baseados na técnica desenvolvida por Ralph Kimball —, são estruturas normalizadas projetadas para recuperar dados de um data warehouse.

10. (FGV/CGE SC/2023) Em relação às diferenças de características técnicas entre um banco de dados planejado para lidar com informações transacionais (operações do dia a dia de uma empresa) e um Data Warehouse, é correto afirmar que

a) a normalização é essencial em um Data Warehouse, sobretudo no modelo dimensional estrela, de forma a evitar dados redundantes.
b) os processos analíticos normalmente usam uma pequena parcela de dados, reservando grandes porções de dados aos processos transacionais.
c) a questão de redundância de dados não é problema para o modelo dimensional (estrela), pois a normalização não é relevante entre fatos e dimensões.

d) os dados transacionais são acessados raramente, ao passo que os dados em um Data Warehouse são acessados frequentemente para o funcionamento operacional de uma empresa.
e) os dados salvos em um Data Warehouse são constantemente atualizados por meio de operações de UPDATE, ao passo que os dados transacionais recebem apenas novos registros (INSERT) e pedidos de leitura (SELECT).

11. (FGV/CGE SC/2023) Assinale a opção que apresenta uma diferença funcional entre um banco de dados planejado para lidar com informações transacionais (operações do dia a dia da empresa) e um Data Warehouse.

a) A finalidade de um banco de dados transacional é ser orientado para uma aplicação de negócio, e a de um Data Warehouse é ser orientado para um assunto de análise.
b) Um Data Warehouse é usado por todos os tipos de colaboradores em uma empresa, e um banco de dados transacional é usado apenas por gestores.
c) Um Data Warehouse deve ser orientado para uma aplicação de negócio, e um banco de dados transacional deve ser orientado para um assunto de análise.
d) A finalidade de um banco de dados transacional e de um Data Warehouse é a mesma: ser orientada para um assunto específico de análise.
e) Um Data Warehouse e um banco de dados transacional são igualmente utilizados por todos os colaboradores em uma empresa no nível operacional.

12. (FGV/CGE SC/2023) Sobre a proposta geral do modelo dimensional em um Data Warehouse, não é correto afirmar que o modelo dimensional

a) cobre tanto dados detalhados quanto dados sumarizados.
b) cobre toda a empresa, e não apenas departamentos.
c) é escalável, podendo entregar relatórios com trilhões de registros.
d) é arquitetado apenas para um uso previsível, geralmente cobrindo os 10 relatórios mais acessados.
e) pode integrar diversas fontes de dados operacionais da empresa, inclusive fontes externas.

13. (CEBRASPE/CNMP/2023) A respeito de data warehouse e data mining, julgue o próximo item.

Em data warehouse, o conceito de granularidade refere-se ao nível de detalhe ou resumo existente em uma unidade de dados, de forma que, quanto mais detalhes, mais alto o nível de granularidade.

14. (FCC/MPE PB/2023) Um Data Warehouse

a) de duas camadas divide as fontes de dados tangíveis em dois segmentos: os dados transacionais, utilizados unicamente por ferramentas analíticas, e os dados operacionais, utilizados exclusivamente por ferramentas de Data Mining.
b) trabalha com metadados, que definem níveis de acesso que permitem aos usuários transferir dados, sendo também usados para dividir as informações por departamento ou dentro de visões necessárias para os diferentes usuários.
c) cujo schema é baseado no modelo Star, permite a decomposição de uma ou mais dimensões que possuem hierarquias, sendo o mais utilizado por ser de baixa complexidade.
d) oferece consultas com nível de granularidade da tabela fato variável, de forma que quanto maior a granularidade, maior o nível de detalhamento e quanto menor a granularidade, menor o nível de detalhamento da informação.
e) possui como característica a não volatilidade, ou seja, os dados ficam disponíveis para consulta, não sendo estes alterados quando novas informações são carregadas.

15. (FGV/TJ RN/2023) Para integrar os dados de diversas fontes, Julia desenvolveu um ETL para executar ações sobre os dados como: extrair, limpar, agregar, transformar e carregar dados em um banco de dados destino visando apoiar análises históricas.

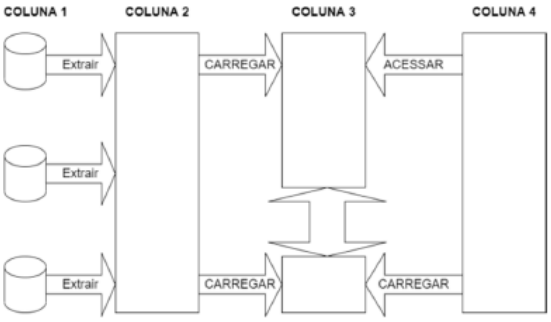
Para implementar as ações sobre os dados em um ETL, Julia utilizou:

a) steps e fluxos de dados;
b) repositório de metadados;
c) sequências temporais;
d) regras de associação;
e) data mining.

16. (FGV/CGE SC/2023) As informações analiticamente úteis das fontes de dados operacionais (das operações do dia a dia do negócio) são carregadas no Data Warehouse por meio do processo de ETL. Um dos recursos úteis em um DW é poder observar um mesmo item de dimensão em vários instantes de tempo (timestamps), como, por exemplo, observar o preço de venda de um produto ao longo dos anos. Assinale a opção que indica a técnica que torna possível a disposição desse recurso.

a) A supressão, no Data Warehouse, das chaves primárias do bando de dados operacional.
b) A criação de chaves primárias compostas por um atributo de chave substituta e um de chave primária do banco de dados operacional.
c) A substituição, e consequente supressão, das chaves primárias do banco de dados operacional por chaves substitutas no Data Warehouse.
d) A criação de chaves primárias substitutas no Data Warehouse, mantendo as chaves primárias do banco de dados operacional como atributos únicos no Data Warehouse.
e) A criação de chaves primárias substitutas no Data Warehouse, mantendo as chaves primárias do banco de dados operacional como atributos não chave no Data Warehouse.

Uma empresa necessita estruturar, melhorar e utilizar cada vez mais recursos, a fim de gerar inteligência para o seu negócio. Nesse sentido, foi desenvolvido o esquema a seguir, a ser utilizado como uma visão dos elementos do seu respectivo data warehouse, que apoiará a inteligência do negócio.



17. (CEBRASPE/FUNPRESP/2022) Dados que são armazenados na área representada pela coluna 2 não podem sofrer nenhum tipo de modificação, ou seja, devem permanecer sem nenhum tipo de alteração ou ajuste.

18. (CEBRASPE/FUNPREP/2022) A área representada pela coluna 2 é alimentada com dados obtidos da coluna 1; ela é uma área acessível aos usuários para efetuarem suas consultas.

19. (CEBRASPE/BNB/2022) Julgue o item a seguir, a respeito do conceito de data lake.
O termo data lake é usado para se referir a uma arquitetura em que os dados são armazenados em vários sistemas de armazenamento de dados e em diferentes formatos, inclusive em sistemas de arquivos, mas podem ser consultados em um único sistema.

20. (CEBRASPE/TCE RJ/2021) A respeito de bancos de dados relacionais e de modelagem dimensional, julgue o item subsequente.

A construção de um data mart antecede a criação de um data warehouse.

21. (CEBRASPE/MPE AP/2021) Tecnologias que recuperam dados de muitas fontes, limpando-os e carregando-os em data warehouse, e que fazem parte de qualquer projeto centrado em dados denominam-se

a) Depósitos de Dados Operacionais (ODS).
b) ETL (Extract, Transform and Load).
c) BPM (Business Performance Management).
d) OLTP (online transaction processing).
e) KPI (Key Performance Indicators).

22. (CEBRASPE/SEFAZ CE/2021) Julgue o próximo item, relativo ao business intelligence .

Um data warehouse (DW), ainda que seja não volátil — ou seja, após os dados serem inseridos nele os usuários não podem alterá-los — é variável no tempo, pois mantém um conjunto de dados históricos que oferecem suporte à tomada de decisões.

23. (CEBRASPE/FUNPRESP/2022) A respeito de modelagem dimensional e data marts, julgue o item subsecutivo.

Na modelagem dimensional, data marts devem ser criados e utilizados para dados de resumo.

24. (CEBRASPE/PETROBRAS/2022) Quanto aos conceitos relativos à arquitetura de dados, julgue o item a seguir.

Data lakes são grandes armazenadores de informações, vindas de diversas fontes, na qual diversos usuários podem ter acesso para fazer a análise e coletar insights importantes para o negócio.

25. (CEBRASPE/TC DF/2023) No que se refere a Big Data, data lake, business intelligence e data warehousing, julgue o item seguinte.

Business intelligence é uma técnica utilizada para organizar dados em tabelas relacionadas a fatos e ocorrências, para otimizar as transações de inclusão.

26. (CESGRANRIO/IPEA/2024) O processo de ingestão de dados é normalmente dividido em três etapas principais:

> 1 - Extração, ou coleta, de dados das fontes disponíveis;
> 2 - Transformação dos dados coletados para que atendam às necessidades específicas de processamento e análise; e
> 3 - Carga dos dados em algum repositório de destino, como um banco de dados relacional ou um data lake. Essas três etapas podem variar dependendo de os dados serem estruturados ou não.

Nesse contexto, verifica-se que, na etapa de

a) carga, os dados estruturados são sempre transferidos diretamente ao repositório de destino, sem necessidade de transformação.
b) carga, os dados não estruturados são sempre convertidos em formatos estruturados antes de serem armazenados.
c) extração, os dados estruturados são coletados exclusivamente através de APIs especializadas.

d) transformação, os dados estruturados podem requerer conversão para um formato não estruturado para facilitar a análise avançada.
e) transformação, os dados não estruturados podem necessitar de processamento de linguagem natural ou de técnicas de reconhecimento de imagens.

27. (CONSULPLAN/SEGER ES/2023) Os processos ELT (Extract, Load, Transform) e ETL (Extract, Transform, Load) lidam com tratamento de dados através da integração de dados de diversas fontes. Sobre os processos ELT e ETL, assinale a afirmativa correta.

a) Assim como no ETL, também para o ELT é na etapa "transformação" que há agregação de valor.
b) ETL é usado especificamente em Data Lake (DL), enquanto o ELT é usado em Data Warehouse (DW).
c) Na abordagem ELT, a transformação é a única etapa automática, ou seja, não é realizada sob demanda.
d) No ELT, o mesmo artefato responsável pela extração não se encarrega da carga dos dados no destino, já que as etapas de Extração e Carga (EL) são distintas.
e) Via de regra, sistemas de ETL são pequenos e simples e não demandam grande esforço, tempo e conhecimentos especializados, para construção e manutenção.

28. (CESGRANRIO/IPEA/2024) Existem várias abordagens para a ingestão de dados, sendo cada uma delas adequada para determinado tipo de necessidade e de cenário.

No caso da ingestão de dados em tempo real, streaming, os dados são

a) coletados e processados em intervalos regulares, por exemplo, diariamente ou semanalmente.
b) capturados e processados continuamente à medida que são gerados.
c) processados em pequenos lotes, com o processamento ocorrendo em intervalos curtos, mas não instantâneos.
d) processados apenas após um evento específico ser acionado, como, por exemplo, uma transação em banco de dados ou um clique de usuário.
e) armazenados em um data lake ou data warehouse, antes de qualquer forma de processamento ou de análise.

29. (FGV/TJ RN/2023) A gestão do TJRN é apoiada por sistemas de informações digitais que estão em produção há mais de dez anos abrangendo diversos contextos, como gestão de pessoal, gestão orçamentária, pedidos de serviço, controle de viaturas etc. Para apoiar a tomada de decisão de alto nível do Tribunal, é necessário o desenvolvimento de um banco de dados analítico que seja orientado a assunto, não volátil e histórico, integrando dados estruturados dos diversos sistemas e contextos.

O banco de dados a ser desenvolvido é um Data:

a) Lake;
b) Mart;
c) Graph;
d) Mining;
e) Warehouse.

30. (CEBRASPE/MCOM/2022) Julgue o item a seguir, a respeito de ETL, ELT e data lake.

O processo ETL (extrair, transformar e carregar) permite analisar grandes volumes de dados de forma rápida; para isso, é necessário duplicar o espaço em disco e triplicar o tempo no carregamento e na transformação de dados em relação ao ELT (extrair, carregar e transformar), que compacta os dados no carregamento.

GABARITO

GABARITO

1. Certo
2. Letra D
3. Letra E
4. Letra E
5. Letra B
6. Letra C
7. Letra B
8. Letra E
9. Errado
10. Letra C
11. Letra A
12. Letra D
13. Errado
14. Letra E
15. Letra A
16. Letra E
17. Errado
18. Errado
19. Certo
20. Errado
21. Letra B
22. Certo
23. Errado
24. Certo
25. Errado
26. Letra E
27. Letra A
28. Letra D
29. Letra E
30. Errado

ESSA LEI TODO MUNDO CONHECE:
PIRATARIA É CRIME.

Mas é sempre bom revisar o porquê e como você pode ser prejudicado com essa prática.

1. 

Professor investe seu tempo para elaborar os cursos e o site os coloca à venda.

2. 

Pirata divulga ilicitamente (grupos de rateio), utilizando-se do anonimato, nomes falsos ou laranjas (geralmente o pirata se anuncia como formador de "grupos solidários" de rateio que não visam lucro).

3. 

Pirata cria alunos fake praticando falsidade ideológica, comprando cursos do site em nome de pessoas aleatórias (usando nome, CPF, endereço e telefone de terceiros sem autorização).

4. 

Pirata compra, muitas vezes, clonando cartões de crédito (por vezes o sistema anti-fraude não consegue identificar o golpe a tempo).

5. 

Pirata fere os Termos de Uso, adultera as aulas e retira a identificação dos arquivos PDF (justamente porque a atividade é ilegal e ele não quer que seus fakes sejam identificados).

6. 

Pirata revende as aulas protegidas por direitos autorais, praticando concorrência desleal e em flagrante desrespeito à Lei de Direitos Autorais (Lei 9.610/98).

7. 

Concurseiro(a) desinformado participa de rateio, achando que nada disso está acontecendo e esperando se tornar servidor público para exigir o cumprimento das leis.

8. 

O professor que elaborou o curso não ganha nada, o site não recebe nada, e a pessoa que praticou todos os ilícitos anteriores (pirata) fica com o lucro.

Deixando de lado esse mar de sujeira, aproveitamos para agradecer a todos que adquirem os cursos honestamente e permitem que o site continue existindo.