Estratégia Concursos

Aula 00

Concursos da Área Fiscal Especialidade TI - Ciência de Dados

Autor:
Equipe Informática e TI, Thiago Rodrigues Cavalcanti

10 de Outubro de 2023

Índice

1) Banco de Dados - Apresentação do Professor 3

2) Business Intelligence, Data Warehouse e Modelagem Dimensional 8

CONCEITOS DE BANCO DE DADOS

APRESENTAÇÃO DO PROFESSOR

Olá,

Sejam bem-vindos a mais um curso de Tecnologia da Informação (TI)! Hoje apresentamos o mais completo curso no que se refere a Banco de Dados para concursos. Gosto sempre de dizer que é um prazer imenso fazer parte desta equipe de professores do Estratégia Concursos e ter a oportunidade de apresentar um pouco do meu conhecimento e experiência em concursos públicos.

Antes de começar de fato o conteúdo teórico desta aula, vou me apresentar de forma rápida. Meu nome é Thiago, sou casado, pernambucano, tenho três filhos, Vinícius (13 anos), Lucas (*in memoriam*) e Júlia (3 anos). Torço pelo Sport Clube do Recife. Sou cristão. Me formei em Ciência da Computação pela UFPE. Tenho mestrado em engenharia de software na mesma instituição. Também tenho doutorado em economia na UnB.

Frequento academia para manter a forma, mas meu hobby mesmo é pedalar! Decidi vender o carro e viver num desafio intermodal de transporte. Ia para o trabalho de *bicicleta* sempre que possível! Ultimamente, com o teletrabalho, a bicicleta é usada apenas nos finais de semana! Agora, uma pergunta: onde eu trabalho? No Banco Central do Brasil!

Fruto de uma trajetória de dois anos de estudos diários. Aposentei as canetas em 2010. Hoje estou lotado no Banco Central em Recife trabalhando com fiscalização das empresas que compõem a infraestrutura do mercado financeiro (IMFs). Hoje, minhas tarefas envolvem verificações e sugestões de TI para melhoria das empresas e do ecossistema financeiro.

Minha experiência com gestão de dados é parte de uma estratégia profissional de alinhar meu trabalho diário como servidor público com minha carreira paralela de professor e consultor de Banco de Dados (BD) e *Business Intelligence* (BI). A ideia é conseguir me especializar cada vez mais no tema desta carreira dentro da TI, que o mercado chama de **cientista dos dados (*Data scientist*)**.

Entrei neste universo de professor de concurso há alguns anos. Desde 2012, tenho me dedicado especificamente ao conteúdo de BD e BI. Minhas experiências em cursos presenciais em Brasília e em diversas partes do Brasil, bem como na gravação sistemática de aulas on-line me ajudaram a desenvolver um conteúdo exclusivo para os alunos do Estratégia Concursos.

A ideia é desenvolver um material completo, recheado de questões e com diversas dicas para ajudar você no seu objetivo: **ser aprovado e nomeado**!

Agora gostaria, humildemente, de fazer um pedido, não deixe de seguir meu perfil no Instagram® (@profthiagocavalcanti), onde eu posto, sistematicamente, questões comentadas e dicas semanais.

Para facilitar sua vida, você pode usar o QR code acima para acessar meu perfil no Instagram. Se precisar falar comigo por e-mail, mande mensagem para:

rcthiago@gmail.com

Por fim, e talvez a dica mais importante relacionada a redes sociais, gostaria de apresentar a vocês o meu canal no Telegram®(https://t.me/profthiagocavalcanti) ... neste canal procuro condensar todas as dicas que apresento nas minhas redes sociais. Na minha opinião, é a melhor forma de acompanhar todas as minhas publicações sem precisar ficar procurando nas redes sociais, lá eu ainda tiro dúvidas (no chat do canal) e interajo diretamente com os alunos. Ou seja, é uma forma de otimizar seus estudos!

Agora que você já me conhece! Vamos seguir em frente com o nosso curso!

PARE TUDO! E PRESTE ATENÇÃO!!

Hoje eu faço parte de uma equipe **SENSACIONAL** de professores: **OS CANETAS PRETAS**! Depois de muita luta conseguimos reunir **um time** de profissionais extremamente **QUALIFICADO** e sobretudo **COMPROMETIDO** em fazer o melhor pelos alunos. Para tal, criamos um conjunto de ações para nos aproximarmos dos alunos, entendermos suas necessidades e evoluirmos nosso material para um patamar ainda mais diferenciado. São 3 as novidades que gostaria de convidá-lo a conhecer:

//estratégia tech ESTRATEGIA CONCURSOS	Nosso podcast alternativo ... livre, descontraído e com dicas rápidas que todo **CANETA PRETA** raiz gosta de ouvir. Já temos alguns episódios disponíveis e vários outros serão gravados... acompanhe em: http://anchor.fm/estrategia-tech
Telegram a new era of messaging	Nosso grupo do Telegram! É um local onde ouvimos os alunos e trocamos ideias. Está crescendo a cada dia. A regra do grupo é: só vale falar sobre concursos. Lá divulgamos nossas aulas ao vivo e falamos sobre os concursos abertos, expectativas de novos concursos, revisões de véspera ... **https://t.me/canetaspretaschat**
Instagram	Criamos um perfil no Instagram ... e qual o objetivo? Fazer com que os alunos percam tempo nas redes sociais? Claro que não!! Estamos consolidando diversos posts dos professores! São dicas especiais, um patrimônio que deve ser explorado por todos os concurseiros de TI ... https://www.instagram.com/canetas.pretas/

MOTIVAÇÃO PARA O CURSO

Preparar esse curso é um desafio! Consolidar de forma amigável o conhecimento de banco de dados, análise de informações ou business Intelligence para concursos não é uma tarefa fácil. Calibrar o nível do teórico associado a uma didática eficiente tem sido minha meta nos últimos anos. Separamos o conteúdo de forma a segmentar e impulsionar seu aprendizado. Para que você entre na primeira aula com um pouco mais de segurança, vou aproveitar para fazer uma rápida apresentação sobre o assunto.

Você já ouviu falar sobre **Data Science ou Ciência dos Dados**? É um conceito relativamente recente que agrupa diversas atividades executadas sobre um conjunto de dados, em especial, sobre grandes conjuntos de dados. Para analisar os dados eles precisam estar **armazenados e organizados** de maneira **conveniente** para os cientistas dos dados. Essa base de dados facilita o trabalho e o entendimento do conteúdo armazenado.

Cientistas de dados são uma nova geração de especialistas em análise que têm habilidades técnicas para resolver problemas complexos e a curiosidade de explorar quais são os problemas que precisam ser resolvidos. A solução desses problemas passa por analisar os dados presentes em um banco de dados. Neste curso veremos o passo-a-passo para a construção de um banco de dados.

Nossa primeira aula deve inserir você no universo dos bancos de dados. Um banco pode ser visto como uma estrutura que armazena algo, por exemplo, um banco de leite guarda leite materno para que possa ser reutilizado de forma adequada em momentos posteriores. Um banco de dados guarda dados que devem ser controlados de forma adequada. É nesse momento que surge um sistema para “cuidar” do acesso consistente aos dados.

Os sistemas de gerenciamento de banco de dados (SGBDs) contribuem para a disponibilidade de um conjunto de informações para diferentes usuários simultaneamente. É preciso decidir quais dados armazenar, estruturar e manter na base de dados. Para controlar esse sistema e todo o desenvolvimento do projeto e da infraestrutura associada ao sistema de banco de dados várias tarefas têm que ser feitas.

Veremos que existem profissionais dedicados a tarefas específicas. Veremos ainda que a construção de um banco de dados, em especial um banco de dados relacional, passa por algumas etapas bem definidas. Essas etapas criam modelos de dados ou esquemas que permitem um melhor entendimento da estrutura de dados da organização ao tentar abstrair a complexidade presente no armazenamento físico dos dados.

Todos esses conceitos serão vistos em detalhes nas próximas páginas. Ao final, teremos nossa tradicional lista de exercícios. Espero conseguir contribuir para a sua aprovação. Vamos em frente?!

Teremos muito trabalho! Por isso, montamos um **curso teórico em PDF,** baseado nas mais diversas bancas, em especial da banca do seu concurso[1], apresentando o conteúdo observando as variadas formas de cobrança do mesmo pelas bancas examinadoras.

Teremos ainda videoaulas que apresentam o conteúdo teórico de forma detalhada para todo o conteúdo deste curso. Caso você não esteja visualizando os vídeos, peço que entre em contato comigo, o mais rápido possível, para que eu possa associá-los às respectivas aulas.

Ao final deste curso, nosso objetivo é garantir que você tenha capacidade e conhecimento para ser aprovado.

Observação importante: este curso é protegido por direitos autorais (copyright), nos termos da Lei 9.610/98, que altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências.

Grupos de rateio e pirataria são clandestinos, violam a lei e prejudicam os professores que elaboram os cursos. Valorize o trabalho de nossa equipe adquirindo os cursos honestamente através do site Estratégia Concursos ;-)

Observação importante II: todo o conteúdo deste curso encontra-se completo em nossos textos escritos. As videoaulas visam reforçar o aprendizado, especialmente para aqueles que possuem maior facilidade de aprendizado com vídeos e/ou querem ter mais uma opção didática.

Motivação I: Para se inspirar, aqui estão algumas frases motivacionais que podem ajudar a manter o foco e a determinação nos estudos:

“As raízes do estudo são amargas, mas seus frutos são doces.” - Aristóteles
“Não deixe seu futuro nas mãos da sorte, comece hoje mesmo a estudar e lutar pelo seu sucesso.” - Provérbio Chinês
“O lucro dos nossos estudos é tornarmo-nos melhores e mais sábios.” - Michel de Montaigne
“Investir em conhecimento rende sempre os melhores juros.” - Benjamin Franklin
“Estudar é crescer em silêncio!”

Agora vamos voltar para a nossa aula. Vamos juntos? Se você tiver alguma dúvida, por favor, não hesite em perguntar.

[1] Sempre que possível a lista de questões comentadas virá com diversas questões da sua banca. Entretanto, é possível que, por questões de didática ou carência de questões, existam questões de outras bancas nas aulas do seu curso.

Índice de figuras 3
Business Intelligence. Data Warehouse. 5
Business Intelligence - Conceitos básicos 5
Business Intelligence 7
Componentes de um sistema de Business Intelligence (BI) 10
SELF Business Intelligence (BI) 13
Governança de dados 15
Questões Business Intelligence Comentadas 20
Data Warehouse 40
Conceitos e características 40
Tipos de DW 45
Processo de DW 49
Arquitetura de DW 52
Kimball x Inmon 56
Questões Data Warehouse Comentadas 59
Data lake 76
Democratização de dados 78
Nível de Maturidade de um Data Lake 79
Criando um Data Lake 80
O pântano de dados (data swamp) 83
Trilha para o Sucesso em Data Lake 84
Arquiteturas de Data Lake 90
Questões 96
Data Mesh 104
Mudanças técnicas e organizacionais 104
Princípios de Data Mesh 106
Visão geral do modelo de malha de dados 111
Questões 112
Modelagem multidimensional 114
Esquemas multidimensionais 119
Processo de design dimensional 125

Revisitando o modelo *127*
Tipos de tabela fato *127*
Questões de Modelagem Comentadas *136*
Resumo **162**
Data Warehouse E Modelagem Dimensional *162*
Data Lake *165*
Data Mesh *167*
Exercícios **168**
Business Intelligence *169*
Data Warehouse *175*
Modelagem Dimensional *185*
Gabarito **200**
Considerações Finais **201**

ÍNDICE DE FIGURAS

Figura 1 - Conceitos associados a Business Intelligence 6
Figura 2 - Objetivos de um Data Warehouse 8
Figura 3 - Processo de construção do ambiente de DW/BI. 9
Figura 4 - Arquitetura de alto nível do BI 11
Figura 5 - Os 9 P´s da governança de dados. 19
Figura 6 - Data Warehouse orientado por assunto. 41
Figura 7 - Data Warehouse integração 42
Figura 8 - Data Warehouse é não volátil. 42
Figura 9 - Conceito de granularidade 44
Figura 10 - Conceito de credibilidade 45
Figura 11 - Orientações para um projeto de Data Warehouse 45
Figura 12 - Processo de Data Warehousing 49
Figura 13 - Quatros componentes do ambiente de DW/BI 52
Figura 14 - Arquitetura de data marts independentes 53
Figura 15 - Arquitetura de barramento de Data Mart (KIMBALL) 54
Figura 16 - Arquitetura Hub-and-spoke (INMON) 55
Figura 17 - Arquitetura de Data Warehouse centralizado 55
Figura 18 - Arquitetura de armazém de dados federada 56
Figura 19 - Arquitetura do Kimball x Inmon 57
Figura 20 - Comparação do Inmon com o Kimball 58
Figura 21 - Os 4 estágios de maturidade de um Data Lake 80
Figura 22 - Um pântano de dados (data swamp) 84
Figura 23 - Diferentes arquiteturas de data lake 86
Figura 24 - Zonas típicas de um data lake 87
Figura 25 - Expectativas de governança por zona 88
Figura 26 - Os quatro estágios da análise 89
Figura 27 - Gerenciando dados no data lake lógico 93
Figura 28 - Criando um conjunto de dados personalizado por meio de uma visualização 94
Figura 29 - Fornecendo metadados por meio de um catálogo 95
Figura 30 - Provisionamento e governança de dados por meio do catálogo 96

Figura 31 - Preocupações do modelo dimensional 114
Figura 32 - Tabela fato de vendas 115
Figura 33 - Tipos de medidas presentes nas tabelas fatos 116
Figura 34 - Exemplo de dimensão produto 117
Figura 35 - Esquema das tabelas fato e dimensões 118
Figura 36 - Modelo estrela (star schema) 119
Figura 37 - Resumo das características do modelo estrela 120
Figura 38 - Constelação com 2 tabelas fato: vendas e inventário. 122
Figura 39 - Esquema floco de neve para vendas 123
Figura 40 - Resumo das características do modelo floco de neve 123
Figura 41 - Resumo dos esquemas multidimensionais 124
Figura 42 - Processo de design dimensional 125
Figura 43 - Exemplo de tabela fato transacional 128
Figura 44 - Estrutura de uma tabela fato 130

BUSINESS INTELLIGENCE. DATA WAREHOUSE.

Começamos hoje um conjunto de aulas relacionadas ao conceito de ***Business Intelligence (BI)*** ou **Sistemas de Suporte à Decisão (SSD)**. A literatura especializada trata esses dois termos de forma bem semelhante, eles estão associados a um conjunto de *buzzwords* ou jargões que estão na moda. Mas será que eles entregam algum valor de fato? Como os elementos que fazem parte destas estruturas se relacionam? É isso que vamos estudar na aula de hoje ...

BUSINESS INTELLIGENCE - CONCEITOS BÁSICOS

O objetivo desta nossa aula é fazer uma introdução, apresentando os termos técnicos associados ao assunto e seus conceitos. Começaremos apresentando os conceitos da arquitetura de BI, em seguida trataremos especificamente de ***Data Warehouse (DW)***. Para construir um DW precisamos modelar sua estrutura de dados, isso é feito com a **modelagem multidimensional**. Assim, logo após apresentarmos os conceitos de DW vamos tratar da modelagem dimensional.

Ok! Vamos começar do início! Pessoas tomam decisões a todo momento. Essas decisões, dentro do contexto organizacional ou corporativo, podem trazer consequências boas ou ruins, por isso devem ser tomadas com cuidado. É aí que entra o suporte da tecnologia. Com ferramentas específicas podemos tomar decisões mais balizadas e consistentes. Diante destes desafios surgem os primeiros sistemas de suporte à decisão.

O conjunto de tecnologias que dão **suporte às decisões gerenciais** por meio de **informações** internas e externas às organizações é o que veremos a partir de agora. Essas tecnologias têm um profundo impacto na **estratégia corporativa**, na **performance** e na **competitividade**. Esse conjunto de tecnologias é coletivamente conhecido como ***BUSINESS INTELLIGENCE* (BI).**

Observem a nuvem de palavras abaixo. Ela nos mostra uma gama de elementos que está associada ao conceito de BI. Caso você ainda não esteja familiarizado com o assunto, tenha paciência, ao longo desta aula apresentaremos os aspectos relevantes de cada um dos termos presentes nesta figura. Neste momento gostaria de comentar sobre as siglas presentes na figura: **DSS, EIS/ESS e OLAP**.

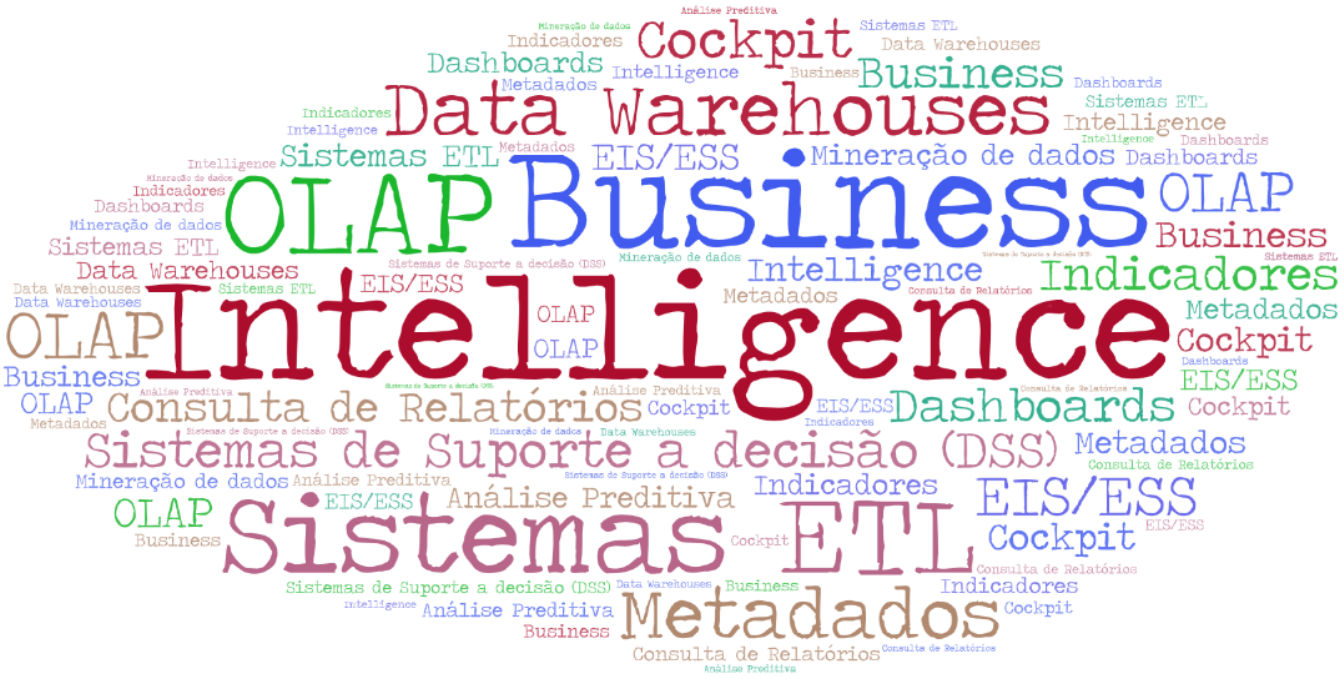


Figura 1 - Conceitos associados a Business Intelligence

À medida que subimos no nível decisório das organizações, as decisões carregam consigo um escopo de influência e impacto financeiro maiores. Isso nos incentiva a tomar a decisões corretas. Neste momento surge a necessidade de sistemas de informação para assistir ou auxiliar os tomadores de decisões. Esses sistemas foram inicialmente chamados de **Sistemas de Suporte à Decisão (DSS)**.

Um sistema de suporte à decisão ou ***Decision Support System*** (DSS) é um sistema de informação que suporta atividades de tomada de decisões empresariais ou organizacionais, resultando tipicamente na ordenação, classificação ou escolha entre alternativas.

Um sistema de informação executiva ou ***Executive Information System*** (EIS), também conhecido como um sistema de suporte executivo ou ***Executive Support System*** (ESS), é um tipo de sistema de informação gerencial que suporta as necessidades de informação e a tomada de decisões dos executivos de uma organização. Ele fornece acesso fácil a informações internas e externas relevantes aos objetivos organizacionais. É considerado uma forma especializada do DSS.

Por fim, OLAP – **On-line Analytical Processsing**, ou seja, processamento analítico online se refere a uma variedade de atividades normalmente executadas por usuários finais. Os produtos OLAP oferecem recursos de modelagem, análise e visualização de grandes conjuntos de dados. Em outras palavras, é uma tecnologia utilizada para integrar e disponibilizar informações gerenciais contidas em diversas bases de dados, em especial em ***Data Warehouses***.

O termo DSS, propriamente dito, tem sido utilizado cada vez menos, tanto em livros e revistas, quanto na internet. Atualmente, podemos dizer que esse conceito foi **modernizado** e esses sistemas são denominados de Sistemas de ***Business Intelligence* (BI).**

BI é um conceito empregado a ferramentas, tecnologias e metodologias, que tem como objetivo fornecer informações estratégicas, que apoiam a tomada de decisão. O termo *Business Intelligence* foi cunhado por **Howard Dresner** em **1989**, e foi descrito como *"conceitos e métodos para melhorar a tomada de decisão de negócio utilizando sistemas de suporte baseados em fatos"*. Simplificando: BI é transformar dados em informações úteis (conhecimento).

Vejamos outra definição de BI: "Um conjunto de **conceitos, métodos e recursos tecnológicos** que habilitam a obtenção e distribuição de informações geradas a partir de dados operacionais, históricos e externos, visando proporcionar **subsídios para a tomada de decisões** gerenciais e estratégicas". A partir desta definição vamos avançar pelo assunto, mas antes vamos fazer uma questão sobre o tema:

CEBRASPE (CESPE) - 2024 - Analista em Ciência e Tecnologia I (CNPq)/Gestão e Acompanhamento de Projetos e Programas em CT&I

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.

Comentário: Certamente! A Inteligência de Negócios (BI) pode ser utilizada pelas empresas para apoiar uma vasta gama de decisões empresariais, desde operacionais até estratégicas. Decisões operacionais básicas incluem o posicionamento ou a precificação de produtos. Já as decisões estratégicas envolvem prioridades, metas e direções em um nível mais abrangente. Em todos os casos, a BI é mais eficaz quando combina dados derivados do mercado em que a empresa atua (dados externos) com dados provenientes de fontes internas, como informações financeiras e operacionais (dados internos). Quando combinados, os dados externos e internos podem fornecer um panorama completo, criando uma "inteligência" que não pode ser obtida de nenhum conjunto singular de dados.

Gabarito: Certo

BUSINESS INTELLIGENCE

Já conhecemos o conceito de BI. Agora vamos continuar nosso estudo. Primeiramente vamos fazer um paralelo entre **sistemas operacionais (OLTP - On-line Transaction Processing) e sistemas analíticos (OLAP - On-line Analytical Processing)**. Os sistemas operacionais tratam das tarefas que fazem **parte do dia a dia** das organizações. Quando

pensamos neste tipo de sistema, o exemplo clássico é um caixa de supermercado. Ele basicamente registra a quantidade de cada produto, calcula o preço, informa para o cliente e recebe o pagamento. São tarefas que fazem parte de um fluxo de compras de cada cliente individualmente.

Quando surge a necessidade de **agregar os dados** das compras dos diferentes clientes visando prover informações aos gestores, entramos numa seara atendida pelos sistemas analíticos. E é neste contexto que surge o armazém de dados ou *Data Warehouse* (DW). Sua função é armazenar os dados de forma padronizada, capturando informações dos diversos sistemas operacionais da empresa.

O Kimball, um dos principais autores de DW, sugere alguns objetivos fundamentais para o DW dentro do contexto de BI. Ele diz que devemos fazer **a informação acessível** mais facilmente, apresentar **a informação consistente** (credibilidade), prover um sistema que seja adaptado a mudanças e apresentar a **informação de forma temporal.** Além disso, o DW pode ser considerado um **bastião da segurança** que protege os ativos de informação, servindo como base de autoridade e de confiança para uma melhor tomada de decisão.

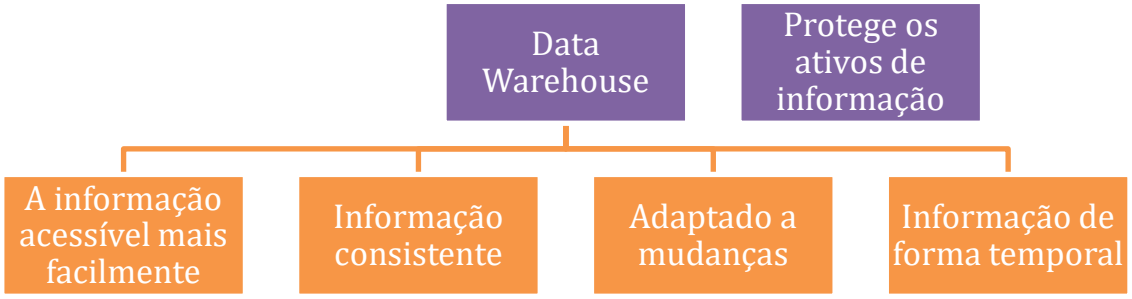


Figura 2 - Objetivos de um Data Warehouse

Para uma implementação de sucesso, um projeto de DW/BI deve ser **bem aceito pela comunidade organizacional**. É sugerido então um passo a passo para atender as demandas e estruturar de forma organizada as necessidades do negócio dentro do sistema.

Primeiramente precisamos (1) **compreender os usuários de negócios**. Nesta etapa é importante entender as responsabilidades de trabalho, metas e objetivos. Devemos então determinar as decisões que os usuários de negócios querem fazer com a ajuda do sistema de DW/BI. Assim podemos identificar os "melhores" usuários, que tomam decisões eficazes e de alto impacto, bem como, encontrar potenciais novos usuários e torná-los conscientes das capacidades do sistema de DW/BI.

Na etapa subsequente, é importante (2) **fornecer informações e análises de alta qualidade, relevante e acessível** para os usuários de negócios. Para atingir esse objetivo, escolha as fontes de dados mais robustas para apresentar no sistema de DW/BI, elas devem ser cuidadosamente selecionadas a partir do universo de possíveis fontes de dados em sua organização.

Neste momento (3) faça **interfaces de usuários e aplicações simples e baseadas em modelos**, explicitamente correspondentes aos perfis de processamento cognitivo dos usuários. Certifique-se que os dados são precisos e podem ser confiáveis, rotulando-os de forma consistente em toda a empresa. É preciso monitorar continuamente a precisão dos dados e o resultado das análises. É necessário ainda, adaptar-se às mudanças de perfis de

usuário, requisitos e prioridades de negócios, juntamente com a disponibilidade de novas fontes de dados.

Depois de desenvolver uma estrutura consistente, a próxima etapa do processo é (4) **sustentar o ambiente DW/BI**. É necessário que se tome uma parte do crédito das decisões de negócios feitas com o auxílio do sistema de DW/BI e se utilize desse sucesso para justificar o custo pessoal e gastos contínuos. Assim é possível **atualizar o sistema em um intervalo de tempo regular**.

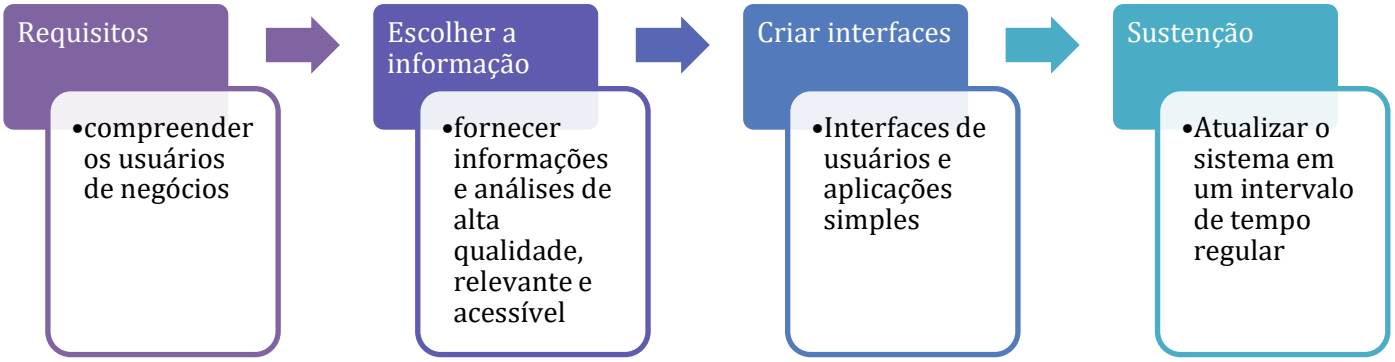


Figura 3 - Processo de construção do ambiente de DW/BI.

Perceba que, de posse dessas informações, percebemos que o objetivo máximo do ambiente de BI é fornecer aos gestores um conjunto de informações consolidadas para tomada de decisão. Isso já foi cobrado em provas anteriores ... quer ver? ...

(Ano: 2020 Banca: AOCP Órgão: MJSP Provas: AOCP - 2020 - MJSP - Cientista de Dados - Big Data) Em um BI, existem grandes desafios para descrição e visualização dos dados. Assinale a alternativa que apresenta o principal desses desafios.

A) A definição de uma estrutura de armazenamento de dados que possibilite a sua recuperação de forma rápida e dinâmica.

B) A captura e o tratamento dos dados de forma que possam ser organizados em estruturas multivariadas para a apresentação das informações de forma clara e objetiva.

C) A definição de resumos apropriados e a exibição desses resumos de forma que a complexidade do processo de negócios seja compreensível para o usuário.

D) A clareza na implementação da camada de apresentação de modo que o usuário compreenda como a informação foi extraída e apresentada.

E) A falta de ferramentas e processos de testes da carga de dados para aprimorar a apresentação das informações ao usuário.

_____

Comentários: Vamos comentar cada uma das alternativas ...

A. Errado. Definir a estrutura do Data Warehouse certamente é um desafio, mas não é considerado o principal dentro do contexto de BI.
B. Errado. A parte do processo ETL é importante, mas ela não constrói estrutura multivariadas ... (A análise multivariada consiste em um conjunto de métodos estatísticos utilizados em situações em que várias variáveis são medidas simultaneamente em cada elemento amostral.)
C. CERTO!! Essa é a nossa resposta!! Veja que o objetivo é facilitar a visão do tomador de decisão, ele precisa enxergar os dados de forma organizada a partir da consolidação ou agregação deles. Logo, é importante que os dados representem resumos apropriados às necessidades dos usuários.
D. Errado. O usuário não quer saber como a informação foi extraída, ele se preocupa apenas com o resultado do processo.
E. Errado. Essa alternativa está confusa ... a falta de ferramentas de teste para aprimorar a apresentação ao usuário? Não faz sentido!!

Gabarito: LETRA C.

COMPONENTES DE UM SISTEMA DE BUSINESS INTELLIGENCE (BI)

Business Intelligence (BI) é um termo abrangente que combina **arquiteturas**, **ferramentas**, **bancos de dados**, **ferramentas de análise**, **aplicações** e **metodologias**. Seus principais objetivos incluem:

1. Permitir o acesso interativo, por vezes em tempo real, aos dados.
2. Permitir a manipulação de dados.
3. Dar aos gestores e analistas a capacidade de realizar análise adequada.

Segundo o Turban[1], BI é baseado na transformação dos dados em informação, em seguida, informações em decisões e, finalmente, em ações. Como você pode observar, BI é um termo "guarda-chuva" que engloba diversos componentes. Para visualizar melhor essa estrutura, vamos descrever a arquitetura e os componentes de BI.

O BI tem **quatro grandes componentes**:

1. Um **data warehouse** (DW) com seus dados-fonte utilizados para a análise de negócios.
2. A **análise de negócio ou business analytics**, uma coleção de ferramentas para manipular e analisar os dados no data warehouse, incluindo data mining.
3. **Business performance management (BPM)** para monitorar e analisar indicadores de desempenho
4. Uma **interface de usuário** fornece uma capacidade visual para os dados solicitados pelos tomadores de decisão. Dashboards, por exemplo, permitem uma compreensão profunda e intuitiva de uma situação complexa ou problemática, muitas vezes revelando soluções inovadoras, conhecidas como **insights**.

[1] Efraim Turban (M.B.A., Ph.D, Universidade da Califórnia, Berkeley) é professor convidado no Pacific Institute for Information System Management, na Universidade do Havaí. Autor de vários livros, incluindo Business Intelligence: Um enfoque gerencial para a inteligência do negócio.

Observe que o ambiente de data warehousing é, sobretudo, de responsabilidade de uma equipe técnica, e o ambiente de análise (também conhecido como análise de negócios) está no âmbito dos usuários de negócios. A figura a seguir tenta apresentar esses conceitos de forma mais organizada dentro da arquitetura de BI.

Dentro das ferramentas de interface com os usuários temos o dashboard. Os **dashboards** fornecem **uma visão abrangente e visual das medidas** (indicadores-chave de desempenho), tendências e exceções do desempenho corporativo provenientes de múltiplas áreas do negócio. Os gráficos mostram o desempenho real em comparação às métricas desejadas, propiciando **uma visão imediata da saúde da organização**. Outras ferramentas que "transmitem" informações são **portais corporativos, cockpits digitais e outras ferramentas de visualização**.

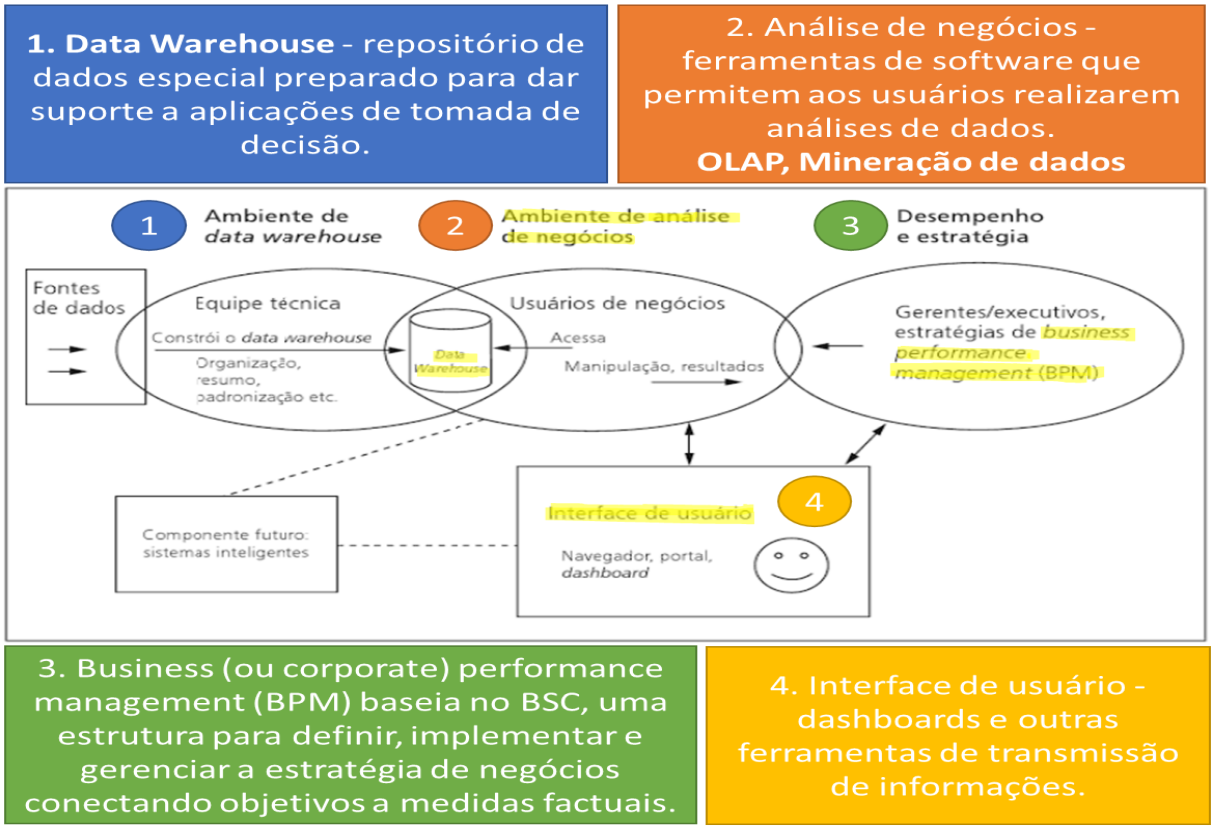


Figura 4 - Arquitetura de alto nível do BI

A figura acima apresenta os componentes de um sistema de BI. Basicamente são quatro os componentes: O *Data Warehouse* (DW) com suas fontes de dados; o *Business Analytics*, uma coleção de ferramentas para manipulação, mineração, análise de dados do DW; o Business performance management (BPM) para monitoramento e análise de performance e a Interface com o usuário (por exemplo, um *dashboard*).

O **armazém de dados e suas variantes são a pedra fundamental (pedra angular)** de qualquer sistema de BI de médio à grande porte. Originalmente, incluiu apenas dados históricos que foram organizados e resumidos, para que os usuários finais pudessem facilmente ver ou manipular dados e informações. Hoje, incluem também dados atuais para que eles possam fornecer apoio à decisão em tempo real.

Os usuários finais podem trabalhar com os dados e informações em um armazém de dados usando uma variedade de ferramentas e técnicas. Estas análises de negócio se enquadram em duas categorias principais:

1. Relatórios e consultas e
2. Mineração de dados.

Análise de negócios inclui relatórios estáticos e dinâmicos, todos os tipos de consultas, a descoberta de informações, visão multidimensional, *drill down*, entre outros. Esses relatórios também estão relacionados com BPM.

A mineração de dados, seja estruturados ou não, e outras ferramentas matemáticas e estatísticas sofisticadas fazem parte do segundo grupo de técnicas relacionadas a análise de negócios. A mineração de dados é um processo de busca de relações desconhecidas ou informações em grandes bases de dados ou armazéns de dados, utilizando ferramentas inteligentes como a computação neural, técnicas de análise preditiva, ou métodos estatísticos avançados.

Também conhecido como *corporate performance management* (CPM), o business performance management (BPM) nos apresenta uma carteira emergente de aplicativos e metodologias que contém a evolução da arquitetura e ferramentas de *BI* em seu núcleo. BPM amplia o monitoramento, medição e comparação das vendas, lucro, custos, rentabilidade e outros indicadores de desempenho, introduzindo o conceito de gestão e feedback. Considera processos, tais como planejamento e previsão, como fundamentais em uma estratégia de negócios.

Em contraste com os tradicionais DSS, EIS, e BI, que suportam a extração *bottom-up* de informação a partir dos dados, BPM proporciona uma aplicação *top-down* da estratégia corporativa. Normalmente sua implementação é combinada com a metodologia *Balanced Scorecard* (BSC) e *dashboards*.

O último componente de um sistema de BI é a interface com os usuários. Os *dashboards* (ou painéis) fornecem uma visão abrangente e visual das medidas de desempenho corporativo (também conhecido como indicadores chave de desempenho - KPI), tendências e exceções. As interfaces integram informações de várias áreas de negócio. Apresentam gráficos que mostram o desempenho real em comparação com as métricas desejadas, assim, um painel apresenta uma visão geral da saúde da organização.

Além dos painéis, outras ferramentas que transmitem informações são portais corporativos, *cockpits* digitais e outras ferramentas de visualização. Muitas ferramentas de visualização, que vão desde apresentação do cubo multidimensional à realidade virtual, são parte integrante dos sistemas de BI. BI surgiu de EIS, portanto, recursos visuais para os executivos foram transformados em software de BI. Tecnologias como sistemas de informações geográficas (SIG) desempenham um papel crescente no apoio à decisão. Vejamos uma questão sobre o assunto:

CEBRASPE (CESPE) - 2023 - Analista (MPE RO)/Sistemas

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.

Comentário: Vamos comentar cada uma das altenativas:

A) **Fontes de dados:** Fontes de dados são os locais ou sistemas de onde as informações são extraídas para serem utilizadas em um sistema de Inteligência de Negócios (BI). Exemplos de fontes de dados incluem bancos de dados, arquivos CSV, APIs, entre outros. Os dashboards, por outro lado, não são fontes de dados; eles são ferramentas para visualizar e interagir com os dados extraídos dessas fontes.

B) **Insights:** Dashboards são projetados para fornecer insights visuais sobre os dados, permitindo que os usuários compreendam rapidamente padrões, tendências e informações relevantes. Eles facilitam a interpretação dos dados de forma intuitiva, ajudando na tomada de decisões informadas.

C) **Usados no ETL:** ETL (Extração, Transformação e Carga) é um processo essencial na preparação dos dados para análise, envolvendo a extração de dados de diversas fontes, a transformação desses dados para um formato adequado e a carga dos dados transformados em um armazém de dados.

D) **Armazéns de dados:** Armazéns de dados são grandes repositórios onde os dados estruturados são armazenados para análise e consulta.

E) **Modelos semânticos de dados:** Modelos semânticos de dados são estruturas que atribuem significado aos dados, facilitando sua interpretação e uso. Eles ajudam a organizar e definir as relações entre diferentes conjuntos de dados.

Gabarito: B

SELF BUSINESS INTELLIGENCE (BI)

Self BI é uma abordagem à análise de dados que permite que os usuários de negócios acessem e trabalhem com dados corporativos, mesmo que não tenham experiência em análise estatística ou mineração de dados. As ferramentas de Self BI permitem que os usuários filtrem, classifiquem, analisem e visualizem dados sem envolver as equipes de BI e de TI da organização.

As empresas orientadas a dados implementam recursos de Self BI para permitir que os usuários corporativos **utilizem e se beneficiem facilmente dos dados coletados** e **gerem resultados comerciais positivos**, como melhor eficiência, ganhos de clientes ou lucros maiores.

Com ferramentas tradicionais de BI, cientistas de dados e equipes de TI controlam o acesso aos dados. Os usuários que solicitam novos relatórios e painéis enviam uma lista de requisitos de negócios que, uma vez aprovado, pode levar semanas para que os dados sejam extraídos, transformados e carregados em um data warehouse. Nesta situação, a equipe de TI ou BI produz o relatório ou o painel.

Por outro lado, uma arquitetura de Self BI é usada por pessoas que talvez não tenham conhecimentos de tecnologia. Portanto, é imperativo que **a interface do usuário (UI) para o software de análise seja intuitiva**. Painéis de controle amigáveis e navegação devem atender às necessidades de usuários ocasionais (aqueles que podem precisar acessar dados, mas não gerar relatórios) e usuários avançados (usuários mais experientes, responsáveis não apenas por acessar e analisar os dados, mas também pela construção de relatórios ad hoc).

Idealmente, o treinamento deve ser fornecido para ajudar os usuários a entender quais dados estão disponíveis e como essas informações podem ser consultadas para tomar decisões baseadas em dados para solucionar problemas de negócios. Uma vez que o departamento de TI tenha configurado o data warehouse e os data marts que suportam o sistema de Self BI, os usuários corporativos devem poder consultar os dados e criar relatórios personalizados com pouco esforço.

O Self BI permite que os usuários de negócios acessem, analisem e modelem dados, o que pode levar a respostas mais rápidas e mais ágeis com o acesso direto as informações de dados quando comparado com o modelo de BI tradicional.

Ao permitir que os usuários finais tomem decisões com base em suas próprias consultas e análises, as organizações liberam as equipes de BI e TI para criar outros relatórios e se concentrarem em outras tarefas que ajudarão a organização a alcançar seus objetivos. A maior agilidade e eficiência podem ajudar os usuários e departamentos de negócios a agir com mais rapidez nas percepções de dados.

No entanto, reunir e analisar solicitações de recursos de ferramentas de Self BI de usuários corporativos pode ser desgastante e demorado. Além disso, embora a Self BI incentive os usuários a basearem suas decisões em dados em vez de intuição, o acesso a dados que ela fornece pode causar problemas, como análises e relatórios imprecisos, se não houver uma política de controle de dados consistente.

Entre outras coisas, a política deve definir as principais medidas para determinar o sucesso, que processos devem ser seguidos para criar e compartilhar relatórios, quais privilégios são necessários para acessar dados confidenciais e como a qualidade, segurança e privacidade dos dados serão mantidas. Os fornecedores de software analítico que já ofereceram ferramentas de BI para analistas agora também oferecem ferramentas de Self BI. Algumas das muitas opções de análise de autoatendimento vêm de fornecedores como Birst, Domo, Google, IBM, Microsoft, Qlik, Salesforce, SAP, Sisense e Tableau.

Facilidade de uso, **sofisticação** e **funcionalidades específicas** diferem para a ferramenta de Self BI de cada fornecedor.

Algumas plataformas podem ser usadas principalmente na construção de painéis e visualizações simples, em vez de tarefas mais complicadas, como preparação de dados, descoberta de dados ou exploração visual interativa. Veja a comparação dos conceitos básicos associados a Self BI e BI tradicional na tabela abaixo:

BI tradicional	Self-service BI
Usuários de negócios reúnem requisitos para um relatório ou dashboard	Time de TI reúne requisitos para a ferramenta de Self-service
Usuários submetem as requisições para a TI	As ferramentas de self-service são implementadas para dar aos usuários de negócio acesso aos dados.
A TI extrai os dados e carrega dentro dos repositórios analíticos (data warehouse) para análise.	Usuários de negócios têm acesso aos dados diretamente.
A TI que cria o modelo de dados.	Usuários de negócios preparam os dados para incluir.
Usuários aprovam os relatórios ou dashboards ou requisitam mudanças.	Usuários de negócio criam os modelos de dados.

GOVERNANÇA DE DADOS

Para darmos início ao estudo de governança de dados, precisamos fazer uma diferenciação entre os conceitos de gestão e governança.

Segundo o Guia DAMA-DMBOK, "**Gestão de Dados** é a função na organização que **cuida do planejamento, controle e entrega de ativos de dados e de informação**. Esta função inclui: as disciplinas do desenvolvimento, execução e supervisão de planos, políticas, programas, projetos, processos, práticas, e procedimentos que controlam, protegem, distribuem e aperfeiçoam o valor dos ativos de dados e informações".

Já a governança pode ser vista como "**o exercício de autoridade e controle** (planejamento, monitoramento e

execução) **sobre o gerenciamento de ativos de dados**. A Governança de Dados é um planejamento e controle de alto nível sobre o gerenciamento de dados".

Segundo a versão atual do guia DAMA-DMBOK, a Gestão de Dados é uma disciplina formada pelo conjunto de onze funções de gerenciamento de dados integradas, que podem ser observadas na figura acima. A integração dessas funções é feita pela função de Governança de Dados, por esta razão ela está localizada como elemento central do framework do DAMA-DMBOK.

Áreas de Conhecimento ou Funções

As 11 áreas de conhecimento (ou funções) de gerenciamento de dados são:

- **Governança de dados**: planejamento, supervisão e controle sobre o gerenciamento de dados e o uso de dados e recursos relacionados a dados.
- **Arquitetura de dados**: a estrutura geral de dados e recursos relacionados a dados como uma parte da arquitetura da empresa.
- **Modelagem e Design de Dados**: análise, projeto, construção, teste e manutenção.
- **Armazenamento e Operações de Dados**: implantação de armazenamento de ativos de dados físicos estruturados e gestão.
- **Segurança de Dados**: garantindo privacidade, confidencialidade e acesso apropriado.
- **Integração de Dados e Interoperabilidade**: aquisição, extração, transformação, movimento, entrega, replicação, federação, virtualização e suporte operacional (uma área de conhecimento nova em DMBOKv2)
- **Documentos e Conteúdo**: armazenar, proteger, indexar e habilitar o acesso aos dados encontrados em fontes não estruturadas (arquivos eletrônicos e registros físicos) e disponibilizando esses dados para integração e interoperabilidade com dados estruturados (banco de dados).
- **Dados Mestre e Referência**: gerenciando dados compartilhados para reduzir a redundância e garantir qualidade de dados através da definição padronizada e uso de valores de dados.
- **Data Warehousing & Business Intelligence**: gerenciamento de processamento de dados analíticos e permitindo acesso a dados de suporte à decisão para relatórios e análises.
- **Metadados**: coletando, categorizando, mantendo, integrando, controlando, gerenciando e entregando de metadados
- **Qualidade dos dados**: definindo, monitorando, mantendo a integridade dos dados e melhorando a qualidade dos dados.

Os objetivos do Guia DAMA-DMBOK2 são:

1. Criar consenso para uma visão geralmente aplicável das áreas de conhecimento de gerenciamento de dados.

2. Fornecer definições padrão para áreas de conhecimento de gerenciamento de dados comumente usadas, entregáveis, funções e outras terminologias, em conjunto com o DAMA Dictionary of Data Management e, assim, promover uma padronização de conceitos e atividades.

3. Identificar princípios orientadores para o gerenciamento de dados.

4. Esclarecer o escopo e os limites das atividades de gerenciamento de dados.

5. Fornecer uma visão geral das boas práticas comumente aceitas, técnicas amplamente adotadas e abordagens alternativas significativas, sem referência a fornecedores de tecnologia específicos ou seus produtos.

6. Apresentar questões organizacionais e culturais comuns.

7. Identificar estratégias para análise de maturidade de gerenciamento de dados.

8. Fornecer recursos adicionais e material de referência para melhor entendimento do gerenciamento de dados

Governança de dados é a organização e implementação de políticas, procedimentos, estrutura, papéis e responsabilidades que delineiam e reforçam regras de comprometimento, direitos decisórios e prestação de contas para garantir o gerenciamento apropriado dos ativos de dados. Governança de dados muito burocrática é um convite à desobediência; já o excesso de flexibilidade pode levar à desgovernança, a uma gestão de dados menos eficiente. Por isso, deve-se começar pela definição dos **princípios**. Abaixo apresentamos uma lista com alguns desses princípios.

Tabela 1 - Princípios de governança dos dados

Princípios	Descrição
Regra de ouro	Todos os dados são tratados como ativo corporativo.
Federação	Há padrões definidos para as estruturas de dados
Eficiência	Dados relevantes devem estar disponíveis no momento certo, no lugar certo, e no formato certo para usuários autorizados.
Qualidade	Dados corporativos são medidos para terem qualidade
Gestão de risco	Manter a conformidade com a legislação, políticas e normativos internos relativos a dados.

Colaboração	Dados corporativos são recursos compartilhados e tornados públicos.
Contextualização	O contexto de uso do dado muda sua forma de armazenamento, tratamento e utilização.
Inovação	Novas técnicas são incentivadas, seguindo-se os demais princípios

Outro conceito importante associado a governança dos dados é o de curadoria. Os curadores (*data stewards*, em inglês) são as pessoas ou grupos de pessoas que têm responsabilidades de cuidar dos dados sob sua alçada de negócio. Essa é uma mudança fundamental, pois compartilha com a TI a missão de cuidar dos dados corporativos.

Como vimos anteriormente, a governança de dados é um conceito que **engloba a gestão de dados de uma organização**, incluindo a forma como eles serão **usados, disponibilizados, compartilhados entre os colaboradores e mantidos de forma segura**. Esse controle é importante para qualquer organização que trabalhe com informação. Trata-se de um processo para garantir que os dados **atendam aos padrões e regras que um negócio precisa**, uma vez que eles estão inseridos em um sistema.

A governança de dados permite que as empresas exerçam **o controle sobre a gestão de ativos de dados**. Ela abrange as **pessoas**, os **processos** e a **tecnologia** que é necessária para garantir que os dados estejam sempre aptos para a sua finalidade. Esse controle é importante para diferentes tipos de organizações, especialmente aquelas que trabalham **em conformidade regulatória**. Para alcançar a conformidade, elas são obrigadas a ter processos de gerenciamento de dados formais, a fim de manter uma gestão de dados eficaz ao longo de seu ciclo de vida.

Resumindo, a governança de dados se refere aos seguintes pontos:

1. A **criação de regras** para a forma que os dados serão adquiridos, mantidos e usados pelos membros da organização. (fluxo de informação)
2. **Aspectos relacionados à segurança** e a forma como os dados serão acessados, para que apenas os colaboradores que precisam acessá-los tenham permissão para isso e no tempo necessário. (sigilo e segurança)
3. Definição de formas para controlar a **qualidade dos dados** registrados, para que não faltem detalhes importantes. (qualidade)
4. Criação da estrutura de dados seguindo todas as **questões regulatórias** envolvidas. (legalidade)

Outra forma de olhar para a governança dos dados é conhecer os 9P´s da governança: Patrocínio, Política de dados, Papéis/Pessoas, Processos, Padrões, Programas, Projetos e Procedimentos. A figura a seguir apresenta de forma organizada esses conceitos.

Figura 5 - Os 9 P´s da governança de dados.

QUESTÕES BUSINESS INTELLIGENCE COMENTADAS

1. FGV - Aud Est (CGE SC)/CGE SC/Ciências da Computação/2023 - TI - Banco de Dados - Data Warehouse e Data Mart

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).

Comentário: A afirmação correta é: 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.

Em um Data Warehouse, a normalização não é aplicada da mesma forma que em bancos de dados transacionais. Dados redundantes podem ser usados para melhorar o desempenho das consultas e facilitar a análise de negócios, uma vez que o foco principal é a capacidade de consulta eficiente em vez de minimizar a redundância de dados. Portanto, a normalização não é uma preocupação principal no contexto de modelos dimensionais, como o modelo estrela.

Gabarito: C

---

2. FGV - Aud Est (CGE SC)/CGE SC/Ciências da Computação/2023 - TI - Banco de Dados - Data Warehouse e Data Mart

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.

Comentário: A diferença funcional entre um banco de dados planejado para lidar com informações transacionais e um Data Warehouse é: 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.

Um banco de dados transacional é projetado principalmente para dar suporte às operações diárias da empresa, registrando transações, atualizando registros e garantindo a integridade dos dados. Por outro lado, um Data Warehouse é projetado para ser usado como uma fonte central de dados para análise de negócios, permitindo consultas complexas e análises de tendências ao longo do tempo em dados consolidados de várias fontes. Portanto, a finalidade e a orientação desses dois tipos de bancos de dados são diferentes.

Gabarito: A

---

3. FGV - Aud Est (CGE SC)/CGE SC/Ciências da Computação/2023- TI - Banco de Dados - Data Warehouse e Data Mart

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**.

Comentário: Todos os itens mencionados fazem parte dos componentes de um Data Warehouse. Portanto, a resposta correta é: e) I, II, III e IV.

Abaixo deixo a descrição de cada um dos itens:

I. **Sistemas de Origem**: São os sistemas que contêm os dados de origem que alimentam o Data Warehouse. Esses sistemas podem incluir bancos de dados

transacionais, sistemas legados, aplicativos de terceiros e qualquer fonte de dados que contenha informações relevantes para análise.

II. **Infraestrutura de ETL (Extração, Transformação e Carga)**: ETL refere-se ao processo de coletar dados de várias fontes (extração), transformá-los em um formato adequado para análise e carregá-los no Data Warehouse. A infraestrutura de ETL inclui ferramentas, servidores e processos necessários para executar essas operações.

III. **Data Warehouse**: É o repositório central onde os dados de origem são armazenados e organizados para fins de análise. O Data Warehouse é projetado para suportar consultas complexas e oferecer um ambiente otimizado para análise de negócios. Pode ser composto por um ou mais data marts, dependendo da arquitetura escolhida.

IV. **Aplicações de Front-end para o Usuário Final**: São as interfaces de usuário e aplicativos que permitem que os usuários finais acessem e interajam com os dados armazenados no Data Warehouse. Isso pode incluir painéis, relatórios, ferramentas de visualização de dados e outras aplicações que facilitam a análise e a tomada de decisões com base nos dados.

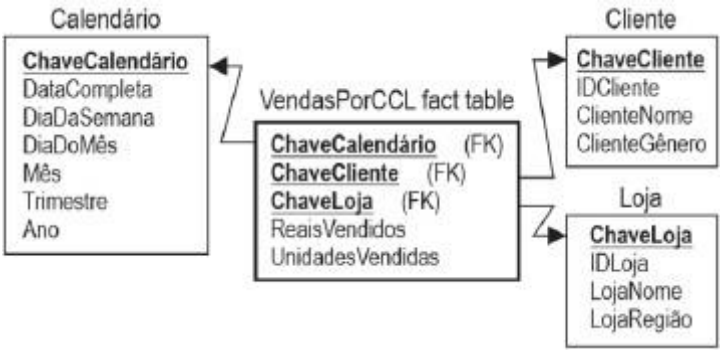
Esses componentes trabalham juntos para criar um ambiente de Business Intelligence (BI) que permite às organizações coletar, organizar, analisar e relatar dados de forma eficaz para fins de análise de negócios e tomada de decisões.

Gabarito: E

---

4. FGV - Aud Est (CGE SC)/CGE SC/Ciências da Computação/2023 - TI - Banco de Dados - Data Warehouse e Data Mart

Uma rede de lojas de departamentos usa o modelo dimensional estrela conforme o seguinte diagrama:



(Os atributos sublinhados denotam chave primária)

Observando o aumento na quantidade de reclamações dos clientes nas lojas, os analistas de BI resolveram incluir as informações analiticamente úteis da base de reclamações no Data Warehouse.

Para que a criação da constelação de fatos (*também chamada de galáxia*) contemple o fato RECLAMAÇÃO, os analistas devem adicionar

a) uma tabela de fato RECLAMAÇÕES, contendo apenas um atributo descritivo, sem a necessidade de conectar a qualquer dimensão.

b) uma tabela de fato RECLAMAÇÕES, contendo um atributo descritivo e três chaves estrangeiras, uma para cada uma das dimensões existentes.

c) uma tabela de fato RECLAMAÇÕES, contendo um atributo descritivo e três atributos que receberão os valores das chaves estrangeiras de Loja, Cliente e RegistroReclamação diretamente do banco de dados operacional.

d) três tabelas de dimensão (CalendárioReclamação, ClienteReclamação e LojaReclamação) mais uma tabela de fato RECLAMAÇÕES, contendo um atributo descritivo e três chaves estrangeiras, uma para cada uma das dimensões recém-criadas.

e) duas tabelas de dimensão (ClienteReclamação e LojaReclamação) mais uma tabela de fato RECLAMAÇÕES, contendo um atributo descritivo e três chaves estrangeiras, duas para cada uma das dimensões recém-criadas e uma para referenciar o registro da reclamação diretamente do banco de dados operacional.

Comentário: Uma constelação de fatos é um modelo de design avançado no contexto de um Data Warehouse. Nesse modelo, várias tabelas de fatos são conectadas a uma ou várias tabelas de dimensões compartilhadas. Essa abordagem é uma estratégia sofisticada para lidar com complexidade analítica e demandas de vários processos de negócios em uma organização.

Aqui estão os principais pontos desse conceito:

Múltiplas Tabelas de Fatos: Uma característica fundamental da constelação de fatos é a presença de várias tabelas de fatos. Cada tabela de fatos é projetada para atender a um processo de negócios específico ou a uma perspectiva analítica particular. Isso significa que diferentes partes da organização podem ter suas próprias tabelas de fatos dedicadas às suas métricas e requisitos exclusivos.

Dimensões Compartilhadas: Embora existam várias tabelas de fatos, elas compartilham um conjunto comum de tabelas de dimensões. Essas tabelas de dimensões fornecem contextos para análise e permitem que os dados das tabelas de fatos sejam relacionados a informações relevantes. Ter dimensões compartilhadas ajuda a manter a consistência e a integridade dos dados analíticos em toda a organização.

Contexto e Flexibilidade: Cada tabela de fatos contém métricas específicas de um processo de negócios ou de uma área de análise. Isso significa que as métricas relevantes para vendas podem estar em uma tabela de fatos separada daquelas relacionadas às operações ou ao marketing. As dimensões comuns permitem que os usuários conectem essas métricas a contextos relevantes, independentemente de qual tabela de fatos estejam consultando. Isso oferece flexibilidade e precisão nas análises.

Gerenciamento da Complexidade: A constelação de fatos é uma estratégia poderosa para gerenciar a complexidade analítica em organizações com diversos processos e necessidades analíticas. Em vez de tentar encaixar todas as métricas em uma única

tabela de fatos, ela reconhece que diferentes áreas da organização têm necessidades distintas e fornece estruturas específicas para atender a essas necessidades.

Em resumo, a constelação de fatos é uma abordagem avançada de modelagem de dados em Data Warehouses oferecem flexibilidade, precisão e gerenciamento eficaz da complexidade analítica. Ela permite que uma organização atenda a diversas demandas de análise, mantendo a consistência dos dados por meio do compartilhamento de dimensões. Isso resulta em uma melhor tomada de decisões e insights mais profundos para todos os setores da empresa.

Sendo assim, A letra B é a única alternativa compatível com a definição de Constelação de Fatos.

Gabarito: B

---

5. FGV - Aud Est (CGE SC)/CGE SC/Ciências da Computação/2023 -TI - Banco de Dados - Data Warehouse e Data Mart

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.

Comentário: A proposta geral do modelo dimensional em um Data Warehouse é projetada para ser flexível e atender às necessidades analíticas da organização. Vamos analisar cada afirmação:

a) **Cobre tanto dados detalhados quanto dados sumarizados**: Correto. O modelo dimensional acomoda tanto dados detalhados (nível granular) quanto dados sumarizados (agregados) para permitir análises em diferentes níveis de detalhe.

b) **Cobre toda a empresa, e não apenas departamentos**: Correto. O modelo dimensional é projetado para abranger toda a empresa, permitindo que diferentes áreas e departamentos acessem dados relevantes para suas análises.

c) **É escalável, podendo entregar relatórios com trilhões de registros**: Geralmente correto. O modelo dimensional é escalável e pode lidar com grandes volumes de dados, embora a capacidade específica dependa da infraestrutura e do hardware do Data Warehouse.

d) **É arquitetado apenas para um uso previsível, geralmente cobrindo os 10 relatórios mais acessados**: Incorreto. O modelo dimensional é projetado para suportar uma ampla gama de consultas e análises, não apenas as 10 mais acessadas. Ele é flexível o suficiente para acomodar uma variedade de consultas imprevisíveis.

e) **Pode integrar diversas fontes de dados operacionais da empresa, inclusive fontes externas**: Correto. O modelo dimensional é capaz de integrar dados de várias fontes, tanto internas quanto externas à empresa, para fornecer uma visão abrangente e unificada.

Portanto, a afirmação incorreta é a letra **d**, pois o modelo dimensional não é limitado apenas aos 10 relatórios mais acessados, mas é projetado para suportar uma ampla gama de consultas e análises.

Gabarito: D

---

6. FGV - ACE (TCE ES)/TCE ES/Tecnologia da Informação/2023 - TI - Banco de Dados - Data Warehouse e Data Mart

Ana está desenvolvendo um banco de dados analítico a partir da integração de dados do sistema de pagamento com o sistema de gestão de pessoal. No sistema de pagamento, os colaboradores são identificados pelo CPF. No sistema de gestão de pessoal, os colaboradores são identificados pelas iniciais do seu nome concatenadas com sua data de nascimento. Ana sabe que essas chaves primárias naturais apresentam diversas desvantagens e riscos para um ambiente de análise de dados integrados, como seu reuso e alteração de regras de composição, além de questões de desempenho.

Com isso, para carregar os dados no banco de dados analítico, Ana desenvolveu um ETL que substituiu as chaves naturais dos sistemas por uma chave artificial contendo inteiros simples sequenciais, utilizando uma:

a) Composite key;

b) Foreing key;

c) Surrogate key;

d) Business Key;

e) Production key.

Comentário: Ana adotou uma abordagem comum em Data Warehouses ao substituir chaves primárias naturais por chaves artificiais. Essas chaves artificiais são chamadas de **surrogate keys**.

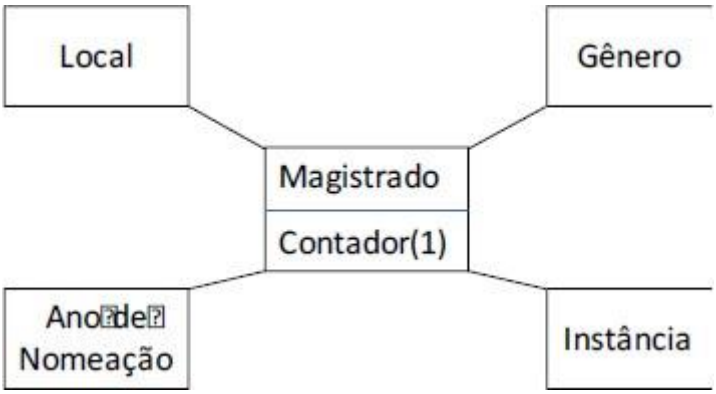
Portanto, a alternativa correta é a letra **c)**, "Surrogate key". Essas chaves são usadas para evitar os problemas associados às chaves primárias naturais ao integrar dados de diferentes fontes em um ambiente de análise de dados. Elas são atribuídas a cada registro de forma única e são usadas como identificadores únicos dentro do Data Warehouse.

Gabarito: C

---

7. FGV - AJ (TJ RN)/TJ RN/Apoio Especializado/Análise de Sistemas/2023 - TI - Banco de Dados - Data Warehouse e Data Mart

Observe o modelo multidimensional de dados de um Data Warehouse a seguir.



O tipo de modelagem multidimensional empregada é:

a) Star Schema;

b) Snowflake Schema;

c) Fact Constellation;

d) Data Lake;

e) Data Mart.

Comentário: Os data warehouses são projetados com um foco especial na criação de modelos que ofereçam características de visões conceituais multidimensionais. Essa abordagem é fundamental para permitir a análise eficaz dos dados. Os modelos multidimensionais têm como objetivo principal aproveitar os relacionamentos naturais e complexos que existem nos dados. Para facilitar esse processo, os dados são organizados em estruturas conhecidas como cubos de dados.

Dois esquemas multidimensionais amplamente utilizados são o esquema estrela (star scheme) e o esquema floco de neve (snowflake scheme). O esquema em estrela é caracterizado pela presença de uma tabela de fatos central que se conecta a várias tabelas de dimensões. Cada tabela de dimensão representa um aspecto específico dos dados e contém informações detalhadas sobre essa dimensão.

Por outro lado, o esquema floco de neve é uma variação do esquema em estrela. Nesse esquema, as tabelas de dimensões são normalizadas, o que significa que elas são divididas em subdimensiones, criando uma hierarquia de tabelas relacionadas. Isso ajuda a economizar espaço de armazenamento, mas pode complicar um pouco a recuperação dos dados, uma vez que é necessário realizar várias junções para acessar informações detalhadas.

Em resumo, esses esquemas multidimensionais desempenham um papel crucial na estruturação e organização dos dados em data warehouses, tornando-os prontos para análises complexas e eficazes. A escolha entre um esquema estrela e um esquema floco de neve dependerá dos requisitos específicos de cada projeto de data warehousing e das necessidades de análise dos dados.

Desta forma, observamos que a figura descreve um modelo estrela ou Star Schema. E a nossa resposta está presente na alternativa A.

Gabarito: A

8. FGV - AJ (TJ RN)/TJ RN/Apoio Especializado/Análise de Sistemas/2023- TI - Banco de Dados - Data Warehouse e Data Mart

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.

Comentário: Para implementar as ações sobre os dados em um ETL (Extração, Transformação e Carga), Julia utilizou:a) steps e fluxos de dados.

Em um processo de ETL, as ações sobre os dados são geralmente divididas em etapas (steps) que envolvem a extração, a limpeza, a agregação, a transformação e, por fim, a carga dos dados em um banco de dados destino. Essas etapas são organizadas em fluxos de dados que representam a sequência lógica das operações. Portanto, "steps" e "fluxos de dados" são conceitos essenciais na implementação de um ETL eficaz. O uso de um repositório de metadados também é comum para gerenciar e documentar o processo de ETL, mas não é especificamente uma parte das ações sobre os dados. Os termos "sequências temporais", "regras de associação" e "data mining" não são diretamente relacionados à implementação das ações típicas de um ETL.

Gabarito: A

---

9. FGV - AJ (TJ RN)/TJ RN/Apoio Especializado/Análise de Suporte/2023 - TI - Banco de Dados - Data Warehouse e Data Mart

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.

Comentário: O banco de dados a ser desenvolvido é um Data Warehouse. Nas características mencionadas, como orientado a assunto, não volátil e histórico, integrando dados estruturados de diversos sistemas e contextos, o tipo de banco de dados que se encaixa é um Data Warehouse (ou Data Warehouse Analítico). Esses bancos de dados são projetados especificamente para armazenar grandes volumes de dados históricos de diferentes fontes, a fim de apoiar a análise e a tomada de decisões de alto nível em uma organização. Portanto, a resposta correta é "e) Warehouse".

Gabarito: E

---

10. FGV - Ana (BBTS)/BBTS/Perfil Tecnológico/2023 - TI - Banco de Dados - Data Warehouse e Data Mart

Com relação ao ETL, a diferença de tempo entre quando os dados são gerados no sistema de origem e quando os dados estão disponíveis para uso no sistema de destino, denomina-se

a) latência.

b) retrocesso.

c) replicação.

d) sobrecarga.

e) anacronismo.

Comentário: A diferença de tempo entre quando os dados são gerados no sistema de origem e quando os dados estão disponíveis para uso no sistema de destino é denominada **latência**.

Portanto, a alternativa correta é "a) latência". A latência representa o atraso ou intervalo de tempo entre a ocorrência de um evento nos sistemas de origem e a disponibilidade desses dados para serem processados ou consultados nos sistemas de destino, como parte do processo de ETL (Extração, Transformação e Carga) em um ambiente de Data Warehouse ou integração de dados.

Gabarito: A

HORA DE PRATICAR!

11. Analista (Prefeitura de Vila Velha)/Desenvolvimento/2020

O processo de pesquisa, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de um negócio é conhecido pela sigla:

a) AFP.

b) SGBD.

c) BI.

d) ERP.

e) GED

Comentário: Questão tranquila! O conceito Business Intelligence se enquadra perfeitamente com o enunciado! Segundo a Wikipédia ... Inteligência de negócios refere-se ao processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de negócios. As demais alternativas se referem a outros conceitos, a saber:

A) *Adaptive Project Framework* (APF), também conhecido como Adaptive Project Management (APM), foi criado para se adaptar constantemente às mudanças no ambiente de um projeto. Assim, nada se fixa com esta abordagem - nem a duração do projeto, nem o orçamento, nem os riscos, e é possível ajustar tudo continuamente de acordo com as mudanças nas características do projeto.

B) Sistema de Gerenciamento de Banco de Dados (SGBD) é um conjunto de ferramentas ou programas que permitem o gerenciamento e a manutenção de um banco de dados.

D) A sigla ERP significa "Enterprise Resource Planning", ou sistema de gestão integrado. Essa tecnologia auxilia o gestor da empresa a melhorar os processos internos e integrar as atividades de diferentes setores, como vendas, finanças, estoque e recursos humanos.

E) O Gerenciamento Eletrônico de Documentos (GED) é uma tecnologia que facilita o controle, armazenamento, compartilhamento e recuperação das informações existentes de determinada Instituição.

Gabarito: C.

HORA DE PRATICAR!

12. Analista Legislativo (ALAP)/Atividade de Tecnologia da Informação/Desenvolvedor de Sistemas/2020

Para construir um Data Warehouse, algumas etapas e processos são necessários. Uma etapa é conhecida como ETL, que compreende as etapas de Extração, Transformação e Armazenagem de dados em Sistemas Específicos ou Armazéns de Dados. Essas etapas são constituídas de várias outras funções, processos e técnicas de data integration. Uma dessas funções chama-se Master Data Management − MDM e é responsável por

a) misturar os dados para criar um panorama virtual.

b) unir os dados para criar uma visão única deles, através de múltiplas fontes. Ela inclui tanto o ETL quanto capacidades de data integration, para misturar as informações e criar o "melhor registro".

c) monitorar e processar fluxos de dados e ajudar a tomar decisões mais rapidamente.

d) fornecer tanto agendamento em lote quanto capacidades em tempo real.

e) criar um ambiente de testes onde os dados possam ser integrados, limpos e padronizados (por exemplo: SP e São Paulo, Masculino e M, Senhora e Sra. etc) além de verificar e remover dados duplicados.

Comentário: Essa questão trata do conceitos de dados mestres. Eles são uma referência para a organização dos dados dentro de uma empresa. O gerenciamento de dados mestre (Master data management - MDM) faz parte da governança de dados e é o processo de unir os dados para criar uma visão única deles, através de múltiplas fontes. Ele inclui várias ações de ETL como a capacidade integração dos dados para misturar as informações e criar um "registro de ouro" ou um "melhor registro" associado a cada instância de um objeto.

Assim, temos nossa resposta na alternativa B.

Gabarito: B.

13. IBFC - Analista de Tecnologia da Informação (EBSERH)/2020

Dado os três conceitos técnicos abaixo, assinale a alternativa que corresponda respectivamente à tecnologia referente a cada um desses conceitos.

1. processo de explorar grandes quantidades de dados à procura de padrões consistentes.

2. refere-se ao processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de negócios.

3. depósito de dados digitais que serve para armazenar informações detalhadas relativamente a uma empresa.

a) 1.Data Warehouse - 2.Business Intelligence - 3.Data Mining

b) 1.Data Mining - 2.Data Warehouse - 3.Business Intelligence

c) 1.Business Intelligence - 2.Data Warehouse - 3.Data Mining

d) 1.Data Mining - 2.Business Intelligence - 3.Data Warehouse

e) 1.Business Intelligence - 2.Data Mining - 3.Data Warehouse

Comentário: Vamos descrever primeiramente a definição de cada um dos termos apresentados acima:

- ***Business Intelligence (BI)***, também chamado de inteligência de negócios ou inteligência empresarial, é o nome que se dá ao processo de coleta, organização, análise e disseminação de informações que auxiliam o empresário na tomada das decisões estratégicas e no planejamento.
- ***Data Mining*** **(Mineração de dados)**: é a prática de examinar dados que já foram coletados – utilizando diversos tipos de algoritmos –, a fim de gerar novas informações e encontrar padrões.

- ***Data Warehouse***: é um depósito de dados digitais que serve para armazenar informações detalhadas relativamente a uma empresa, criando e organizando relatórios através de históricos que são depois usados pela empresa para ajudar a tomar decisões importantes com base nos fatos apresentados.

Assim, podemos encontrar nossa resposta na alternativa D.

Gabarito: D.

14. FAEPESUL - Assistente (CRC SC)/Suporte em Informática/2019

É correto afirmar que Business Intelligence é:

a) O processo de coleta, organização, análise, compartilhamento e monitoramento de informações para a gestão de negócios.

b) Um software.

c) O mesmo que inteligência artificial.

d) O nome dado a um algoritmo de pesquisa.

e) Um padrão de projetos.

Comentário: Questão de definição ... BI é, segundo a Wikipédia, processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de negócios. Ou seja, nossa resposta está na alternativa A.

Gabarito: A.

15. CEBRASPE (CESPE) - Auditor Municipal de Controle Interno (CGM João Pessoa)/Tecnologia da Informação/Desenvolvimento de Sistemas/2018

A respeito de business intelligence, julgue o item.

Business intelligence pode ser definido como um processo inteligente de coleta, organização, análise, compartilhamento e monitoração de dados que, depois de processados, geram informações para o suporte e para a tomada de decisões no ambiente de negócios.

Comentário: Inteligência de negócios (ou Business Intelligence, em inglês) refere-se ao processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de negócios. É um conjunto de técnicas e ferramentas para auxiliar na transformação de dados brutos em informações significativas e uteis a fim de analisar o negócio. As tecnologias BI são capazes de suportar uma grande quantidade de dados

estruturados para ajudar a identificar, desenvolver e até mesmo criar uma oportunidade de negócios. O objetivo do BI é permitir uma fácil interpretação do grande volume de dados. Identificando novas oportunidades e implementando uma estratégia efetiva baseada nos dados, também pode promover negócios com vantagem competitiva no mercado e estabilidade a longo prazo.

Gabarito: C.

HORA DE PRATICAR!

16. Ano: 2018 Banca: CESGRANRIO Órgão: TRANSPETRO Cargo: Analista de processo de negócio Questão: 54

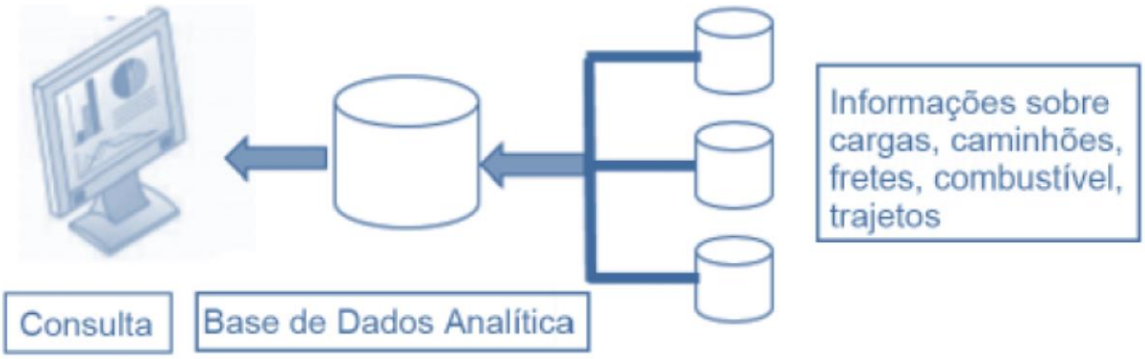
Determinada empresa de transporte possui uma frota de caminhões que movimenta diversos tipos de carga, tais como eletrônicos, brinquedos e eletrodomésticos. Um Sistema de Informações proprietário calcula detalhes financeiros e técnicos das viagens dessa frota. Os cálculos financeiros incluem, entre outros, custos de combustível, mão de obra e valor de frete. Os detalhes técnicos são inúmeros, como tipo e volume da carga, capacidade, consumo e velocidade dos caminhões, restrições dos trajetos, distâncias aos destinos e outros.

O sistema responde a perguntas, tais como:

i) dada uma especificação de carga, uma escala de entrega e preços de frete, quais caminhões e motoristas devem ser alocados para maximizar o lucro?

ii) qual conjunto (velocidade, trajeto) deve ser utilizado por determinado caminhão para otimizar o lucro e garantir as datas de entrega?

A Figura resume a configuração do sistema.



Adaptado de Laudon and Laudon. Management Information Systems: Managing the digital firm. 13 ed; Pearson 2014.

Com base na descrição acima, o tipo de Sistema de Informação utilizado por essa empresa é o

(A) CRM

(B) SIG

(C) Sistema Especialista

(D) Sistema de Suporte à Decisão

(E) Sistema de Processamento de Transações

Comentário: Neste caso a figura nos ajuda a responder a questão. Veja que a base de dados analítica é utilizada no contexto de suporte a decisão. Desta forma, podemos marcar nossa resposta na alternativa D.

Gabarito: D

17. BANCA: FCC ANO: 2016 ÓRGÃO: TRT - 14ª REGIÃO (RO E AC) PROVA: TÉCNICO JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

[35] Quando uma empresa utiliza Data Warehouse (DW) é necessário carregá-lo para permitir a análise comercial. Para isso, os dados de um ou mais sistemas devem ser extraídos e copiados para o DW em um processo conhecido como

a) ERP.

b) BI.

c) CRM.

d) ETL.

e) Data Mart.

Comentário: Essa questão é interessante pois apresenta sistemas ou soluções tecnológicas que fazem parte do ecossistema de Business Intelligence. Já falamos sobre eles na primeira parte da nossa aula. Os sistemas integrados (ERPs) são utilizados na automação das diversas áreas de uma organização, centralizando as informações em uma única base de dados.

CRM é um sistema que se preocupa com o relacionamento com o cliente. Todo contato do cliente com a empresa deve ser registrado, de forma que o cliente seja atendido de acordo com suas características e necessidades, baseado no seu comportamento anterior.

ETL e Data Mart fazem parte do fluxo básico de dados e informações dentro do contexto de BI. O ETL é responsável pela extração, transformação e carga dos dados das bases operacionais para a base de dados analítico. Sendo assim, podemos marcar nossa resposta na alternativa D.

Gabarito: D

HORA DE PRATICAR!

18. ANO: 2013 BANCA: ESAF ÓRGÃO: DNIT PROVA: ANALISTA ADMINISTRATIVO - TECNOLOGIA DA INFORMAÇÃO

O componente final do processo de Business Intelligence é

A Business balance management (BBM).

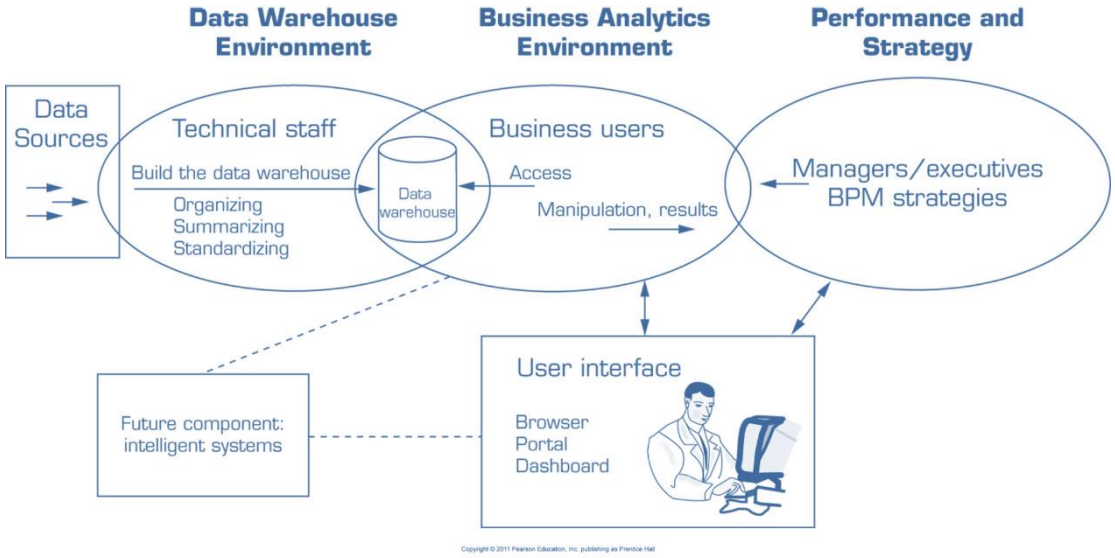
B Executive office team (EOT).

C Business performance management (BPM).

D Priority statement board (PSB).

E Business advisory management (BAM).

Comentários: Essa questão segue os conceitos descritos no livro do Turban que divide o processo de BI em quatro partes: 1. O ambiente do DW, com as fontes de dados, que define como os dados são armazenados; 2. O ambiente de análise de negócio que fornece um conjunto de ferramentas para minerar, manipular e analisar os dados; 3. A interface com o usuário que diz como apresentar e distribuir os dados consolidados; e 4. Performance e estratégia para monitorar e contribuir com a gestão da organização.



Aproveitando para falar um pouco mais sobre BPM, ele também é chamado de Corporate Performance Management (CPM). BPM é um conceito que veio ratificar a importância de ter sempre o alinhamento das informações com a estratégia da empresa. BPM (Business Performance Management), é um conjunto de software, processos de negócios e medidas de sucesso dos negócios (métricas e KPI's - key performance indicators) que, quando combinados, permitem a uma organização entender, agir e influenciar a performance de seus negócios.

Gabarito: C

19. ANO: 2010 BANCA: ESAF ÓRGÃO: MPOG PROVA: ANALISTA - TECNOLOGIA DA INFORMAÇÃO

BI – Business Intelligence

A é uma técnica de otimização da árvore de decisão.

B é um método de formação avançada de gestores.

C compreende ferramentas de análise de dados para otimizar os processos produtivos de uma empresa.

D são técnicas, métodos e ferramentas para mineração de dados na área de negócios de uma empresa.

E são técnicas, métodos e ferramentas de análise de dados para subsidiar processos de decisão de uma empresa.

Comentário: Vejam que a alternativa que converge para um dos conceitos de BI que vimos ao longo da nossa aula é a alternativa presente na letra E. Apenas para lembramos do conceito, vamos exibi-lo abaixo:

"BI representa a habilidade de se estruturar, acessar e explorar informações, normalmente guardadas em um DW/DM (Data Warehouse/Data Mart), com o objetivo de desenvolver percepções, entendimentos, conhecimento, os quais podem produzir um melhor processo de tomada de decisão". Essa definição é do autor brasileiro Carlos Barbieri.

Gabarito: E.

20. ANO: 2015 BANCA: FCC ÓRGÃO: CNMP PROVA: ANALISTA DO CNMP - DESENVOLVIMENTO DE SISTEMAS

Soluções informatizadas de Business Intelligence (BI) geralmente contêm sistemas que podem ser de diversos tipos, dependendo do objetivo das análises e do perfil do usuário, como:

A Online Analytical Processing (OLAP), também conhecidos como sintéticos, que baseiam-se em transações, como: Sistemas Contábeis; Aplicações de Cadastro; Sistemas de Compra, Estoque, Inventário; ERPs; CRMs.

B Decision Support Systems (DSS) ou Sistemas de Apoio a Decisão, voltados para profissionais que atuam no nível estratégico das empresas, como diretoria e presidência. Oferecem, para tanto, um conjunto de indicadores chave de desempenho como o CMMI.

C Management Information Systems (MIS) ou Sistemas de Informações Gerenciais, que permitem análises mais profundas, com a realização de simulações de cenários. Por vezes, utilizam-se de ferramentas de Data Mining para identificação de cruzamentos não triviais. São utilizados por analistas de negócio no nível tático.

D Online Transactional Processing (OLTP) ou Sistemas transacionais, que fornecem subsídio para tomadas de decisão a partir de análises realizadas sobre bases de dados históricas, por vezes com milhões de registros a serem totalizados.

E Executive Information Systems (EIS) ou Sistemas de Informações Executivas, que são baseados em relatórios analíticos, normalmente utilizados por usuários de nível operacional.

Comentários: Vamos ver o que tem de errado nas alternativas distintas a resposta da questão. Na alternativa A é listado um conjunto de sistemas operacionais e associado essa lista a sistemas OLAP, que se trata da sigla para sistemas analíticos.

Sabemos que CMMI não é um conjunto de indicadores básicos de desempenho. É um certificado do nível de maturidade no desenvolvimento de software de uma determinada organização. Sendo assim a alternativa B também está errada.

A alternativa C é a nossa resposta, trata dos Sistemas de informações gerenciais de forma correta.

Na alternativa D o erro está em associar os sistemas transacionais com subsídio para tomada de decisões.

Por fim, a alternativa E diz que relatórios analíticos são utilizados por usuários operacionais, o que está incorreto.

Gabarito: C

HORA DE PRATICAR!

21. ANO: 2010 BANCA: FCC ÓRGÃO: TCE-SP PROVA: AGENTE DA FISCALIZAÇÃO FINANCEIRA - CONHECIMENTOS BÁSICOS

Os conceitos de inteligência empresarial ou organizacional estão intimamente relacionados com o PETI que considera

A o planejamento de sistemas de informação, apenas.

B o planejamento de sistemas de informação e conhecimentos, apenas.

C a informática e os conhecimentos, apenas.

D a informática, apenas.

E o planejamento de sistemas de informação, conhecimentos e informática.

Comentários: O Planejamento Estratégico da Tecnologia da Informação (PETI) é um processo dinâmico e interativo para estruturar estratégica, tática e operacionalmente as informações

organizacionais, a TI (e seus recursos: hardware, software, sistemas de telecomunicação, gestão de dados e informação), os sistemas de informação e do conhecimento, as pessoas envolvidas e a infraestrutura necessária para o atendimento de todas as decisões, ações e respectivos processos da organização.

O PETI deve estar alinhado aos negócios. Para que esse alinhamento aconteça, o maior desafio dos gestores ainda é fazer com que a TI desempenhe seu relevante papel estratégico nas organizações, agregando valores aos seus produtos e/ou serviços e auxiliando a promoção das inteligências competitiva e empresarial. É fundamental, portanto, que os recursos computacionais da TI disponibilizem informações oportunas e conhecimentos personalizados que possibilitem a geração de cenários decisórios.

Gabarito: E

---

22. ANO: 2013 BANCA: FCC ÓRGÃO: MPE-CE PROVA: ANALISTA MINISTERIAL - CIÊNCIAS DA COMPUTAÇÃO

Em relação ao entendimento do significado do termo Business Intelligence (BI) e da solução que provê, a definição que **NÃO** é coerente com o termo Business Intelligence é a que

A consiste em uma metodologia que fornece objetivos de negócios ligados a objetivos de TI, provendo métricas e modelos de maturidade para medir a sua eficácia e identificando as responsabilidades relacionadas dos donos dos processos de negócios e de TI.

B se refere à aplicação de técnicas analíticas para informações sobre condições de negócio no sentido de melhorá-las, de uma maneira automatizada, mas com a interpretação e respostas humanas, de forma a melhorar a tomada de decisões.

C reúne recursos que provêm a habilidade para que a pessoa certa receba a informação adequada e no momento correto para tomar a melhor decisão.

D consiste em um sistema de negócios que inclui uma estrutura de busca efetiva e acessível, acurada, em tempo real, com informações e relatórios que permitam aos líderes das áreas de negócio se manterem informados para tomar decisões.

E é uma solução fácil de dizer, mas difícil de fazer corretamente pois envolve mudanças na forma como a organização conduz uma busca efetiva, bem como, a necessidade de se possuir uma base de dados de qualidade para que se possa tomar ações com o objetivo de otimizar a performance corporativa.

Comentário: Percebam primeiramente que a questão solicita a alternativa incorreta. Ou, conforme está descrito no enunciado, aquela que não tem relação com o termo inteligência de negócio. Analisando as alternativas percebemos que na alternativa A temos uma descrição alinhadas com os termos apresentados pelo framework do COBIT.

O COBIT tem como objetivo principal **o alinhamento entre os objetivos do negócio e os objetivos da TI**, fazendo com que a TI atenda às necessidades de negócio (requisitos de negócios) da maneira mais eficiente possível.

Em muitas organizações, a TI parece ter uma vida independente da empresa em que está inserida, muitas vezes sendo difícil para a alta direção compreender, por exemplo, como os investimentos aplicados nesta área ajudam a organização a atingir seus objetivos, suas metas.

O COBIT vai ao encontro desta e outras necessidades, ajudando a guiar os investimentos na área de TI, analisar riscos e atender a legislação pertinente.

Logo em seguida a alternativa ainda fala do dono do processo que é um conceito intimamente relacionado com o ITIL.

O **dono do processo** deve responder por um processo, garantindo que seja executado conforme acordado e documentado, atingindo os objetivos definidos independentemente de onde estão e quais são as tecnologias, os serviços e os profissionais envolvidos. É de responsabilidade (***accountable***) do dono do processo patrocinar, desenhar, gerenciar mudanças e focar na melhoria contínua do processo e suas métricas. Suas principais atribuições de Planejamento são definir a estratégia e desenho, documentação, políticas, padrões e publicação do processo, incluindo a definição dos Indicadores Chave de Performance (KPIs).

Desta forma, ela é a nossa resposta.

Gabarito: A

23. ANO: 2012 BANCA: FCC ÓRGÃO: TST PROVA: ANALISTA JUDICIÁRIO - ANALISTA DE SISTEMAS

Em Business Intelligence (BI), as consultas de dados que NÃO estão disponíveis em relatórios periódicos, ou seja, consultas criadas sob demanda especificamente para um conteúdo, layout ou cálculo, agilizando ou facilitando a tomada de decisão, são chamadas de consultas

A evolutivas.

B multidimensionais.

C single shot.

D data mining.

E ad hoc.

Comentário: Uma das características das ferramentas de BI é permitir aos usuários, em especial aos tomadores de decisão, a capacidade de realizar consultas sobre os dados. Essas consultas são baseadas em perguntas que surgem durante a elaboração de uma solução para um problema ou durante o planejamento de uma nova ação estratégica. Neste momento, os dados organizados em modelos multidimensionais facilitam o cruzamento de informações das diferentes dimensões por meio das consultas ***AD HOC***. Esse termo é derivado do latim e significa aleatória ou não prevista anteriormente. Logo, nossa resposta encontra-se na alternativa E.

Gabarito: E.

DATA WAREHOUSE

CONCEITOS E CARACTERÍSTICAS

Vamos apresentar a definição de DW de três diferentes autores:

> **Kimball**: Um data warehouse é uma **cópia de dados transacionais** especificamente estruturada para consulta e análise. Kimball também define o processo de data warehousing que é um conjunto de ferramentas e técnicas de projeto, que quando aplicadas às necessidades específicas dos usuários e aos bancos de dados específicos permitirá que planejem e construam um Data Warehouse.
>
> **Laudon&Laudon**: Banco de dados, com ferramentas de consulta e relatório, que armazena dados atuais e históricos extraídos de vários sistemas operacionais e consolidados para fins de análises e relatórios administrativos.
>
> **Inmon**: É uma coleção de dados orientados por assunto, integrados, variáveis com o tempo e não voláteis, para dar suporte ao processo de tomada de decisão.

Um primeiro comentário interessante sobre essas definições, em especial sobre as definições do Inmon e do Kimball, perceba que o Kimball apresenta dentro do conceito de data warehousing um conjunto de ferramentas e técnicas, ou seja, o processo de estruturar um banco de dados conhecido como DW é incluso dentro da definição. Inmon, por sua vez, considera apenas a coleção de dados.

A última definição acima traz consigo quatro características que são de suma importância para o entendimento do assunto. Juntam-se a elas a granularidade e a credibilidade dos dados. Essas características precisam ser entendidas para a compreensão do assunto.

Ser **orientado por assunto** refere-se ao fato do *Data Warehouse* armazenar informações sobre temas específicos importantes para o negócio da empresa. São exemplos típicos de temas: produtos, atividades, contas, clientes. Em contrapartida, quando observamos o ambiente operacional, percebemos que ele é organizado por aplicações **funcionais**. Por exemplo, em uma organização bancária, estas aplicações incluem empréstimos, investimentos e seguros. Observe um exemplo na figura abaixo.

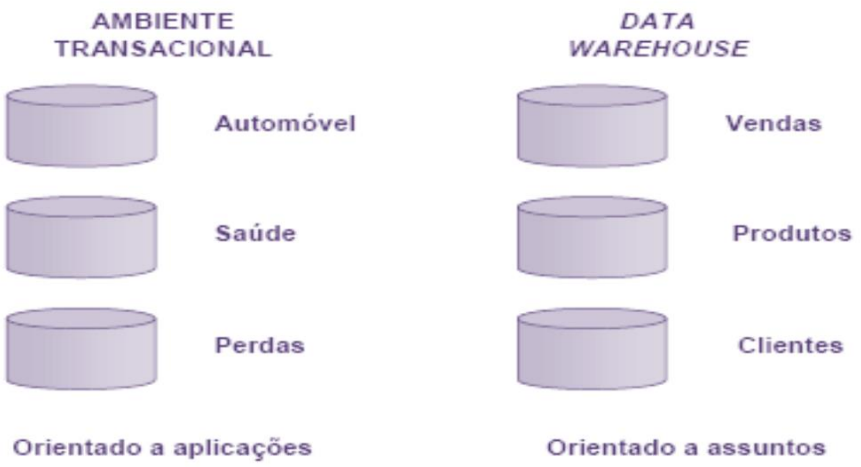


Figura 6 - Data Warehouse orientado por assunto.

Ser **integrado** refere-se à consistência de nomes, das unidades, das variáveis etc. É importante que os dados armazenados sejam transformados até um estado uniforme. Por exemplo, considere sexo como um elemento de dado. Uma aplicação pode codificar sexo como M/F, outra como 1/0 e uma terceira como H/M. Conforme os dados são inseridos ou repassados para o *Data Warehouse*, eles são convertidos para um mesmo padrão. O atributo Sexo, portanto, seria codificado apenas de uma forma.

Da mesma maneira, se um elemento de dado é medido em centímetros em uma aplicação, em polegadas em outra, ele será convertido para uma representação única ao ser colocado no *Data Warehouse*. Vejam na figura abaixo a ideia por trás do conceito de integração.

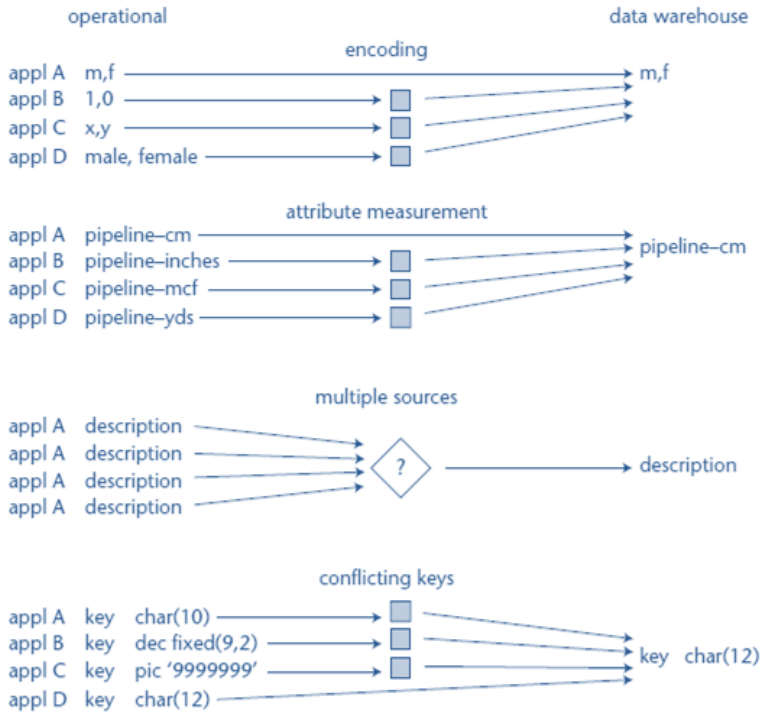


Figura 7 - Data Warehouse integração

O fato de ser **não volátil** significa que o *Data Warehouse* permite apenas a carga inicial dos dados e consultas a estes dados. Após serem integrados e transformados, os dados são carregados em bloco para o DW, para que estejam disponíveis aos usuários para acesso.

No ambiente operacional, ao contrário, os dados são, em geral, atualizados registro a registro, em múltiplas transações. Esta volatilidade requer um trabalho considerável para assegurar integridade e consistência através de atividades de *rollback*, recuperação de falhas, *commits* e bloqueios. Vejam abaixo uma figura que representa os diferentes ambientes e suas respectivas operações sobre os dados.

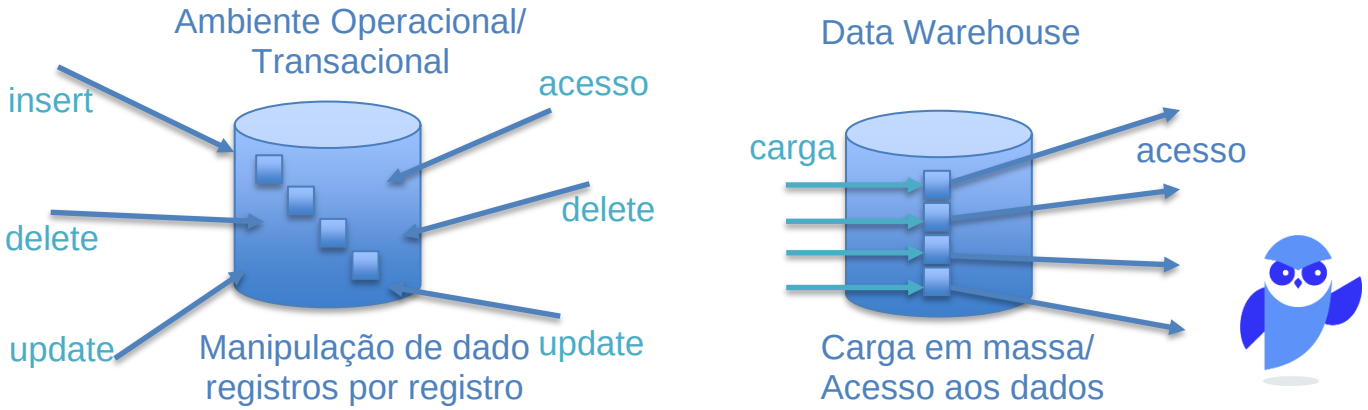


Figura 8 - Data Warehouse é não volátil.

Ser **variante no tempo** trata do fato de um registro em um *Data Warehouse* referir-se a algum momento específico, significando que ele não é atualizável. Enquanto o dado de produção é

atualizado de acordo com mudanças de estado do objeto em questão, refletindo, em geral, o estado do objeto no momento do acesso, em um DW, **a cada ocorrência de uma mudança, uma nova entrada é criada para marcar esta mudança**. Vejamos como isso já foi cobrado em provas anteriores:

CEBRASPE (CESPE) - 2023 - Analista de Tecnologia da Informação (DATAPREV)/Análise de Negócios

No que se refere à análise de dados e informações, julgue o item a seguir.

Data warehouse é um repositório de dados para análises e, portanto, deve conter apenas dados atualizados a fim de não enviesar a análise.

Comentário: Um data warehouse é um repositório de dados utilizado para análise e suporte à tomada de decisões, projetado para armazenar grandes volumes de **dados históricos** de várias fontes, oferecendo aos usuários uma visão consolidada e consistente das informações.

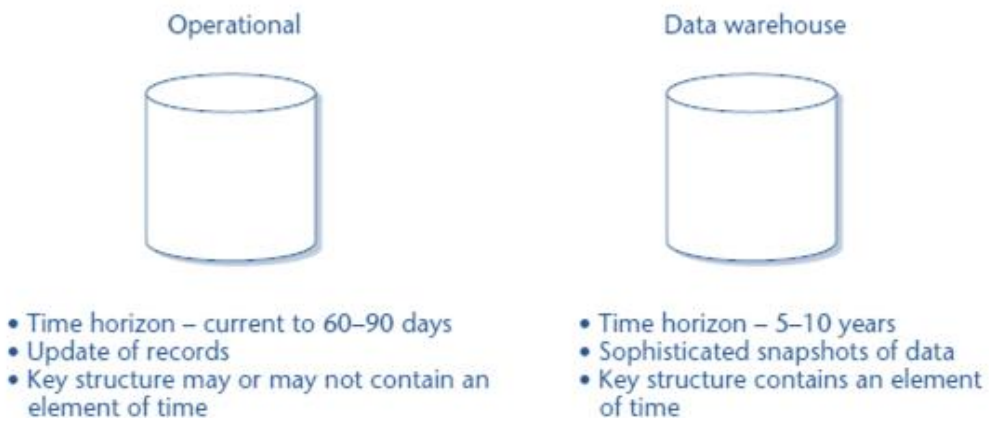
É **incorreto** afirmar que um data warehouse deve conter apenas dados atualizados. Na verdade, um data warehouse é tipicamente **carregado com dados históricos, frequentemente provenientes de diferentes fontes e sistemas transacionais**. Esses dados históricos são essenciais para análises de tendências ao longo do tempo, identificação de padrões e obtenção de insights sobre o desempenho passado da organização.

Embora seja crucial que os dados no data warehouse sejam atualizados periodicamente para garantir que as análises estejam baseadas em informações recentes, os dados históricos permanecem fundamentais. Eles são utilizados para diversos tipos de análises, como análise de tendências, previsões e modelagem estatística, proporcionando uma compreensão profunda e abrangente do desempenho e comportamento da organização ao longo do tempo.

Gabarito: Errado

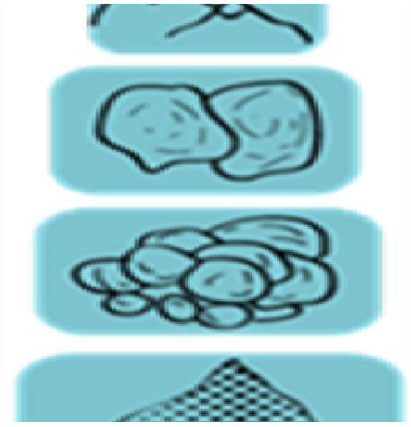
O tratamento de séries temporais apresenta características específicas, que adicionam complexidade ao ambiente do *Data Warehouse*. Deve-se considerar não apenas que os dados tenham uma característica temporal, mas também os metadados, que incluem definições dos itens de dados, rotinas de validação, algoritmos de derivação, etc. Sem a manutenção do histórico dos metadados, as mudanças das regras de negócio que afetam os dados no DW são perdidas, invalidando dados históricos.

Vejam na figura abaixo uma comparação entre os ambientes operacional e analítico no que se refere a variação dos dados ao longo do tempo.



A próxima característica que temos nos DW é o tratamento da **granularidade dos dados**. A granularidade de dados refere-se ao nível de sumarização dos elementos e de detalhe disponíveis nos dados, considerado **o mais importante aspecto** do projeto de um *Data Warehouse*. Em um nível de granularidade muito alto, o espaço em disco e o número de índices necessários se tornam bem menores, há, porém, uma diminuição da possibilidade de utilização dos dados para atender a consultas detalhadas.

Em outras palavras a granularidade diz respeito ao nível de detalhe ou de resumo contido nas unidades de dados existentes no DW. Quanto maior o nível de detalhes, menor o nível de granularidade. O nível de granularidade afeta diretamente o volume de dados armazenado no DW e ao mesmo tempo o tipo de consulta que pode ser respondida. Veja a figura abaixo:



A **granularidade de dados** refere-se ao nível de sumarização dos elementos e de detalhe disponíveis nos dados, considerado **o mais importante aspecto** do projeto de um *Data Warehouse*. Em um nível de granularidade muito alto, o espaço em disco e o número de índices necessários se tornam bem menores, há, porém, uma diminuição da possibilidade de utilização dos dados para atender a consultas detalhadas.

Figura 9 - Conceito de granularidade

Antes de serguirmos em frente, vamos fazer uma questão sobre o assunto:

CEBRASPE (CESPE) - 2023 - Analista (CNMP)/Tecnologia da Informação e Comunicação/Suporte e Infraestrutura

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.

Comentário: No contexto de data warehouse, o conceito de granularidade se refere ao nível de detalhe presente em uma unidade de dados. Quando uma unidade de dados possui muitos detalhes,

ela é considerada de baixa granularidade. Em contraste, se a unidade de dados é mais agregada ou resumida, ela é considerada de alta granularidade.

Gabarito: Errado

A última característica diz respeito a confiança ou **credibilidade dos dados**.

É necessário, para que as análises sejam consideradas corretas, que os dados tenham baixa dispersão e estejam consistentes com os valores observados. É importante lembrar que serão feitas várias manipulações nos dados, inclusive agregações e cálculos estatísticos. A depender dos desvios presentes nos dados, as análises podem ser inócuas ou incoerentes com a realidade.

Figura 10 - Conceito de credibilidade

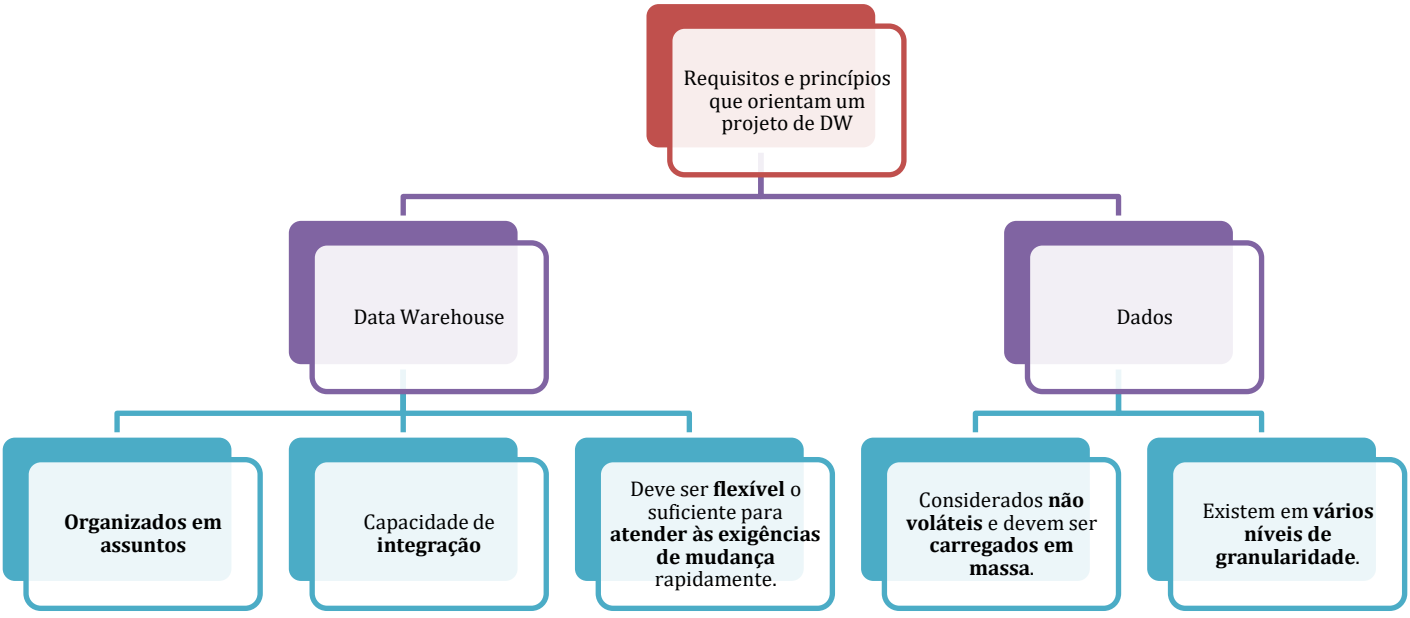


Figura 11 - Orientações para um projeto de Data Warehouse

TIPOS DE DW

A indústria atualmente reconhece pelo menos três tipos diferentes de *Data Warehouses*:

- ✓ Data Mart (DM)
- ✓ Data Warehouse Empresarial (EDW)
- ✓ Armazenamento de Dados Operacionais (ODS).

O qualificador "Empresarial" implica em ser grande e abrangente. Esta é a ideia tradicional de um armazém de dados. Algumas organizações nunca vão conseguir concluir um projeto de EDW, pois ele requer um compromisso enorme de recursos. Ainda assim, se bem feito, um sistema abrangente acrescenta toneladas de valor e produz um retorno considerável sobre o investimento.

É conveniente, por vezes, a criação de uma coleção menor de dados conhecida como Data Mart. Esta tem um público mais focado e normalmente consiste em um subconjunto do EDW. Esse subconjunto pode ser definido pela geografia (por exemplo, apenas os dados da Alemanha), linha de produtos (por exemplo, apenas produtos para os cabelos), ou área funcional (por exemplo, de fabricação).

O armazenamento de dados operacionais (ODS) se concentra em um prazo mais curto para a análise e assim é um subconjunto definido por tempo (por exemplo, apenas os dados desta semana). Há um processo bem definido para a criação e manutenção de um armazém de dados. Falaremos sobre ele mais adiante neste curso.

Considerando que um armazém de dados combina bases de dados de toda a empresa, um Data Mart (DM) é geralmente menor e se concentra em um assunto ou departamento específico. Um DM é um subconjunto de um armazém de dados, geralmente constituídos por uma única área temática (por exemplo, marketing, operações).

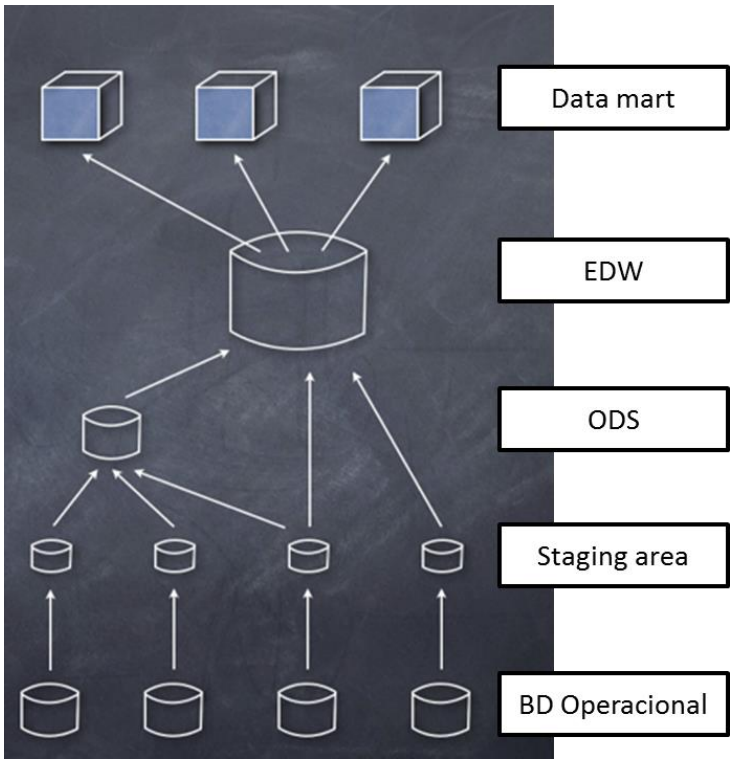
O DM pode ser **dependente** ou **independente**. Ser dependente é ser um subconjunto que é criado diretamente a partir do armazém de dados. Tem as vantagens de usar um modelo de dados consistente e o fornecimento de dados de qualidade. DM dependentes apoiam o conceito de um único modelo de dados em toda a empresa, mas o *data warehouse* deve ser construído primeiro.

O DM dependente garante que o usuário final está vendo a mesma versão dos dados que é acessada por todos os outros usuários de data warehouse. O alto custo de dados armazenados em armazéns (DW) limita seu uso para grandes empresas. Como uma alternativa, muitas empresas usam uma versão em escala reduzida de um armazém de dados referida como uma DM independente.

Um *Data Mart* **independente** é um pequeno armazém concebido para uma unidade estratégica de negócios ou um departamento, mas sua origem não é um EDW. Percebam que o DM tem como sua principal característica o escopo reduzido do projeto.

Outra definição importante para Data Mart vem do Date. Ele diz que um Data Mart deve ser especializado e volátil. Por especializado queremos dizer que ele possui uma estrutura baseada em um ambiente, tema, situação, área, setor ou aplicação específica. Enquanto o EDW se baseia em várias fontes de diversas aplicações, fontes e situações para facilitar um suporte a decisão gerencial.

Quando falamos de volátil, segundo o Date, queremos transmitir a ideia de que os dados são alterados frequentemente. Enquanto os dados do DW, por guardarem histórico, só são alterados quando uma carga foi feita de forma errada, o DM por ser baseado em aplicações é mais frequentemente modificado.



O armazenamento de dados operacionais (ODS) fornece uma forma relativamente recente dos arquivos de informações dos clientes. Este tipo de banco de dados é muitas vezes usado como uma área de estágio provisório para um *data warehouse*. Ao contrário dos conteúdos estáticos de um armazém de dados, o conteúdo de um ODS é atualizado durante todo o curso das operações de negócios.

Um ODS é usado para decisões de curto prazo envolvendo aplicações de missão crítica, para o médio e longo prazo as decisões devem estar associadas com o EDW. O ODS é semelhante à memória de curto prazo, na medida em que armazena apenas as informações mais recentes. Em comparação, um armazém de dados é como uma memória de longo prazo, porque ele armazena informações permanentes.

Neste ponto é importante abrir um parêntese para uma explicação relevante. Pense um pouco: **existe alguma diferença entre ODS e o *staging area* (SA)?** Alguns autores acham que sim. Eles argumentam que o SA é utilizado pelas ferramentas ETL para manipulação dos dados e padronização. Já o ODS é um armazenamento temporário de curto prazo para permitir o agrupamento de dados do período mais recente.

Outros autores misturam os dois conceitos. Dizem que o ODS está dentro do *staging area*. Para eles o *staging* é todo o aparato que envolve as tarefas de extração e carga e o ODS é o banco de dados que armazena as manipulações intermediárias antes da carga no DW. Em alguns momentos, os dois são usados de forma intercambiáveis, como sinônimos.

O EDW por sua vez é um DW de larga escala usado pela organização como um todo. Congrega informações de diversas fontes de dados. Vejam na figura a seguir uma hierarquia dos diferentes tipos de DW e sua relação com as demais bases existentes dentro de uma organização.

Vejamos uma questão sobre o assunto:

CEBRASPE (CESPE) - 2024 - Tecnologista Pleno 2 (CTI)/Tecnologias Habilitadoras/Inteligência Artificial e Ciência de Dados (e mais 1 concurso)

Considerando processos de análise e mineração de dados, julgue o item subsecutivo.

Data mart e data warehouse são termos sinônimos que se referem igualmente a um local onde é armazenada uma grande quantidade de dados.

Comentário: Lembre-se que Data Warehouse e Data Mart são conceitos distintos. Um **data warehouse (armazém de dados)** é um grande repositório central que coleta e armazena dados de várias fontes, como sistemas operacionais, bancos de dados transacionais e fontes externas. Esses dados são organizados de maneira estruturada para facilitar a análise, apoiar a tomada de decisões e gerar relatórios. Ele geralmente armazena dados históricos e utiliza um esquema dimensional otimizado para consultas analíticas complexas.

Um **data mart** é uma parte específica de um data warehouse, focada em um tema ou área de negócios particular. É um repositório menor e mais especializado, geralmente extraído do data warehouse principal (neste caso chamado de dependente), contendo apenas os dados necessários para uma equipe, departamento ou função específica dentro da organização. Data marts são projetados para atender às necessidades analíticas específicas de um grupo de usuários.

Gabarito: Errado

PROCESSO DE DW

Apresentamos abaixo uma figura que descreve o processo de DW/BI. Essa figura exibe os **componentes** que fazem parte do sistema.

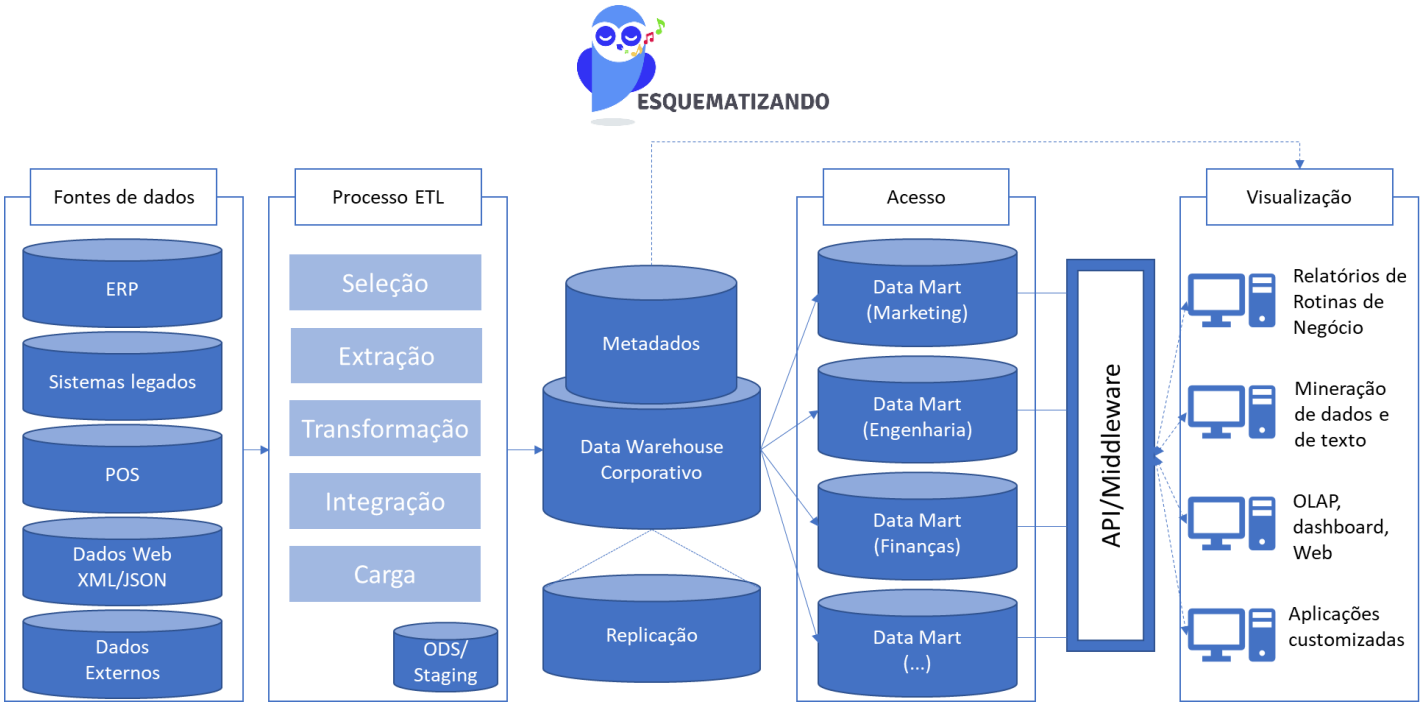


Figura 12 - Processo de Data Warehousing

Vejamos a descrição sucinta de cada um dos principais **componentes**:

TOME NOTA!

As fontes de dados (*Data source*) - Os dados são provenientes de vários sistemas legados[2] independentes e, possivelmente, a partir de provedores de dados externos. Os dados também podem vir de um processamento de transações online (OLTP) ou sistema de ERP. Dados da Web na forma de logs também podem alimentar um data warehouse.

Extração e transformação de dados (*Data extraction and transformation*) - Os dados são extraídos e devidamente transformados usando software personalizado ou comercial chamado de ferramenta de ETL.

Carregamento dos dados. (*Data loading*) – Depois que os dados são carregados para uma área de preparação (*staging area*), em que eles são transformados e limpos, eles estão, então, prontos para serem carregados no *data warehouse (EDW)* e/ou *data marts*.

[2] Sistema legado é o termo utilizado em referência aos sistemas computacionais de uma organização que, apesar de serem bastante antigos, fornecem serviços essenciais. (Wikipédia)

Banco de dados abrangente. - Essencialmente, este é o EDW para apoiar todas as análises de suporte à decisão, fornecendo uma visão resumida relevante e informações detalhadas provenientes de fontes diferentes.

Metadados - Os metadados são mantidos para que possam ser avaliados pelo pessoal de TI e pelos usuários. Metadados incluem programas sobre os dados e as regras para a organização de resumos de dados que são fáceis de indexar e pesquisar, especialmente com as ferramentas da Web.

Ferramentas de middleware - Ferramentas de middleware habilitam o acesso ao DW. Usuários avançados, como os analistas de BI, podem escrever suas próprias consultas SQL. Outros podem empregar um ambiente gerenciador de consulta, como *Business Objects*, para acessar os dados. Existem muitas aplicações *front-end* que os usuários podem usar para interagir com os dados armazenados nos repositórios de dados, incluindo a mineração de dados, OLAP, ferramentas de relatórios e ferramentas de visualização de dados.

Duas observações são importantes sobre a figura anterior. Primeiramente, você deve ter percebido a presença de um elemento denominado **ODS (*operational data storage*)/Staging**. Essa base de dados ajuda as ferramentas ETL a trabalharem com os dados no processo de transformação e integração dos dados. O outro ponto seria o **Data Mart (DM)**.

Considerando que um armazém de dados (DW) combina bases de dados de toda a empresa**, um Data Mart (DM)** é geralmente menor e se concentra em um assunto ou departamento específico. Um DM é um subconjunto de um armazém de dados, geralmente constituídos por uma única área temática (por exemplo, marketing, operações). O DM pode ser **dependente** ou **independente**.

Ser **dependente** é ser um subconjunto que é criado diretamente a partir do armazém de dados. Esse modelo tem como vantagem usar um modelo de dados consistente e o fornecimento de dados de qualidade. DM dependentes apoiam o conceito de um único modelo de dados em toda a empresa, mas o *data warehouse* deve ser construído primeiro. O DM **dependente** garante que o usuário final está vendo a mesma versão dos dados que é acessada por todos os outros usuários de data warehouse.

O alto custo de dados armazéns limita seu uso para grandes empresas. Como uma alternativa, muitas empresas usam uma versão em escala reduzida de um armazém de dados referida como uma DM **independente**. Um *Data Mart* **independente** é um pequeno armazém concebido para uma unidade estratégica de negócios ou um departamento, mas sua origem não é um EDW. Percebam que o DM tem como sua principal característica o escopo reduzido do projeto.

Qualquer organização comercial reúne rotineiramente muitos bancos de dados para várias funções de análise de marketing e negócios. A tarefa é correlacionar informações de bancos de dados diferentes, identificando indivíduos que aparecem em vários bancos de dados diferentes, normalmente de maneira inconsistente e frequentemente incorreta. O problema bastante estudado pelos analistas de dados é a tarefa de mesclar dados de fontes múltiplas da maneira mais eficiente possível, enquanto maximiza a precisão do resultado. Chamamos isso de problema de **merge/purge, que também é responsável por eliminar dados duplicados**.

Outro ponto importante é que essas ações são realizadas durante o processo de ETL, ou seja, antes dos dado serem carregados nos ambiente analítico.

ARQUITETURA DE DW

Há algumas arquiteturas básicas de DW. As arquiteturas de 2 e 3 camadas são mais comuns. Hoffer as distingue dividindo o DW em três partes:

1. O **próprio DW**, que contém os dados e o software associados,
2. **Software de aquisição de dados** (*back-end*), que extrai dados de sistemas legados e fontes externas, os consolida e resume, e depois os carrega e
3. **Software cliente** (*front-end*), que permite aos usuários acessar e analisar dados a partir do DW.

Há quatro componentes separados e distintos a serem considerados no ambiente de DW/BI: (1) Sistemas operacionais de origem, (2) Sistema de ETL, (3) Área de apresentação de dados e (4) Aplicações de *Business Intelligence* ou ferramentas de acesso a dados. Aqui vale uma ressalva importante. Quando estamos falando das aplicações de BI, estamos tratando de uma **camada de apresentação.** Na figura abaixo podemos visualizar esses 4 componentes de forma clara.

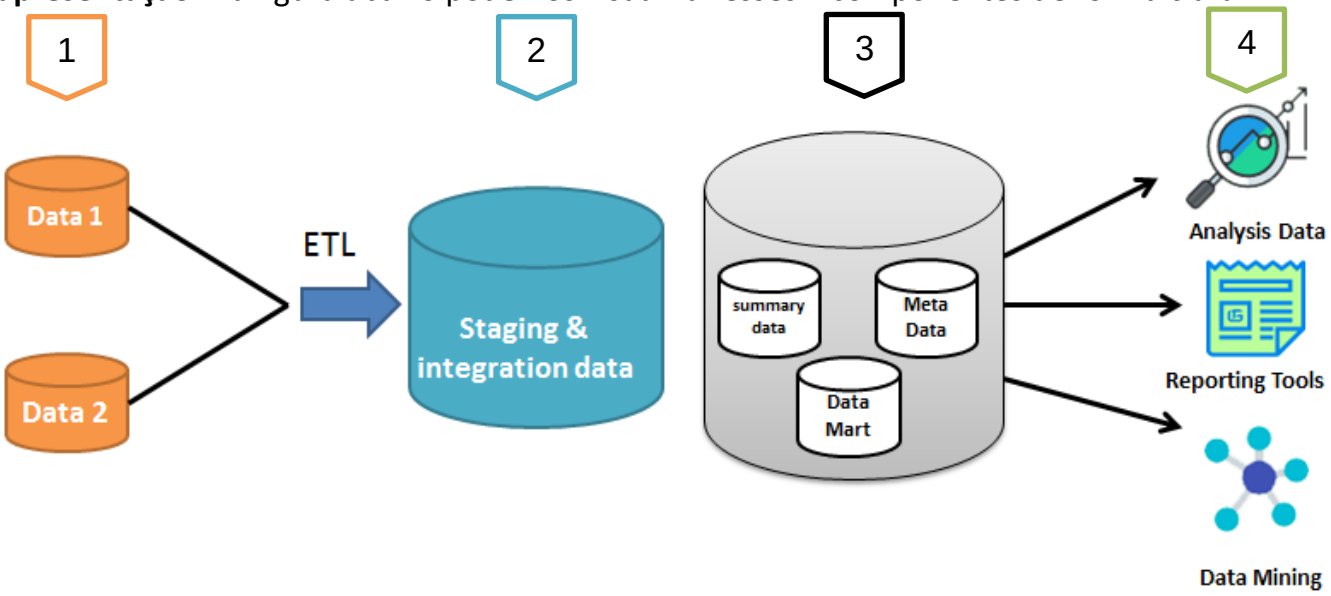


Figura 13 - Quatros componentes do ambiente de DW/BI

Segundo o Turban, quando juntamos os componentes dentro das três camadas da arquitetura temos:

1. Sistemas operacionais de origem + Sistema de ETL = Primeira camada: sistema operacional com os dados e o software para aquisição.
2. **Área de apresentação** de dados - Segunda camada: data warehouse.
3. Aplicações de BI - Terceira camada: servidor de aplicação e cliente, também chamada de **camada de apresentação**.

Vejamos uma questão sobre o assunto:

(Ano: 2019 Banca: CESPE Órgão: SEFAZ-RS Prova: Auditor Assunto: Business Intelligence) A respeito do BI (business intelligence), assinale a opção correta.

a) O BI consiste na transformação metódica e consciente das informações exclusivamente prestadas pelos tomadores de decisão em novas formas de conhecimento, para evolução dos negócios e dos resultados organizacionais.

b) ETL é o processo de análise de dados previsto pela arquitetura de BI.

c) As técnicas do BI objetivam definir regras para a formatação adequada dos dados, com vista a sua transformação em depósitos estruturados de informações, sem considerar a sua origem.

d) O repositório de dados analíticos de BI é representado pelas diversas bases de dados relacionais e por repositórios de dados que utilizem modelagens relacionais.

e) A camada de apresentação de uma arquitetura de BI é aquela em que as informações são organizadas e centralizadas.

Comentário: Vamos comentar cada uma das alternativas acima:

A) **ERRADO.** Os tomadores de decisão são, na maioria das vezes, consumidores de informações disponibilizadas pelos ambientes de business Intelligence e não fornecedores de informação.

B) **ERRADO.** ETL é o processo de extração, transformação e carga previsto na arquitetura de BI.

C) **CERTO.** A Arquitetura BI/DW realmente define regras para transformação de dados de diversas fontes que serão carregados em depósitos de dados (Data Warehouse). A questão que você deve estar se perguntando é: mas não consideram a origem dos dados? Não consideram! Veja que estamos falando da definição de regras para a formatação adequada do dados! Ou seja, eu quero definir um processo de tratamento de um determinado tipo de dado, por exemplo, um registro de cliente. Não estou muito preocupado qual sistema vai prover os dados que preciso no momento da definição das regras.

D) **ERRADO.** O repositório de dados analíticos é, geralmente, representado por uma única base de dados centralizada que utiliza uma modelagem multidimensional e não relacional.

E) **ERRADO.** A camada de apresentação de dados é onde os usuários podem interagir com os dados armazenados no data warehouse. Consultas e diversas ferramentas serão empregadas para obter diferentes tipos de informações com base nos dados. A informação pode chegar ao usuário por meio da representação gráfica dos dados, usando dashboards ou cockpits. Ferramentas de relatórios são usadas para obter dados de negócios e a lógica de negócios também é aplicada para reunir vários tipos de informações.

Gabarito: C

Data Marts (DM) independentes. Esta é sem dúvida a arquitetura mais simples e menos onerosa. Os DM são desenvolvidos para operar de forma independente visando atender às necessidades das unidades organizacionais individuais. Por causa da independência, eles podem ter definições de dados inconsistentes, diferentes dimensões e medidas, o que torna difícil analisar os dados através dos *Data Marts*.

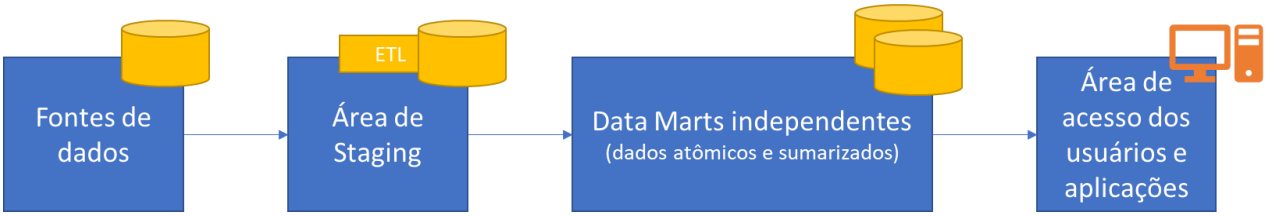


Figura 14 - Arquitetura de data marts independentes

Arquitetura de barramento de Data Mart (KIMBALL). Esta arquitetura é uma alternativa viável para os *data marts* independentes. Nela os *Data Marts* individuais são ligadas entre si através de algum tipo de middleware. Como os dados são ligados entre os DM individuais, há uma melhor chance de manter a consistência entre os dados de toda a empresa (pelo menos no nível de metadados). Mesmo que se permita que consultas complexas utilizem dados de diversos DMs, o desempenho destes tipos de análise pode não ser satisfatório.

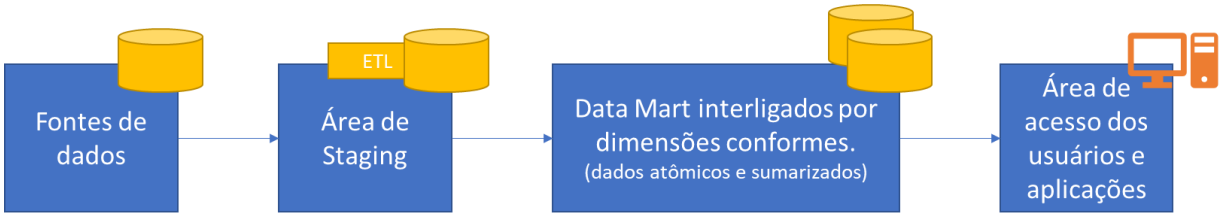
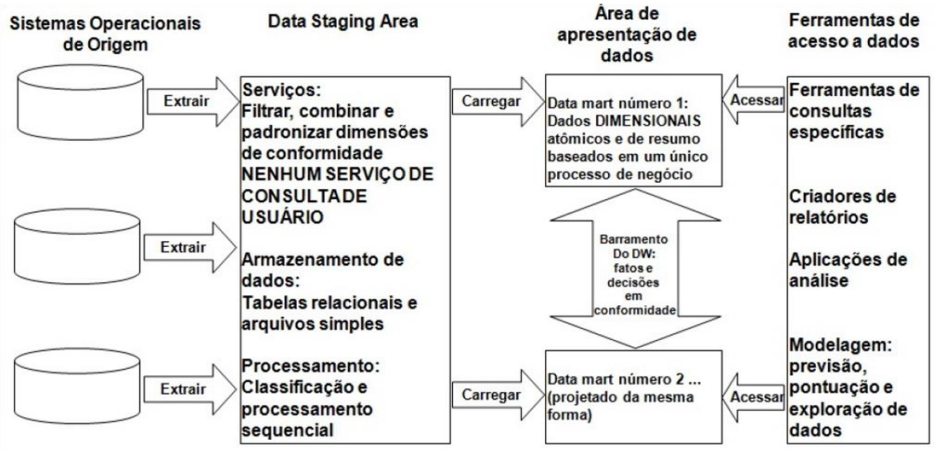


Figura 15 - Arquitetura de barramento de Data Mart (KIMBALL)

Na figura abaixo, observamos um fluxo dos dados entre os elementos de Business Intelligence. Perceba que a **apresentação dos dados** é o local onde **os dados ficam organizados, armazenados e tornam-se disponíveis para serem consultados diretamente pelos usuários**, por criadores de relatórios e por outras aplicações de análise. Kimball se refere à área de apresentação como uma série de data marts integrados. Um **Data Mart** é uma parte do todo que compõe a área de apresentação.



Hub-and-spoke (Inmon). Esta é talvez a mais famosa arquitetura de DW. Aqui, a atenção está focada na construção de uma infraestrutura escalável e sustentável (muitas vezes desenvolvidas de forma iterativa, assunto por assunto) que inclui um *data warehouse* centralizado e vários *data marts* dependentes (um para cada unidade organizacional). Esta arquitetura permite uma customização fácil de interfaces de usuário e relatórios. No lado negativo, esta arquitetura não tem a holística, e pode levar a redundância e a latência de dados. A latência dos dados refere-se ao tempo máximo permitido para disponibilização dos dados através do sistema de BI.

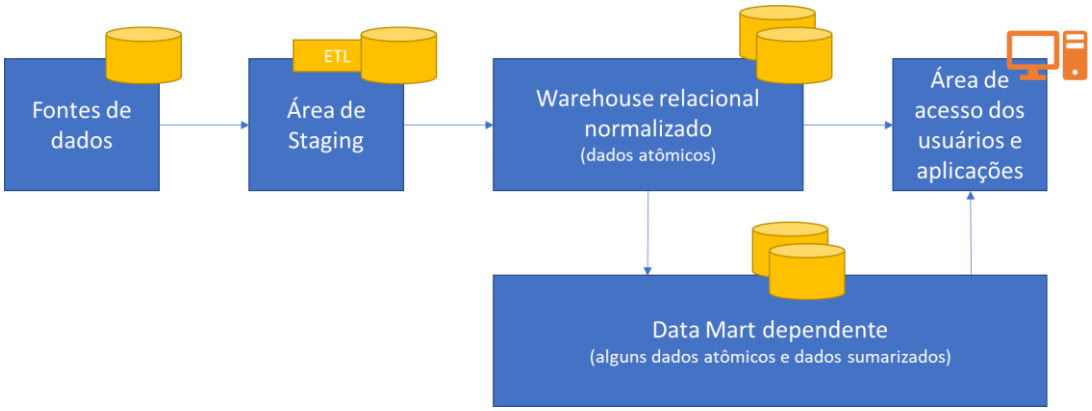


Figura 16 - Arquitetura Hub-and-spoke (INMON)

Data warehouse centralizado. A arquitetura centralizada de *data warehouse* é similar à arquitetura *hub-and-spoke*, exceto que não há *Data Marts* dependentes, em vez disso, há um *data warehouse* da empresa inteira que serve para a necessidade de todas as unidades organizacionais. Esta abordagem centralizada fornece aos usuários acesso a todos os dados no warehouse em vez de limitar-lhes a DM de dados. Adicionalmente, reduz-se a quantidade de dados que a equipa técnica tem de transferir, portanto, simplifica-se o gerenciamento e administração de dados. Se projetado e implementado corretamente, essa arquitetura oferece uma visão oportuna e holística dos dados de uma organização. Dentro de um EDW teremos informações sobre os fatos e acontecimentos relevantes, saberemos, por exemplo, quem, quando e onde os eventos aconteceram. A arquitetura de armazéns de dados central, que é defendida principalmente pela Teradata Corporation, aconselha usar armazéns de dados sem data marts.

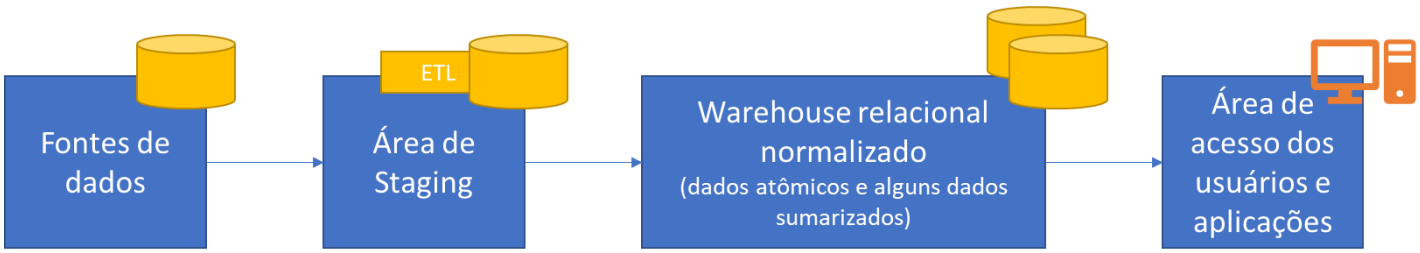


Figura 17 - Arquitetura de Data Warehouse centralizado

Armazém de dados federado. A abordagem federada é uma concessão as forças naturais que minam os melhores planos para o desenvolvimento de um sistema perfeito. Ela usa todos os meios possíveis para integrar recursos analíticos de várias fontes para atender evolução das necessidades e condições do negócio. Essencialmente, a abordagem envolve a integração de sistemas distintos. Em uma arquitetura federada de apoio à decisão, estruturas existentes são deixadas no lugar, e os dados são cedidos a partir dessas fontes, conforme necessário. A abordagem federada é suportada por fornecedores de middleware distribuídos que propõem consultar e juntar capacidades. Estas ferramentas baseadas em *eXtensible Markup Language* (XML) oferecem aos usuários uma visão global das fontes de dados distribuídas, incluindo armazéns de DM, sites, documentos e sistemas operacionais.

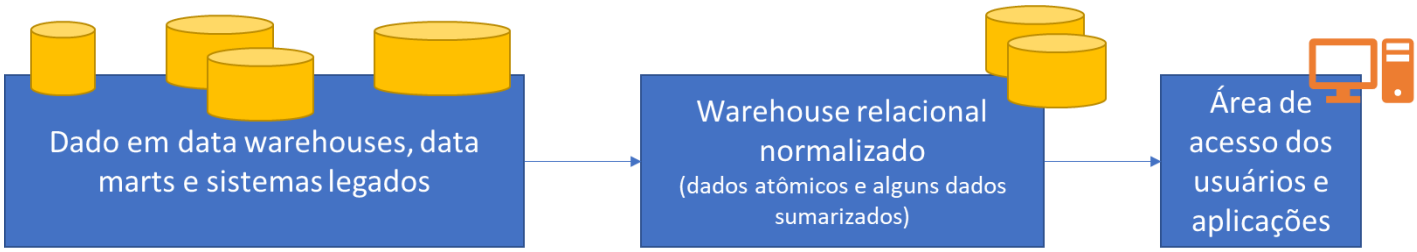


Figura 18 - Arquitetura de armazém de dados federada

Quando os usuários escolherem objetos de consulta a partir deste ponto de vista e pressionarem o botão de enviar, a ferramenta de consulta automaticamente junta os resultados das fontes distribuídas, e apresenta-os ao usuário. Devido a questões de desempenho e qualidade de dados, a maioria dos especialistas concorda que os armazéns de dados federados devem trabalhar bem de forma complementar aos *data warehouses*, e não os substituir.

KIMBALL X INMON

Considerados os dois principais autores de *Data Warehoure*, **Bill Inmon** e **Ralph Kimball** travam há décadas uma batalha teórica no campo de BI.

Ralph Kimball é um defensor da teoria de que o DW deve ser dividido para depois ser conquistado, ou seja, que o mais viável para as empresas é desenvolver vários *Data Marts* para posteriormente integrá-los e, assim, chegar-se ao EDW. Na sua avaliação, as empresas devem construir Data Marts orientados por assuntos. Ao final, teríamos uma série de pontos de conexão entre eles, que seriam as tabelas Fato e Dimensões em conformidade. Dessa forma, informações entre os diferentes *Data Marts* poderiam ser geradas de maneira íntegra e segura. Kimball batizou esse conceito de *Data Warehouse Bus Architeture*.

Bill Inmon rebate essa teoria e propõe justamente o contrário. Na sua avaliação deve-se construir primeiramente um *Data Warehouse*, modelando toda a empresa para se chegar a um único modelo corporativo, partindo posteriormente para os *Data Marts* construídos por assuntos ou departamentais. Inmon defende a ideia de que o ponto de partida seriam os CIF – *Corporate Information Factory* – uma infraestrutura ideal para ambientar os dados da empresa. O CIF seria alimentado pelos sistemas transacionais. A construção de um ODS (*Operational Data Store*) seria facultativa, mas essa iniciativa ajudaria a reduzir a complexidade da construção de um DW, uma vez que todo o esforço de integração entre os sistemas transacionais da empresa seria depositado nele.

Inmon é considerado o **pai do conceito de DW** e sustenta a tese de que a melhor estratégia seria a construção de um DW de forma TOP-DOWN. A sua ênfase sempre foi em um grande depósito central de informações. O Kimball é considerado o criador do conceito de star schema. Ele propõe uma abordagem BOTTOM-UP para construção do DW, sendo considerado um estilo mais simples com uma abordagem incremental. Vejam na figura abaixo a diferença entre as duas arquiteturas. A esquerda temos a abordagem do Kimball e a direita do Inmon. Veja que TOP-DOWN termina com N assim como Inmon. Apenas uma dica para memorização da diferença entre os modelos.

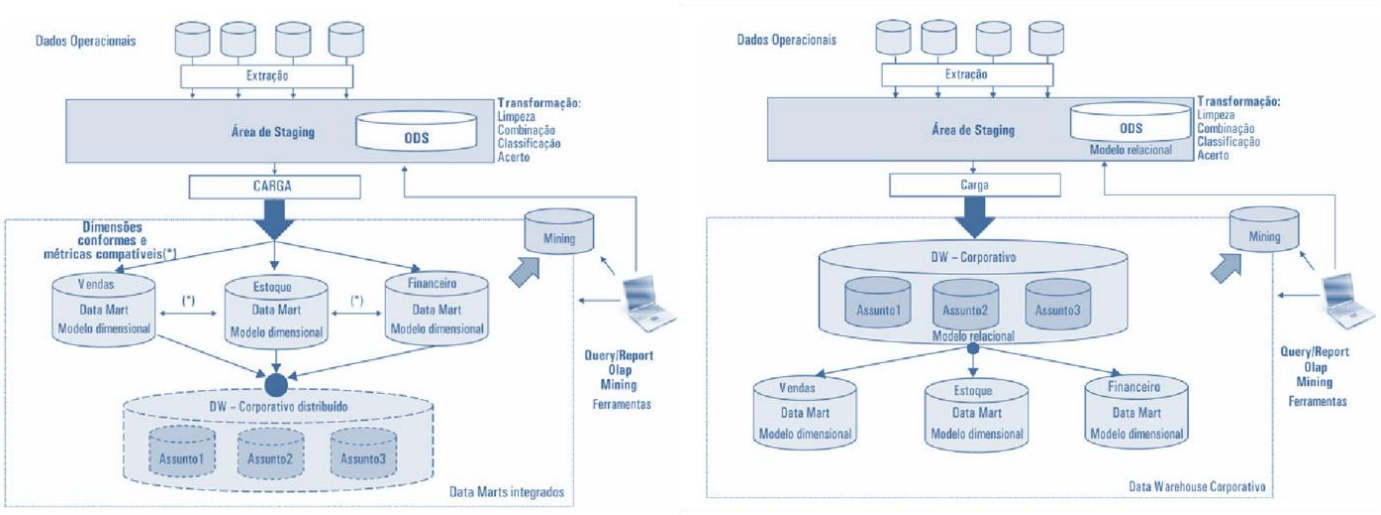


Figura 19 - Arquitetura do Kimball x Inmon

O esquema a seguir mostra um resumo dos principais pontos divergentes entres os dois autores nessa batalha de gigantes:

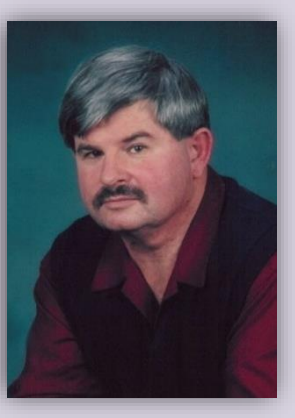
Característica	Bill **Inmon** 	Ralph **Kimball** 
Arquitetura	De cima para baixo (top-down)	De baixo para cima (bottom-up)
Estrutura arquitetônica	Data Warehouse Corporativo (EDW) alimenta os data marts departamentais (EDW -> DM)	Data Mart modelam processos de negócio e a consistência é obtida por meio de um barramento de dados. (DM -> DW)
Modelagem de dados	Normalizados até a terceira forma normal. (Snowflake Schema)	Não normalizados. (Star Schema)
Orientação dos dados	Por assunto ou por dados	Por processos
Acessibilidade aos usuários finais	Baixa, mais focada nos profissionais de TI	Alta
Metodologia	Adapta metodologias e ferramentas de banco de dados relacionais.	Usa modelagem dimensional
Consultas	São feitas diretamente nos Data Marts	São realizadas sobre o Data Warehouse
Pontos fortes	Reduz o espaço de armazenamento e fornece um conjunto de ações baseadas em técnicas já consolidadas de BD.	Simplifica o modelo para o usuário e torna mais rápido o acesso aos dados.

Figura 20 - Comparação do Inmon com o Kimball

QUESTÕES DATA WAREHOUSE COMENTADAS

Neste bloco apresentaremos questões que exploram esses conceitos associados a *data warehouse*.

1. FGV - Aud Est (CGE SC)/CGE SC/Ciências da Computação/2023 - TI - Banco de Dados - Conceitos e Fases de Projeto e Modelagem de Dados

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*.

Comentário: A técnica que torna possível a disposição do recurso de observar um mesmo item de dimensão em vários instantes de tempo (timestamps) em um Data Warehouse é:

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.

Essa abordagem permite manter os atributos de chave primária do banco de dados operacional como informações não chave no Data Warehouse, enquanto são criadas chaves primárias substitutas para fins de consulta e análise temporais.

Gabarito: E

---

2. FGV - Aud Est (CGE SC)/CGE SC/Ciências da Computação/2023 - TI - Banco de Dados - Normalização

Um Sistema de Gerenciamento de Banco de Dados Relacional (SGBDR) permite manipular bancos de dados sobre uma camada de software, dispondo os dados em formato de tabelas ao invés de arquivos em pastas. Para servir à finalidade de aplicações transacionais, as boas práticas apontam o uso do conceito de normalização.

Assinale a afirmativa ***incorreta*** em relação às vantagens da normalização.

a) Melhora a performance de consultas analíticas em um *Data Warehouse*, pois o modelo dimensional estrela depende da normalização.

b) A metodologia em etapas (1FN -> 2FN -> 3FN) facilita o processo de eliminação de dependências funcionais.

c) Diminui o esforço computacional de operações de UPDATE, pois as atualizações ocorrem apenas onde necessário.

d) Economiza espaço em disco, pois evita repetições de dados.

e) Melhora o desempenho geral sistêmico de uma aplicação, sobretudo com grandes volumes de dados, pois as transações ocorrem sob escopos específicos.

Comentário: A afirmativa incorreta em relação às vantagens da normalização é:

a) Melhora a performance de consultas analíticas em um Data Warehouse, pois o modelo dimensional estrela depende da normalização.

Na verdade, em um Data Warehouse, o modelo dimensional, como o esquema estrela ou o esquema floco de neve, é frequentemente utilizado, e esses modelos tendem a desnormalizar os dados para melhorar o desempenho das consultas analíticas. Portanto, a normalização não é uma prática comum em Data Warehouses devido à sua ênfase em otimizar as consultas de análise. A normalização é mais comumente aplicada em bancos de dados transacionais onde a ênfase está na integridade dos dados e na economia de espaço.

As demais opções estão corretas em relação às vantagens da normalização.

Gabarito: A

---

3. FCC - Auditor Fiscal (SEFAZ-BA)/Administração, Finanças e Controle Interno/2019

Nos sistemas transacionais, os dados sofrem diversas alterações como inclusão, alteração e exclusão. Antes de serem carregados no ambiente de um Data Warehouse, os dados são filtrados e limpos, de forma a gerarem informação útil. Após esta etapa, esses dados

a) ficam disponíveis para a mineração em tempo real, pois tais dados são constantemente atualizados a partir da chave de tempo que indica o dia em que foram extraídos dos sistemas transacionais.

b) podem sofrer operações de consulta, mas, devido a sua não volatilidade, não podem ser alterados, não havendo necessidade de bloqueio por concorrência de usuários ao seu acesso.

c) são reunidos a partir de diversas fontes de dados, o que facilita muito o trabalho do analista, embora este tenha que lidar com a grande redundância das informações.

d) ficam ordenados pela data da extração do sistema transacional, sendo necessárias técnicas de data mining para fazer a sua recuperação orientada por assunto.

e) são classificados somente pelo assunto principal de interesse da organização. Por exemplo, em uma organização de arrecadação de impostos, os dados são organizados pelo cadastro de contribuintes que possuem impostos a recolher.

Comentário: Vamos comentar cada uma das alternativas acima:

a) Errado. Após a carga dos dados no DW é possível extrair os dados por meio de diversas técnicas ou ferramentas como a mineração de dados. Contudo, uma vez armazenados no DW os dados se tornam não voláteis, ou seja, não sofrem modificações.

b) Certo. Um DW tem a característica de ser não volátil, o que significa que os dados não são atualizados. Assim, não existe a necessidade de bloqueio da concorrência no acesso aos dados, o que aumenta a capacidade de paralelismo do DW.

c) Errado. O analista não tem que se preocupar com a redundância da informação. Inclusive, é importante salientar que o modelo estrela possui redundância nas tabelas para facilitar o entendimento do modelo por parte do usuário/analista.

d) Errado. Um modelo dimensional sempre vai possuir uma dimensão tempo. Por meio dessa dimensão, é possível ordenar os dados de forma cronológica dos acontecimentos e não da data de extração.

e) Errado. Um DW é abrangente. Ele cobre toda a organização. Quando há um nicho especializado cria-se um Data Mart.

Gabarito: B.

HORA DE PRATICAR!

4. FCC - Auditor Fiscal (SEFAZ-BA)/Tecnologia da Informação/2019

Um Auditor da SEFAZ-BA, observando as necessidades da organização, propôs um Data Warehouse (DW) com as seguintes características:

− na camada de dados resumidos ficam os dados que fluem do armazenamento operacional, que são resumidos na forma de campos que possam ser utilizados pelos gestores de forma apropriada.

− na segunda camada, ou no nível de dados históricos, ficam todos os detalhes vindos do ambiente operacional, em que se concentram grandes volumes de dados.

Com esta organização, os tipos de consulta analítica de maior frequência acessariam os dados resumidos, mais compactos e de mais fácil acesso e, em situações em que seja necessário um maior nível de detalhe, utilizar-se-iam os dados históricos.

O Auditor propôs um DW

a) que oferece maior nível de detalhes, ou seja, alto nível de granularidade.

b) que oferece menor nível de detalhes, ou seja, baixo nível de granularidade.

c) com nível duplo de granularidade.

d) com OLAP integrado.

e) com data marts geminados.

Comentário: A questão apresenta uma situação em que você pode usar um ambiente analítico com diferentes níveis de granularidade. Essa situação é importante quando queremos ter os dados em um nível de granularidade mais baixo ou um maior nível de detalhes para responder a um conjunto maior de consultas Ad Hoc, e, ao mesmo tempo, queremos ter acesso rápido a um determinado conjunto de dados devidamente sumarizados. Assim, usamos um DW com duplo nível de granularidade.

Gabarito: C.

5. FEPESE - Analista (CELESC)/Sistemas/Desenvolvimento/2019

Assinale a alternativa que apresenta características de um Data Warehouse.

a) Orientado por assunto, integrado, volátil, variável no tempo.

b) Orientado por assunto, integrado, volátil, invariante no tempo.

c) Orientado por assunto, integrado, não volátil, variável no tempo.

d) Orientado por departamento, integrado, volátil, invariante no tempo.

e) Orientado por departamento, integrado, volátil, variável no tempo.

Comentário: A questão cobra as 4 propriedades ou características definidas pelo Inmon para um Data Warehouse, que deve ser: orientado por assunto, integrado, não volátil e variável no tempo. Desta forma, temos nossa resposta na alternativa C.

Gabarito: C.

6. CCV UFC - Técnico (UFC)/Tecnologia da Informação/"Sem Especialidade"/2019

É considerado um conjunto de informações associadas um sistema de apoio a decisão, de forma que suas operações são prioritariamente de consultas para a obtenção de dados para embasar a tomada de decisão.

Marque o item que está associado ao conceito descrito.

a) data warehousing

b) mineração de dados (data mining)

c) extração, transformação e carga (ETL)

d) processamento analítico on-line (OLAP)

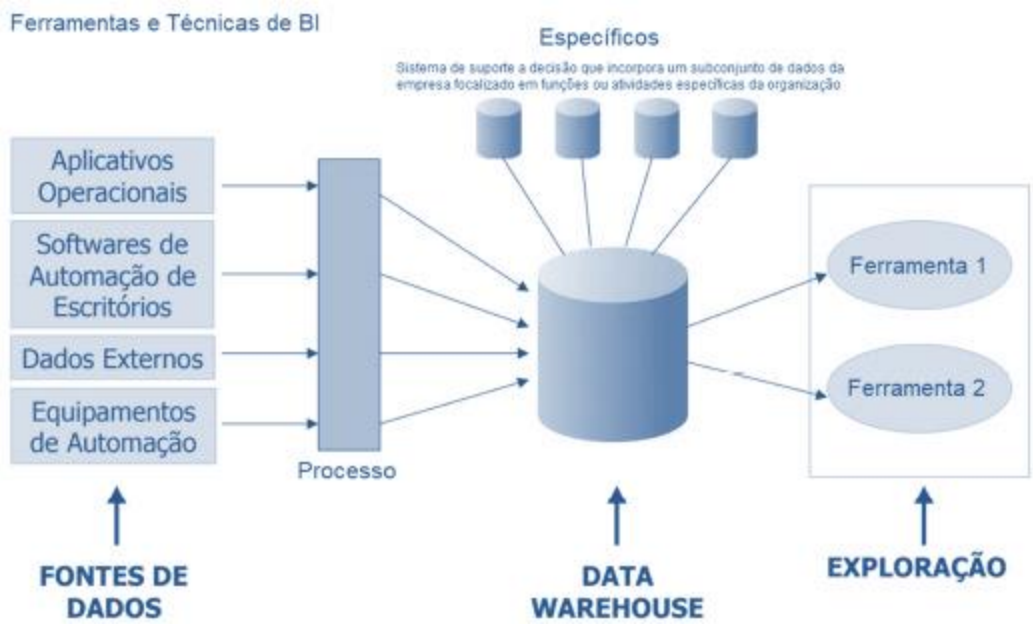
e) processamento de transações on-line (OLTP)

Comentário: O conjunto de informações armazenadas para dar suporte a decisão é encontrado no data warehouse. O processo de construção de um DW é conhecido como data warehousing. Logo, temos nossa resposta na alternativa A.

Gabarito: A.

HORA DE PRATICAR!

7. FCC - Analista de Tecnologia da Informação (SANASA)/Análise e Desenvolvimento/2019



O sistema de suporte a decisão representado em cada um dos cilindros do conjunto denominado Específicos, na imagem, é um

a) Catálogo de Metadados.

b) Schema.

c) Drill.

d) OLTP.

e) Data Mart.

Comentário: Essa questão aborda elementos de armazenamentos de dados analíticos que são considerados subconjuntos do DW. Eles geralmente são usados em um contexto departamental e são conhecidos como data marts. Desta forma, temos nossa resposta na alternativa E.

Gabarito: E.

8. VUNESP - Analista de Tecnologia da Informação (Campinas)/2019

No contexto de armazéns de dados (data warehouse), a área intermediária na qual os dados coletados pelo processo de ETL são armazenados antes de serem processados e transportados para o seu destino é chamada de

a) cubo OLAP.

b) dicionário de dados.

c) staging.

d) data vault.

e) data mart.

Comentário: A **Staging Area** é um local de armazenamento **intermediário**, situado dentro do processo de ETL (Extração, Transformação e Carga). Sua função é auxiliar a **transição dos dados das origens seu o destino no Data Warehouse**

Gabarito: C.

9. IBADE - Técnico em Informática (Pref Jaru)/2019

Todo dado é relevante. Baseado nessa premissa algumas empresas acumulam e mantém grandes quantidades de dados que, organizados e analisados, fornecem informações relevantes para os processos de decisão. A esses “depósitos” de dados chamamos Data:

a) Check.

b) Mart.

c) Pool.

d) Mining.

e) Warehouse.

Comentário: O Data Warehouse (DW) é um **grande repositório** de dados que tem como objetivo permitir consultas rápidas e complexas, principalmente utilizando ferramentas OLAP, a fim de auxiliar **na manutenção do negócio e na tomada de decisão estratégica**. Já um data mart é considerado um subconjunto das informações analíticas. Como a questão afirma que todo o dado é importante, não devemos restringir o conteúdo do ambiente analítico. Desta forma, podemos marcar a nossa resposta na alternativa E.

Gabarito: E.

HORA DE PRATICAR!

10. CEBRASPE (CESPE) - Assistente Judiciário (TJ AM)/Programador/2019

Com relação a arquitetura e tecnologias de sistemas de informação, julgue o próximo item.

Data warehouse, o principal dispositivo de armazenamento de um computador, é formado pelo processador, pela entrada e pela saída de dados.

Comentário: Data warehouse não é uma dispositivo de armazenamento (memória), mas sim uma base de dados analítica.

Gabarito: Errado.

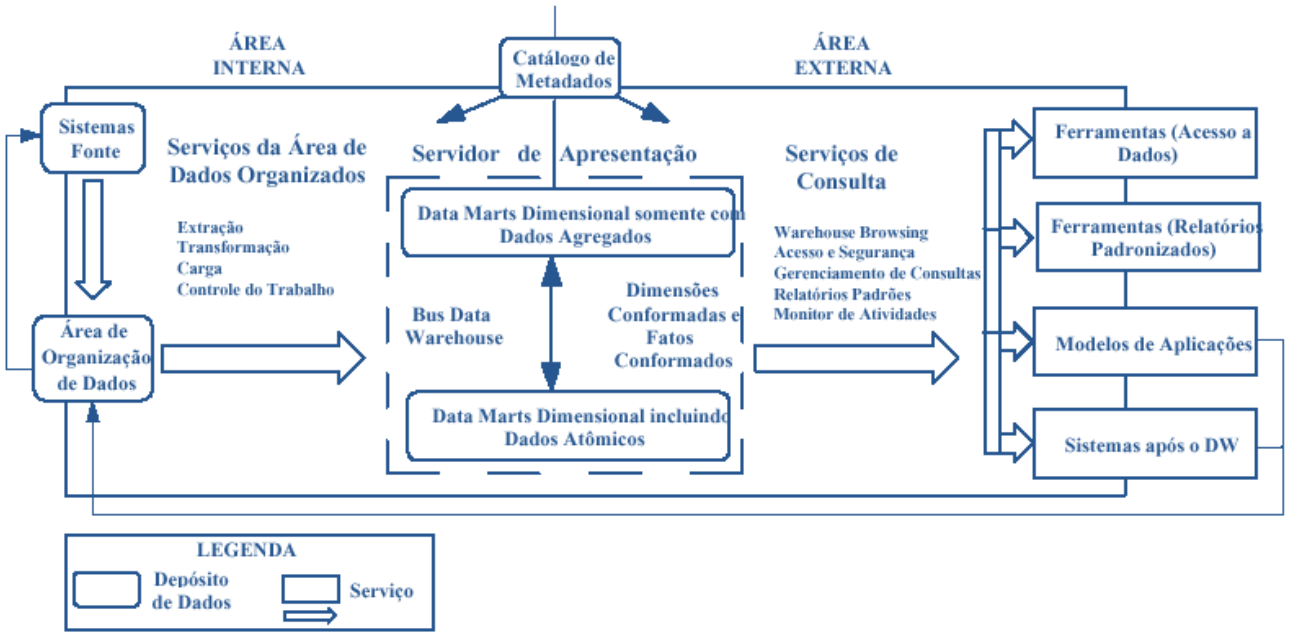
HORA DE PRATICAR!

11. CEBRASPE (CESPE) - Assistente Judiciário (TJ AM)/Suporte ao Usuário de Informática/2019

A respeito de data warehouse e data mining, julgue o item que se segue.

Chamados de data mart, os servidores de apresentação de data warehouse permitem consultas.

Comentário: Essa questão foi um pouco mais profunda do que estamos acostumados ... e para respondê-la, vamos recapitular alguns conceitos. A figura abaixo apresenta uma arquitetura funcional para um data warehouse:



Perceba a presença de um **servidor de apresentação** no meio da figura. Ele é formado por um conjunto de Data Marts interligados por dimensões conformes. Esse aspecto está relacionado a arquitetura do Kimball para Data Warehouses. Veja que o servidor de apresentação está provendo serviços para a área externa, que são exatamente os serviços de consulta. Logo, temos uma alternativa correta.

Gabarito: C.

HORA DE PRATICAR!

12. DECEx - Curso de Formação de Oficiais do Quadro Complementar (EsFCEx)/Informática/2019/CA CFO-QC 2020

Nas palavras de LAUDON e LAUDON, em sistemas de informação, existem quatro tipos de sistemas que apoiam os diferentes níveis e tipos de decisão. Os Sistemas de Informações Gerenciais (SIG) fornecem resumos e relatórios de rotina com dados no nível de transação para a gerência de nível operacional e médio. Sistemas de Apoio à Decisão (SAD) fornecem ferramentas ou modelos analíticos para analisar grandes quantidades de dados, além de consultas interativas de apoio para gerentes de nível médio que enfrentam situações de decisões semiestruturadas. Sistemas de Apoio ao Executivo (SAE) são sistemas que fornecem à gerência sênior, envolvida em decisões não estruturadas, informações externas e resumos de alto nível. Sistemas de Apoio à Decisão em Grupo (SADG) são sistemas especializados que oferecem um ambiente eletrônico no qual gerentes e equipes podem coletivamente tomar decisões e formular soluções.

Considerando os conceitos do universo dos Sistemas de Apoio à Decisão (SAD), avalie as seguintes asserções e a relação proposta entre elas.

I. DATA WAREHOUSE é uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo, voláteis, para dar suporte ao processo de tomada de decisão. É um repositório de grande volume de dados tratados, objetivando levar informação a partir dos dados.

OBTIDOS POR MEIO DE

II. Ambientes heterogêneos, geralmente de bancos transacionais, utilizando técnica de ETL para extrair, transformar e carregar, com objetivo de processamento analítico, de modo a permitir a criatividade das pessoas envolvidas, também denominado de OLAP.

A respeito dessas asserções, assinale a alternativa correta.

a) As asserções I e II são proposições verdadeiras, e a II complementa a I.

b) As asserções I e II são proposições verdadeiras, mas a II não complementa a I.

c) A asserção I é uma proposição verdadeira, e a II é uma proposição falsa.

d) A asserção I é uma proposição falsa, e a II é uma proposição verdadeira.

e) As asserções I e II são proposições falsas.

Comentário: Perceba que existe um erro na primeira afirmação, DW é não volátil! A afirmação I fala o contrário. Já a afirmação II está correta. Desta forma, temos o gabarito na alternativa D.

Gabarito: D.

HORA DE PRATICAR!

13. Ano: 2018 Banca: FCC Órgão: DPE-AM Prova: Analista em Gestão Especializado de Defensoria - Analista de Banco de Dados

Uma das características fundamentais de um ambiente de data warehouse está em

a) servir como substituto aos bancos de dados operacionais de uma empresa, na eventualidade da ocorrência de problemas com tais bancos de dados.

b) ser de utilização exclusiva da área de aplicações financeiras das empresas.

c) proporcionar um ambiente que permita realizar análise dos negócios de uma empresa com base nos dados por ela armazenados.

d) ser de uso prioritário de funcionários responsáveis pela área de telemarketing das empresas.

e) armazenar apenas os dados mais atuais (máximo de 3 meses de criação), independentemente da área de atuação de cada empresa.

Comentário: Essa questão é interessante pois aborda pontos importantes do assunto. O primeiro deles aparece logo na alternativa "a" quando o examinador fala em **bancos de dados operacionais**. Esses bancos de dados são, geralmente, relacionais e são usados no dia a dia das operações de uma organização. Nele são registradas as vendas, contratações e ações de marketing ou produção. Esse tipo de banco de dados contrasta com os bancos de dados analíticos que permitem analisar contextos mais amplos e os dados de forma mais agregada. O data warehouse é o principal representante dessas bases analíticas.

Os data warehouse são utilizados por toda as áreas da empresa. Não existe uma restrição ao departamento financeiro. Logo, a alternativa B também está incorreta.

A letra "c" é a nossa resposta. O armazém de dados vai prover uma estrutura de armazenamento que possa ser usada pelos gestores de uma empresa para extrair informações que subsidiem a tomada de decisão.

A alternativa "d" fala um absurdo. Dizer que o data warehouse é de uso exclusivo do telemarketing é uma afirmação, no mínimo absurda.

Por fim, a alternativa "e" vai de encontro a uma das características dos DW, que é o armazenamento de dados históricos.

Gabarito: **C.**

HORA DE PRATICAR!

14. Ano: 2017 Banca: CESPE Órgão: SEDF Cargo: Analista de gestão educacional – Especialidade: tecnologia da informação Questão: 118

Com relação a data mining e data warehouse, julgue os itens que se seguem.

[118] Comparados aos bancos de dados transacionais, os data warehouses são mais voláteis porque, para que se mantenham consistentes, são atualizados em tempo real a cada atualização que ocorrer em qualquer uma das bases originais de dados que o componham.

Comentário: A característica de **não volatilidade** está relacionada ao fato de que o conteúdo do **Data Warehouse** permanece estável por **longos períodos** de tempo. Basicamente duas operações são efetuadas no Data Warehouse. A primeira é a **transação de manutenção**, onde o objetivo é a **carga** dos dados provenientes dos provedores de informação. A segunda é relacionada à **leitura** dos dados para **geração de relatórios** de tomadas de decisão.

Analisando as informações podemos concluir que alternativa está **incorreta**.

Gabarito: E.

HORA DE PRATICAR!

15. Ano: 2018 Banca: CESGRANRIO Órgão: TRANSPETRO Cargo: ANALISTA DE PROCESSO DE NEGÓCIO Questão: 22

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

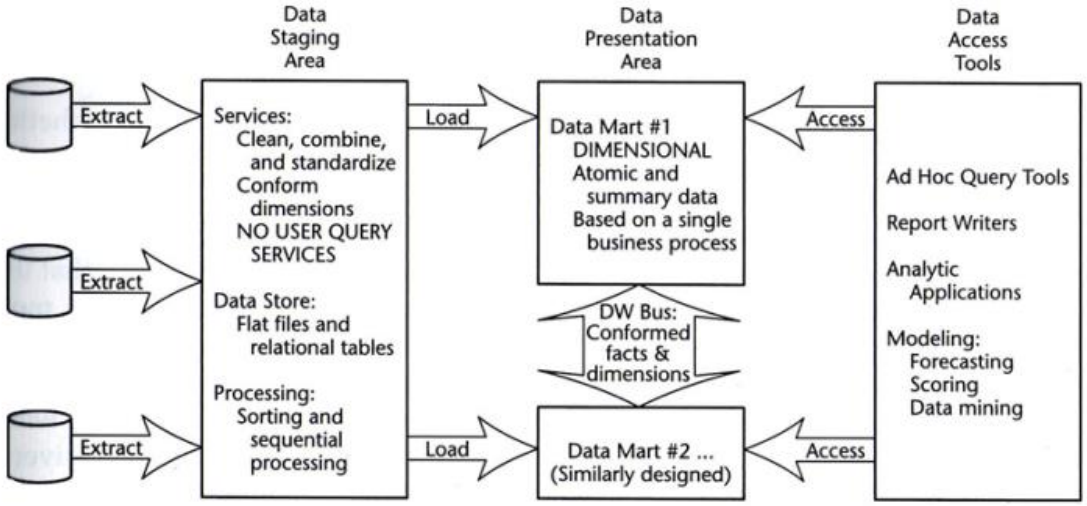
(B) Data Staging Area

(C) Dimensional Model Area

(D) Presentation Area

(E) Living Sample Area

Comentário: A data staging área pode ser considerada uma área de trabalho na qual as ferramentas ETL fazem o trabalho de limpeza e transformação dos dados. Pode ser vista como a parte do Data Warehouse responsável pelo armazenamento e a execução de um conjunto de processos normalmente denominados como extração, transformação e carga (ETL – extract, transformation, load) dos dados. A área de Staging encontra-se entre as bases operacionais e a camada de apresentação. É considerada a "cozinha do restaurante" que está fora do acesso dos usuários. Veja esse contexto na figura abaixo.



Sendo assim, podemos marcar a resposta na alternativa B.

Gabarito: B

HORA DE PRATICAR!

16. Ano: 2016 Banca: FCC Órgão: TRT-20 Cargo: Técnico de TI – Q. 38

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) tem como vantagem a criação de legamarts ou DMs legados que facilitam e agilizam futuras integrações.

(B) é conhecida como top down.

(C) permite um retorno de investimento apenas em longo prazo, ou seja, um slower pay back.

(D) tem como objetivo a construção de um sistema OLAP incremental a partir de DMs independentes.

(E) não garante padronização dos metadados, podendo criar inconsistências de dados entre os DMs.

Comentário: Vejam que ele descreveu no enunciado o modelo de desenvolvimento ***bottom-up*** descrito por Kimball. Nesta abordagem, temos como vantagem a percepção mais rápida do retorno sobre o investimento. Outro ponto é a estruturação do DW a partir dos diversos DM já existentes. Essa carga de informações pode, por trazer dados de diferentes fontes, apresentar incompatibilidade de informações ou inconsistências. Sendo assim, a nossa resposta encontra-se na alternativa E.

Gabarito: E.

17. ANO: 2015 BANCA: CESPE ÓRGÃO: MEC PROVA: TÉCNICO DE NÍVEL SUPERIOR - ADMINISTRADOR DE DADOS

No que se refere a bancos de dados transacionais (OLTP) e a banco de dados analíticos (OLAP), julgue os itens que se seguem.

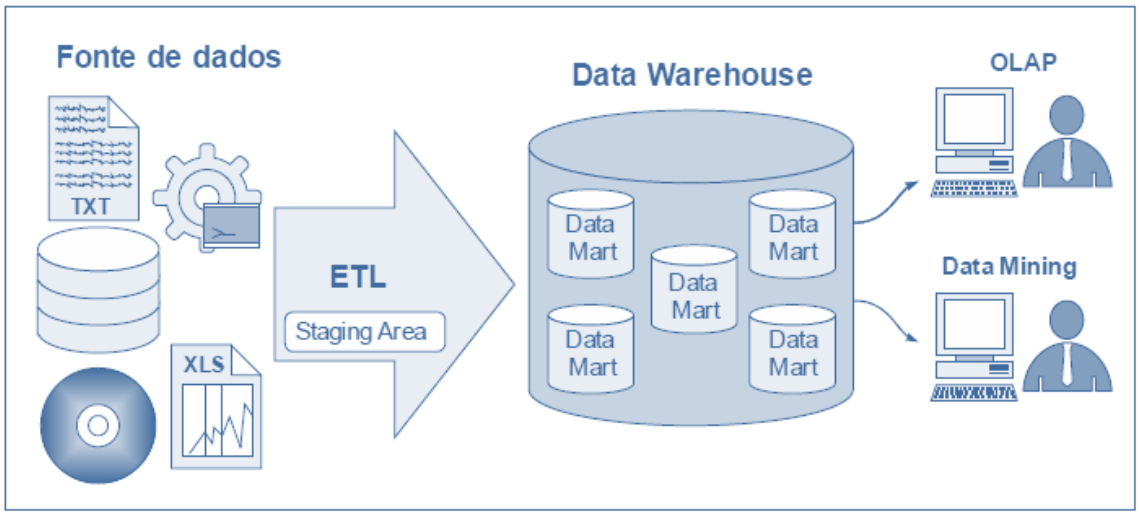
[1] Para melhor manter o controle sobre identificadores de registro de ambientes de data warehouse (armazém de dados), em geral recomenda-se a geração de chaves substitutas (surrogate keys). Assim, cada junção entre as tabelas de dimensão e tabelas fato em um ambiente de data warehouse deve se basear nessas chaves substitutas, e não nas chaves naturais existentes.

Comentários: Existem vários motivos para a utilização de chaves substitutas ou artificiais dentro dos nossos modelos de DW. A primeira seria a que a mudança de estado de um tupla nos modelos operacionais pode gerar uma nova entrada na tabela de DW. Se estivermos usando a mesma chave primária do objeto no modelo operacional, nós não conseguiríamos incluir essa nova linha. É justamente sobre isso que a questão fala. O que torna a alternativa correta.

Gabarito: C.

18. ANO: 2015 BANCA: FCC ÓRGÃO: TRT - 4ª REGIÃO (RS) PROVA: ANALISTA JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

Considere a arquitetura geral de um sistema de BI - Business Intelligence mostrada na figura abaixo.



Nesta arquitetura

A Data Mining se refere ao processo que, na construção do Data Warehouse, é utilizado para composição de análises e relatórios, armazenando dados descritivos e qualificando a respectiva métrica associada.

B Data Marts representam áreas de armazenamento intermediário criadas a partir do processo de ETL. Auxiliam na transição dos dados das fontes OLTP para o destino final no Data Warehouse.

C OLAP é um subconjunto de informações extraído do Data Warehouse que pode ser identificado por assuntos ou departamentos específicos. Utiliza uma modelagem multidimensional conhecida como modelo estrela.

D os dados armazenados no Data Warehouse são integrados na base única mantendo as convenções de nomes, valores de variáveis e outros atributos físicos de dados como foram obtidos das bases de dados originais.

E o Data Warehouse não é volátil, permite apenas a carga inicial dos dados e consultas a estes dados. Além disso, os dados nele armazenados são precisos em relação ao tempo, não podendo ser atualizados.

Comentários: Vamos analisar cada uma das alternativas. Na letra A temos uma definição equivocada de Data Mining. Apenas para apresentar um conceito correto, podemos definir "Data mining (mineração de dados) é o processo de extração de conhecimento de grandes bases de dados, convencionais ou não." Utiliza técnicas de inteligência artificial que procuram relações de similaridade ou discordância entre dados. Seu objetivo é encontrar padrões, anomalias e regras com o propósito de transformar dados, aparentemente ocultos, em informações úteis para a tomada de decisão e/ou avaliação de resultados.

A alternativa B trata das áreas de armazenamento temporárias, geralmente conhecidas como Staging Area. Elas são utilizadas para manipulação dos dados durante a transição do ambiente operacional para os DW. Vejam que Staging Area não se assemelha de forma alguma com um Data Mart.

A alternativa C define Data Mart, mas associa a definição a OLAP.

A alternativa D trata da manutenção dos atributos como foram obtidos nas bases originais. Se você se lembrar da transformação de chaves naturais em chaves artificiais tratada durante a aula vai visualizar o erro da alternativa.

Por fim, temos a letra E que se trata da nossa resposta.

Gabarito: E.

19. ANO: 2015 BANCA: FCC ÓRGÃO: TRT - 3ª REGIÃO (MG) PROVA: TÉCNICO JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

Um técnico de TI precisa utilizar um subconjunto de dados de um Data Warehouse direcionado à área administrativa de um Tribunal. Esses dados serão armazenados em um banco de dado modelado multidimensionalmente, que será criado capturando-se dados diretamente de sistemas transacionais, buscando as informações relevantes para os processos de negócio da área administrativa. Esse banco de dados será um

A Big Data.

B Data Mart.

C OLAP.

D MOLAP.

E Data Mining.

Comentário: Durante nossa explicação teórica sobre o assunto, analisamos a diferença em Data Warehouse e Data Mart. O segundo está geralmente associado a uma parte de organização, por exemplo, uma área ou divisão de negócio. Este tipo de banco de dados facilita a implementação de projetos de BI em uma organização usando a técnica de dividir para conquistar.

Uma outra definição possível para Data Mart é um repositório de dados, devidamente agregados e sumarizados, com o objetivo de atender a interesses de uma área específica de negócios de uma organização. Há quem o defina como um subconjunto lógico de um Data Warehouse, ou seja, um DW setorial.

Sendo assim, podemos perceber que nossa resposta se encontra na alternativa B.

Gabarito: B

20. ANO: 2014 BANCA: FCC ÓRGÃO: TCE-RS PROVA: AUDITOR PÚBLICO EXTERNO - TÉCNICO EM PROCESSAMENTO DE DADOS

A granularidade de dados é uma questão crítica no projeto de um Data Warehouse (DW), pois afeta o volume de dados que reside no DW e, ao mesmo tempo, afeta o tipo de consulta que pode ser atendida. Considere:

I. Quanto mais detalhe existir, mais baixo será o nível de granularidade. Quanto menos detalhe existir, mais alto será o nível de granularidade.

II. Quando há um nível de granularidade muito alto, o espaço em disco e o número de índices necessários se tornam bem menores, mas há uma correspondente diminuição da possibilidade de utilização dos dados para atender a consultas detalhadas.

É correto afirmar que a afirmativa I

A é equivalente a: quanto menos detalhes há nos dados, menor é a granularidade, consequentemente, quanto mais detalhes existem, maior é a granularidade.

B e a afirmativa II estão corretas e coerentes em relação ao nível de granularidade, espaço em disco e tipos de consultas em um DW.

C está correta. A afirmativa II está incorreta, pois apresenta incoerência em relação ao nível de granularidade, espaço em disco e tipos de consultas em um DW.

D e a afirmativa II estão incorretas. Ambas apresentam incoerência em relação ao nível de granularidade, espaço em disco e tipos de consultas em um DW.

E está incorreta. A afirmativa II está correta, pois é coerente em relação ao nível de granularidade, espaço em disco e tipos de consultas em um DW.

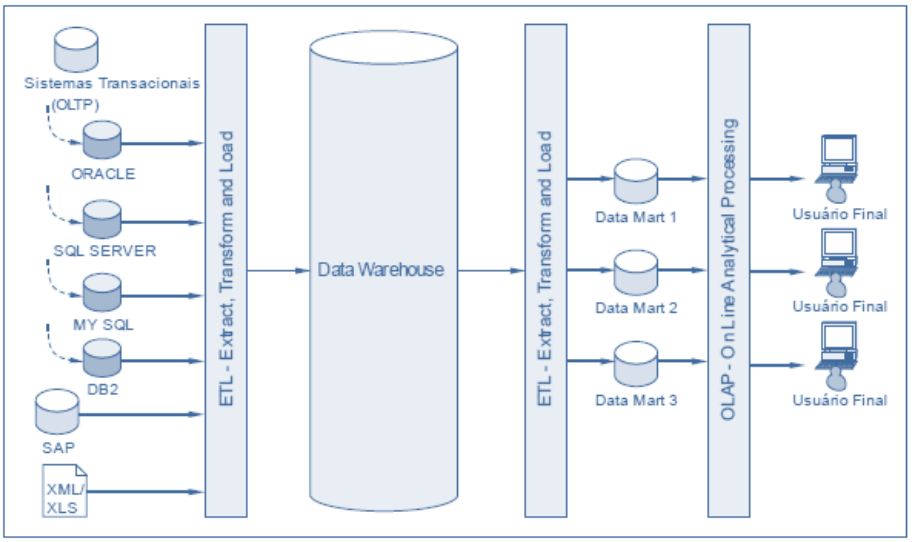
Comentário: Um importante aspecto no projeto de um Data Warehouse é a questão da granularidade: quanto **menos detalhe** uma unidade de dados apresentar, **mais alto será o nível de granularidade**. Esse conceito tem a ver com o volume de dados possível de armazenar e afeta o tipo de consulta que pode ser feita sobre os dados.

Vejam que os textos das afirmativas I e II estão corretos e são coerentes com a afirmação teórica que fizemos no parágrafo anterior. Desta forma, podemos confirmar que a alternativa B é a nossa resposta.

Gabarito: B.

21. ANO: 2015 BANCA: FCC ÓRGÃO: CNMP PROVA: ANALISTA DO CNMP - DESENVOLVIMENTO DE SISTEMAS

Considere que a equipe de Analistas de Desenvolvimento de Sistemas do CNMP está projetando a arquitetura para o Data Warehouse (DW) da instituição, conforme mostra a figura abaixo:



correto afirmar que esta arquitetura

A é bottom-up, pois primeiro a equipe cria um DW e depois parte para a segmentação, ou seja, divide o DW em áreas menores gerando pequenos bancos orientados por assuntos aos departamentos.

B é bottom-up. Permite um rápido desenvolvimento, pois a construção dos Data Marts é altamente direcionada. Normalmente um Data Mart pode ser colocado em produção em um período de 2 a 3 meses.

C é top-down. A partir do DW são extraídos os dados e metadados para os Data Marts. Nos Data Marts as informações estão em maior nível de sumarização e, normalmente, não apresentam o nível histórico encontrado no DW.

D é top-down, pois possui um retorno de investimento muito rápido ou um faster pay back. O propósito desta arquitetura é a construção de um DW incremental a partir de Data Marts independentes.

E é bottom-up. Garante a existência de um único conjunto de aplicações para ETL, ou seja, extração, limpeza e integração dos dados, embora os processos de manutenção e monitoração fiquem descentralizados.

Comentário: Uma perspectiva ***top-down*** considera que um Data Warehouse completo, centralizado, deve ser desenvolvido antes que partes dele venham a ser derivadas sob a forma de Data Mart. Uma perspectiva ***bottom-up*** considera que um Data Warehouse possa ser composto a partir da agregação de Data Mart previamente desenvolvidos.

Vejamos as vantagens de cada perspectiva:

Top-Dow (DW → Data Mart)

Evita duplicação de dados.

Controla tendência à dispersão de dados.

Acarreta menor sobrecarga administrativa.

Acarreta menor custo para aplicações múltiplas.

Bottom-Up (Data Mart → DW)

Apresenta menor custo inicial.

Oferece resultados mais imediatos.

Tem justificativa mais fácil.

Possibilita uma análise de resultado prematura.

Analisando o texto apresentado podemos marcar nossa resposta na alternativa C.

Gabarito: C.

22. ANO: 2010 BANCA: FCC ÓRGÃO: TRT - 22ª REGIÃO (PI) PROVA: ANALISTA JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

No âmbito dos DWs, uma outra concepção do ODS (Staging Area) está sendo estabelecida por alguns autores. Trata-se de

A OLAP.

B Drill throught.

C ETL.

D Data Mining.

E Dynamic Data Storage.

Comentário: Um ***Operational Data Storage*** (**ODS**) ou ***Staging Area*** (SA) representa um armazenamento intermediário dos dados, promovendo a integração dos dados do ambiente operativo antes de sua atualização no DW. Inicialmente, um ODS era considerado um repositório temporário que armazenava apenas informações correntes antes de serem carregadas para o DW, similar a uma cópia dos ambientes de sistemas transacionais em uma empresa. Atualmente, alguns autores passaram a denominá-lo ***Dynamic Data Storage*** **(DDS)**.

Essa concepção se diferencia do conceito original pela sua periodicidade de armazenamento e pelo fato de não somente armazenar dados temporários para a carga do DW. Por não ser volátil, seus dados são armazenados ao longo do tempo e passam por alterações incrementais podendo se tornar um DW. Sendo assim, podemos marcar a resposta na alternativa E.

Gabarito: E.

DATA LAKE

A tomada de decisão orientada por dados está mudando a forma como trabalhamos e vivemos. Desde a ciência de dados, aprendizado de máquina e análises avançadas até painéis de controle em tempo real, os tomadores de decisão estão exigindo dados para ajudar nas suas escolhas. Empresas como Google, Amazon e Facebook são gigantes orientados por dados que estão dominando os negócios tradicionais ao alavancar informações.

Organizações de serviços financeiros e seguradoras sempre foram orientadas por dados, com negociações automatizadas liderando o caminho. A Internet das Coisas (IoT) está mudando a manufatura, transporte, agricultura e saúde. De governos a corporações em todos os setores até organizações sem fins lucrativos e instituições de ensino, **os dados estão sendo vistos como um fator transformador**. Inteligência artificial e aprendizado de máquina estão permeando todos os aspectos de nossas vidas. O mundo está "devorando" dados devido ao seu potencial. Até temos um termo para esse "banquete": **big data**, definido por Doug Laney da Gartner em termos dos três Vs (volume, variedade e velocidade), aos quais ele posteriormente adicionou um quarto V — veracidade.

Com tanta variedade, volume e velocidade, os sistemas e processos antigos não são mais capazes de suportar as necessidades de dados da empresa. A veracidade é um problema ainda maior para análises avançadas e inteligência artificial, onde o princípio do "GIGO" (*garbage in = garbage out*) é ainda mais crítico, pois é virtualmente impossível dizer se os dados estavam ruins e causaram decisões ruins em modelos estatísticos e de aprendizado de máquina ou se o modelo estava ruim.

Para apoiar esses empreendimentos e enfrentar esses desafios, uma revolução está ocorrendo na gestão de dados em torno de como os dados são armazenados, processados, gerenciados e fornecidos aos tomadores de decisão. A tecnologia de big data está permitindo **escalabilidade e eficiência de custo** em ordens de magnitude maiores do que o possível com a infraestrutura tradicional de gerenciamento de dados. O **autoatendimento** está tomando o lugar das abordagens cuidadosamente elaboradas e intensivas em mão de obra do passado, onde exércitos de profissionais de TI criavam data warehouses e data marts bem governados, mas levavam meses para fazer qualquer alteração.

O data lake é uma abordagem inovadora que aproveita o poder da tecnologia de big data e une-o à agilidade do autoatendimento. A maioria das grandes empresas hoje já implantou ou está em processo de implantação de data lakes.

Vejamos uma questão sobre o assunto antes de continuar:

CEBRASPE (CESPE) - 2023 - Analista de Processamento (DATAPREV)

No que se refere a conceitos de business intelligence, data lake, inteligência artificial e machine learning, julgue o item a seguir.

Um data lake é um excelente recurso para solucionar problemas e encontrar alternativas no momento da tomada de decisão.

Comentário: Um data lake pode ser uma ferramenta poderosa para fornecer insights valiosos e suportar a tomada de decisões, mas seu verdadeiro valor está diretamente

relacionado à capacidade de explorar e analisar eficientemente os dados armazenados. Dados brutos e não estruturados em um data lake, se não forem adequadamente organizados, catalogados e analisados, podem dificultar a extração de informações significativas.

A eficácia de um data lake na tomada de decisões depende da implementação de processos adequados de governança de dados, qualidade dos dados e segurança. Além disso, a utilização de ferramentas analíticas avançadas, como business intelligence (BI), inteligência artificial (IA) e machine learning (ML), é crucial para explorar e extrair insights dos dados armazenados.

Portanto, embora um data lake ofereça um grande potencial para apoiar a tomada de decisões, é fundamental uma gestão efetiva dos dados e a aplicação de técnicas analíticas avançadas para obter benefícios significativos. Vale ressaltar que um data lake, por si só, não resolve problemas nem encontra alternativas; ele é apenas um repositório de dados. A solução de problemas e a identificação de alternativas dependem do uso de ferramentas de análise de dados, como BI, IA e ML, que exploram os dados do data lake para identificar padrões e tendências úteis na tomada de decisões.

Gabarito: Errado

Esta parte da nossa aula tem por objetivo explicar a importância dos data lakes e seus conceitos básicos. Nas próximas páginas, procuraremos responder algumas perguntas como: o que é um data lake? Por que precisamos disso? Como é diferente do que já temos? Tentaremos também apresentar os tópicos que sejam suficientes para que você responda qualquer questão que venha a aparecer no seu concurso sobre esse assunto. Assim, vamos oferecer uma visão geral.

A tomada de decisão orientada por dados está mudando a forma como trabalhamos e vivemos. Esta demanda por dados precisa de um lugar para residir, e o data lake é a solução preferida para criar essa moradia. O termo foi inventado e primeiro descrito por **James Dixon, CTO da Pentaho**, que escreveu em seu blog: “Se você pensar em um Data Mart como uma loja de água engarrafada — limpa, embalada e estruturada para fácil consumo — o data lake é um grande corpo de água em um estado mais natural. O conteúdo do data lake flue de uma fonte para encher o lago, e vários usuários do lago podem vir para examinar, mergulhar ou tirar amostras.” Os pontos críticos, são:

- Os dados estão em sua forma e formato originais (dados naturais ou brutos).
- Os dados são usados por vários usuários (ou seja, acessados e acessíveis por uma grande comunidade de usuários).

Mas, como construir um Data Lake que traz dados brutos (bem como dados processados) para uma grande comunidade de usuários de análise de negócios, em vez de apenas usá-lo para projetos orientados por TI? A razão para disponibilizar dados brutos aos analistas é para que eles possam realizar análises de autoatendimento. O autoatendimento tem sido uma importante tendência em direção à democratização de dados. Começou com ferramentas de visualização de autoatendimento como Tableau e Qlik (às vezes chamadas de ferramentas de descoberta de dados) que permitem aos usuários de dados analisarem os mesmos sem precisar de ajuda especializada da equipe de TI.

A tendência de autoatendimento continua com ferramentas de preparação de dados que ajudam a moldar os dados para análises e ferramentas de catálogo que ajudam a encontrar os dados de que precisam, além de ferramentas de ciência de dados que ajudam a realizar análises avançadas. Para análises ainda mais avançadas, geralmente referidas como ciência de dados, uma nova classe de usuários chamada cientistas de dados também costuma fazer do data lake sua principal fonte de dados.

Claro, um grande desafio com o autoatendimento é **a governança e a segurança dos dados**. Todos concordam que os dados precisam ser mantidos seguros, mas em muitos setores regulados, existem políticas de segurança de dados prescritas que precisam ser implementadas e é ilegal dar acesso aos dados a todos. Mesmo em alguns setores não regulamentados, isso é considerado uma má ideia. A questão é: como disponibilizamos dados sem violar regulamentos de conformidade de dados internos e externos? Isso às vezes é chamado de democratização de dados. Vamos começar nossa aula abordando esse conceito.

DEMOCRATIZAÇÃO DE DADOS

A democratização dos dados refere-se ao processo de tornar os dados acessíveis e compreensíveis para um público mais amplo dentro de uma organização, além dos especialistas em tecnologia da informação (TI) ou ciência de dados. Este movimento tem como objetivo capacitar os usuários finais, como analistas de negócios, gerentes e tomadores de decisão, a explorar e utilizar os dados para obter insights e orientar ações, sem dependerem exclusivamente de equipes técnicas para acessar e interpretar as informações.

A democratização dos dados envolve diversos aspectos, incluindo:

1. **Acesso simplificado:** Significa tornar os dados facilmente acessíveis para os usuários finais, por meio de interfaces intuitivas e ferramentas de autoatendimento. Isso pode incluir a implementação de plataformas de visualização de dados, dashboards interativos e ferramentas de consulta simples.
2. **Compreensão dos dados:** Envolve fornecer treinamento e suporte para garantir que os usuários finais entendam os conceitos básicos de dados, como tipos de dados, métricas-chave e técnicas de análise. Isso ajuda a promover uma cultura organizacional baseada em dados, onde todos têm a capacidade de interpretar e usar as informações de forma eficaz.
3. **Governança e segurança:** Embora seja importante democratizar o acesso aos dados, também é crucial garantir a segurança e a privacidade das informações. Isso requer a implementação de políticas e controles adequados para proteger os dados confidenciais e garantir a conformidade com regulamentações de privacidade, como LGPD.
4. **Colaboração e compartilhamento:** A democratização dos dados promove a colaboração entre diferentes equipes e departamentos, permitindo que compartilhem informações e insights de maneira transparente e eficiente. Isso pode ser facilitado

por meio de plataformas de colaboração e comunicação, onde os usuários podem compartilhar análises, comentários e descobertas.

Em resumo, a democratização dos dados é essencial para capacitar as organizações a aproveitarem todo o potencial de seus dados, promovendo uma cultura orientada por dados e permitindo que todos os membros da equipe contribuam para a tomada de decisões informadas e baseadas em evidências.

Agora vamos voltar nossa atenção aos conceitos de Data Lake, sempre tentando entender como ele pode contribuir para a democratização dos dados.

NÍVEL DE MATURIDADE DE UM DATA LAKE

Os níveis de maturidade de um data lake são uma maneira de entender e definir as diferentes etapas pelas quais uma organização pode passar ao adotar essa tecnologia. Aqui estão os estágios de maturidade conforme descritos no texto de referência:

1. **Data Puddle:** Um "data puddle" é essencialmente um data mart ou repositório de dados de propósito único, geralmente construído para um projeto específico ou equipe. Ele representa o primeiro passo na adoção da tecnologia de big data e é caracterizado por conter dados bem conhecidos e compreendidos. O principal motivo para o uso de tecnologia de big data é a redução de custos e o aumento de desempenho em comparação com abordagens tradicionais de armazenamento de dados.
2. **Data Pond:** Um "data pond" é uma coleção de data puddles, semelhante a um armazém de dados mal projetado. Pode ser uma agregação de data marts ou um offload de um armazém de dados existente para uma plataforma de big data. Embora ofereça benefícios como custos mais baixos e melhor escalabilidade, ainda requer alto envolvimento da equipe de TI. Além disso, os dados são limitados ao escopo do projeto e não contribuem para a democratização do uso de dados ou a tomada de decisões orientada por dados para os usuários de negócios.
3. **Data Lake:** Um data lake difere de um data pond em dois aspectos importantes. Primeiro, ele suporta o autoatendimento, permitindo que os usuários de negócios encontrem e usem conjuntos de dados sem depender da equipe de TI. Em segundo lugar, visa conter dados que os usuários de negócios possam vir a precisar, mesmo que não haja um projeto específico exigindo esses dados no momento.
4. **Data Ocean:** Um "data ocean" expande o autoatendimento e a tomada de decisões orientada por dados para todos os dados corporativos, independentemente de terem sido carregados no data lake ou não.

À medida que a maturidade cresce de um puddle para um pond, de um pond para um lake e de um lake para um ocean, a quantidade de dados e o número de usuários aumentam, às vezes de forma significativa. O padrão de uso evolui de um envolvimento intenso da TI para o autoatendimento, e os dados se expandem além do necessário para projetos imediatos.

Esses estágios de maturidade refletem a evolução contínua das capacidades e do valor proporcionado pelos data lakes em uma organização, destacando a importância de uma abordagem estratégica e progressiva para sua implementação e adoção.

A figura abaixo ilustra as diferenças entre esses conceitos. À medida que a maturidade cresce, de uma poça a um lago, a um lago e a um oceano, a quantidade de dados e o número de usuários aumentam – às vezes de forma bastante dramática. O padrão de uso passa de um envolvimento de TI de alto contato para o autoatendimento, e os dados se expandem além do que é necessário para projetos imediatos.

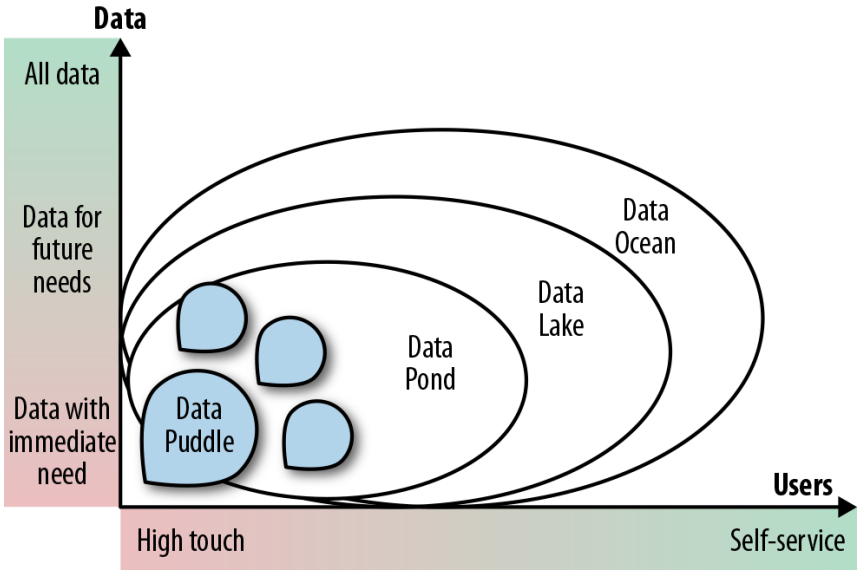


Figura 21 - Os 4 estágios de maturidade de um Data Lake

CRIANDO UM DATA LAKE

Para ter um data lake bem-sucedido, é essencial alinhá-lo com a estratégia de negócios da empresa e contar com o patrocínio executivo e o apoio amplo. Além disso, com base em discussões com dezenas de empresas que implementam data lakes com diversos níveis de sucesso, três pré-requisitos-chave podem ser identificados:

- A plataforma certa
- Os dados certos
- As interfaces certas

A plataforma certa

Uma decisão crucial na jornada rumo a um data lake bem-sucedido é a escolha da plataforma adequada. As tecnologias de big data, como Hadoop e soluções em nuvem como Amazon Web Services (AWS), Microsoft Azure e Google Cloud Platform, emergiram como as principais opções para hospedar um data lake. Essas plataformas oferecem uma série de vantagens que as tornam ideais para lidar com os desafios e demandas de um ambiente de big data:

Escalabilidade: Uma característica fundamental das plataformas de big data é a capacidade de escalar horizontalmente, permitindo que os recursos sejam expandidos conforme necessário. Isso significa que, à medida que a quantidade de dados cresce, a plataforma pode se adaptar automaticamente, garantindo desempenho consistente sem comprometer a qualidade do serviço.

Custo-Eficiência: As tecnologias de big data oferecem uma alternativa muito mais econômica em comparação com os sistemas tradicionais de armazenamento e processamento de dados. Ao aproveitar recursos como armazenamento em nuvem e distribuição de carga, essas plataformas permitem que as empresas reduzam significativamente seus custos operacionais, sem sacrificar o desempenho ou a capacidade de processamento.

Versatilidade: As plataformas de big data são altamente versáteis e podem lidar com uma ampla variedade de tipos de dados, desde estruturados até não estruturados. Isso permite que as empresas armazenem e processem dados de diferentes fontes, incluindo redes sociais, dispositivos IoT e logs de servidores, entre outros. Essa flexibilidade é essencial para empresas que buscam obter insights valiosos a partir de uma variedade de fontes de dados.

Prova de futuro (future-proofing): Uma das maiores vantagens das plataformas de big data é sua capacidade de se adaptar e evoluir com o tempo. À medida que novas tecnologias e técnicas de processamento de dados surgem, essas plataformas podem incorporá-las facilmente, garantindo que os dados armazenados permaneçam relevantes e utilizáveis no futuro. Isso proporciona às empresas uma sensação de segurança e confiança em seus investimentos em dados a longo prazo.

Em resumo, ao escolher a plataforma certa para o seu data lake, é essencial considerar fatores como escalabilidade, custo-eficiência, versatilidade e capacidade de adaptação. Ao optar por tecnologias de big data comprovadas, as empresas podem criar um ambiente robusto e flexível que atenda às suas necessidades de dados atuais e futuras.

Os dados certos

Além da escolha da plataforma adequada, a seleção dos dados certos desempenha um papel crucial no sucesso de um data lake. Aqui estão alguns pontos-chave a serem considerados ao decidir quais dados incluir:

Amplitude: O objetivo de um data lake é armazenar uma ampla variedade de dados, tanto estruturados quanto não estruturados. Isso inclui dados transacionais, dados de logs, dados de sensores IoT, feeds de redes sociais e muito mais. Ao capturar uma ampla gama de dados, as organizações podem obter uma visão completa e holística de suas operações e do ambiente externo em que operam.

Profundidade: Além da amplitude, a profundidade dos dados também é importante. Isso se refere à quantidade de detalhes e granularidade dos dados armazenados. Ter acesso a dados detalhados e de alta granularidade pode ser crucial para análises mais avançadas e insights mais precisos. No entanto, é essencial equilibrar a profundidade dos dados com considerações de custo e desempenho.

Integridade: Os dados incluídos no data lake devem ser precisos, confiáveis e consistentes. Isso significa garantir que os dados sejam limpos, normalizados e devidamente documentados antes de serem carregados no data lake. A integridade dos dados é fundamental para garantir a confiabilidade e a precisão das análises realizadas no data lake.

Relevância: É importante priorizar dados que sejam relevantes para os objetivos de negócios da organização. Isso envolve identificar os principais indicadores de desempenho (KPIs) e fontes de dados que possam fornecer insights valiosos para informar a tomada de

decisões. Ao focar nos dados mais relevantes, as empresas podem maximizar o valor obtido do data lake e evitar sobrecarregar o sistema com dados desnecessários.

Governança: A governança dos dados desempenha um papel fundamental na seleção dos dados certos para o data lake. Isso envolve estabelecer políticas e procedimentos para garantir a qualidade, segurança e conformidade dos dados armazenados. A implementação de práticas de governança eficazes ajuda a garantir que apenas os dados adequados e autorizados sejam incluídos no data lake, mitigando o risco de problemas de qualidade ou conformidade.

Ao considerar esses fatores ao selecionar os dados para o seu data lake, as organizações podem garantir que estão construindo um ambiente de dados robusto e eficaz que atenda às suas necessidades de análise e tomada de decisão. A seleção cuidadosa dos dados certos é essencial para extrair o máximo valor do data lake e impulsionar o sucesso dos negócios.

A interface correta

A interface de um data lake desempenha um papel crucial na facilitação do acesso e uso dos dados por parte dos usuários finais. Aqui estão algumas considerações importantes ao selecionar a interface correta:

Self-Service: Uma interface eficaz de data lake deve capacitar os usuários a realizarem suas próprias análises e consultas sem dependerem da intervenção constante da equipe de TI. Isso significa fornecer ferramentas e recursos que permitam aos usuários explorarem, visualizar e analisar os dados de forma independente, de acordo com suas necessidades e habilidades.

Facilidade de Uso: A interface do data lake deve ser intuitiva e fácil de usar, mesmo para usuários sem experiência técnica avançada. Isso inclui recursos como uma interface gráfica amigável, recursos de pesquisa poderosos e navegação simplificada para ajudar os usuários a encontrarem e acessar os dados de que precisam de forma rápida e eficiente.

Personalização: Uma interface flexível e personalizável permite que os usuários adaptem o ambiente de trabalho de acordo com suas preferências e necessidades individuais. Isso pode incluir a capacidade de criar painéis personalizados, salvar consultas frequentes e configurar alertas para monitorar métricas importantes.

Governança: Embora a autonomia dos usuários seja importante, é igualmente essencial garantir que a interface do data lake esteja em conformidade com as políticas e regulamentos de governança de dados da organização. Isso pode incluir recursos como controle de acesso granular, rastreamento de auditoria e funcionalidades de governança de dados integradas para garantir a segurança e a conformidade dos dados.

Integração: Uma interface de data lake eficaz deve ser capaz de se integrar facilmente a outras ferramentas e sistemas utilizados pela organização. Isso pode incluir integrações com ferramentas de visualização de dados, plataformas de business intelligence (BI), sistemas de gerenciamento de conteúdo e muito mais. A integração perfeita ajuda a garantir uma experiência de usuário contínua e uma adoção mais ampla do data lake em toda a organização.

Suporte e Treinamento: Por fim, é importante fornecer suporte adequado e recursos de treinamento para os usuários que estão utilizando a interface do data lake. Isso pode incluir documentação detalhada, tutoriais online, sessões de treinamento presenciais e suporte técnico para ajudar os usuários a aproveitarem ao máximo as capacidades do data lake.

Ao considerar esses aspectos ao escolher a interface correta para o seu data lake, as organizações podem garantir uma experiência de usuário superior, promover a adoção generalizada do data lake e capacitar os usuários a obterem insights valiosos a partir dos dados. Uma interface bem projetada e funcional é fundamental para o sucesso contínuo do data lake e para o apoio às iniciativas de análise de dados da organização.

O PÂNTANO DE DADOS (DATA SWAMP)

O conceito de "pântano de dados" é frequentemente mencionado em contraste com o conceito de "data lake". Enquanto o data lake é estruturado e organizado para permitir o acesso e a análise eficientes dos dados, o pântano de dados refere-se a uma situação em que os dados não são gerenciados de forma eficaz ou estruturada. São características do Pântano de Dados:

1. **Desorganização:** No pântano de dados, os dados geralmente são armazenados sem organização ou estruturação adequada. Isso pode levar a uma falta de clareza sobre quais dados estão disponíveis, onde estão localizados e como podem ser acessados.
2. **Baixa Qualidade dos Dados:** Os dados no pântano de dados podem ser de baixa qualidade devido à falta de controle de qualidade e validação. Isso pode incluir dados duplicados, inconsistentes ou incompletos, o que pode prejudicar a precisão das análises e insights derivados desses dados.
3. **Falta de Governança:** A governança dos dados costuma ser ausente ou inadequada no pântano de dados. Isso significa que não há políticas, processos ou controles estabelecidos para garantir a segurança, privacidade, conformidade e integridade dos dados.
4. **Acesso Limitado:** No pântano de dados, o acesso aos dados pode ser restrito ou desigual. Isso pode resultar em dificuldades para os usuários encontrarem e acessarem os dados de que precisam para suas análises e tomadas de decisão.
5. **Custo Elevado:** Embora os dados sejam armazenados, sua falta de organização e qualidade pode resultar em custos operacionais mais altos. Isso ocorre devido à necessidade de limpar, preparar e validar os dados antes que possam ser utilizados de forma eficaz.

Como Evitar o Pântano de Dados

1. **Planejamento Adequado:** Um planejamento cuidadoso é essencial para evitar a criação de um pântano de dados. Isso inclui a definição de requisitos de dados claros, a identificação de fontes de dados relevantes e a elaboração de uma estratégia de gestão de dados abrangente.
2. **Governança Efetiva:** Estabelecer políticas robustas de governança de dados é fundamental para garantir a qualidade, segurança e conformidade dos dados. Isso

inclui a implementação de processos de controle de qualidade, políticas de segurança de dados e mecanismos de conformidade regulatória.

3. **Estruturação e Organização:** Os dados devem ser estruturados e organizados de forma lógica e coerente para facilitar o acesso e a análise. Isso pode envolver a criação de modelos de dados consistentes, a definição de metadados claros e a implementação de hierarquias de dados eficazes.
4. **Colaboração e Comunicação:** Promover a colaboração e a comunicação entre as equipes de dados, TI e negócios é essencial para garantir o alinhamento de objetivos e a eficácia das iniciativas de gestão de dados. Isso ajuda a evitar silos de dados e promove uma cultura de dados compartilhada dentro da organização.

Ao adotar uma abordagem proativa para a gestão de dados e implementar as práticas recomendadas acima, as organizações podem evitar os desafios associados ao pântano de dados e criar um ambiente de dados mais eficiente, estruturado e eficaz. Isso permite que eles aproveitem ao máximo o valor de seus ativos de dados e impulsionem melhores insights e decisões de negócios.

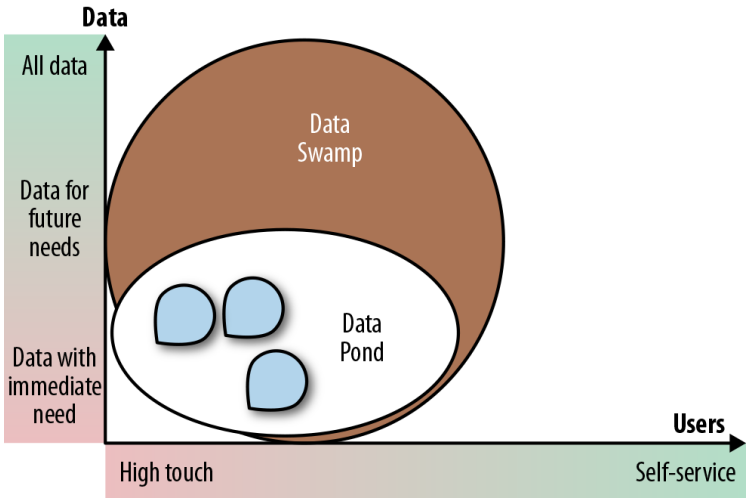


Figura 22 - Um pântano de dados (data swamp)

TRILHA PARA O SUCESSO EM DATA LAKE

Agora que sabemos o que é necessário para que um data lake seja bem-sucedido e quais armadilhas devemos observar, como faremos para construí-lo? Normalmente, as empresas seguem este processo:

1. Levantar a infraestrutura (colocar o cluster Hadoop em funcionamento).
2. Organizar o data lake (criar zonas para uso de diversas comunidades de usuários e ingerir os dados).
3. Configurar o data lake para autoatendimento (criar um catálogo de ativos de dados, configure permissões e forneça ferramentas para uso dos analistas).
4. Abrir o data lake para os usuários.

Vamos passar por cada uma dessas etapas:

Colocando o Data Lake em funcionamento

Construir um data lake envolve um processo de várias etapas para garantir o sucesso da implementação e uso eficaz dos dados. A primeira etapa desse processo é levantar a infraestrutura necessária para suportar o data lake, incluindo a configuração de um cluster Hadoop funcional.

Nesta fase inicial, é essencial provisionar os recursos de hardware e software necessários para hospedar o cluster Hadoop. Isso pode envolver a seleção de uma plataforma de nuvem adequada, como Amazon Web Services (AWS), Microsoft Azure ou Google Cloud Platform, ou a configuração de servidores físicos em um data center local.

Uma vez escolhida a plataforma, é preciso configurar o ambiente do cluster Hadoop, incluindo a instalação e configuração dos componentes principais, como o Hadoop Distributed File System (HDFS), YARN (Yet Another Resource Negotiator) e MapReduce. Além disso, é necessário garantir a conectividade de rede adequada entre os nós do cluster e configurar os parâmetros de segurança, como firewalls, autenticação e autorização.

Após a conclusão da configuração inicial, o cluster Hadoop estará pronto para receber e processar os dados que serão armazenados no data lake. No entanto, é importante ressaltar que a infraestrutura do data lake é escalável e pode ser expandida conforme necessário para lidar com o crescimento futuro dos dados e das demandas de processamento. Essa capacidade de escalabilidade é uma das vantagens-chave do uso de tecnologias de big data como o Hadoop para construir data lakes.

Construir um data lake oferece diversas abordagens de arquitetura, cada uma com suas próprias vantagens e considerações. As três abordagens mais comuns são: On Premises (local), Cloud (nuvem) e Logical Data Lake (data lake lógico).

1. **On Premises (Local):** Nesta abordagem, o data lake é construído e mantido dentro das instalações da empresa, utilizando hardware e software próprios. Isso oferece controle total sobre a infraestrutura e os dados, ideal para organizações que têm requisitos rigorosos de segurança ou conformidade regulatória. No entanto, pode exigir investimentos significativos em hardware, software e equipe de TI para configurar e manter o data lake.
2. **Cloud (Nuvem):** Com essa abordagem, o data lake é implantado em uma infraestrutura de nuvem fornecida por provedores como Amazon Web Services (AWS), Microsoft Azure ou Google Cloud Platform. Isso oferece escalabilidade instantânea, flexibilidade e elimina a necessidade de gerenciar hardware físico. As empresas pagam apenas pelos recursos que usam, o que pode resultar em custos mais baixos e maior agilidade. No entanto, a segurança e a conformidade ainda são preocupações, e a transferência de grandes volumes de dados para a nuvem pode ser demorada e cara em algumas situações.
3. **Logical Data Lake (Data Lake Lógico):** Essa abordagem trata o data lake como uma camada virtual sobre múltiplos sistemas de dados heterogêneos, como Hadoop, bancos de dados relacionais ou NoSQL, tanto on premises quanto na nuvem. Isso permite que as empresas centralizem o acesso aos dados, independentemente de onde eles residam, simplificando a integração e o acesso aos dados para os usuários. O data lake lógico é altamente flexível e escalável, permitindo que as empresas se

adaptem rapidamente a novas fontes de dados e requisitos de negócios em constante mudança. No entanto, a complexidade de gerenciar múltiplos sistemas de dados e garantir a consistência e a integridade dos dados pode ser um desafio.

Cada abordagem de arquitetura tem suas próprias considerações de custo, desempenho, segurança e conformidade, e a escolha entre elas dependerá das necessidades e objetivos específicos de cada organização. A figura abaixo tenta mostrar as três soluções:

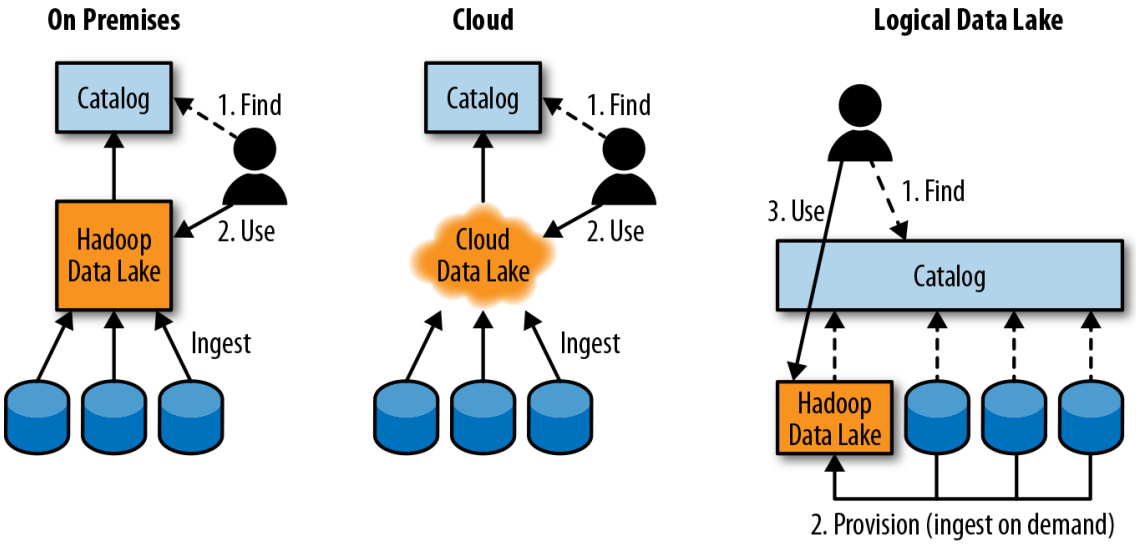


Figura 23 - Diferentes arquiteturas de data lake

Organizando um Data Lake

Um data lake é geralmente organizado em diferentes zonas para facilitar o gerenciamento e o acesso aos dados. Cada zona desempenha um papel específico no ciclo de vida dos dados e na utilização pelos usuários. As principais zonas presentes em um data lake incluem:

1. **Zona Bruta ou de Aterrissagem:** Esta é a primeira parada para os dados recém-ingressados no data lake. Também conhecida como "zona de aterrissagem", esta área contém os dados em sua forma bruta e original, sem qualquer processamento ou transformação. Os dados são carregados nesta zona diretamente de suas fontes de origem, como sistemas transacionais, sensores ou feeds de streaming. A zona bruta serve como um repositório centralizado para todos os dados brutos, garantindo que nada seja perdido e permitindo que os usuários realizem análises detalhadas e refinem os dados conforme necessário.
2. **Zona de Ouro ou Produção:** Na zona de ouro, os dados são refinados, limpos e transformados em uma forma pronta para uso em análises avançadas e aplicativos de produção. Esta área é onde os dados são enriquecidos, harmonizados e agregados para criar conjuntos de dados de alta qualidade e confiáveis. A zona de ouro é frequentemente governada por políticas de qualidade de dados e segurança, garantindo a integridade e a precisão dos dados disponíveis para os usuários finais.
3. **Zona de Desenvolvimento ou Trabalho:** Esta zona é reservada para atividades de desenvolvimento, teste e experimentação. Os cientistas de dados e analistas podem

usar esta área para explorar novas fontes de dados, criar modelos de machine learning, realizar experimentos e desenvolver novos pipelines de dados. Como é uma área de trabalho flexível e experimental, os dados aqui podem ser menos governados e estruturados, permitindo uma rápida iteração e inovação.

4. **Zona Sensível:** Esta zona é designada para dados sensíveis que requerem níveis mais elevados de segurança e conformidade. Dados pessoais, informações financeiras ou outros dados confidenciais são armazenados e acessados nesta zona com medidas de segurança rigorosas, como criptografia, controle de acesso e auditoria. A zona sensível é estritamente controlada e acessada apenas por usuários autorizados com permissões adequadas.

Cada zona em um data lake desempenha um papel fundamental na garantia da qualidade, segurança e acessibilidade dos dados, permitindo que as organizações aproveitem ao máximo seu ambiente de dados para impulsionar a inovação e a tomada de decisões baseadas em dados.

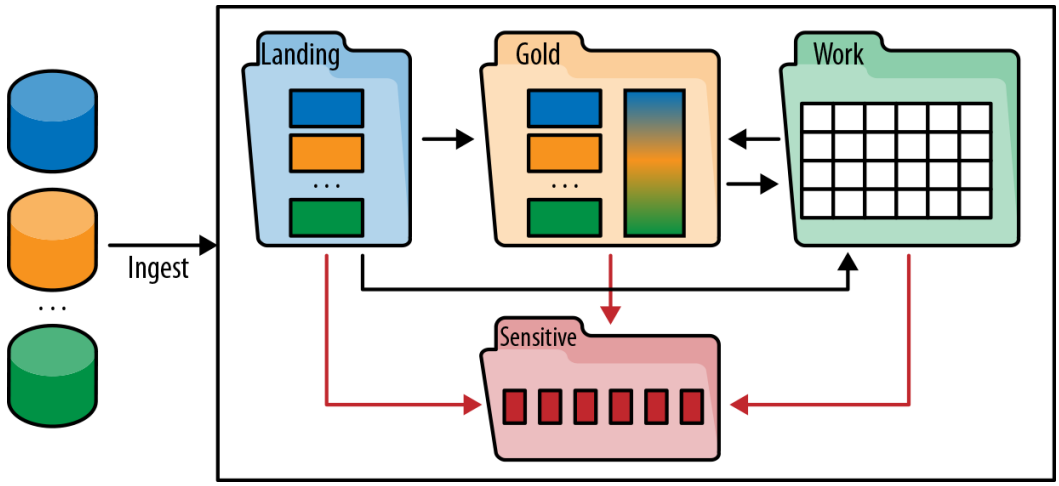


Figura 24 - Zonas típicas de um data lake

Durante muitos anos, a sabedoria predominante para as equipes de governança de dados era que os dados deveriam estar sujeitos à mesma governança, independentemente de sua localização ou propósito. No entanto, nos últimos anos, analistas do setor, como os da Gartner, vêm promovendo o conceito de TI multimodal - basicamente, a ideia de que a governança deve refletir o uso dos dados e os requisitos da comunidade de usuários. Esta abordagem tem sido amplamente adotada pelas equipes de data lake, com diferentes zonas tendo diferentes níveis de governança e acordos de nível de serviço (SLAs).

Por exemplo, os dados na zona de ouro geralmente são fortemente governados, são bem curados e documentados, e possuem SLAs de qualidade e frescor, enquanto os dados na área de trabalho têm uma governança mínima (principalmente garantindo que não haja dados sensíveis) e SLAs que podem variar de projeto para projeto.

Diferentes comunidades de usuários naturalmente gravitam para diferentes zonas. Analistas de negócios usam dados principalmente na zona de ouro, engenheiros de dados trabalham com dados na zona bruta (convertendo-os em dados de produção destinados à zona de ouro), e cientistas de dados executam seus experimentos na zona de trabalho. Embora alguma governança seja necessária para cada zona para garantir que dados sensíveis

sejam detectados e protegidos, os administradores de dados geralmente se concentram nos dados nas zonas sensíveis e de ouro, para garantir que estejam em conformidade com as regulamentações da empresa e do governo. A figura a seguir ilustra os diferentes níveis de governança e as diferentes comunidades de usuários para diferentes zonas.

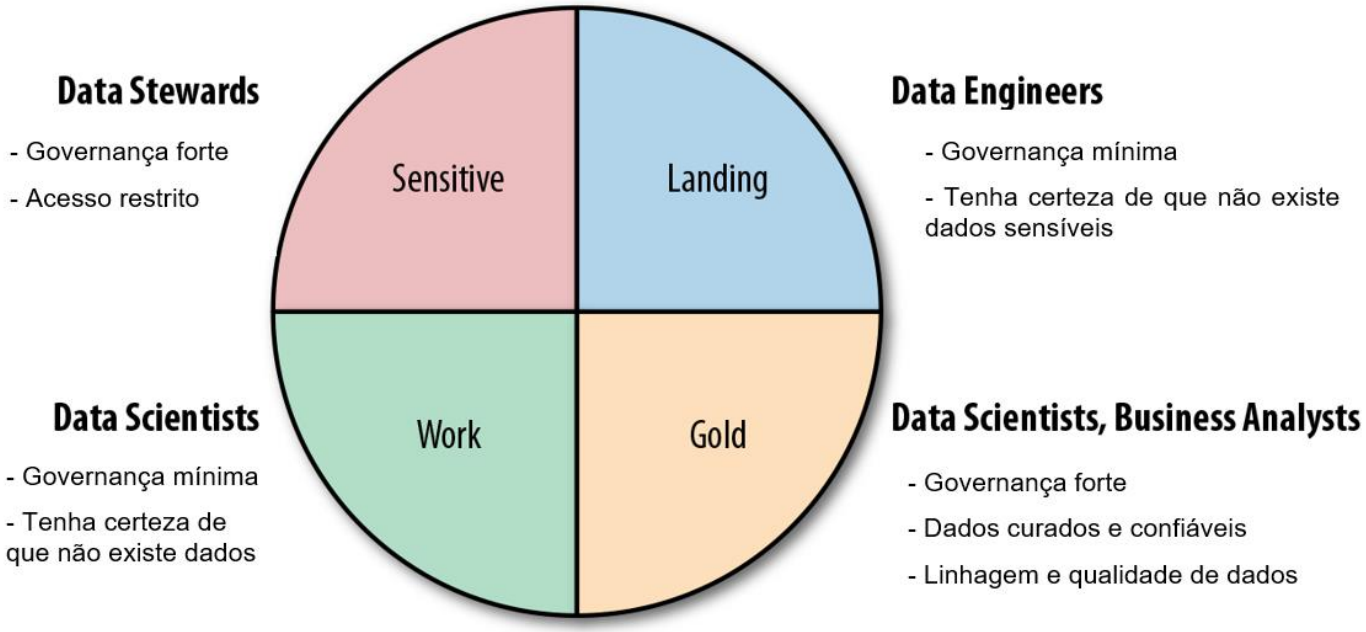


Figura 25 - Expectativas de governança por zona

Configurar o Data Lake para autoatendimento

Configurar o data lake para autoatendimento é uma etapa crucial para garantir que os usuários possam acessar, entender e utilizar os dados de forma independente, sem depender constantemente da equipe de TI. Isso envolve a implementação de várias ferramentas, processos e políticas que capacitam os usuários a explorarem e analisarem os dados de maneira eficaz. Aqui estão algumas das principais considerações ao configurar o data lake para autoatendimento:

1. **Catálogo de ativos de dados**: Implemente um catálogo abrangente que liste todos os ativos de dados disponíveis no data lake, incluindo metadados detalhados sobre cada conjunto de dados. Isso permite que os usuários pesquisem e descubram facilmente os dados de que precisam.
2. **Permissões e segurança**: Estabeleça políticas de segurança robustas para controlar o acesso aos dados e garantir que apenas usuários autorizados possam visualizar e manipular informações sensíveis. Isso pode envolver a implementação de controle de acesso baseado em função e criptografia de dados.
3. **Ferramentas de análise e visualização**: Forneça aos usuários uma variedade de ferramentas de análise e visualização que sejam fáceis de usar e ofereçam recursos avançados para explorar e interpretar os dados. Isso pode incluir ferramentas de BI (Business Intelligence), visualização de dados e análise preditiva.
4. **Integração de dados**: Garanta que os dados sejam integrados e harmonizados de maneira adequada, para que os usuários possam acessar e combinar informações

de várias fontes sem dificuldade. Isso pode envolver o uso de ferramentas de preparação de dados e pipelines de ingestão de dados automatizados.

5. **Treinamento e suporte**: Ofereça treinamento abrangente e suporte contínuo aos usuários para ajudá-los a aproveitar ao máximo as ferramentas e recursos do data lake. Isso pode incluir sessões de treinamento presenciais ou online, documentação detalhada e suporte técnico dedicado.

Ao configurar o data lake para autoatendimento, as organizações podem capacitar seus usuários a explorarem e extrair insights valiosos dos dados de forma rápida e eficiente, promovendo uma cultura de análise de dados e tomada de decisões baseada em dados em toda a empresa.

Abrir o data lake para os usuários

Os analistas, sejam eles analistas de negócios, analistas de dados ou cientistas de dados, geralmente passam por quatro etapas para realizar seu trabalho. Essas etapas são ilustradas na figura a seguir.

Figura 26 - Os quatro estágios da análise

O processo de preparação e análise de dados realizado pelos analistas, é dividido em quatro etapas distintas:

1. **Encontrar e entender os dados**: Esta etapa envolve a identificação e compreensão dos conjuntos de dados relevantes. Os analistas muitas vezes enfrentam dificuldades para localizar os dados devido à sua variedade e complexidade. Ferramentas de crowdsourcing de analistas estão sendo desenvolvidas para abordar esse problema, permitindo que os analistas documentem conjuntos de dados usando descrições simples compostas por termos de negócios.
2. **Acessar e provisionar os dados**: Uma vez identificados os conjuntos de dados corretos, os analistas precisam acessá-los. Tradicionalmente, o acesso é concedido a partir do início de um projeto. Uma abordagem mais prática é publicar informações sobre todos os conjuntos de dados em um catálogo de metadados, permitindo que os analistas encontrem conjuntos de dados úteis e solicitem acesso conforme necessário.

3. **Preparar os dados**: Muitas vezes, os dados precisam ser preparados antes da análise. Isso pode envolver operações como modelagem, limpeza e combinação de dados de várias fontes. Ferramentas de preparação de dados, como Trifacta, são usadas para automatizar e simplificar esse processo. Vejamos mais algumas informações relacionadas as operações sobre os dados:
    - **Moldagem**: Esta operação envolve a seleção de um subconjunto de campos e registros para trabalhar, combinando múltiplos arquivos e tabelas em um (junção), transformando e agregando dados, criando intervalos ou "buckets" para variáveis discretas e convertendo variáveis em recursos úteis para análise.
    - **Limpeza**: Durante a limpeza dos dados, são realizadas ações para preencher valores ausentes, corrigir valores inconsistentes, resolver conflitos entre dados, normalizar unidades de medida e códigos, entre outras atividades. Isso garante que os dados estejam consistentes e prontos para análise.
    - **Combinação**: A operação de combinação envolve a harmonização de diferentes conjuntos de dados para o mesmo esquema, as mesmas unidades de medida, os mesmos códigos, etc. Isso facilita a comparação e análise de diferentes conjuntos de dados, garantindo a consistência e integridade dos resultados.

    Essas operações são essenciais para preparar os dados para análise e garantir que os resultados sejam confiáveis e precisos. Automatizar essas operações sempre que possível pode economizar tempo e minimizar erros humanos.
4. **Análise e visualização**: Uma vez preparados, os dados podem ser analisados e visualizados. Isso pode variar desde a criação de relatórios simples até análises avançadas e aprendizado de máquina. Existem inúmeras ferramentas disponíveis para realizar análises e visualizações, incluindo soluções específicas para data lakes baseados em Hadoop, como Arcadia Data e AtScale.

ARQUITETURAS DE DATA LAKE

Inicialmente, muitas empresas acreditavam que teriam um único e enorme data lake local, que conteria todos os seus dados. Conforme a compreensão e as melhores práticas evoluíram, a maioria das empresas percebeu que um único ponto central não era o ideal. Entre regulamentações de soberania de dados (por exemplo, não é permitido retirar dados da Alemanha) e pressões organizacionais, geralmente se constatou que vários lagos de dados eram uma solução melhor.

Além disso, à medida que as empresas percebiam a complexidade de suportar um cluster massivamente paralelo e experimentavam a frustração por sua incapacidade de encontrar e contratar administradores experientes para plataformas Hadoop e outras de big data, elas começaram a optar por data lakes baseados em nuvem, onde a maioria dos componentes de hardware e plataforma são gerenciados por especialistas que trabalham para Amazon, Microsoft, Google e outros.

Lago de dados em nuvens públicas

O lago de dados em nuvem pública oferece uma abordagem altamente atraente para implementar uma infraestrutura de big data. Além dos benefícios do acesso à expertise em tecnologia de big data e dos curtos tempos de implantação, o baixo custo de armazenamento e a natureza elástica da computação em nuvem tornam essa opção extremamente atrativa. Dado que muitos dados estão sendo armazenados para uso futuro, faz sentido armazená-los da maneira mais econômica possível. Isso se alinha bem com as possibilidades de otimização de custos oferecidas por várias camadas de armazenamento fornecidas pela Amazon e outros provedores de nuvem: o acesso varia de alta velocidade a glacial, sendo os meios de acesso mais lentos significativamente mais baratos.

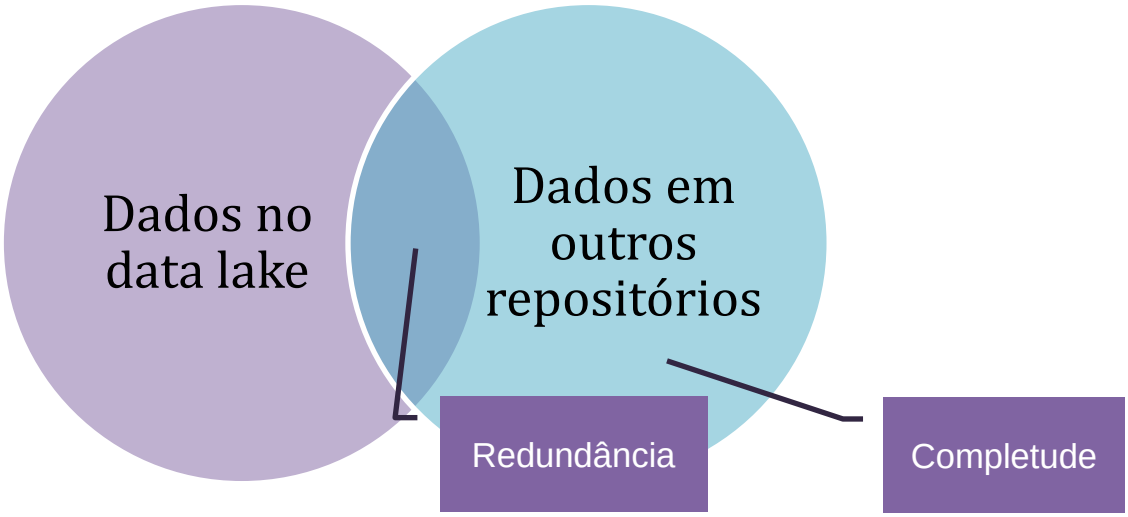
Além disso, a **elasticidade** da computação em nuvem permite a criação de um cluster muito grande conforme a demanda, quando necessário. Isso é comparado a um cluster local, que tem um tamanho fixo e armazena seus dados em armazenamento anexado (embora novas arquiteturas com armazenamento anexado à rede estejam sendo exploradas). Isso significa que, à medida que os nós são preenchidos com dados, novos nós precisam ser adicionados apenas para armazenamento. Além disso, se as cargas analíticas exigirem muito da CPU e precisarem de mais poder de computação, será necessário adicionar nós, mesmo que eles sejam usados apenas por um curto período.

Na nuvem, você paga apenas pelo armazenamento que precisa (ou seja, não é necessário comprar nós de computação extras apenas para obter mais armazenamento) e pode criar clusters enormes por curtos períodos. Por exemplo, se você tiver um cluster local de 100 nós e um trabalho que leve 50 horas, não é prático comprar e instalar 1.000 nós apenas para fazer esse único trabalho ser executado mais rápido. Na nuvem, no entanto, você pagaria aproximadamente o mesmo pela potência de computação de 100 nós por 50 horas do que pagaria por 1.000 nós por 5 horas. Essa é a enorme vantagem da **computação elástica.**

Data lakes lógicos

Uma vez que as empresas perceberam que ter um único lago de dados centralizado não era uma boa solução, a ideia do **lago de dados lógico** ganhou força. Com essa abordagem, em vez de carregar todos os dados no lago de dados apenas no caso de alguém eventualmente precisar deles, eles são disponibilizados para os analistas por meio de um catálogo central ou por meio de software de virtualização de dados.

Os lagos de dados lógicos abordam as questões de completude e redundância, como ilustrado na figura a seguir.



Essas questões podem ser resumidas da seguinte forma:

Completude Como os analistas encontram o melhor conjunto de dados? Se os analistas só podem encontrar dados que já estão no lago de dados, outros dados que não foram ingeridos no lago de dados não serão encontrados ou usados.

Redundância Se ingerirmos todos os dados no lago de dados, teremos redundância entre as fontes de dados e o lago de dados (ilustrado como a área de sobreposição entre os dois círculos na figura acima). Com vários lagos de dados, para alcançar a completude, precisaríamos ingerir os mesmos dados em cada lago de dados.

Para piorar, já existe muita redundância na empresa. Tradicionalmente, quando um novo projeto é iniciado, a abordagem mais rápida e politicamente simples é para a equipe do projeto criar um depósito de dados, copiar dados de outras fontes ou do data warehouse e adicionar seus próprios dados exclusivos. Isso é muito mais fácil do que estudar os depósitos de dados existentes e negociar o uso compartilhado com os proprietários e usuários atuais. Como resultado, há uma proliferação de depósitos de dados que são em sua maioria idênticos. Se carregarmos cegamente todos os dados desses depósitos de dados no lago de dados, teremos níveis extremamente altos de redundância em nosso lago.

A melhor abordagem para os desafios de completude e redundância que eu vi envolve algumas regras simples:

- Para resolver o problema de completude, crie um catálogo de todos os ativos de dados, para que os analistas possam encontrar e solicitar qualquer conjunto de dados que esteja disponível na empresa.
- Para resolver o problema de redundância, siga o processo mostrado na figura abaixo:
    - Armazene dados que não estão armazenados em nenhum outro lugar no lago de dados.
    - Traga dados armazenados em outros sistemas para o lago de dados se e quando necessário e mantenha-os sincronizados enquanto necessário.
    - Traga cada conjunto de dados apenas uma vez para todos os usuários.

Dados de lagos de dados e outros sistemas

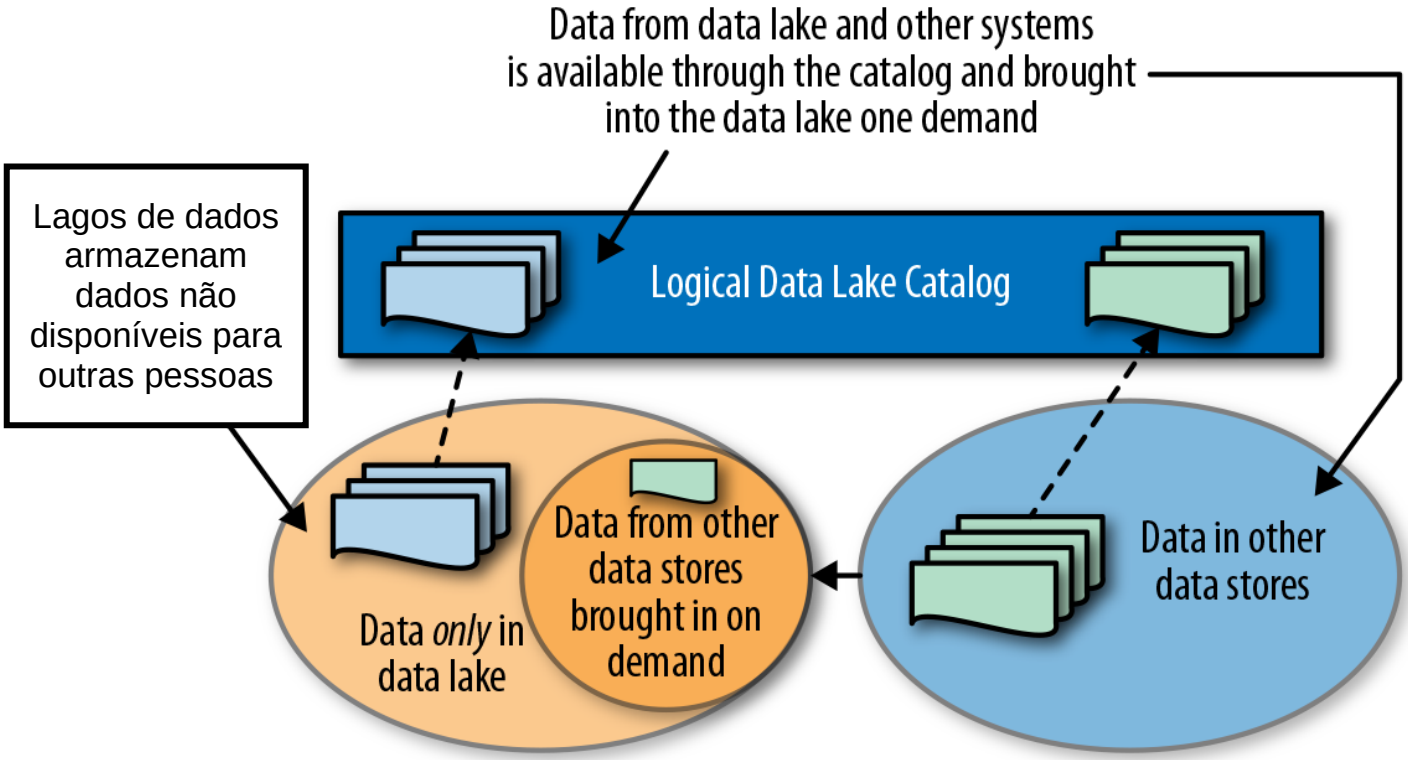


Figura 27 - Gerenciando dados no data lake lógico

Virtualização versus um data lake lógico baseado em catálogo

A virtualização (às vezes também chamada de *federação* ou EII, para *integração de informações empresariais*) é uma tecnologia desenvolvida na década de 1980 e aprimorada ao longo de várias gerações até a década de 2010. Basicamente, cria uma visualização ou tabela virtual que oculta a localização e implementação das tabelas físicas. Na figura a seguir, uma visão é criada juntando duas tabelas de bancos de dados diferentes. A consulta então consultaria essa visualização e deixaria para o sistema de virtualização de dados descobrir como acessar e unir os dados nos dois bancos de dados.

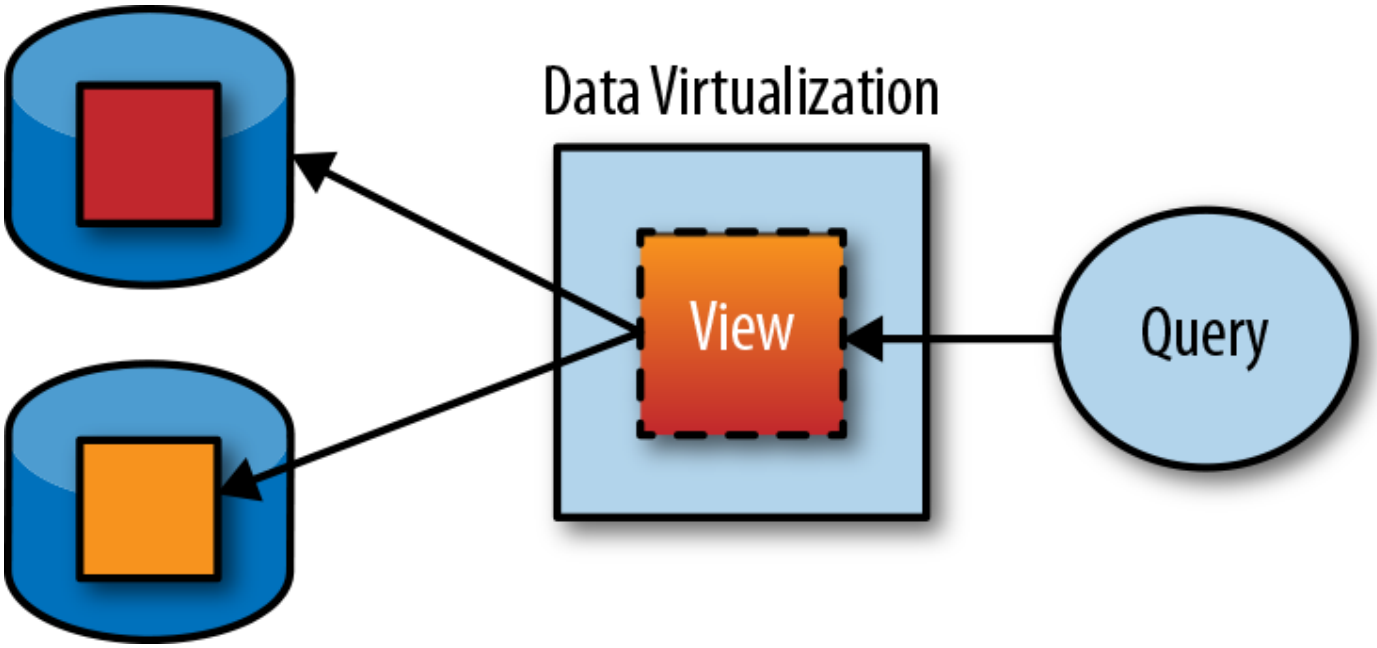


Figura 28 - Criando um conjunto de dados personalizado por meio de uma visualização

Embora esta tecnologia funcione bem para alguns casos de utilização, num data lake lógico, para atingir a integridade, seria necessário que cada conjunto de dados fosse publicado como uma tabela virtual e mantido atualizado à medida que os esquemas da tabela subjacente mudam.

Mesmo que o problema inicial de publicação de todos os ativos de dados fosse resolvido, as visualizações ainda apresentam problemas significativos:

- Criar uma visão virtual não facilita a localização dos dados.
- Unir dados de vários sistemas heterogêneos é complexo e exige muita computação, muitas vezes causando cargas massivas nos sistemas e longos ciclos de execução. Essas chamadas *junções distribuídas* de tabelas que não cabem na memória consomem muitos recursos.

Por outro lado, na abordagem orientada por catálogo, apenas os metadados sobre cada conjunto de dados é publicado, a fim de torná-lo localizável. Os conjuntos de dados são então provisionados para o mesmo sistema (por exemplo, cluster Hadoop) para serem processados localmente, conforme demonstrado na figura a seguir.

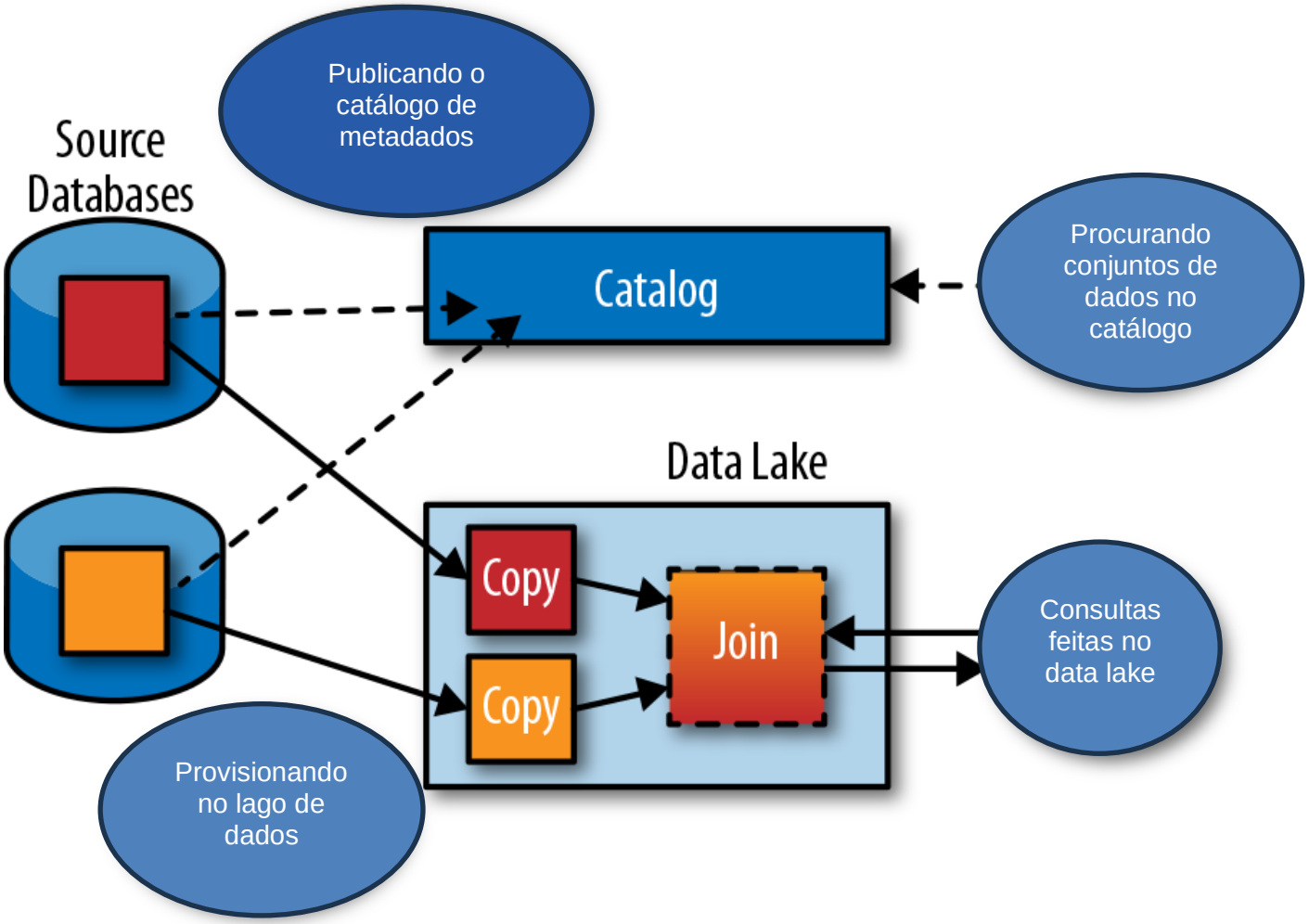


Figura 29 - Fornecendo metadados por meio de um catálogo

Além de tornar todos os dados localizáveis e acessíveis aos analistas, um catálogo corporativo pode servir como um único ponto de acesso, governança, e auditoria, conforme mostrado na figura abaixo. No topo, sem um catálogo centralizado, o acesso aos ativos de dados está em todos os lugares e é difícil de gerenciar e rastrear. Na parte inferior, com o catálogo centralizado, todas as solicitações de acesso passam pelo catálogo. O acesso é concedido sob demanda por um período específico e é auditado pelo sistema.

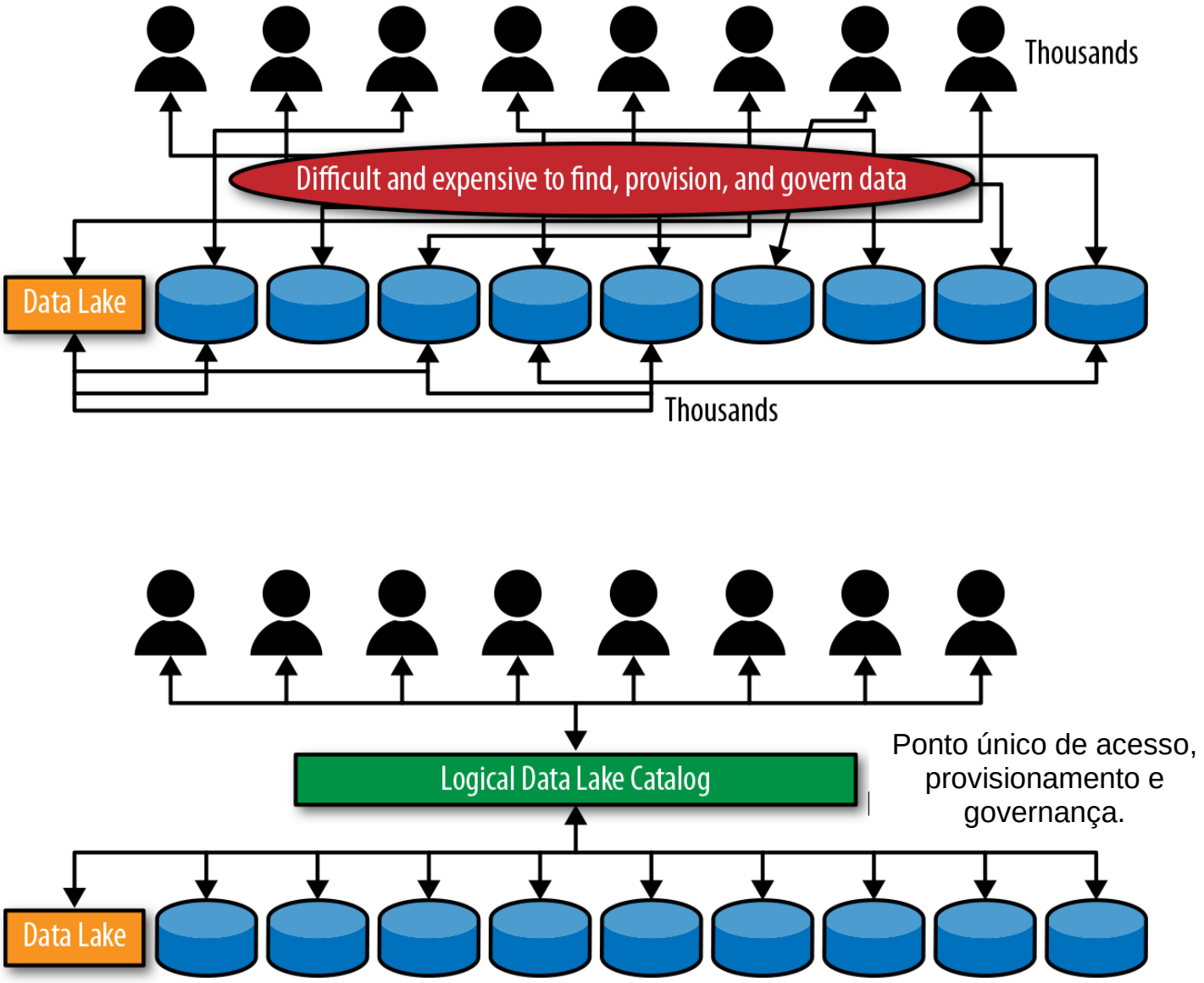


Figura 30 - Provisionamento e governança de dados por meio do catálogo

Em resumo, o processo de criação de um data lake passa por obter a plataforma certa, carregá-la e organizá-la com os dados corretos, configurá-la para autoatendimento com uma interface apropriada às habilidades e necessidades. Essas são as chaves para criar um data lake de sucesso.

QUESTÕES

Instituto Consulplan - 2023 - Analista do Executivo (SEGER ES)/Tecnologia da Informação

Considerando Data Lake, que geralmente é o armazenamento único de todos os dados corporativos, incluindo cópias brutas de dados do sistema de origem, assinale a afirmativa correta.

A Uma de suas maiores vantagens é que ele somente utiliza dados estruturados.

B Suas maiores desvantagens são: lentidão e não permitir configuração mais de uma vez.

C É um repositório de dados 100% seguro em relação à segurança dos dados e ao modelo de acesso.

D Com ele, é possível executar diversos tipos de análises em tempo real, processamentos de Big Data e aprendizado de máquina.

E Trata-se de uma ferramenta mais rápida que o Data Warehouse (DW), pois gera o esquema de dados no momento da gravação e não no momento de leitura.

Comentário: A afirmativa correta é: D. Esta é uma descrição precisa de uma das principais vantagens de um Data Lake. Ele é projetado para lidar com uma ampla variedade de dados, incluindo dados brutos e não estruturados, permitindo análises em tempo real, processamento de Big Data e aplicação de algoritmos de aprendizado de máquina em larga escala.

Gabarito: D

Instituto Consulplan - 2023 - Analista Técnico (MPE BA)/Sistemas de Informação

A ideia básica do Data Lake é simples; todos os dados emitidos pela organização serão armazenados em uma única estrutura de dados chamada Data Lake. Assinale, a seguir, um dos estágios para a implementação de um Data Lake.

A Descarregamento para Data Mining.

B Descarregamento para Data Warehouse.

C Transformação dos dados brutos em dados estruturados.

D Transformação dos dados brutos em dados semiestruturados.

E Indexação de todos os registros (criação de índices no banco para todas as tabelas).

Comentário: O estágio de "Descarregamento para Data Warehouse" como parte da implementação de um Data Lake refere-se ao processo de transferir os dados brutos ou semiestruturados para um Data Warehouse após sua ingestão inicial no Data Lake.

Embora os Data Lakes sejam projetados para armazenar dados brutos em sua forma original, muitas organizações optam por transferir parte desses dados para um Data Warehouse, onde podem ser mais processados, limpos, estruturados e otimizados para análises específicas. A transferência para um Data Warehouse geralmente ocorre quando há necessidade de:

- Realizar análises mais detalhadas e específicas que exigem dados mais estruturados e refinados.
- Implementar modelos de dados dimensionais para suportar relatórios e análises de negócios.
- Facilitar o acesso aos dados por meio de ferramentas de BI (Business Intelligence) e SQL.
- Atender a requisitos de conformidade e governança de dados que exigem metadados detalhados e históricos de alterações.
- Melhorar o desempenho das consultas ao fornecer dados pre-agregados e otimizados.

Em resumo, o estágio de "Descarregamento para Data Warehouse" representa uma etapa adicional na jornada dos dados, onde os dados brutos do Data Lake são refinados e movidos para um ambiente mais estruturado e otimizado para análises específicas e relatórios de negócios. Logo, nosso resposta encontra-se na alternativa B.

Gabarito: B

CEBRASPE (CESPE) - 2023 - Analista (SERPRO)/Tecnologia

Acerca de armazenamento e processamento de dados, julgue o item a seguir.

Uma das características do data lake é armazenar um grande volume de dados brutos e heterogêneos, oriundos de diversas fontes distintas.

Comentário: O item está correto. Uma das características fundamentais de um Data Lake é armazenar um grande volume de dados brutos e heterogêneos, provenientes de diversas fontes diferentes. O objetivo é manter os dados em sua forma original, sem a necessidade de estruturação prévia, permitindo uma ampla variedade de análises e insights posteriormente. Portanto, o item está de acordo com as características típicas de um Data Lake.

Gabarito: Certo

CEBRASPE (CESPE) - 2023 - Analista de Infraestrutura e Suporte (EMPREL)/Banco de Dados

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.

Comentário: A primeira etapa que deve estar prevista em um projeto de arquitetura de um data lake é a criação de um ambiente virtual de captura de dados. Portanto, a opção correta é a alternativa A. Isso envolve estabelecer a infraestrutura necessária para capturar e armazenar os dados brutos de diversas fontes, garantindo que o data lake tenha capacidade para lidar com grandes volumes de dados heterogêneos.

Gabarito: A

CESGRANRIO - 2023 - Profissional Transpetro de Nível Superior (TRANSPETRO)/Análise de Sistemas/Processos de Negócios

A respeito do uso de Data Lakes como solução para o gerenciamento e análise de Big Data, constata-se que eles

A convertem dados não estruturados em estruturados durante o processo de ingestão.

B permitem o armazenamento de dados em formatos estruturados, semiestruturados e não estruturados, oferecendo flexibilidade na análise e no processamento de diferentes tipos de dados.

C são geralmente implementados usando apenas sistemas de arquivos convencionais devido à sua eficiência em armazenar grandes volumes de dados.

D são adequados apenas para armazenar grandes volumes de dados não estruturados.

E são inadequados em soluções de Big Data por não suportarem alguns tipos de dados.

Comentário: A respeito do uso de Data Lakes como solução para o gerenciamento e análise de Big Data, constata-se que eles: B. Permitem o armazenamento de dados em formatos estruturados, semiestruturados e não estruturados, oferecendo flexibilidade na análise e no processamento de diferentes tipos de dados.

Data Lakes são projetados para armazenar uma ampla variedade de dados, incluindo dados estruturados, semiestruturados e não estruturados. Isso oferece flexibilidade na análise e no processamento de diferentes tipos de dados, o que é uma característica fundamental dos Data Lakes.

Gabarito: B

CEBRASPE (CESPE) - 2023 - Auditor de Controle Externo (TC /DF)/Especializada/ Sistemas de TI

No que se refere a Big Data, data lake, business intelligence e data warehousing, julgue o item seguinte.

Data lake é um repositório onde os dados podem ser armazenados em vários formatos, incluindo-se registros estruturados e formatos de arquivo não estruturados.

Comentário: Um data lake é de fato um repositório que pode armazenar uma grande variedade de dados em vários formatos, incluindo registros estruturados, semiestruturados e não estruturados. Portanto, a afirmação é verdadeira.

Gabarito: Certo

IBADE - 2022 - Analista de Informática (SEA SC)

Há um tipo de repositório de dados que centraliza e armazena todos os tipos de dados gerados pela e para a empresa. Eles são depositados ali ainda em estado bruto, sem o processamento e análise. A esse repositório chamamos:

A Data Storage.

B Data Wharehouse.

C Data Mining.

D Data Lake.

E Data Trash.

Comentário: Um Data Lake é um tipo de repositório de dados que armazena uma grande quantidade e variedade de dados brutos e não processados, incluindo dados estruturados, semiestruturados e não estruturados. Diferentemente de um Data Warehouse, que é projetado para armazenar dados estruturados e já processados para análise, o Data Lake

aceita dados em seu estado bruto original, sem a necessidade de definir sua estrutura antecipadamente.

Essa característica do Data Lake permite que as organizações armazenem todos os tipos de dados, desde transações comerciais até dados de redes sociais, logs de servidores, imagens e vídeos, entre outros, sem a necessidade imediata de definir sua utilização ou estrutura. Posteriormente, esses dados podem ser processados e analisados conforme necessário para atender às necessidades específicas da organização, o que proporciona flexibilidade e escalabilidade na análise de grandes volumes de dados de diferentes fontes.

Logo, nossa resposta encontra-se na alternativa D.

Gabarito: D

CEBRASPE (CESPE) - 2022 - Especialista Técnico (BNB)/Analista de Sistemas/Desenvolvimento de Sistemas

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.

Comentário: Para mim, Thiago, a afirmação está incorreta! O conceito de Data Lake refere-se a um único repositório centralizado de dados, onde os dados são armazenados em seu estado bruto e original, independentemente de seu formato, estrutura ou origem. Esses dados podem incluir registros estruturados, como tabelas de bancos de dados, bem como arquivos de texto não estruturados, imagens, vídeos e outros tipos de dados. Portanto, o Data Lake não se refere à dispersão de dados em vários sistemas de armazenamento, mas sim a um único local onde todos os dados são centralizados e podem ser acessados para análise e processamento.

Todavia o CESPE deu a questão como correta! 😟 É verdade que algumas arquiteturas possam conviver com vários sistemas de armazenamento, mas isso não é um afirmação que pode ser generalizada para qualquer data lake. Falha da banca, na minha visão, deixar esse gabarito como correto.

Gabarito: Banca: Certo/Professor: Errado

CEBRASPE (CESPE) - 2022 - Especialista Técnico (BNB)/Analista de Sistemas/Infraestrutura e Segurança da Informação

A respeito de business intelligence, julgue o próximo item.

Data lake é um sistema que permite o armazenamento em uma base de dados os quais tenham sido refinados em processos anteriores.

Comentário: A afirmação está incorreta. Um Data Lake é um repositório de dados que armazena uma grande quantidade de dados brutos e não processados, independentemente de sua estrutura ou formato original. Ao contrário de um Data Warehouse, onde os dados são refinados e transformados antes de serem armazenados, um Data Lake mantém os dados em seu estado original, permitindo uma maior flexibilidade na análise e processamento posterior. Portanto, o Data Lake não requer que os dados sejam refinados em processos anteriores; ele armazena os dados brutos para análise futura.

Gabarito: Errado

FEPESE - 2022 - Analista de Informática (FAPESC)

Assinale a alternativa correta com relação ao assunto Data Lake.

A Os Data Lakes que se tornam inacessíveis para os usuários são chamados de “Datawarehouses”.

B Os Data Lakes exigem governança e manutenção contínuas para que os dados possam ser usados e acessados. Sem esse controle, há o risco de eles se tornarem lixo eletrônico – inacessíveis, pesados, caros e inúteis.

C Em um Data Lake, os dados são transformados apenas quando inseridos no banco de dados, por meio da aplicação de esquemas. Esse processo é chamado de “esquema para gravação” porque os dados são convertidos em estado apropriado para armazenamento.

D O Data Lake é um tipo de repositório que armazena conjuntos grandes e variados de dados processados para uma finalidade específica. Com os Data Lakes, você tem uma visão refinada dos dados.

E O Data Lake oferece um modelo de dados estruturados projetado para a geração de relatórios. Essa é a principal diferença entre ele e o data warehouse. Já o data warehouse armazena dados brutos não estruturados que não têm uma finalidade definida.

Comentário: A alternativa correta é a letra B: os Data Lakes exigem governança e manutenção contínuas para que os dados possam ser usados e acessados. Sem esse controle, há o risco de eles se tornarem lixo eletrônico – inacessíveis, pesados, caros e inúteis. Esta afirmativa destaca a importância da governança e manutenção dos Data Lakes para garantir que os dados armazenados permaneçam acessíveis, úteis e relevantes ao longo do tempo. Sem uma gestão adequada, os Data Lakes correm o risco de se tornarem ineficazes e caros para a organização.

Gabarito: B

CEBRASPE (CESPE) - 2022 - Atividades Técnicas de Complexidade Gerencial (MCom)/Tecnologia da Informação e de Engenharia Sênior

Julgue o item a seguir, a respeito de ETL, ELT e data lake.

Data lake é um tipo de repositório que armazena grandes volumes de dados, sob um esquema de banco de dados comum, unificado, visando responder perguntas específicas do negócio; esse sistema de armazenamento também oferece uma visão multidimensional dos dados atômicos e resumidos.

Comentário: O item refere-se a características que não são atribuídas ao Data Lake. A definição apresentada parece mais relacionada ao conceito de um Data Warehouse. Portanto, o item está incorreto.

O conceito de Data Lake não implica necessariamente em um esquema de banco de dados comum e unificado, nem oferece uma visão multidimensional dos dados. Em vez disso, um Data Lake é um repositório centralizado que armazena uma grande variedade de dados brutos, estruturados e não estruturados, em seu formato original, sem exigir um esquema de dados pré-definido. Esses dados podem ser usados para uma variedade de finalidades, incluindo análises avançadas e aprendizado de máquina. Portanto, o item está incorreto.

Gabarito: Errado

CEBRASPE (CESPE) - 2021 - Analista Judiciário (TJ RJ)/Tecnologia da Informação/Analista de Gestão de TIC

Construído(a) em arquitetura distribuída em grande escala, com capacidade de armazenar e processar conjuntos de dados não estruturados, a fim de agrega-los sobre clientes de diferentes fontes, enriquecê-los, limpá-los e analisa-los para entender melhor às jornadas dos clientes caracteriza um(a)

A virtualização de dados.

B storage de objetos.

C data lake.

D desktop como serviço (DaaS).

E software como serviço (SaaS).

Comentário: A descrição fornecida caracteriza um Data Lake (opção C). Um Data Lake é um repositório de dados que permite armazenar grandes volumes de dados brutos, estruturados e não estruturados, provenientes de diversas fontes, em sua forma original. Esses dados podem ser agregados, enriquecidos, limpos e analisados para obter insights valiosos e entender melhor o comportamento dos clientes, entre outras finalidades. Portanto, a opção correta é a letra C: data lake.

Gabarito: C

COMPERVE (UFRN) - 2020 - Analista de Suporte (TJ RN)/Pleno/Banco de Dados/Servidores Temporários de TIC

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

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

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

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

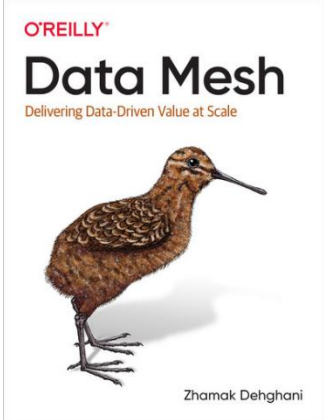
D é projetado para armazenar dados de diversas fontes e formatos, não havendo a necessidade da definição de um esquema de dados para inserir novos itens.

Comentário: A definição que melhor descreve um Data Lake é a opção D: é projetado para armazenar dados de diversas fontes e formatos, não havendo a necessidade da definição de um esquema de dados para inserir novos itens.

Um Data Lake é um repositório de dados que permite armazenar uma grande variedade de dados, tanto estruturados quanto não estruturados, em sua forma original, sem a necessidade de impor uma estrutura de dados rígida ou um esquema definido antecipadamente. Isso proporciona flexibilidade para lidar com diferentes tipos e formatos de dados, facilitando a integração de diversas fontes de dados.

Gabarito: D

DATA MESH



O conceito de Data Mesh surgiu como uma resposta aos desafios enfrentados pelas organizações na gestão de dados em larga escala. Em um mundo cada vez mais digitalizado, a quantidade e a complexidade dos dados têm crescido exponencialmente, levando à necessidade de abordagens inovadoras para lidar com essa avalanche de informações. O termo "Data Mesh" foi cunhado por Zhamak Dehghani em 2019, durante sua apresentação na conferência QCon, onde ela propôs uma nova abordagem para gerenciar dados em ambientes empresariais complexos e distribuídos.

A essência do Data Mesh pode ser descrita como um paradigma sociotécnico descentralizado, extraído da arquitetura distribuída moderna. Essa abordagem revoluciona a forma como as organizações obtêm, compartilham, acessam e gerenciam dados analíticos em larga escala. Em contraste com abordagens tradicionais, centradas em torno de monólitos de dados centralizados, o Data Mesh propõe uma estrutura distribuída na qual os dados são tratados como produtos, com responsabilidade e propriedade claramente definidas. Isso permite uma maior escalabilidade, agilidade e resiliência na gestão de dados, capacitando as organizações a extrair insights valiosos de seus recursos de dados.

MUDANÇAS TÉCNICAS E ORGANIZACIONAIS

A malha de dados introduz uma série de mudanças significativas em comparação com abordagens anteriores, como ilustrado na figura a seguir. Essas mudanças abrangem vários aspectos organizacionais, arquitetônicos, tecnológicos, operacionais, de valores e infraestruturais.

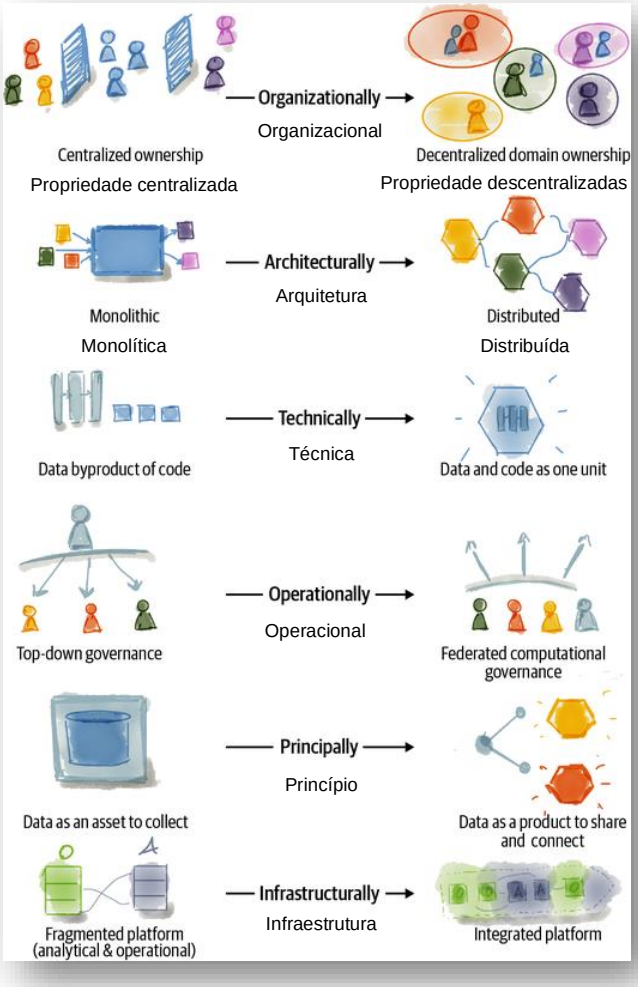
Organizacionalmente, ocorre uma transição da propriedade centralizada dos dados por especialistas que administram as tecnologias da plataforma de dados para um modelo **descentralizado de propriedade de dados**. Isso implica em transferir a propriedade e a responsabilidade dos dados de volta aos domínios de negócios onde os dados são produzidos ou usados.

Arquitetonicamente, a malha de dados deixa de coletar dados em armazéns e lagos monolíticos, passando **a conectar dados por meio de uma malha distribuída** de produtos de dados acessados por protocolos padronizados. **Tecnologicamente**, há uma mudança de soluções que tratam os dados como **um subproduto da execução de código de pipeline** para soluções que tratam **dados e códigos** que os mantêm como uma unidade autônoma e ativa.

Operacionalmente, a governança de dados muda de um sistema centralizado de cima para baixo, com intervenções humanas, para um **modelo federado com políticas computacionais embutidas nos nós da malha**. Em **termos de valores**, há uma transição

dos dados como um ativo a ser coletado para os **dados como um produto para servir e encantar os usuários**, tanto internos quanto externos à organização.

Por fim, **infraestruturalmente**, ocorre uma mudança de dois conjuntos de serviços de infraestrutura fragmentados e integrados ponto a ponto - um para dados e análises e outro para aplicações e operações - para **um conjunto bem integrado de infraestrutura para sistemas operacionais e de dados**.



Percebemos que a implementação da data mesh traz uma série de mudanças significativas na forma como as organizações lidam com seus dados. Ao descentralizar a propriedade e responsabilidade dos dados, promove-se uma maior integração entre os domínios de negócios onde os dados são produzidos ou utilizados. Isso não apenas permite uma governança mais eficaz e adaptável, mas também incentiva **uma cultura de colaboração e responsabilidade compartilhada em torno dos dados**.

Além disso, a transição para uma arquitetura distribuída de dados, conectando produtos de dados por meio de protocolos padronizados, facilita o acesso e a utilização dos dados em toda a organização. Essas mudanças não apenas moldam o futuro do trabalho com dados, mas também redefinem a maneira como as empresas abordam a inovação, a tomada de decisões e a criação de valor por meio da análise de dados. O trabalho com dados se torna

mais ágil, colaborativo e orientado para resultados, refletindo uma mudança fundamental na forma como as organizações percebem e alavancam seu ativo mais valioso: os dados.

PRINCÍPIOS DE DATA MESH

Data Mesh é uma abordagem revolucionária para a gestão de dados em larga escala, que busca superar os desafios tradicionais associados à centralização e monolitismo das infraestruturas de dados. No centro dessa abordagem estão quatro princípios fundamentais que orientam a maneira como os dados são gerenciados, compartilhados e utilizados em uma organização. Esses princípios representam uma mudança radical na forma como as empresas abordam o seu ecossistema de dados, promovendo uma maior agilidade, autonomia e eficiência em todo o ciclo de vida dos dados.

1. **Princípio de Propriedade de Domínio:** Reconhecendo a expertise única de cada equipe de negócios em relação aos dados que elas geram e consomem, este princípio atribui à propriedade dos dados aos domínios de negócios. Isso permite uma maior autonomia e agilidade nas operações de dados, capacitando as equipes de domínio a tomar decisões mais informadas e ágeis com base nos dados que conhecem melhor.
2. **Princípio de Dados como Produto:** Este princípio propõe uma mudança de mentalidade em relação aos dados, tratando-os como produtos de valor que são projetados, desenvolvidos e operacionalizados para atender às necessidades e requisitos dos usuários finais. Os dados são disponibilizados por meio de interfaces bem definidas e acessíveis, promovendo uma experiência de uso intuitiva e eficiente.
3. **Princípio da Plataforma de Dados de Autoatendimento:** Buscando capacitar as equipes de negócios a acessarem e gerenciarem os dados de forma independente, este princípio promove a implementação de uma plataforma de dados de autoatendimento. Essa plataforma fornece às equipes de domínio as ferramentas e recursos necessários para realizar tarefas de coleta, limpeza, transformação e análise de dados sem depender de conhecimento especializado em tecnologia.
4. **Princípio de Governança Computacional Federada:** Em contraste com a abordagem tradicional de governança de dados centralizada, este princípio propõe a distribuição da governança de dados entre os nós da malha de dados. As políticas e diretrizes de governança são incorporadas nos próprios produtos de dados, garantindo a conformidade e a segurança em toda a organização de forma transparente e responsável.

Juntos, esses quatro princípios formam a base da abordagem de Data Mesh, promovendo uma gestão mais eficaz, colaborativa e descentralizada dos dados em larga escala. Ao adotar esses princípios, as organizações podem transformar sua abordagem aos dados, promovendo uma cultura de dados mais ágil, inovadora e orientada para resultados. Vamos tentar expandir o assunto e apresentar mais características que permeiam cada um dos princípios.

Princípio de Propriedade de Domínio

Este princípio reconhece a importância da expertise de cada equipe de negócios em relação aos dados que eles geram e consomem. Cada equipe de domínio é responsável por todos os aspectos relacionados aos dados em seu domínio, desde a coleta e armazenamento até o processamento e disponibilização para uso.

Isso promove uma maior agilidade e autonomia, permitindo que as equipes de negócios atuem de forma independente na gestão e tomada de decisões relacionadas aos dados que conhecem melhor. Ao atribuir a propriedade dos dados aos domínios de negócios, cria-se uma estrutura organizacional mais flexível e adaptável, capaz de responder rapidamente às necessidades em constante mudança do mercado e dos clientes.

As motivações da propriedade de domínio são:

- A capacidade de expandir o compartilhamento de dados alinhado aos eixos de crescimento organizacional: aumento do número de fontes de dados, aumento do número de consumidores de dados e aumento da diversidade de casos de uso de dados
- Otimização para mudanças contínuas, localizando as mudanças nos domínios de negócios
- Permitindo agilidade reduzindo sincronizações entre equipes e removendo gargalos centralizados de equipes de dados, armazéns e arquitetura de lago
- Aumentar a veracidade dos negócios dos dados, eliminando a lacuna entre a origem real dos dados e onde e quando eles são usados para casos de uso analíticos
- Aumentar a resiliência das soluções de análise e aprendizado de máquina, removendo pipelines de dados intermediários complexos

Princípio de Dados como Produto

Esse princípio se baseia na ideia de tratar os dados como produtos, com uma abordagem centrada no usuário. Os dados são projetados, desenvolvidos e operacionalizados como produtos de valor, disponibilizados para consumo por meio de interfaces bem definidas e acessíveis, como APIs.

Os produtos de dados devem atender às necessidades e requisitos dos usuários finais, proporcionando uma experiência de uso intuitiva e eficiente. Isso promove uma mentalidade de produto em relação aos dados, incentivando a inovação, a qualidade e a excelência no design e entrega dos produtos de dados.

Os dados como produto aderem a um conjunto de características de usabilidade:

Um produto de dados fornece um conjunto de contratos de compartilhamento de dados explicitamente definidos e fáceis de usar. Cada produto de dados é autônomo e seu ciclo de vida e modelo são gerenciados independentemente dos demais.

Dados como um produto introduz uma nova unidade de arquitetura lógica chamada *data quantum*, controlando e encapsulando todos os componentes estruturais necessários para compartilhar dados como um produto – dados, metadados, código, política e declaração de dependências de infraestrutura – de forma autônoma.

As motivações dos dados como produto são:

- Eliminar a possibilidade de criar silos de dados orientados ao domínio, alterando o relacionamento das equipes com os dados. Os dados se tornam um produto que as equipes compartilham, em vez de coletar e armazenar em silos.
- Criar uma cultura de inovação orientada por dados, simplificando a experiência de descobrir e usar dados de alta qualidade, ponto a ponto, sem atrito.
- Criar resiliência às mudanças com isolamento integrado e em tempo de execução entre produtos de dados e contratos de compartilhamento de dados explicitamente definidos, para que a alteração de um não desestabilize outros.

Obtenha maior valor dos dados compartilhando e usando dados além das fronteiras organizacionais.

Princípio da Plataforma de Dados de Autoatendimento

Esse princípio busca capacitar as equipes de negócios a acessarem e gerenciarem os dados de forma independente, sem depender de uma equipe central de dados. Uma plataforma de dados de autoatendimento fornece às equipes de domínio as ferramentas e recursos necessários para realizar tarefas como coleta, limpeza, transformação e análise de dados sem a necessidade de conhecimento especializado em tecnologia.

Isso permite uma maior agilidade e eficiência nas operações de dados, reduzindo a sobrecarga sobre os profissionais de TI e acelerando o tempo de obtenção de insights e valor a partir dos dados.

A plataforma simplifica a experiência dos usuários de dados para descobrir, acessar e usar produtos de dados. Ele simplifica a experiência dos provedores de dados para criar, implantar e manter produtos de dados.

As motivações da plataforma de dados de autoatendimento são:

- Reduzir o custo total da propriedade descentralizada de dados.
- Abstrair a complexidade do gerenciamento de dados e reduza a carga cognitiva das equipes de domínio no gerenciamento do ciclo de vida ponta a ponta de seus produtos de dados.
- Mobilizar uma população maior de desenvolvedores – generalistas em tecnologia – para embarcar no desenvolvimento de produtos de dados e reduzir a necessidade de especialização.
- Automatizar políticas de governança para criar padrões de segurança e conformidade para todos os produtos de dados.

Princípio de Governança Computacional Federada

Este princípio diz respeito à distribuição da governança de dados entre os nós da malha de dados, em vez de ser centralizada em uma única entidade. As políticas e diretrizes de governança são incorporadas nos próprios produtos de dados, garantindo a conformidade e a segurança em toda a organização.

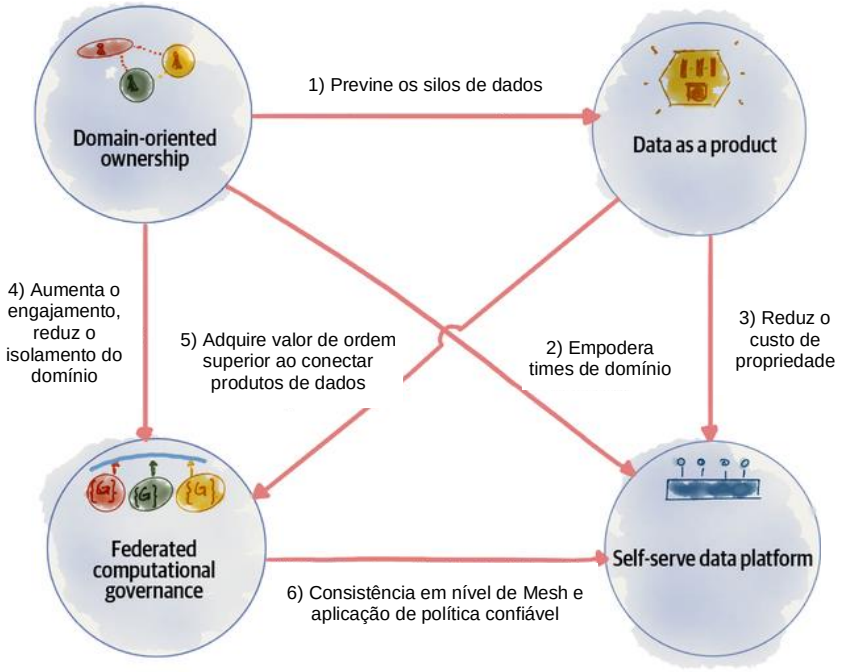
Isso permite uma maior flexibilidade e adaptabilidade na implementação e execução das políticas de governança, adaptando-as às necessidades e contextos específicos de cada domínio de negócios. Além disso, promove a transparência e responsabilidade no uso e gerenciamento dos dados, incentivando uma cultura de governança e conformidade em toda a organização.

O modelo operacional cria uma estrutura de incentivo e responsabilidade que equilibra a autonomia e agilidade dos domínios, com a interoperabilidade global da malha. O modelo de execução da governação depende fortemente da codificação e automatização das políticas a um nível minucioso, para cada produto de dados, através dos serviços da plataforma.

As motivações da governança computacional federada são:

- A capacidade de obter valor de ordem superior a partir da agregação e correlação de produtos de dados independentes, porém interoperáveis
- Combater as consequências indesejáveis das descentralizações orientadas para os domínios: incompatibilidade e desconexão de domínios
- Tornando viável a incorporação de requisitos de governança transversais, como segurança, privacidade, conformidade legal, etc., em uma malha de produtos de dados distribuídos
- Reduzindo a sobrecarga da sincronização manual entre domínios e a função de governança

Integração entre os princípios

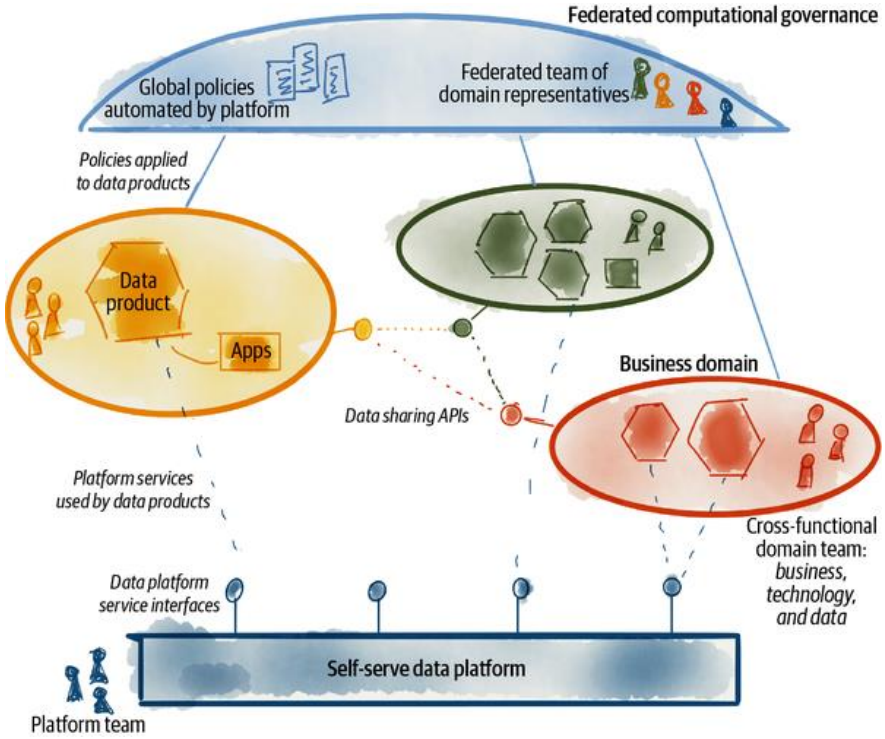


***A direção das setas mostra a dependência de um princípio em relação ao outro, ao implementar o princípio de origem cria-se o desafio que o princípio de destino vai endereçar.**

A integração entre os princípios de Data Mesh é fundamental para mitigar desafios como a formação de silos de dados e a duplicação de esforços. A figura acima tenta organizar essa integração de forma lógica. Por exemplo, a propriedade descentralizada dos dados orientada para o domínio pode levar à fragmentação dos dados dentro dos domínios, impedindo o compartilhamento eficiente entre diferentes partes da organização. Nesse cenário, o princípio de Dados como Produto entra em jogo, exigindo que os domínios tratem seus dados como produtos de alta qualidade, prontos para serem compartilhados com outros domínios. Isso não apenas incentiva a padronização e a qualidade dos dados, mas também promove uma cultura de compartilhamento colaborativo em toda a organização.

Da mesma forma, a propriedade do domínio dos produtos de dados pode levar à duplicação de esforços e aumentar os custos de propriedade. Para lidar com essa questão, a plataforma de dados de autoatendimento capacita as equipes multifuncionais do domínio a compartilharem e utilizar os produtos de dados de forma eficiente. Ao oferecer ferramentas e recursos que simplificam o acesso e o uso dos dados, a plataforma reduz a carga cognitiva das equipes de domínio, aumenta a produtividade e diminui o custo total de propriedade dos produtos de dados. Assim, a integração harmoniosa entre os princípios de Data Mesh garante uma gestão eficaz e colaborativa dos dados em toda a organização, promovendo a inovação e o crescimento sustentável.

VISÃO GERAL DO MODELO DE MALHA DE DADOS



A malha de dados é um modelo operacional inovador que revoluciona a maneira como os dados são gerenciados, compartilhados e utilizados em organizações modernas. Este modelo se baseia em uma série de princípios fundamentais para promover a descentralização, a colaboração e a eficiência no tratamento de dados. Vamos explorar mais a fundo os elementos-chave desse modelo.

Domínios com Equipes Multifuncionais: A malha de dados parte do princípio de que cada domínio de negócio possui equipes multifuncionais dedicadas a alcançar as metas específicas do domínio. Essas equipes têm a responsabilidade de produzir e utilizar aplicativos digitais e produtos de dados para impulsionar o crescimento e a inovação dentro da organização.

Compartilhamento de Dados e Serviços: Dentro da malha de dados, os domínios compartilham seus dados e serviços por meio de contratos bem definidos. Isso permite uma colaboração mais eficaz entre os diferentes setores da organização, facilitando a troca de informações e o desenvolvimento de soluções integradas.

Produtos de Dados Compostos: Os produtos de dados na malha de dados podem ser compostos e pertencer a vários domínios simultaneamente. Isso significa que os dados são tratados como ativos reutilizáveis, que podem ser combinados e adaptados para atender às necessidades específicas de diferentes partes da organização.

Políticas Globais e Governança Federada: As políticas globais na malha de dados são definidas por uma entidade federada composta por representantes de diversos domínios. Essas políticas, juntamente com outros serviços de plataforma, são oferecidas como recursos automatizados para garantir a conformidade, a segurança e a eficácia do modelo de malha de dados como um todo.

Em resumo, o modelo de malha de dados representa uma abordagem inovadora e colaborativa para o gerenciamento de dados em organizações modernas. Ao promover a descentralização, o compartilhamento e a reutilização de dados, esse modelo capacita as empresas a maximizarem o valor de seus ativos de dados, impulsionar a inovação e alcançar o sucesso em um ambiente empresarial cada vez mais digitalizado.

Antes de passar para o próximo tópico vamos fazer algumas questões sobre o assunto:

QUESTÕES

Questão inédita/2024.

O que é Data Mesh em ciência de dados?

a) Uma técnica para limpar dados inconsistentes em um conjunto de dados.

b) Uma abordagem para centralizar todos os dados em um único data lake.

c) Um framework para descentralizar o armazenamento e gerenciamento de dados em organizações.

d) Um modelo de machine learning para análise de séries temporais.

e) Uma técnica de visualização de dados em forma de malha.

Comentário: Data Mesh é uma abordagem para descentralizar o gerenciamento de dados, permitindo que cada equipe ou domínio de negócios seja responsável por seus próprios dados.

Gabarito: C

Questão inédita/2024.

Julgue as afirmativas a seguir:

[1] A abordagem Data Mesh é uma solução centralizada para gerenciamento de dados em ambientes complexos e de grande escala.

[2] O Data Mesh promove a centralização do gerenciamento de dados, facilitando o controle por uma única equipe de TI.

[3] Um dos princípios fundamentais do Data Mesh é tratar os dados como um subproduto da execução do código de pipeline.

[4] O Data Mesh é classificado como uma arquitetura no contexto de gestão de dados.

[5] O princípio de governança computacional federada no Data Mesh é implementado através de políticas de governança centralizadas com intervenções humanas.

Comentário: [1] Data Mesh é uma abordagem descentralizada para o gerenciamento de dados, oposta à abordagem centralizada.

[2] O Data Mesh descentraliza o gerenciamento de dados, empurrando a responsabilidade de volta para os domínios de negócios.

[3] O princípio fundamental do Data Mesh é tratar os dados e o código que os mantém como uma unidade autônoma e viva.

[4] O Data Mesh é classificado como um paradigma sociotécnico, reconhecendo as interações entre pessoas e a arquitetura técnica em organizações complexas.

[5] O princípio de governança computacional federada no Data Mesh é implementado através de políticas de governança codificadas e automatizadas, sem intervenções humanas centralizadas.

Gabarito: [1] Errado [2] Errado [3] Errado [4] Errado [5] Errado

Questão inédita/2024.

Julgue as afirmativas a seguir:

[1] O Data Mesh promove a descentralização do gerenciamento de dados, empurrando a responsabilidade de volta para os domínios de negócios.

Comentário: Certo, no Data Mesh, a propriedade dos dados é descentralizada, com a responsabilidade sendo empurrada de volta para os domínios de negócios.

Gabarito: Certo

MODELAGEM MULTIDIMENSIONAL

Um modelo dimensional contém **as mesmas informações que um modelo normalizado**. Essa é uma frase importante para começar a explicar o assunto. Foi dita pelo Kimball e pode ser usada em questões de prova ... se ligue!! E o que muda nos modelos dimensionais? Eles vão organizar os dados com um propósito diferente. Geralmente as ferramentas analíticas possuem as seguintes preocupações: **facilidade de compreensão ao usuário**, **desempenho da consulta** e **resiliência às mudanças**.

Figura 31 - Preocupações do modelo dimensional

Imagine um executivo que descreve o seu negócio como, "Nós vendemos produtos em vários mercados e medimos o nosso desempenho ao longo do tempo." Projetistas multidimensionais devem ouvir atentamente a ênfase no **produto**, **mercado** e **tempo**.

A modelagem multidimensional, ou dimensional como às vezes é chamada, é a técnica de modelagem de banco de dados para o auxílio às **consultas em um *Data Warehouse*** nas mais **diferentes perspectivas**. A visão multidimensional permite o uso **mais intuitivo** para o processamento analítico pelas ferramentas OLAP (*On-line Analytical Processing*).

Toda modelagem dimensional possui dois elementos imprescindíveis: **Fatos** e **Dimensões**. Ambos **são obrigatórios** e possuem características complementares dentro de um *Data Warehouse*. As **Dimensões** são os **descritores** dos dados oriundos dos Fatos. Possui o **caráter qualitativo** da informação. É a Dimensão que permite a visualização das informações por diversos aspectos e perspectivas.

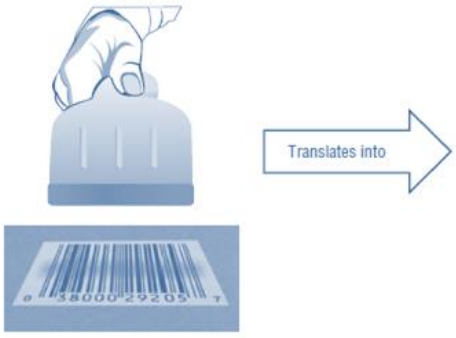
Os fatos servem para o armazenamento dos registros e medidas (quase sempre) numéricas associadas a eventos de negócio. Perceba que até aqui não falei em tabela, mas por quê? Fatos e dimensões são elementos genéricos que existem nos modelos dimensionais, mas estamos

acostumados a falar da implementação ou estruturação relacional deles. Neste contexto, aparecem os conceitos tabela fato e tabelas dimensões ...

As dimensões possuem **um relacionamento de "um para muitos" com as tabelas fatos**. Ou seja, cada linha da tabela dimensão ligada diretamente a tabela fato pode estar associada a várias linhas da tabela dimensão.

Uma **tabela fato** armazena as **medições de desempenho** decorrentes de eventos dos processos de negócios de uma organização. Basicamente, representa uma medida de negócios. Uma tabela fato contém **vários fatos**, correspondentes a cada uma das suas **linhas**. Cada linha corresponde a um evento de medição. Os dados em cada linha estão a um nível específico de detalhe, referido como o **grão**, por exemplo, uma linha por produto vendido em determinada loja em um dia específico.

Uma única linha da tabela fato tem uma relação um-para-um com o evento de medição, como descrito pela granularidade da tabela fato. Veja que estamos falando de um evento de medição, ou seja, algo que aconteceu e precisa ser armazenado para análise posterior. **Uma tabela fato corresponde a um evento físico observável, e não às exigências de um relatório específico**. Dentro de uma tabela fatos, apenas fatos consistentes com a granularidade definida são permitidos. Por exemplo, em uma transação de vendas no varejo, a quantidade de um produto vendido e seu preço são bons fatos. Veja a figura abaixo que representa uma tabela fato de vendas:



Fato_Vendas
•Data_Key (FK) •Produto_Key (FK) •Loja_Key (FK) •Promocao_Key (FK) •Trasaction # •ValorVendido •QuantidadeVendida

Figura 32 - Tabela fato de vendas

A ideia de que um evento de medição no mundo físico tenha uma relação de um-para-um com uma única linha na tabela fato é um princípio fundamental para a modelagem dimensional. Todo o resto da teoria se constrói a partir deste conceito. Você vai ver que os fatos são, por vezes aditivos, semi-aditivos ou mesmo não-aditivos. Antes de falarmos sobre os tipos de medidas vamos fazer uma questão sobre o assunto:

(Ministério da Economia – Especialista em Ciência de Dados - 2020) Julgue os itens a seguir, relativos a conceitos de modelagem dimensional.

Uma tabela de fatos registra dados dimensionais que explicam os fatos registrados.

_____

Comentários: As tabelas dimensões guardam informações de contexto. A tabela fato registra os eventos e as medidas a eles associadas.

Gabarito: ERRADO.

Tipos de fatos ou medidas

Fatos aditivos podem ser agrupados em qualquer uma das dimensões. Fatos semi-aditivos, tais como saldos de conta, não podem ser resumidos por meio da dimensão de tempo. Fatos não-aditivos, tais como preços unitários, taxas e percentuais, nunca podem ser adicionados ou somados. É teoricamente possível que um fato medido seja textual, no entanto, essa condição raramente aparece.

Aditivos	Semi-aditivos	Não aditivos
• Podem ser agrupadas em uma qualquer das dimensões associadas à tabela de fatos • Os fatos mais flexíveis e úteis. • Lucro líquido	• Podem ser agrupadas em algumas dimensões, **mas não todas**. • Ex.: Saldo em conta - não podem ser resumidos por meio da dimensão de tempo	• Nunca podem ser adicionados ou somados • Taxas e percentuais

Figura 33 - Tipos de medidas presentes nas tabelas fatos

Perceba que alguns atributos, mesmo sendo numéricos, não fazem sentido quando agregados ou somados. Outro ponto é que algumas dimensões não numéricas podem ser eventualmente um fato. Nestes casos as informações textuais só permitem contagem e estatísticas associadas a quantidade de eventos com a mesma descrição.

Na maioria dos casos, uma medição textual é uma descrição de algo, e é traçada a partir de uma lista de valores discretos. O designer deve fazer todos os esforços para colocar os dados textuais em dimensões onde podem ser correlacionados de forma mais eficaz com os outros atributos de dimensão e consumir menos espaço em disco.

Você não deve armazenar informações textuais redundantes em tabelas de fatos. A menos que o texto seja exclusivo para cada linha na tabela de fato, ele deve pertencer a uma tabela de dimensão. Um fato “texto verdadeiramente exclusivo” é raro, porque se o conteúdo do fato for imprevisível, como um texto de comentário de forma livre, torna quase impossível de se analisar.

Todas as tabelas fato têm duas ou mais chaves estrangeiras que ligam para as chaves primárias das tabelas de dimensão. Por exemplo, a chave do produto na tabela coincide com o fato de sempre uma chave de produto específico na tabela de dimensão de produto. Quando todas as chaves na tabela fato corretamente coincidirem com suas respectivas chaves primárias das tabelas dimensão correspondentes, as tabelas satisfazem a integridade referencial. Você pode acessar a tabela de fatos através das tabelas de dimensões por meio de *join*.

As tabelas dimensões apresentam o contexto descritivo. São companheiros integrais para uma tabela de fatos e contém o contexto textual associado a um evento de medição dos processos de negócios. Elas descrevem o "quem, o que, onde, quando, como e por que" associado ao evento.

As dimensões contêm os atributos descritivos usados pelas aplicações de BI para filtrar e agrupar os fatos. Com a granularidade de uma tabela de fato definida, todas as possíveis dimensões podem ser identificadas. Sempre que possível, a dimensão deve ter um valor único associado a uma determinada linha da tabela fato. Tabelas dimensão são chamadas a "alma" do DW, pois elas contêm os pontos de entrada e rótulos descritivos que permitem ao sistema de DW/BI ser aproveitado para a análise de negócios.

Um esforço é necessário para o desenvolvimento e o gerenciamento das tabelas de dimensão, pois elas são os condutores da experiência de BI do usuário. Dimensões fornecem os pontos de entrada para os dados, e as etiquetas finais e os agrupamentos em todas as análises de DW/BI. Vejam um exemplo de uma tabela dimensão abaixo.

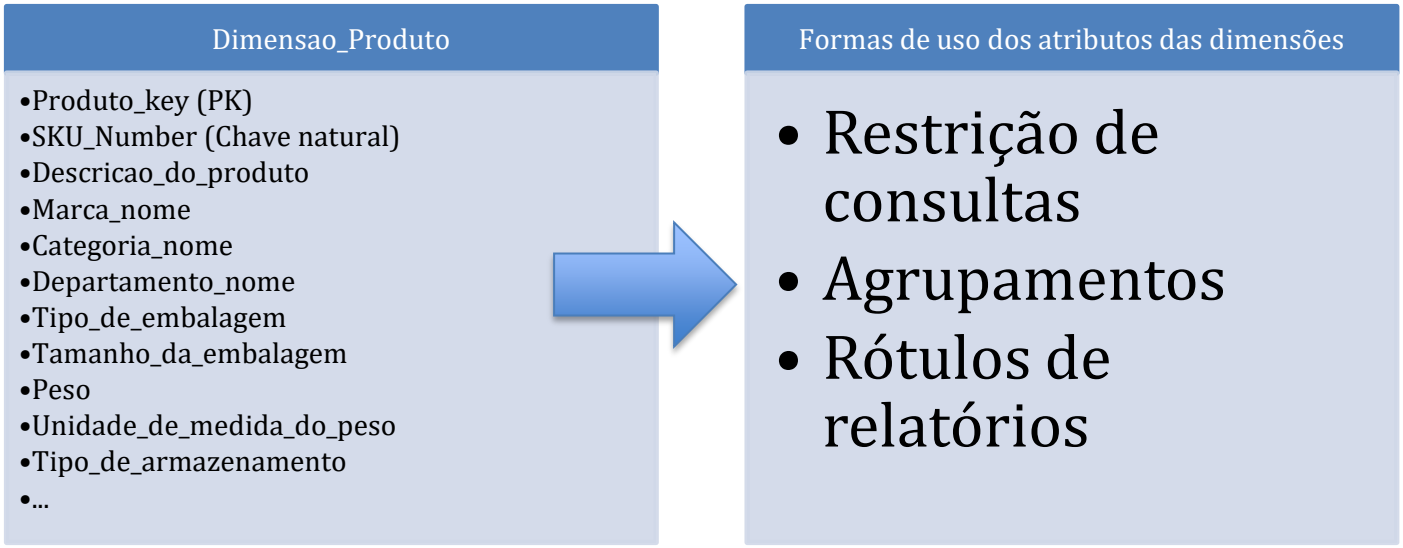


Figura 34 - Exemplo de dimensão produto

Ok! Antes de continuarmos nosso estudo e juntarmos as tabelas dimensões e fatos vamos refletir sobre a seguinte pergunta: Uma quantidade numérica é um fato ou um atributo de dimensão? Geralmente a resposta a esse questionamento segue a seguinte lógica: observações numéricas continuamente valoradas são quase sempre medidas da tabela fato; e observações numéricas discretas e extraídas de uma lista pequena, quase sempre são atributos de dimensão.

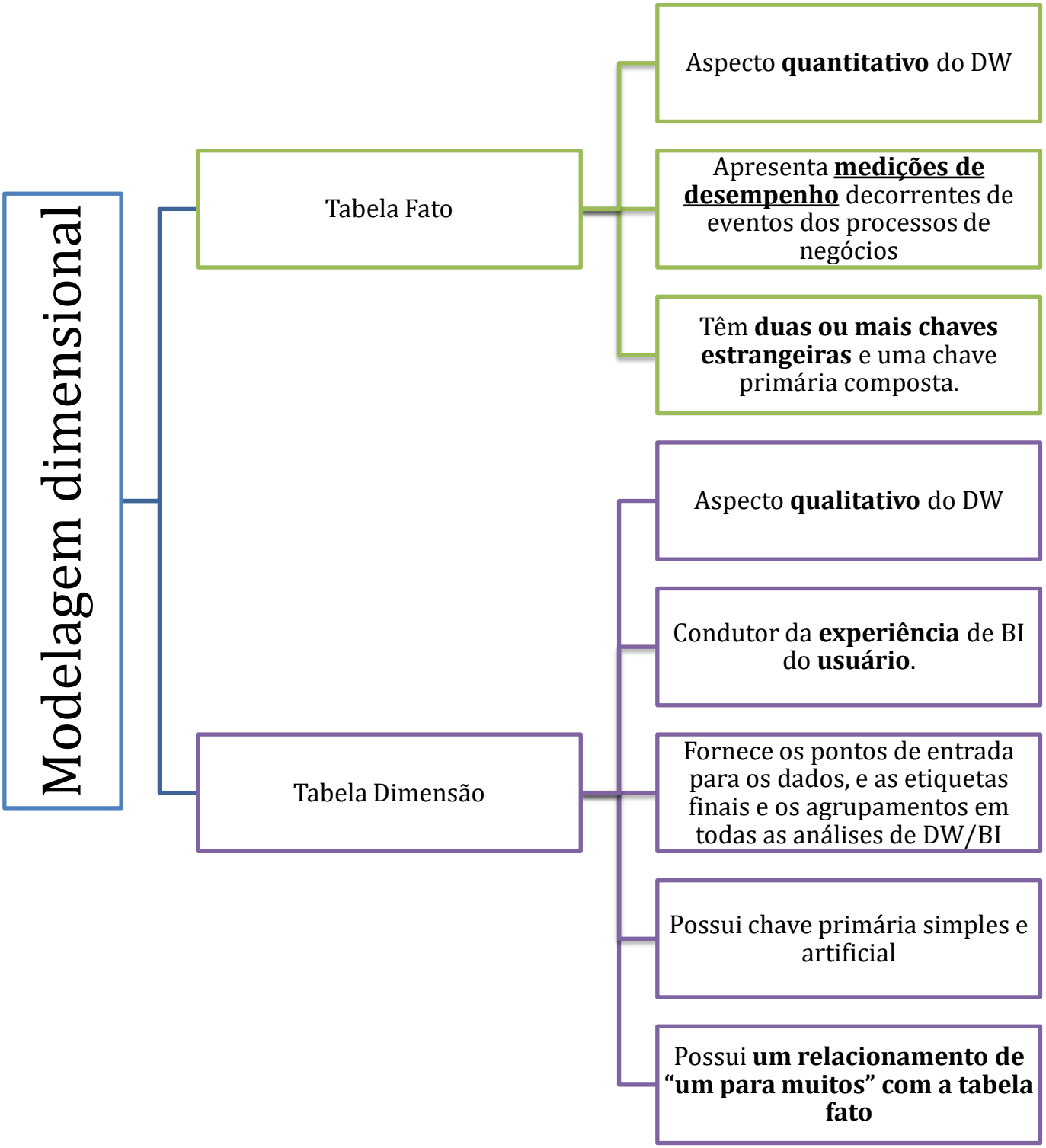
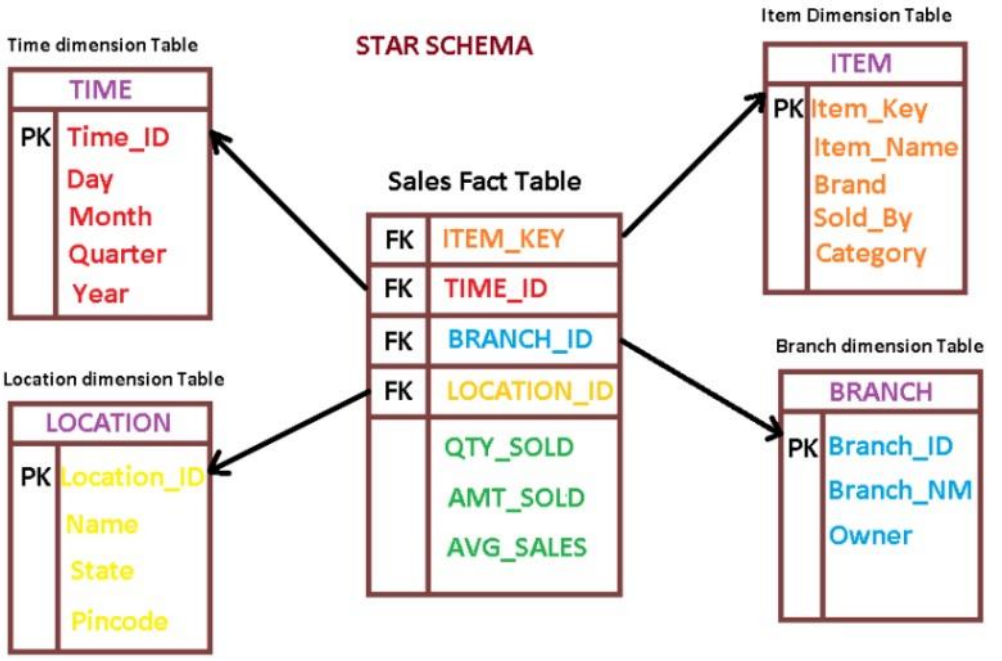


Figura 35 - Esquema das tabelas fato e dimensões

ESQUEMAS MULTIDIMENSIONAIS

Star Schema

Cada processo de negócio é representado por um modelo dimensional que consiste em uma tabela fato contendo medições numéricas do evento e, cercada por um conjunto de tabelas dimensão que contêm o contexto quando ocorreu o evento. Esta característica de estrutura estrela (*star schema*) é muitas vezes chamada de junção estrela (*star-join*), um termo que remonta aos primórdios de bancos de dados relacionais. Vejamos um exemplo de um modelo:



Chave **primária** da tabela fato:

Composta pelos atributos que "apontam" para as chaves primárias das tabelas dimensão.

No exemplo, temos: **ITEM_KEY, BRANCH_ID, Location_ID e Time_ID**

Figura 36 - Modelo estrela (star schema)

A primeira coisa a notar sobre o esquema dimensional é a sua simplicidade e simetria. Obviamente, os usuários de negócios se beneficiam da simplicidade, pois os dados são mais fáceis de compreender e navegar. O charme do desenho na figura acima é que é altamente reconhecível para usuários corporativos. Segundo Kimball: "Temos observado, literalmente, centenas de casos em que os usuários imediatamente concordam que o modelo dimensional é o seu negócio".

Além disso, a redução do número de tabelas e da utilização de descrições gerenciais torna-o mais fácil de navegar e menos suscetível a erros. A simplicidade de um modelo multidimensional também tem benefícios de desempenho. Otimizadores de banco de dados processam esses esquemas de forma mais simples, com menos joins e, portanto, de forma mais eficiente.

Um motor de banco de dados pode fazer suposições fortes sobre a primeira restrição das tabelas de dimensão que são fortemente indexadas e, em seguida, atacar a tabela de fatos de uma só vez com o produto cartesiano das chaves das tabelas dimensões que satisfazem as restrições do usuário. Surpreendentemente, usando essa abordagem, o otimizador pode avaliar arbitrariamente uma junção para uma tabela fato em uma única passagem através do índice da tabela fato.

Finalmente, os modelos dimensionais são graciosamente extensíveis para acomodar a mudança. O framework previsível de um modelo dimensional resiste a mudanças inesperadas no comportamento do usuário.

Antes de continuarmos vamos tentar resumir os pontos associados aos conceitos presentes no modelo estrela.

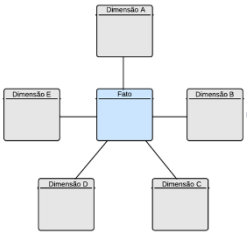
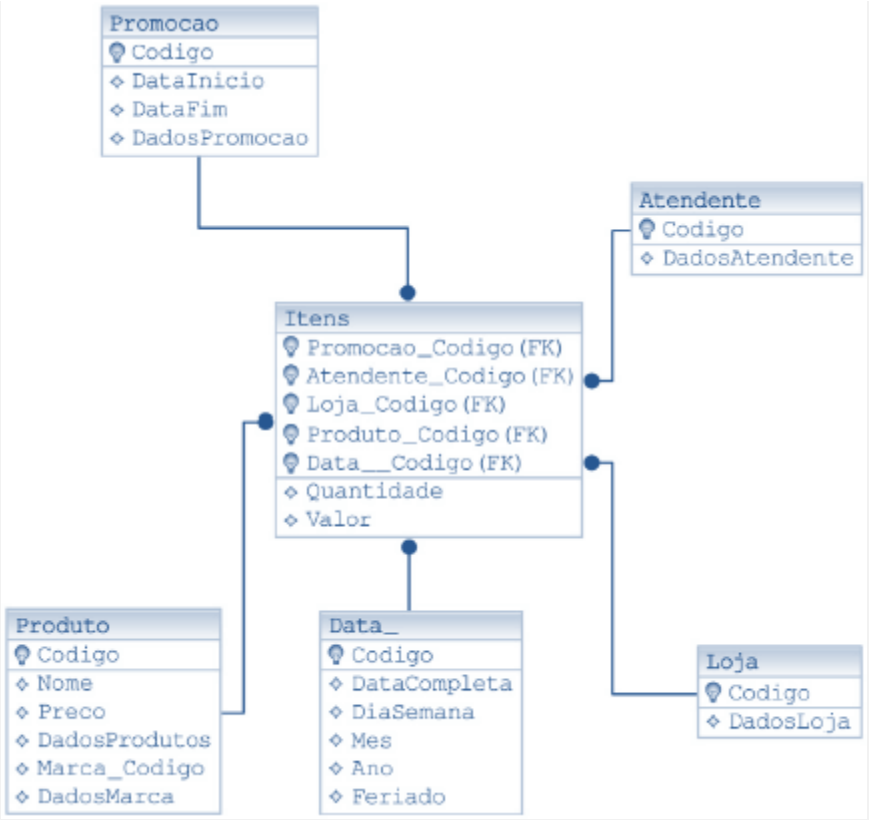
Modelo estrela	
	Uma tabela fato central e dimensões respresentadas em apenas 1 tabela.
	Todas as **dimensões** são **ligadas diretamente** a tabela fato.
	As tabelas dimensões são **desnormalizadas.**
	Mais fácil de ser entendido pelos usuários finais. (consultas mais simples)
	Performance melhor do que o modelo snowflake

Figura 37 - Resumo das características do modelo estrela

(CEBRASPE (CESPE) - Auditor de Finanças e Controle de Arrecadação da Fazenda Estadual (SEFAZ-AL)/2020) Com relação a banco de dados, julgue o item seguinte.

Considerando-se o modelo multidimensional a seguir, é correto afirmar que Quantidade e Valor são dimensões de Itens.



Comentário: Perceba que a tabela central fato, denominada Itens, possui 5 dimensões associadas a ela. Observe que Quantidade e Valor são 2 atributos da tabela fato classificados como medidas. Logo, temos uma alternativa **incorreta**.

Gabarito: Errado.

Para finalizar, vamos apresentar de forma gráfica os esquemas conhecidos como *snowflake* e multiestrela (ou constelação). Já falamos sobre o modelo estrela. Você deve se lembrar dele com uma tabela fato no meio e várias tabelas dimensões ao redor.

Esquema constelação ou multiestrela

O esquema estrela da figura acima representa **um único processo de negócio** para o rastreamento de vendas. Outros esquemas podem ser necessários para outros processos, como **entregas e compras**. No caso de processos de negócio relacionados que **compartilhem algumas das tabelas dimensionais**, o esquema estrela pode ser estendido para um **esquema constelação**, com vários tipos de entidades de fatos, como mostra a figura a seguir.

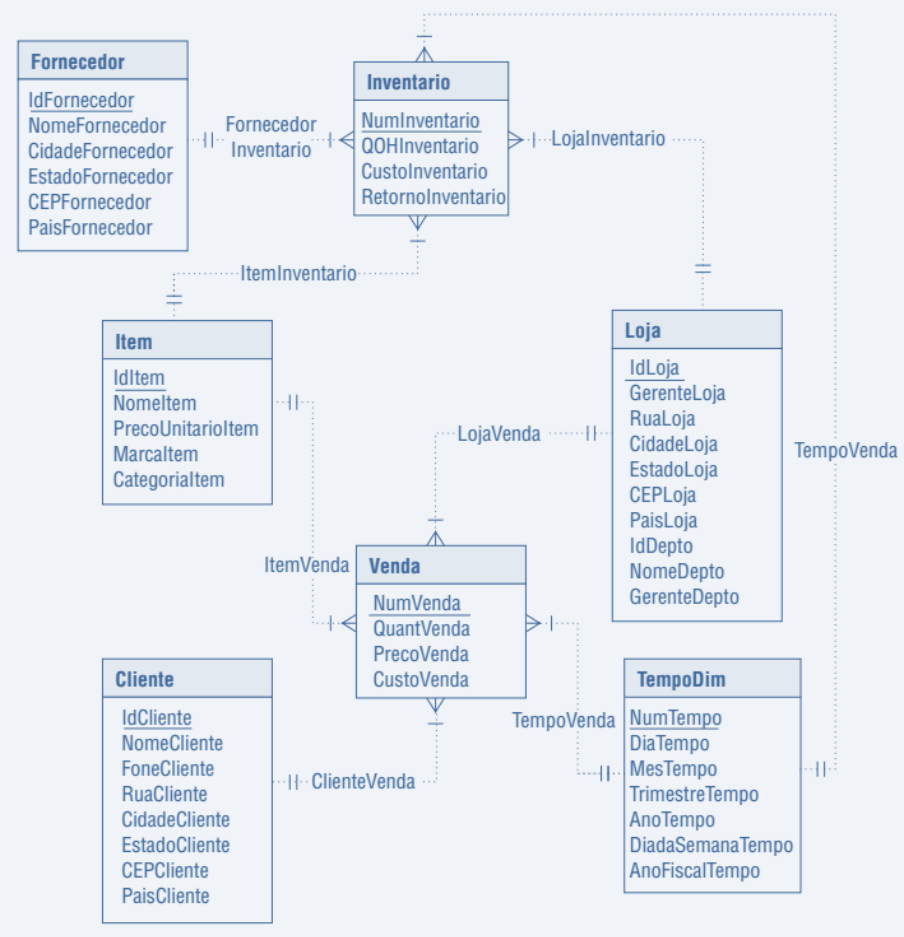


Figura 38 - Constelação com 2 tabelas fato: vendas e inventário.

Veja que **Inventário se torna uma tabela de fatos** e os relacionamentos 1-N tornam-se chaves estrangeiras na tabela de fatos. O tipo de entidade Inventário adiciona inúmeras medidas, incluindo a quantidade ofertada de um item, o custo desse item e a quantidade devolvida. **Todas as tabelas dimensionais são compartilhadas entre ambas as tabelas de fatos, exceto as tabelas Fornecedor e Cliente**. A essas tabelas compartilhadas dá-se o nome de **dimensões conformes.** Uma desvantagem que podemos observar no modelo multiestrela é que ele acaba limitando as consultas feitas sobre o data warehouse.

Esquema floco de neve

Neste tipo de modelo de dados algumas dimensões podem ser normalizadas até a terceira forma normal (3FN). Essa normalização cria uma hierarquia entre as tabelas dimensões, fazendo com que algumas tabelas não estejam conectadas diretamente à tabela fato. Esse modelo reduz o espaço de armazenamento dos dados, pois a normalização elimina parte da redundância do modelo. Contudo, esse benefício não melhora a performance e acaba criando um modelo mais complexo para o usuário final. Essa complexidade é refletida em consultas mais complexas e difíceis de entender.

Em geral não é necessário normalizar as tabelas dimensão para impedir anomalias no armazenamento, porque normalmente elas são estáveis e pequenas. A natureza de um data warehouse indica que as **tabelas de dimensão devem ser projetadas para recuperação, não para atualização**. O desempenho associado à recuperação de dados é melhorado eliminando-se as

operações de junção que seriam necessárias para combinar tabelas dimensionais totalmente normalizadas.

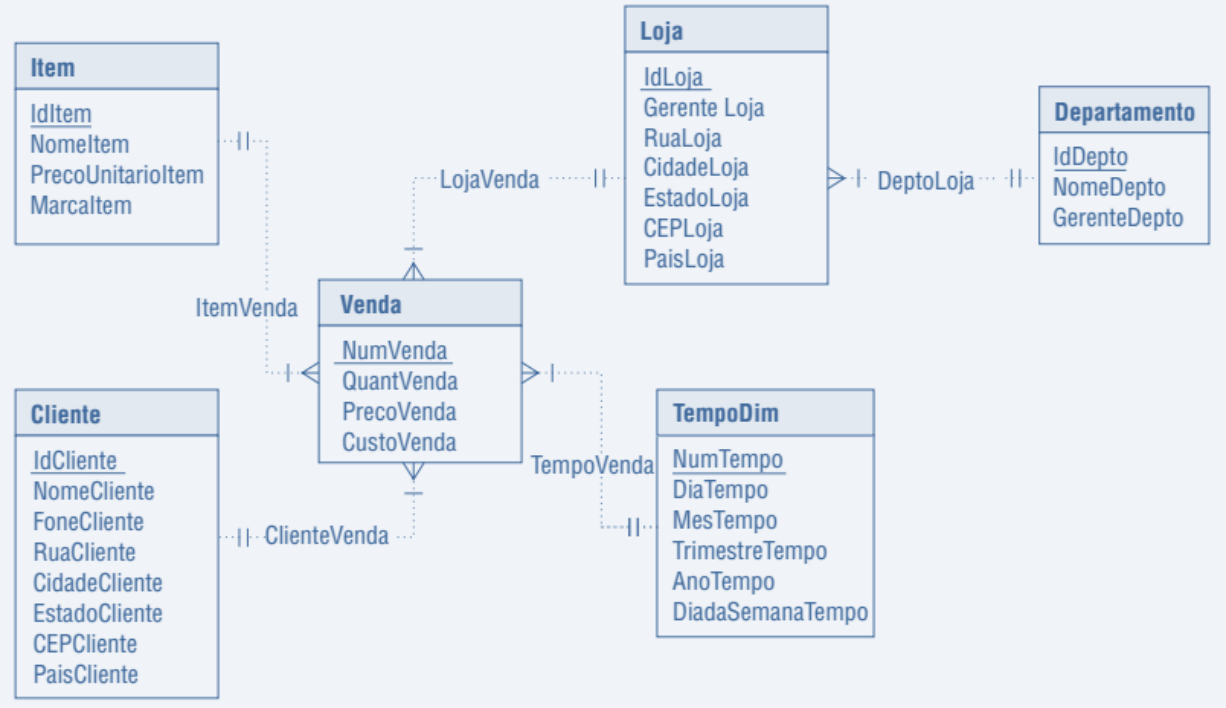


Figura 39 - Esquema floco de neve para vendas

Floco de neve	Dimensões normalizaadas (pelo menos uma) até a 3FN.
	Apresenta hieraquia nas dimensões.
	Algumas dimensões não estão ligadas diretamente a tabela fato.
	Modelo mais complexo, dificulta o entendimento por usuários finais.
	Ocupa menos espaço de armazenamento.
	Consultas mais complexas e lentas quando comparadas com o modelo estrela.

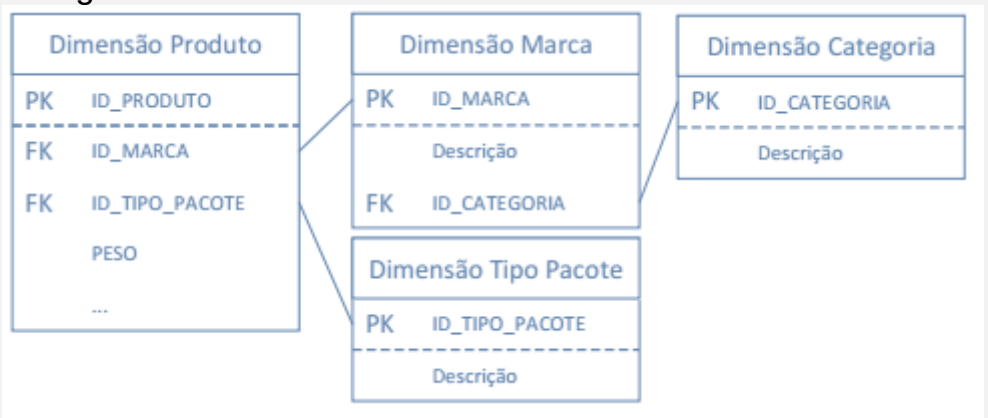
Figura 40 - Resumo das características do modelo floco de neve

Resumido de forma rápida os esquemas que estudamos até agora temos:

Figura 41 - Resumo dos esquemas multidimensionais

Vamos fazer uma questão para fixação do conteúdo:

(Ano: 2017 Banca: FGV Órgão: Alerj Cargo: Analista de Tecnologia da Informação Q. 43)
Observe o seguinte Modelo Multidimensional de Dados.



A técnica de modelagem multidimensional utilizada para normalizar a dimensão, movendo os campos de baixa cardinalidade para tabelas separadas e ligadas à tabela original através de chaves artificiais, é:
(A) Slowly Changing Dimension;
(B) Conformed Dimension;
(C) Degenerated Dimension;
(D) Snowflaked Dimension;
(E) Role-Playing Dimension.

Comentário: A questão apresenta o conceito de modelo de dados floco de neve de forma um pouco mais elaborada. Contudo, você pode perceber que a característica determinante para esse tipo de modelagem multidimensional que é ter **dimensões normalizadas** está presente, tanto na figura quanto na descrição.

Gabarito: D.

PROCESSO DE DESIGN DIMENSIONAL

O processo de desenvolvimento ou projeto de sistemas multidimensionais segue basicamente quatro etapas: selecionar o processo de negócio, definir a granularidade, identificar as dimensões e identificar os fatos. Esse processo é alimentado por informações a respeito dos requisitos de negócio e da realidade dos dados. Vamos então passar pelas etapas e entender o que deve ser realizado em cada uma delas.

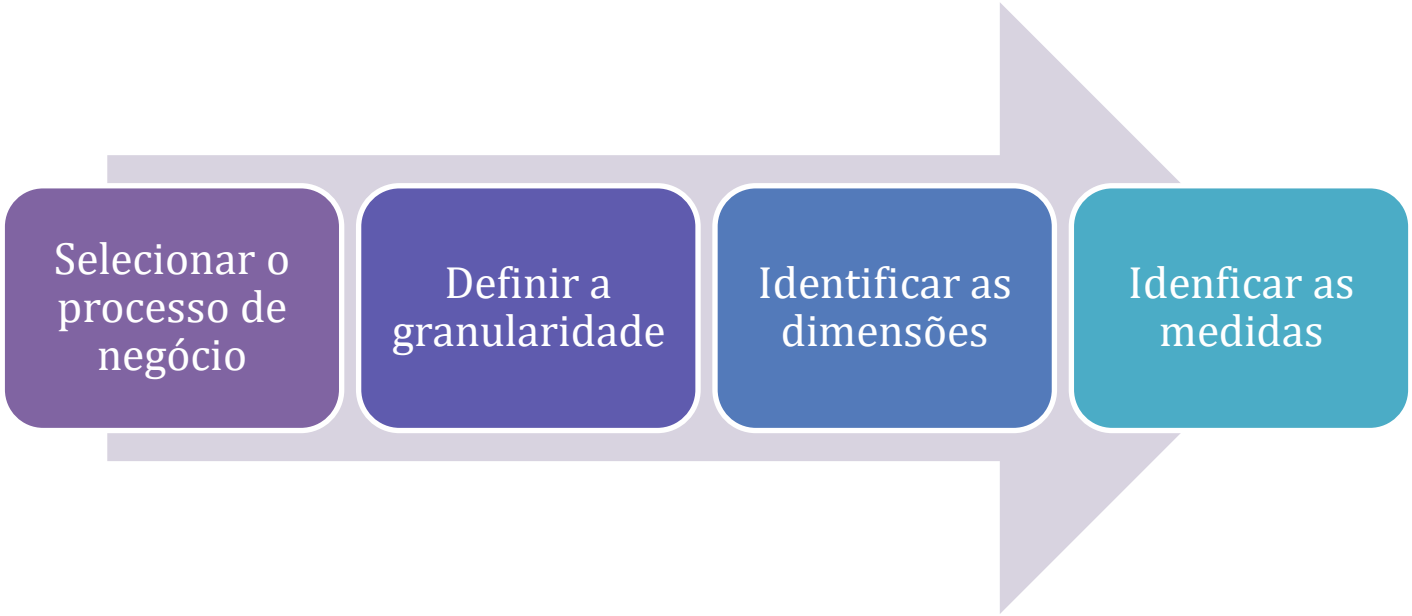


Figura 42 - Processo de design dimensional

Passo 01: Selecionando o processo de negócio

O primeiro passo é decidir o processo de negócio para o modelo, combinando a compreensão dos requisitos de negócio com a análise dos dados de origem disponíveis. O primeiro projeto de BI/DW deve se concentrar no processo de negócio que é o mais crítico para os usuários de negócios

A viabilidade do projeto abrange uma série de considerações, incluindo a disponibilidade e qualidade dos dados, e ainda a capacidade organizacional. Suponha que os gestores querem entender melhor as compras dos clientes, e o processo de negócio que você está modelando é baseado nas transações de vendas de varejo. Esta informação permite que os usuários de negócios analisem, por exemplo, quais produtos estão vendendo em quais lojas, em quais dias e em quais condições promocionais.

Passo 02: Declare a granularidade

Declarar a granularidade é um passo fundamental em um projeto dimensional. Estabelece exatamente o que uma única linha da tabela fato representa. Torna-se um contrato vinculativo sobre o design. Deve ser declarada antes de escolher dimensões ou fatos, porque cada dimensão candidata ou fato relevante deve ser consistente com a granularidade.

A atomicidade do grão refere-se ao nível mais baixo no qual os dados são capturados por um determinado processo de negócio. Kimball encoraja a começar pelos dados de grãos atômicos, porque resiste ao ataque de consultas de usuários imprevisíveis. A sumarização realizada pela

operação de roll-up é importante para o ajuste de desempenho, mas ela pressupõe perguntas comuns do negócio. Cada proposta de granularidade da tabela fato resultada em uma tabela física separada, diferentes granularidades não devem ser misturadas na mesma tabela fato.

Mas qual o nível de detalhe dos dados que devem ser disponibilizados no modelo dimensional? Você deve desenvolver modelos dimensionais que representem as informações, capturadas por um processo de negócio, **no nível mais detalhado ou atômico possível**. Um sistema de DW/BI exige quase sempre os dados expressos no mais baixo nível de granularidade. Não porque as consultas querem ver linhas individuais, mas porque as consultas precisam cortar os detalhes de formas muito precisas.

Passo 03: Identificando as dimensões

Após a granularidade da tabela fato ter sido escolhida, a escolha das dimensões é simples. A declaração cuidadosa da granularidade determina a dimensão primária da tabela fato. Em seguida, adiciona-se mais dimensões para a tabela fato, se essas dimensões adicionais tiverem apenas um valor para cada combinação das dimensões principais.

Se a dimensão adicional viola a granularidade, fazendo com que linhas adicionais da tabela fato sejam geradas, a dimensão precisa ser desclassificada ou a granularidade precisa ser revista. Algumas dimensões descritivas geralmente se aplicam a um modelo de loja: data, produto, loja, promoção, caixa, e forma de pagamento. Além disso, o número do bilhete da transação é incluído como uma dimensão especial, dita dimensão degenerada para números de transação.

Mas o que seria uma dimensão degenerada? Uma chave de dimensão, como o número de uma transação, número de fatura ou de ticket que não tenha nenhum atributo associado, portanto não se constitui com uma tabela de dimensão. Ela aparece apenas como uma das colunas da tabela fato.

Passo 04: Identificando os fatos

A quarta e última etapa do projeto é a determinação cuidadosa dos fatos que aparecerão na tabela fato. Mais uma vez, a declaração de granularidade ajuda a ancorar o raciocínio. Ao considerar os fatos em potencial, você pode descobrir novamente ajustes que precisam ser feitos tanto na granularidade quanto na escolha das dimensões.

Os dados coletados pelo sistema incluem, por exemplo, a quantidade de vendas por unidade regular, o desconto, preços líquidos pagos, valores de vendas em dólares. O valor de vendas é igual a quantidade de vendas multiplicado pelo preço unitário. Da mesma forma, o valor do desconto é a quantidade de vendas multiplicada pelo valor do desconto unitário.

Vejamos uma questão sobre o assunto:

(VUNESP - Analista de Sistemas (CM Piracicaba)/2019) Em um modelo dimensional de dados, definir a granularidade significa definir o
a) nível de detalhamento dos dados a serem inseridos nesse modelo.
b) número máximo de tabelas dimensão a ser suportado pelo modelo.
c) número máximo de usuários suportados pelo sistema.
d) tamanho dos registros suportados pelas tabelas dimensão.

e) tipo de relacionamento a ser estabelecido entre as tabelas dimensão e as tabelas fato.

_____

Comentários: A granularidade estabelece exatamente o que uma única linha da tabela fato representa. Ela vai estabelecer o nível mais baixo de detalhamento possível para as consultas no cubo de dados. A partir da definição, podemos marcar nossa resposta na alternativa A.

Gabarito: A.

REVISITANDO O MODELO

Vamos agora revisitar as estruturas das tabelas fato e dimensões para entender alguns conceitos relacionados com cada uma delas. Primeiramente começaremos pelas tabelas fatos.

Uma tabela fato contém as medidas numéricas produzidas por um evento de medição operacional no mundo real. No nível mais baixo de granularidade, uma linha da tabela fato **corresponde a um evento de medição** (fato) e vice-versa.

Além das medidas numéricas, uma tabela fato sempre contém **chaves estrangeiras** para cada uma das suas **dimensões** associadas, bem como as chaves de dimensão degenerados opcionais. As tabelas fatos são o principal alvo de computações e agregações dinâmicas decorrentes das consultas.

As medidas numéricas em uma tabela de fatos podem ser divididas em três categorias. Os fatos mais flexíveis e úteis são **totalmente aditivos**. Medidas aditivas podem ser agrupadas com qualquer das dimensões associadas à tabela de fatos. Medidas **semi-aditivas** podem ser agrupadas em algumas dimensões, mas não todas. Valores de saldo são fatos comuns semi-aditivos, porque eles são aditivos em todas as dimensões, exceto tempo. Finalmente, algumas medidas são completamente **não-aditivas**, como índices.

Medidas com **valores nulos se comportam normalmente em tabelas de fatos**. As funções de agregação (SUM, COUNT, MIN, MAX e AVG) todas fazem a "coisa certa" com fatos nulos. No entanto, **os valores nulos devem ser evitados em chaves estrangeiras da tabela de fatos**, porque esses valores nulos iriam provocar automaticamente uma violação de integridade referencial. Ao invés de uma chave estrangeira nula, a tabela de dimensão associado **deve ter uma linha padrão** (com uma *surrogate key*) que representa a condição desconhecida ou não aplicável.

Se a **mesma medida aparece em tabelas fatos separados**, os cuidados devem ser tomados para garantir que as definições técnicas dos fatos são idênticas, se elas podem ser comparadas ou calculadas juntas. Se as definições de fato separadas são consistentes, os fatos devem ser **conformados com nomes idênticos**, mas se eles são incompatíveis, devem ser nomeados de forma diferente para alertar os usuários de negócios e aplicações de BI.

TIPOS DE TABELA FATO

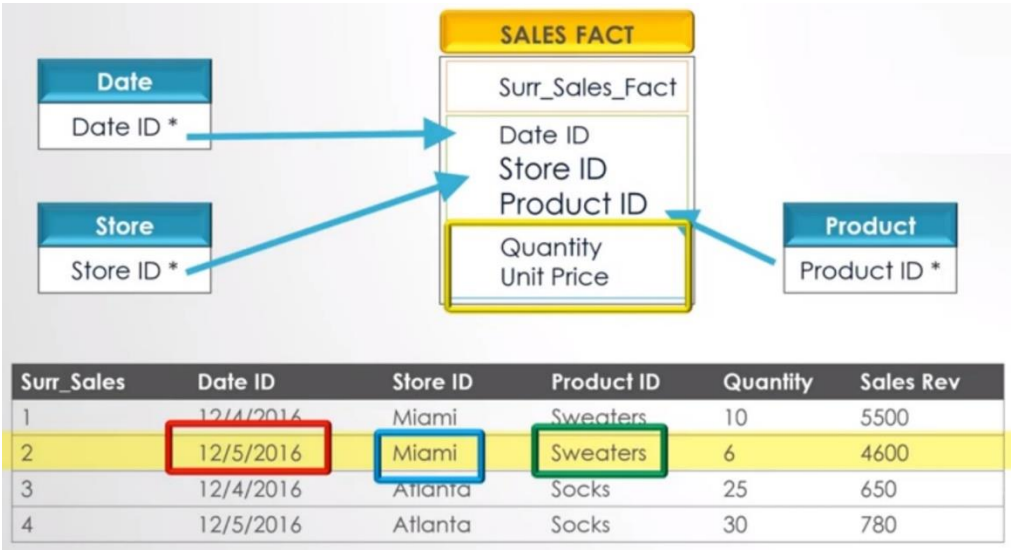
Vamos agora falar dos tipos de tabelas fatos. Existem na literatura seis tipos de fatos:

1. Fato transacional

2. Fato agregada
3. Fato consolidada
4. Fato snapshot periódico
5. Fato de snapshot acumulado
6. Fato sem fato

Tabela Fato Transacional

A linha em uma tabela de fatos de transações corresponde a **um evento de medição em um ponto no espaço e no tempo**. As tabelas fato de transações de **granularidade atômicas são as tabelas mais expressivas e dimensionais**. Essa dimensionalidade robusta permite o máximo de **operações de *slice and dice* sobre os dados**. Segundo o especialista Rafael Piton, a maioria dos bilhões de linhas que temos no Data Warehouse são de tabelas fato transacional. Elas geralmente utilizam métricas aditivas, aquelas métricas que a gente pode somar por todas as dimensões. A figura abaixo apresenta um exemplo de tabela fato transacional que armazena informações sobre vendas (sales fact).



Surr_Sales	Date ID	Store ID	Product ID	Quantity	Sales Rev
1	12/4/2016	Miami	Sweaters	10	5500
2	12/5/2016	Miami	Sweaters	6	4600
3	12/4/2016	Atlanta	Socks	25	650
4	12/5/2016	Atlanta	Socks	30	780

Figura 43 - Exemplo de tabela fato transacional

Tabela Fato Agregada

A tabela fato agregada tem **a função de acelerar o desempenho das consultas**. Ela sumariza os dados de outra tabela fato. Geralmente é construída para **armazenar o resultado de consultas agregadas muito utilizadas**. Por exemplo, se você consulta todos os dias o valor de vendas nos meses anteriores esse pode ser um bom dado para ser armazenado em uma tabela agregada. Assim, a consulta a essa informação será bem mais eficiente. Entretanto, temos que perceber que existe um esforço adicional para construção e manutenção dos dados nesta tabela.

Tabela Fato Consolidada

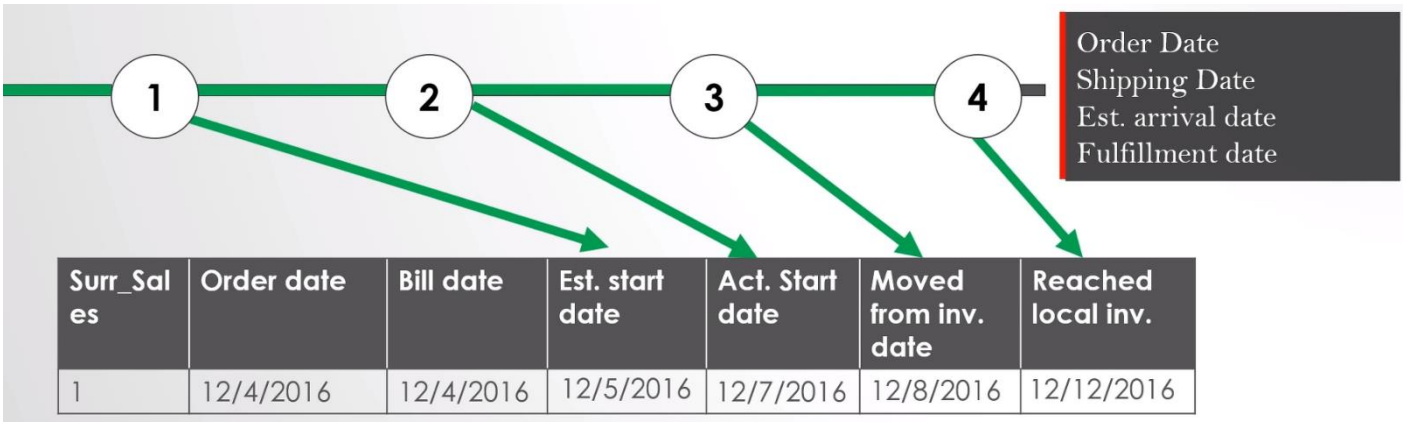
Muitas vezes é conveniente **combinar fatos de vários processos em uma única tabela fato consolidada** caso possam ser expressos na mesma granularidade. Por exemplo, os valores reais de vendas podem ser consolidados com as previsões de vendas em uma única tabela fato para tornar a tarefa de analisar os valores reais contra previsões simples e rápida.

Tabela Fato de Snapshot Periódico

A linha em **uma tabela fato de snapshot periódico** resume muitos eventos de medição ou estados atuais que ocorrem ao longo de **um período normal**, como um dia, uma semana ou um mês. **A granularidade é o período**, e não a transação individual. **Snapshots periódicos** muitas vezes contêm muitos fatos porque qualquer evento de medição de acordo com a granularidade da tabela de fatos é permitido. Estas tabelas de fatos são **uniformemente densas** em suas chaves estrangeiras, porque mesmo que nenhuma atividade ocorra durante o período, uma linha é tipicamente inserida na tabela fato contendo um zero ou o valor nulo para cada fato. Um uso interessante para esse tipo de tabela fato é para controle de estoque, fazendo uma fotografia, de tempos em tempos da quantidade de produtos ou mercadorias armazenadas.

Tabela Fato Snapshot Acumulado

Uma linha numa **tabela de *snapshot* de fatos acumulados** resume os eventos que ocorrem em etapas de medição previsíveis entre o início e o fim de um processo, pipeline ou fluxo de trabalho, tais como cumprimento de ordem ou processamento de pedidos, que têm um ponto definido de início, etapas intermediárias padrões, e um ponto final definido podem ser modelados com este tipo de tabela de fatos. Existe uma chave estrangeira data na tabela original para cada etapa crítica no processo. Uma linha individual em uma tabela de snapshot acumulado, correspondente, por exemplo, a uma ordem ou pedido de um determinado produto. Veja a figura abaixo.



Surr_Sales	Order date	Bill date	Est. start date	Act. Start date	Moved from inv. date	Reached local inv.
1	12/4/2016	12/4/2016	12/5/2016	12/7/2016	12/8/2016	12/12/2016

Tabela Fato sem Fato

As tabelas **fato sem fato** também podem ser usadas para analisar o que não aconteceu. Essas consultas têm sempre duas partes: uma tabela fato de cobertura, que contém todas as possibilidades de eventos que podem acontecer, e uma tabela fato de atividade, que contém os eventos que aconteceram. Quando a atividade é subtraída da cobertura, o resultado é o conjunto de acontecimentos que não aconteceram.

A seguir apresentamos uma figura que mostra a composição básica de uma tabela fato:

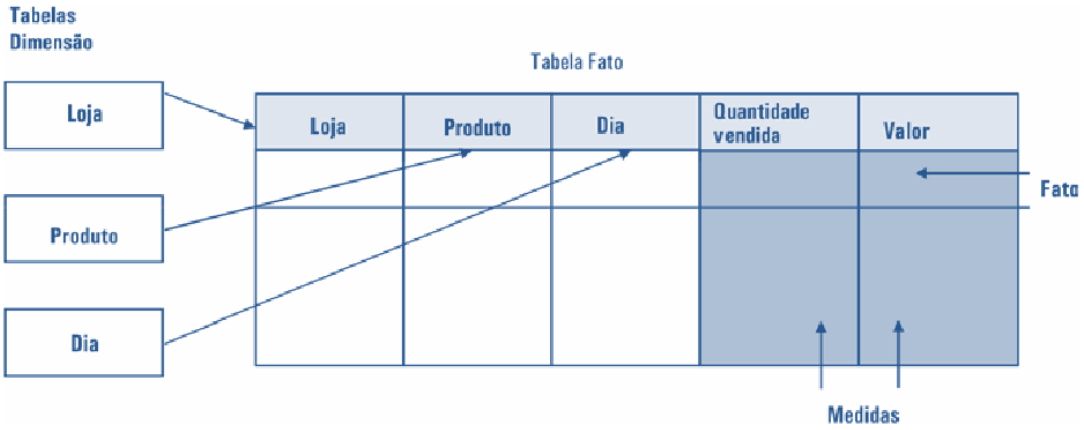


Figura 44 - Estrutura de uma tabela fato

Agora vamos fazer algumas questões para ajudar na fixação do assunto.

(VUNESP - Curso de Formação de Oficiais do Quadro Complementar (EsFCEx)/Informática/2020/CA CFO-QC 2021) Na modelagem de Armazéns de Dados (Data Warehouse), uma tabela de fato sem fato (factless fact tabla) é uma tabela que registra
a) os dados consolidados de outras tabelas de fato.
b) a intersecção entre dimensões, mas não possui métricas de medição.
c) os cálculos realizados sobre as métricas, como contagem, soma e média.
d) as métricas de um processo.
e) as estruturas que permitem categorizar os fatos e as métricas de medição.

_____

Comentários: Vamos analisar cada uma das alternativas e associas as respectivas definições:

a) Errada. A definição da alternativa refere-se a uma tabela fato consolidada.
b) Certa. Conforme descrevemos as tabelas fato sem fato funcionam como uma espécie de chamada, verificando se alguns eventos aconteceram. Esses eventos podem ser comparados com o domínio de eventos possíveis.
c) Errado. Tal definição está associada a funções de agregação que podem ser usadas na sumarização dos dados.
d) Errado. As métricas de um processo são as medidas da tabela fato.
e) Errado. As tabelas fatos são categorizadas ou contextualizadas pelo conteúdo das tabelas dimensões.

Gabarito: B.

(COVEST-COPSET - Analista (UFPE)/Tecnologia da Informação/2019) A tabela de fato de um Data Warehouse sobre estoque deve ser do tipo:
a) Instantâneo Transacional.
b) Instantâneo Analítico.
c) Instantâneo Periódico.
d) Instantâneo Acumulado.
e) Instantâneo Fragmentado.

_____

Comentário: De acordo com o Ralph Kimball, os modelos dimensionais para estoques são divididos em 3 tipos e são **complementares** (podem ser usados juntos):

- **Instantâneo periódico (snapshot periódico)**: Medição periódica dos níveis de estoque de cada produto e inserção desses dados em linhas separadas em uma tabela de fatos;
- **Transacional:** Registro de cada transação que afeta os níveis de estoque quando os produtos são colocados no warehouse;
- **Instantâneo acumulativo**: Construção de uma linha de fatos para cada entrega do produto e atualização da linha até o produto sair do warehouse, perceba que essa linha pode apresentar datas ou valores que representam valores acumulados.

Gabarito: C

(CESGRANRIO - 2008 - PETROBRÁS - ANALISTA DE SISTEMAS JÚNIOR - PROCESSOS DE NEGÓCIOS)

A empresa passou a sortear cupons de desconto para alguns clientes, os quais dão direito a um desconto nas compras em uma determinada data. A informação sobre que clientes possuem cupons para que datas é mantida de forma independente e consolidada no processo de extração, transformação e carga, resultando em um campo, na tabela fato, indicando se a venda foi realizada com o desconto ou não. A solução parecia atender bem às demandas dos usuários do data warehouse, até que um deles tentou realizar uma consulta para saber quais clientes não haviam realizado compras, mesmo tendo um cupom de desconto para a data. Este tipo de demanda tipicamente será resolvido introduzindo, no data warehouse, uma

(a) tabela de fatos complementares (complimentary fact table).
(b) tabela de fatos sem dimensão (dimensionless fact table).
(c) tabela de fatos sem fatos (factless fact table).
(d) dimensão multivalorada (multivalued dimension).
(e) dimensão degenerada (degenerated dimension).

Comentários: Vamos analisar cada uma das alternativas:

Tabela de fatos - Num esquema estrela, a tabela central com medidas numéricas de desempenho caracterizadas por uma chave composta, cada um dos elementos é uma chave estrangeira trazida de uma tabela de dimensão.

Tabela de fatos sem fatos (factless fact table) - Uma tabela de fatos que não tem fatos, mas captura alguns relacionamentos muitos-para-muitos entre as chaves de dimensões. Mais frequentemente usada para representar eventos ou prover informação de cobertura que não aparece em outras tabelas de fatos.

Dimensão degenerada (degenerated dimension) - Uma chave de dimensão, como o número de uma transação, número de fatura, de ticket, ou de bill-of-lading, que não tenha nenhum atributo associado, portanto não se constitui com uma tabela de dimensão.

Dimensão multivalorada (multivalued dimension) - Normalmente, uma tabela de fatos possui conexões somente para dimensões representando um valor simples, como uma data ou produto. Mas ocasionalmente, é válido conectar um registro de fato a uma dimensão representando um número aberto de valores, como o número de diagnósticos simultâneos que um paciente pode ter num momento de um mesmo tratamento. Neste caso, dizemos que a tabela de fatos tem uma dimensão multivalorada. Tipicamente manipulada por uma tabela ponte.

Baseado no exposto, podemos concluir que a resposta se encontra na alternativa C.

Gabarito: C

Vamos agora apresentar algumas características e taxonomia para as **tabelas de dimensões.**

Cada tabela de dimensão tem uma única coluna de chave primária. Esta chave primária é incorporada como uma chave estrangeira em qualquer tabela de fatos onde a descrição textual presente na linha da dimensão é exatamente a correta para a linha da tabela de fatos. Tabelas de dimensão são geralmente grandes, desnormalizadas, com muitos atributos de texto de baixa cardinalidade.

Embora os códigos operacionais e os indicadores possam ser tratados como atributos, os atributos de dimensão mais poderosos são preenchidos com descrições verbais. Os atributos da tabela de dimensão são o principal alvo de especificações de restrições e agrupamento de consultas em aplicações de BI. Os rótulos descritivos sobre os relatórios são tipicamente valores de domínio do atributo da dimensão.

A tabela de dimensão é projetada com uma coluna que funciona como uma chave primária única. Esta chave primária não pode ser a chave natural do sistema operacional, pois haverá várias linhas de dimensão para a chave natural quando as alterações forem feitas ao longo do tempo. Além disso, chaves naturais para uma dimensão podem ser criadas por mais do que um sistema, e estas chaves naturais podem ser incompatíveis ou mal administradas.

O sistema de DW/BI precisa reivindicar o controle das chaves primárias de todas as dimensões, ao invés de usar chaves naturais explícitas ou chaves naturais com datas concatenadas, você deve criar as chaves primárias inteiras (numéricas) anônimas para cada dimensão. Essas chaves são conhecidas como chaves artificiais.

Estas chaves substitutas para dimensão são números inteiros simples, atribuídos em sequência, começando com o valor um, a cada vez que uma nova chave é necessária. A dimensão data é isenta da regra fundamental da chave substituta. Esta dimensão altamente previsível e estável pode usar uma chave primária mais significativa.

Chaves naturais criadas pelos sistemas operacionais de origem estão sujeitos às regras de negócios fora do controle do sistema de DW/BI. Por exemplo, um número de funcionário (chave natural) pode ser alterado se o empregado se demite e depois é recontratado. Quando o armazém de dados quer ter uma chave única para o empregado, uma nova **chave durável** deve ser criada que seja persistente e não se altere nesta situação.

Esta chave é muitas vezes referida como uma **chave sobrenatural durável**. As melhores chaves duráveis têm um formato que é independente do processo de negócio original e, portanto, devem ser inteiros simples atribuídos em sequência começando de um. Enquanto várias **chaves de substituição** podem ser associadas com um funcionário ao longo do tempo com as suas alterações do perfil, a chave durável nunca muda.

Às vezes uma dimensão é definida e não tem conteúdo, exceto a sua chave primária. Por exemplo, quando uma nota fiscal tem vários itens de linha, cada linha da tabela fato herda dados de todas as dimensões descritivas por meio das chaves estrangeiras da nota fiscal, e nota fica, portanto, sem conteúdo exclusivo. Mas o número de fatura continua a ser uma chave de dimensão válida para as tabelas fatos no nível de item de linha. Esta dimensão é degenerada e colocada na tabela fato, com o reconhecimento explícito que não há tabela de dimensão associada. Dimensões degeneradas são mais comuns com tabelas fato de snapshots e tabelas fato de acumulação.

Em geral, os designers de modelos dimensionais devem resistir à normalização, causada por anos de projetos de banco de dados operacionais, e desnormalizar as hierarquias de profundidade fixa em atributos separados em uma linha de dimensão achatada. A desnormalização da dimensão apoia os objetivos individuais de modelagem dimensional de simplicidade e velocidade.

Muitas dimensões contêm mais de uma hierarquia natural. Por exemplo, a dimensão data de calendário pode ter dias úteis por semana na hierarquia de período fiscal, assim como uma hierarquia para dia, mês e ano. Dimensões de localização podem ter várias hierarquias geográficas. Em todos estes casos, as hierarquias separadas podem graciosamente coexistir na mesma tabela dimensão.

Abreviaturas, flags de verdadeiro/falso, e indicadores operacionais devem ser complementados nas tabelas de dimensão com palavras de texto completo que têm significado quando vistos de forma independente. Códigos operacionais com significado embutido no valor do código devem ser divididos e cada parte do código deve possuir uma dimensão descritiva separada para seu próprio atributo (significado). Imagine o código 101.2012.1342.23-1 nele cada subparte descrever uma característica do produto.

Valores nulos nos atributos da dimensão são resultado de uma determinada linha de dimensão que não foi totalmente preenchida, ou quando existem atributos que não são aplicáveis a todas as linhas da dimensão. Em ambos os casos, recomenda-se a substituição por uma sequência descritiva, como “desconhecido” ou “não se aplica” no lugar do valor nulo. Nulos em atributos de dimensão devem ser evitados, porque lidar com bancos de dados diferentes, agrupando e restringindo os nulos, é, muitas vezes, inconsistente.

Dimensões calendário ou data estão ligadas a praticamente todas as tabelas de fatos para permitir a navegação da tabela de fatos através de datas familiares, meses, períodos fiscais, e dias especiais no calendário. Você nunca iria querer calcular o feriado de páscoa usando SQL, mas sim querer procurá-lo na dimensão data do calendário. A dimensão data do calendário normalmente tem muitos atributos que descrevem as características tais como número da semana, o nome do mês, período fiscal, e um indicador de feriado nacional. Para facilitar o particionamento, a chave primária de uma dimensão de data pode ser mais significativa, como um inteiro representado por AAAAMMDD, em vez de uma chave substituta sequencialmente atribuída.

Uma única dimensão física pode ser referenciada várias vezes em uma tabela de fato, com cada referência ligando para um papel logicamente distinto para a dimensão. Por exemplo, uma tabela de dados pode ter várias datas, cada uma delas representada por uma chave estrangeira para a dimensão de data. É essencial que cada chave estrangeira se refira a uma visão separada da dimensão data de modo que as referências sejam independentes. Estas dimensões separadas (com nomes exclusivos de colunas de atributo) são chamadas de papéis.

Processos de negócios transacionais normalmente produzem números variados e uma baixa cardinalidade de bandeiras (flags) e indicadores. Ao invés de fazer dimensões diferentes para cada bandeira (flag) e atributo, você pode criar uma única **dimensão junk** e combiná-las. Esta dimensão, muitas vezes, é rotulada como uma dimensão de perfil de transação em um esquema, não precisa ser o produto cartesiano de todos os valores possíveis dos atributos, mas deve conter apenas a combinação de valores que ocorrem realmente nos dados de origem.

Quando uma relação hierárquica em uma tabela de dimensão é normalizada, os atributos de baixa cardinalidade aparecem como tabelas secundárias ligadas à tabela de dimensão base, por uma chave de atributo. Quando este processo é repetido com todas as hierarquias da tabela de dimensão, uma estrutura característica com vários níveis é criada. Chamamos esse novo **modelo de floco de neve**.

Embora o floco de neve represente dados hierárquicos com precisão, você deve evitar os flocos de neve, porque é difícil para os usuários de negócios compreenderem e navegarem sobre os dados. Eles também podem afetar negativamente o desempenho da consulta. A tabela de dimensão desnormalizada contém exatamente a mesma informação como uma ***Dimensão Snowflaked***.

A dimensão pode conter uma referência para outra tabela dimensão. Por exemplo, uma dimensão conta bancária pode fazer referência a uma dimensão separada representando a data em que a conta foi aberta. Estas referências dimensão secundária são chamadas ***Dimensões Outrigger***. Dimensões *outrigger* são permitidas, mas devem ser usadas com moderação. Na maioria dos casos, as correlações entre as dimensões devem ser rebaixadas para uma tabela fato, quando as duas dimensões são representadas como chaves estrangeiras separadas.

Dimensões grandes e minidimensões.

As tabelas de dimensão geralmente são pequenas quando comparadas com as tabelas fatos. Contudo, existe situações em que as mudanças nas dimensões são muito frequentes e precisam ser registradas na tabela.

O Kimball analisa essa questão do crescimento das dimensões (large dimensions) da seguinte forma: *O que acontece quando a taxa de mudança acelera, especialmente em uma grande tabela* ***de dimensões de vários milhões de linhas?***

As grandes dimensões apresentam ***dois desafios que requerem tratamento especial.***

1. *O tamanho dessas dimensões pode afetar negativamente o desempenho da navegação e do filtro de consultas.*
2. *Além disso, a técnica comprovada do SCD tipo 2 (cria-se uma linha para capturar a alteração nas dimensões) para rastreamento de alterações não é atraente, porque não queremos adicionar mais linhas a uma dimensão que já tenha milhões de linhas, principalmente se as alterações ocorrerem com frequência.*

Felizmente, uma única técnica vem ao resgate para abordar o desempenho da navegação e os desafios de rastreamento de alterações. ***A solução é dividir os atributos frequentemente analisados ou que mudam com frequência em uma dimensão separada, denominada minidimensão.***

(COVEST-COPSET - Analista (UFPE)/Tecnologia da Informação/2019) Qual técnica de modelagem de Data warehouse ajuda a reduzir o volume de dados de dimensões naturalmente enormes?

a) Minidimensões
b) Microdimensões
c) Dimensões Reduzidas
d) Dimensões Multifacetadas
e) Dimensões Materializadas

_____

Comentário: Perceba que acabamos de comentar sobre o assunto. Uma forma de reduzir o volume das dimensões é criar as chamadas minidimensões.

Gabarito: A

QUESTÕES DE MODELAGEM COMENTADAS

1. FGV - AFRE MG/SEF MG/Tecnologia da Informação/2023 - TI - Banco de Dados - Conceitos de Modelagem Dimensional e Business Intelligence

Kimball elenca uma série de conceitos fundamentais para a elaboração de um modelo dimensional.

Em relação a esses conceitos, assinale a afirmativa ***incorreta***.

a) O estabelecimento da granularidade mostra exatamente o que é representado por uma linha na tabela de fato.

b) A análise de requisitos de negócio e o conhecimento da realidade em relação aos dados disponíveis nos sistemas de origem é uma etapa essencial.

c) Todo o contexto descritivo de um modelo dimensional está pautado nas tabelas de fato, revelando quem, o que, onde, quando, por que, e como.

d) Os processos de negócio são eventos usados para gerar métricas de performance que são traduzidas em fatos em uma tabela de fato.

e) O processo de design do modelo dimensional em quatro etapas contempla: 1) selecionar os processos de negócio; 2) estabelecer a granularidade; 3) identificar as dimensões; e 4) identificar os fatos.

Comentário: A afirmativa incorreta é: c) Todo o contexto descritivo de um modelo dimensional está pautado nas tabelas de fato, revelando quem, o que, onde, quando, por que e como.

Na verdade, o contexto descritivo de um modelo dimensional é principalmente pautado nas **tabelas de dimensão**. As tabelas de dimensão contêm informações descritivas que respondem a perguntas como "quem, o que, onde, quando, por que e como". As tabelas de fato, por outro lado, contêm métricas ou medidas quantitativas. Portanto, as tabelas de dimensão desempenham um papel fundamental na contextualização dos dados nas tabelas de fato.

Gabarito: C

---

2. FGV - Aud Est (CGE SC)/CGE SC/Ciências da Computação/2023- TI - Banco de Dados - Conceitos de Modelagem Dimensional e Business Intelligence

Uma rede de lojas de departamentos planeja configurar uma tabela de fato (VENDAS) que favoreça a integridade para análise de vendas no mesmo carrinho (*na mesma transação de venda, tal qual a associação conhecida entre fralda e cerveja*). A tabela de fato possui os seguintes atributos: ChaveCalendário(FK), ChaveLoja(FK), ChaveProduto(FK), ChaveCliente(FK), IDTransação, HoraMinVenda, ReaisVendidos e QuantidadeVendida.

Assinale a opção que indica o(s) atributo(s) que deve(m) ser a chave primária da tabela de fato VENDAS.

(*FK = Foreign Key/chave estrangeira*)

a) Apenas IDTransação.

b) Apenas ChaveProduto.

c) Chave composta por ChaveProduto e IDTransação.

d) Chave composta por ChaveCliente e IDTransação.

e) IDTransação, ChaveProduto e ChaveCliente.

Comentário: A resposta correta é:c) Chave composta por ChaveProduto e IDTransação.

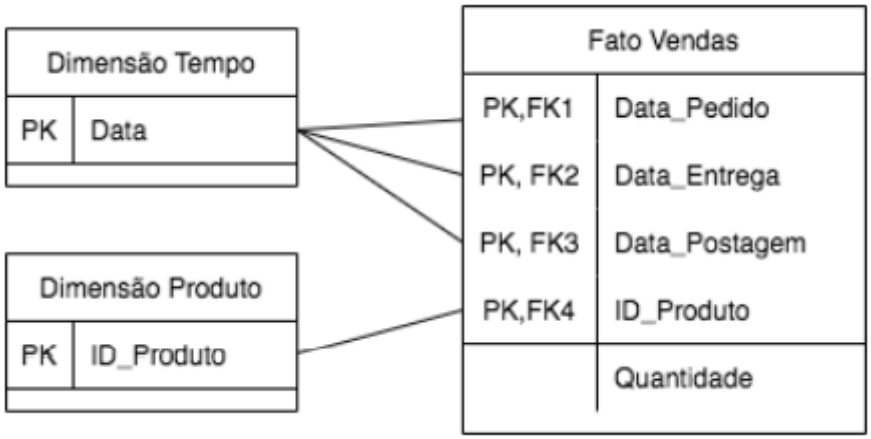
A chave primária deve ser composta por ChaveProduto e IDTransação, garantindo que cada transação de venda seja identificada de forma única e permitindo análises que favoreçam a integridade para vendas no mesmo carrinho (na mesma transação de venda).

Gabarito: C

---

3. FGV – Auditor de Controle Externo – Tecnologia da Informação (TCE-TO)/2022

Observe a seguinte modelagem dimensional.



A técnica utilizada para implementar a dimensão tempo e seus diferentes relacionamentos com a tabela fato é:

a) Factless table;

b) Fact Constellation;

c) Role playing dimension;

d) Degenerated dimension;

e) Slowly changing dimension.

Comentários:

Primeiramente, vamos explicar cada uma das definições acima ...

vou explicar cada uma das definições:

(a) Factless table: Uma factless table é uma tabela em um modelo dimensional de banco de dados que contém apenas chaves estrangeiras, sem medidas ou fatos

associados. Em outras palavras, é uma tabela que registra ocorrências de eventos ou relacionamentos, mas não possui valores numéricos associados a essas ocorrências. Ela é útil quando precisamos rastrear eventos ou ocorrências que não possuem medidas quantitativas associadas, como registros de atividades ou eventos de um processo de negócio.

(b) Fact Constellation: Uma fact constellation, também conhecida como galaxy schema, é um modelo dimensional que consiste em várias tabelas de fatos (fact tables) que compartilham dimensões em comum. Esse padrão é comum quando um conjunto de fatos distintos está relacionado a um mesmo conjunto de dimensões. Cada tabela de fatos na fact constellation representa uma perspectiva ou visão diferente dos dados e, juntas, formam uma constelação.

(c) Role playing dimension: Uma role playing dimension é uma dimensão que é utilizada em diferentes funções ou "papéis" em um modelo dimensional. Geralmente, isso ocorre quando a mesma dimensão é utilizada várias vezes em uma tabela de fatos para representar diferentes aspectos ou contextos. Por exemplo, em um modelo de vendas, uma dimensão de data pode ser usada para representar a data da venda, a data de envio e a data de pagamento, cada uma em um "papel" diferente.

(d) Degenerated dimension: Uma degenerated dimension é uma dimensão que é representada apenas por uma única coluna em uma tabela de fatos. Em vez de criar uma tabela de dimensão separada para essa informação, ela é incorporada diretamente na tabela de fatos. Um exemplo comum de degenerated dimension é o número de pedido ou o número de identificação exclusivo associado a uma transação, que é armazenado na tabela de fatos em vez de criar uma tabela separada para os pedidos.

(e) Slowly changing dimension: Uma slowly changing dimension (dimensão com mudança lenta) é uma dimensão em um modelo dimensional que pode sofrer alterações ao longo do tempo. Existem três tipos principais de slowly changing dimensions: Tipo 1 (SCD1), onde os dados antigos são substituídos pelos novos; Tipo 2 (SCD2), onde novos registros são inseridos para representar alterações ao longo do tempo, criando um histórico; e Tipo 3 (SCD3), onde é mantido um registro do estado anterior e atual, sem criar um histórico completo. Essa técnica é útil para lidar com mudanças em atributos ou características de dimensões que precisam ser rastreadas em análises e relatórios.

Perceba que o problema apresentado é um caso de role play dimension. A dimensão tempo está participando da tabela fato em diferentes contextos. Logo, nossa resposta encontra-se na alternativa C.

Gabarito: Letra C

4. Analista Legislativo (ALAP)/Atividade de Tecnologia da Informação/Desenvolvedor de Banco de Dados/2020

Duas definições de estruturas de dados estão determinadas para um projeto de datamart de uma loja de varejo: uma delas (tabela A) contém a data da venda, a identificação do produto vendido, a quantidade vendida do produto no dia e o valor total das vendas do produto no dia; a outra (tabela B) contém a identificação do produto, nome do produto, marca, modelo, unidade de medida de peso, largura, altura e profundidade da embalagem.

Considerando os conceitos de modelagem multidimensional de data warehouse, as tabelas A e B são, respectivamente:

a) Query e Réplica

b) Fato e Dimensão

c) Dimensão e Réplica

d) Fato e ETL

e) ETL e Query

Comentário: A questão fala que a tabela A contém uma medida numérica, a quantidade de produtos vendidos, logo ela pode ser vista como uma tabela fato. Já a tabela B possui a características dos produtos ou o contexto, o que está diretamente associado a uma dimensão. Desta forma, temos a nossa resposta na alternativa B.

Gabarito: B.

HORA DE PRATICAR!

5. Analista de Tecnologia da Informação (EBSERH HC-UFU)/2020

Em um banco de dados de um data warehouse baseado em modelagem multidimensional, encontrou-se três tabelas: Venda, Vendedor e Produto, entre outras. Um subconjunto da estrutura referente a este trecho do modelo é o seguinte:

As tabelas **Venda**, **Vendedor** e **Produto** são classificadas, respectivamente, como:

a) fato, fato e dimensão.

b) fato, dimensão e dimensão.

c) dimensão, fato e fato.

d) dimensão, fato e dimensão.

e) dimensão, dimensão e fato.

Comentário: Essa questão tem uma pegadinha. Perceba que a ordem das tabelas na figura não é mesma ordem listada no enunciado. Sabemos que tabela central do diagrama (Vendas) é uma tabela fato, já as outras duas são tabelas de dimensão. Assim, temos a resposta na alternativa B.

Gabarito: B.

6. (Ministério da Economia – Especialista em Ciência de Dados - 2020) Julgue os itens a seguir, relativos a conceitos de modelagem dimensional.

Em um processo de modelagem dimensional, a operação de merge/purge agrega informações das dimensões para diminuir a tabela de fatos.

Comentário: A operação de merge/purge vai combinar dados de duas ou mais fontes, identificando e/ou combinando duplicatas e eliminando (purgando) registros indesejados.

Gabarito: Errado.

7. CEBRASPE (CESPE) - Profissional de Tecnologia da Informação (ME)/Atividades Técnicas de Complexidade Gerencial, de Tecnologia da Informação e de Engenharia Sênior/Desenvolvimento de Software/2020

No que se refere a conceitos de modelagem de dados relacional e dimensional, julgue o item a seguir.

Na modelagem dimensional, a tabela fatos armazena as dimensões e os detalhes dos valores descritivos do armazém de dados.

Comentário: A tabela fato armazena os fatos e as respectivas medidas associadas a esses fatos. O contexto ou descrição é armazenado nas dimensões.

Gabarito: Errado.

8. Ano: 2019 Banca: CESPE Órgão: SEFAZ-RS Prova: Auditor Assunto: Modelagem dimensional

Com relação aos modelos de dados multidimensionais, assinale a opção correta.

A A principal característica da tabela de fatos é a ausência de dados redundantes, o que melhora o desempenho nas consultas.

B Esses modelos são cubos de dados, sendo cada cubo representado por uma única tupla com vários atributos.

C Esses modelos proporcionam visões hierárquicas, ou seja, exibição roll-up ou drill-down.

D Os modelos de dados multidimensionais dão ênfase à coleta e às transações de dados.

E Esses modelos não utilizam processos de transferência de dados, mas sim acessos nativos do próprio SGBD utilizado.

Comentário: Vamos comentar cada uma das alternativas:

A) A principal característica da tabela fato é representar os eventos que aconteceram e, geralmente, as medidas associadas a esses eventos. O fato de ter ou não dados duplicados não é uma característica relevante, vai depender do grau de normalização dos dados armazenados na tabela fato. Lembrando que, segundo o Kimball, as tabelas fatos são **normalizadas**.

B) Muitas vezes os modelos dimensionais são representados por uma estrutura denominada cubo de dados. Os cubos podem ser vistos como matrizes que permite que os dados sejam modelados e visualizados em várias dimensões. Não faz sentido que nesta estrutura seja armazenada apenas uma tupla ou registro. Logo, temos uma alternativa errada.

C) Sobre drill-down e roll-up são operações para movimentar a visão dos dados ao longo dos níveis hierárquicos de uma dimensão. Na operação de drill-down o usuário navega de um nível mais alto de detalhe até um nível mais baixo (diminui-se a granularidade). Já na operação de roll-up o usuário navega de um nível mais baixo de detalhe até o nível mais alto (aumenta-se a granularidade). Logo, temos a nossa resposta na alternativa C.

D) Quem dá ênfase a coleta dos dados são os sistemas ETL, sendo tal ação executada na fase inicial do processo. Já os sistemas transacionais (OLTP) dão ênfase às transações. Por outro lado, os modelos multidimensionais focam na análise de dados e informações.

E) Os modelos multidimensionais usam um processo de ETL para carga dos dados nas bases analíticas.

Gabarito: C.

HORA DE PRATICAR!

9. FCC - Auditor Fiscal (SEFAZ-BA)/Tecnologia da Informação/2019

Suponha que uma Auditora Fiscal da área de TI tenha proposto a seguinte modelagem multidimensional para a SEFAZ−BA:

Fato central: Controle de Receitas e Despesas

A partir do Fato Controle de Receitas e Despesas:

Dimensão Tempo

Dimensão Receitas

Dentro da dimensão Receitas: Dimensão Receitas de Impostos

Dentro da dimensão Receitas: Dimensão Receitas de Taxas

Dimensão Despesas

Dentro da dimensão Despesas: Dimensão Tipo de Despesa

Dimensão Cidade

Dentro da dimensão Cidade: Dimensão NF-e

A modelagem multidimensional proposta

a) é o resultado da decomposição de mais de uma dimensão que possui hierarquias entre seus membros, caracterizando o modelo snowflake, a partir de um fato central.

b) tem como característica um fato central, a partir do qual estão dispostas as dimensões que dele participam, em um formato simétrico, característico do modelo star.

c) parte de um elemento central, denominado pivot, a partir do qual são realizadas operações OLAP como roll up, em que busca-se aumentar o nível de detalhe ou diminuir a granularidade da consulta.

d) possui um fato central, a partir do qual estão dispostas as dimensões que dele participam e seus membros, sob uma única estrutura hierárquica, facilitando a inclusão de dados por digitação nas tabelas do DW.

e) não é um modelo normalizado, por isso evita a redundância de valores textuais em cada uma das tabelas, representadas pelas dimensões denominadas dimension tables.

Comentário: Essa questão descreve um modelo de dados dimensional onde as dimensões possuem uma hierarquia entre as tabelas. Perceba que isso nos leva rapidamente ao modelo snowflake! Nele temos uma tabela central fato e tabelas dimensões que não estão ligadas diretamente a tabela fato. Assim, a resposta da questão pode ser vista na alternativa A. Vejamos os erros das demais opções:

B) Não temos um modelo estrela, mas sim o SNOWFLAKE.

C) O elemento central é denominado tabela fato. Pivot é uma operação que pode ser feita sobre o cubo de dados que rotaciona o cubo mudando a perspectiva de análise.

D) A alternativa vinha bem ... até que falou que a existência de um modelo facilita a inclusão de dados por digitalização!! Imagine ter que escanear documentos para inclusão de dados em um DW ... não faz muito sentido!

E) O modelo snowflake possui hierarquia nas dimensões devido a normalização delas ... logo, temos um modelo normalizado.

Gabarito: A.

HORA DE PRATICAR!

10. VUNESP - Programador (CM Piracicaba)/2019

No modelo dimensional, composto por tabelas fato e tabelas dimensão,

a) as tabelas fato não admitem chaves estrangeiras.

b) as tabelas dimensão comportam apenas atributos multivalorados.

c) o relacionamento de cada tabela dimensão para a tabela fato é de “um para muitos”.

d) nas tabelas dimensão há apenas atributos numéricos.

e) não há atributos numéricos nas tabelas fato.

Comentário: Essa questão apresenta as principais características das tabelas fato e dimensões. A tabela fato é composta por colunas (chaves estrangeiras) que apontam para as chaves primárias das respectivas dimensões. Além disso, geralmente, possuem outros atributos representando as medidas associadas a cada fato registrado. Já as tabelas dimensões possuem um relacionamento de um para muitos com a tabela fato, ou seja, cada linha da dimensão pode estar associada a várias linhas da tabela fato, e cada linha da tabela fato está associada a apenas uma linha de cada dimensão. As dimensões também apresentam atributos textuais que descrevem o contexto dos fatos.

Assim, podemos encontrar nossa resposta na alternativa C.

Gabarito: C.

11. CEBRASPE (CESPE) - Analista Judiciário (TJ-AM)/Analista de Sistemas/2019

A respeito de bancos de dados relacionais, julgue o item a seguir.

O esquema multidimensional estrela de data warehouse é composto por uma tabela de fatos associada com uma única tabela para cada dimensão.

Comentário: Sabemos que o esquema estrela tem mais de uma tabela dimensão ... o modelo multidimensional, precisa, pela lógica, de mais de uma dimensão para analisar os dados. Cada uma dessas dimensões é desnormalizada, representada por uma única tabela e ligada diretamente a tabela dimensão. Lembre-se que a dimensão tempo aparece sempre nos modelos dimensionais.

Gabarito: Certo.

12. IDECAN - Professor de Ensino Básico, Técnico e Tecnológico (IF-Baiano)/Informática/2019

A modelagem de data warehouses pode ser feita seguindo diferentes esquemas. Sobre esse tópico, analise as afirmativas:

I. No esquema estrela, os dados são organizados em uma tabela de dimensão e muitas tabelas de fatos.

II. O esquema floco de neve é uma variação do esquema estrela, onde algumas tabelas de fatos são normalizadas, dividindo, assim, os dados em tabelas adicionais.

III. Quando várias tabelas de fatos compartilham tabelas de dimensão, temos o chamado esquema de constelação de fatos ou galáxia, pois podem ser considerados como coleções de estrelas.

a) se somente as afirmativas I e II estiverem corretas.

b) se somente as afirmativas II e III estiverem corretas.

c) se somente a afirmativa I estiver correta

d) se somente a afirmativa II estiver correta.

e) se somente a afirmativa III estiver correta.

Comentário: Vamos analisar as afirmações:

I. **Errado**. O modelo estrela possui uma tabela fato e várias dimensões.

II. **Errado**. A tabelas normalizadas do esquema floco de neve são as tabelas dimensões.

III. **Certo**. Inclusive as dimensões compartilhadas são denominadas dimensões conformes.

Gabarito: E.

13. FUNDATEC - Auditor Fiscal da Receita Municipal (Pref-POA)/2019/"Sem Edição"

A questão baseia-se na Figura 7, que mostra uma modelagem multidimensional, elaborada no Microsoft Access 365 (MS Access 365).

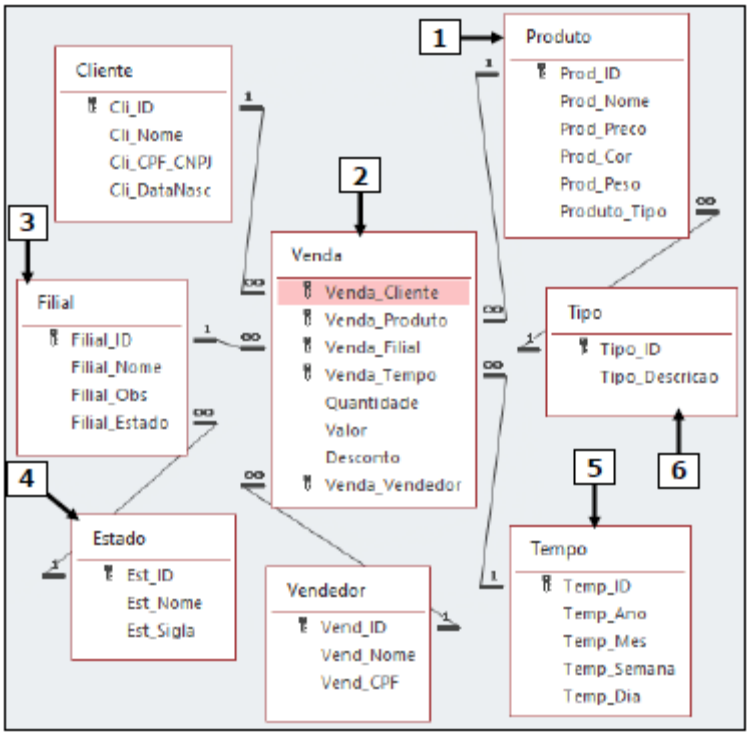


Figura 7 – Modelagem multidimensional

Após observar a Figura 7, analise as seguintes assertivas:

I. A tabela fato, dessa modelagem, é "Venda", apontada pela seta nº 2.

II. As tabelas "Produto", "Filial", "Estado", "Tempo" e "Tipo", apontados, respectivamente pelas setas nº 1, 3, 4, 5 e 6, são tabelas "Dimensão".

III. O esquema multidimensional exibido na Figura 7 é chamado de esquema "Estrela".

Quais estão corretas?

a) Apenas I.

b) Apenas III.

c) Apenas I e II.

d) Apenas II e III.

e) I, II e III.

Comentário: Vamos comentar cada uma das afirmações da questão:

I. Sim! De fato! A tabela fato do esquema em questão é Venda! Veja pela quantidade de atributos que compõe a chave ... já é um forte indicativo! :)

II. Ok! São dimensões! Algumas não estão ligadas diretamente a tabela fato, o que caracteriza um esquema Snowflake.

III. Errado! O esquema em questão é Snowflake.

Gabarito: C.

HORA DE PRATICAR!

14. CEBRASPE (CESPE) - Auditor Municipal de Controle Interno (CGM-João Pessoa)/Tecnologia da Informação/Desenvolvimento de Sistemas/2018

Com relação à modelagem dimensional e à otimização de bases de dados para business intelligence, julgue o item subsequente.

O modelo snowflake acrescenta graus de normalização às tabelas de dimensões, eliminando redundâncias; em termos de eficiência na obtenção de informações, seu desempenho é melhor que o do modelo estrela, o qual, apesar de possuir um único fato, possui tamanho maior que o do snowflake, considerando-se a desnormalização das tabelas de dimensões.

Comentário: A eficiência do modelo estrela é melhor do que o snowflake. Lembre-se: o modelo estrela é mais simples de entender e mais rápido. Por outro lado, o modelo snowflake ocupa menos espaço de armazenamento.

Gabarito: Errado.

HORA DE PRATICAR!

15. CEBRASPE (CESPE) - Auditor Municipal de Controle Interno (CGM-João Pessoa)/Tecnologia da Informação/Desenvolvimento de Sistemas/2018

Com relação à modelagem dimensional e à otimização de bases de dados para business intelligence, julgue o item subsequente.

Na modelagem multidimensional utilizada em data warehouses para se prover melhor desempenho, a tabela fato central deve relacionar-se às suas dimensões por meio da chave primária oriunda da fonte de dados original. O valor dessa chave deve ser idêntico ao da fonte, para que tenha valor semântico e garanta que o histórico das transações seja mantido.

Comentário: A chave primária das tabelas fatos é uma composição dos atributos que apontam para cada uma das dimensões. Ou seja, cada atributo que faz parte da chave é uma chave estrangeira que aponta para a chave primária da sua respectiva dimensão. Já a dimensão possui uma chave primária artificial que é diferente da chave natural usada nos modelos transacionais. Tal mudança permite o armazenamento dos histórico de alterações que acontecem no modelo operacional/transacional dentro do data warehouse. Assim, as chaves primárias oriundas dos dados originais não são usadas como chaves primárias nas dimensões dos modelos analíticos. Em vez disso, uma nova chave artificial é criada. Logo, temos uma alternativa **incorreta**.

Gabarito: Errado.

HORA DE PRATICAR!

16. Ano: 2018 Banca: CESPE Órgão: TCM-BA Cargo: Auditor de Contas Questão: 11

Acerca de modelagem dimensional, assinale a opção correta.

A As granularidades fundamentais para classificar todas as tabelas fato de um modelo dimensional são: transacional, snapshot periódico e snapshot acumulado.

B Os fatos e dimensões não são tabelas do banco de dados, pois, no modelo dimensional, são componentes do cubo de um data warehouse.

C No modelo estrela, as dimensões são normalizadas para tornar mais ágeis as consultas analíticas.

D O modelo floco de neve (SnowFlake) aumenta o espaço de armazenamento dos dados dimensionais, pois acrescenta várias tabelas ao modelo, todavia torna mais simples a navegação por software que utilizarão o banco de dados.

E Os códigos e as descrições associadas, usadas como nomes de colunas em relatórios e como filtros em consultas, não devem

ser gravados em tabelas dimensionais.

Comentário: Vamos comentar cada uma das alternativas acima.

A) As tabelas fatos podem ser estruturadas de três forma distintas que representam a forma como queremos armazenar as informações: transacional, snapshot periódico e snapshot acumulado. Essa classificação vai influenciar a escolha da granularidade da tabela fato. Assim, temos nossa resposta na alternativa A.

B) Os fatos e dimensões são tabelas dos bancos de dados multidimensionais estruturados em uma base relacional. Assim, a alternativa A está **incorreta**.

C) No modelo estrela as dimensões não são normalizadas. A normalização aparece nos modelos floco de neve e tem por objetivo reduzir a redundância dos dados e não a melhora do desempenho das consultas. Temos, portanto, uma afirmação **incorreta**.

D) Pela justificativa da alternativa anterior, podemos observar que o modelo floco de neve reduz o espaço de armazenamento quando reduz a redundância dos dados. Assim, a alternativa C está **errada**.

E) Cada dimensão deve conter os atributos descritivos sobre os dados armazenados na tabela fato. Logo, alternativa D está **errada**.

Gabarito: A.

HORA DE PRATICAR!

17. Ano: 2018 Banca: Cesgranrio Órgão: Petrobras Cargo: Analista de Processo de Negócio Questão: 44

Ao construir um modelo de dados para um data warehouse de sua empresa, um desenvolvedor viu-se às voltas com três tabelas relacionais: venda, cliente e vendedor. Ao fazer uma transformação para o modelo estrela, ele deve organizar:

(A) venda, como tabela fato; cliente e vendedor, como tabelas dimensão

(B) cliente e vendedor, como tabelas fato; venda, como tabela dimensão

(C) cliente, como tabela fato; venda e vendedor, como tabelas dimensão

(D) vendedor e venda, como tabelas fato; cliente, como tabela dimensão

(E) vendedor, como tabela fato; cliente e venda, como tabelas dimensão

Comentário: Observem que pelas definições vistas ao longo da aula, as vendas representam as medidas que queremos analisar. Já as tabelas cliente e vendedor são descrições associadas as vendas realizadas no sistema. Logo, venda, pode ser descrita como tabela fato; cliente e vendedor, como tabelas dimensão.

Gabarito: A

18. Ano: 2017 Banca: CESPE Órgão: SEDF Cargo: Analista de gestão educacional – Especialidade: tecnologia da informação

Com relação aos conceitos de modelagem multidimensional de dados para inteligência computacional, julgue os seguintes itens.

[104] Diferentemente da estrutura relacional, a estrutura multidimensional oferece baixa redundância de dados e suporte a normalização até a segunda forma normal.

[106] Ao se modelar uma tabela-fato, deve-se considerar que a chave primária é composta e que a dimensão tempo sempre será parte integrante dessa chave.

Comentário: Mais uma vez vamos comentar todas as alternativas, desta que foi a primeira prova do CESPE de 2017.

Na alternativa 104 temos uma falha na definição da modelagem multidimensional. Sabemos que o modelo em estrela, mais utilizado no desenho ou projeto de bases de dados analíticas, utiliza-se de alta redundância e baixa normalização para apresentar um modelo de dados numa estrutura mais compreensiva para os usuários finais. Essa construção facilita ainda a navegação entre as diversas dimensões do modelo, facilitando a construção de relatórios. Podemos, então, concluir que a afirmação está incorreta.

Vamos agora comentar a última alternativa (106). A questão apresenta uma sugestão de projeto presente na literatura especializada: a criação da dimensão tempo. Outro ponto, também descrito na questão é a composição da chave primária pela união das chaves artificiais das dimensões a ela relacionadas. Sendo assim, podemos marcar a alternativa como correta.

Gabarito: E C

HORA DE PRATICAR!

19. BANCA:CESPE | CEBRASPE - ANO: 2016 - CONCURSO:FUNPRESP – CARGO 8: ESPECIALISTA - ÁREA: TECNOLOGIA DA INFORMAÇÃO (TI)

Acerca dos modelos de dados relacional e dimensional em engenharia de software, julgue os itens que se seguem.

63 Na modelagem dimensional, as tabelas dimensão estão menos sujeitas ao processo de desnormalização que as tabelas fato.

64 Em um modelo de dados relacional, a integridade referencial assegura que os valores dos campos presentes na chave estrangeira apareçam na chave primária da mesma tabela, a fim de garantir a integridade dos dados.

Comentário: Como de praxe vamos comentar cada uma das alternativas da questão.

63 Os modelos conhecidos nos projetos de DW são o *star schema* e o *snowflake schema*. No primeiro, tanto as tabelas dimensões quanto a fato são desnormalizadas. No segundo, podemos optar pela normalização das dimensões até a terceira forma normal. Observem que em ambos os casos a tabela fato se mantem desnormalizada. Logo a alternativa está incorreta.

64 Essa questão está com o texto um pouco esquisito. O CESPE considerou a alternativa correta, mas eu tenho algumas considerações que me deixaram com dúvida sobre o gabarito da questão. Vamos pensar no caso concreto, uma tabela aluno e outra tabela responsável. Para garantir a integridade entre os dois, o CPF do responsável deve aparecer na tabela aluno como chave estrangeira. CPF é chave primária da relação responsável e chave estrangeira na relação aluno. Vejam que neste caso elas não são a mesma entidade.

Agora vamos analisar uma relação funcionário com um atributo gerente que é o CPF do funcionário que gerencia. Veja que a coluna GERENTE vai receber o valor do CPF de outro funcionário da mesma tabela. Neste caso temos uma chave estrangeira composta por um atributo da mesma relação. Que ficaria coerente com a questão.

Na minha humilde opinião o texto da questão pode ser facilmente derrotado com o contraexemplo do primeiro parágrafo do comentário. Por isso, deixei uma interrogação ao lado do gabarito.

Gabarito: E C(?)

HORA DE PRATICAR!

20. ANO: 2015 BANCA: CESPE ÓRGÃO: MEC PROVA: TÉCNICO DE NÍVEL SUPERIOR - ANALISTA DE SISTEMAS

Com relação aos passos do processo de projeto de bancos de dados e de modelagem de dados relacional e dimensional, julgue os itens subsequentes.

[1] Na modelagem dimensional, implementada em sistemas de data warehouse, o esquema snowflake caracteriza-se por possuir diversas tabelas de fatos e de dimensões, sendo estas últimas organizadas hierarquicamente na terceira forma normal (3FN).

Comentários: O modelo snowflake caracteriza-se por possuir uma tabela fatos e um conjunto de tabelas normalizadas para representação de cada dimensão. Sendo assim, a alternativa está errada por dizer que o modelo possui diversas tabelas fato.

Gabarito: E

21. Ano: 2018 Banca: FCC Órgão: SABESP Cargo: Analista de Gestão Área: Tecnologia da Informação Questão: 42

Um Analista está trabalhando em um Data Warehouse – DW que utiliza no centro do modelo uma única tabela que armazena as métricas e as chaves para as tabelas ao seu redor (que descrevem os dados que estão na tabela central) às quais está ligada. O esquema de modelagem utilizado pelo DW, a denominação da tabela central e a denominação das tabelas periféricas são, respectivamente,

(A) floco de neve, base, granulares.

(B) estrela, fato, dimensões.

(C) constelação, fato, granulares.

(D) atomic, base, branches.

(E) anel, base, dimensões.

Comentário: Essa questão trata os conceitos básicos de um modelo dimensional organizados em tabelas. Neste caso temos uma tabela central, conhecida como tabela **fato,** que vai permitir medir e estabelecer links entre as diversas tabelas periféricas, denominadas **dimensões**. O modelo de dados padrão que utiliza essa estrutura é conhecido como **estrela.** Nele as dimensões se apresentam **desnormalizadas** e ligadas diretamente a tabela fato com cardinalidade 1 – N. Ou seja, cada linha da dimensão pode ter associada a ela várias linhas da tabela fato.

Desta forma, podemos marcar nossa resposta na alternativa B.

Gabarito: B

22. BANCA: FCC ANO: 2017 ÓRGÃO: DPE-RS PROVA: ANALISTA – BANCO DE DADOS

[42] Um dos modelos mais utilizados no projeto e implementação de um data warehouse é o modelo dimensional ou multidimensional. Em um modelo dimensional (composto por uma tabela fato e várias tabelas dimensão),

a) as tabelas dimensão devem conter apenas atributos do tipo literal.

b) a tabela fato tem uma cardinalidade de mapeamento de um para um com cada tabela dimensão.

c) a tabela fato deve conter atributos numéricos, visando proporcionar dados para uma análise de atividades da empresa.

d) há um número teórico mínimo de 3 e máximo de 15 tabelas dimensão.

e) as tabelas dimensão comportam um número máximo teórico de atributos.

Comentário: Essa é uma questão interessante, pois trata de características presentes nas tabelas fatos e nas tabelas dimensões. As tabelas de dimensões **fornecem o contexto para tabelas de fatos** e, portanto, para todas as medidas apresentadas no data warehouse. Embora as tabelas de dimensões sejam geralmente muito menores que as tabelas de fatos, elas são o coração e a alma do data warehouse, pois **fornecem pontos de entrada aos dados**. Costuma-se dizer que um data warehouse é tão bom quanto suas dimensões.

As tabelas fato guardam as medidas sobre os dados, nela as **métricas** são armazenadas, junto com as ***surrogate keys*** que ligam às dimensões que descrevem essa métrica. Com essa rápida lembrança sobre o assunto, podemos analisar as alternativas acima.

Primeiramente, não existe restrição quanto ao tipo do atributo de uma tabela dimensão. Ele precisa apenas ter a característica de descrever o contexto que estamos analisando e fazer sentido dentro do escopo da dimensão e da sua granularidade. Logo, a alternativa “a” está errada!

O relacionamento entre uma tabela fato e suas dimensões é de cardinalidade 1:N. Cada linha da tabela fato pode estar associada a 1 linha de cada tabela dimensão. Contudo, cada linha da dimensão pode estar associada a várias linhas da tabela fato. Assim a letra “b” também está incorreta.

As medidas ou valores numéricos existentes na tabela fato vão proporcionar um melhor entendimento do negócio em questão. Logo a alternativa “c” é a nossa resposta.

Sobre as letras “d” e “e”. Primeiramente não existe limite teórico mínimo e máximo para quantidade de tabelas. Da mesma forma não existe um número máximo teórico dos atributos da tabela fato. O que pode acontecer é, em algumas ferramentas a quantidade é limitada por uma questão técnica.

Gabarito: C.

HORA DE PRATICAR!

23. BANCA: FCC ANO: 2017 ÓRGÃO: TRT - 24ª REGIÃO (MS) PROVA: ANALISTA JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

[42] Uma das formas de apresentação de um banco de dados multidimensional é através do modelo estrela. No centro de um modelo estrela encontra-se a tabela de

a) dimensão e, ao seu redor, as tabelas de fatos.

b) dimensão, cuja chave primária deve ser composta.

c) núcleo e, ao seu redor, as tabelas de nível.

d) fatos, cuja chave primária deve ser simples.

e) fatos e, ao seu redor, as tabelas de dimensões.

Comentário: A modelagem dimensional recebe seu nome das dimensões de negócios que precisamos incorporar ao modelo de dados lógicos. É uma técnica de design lógico que visa estruturar as dimensões de negócios e as métricas analisadas ao longo dessas dimensões. Essa técnica de modelagem tenta ser intuitiva para esse propósito. O modelo também fornece um alto desempenho para consultas e análises. Um modelo dimensional descreve os dados usando dimensões e fatos, que se tornam tabelas reais no banco de dados.

A tabela fato contém informações factuais e, geralmente, é a maior tabela do data warehouse. As tabelas de fatos são normalmente onde todos os dados detalhados (no menor nível de granularidades), que você deseja manter no seu data warehouse, são armazenados, como todas as chamadas telefônicas feitas por um cliente ou os pedidos feitos pelo cliente.

As tabelas de dimensões podem ser vistas como uma tabela de referência para a tabela de fatos, em que **as descrições e outras informações estáticas** sobre uma parte específica dos dados são mantidas. Por exemplo, o produto é considerado uma dimensão porque, nessa tabela, tudo sobre o produto é mantido, como nome completo do produto, fornecedores e tamanho.

Sendo assim, podemos encontrar no nosso modelo uma tabela fato e, ao seu redor, as tabelas dimensões. E, portanto, a nossa reposta pode ser visualizada na alternativa E.

Gabarito: E.

24. BANCA: FCC ANO: 2016 ÓRGÃO: TRT - 20ª REGIÃO (SE) PROVA: TÉCNICO JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

[37] 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.

Comentário: Primeiramente, a abordagem que começa pelos data marts e depois parte para o *data warehouse* é denominada bottom-up. Por ser interativo e incremental esse modelo permite o retorno sobre o investimento mais rapidamente quanto comparado com a abordagem top-down. O objetivo é construir uma estrutura de DW utilizando DMs dependentes.

A questão dos metadados precisa ser trabalhada com cuidado, pois, apesar de alguma dependência entre os DM, não existe garantia que os dados não aparecerão em mais de uma relação estruturados de forma distintas. Isso poderia gerar inconsistência entre DM. Logo, nossa resposta está na alternativa D.

Legamarts ou DM legados é um conceito que existe de fato. Nessa arquitetura os data marts não são integrados. Os data marts legados, ou "legamarts" são marcados por múltiplos processos de extração, múltiplas regras de negócios, falta de arquitetura e múltiplas versões de informações de clientes. Sendo assim, temos um erro na alternativa “e”

Gabarito: D

HORA DE PRATICAR!

25. BANCA: FCC ANO: 2016 ÓRGÃO: PREFEITURA DE TERESINA - PI PROVA: TÉCNICO DE NÍVEL SUPERIOR - ANALISTA DE SISTEMAS

[57] Em um Star Schema de um Data Warehouse – DW, a tabela Dimensão possui característica

a) descritiva dentro do DW. Ela qualifica as informações provenientes da tabela Fato; A tabela Fato possui característica quantitativa dentro do DW. A partir dela são extraídas as métricas que são cruzadas com os dados das Dimensões. Dimensões são ligadas entre si e qualquer uma delas se liga diretamente a tabela Fato. Os dados devem ser normalizados.

b) quantitativa dentro do DW. Ela quantifica as informações provenientes da tabela Fato; A tabela Fato possui característica descritiva dentro do DW. A partir dela são extraídas as nomenclaturas que são quantificadas com os dados das Dimensões. Dimensões são ligadas entre si. Os dados devem ser desnormalizados.

c) quantitativa dentro do DW. Ela quantifica as informações provenientes da tabela Fato; A tabela Fato possui característica descritiva dentro do DW. A partir dela são extraídas as

nomenclaturas que são quantificadas com os dados das Dimensões. Dimensões não são ligadas entre si. Os dados devem ser normalizados.

d) quantitativa dentro do DW. Ela quantifica as informações provenientes da tabela Fato; A tabela Fato possui característica descritiva dentro do DW. A partir dela são extraídas as nomenclaturas que são quantificadas com os dados das Dimensões. Dimensões são ligadas entre si. Os dados devem ser normalizados.

e) descritiva dentro do DW. Ela qualifica as informações provenientes da tabela Fato; A tabela Fato possui característica quantitativa dentro do DW. A partir dela são extraídas as métricas que são cruzadas com os dados das Dimensões. Dimensões são ligadas diretamente a tabela Fato. Outra característica marcante é que os dados são desnormalizados.

Comentário: É importante perceber que a tabela fato possui **característica quantitativa dentro do DW**. A partir dela são **extraídas as métricas,** que são cruzadas com os dados das tabelas de dimensões, concebendo, assim, informações significativas para a análise do usuário. A **tabela fato armazena as medições** necessárias para avaliar o assunto pretendido. O conteúdo histórico no DW, contendo longo período, fica depositado na tabela fato.

A estrutura dimensional normalmente é desenhada no formado do esquema estrela (star schema). Nesse modelo, as tabelas de dimensões são ligadas diretamente a tabela Fato. Outra característica marcante é que **os dados são desnormalizados**, pois a redundância resultante gera benefícios para a otimização das consultas e navegação das informações.

Desta forma, se compararmos o texto acima com as alternativas, vamos encontrar nossa resposta na alternativa E.

Gabarito: E

HORA DE PRATICAR!

26. BANCA: FCC - ANO: 2016 ÓRGÃO: PREFEITURA DE TERESINA - PI PROVA: ANALISTA TECNOLÓGICO - ANALISTA DE SUPORTE TÉCNICO

[35] O modelo dimensional utilizado na modelagem de data warehouse tem como característica:

a) Todas as tabelas dimensão de um mesmo modelo devem possuir o mesmo número de atributos.

b) A tabela fato possui pelo menos 4 atributos numéricos, além das chaves estrangeiras.

c) Poder ter quantas tabelas dimensionais, quantas forem necessárias para representar o negócio sob análise.

d) As tabelas dimensão não necessitam ter atributos que sirvam como chave primária.

e) A cardinalidade de relacionamento da tabela fato para as tabelas dimensão é de um para um.

Comentário: Já vimos que não existe restrições quanto a quantidade de dimensões presentes em um modelo. Outro ponto é que não existe restrição quanto ao número de atributos do modelo, pelo menos do ponto de vista teórico. Logo, a alternativa C é a nossa resposta.

Veja que o erro da letra "d" é afirmar que a tabela dimensão não possui atributo que possa ser usado como chave primária. Mesmo que isso aconteça é possível criar uma chave artificial para a relação. Por fim, a alternativa "e" está errada pois a cardinalidade da tabela fato e das dimensões é 1:N.

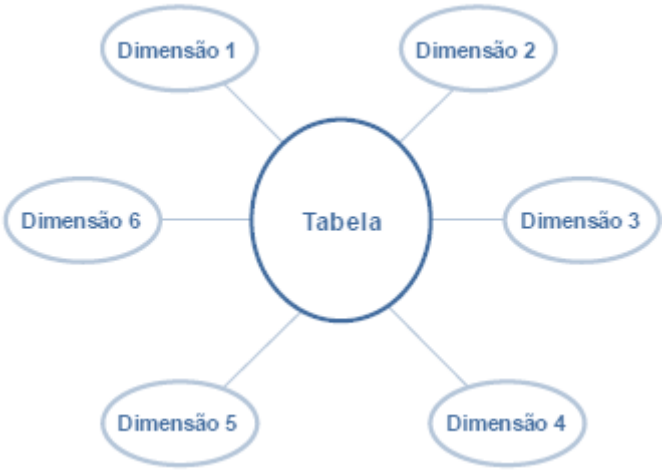
Logo, nossa reposta pode ser vista na alternativa C.

Gabarito: C

HORA DE PRATICAR!

27. BANCA: FCC ANO: 2016 ÓRGÃO: ELETROBRAS-ELETROSUL PROVA: INFORMÁTICA

[57] Considere a figura abaixo que ilustra um modelo multidimensional na forma de modelo relacional em esquema estrela. Há uma tabela central que armazena as transações que são analisadas e ao seu redor há as tabelas look up, denominadas dimensões.



De acordo com o modelo estrela da figura e sua relação com um Data Warehouse, é correto afirmar:

a) Uma das candidatas à chave primária da tabela central, denominada star table, seria uma chave composta pelas chaves primárias de todas as dimensões.

b) A tabela fato armazena os indicadores que serão analisados e as chaves que caracterizam a transação. Cada dimensão registra uma entidade que caracteriza a transação e os seus atributos.

c) As dimensões devem conter todos os atributos associados à sua chave primária. Por causa disso, o modelo multidimensional estrela está na 3ª Forma Normal.

d) O modelo estrela é derivado do modelo snowflake, ou seja, é o resultado da aplicação da 1ª Forma Normal sobre as entidades dimensão.

e) Um Data Waherouse, por permitir a inclusão de dados por digitação, necessita da aplicação de normalização para garantir a unicidade de valores.

Comentário: Vamos analisar cada uma das alternativas. Primeiramente a letra "a" cria um conceito que não existe na literatura denominado "star table", logo está incorreta.

A letra b é a nossa resposta. E está perfeitamente condizente com o que vimos ao longo do curso. A tabela fato está associada às medidas ou indicadores que podem ser quantificados com o cruzamento das diversas dimensões. Já as dimensões representam as descrições ou o contexto associado às diversas perspectivas da tabela fato.

Quanto às alternativas "c", "d" e "e", podemos afirmar o seguinte: (c) o fato dos atributos estarem associados a chave primária não elimina da dependência transitiva (3FN); (d) na realidade o modelo snowflake é derivado do star schema; e (e) o DW deve usar ferramentas ETL para viabilizar a carga dos dados, não temo como fazer isso por meio de digitação.

Gabarito: B

28. BANCA: FCC ANO: 2016 ÓRGÃO: TRT - 23ª REGIÃO (MT) PROVA: ANALISTA JUDICIÁRIO – TECNOLOGIA DA INFORMAÇÃO

[34] Na abordagem Star Schema, usada para modelar data warehouses, os fatos são representados na tabela de fatos, que normalmente

a) é única em um diagrama e ocupa a posição central.

b) está ligada com cardinalidade n:m às tabelas de dimensão.

c) está ligada às tabelas de dimensão, que se relacionam entre si com cardinalidade 1:n.

d) tem chave primária formada independente das chaves estrangeiras das tabelas de dimensão.

e) está ligada a outras tabelas de fatos em um layout em forma de estrela.

Comentário: Os esquemas de um data warehouse às vezes são chamados de esquemas em estrela. O **ponto central é a tabela fato e as dimensões ficam em torno da tabela fato como pontos da nossa estrela**. Sendo assim, podemos marcar nossa resposta na alternativa A.

As demais alternativas estão erradas pelos seguintes motivos: (b) a cardinalidade entre a tabela fato e as tabelas dimensões é 1:N; (c) as tabelas dimensões, geralmente, não se relacionam entre si. Existem alguns casos das dimensões *outtriggers*, por exemplo, que aceitam que uma dimensão faça referência a outra; (d) a chave primária da tabela fato é formada pela composição das chaves estrangeiras de cada tabelas dimensão e; por fim (e) quando uma tabela fato se liga a outra, o layout recebe o nome de constelação.

Gabarito: A

29. ANO: 2014 BANCA: FCC ÓRGÃO: TJ-AP PROVA: ANALISTA JUDICIÁRIO - BANCO DE DADOS - DBA

Os sistemas de Data Warehouse utilizam-se de um modelo de dados diferente dos bancos de dados tradicionais, que proporciona ganhos de desempenho nas consultas. Esse modelo é conhecido como modelagem

A dinâmica.

B dimensional.

C fixa.

D online.

E transacional.

Comentário: A **modelagem multidimensional**, ou dimensional como às vezes é chamada, é a técnica de modelagem de banco de dados para o auxílio às consultas do *Data Warehouse* nas mais diferentes perspectivas. A visão multidimensional permite o uso mais intuitivo para o processamento analítico pelas ferramentas OLAP (*On-line Analytical Processing*).

Gabarito: B.

30. ANO: 2015 BANCA: FCC ÓRGÃO: TRT - 3ª REGIÃO (MG) PROVA: TÉCNICO JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

A modelagem multidimensional é utilizada especialmente para sumarizar e reestruturar dados e apresentá-los em visões que suportem a análise dos valores desses dados. Um modelo multidimensional é formado por dimensões, e por uma coleção de itens de dados composta de dados de medidas e de contexto, denominada

A schema.

B pivot.

C slice.

D fato.

E versão.

Comentário: Um modelo multidimensional é formado por 3 elementos básicos: **fatos, dimensões e medidas (variáveis).**

Fatos: Uma tabela fato é uma coleção de itens de dados, composta de dados de medidas e de contexto. Cada fato representa um item, uma transação ou um evento de negócio e é utilizado

para analisar o processo de negócio de uma empresa. É tudo aquilo que reflete a evolução dos negócios do dia a dia de uma organização.

Dimensões: São elementos que participam de um fato, assunto de negócios. São possíveis formas de visualizar os dados, por exemplo: “por mês”, “por país”, “por produto”, “por região”, “por funcionário”, e por aí vai. Dimensões normalmente não possuem atributos numéricos, pois são somente descritivas e classificatórias dos elementos que participam de um fato.

Medidas (variáveis): São os atributos numéricos que representam um fato, a performance de um indicador de negócios relativo às dimensões que participam desse fato, tais números são denominados de variáveis.

Analisando as definições acima, podemos encontrar nossa resposta na alternativa D.

Gabarito: D.

31. ANO: 2015 BANCA: FCC ÓRGÃO: TCM-GO PROVA: AUDITOR DE CONTROLE EXTERNO - INFORMÁTICA

Quando o modelo de dados multidimensionais começa a ser definido, elementos básicos de representação precisam ter sido estabelecidos, de modo a se criar um padrão de modelagem. Considere um modelo em que as dimensões e fatos são representados em tabelas, podendo haver múltiplas dimensões e múltiplas tabelas de fatos.

Ao modelar cada tabela _____I_____ devem ser considerados os seguintes pontos:

– A chave primária é composta, sendo um elemento da chave para cada dimensão;

– Cada elemento chave para a dimensão deve ser representado e descrito na tabela _____II_____ correspondente (para efetuar a junção);

– A dimensão tempo é sempre representada como parte da chave primária.

Deve haver uma tabela _____III_____ para cada dimensão do modelo, contendo

– Uma chave artificial (ou gerada) genérica;

– Uma coluna de descrição genérica para a dimensão;

– Colunas que permitam _____IV_____;

– Um indicador nível que indica o nível da hierarquia a que se refere a linha da tabela.

As lacunas de I a IV são corretas, e respectivamente, preenchidas com:

A dimensão – de fatos – de tempo – efetuar os filtros.

B dimensão – de fatos – de fatos – a junção com as tabelas de fatos.

C de fatos – de tempo – dimensão – sinalizar a presença de fatos para o período de tempo indicado na linha.

D de fatos – dimensão – dimensão – efetuar os filtros.

E de tempo – dimensão – de fatos – a junção com as tabelas de dimensão.

Comentário: Vamos novamente fazer uma revisão sobre os termos básicos da modelagem dimensional, em especial as tabelas fato e dimensões:

Tabela fato: contém uma grande quantidade de linhas que correspondem a fatos observados e links externos. Contém atributos descritivos necessários à análise de decisão e reporte de consultas. Servem para o armazenamento, medidas (quase sempre) numéricas associadas a eventos de negócio. Ao modelar a (s) tabela (s) de fatos (ou apenas tabela fato), deve-se ter em mente os seguintes pontos:

- A chave primária é composta, sendo um elemento da chave para cada dimensão.
- Cada elemento chave para a dimensão deve ser representado e descrito na "tabela dimensão" correspondente (para efetuar a junção).
- A dimensão tempo é sempre representada como parte da chave primária.

Tabela dimensão: contém informação de classificação e agregação sobre as linhas da tabela fato. Contém atributos que descrevem os dados contidos em uma tabela fato. Representa entidades de negócios e constituem as estruturas de entrada que servem para armazenar informações como tempo, geografia, produto, cliente. Deve haver uma "tabela dimensão" para cada dimensão do modelo, contendo:

- Um indicador NÍVEL que indica o nível da hierarquia a que se refere a linha da tabela.
- Colunas que permitam efetuar os filtros.
- Uma coluna de descrição genérica para a dimensão.
- Uma chave artificial (ou gerada) genérica.

Pelo texto exposto, acima podemos chegar à conclusão de que a nossa resposta se encontra na alternativa D.

Gabarito: D

HORA DE PRATICAR!

32. ANO: 2013 BANCA: FCC ÓRGÃO: MPE-MA PROVA: ANALISTA JUDICIÁRIO - BANCO DE DADOS

Na modelagem de um data warehouse, pode ser feito o snowflaking, que significa

A criptografar as tabelas fato e dimensão.

B normalizar as tabelas dimensão.

C excluir atributos do tipo binário.

D indexar as tabelas dimensão por todos seus atributos.

E duplicar a tabela fato.

Comentário: O modelo floco de neve reduz o espaço de armazenamento e a redundância das tabelas dimensão por meio do processo de normalização delas até a terceira forma normal. Sabemos, porém, que esse processo carrega consigo um custo adicional da operação de junção no momento da consulta sobre os dados do modelo.

A nossa resposta pode ser vista na alternativa B, que afirma que o snowflaking significa normalizar as tabelas dimensão.

Gabarito: B

33. ANO: 2013 BANCA: FCC ÓRGÃO: MPE-MA PROVA: ANALISTA JUDICIÁRIO - BANCO DE DADOS

Na modelagem dimensional de um data warehouse voltado para vendas, o tipo de tabela fato que inclui pares de produtos adquiridos em uma mesma compra recebe a denominação de

A cesta de mercado.

B tabela de degeneração.

C data mart.

D outrigger.

E pacote de integralização.

Comentário: Essa questão apresenta um conceito que é visto com maiores detalhes quando falamos das regras de associação em mineração de dados. Contudo, quando tratamos de vendas de supermercados, por exemplo, podemos trazer a lembrança do carrinho ou cesta de compras.

O termo "**cesta de mercado**" (do inglês ***Market-Basket Model***) é usado para descrever um modelo de descoberta de associações usado no processo de ***Data Mining***. O objetivo desta técnica é identificar quais produtos são mais prováveis de serem consumidos em conjunto, a fim de determinar a disposição deles nas lojas.

Desta forma, podemos marcar o gabarito na alternativa A.

Gabarito: A.

34. ANO: 2009 BANCA: FCC ÓRGÃO: TRT - 15ª REGIÃO (CAMPINAS-SP) PROVA: ANALISTA JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

No contexto OLAP:

I. As visões materializadas agregadas a partir de uma tabela de fatos podem ser identificadas exclusivamente pelo nível de agregação para cada dimensão.

II. Quando aplicada a configuração star schema as tabelas de fatos e as de dimensão são idênticas quanto à totalidade dos atributos que contêm e também quanto ao grau de granularidade.

III. O esquema snow flake é uma variação do star schema.

Está correto o que consta em

A I, somente.

B I e III, somente.

C II e III, somente.

D III, somente.

E I, II e III.

Comentário: Primeiro essa questão foi tirada de um material da oracle: http://cdn.ttgtmedia.com/searchOracle/downloads/Teorey08.pdf.

Na realidade ser você colocar esse trecho na internet: "*Materialized views aggregated from a fact table can be uniquely identified by the aggregation level for each dimension.*" Vai encontrar ele em uma meia dúzia de artigos. Mas o que ele quer dizer com isso? Bom vou colocar aqui a explicação do artigo para não ficar traduzindo.

“Given a hierarchy along a dimension, let 0 represent no aggregation, 1 represent the first level of aggregation, and so on. For example, if the Invoice Date dimension has a hierarchy consisting of date id, month, quarter, year and “all” (i.e., complete aggregation), then date id is level 0, month is level 1, quarter is level 2, year is level 3, and “all” is level 4. If a dimension does not explicitly have a hierarchy, then level 0 is no aggregation, and level 1 is “all.”

Basicamente ele associa a um nível da hierarquia de valores um número e se existir uma visão materializada para aquele nível ele pode ser identificado pelos seus níveis de hierarquia.

Suponha duas dimensões: Tempo(dia(0), mês(1), ano(2)), Produto(Suco(0), Sanduíche(1)) ... O nível de agregação (1,1) identifica uma visão materializada no nível de (mes, Sanduiche). Com isso podemos avaliar a primeira alternativa como correta.

A alternativa II, apresenta uma confusão sobre os atributos e a granularidade das tabelas fatos e dimensões, estando, portanto, incorreta.

Por fim, a alternativa III fala que snowflake é uma variação do star schema, onde as dimensões são normalizadas. Isso está correto.

Juntando as análises das afirmações temos nosso gabarito na alternativa B.

Gabarito: B

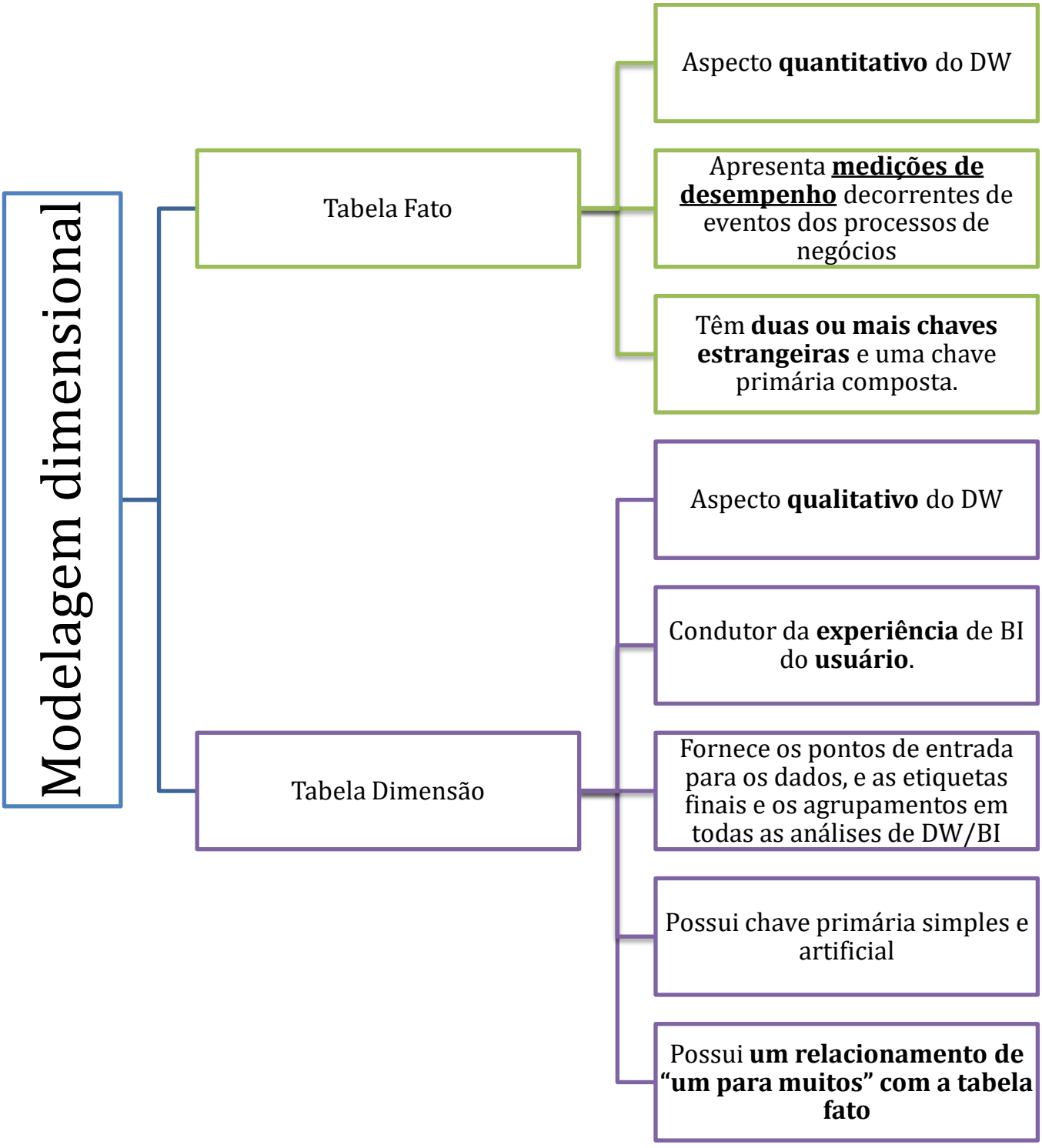
RESUMO

DATA WAREHOUSE E MODELAGEM DIMENSIONAL

Um data warehouse é um sistema de gerenciamento de dados projetado para possibilitar e apoiar atividades de business intelligence (BI), especialmente análise de dados. É uma centralização de dados coletados de diferentes fontes heterogêneas, como bancos de dados transacionais, sistemas de ERP, e outras fontes externas e internas. Esses dados são integrados, consolidados, organizados e armazenados em um formato que facilita a consulta e a análise.

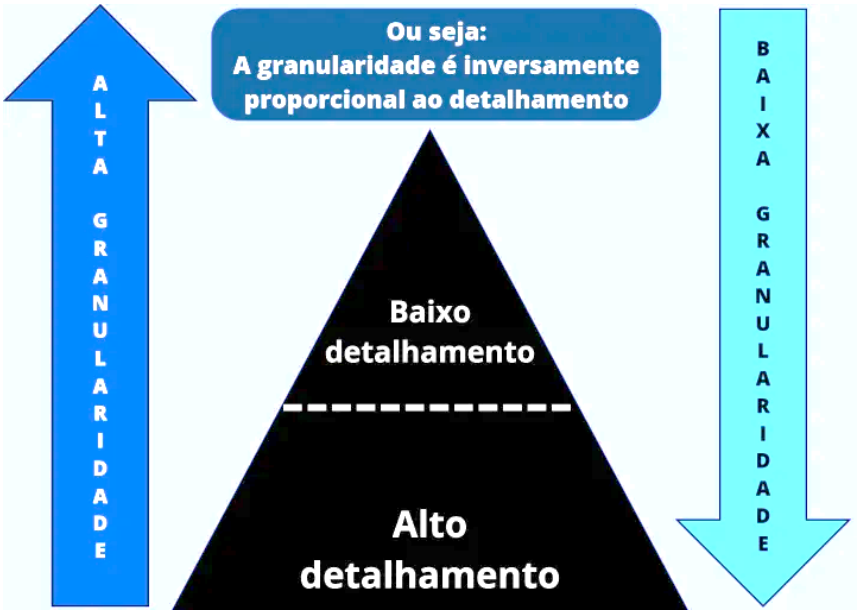
Principais Conceitos e Definições:

1. **ETL (Extract, Transform, Load)**:
    - **Extract (Extração)**: O processo de coleta de dados de diversas fontes de dados. Essas fontes podem ser bancos de dados, planilhas, arquivos de log, ou sistemas de ERP.
    - **Transform (Transformação)**: Dados extraídos são limpos, formatados e transformados conforme as necessidades de análise. Esta etapa pode incluir filtragem, agregação, padronização e sumarização dos dados.
    - **Load (Carregamento)**: Os dados transformados são carregados no data warehouse. Esse processo pode ser realizado em tempo real (near real-time) ou em batch (lote).
2. **Modelagem Dimensional**:
    - **Fato**: Representa eventos ou transações que ocorrem na organização. Contém dados **quantitativos**, como vendas, receitas, etc. As tabelas de fato geralmente possuem **muitas linhas e poucas colunas**.
    - **Dimensão**: Fornece contexto às medidas armazenadas nas tabelas de fato. As dimensões contêm atributos **descritivos e qualitativos**, como tempo, localização, produto, etc. As tabelas dimensionais possuem muitas colunas e relativamente poucas linhas.
    - **Esquemas Estrela e Floco de Neve**: Modelos usados para organizar dados no data warehouse. O esquema estrela tem uma tabela de fato central conectada diretamente a várias tabelas dimensionais. O esquema floco de neve normaliza as tabelas dimensionais, quebrando-as em tabelas menores e mais detalhadas.



3. **Granularidade**:

- Refere-se ao **nível de detalhe** ou a finura dos dados armazenados no data warehouse. A granularidade pode variar de dados altamente **detalhados** (como transações individuais) a dados altamente **agregados** (como totais mensais de vendas). Decisões sobre granularidade afetam o design, a performance e a capacidade de armazenamento do data warehouse.



4. **Data Mart**:
    - Um subconjunto do data warehouse voltado para uma área específica de negócio ou departamento. Data marts são menores e mais focados, facilitando o acesso e a análise de dados relevantes para usuários específicos.
5. **Armazenamento e Particionamento**:
    - O data warehouse pode ser particionado para melhorar a performance e a gerenciabilidade. Particionamento divide grandes tabelas em partes menores e gerenciáveis, com base em critérios como data, geografia, ou outro atributo relevante.
6. **Metadados**:
    - Dados sobre dados. Metadados descrevem a estrutura, as regras, as transformações e o contexto dos dados armazenados no data warehouse. Eles ajudam a entender, gerenciar e usar o data warehouse de forma eficaz.
7. **Qualidade dos Dados**:
    - A precisão, a consistência e a integridade dos dados são cruciais para a confiança nas análises feitas com base no data warehouse. Processos de limpeza e validação de dados são essenciais para manter a alta qualidade dos dados.

Vantagens de um Data Warehouse:

- **Integração de Dados**: Consolida dados de várias fontes, proporcionando uma visão única e integrada das operações da organização.
- **Performance**: Projetado para consultas e análises rápidas, diferentemente dos sistemas transacionais que são otimizados para operações de inserção e atualização.

- **Histórico de Dados**: Armazena dados históricos, permitindo análises de tendências e comparações ao longo do tempo.
- **Tomada de Decisão Informada**: Fornece informações detalhadas e precisas para suporte à decisão em todos os níveis da organização.

Desafios:

- **Custo**: Implementar e manter um data warehouse pode ser caro em termos de infraestrutura, software e mão-de-obra.
- **Complexidade**: A construção e a gestão de um data warehouse envolvem processos complexos de ETL, modelagem de dados e garantia de qualidade.
- **Escalabilidade**: À medida que o volume de dados cresce, o data warehouse deve ser capaz de escalar eficientemente para continuar a atender às necessidades de performance.

Em resumo, um data warehouse é um componente essencial de uma estratégia de BI, proporcionando uma base sólida para a análise e a tomada de decisões informadas. Ele integra dados de várias fontes, armazena-os em um formato adequado para análise e facilita a extração de insights valiosos para a organização.

DATA LAKE

Um **data lake** é como um vasto reservatório digital, capaz de armazenar grandes volumes de dados em seu formato original, sem a necessidade de pré-processamento. Essa flexibilidade permite que empresas capturem e armazenem dados estruturados (como dados de bancos de dados relacionais), semi-estruturados (como arquivos JSON ou XML) e não estruturados (como logs, imagens e vídeos).

Por que usar um data lake?

- **Flexibilidade:** Armazena dados em diversos formatos, permitindo análises futuras e imprevistas.
- **Escalabilidade:** Adapta-se a volumes de dados crescentes sem a necessidade de reestruturação.
- **Custo-benefício:** Geralmente mais econômico para armazenamento de grandes volumes de dados.
- **Centralização:** Unifica dados dispersos em diversas fontes, facilitando a gestão e a análise.

Tipos de Data Lakes

Existem diferentes tipos de data lakes, cada um com suas características e aplicações:

- **Data Lake House:** Combina as melhores características de data warehouses e data lakes, oferecendo tanto estrutura quanto flexibilidade.
- **Data Lake on Cloud:** Hospedado em uma plataforma de nuvem, proporcionando escalabilidade e elasticidade.
- **Data Lake on Premise:** Instalado localmente na infraestrutura da empresa, oferecendo maior controle e segurança.

Zonas em um Data Lake

Para organizar e gerenciar os dados de forma eficiente, os data lakes são divididos em zonas:

- **Zona crua:** Armazena os dados brutos, sem qualquer tipo de processamento ou limpeza.
- **Zona refinada:** Contém os dados que passaram por um processo de limpeza e transformação, tornando-os mais adequados para análise.
- **Zona curável:** Armazena os dados que foram processados e preparados para serem utilizados em modelos de machine learning ou outras aplicações.

O Pântano de Dados

Um **pântano de dados** é um termo utilizado para descrever um data lake que se tornou desorganizado e difícil de gerenciar. Isso ocorre quando os dados não são devidamente organizados, documentados e governados. Um pântano de dados pode comprometer a qualidade dos dados e dificultar a obtenção de insights valiosos.

Como evitar o pântano de dados?

- **Governança de dados:** Estabelecer regras e políticas para a gestão dos dados.
- **Metadados:** Documentar os dados para facilitar a compreensão e o acesso.
- **Qualidade dos dados:** Garantir a precisão e a consistência dos dados.
- **Limpeza de dados:** Remover dados duplicados, inconsistentes e inválidos.

Em resumo, o data lake é uma ferramenta poderosa para empresas que desejam armazenar e analisar grandes volumes de dados de forma eficiente. Ao entender os conceitos de data lake, os diferentes tipos, as zonas e como evitar o pântano de dados, as empresas podem tirar o máximo proveito dessa tecnologia e obter insights valiosos para a tomada de decisões.

DATA MESH

Data Mesh é uma abordagem moderna de arquitetura de dados que visa resolver problemas comuns em grandes organizações, onde a centralização de dados pode levar a gargalos, ineficiências e desafios de escalabilidade. A ideia central do Data Mesh é distribuir a responsabilidade dos dados para os domínios de negócios, tratando os dados como um produto e promovendo a autonomia dos times. Aqui estão os principais conceitos e definições do Data Mesh:

1. **Domínios de Dados (Data Domains)**:
    - O Data Mesh propõe a decomposição da arquitetura de dados em domínios de negócios específicos, cada um responsável pelos seus próprios dados. Isso reflete uma abordagem descentralizada onde cada domínio gerencia, possui e serve seus dados, semelhante ao funcionamento dos microserviços em engenharia de software.
2. **Dados como Produto (Data as a Product)**:
    - Os dados devem ser tratados como produtos, com foco na satisfação dos "clientes" dos dados, que podem ser outros domínios, analistas de dados, ou sistemas de BI. Cada domínio deve garantir a qualidade, acessibilidade e usabilidade dos dados que produz e publica.
3. **Equipes Multifuncionais e Autônomas**:
    - Cada domínio de dados deve ter equipes multifuncionais que possuem todas as habilidades necessárias para desenvolver e operar suas próprias pipelines de dados. Essas equipes têm autonomia para decidir como modelar, armazenar e compartilhar seus dados, permitindo maior agilidade e inovação.
4. **Interoperabilidade e Padrões Federados**:
    - Embora os domínios sejam autônomos, é crucial estabelecer padrões comuns para garantir a interoperabilidade entre eles. Esses padrões podem incluir esquemas de dados, contratos de API, governança de dados, segurança e conformidade. A federação de padrões permite que os dados sejam facilmente compreendidos e integrados entre diferentes domínios.
5. **Plataforma de Dados Autosserviço (Self-Serve Data Platform)**:
    - Para suportar a autonomia dos domínios, uma plataforma de dados autosserviço deve ser disponibilizada. Esta plataforma fornece ferramentas e infraestrutura necessárias para que os domínios possam facilmente ingerir, processar, armazenar e compartilhar dados. Ela inclui capacidades de gerenciamento de dados, segurança, monitoramento e catalogação.
6. **Governança Federada**:

- A governança de dados no Data Mesh é federada, o que significa que há um equilíbrio entre governança centralizada e descentralizada. As políticas e normas são definidas de forma central, mas a execução e a adaptação dessas políticas são realizadas pelos próprios domínios, permitindo flexibilidade e conformidade simultaneamente.

Vantagens do Data Mesh

- **Escalabilidade**: A descentralização dos dados permite que a organização escale suas operações de dados sem criar gargalos comuns em arquiteturas centralizadas.
- **Agilidade**: Equipes autônomas podem desenvolver e implementar soluções de dados mais rapidamente, respondendo às necessidades do negócio em tempo hábil.
- **Qualidade de Dados**: Tratando dados como um produto, há um maior foco na qualidade e na satisfação dos usuários finais dos dados.
- **Inovação**: Com a autonomia para experimentar e inovar, os domínios podem adotar novas tecnologias e abordagens sem estar limitados por uma infraestrutura centralizada.

Desafios do Data Mesh

- **Coordenação**: Estabelecer e manter padrões comuns entre domínios pode ser desafiador.
- **Mudança Cultural**: Requer uma mudança significativa na cultura organizacional, com ênfase na colaboração e na responsabilidade dos dados.
- **Complexidade Operacional**: A descentralização pode aumentar a complexidade operacional, exigindo novas habilidades e ferramentas para gerenciar um ambiente distribuído.

Data Mesh é uma abordagem inovadora que promove a descentralização e a autonomia no gerenciamento de dados, tratando-os como produtos e utilizando equipes multifuncionais para garantir a qualidade e a acessibilidade dos dados. Com uma plataforma de dados autosserviço e governança federada, o Data Mesh busca proporcionar escalabilidade, agilidade e inovação, ao mesmo tempo em que enfrenta desafios de coordenação e mudança cultural.

EXERCÍCIOS

Apresentamos abaixo uma lista de exercícios das mais variadas bancas para que você possa praticar um pouco. Qualquer dúvida estou às ordens!

BUSINESS INTELLIGENCE

1. Analista (Prefeitura de Vila Velha)/Desenvolvimento/2020

O processo de pesquisa, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de um negócio é conhecido pela sigla:

a) AFP.

b) SGBD.

c) BI.

d) ERP.

e) GED

2. Analista Legislativo (ALAP)/Atividade de Tecnologia da Informação/Desenvolvedor de Sistemas/2020

Para construir um Data Warehouse, algumas etapas e processos são necessários. Uma etapa é conhecida como ETL, que compreende as etapas de Extração, Transformação e Armazenagem de dados em Sistemas Específicos ou Armazéns de Dados. Essas etapas são constituídas de várias outras funções, processos e técnicas de data integration. Uma dessas funções chama-se Master Data Management – MDM e é responsável por

a) misturar os dados para criar um panorama virtual.

b) unir os dados para criar uma visão única deles, através de múltiplas fontes. Ela inclui tanto o ETL quanto capacidades de data integration, para misturar as informações e criar o "melhor registro".

c) monitorar e processar fluxos de dados e ajudar a tomar decisões mais rapidamente.

d) fornecer tanto agendamento em lote quanto capacidades em tempo real.

e) criar um ambiente de testes onde os dados possam ser integrados, limpos e padronizados (por exemplo: SP e São Paulo, Masculino e M, Senhora e Sra. etc) além de verificar e remover dados duplicados.

3. IBFC - Analista de Tecnologia da Informação (EBSERH)/2020

Dado os três conceitos técnicos abaixo, assinale a alternativa que corresponda respectivamente à tecnologia referente a cada um desses conceitos.

1. processo de explorar grandes quantidades de dados à procura de padrões consistentes.

2. refere-se ao processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de negócios.

3. depósito de dados digitais que serve para armazenar informações detalhadas relativamente a uma empresa.

a) 1.Data Warehouse - 2.Business Intelligence - 3.Data Mining

b) 1.Data Mining - 2.Data Warehouse - 3.Business Intelligence

c) 1.Business Intelligence - 2.Data Warehouse - 3.Data Mining

d) 1.Data Mining - 2.Business Intelligence - 3.Data Warehouse

e) 1.Business Intelligence - 2.Data Mining - 3.Data Warehouse

4. FAEPESUL - Assistente (CRC SC)/Suporte em Informática/2019

É correto afirmar que Business Intelligence é:

a) O processo de coleta, organização, análise, compartilhamento e monitoramento de informações para a gestão de negócios.

b) Um software.

c) O mesmo que inteligência artificial.

d) O nome dado a um algoritmo de pesquisa.

e) Um padrão de projetos.

5. CEBRASPE (CESPE) - Auditor Municipal de Controle Interno (CGM João Pessoa)/Tecnologia da Informação/Desenvolvimento de Sistemas/2018

A respeito de business intelligence, julgue o item.

Business intelligence pode ser definido como um processo inteligente de coleta, organização, análise, compartilhamento e monitoração de dados que, depois de processados, geram informações para o suporte e para a tomada de decisões no ambiente de negócios.

HORA DE PRATICAR!

6. Ano: 2018 Banca: CESGRANRIO Órgão: TRANSPETRO Cargo: Analista de processo de negócio Questão: 54

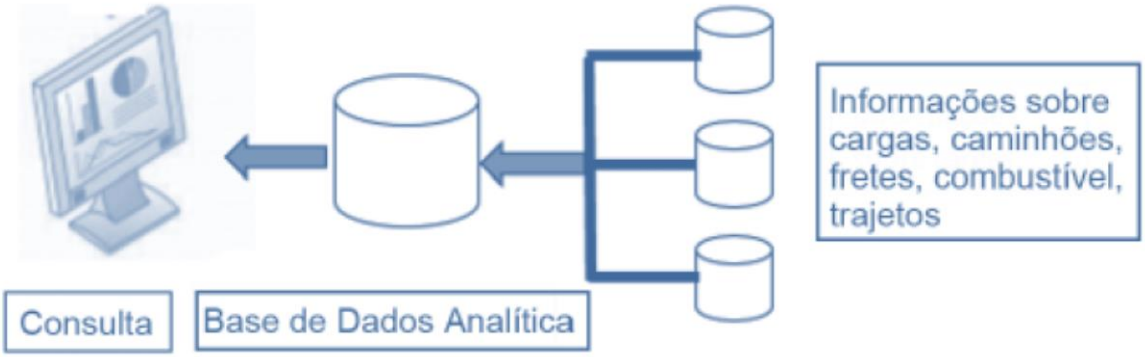
Determinada empresa de transporte possui uma frota de caminhões que movimenta diversos tipos de carga, tais como eletrônicos, brinquedos e eletrodomésticos. Um Sistema de Informações proprietário calcula detalhes financeiros e técnicos das viagens dessa frota. Os cálculos financeiros incluem, entre outros, custos de combustível, mão de obra e valor de frete. Os detalhes técnicos são inúmeros, como tipo e volume da carga, capacidade, consumo e velocidade dos caminhões, restrições dos trajetos, distâncias aos destinos e outros.

O sistema responde a perguntas, tais como:

i) dada uma especificação de carga, uma escala de entrega e preços de frete, quais caminhões e motoristas devem ser alocados para maximizar o lucro?

ii) qual conjunto (velocidade, trajeto) deve ser utilizado por determinado caminhão para otimizar o lucro e garantir as datas de entrega?

A Figura resume a configuração do sistema.



Adaptado de Laudon and Laudon. Management Information Systems: Managing the digital firm. 13 ed; Pearson 2014.

Com base na descrição acima, o tipo de Sistema de Informação utilizado por essa empresa é o

(A) CRM

(B) SIG

(C) Sistema Especialista

(D) Sistema de Suporte à Decisão

(E) Sistema de Processamento de Transações

7. BANCA: FCC ANO: 2016 ÓRGÃO: TRT - 14ª REGIÃO (RO E AC) PROVA: TÉCNICO JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

[35] Quando uma empresa utiliza Data Warehouse (DW) é necessário carregá-lo para permitir a análise comercial. Para isso, os dados de um ou mais sistemas devem ser extraídos e copiados para o DW em um processo conhecido como

a) ERP.

b) BI.

c) CRM.

d) ETL.

e) Data Mart.

8. ANO: 2013 BANCA: ESAF ÓRGÃO: DNIT PROVA: ANALISTA ADMINISTRATIVO - TECNOLOGIA DA INFORMAÇÃO

O componente final do processo de Business Intelligence é

A Business balance management (BBM).

B Executive office team (EOT).

C Business performance management (BPM).

D Priority statement board (PSB).

E Business advisory management (BAM).

9. ANO: 2010 BANCA: ESAF ÓRGÃO: MPOG PROVA: ANALISTA - TECNOLOGIA DA INFORMAÇÃO

BI – Business Intelligence

A é uma técnica de otimização da árvore de decisão.

B é um método de formação avançada de gestores.

C compreende ferramentas de análise de dados para otimizar os processos produtivos de uma empresa.

D são técnicas, métodos e ferramentas para mineração de dados na área de negócios de uma empresa.

E são técnicas, métodos e ferramentas de análise de dados para subsidiar processos de decisão de uma empresa.

10. ANO: 2015 BANCA: FCC ÓRGÃO: CNMP PROVA: ANALISTA DO CNMP - DESENVOLVIMENTO DE SISTEMAS

Soluções informatizadas de Business Intelligence (BI) geralmente contêm sistemas que podem ser de diversos tipos, dependendo do objetivo das análises e do perfil do usuário, como:

A Online Analytical Processing (OLAP), também conhecidos como sintéticos, que baseiam-se em transações, como: Sistemas Contábeis; Aplicações de Cadastro; Sistemas de Compra, Estoque, Inventário; ERPs; CRMs.

B Decision Support Systems (DSS) ou Sistemas de Apoio a Decisão, voltados para profissionais que atuam no nível estratégico das empresas, como diretoria e presidência. Oferecem, para tanto, um conjunto de indicadores chave de desempenho como o CMMI.

C Management Information Systems (MIS) ou Sistemas de Informações Gerenciais, que permitem análises mais profundas, com a realização de simulações de cenários. Por vezes, utilizam-se de ferramentas de Data Mining para identificação de cruzamentos não triviais. São utilizados por analistas de negócio no nível tático.

D Online Transactional Processing (OLTP) ou Sistemas transacionais, que fornecem subsídio para tomadas de decisão a partir de análises realizadas sobre bases de dados históricas, por vezes com milhões de registros a serem totalizados.

E Executive Information Systems (EIS) ou Sistemas de Informações Executivas, que são baseados em relatórios analíticos, normalmente utilizados por usuários de nível operacional.

11. ANO: 2010 BANCA: FCC ÓRGÃO: TCE-SP PROVA: AGENTE DA FISCALIZAÇÃO FINANCEIRA - CONHECIMENTOS BÁSICOS

Os conceitos de inteligência empresarial ou organizacional estão intimamente relacionados com o PETI que considera

A o planejamento de sistemas de informação, apenas.

B o planejamento de sistemas de informação e conhecimentos, apenas.

C a informática e os conhecimentos, apenas.

D a informática, apenas.

E o planejamento de sistemas de informação, conhecimentos e informática.

12. ANO: 2013 BANCA: FCC ÓRGÃO: MPE-CE PROVA: ANALISTA MINISTERIAL - CIÊNCIAS DA COMPUTAÇÃO

Em relação ao entendimento do significado do termo Business Intelligence (BI) e da solução que provê, a definição que **NÃO** é coerente com o termo Business Intelligence é a que

A consiste em uma metodologia que fornece objetivos de negócios ligados a objetivos de TI, provendo métricas e modelos de maturidade para medir a sua eficácia e identificando as responsabilidades relacionadas dos donos dos processos de negócios e de TI.

B se refere à aplicação de técnicas analíticas para informações sobre condições de negócio no sentido de melhorá-las, de uma maneira automatizada, mas com a interpretação e respostas humanas, de forma a melhorar a tomada de decisões.

C reúne recursos que provêm a habilidade para que a pessoa certa receba a informação adequada e no momento correto para tomar a melhor decisão.

D consiste em um sistema de negócios que inclui uma estrutura de busca efetiva e acessível, acurada, em tempo real, com informações e relatórios que permitam aos líderes das áreas de negócio se manterem informados para tomar decisões.

E é uma solução fácil de dizer, mas difícil de fazer corretamente pois envolve mudanças na forma como a organização conduz uma busca efetiva, bem como, a necessidade de se possuir uma base de dados de qualidade para que se possa tomar ações com o objetivo de otimizar a performance corporativa.

13. ANO: 2012 BANCA: FCC ÓRGÃO: TST PROVA: ANALISTA JUDICIÁRIO - ANALISTA DE SISTEMAS

Em Business Intelligence (BI), as consultas de dados que NÃO estão disponíveis em relatórios periódicos, ou seja, consultas criadas sob demanda especificamente para um conteúdo, layout ou cálculo, agilizando ou facilitando a tomada de decisão, são chamadas de consultas

A evolutivas.

B multidimensionais.

C single shot.

D data mining.

E ad hoc.

DATA WAREHOUSE

14. FGV - Aud Est (CGE SC)/CGE SC/Ciências da Computação/2023 - TI - Banco de Dados - Conceitos e Fases de Projeto e Modelagem de Dados

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*.

15. FGV - Aud Est (CGE SC)/CGE SC/Ciências da Computação/2023 - TI

Um Sistema de Gerenciamento de Banco de Dados Relacional (SGBDR) permite manipular bancos de dados sobre uma camada de software, dispondo os dados em formato de tabelas ao invés de arquivos em pastas. Para servir à finalidade de aplicações transacionais, as boas práticas apontam o uso do conceito de normalização.

Assinale a afirmativa ***incorreta*** em relação às vantagens da normalização.

a) Melhora a performance de consultas analíticas em um *Data Warehouse*, pois o modelo dimensional estrela depende da normalização.

b) A metodologia em etapas (1FN -> 2FN -> 3FN) facilita o processo de eliminação de dependências funcionais.

c) Diminui o esforço computacional de operações de UPDATE, pois as atualizações ocorrem apenas onde necessário.

d) Economiza espaço em disco, pois evita repetições de dados.

e) Melhora o desempenho geral sistêmico de uma aplicação, sobretudo com grandes volumes de dados, pois as transações ocorrem sob escopos específicos.

16. FCC - Auditor Fiscal (SEFAZ-BA)/Administração, Finanças e Controle Interno/2019

Nos sistemas transacionais, os dados sofrem diversas alterações como inclusão, alteração e exclusão. Antes de serem carregados no ambiente de um Data Warehouse, os dados são filtrados e limpos, de forma a gerarem informação útil. Após esta etapa, esses dados

a) ficam disponíveis para a mineração em tempo real, pois tais dados são constantemente atualizados a partir da chave de tempo que indica o dia em que foram extraídos dos sistemas transacionais.

b) podem sofrer operações de consulta, mas, devido a sua não volatilidade, não podem ser alterados, não havendo necessidade de bloqueio por concorrência de usuários ao seu acesso.

c) são reunidos a partir de diversas fontes de dados, o que facilita muito o trabalho do analista, embora este tenha que lidar com a grande redundância das informações.

d) ficam ordenados pela data da extração do sistema transacional, sendo necessárias técnicas de data mining para fazer a sua recuperação orientada por assunto.

e) são classificados somente pelo assunto principal de interesse da organização. Por exemplo, em uma organização de arrecadação de impostos, os dados são organizados pelo cadastro de contribuintes que possuem impostos a recolher.

17. FCC - Auditor Fiscal (SEFAZ-BA)/Tecnologia da Informação/2019

Um Auditor da SEFAZ-BA, observando as necessidades da organização, propôs um Data Warehouse (DW) com as seguintes características:

− na camada de dados resumidos ficam os dados que fluem do armazenamento operacional, que são resumidos na forma de campos que possam ser utilizados pelos gestores de forma apropriada.

− na segunda camada, ou no nível de dados históricos, ficam todos os detalhes vindos do ambiente operacional, em que se concentram grandes volumes de dados.

Com esta organização, os tipos de consulta analítica de maior frequência acessariam os dados resumidos, mais compactos e de mais fácil acesso e, em situações em que seja necessário um maior nível de detalhe, utilizar-se-iam os dados históricos.

O Auditor propôs um DW

a) que oferece maior nível de detalhes, ou seja, alto nível de granularidade.

b) que oferece menor nível de detalhes, ou seja, baixo nível de granularidade.

c) com nível duplo de granularidade.

d) com OLAP integrado.

e) com data marts geminados.

18. FEPESE - Analista (CELESC)/Sistemas/Desenvolvimento/2019

Assinale a alternativa que apresenta características de um Data Warehouse.

a) Orientado por assunto, integrado, volátil, variável no tempo.

b) Orientado por assunto, integrado, volátil, invariante no tempo.

c) Orientado por assunto, integrado, não volátil, variável no tempo.

d) Orientado por departamento, integrado, volátil, invariante no tempo.

e) Orientado por departamento, integrado, volátil, variável no tempo.

19. CCV UFC - Técnico (UFC)/Tecnologia da Informação/"Sem Especialidade"/2019

É considerado um conjunto de informações associadas um sistema de apoio a decisão, de forma que suas operações são prioritariamente de consultas para a obtenção de dados para embasar a tomada de decisão.

Marque o item que está associado ao conceito descrito.

a) data warehousing

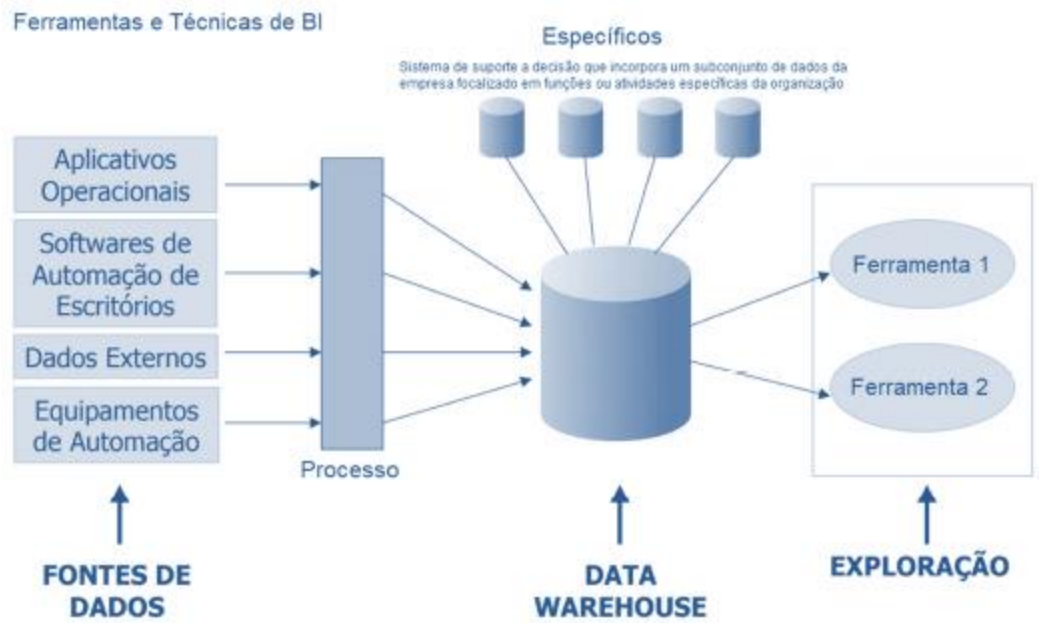
b) mineração de dados (data mining)

c) extração, transformação e carga (ETL)

d) processamento analítico on-line (OLAP)

e) processamento de transações on-line (OLTP)

20. FCC - Analista de Tecnologia da Informação (SANASA)/Análise e Desenvolvimento/2019



O sistema de suporte a decisão representado em cada um dos cilindros do conjunto denominado Específicos, na imagem, é um

a) Catálogo de Metadados.

b) Schema.

c) Drill.

d) OLTP.

e) Data Mart.

HORA DE PRATICAR!

21. VUNESP - Analista de Tecnologia da Informação (Campinas)/2019

No contexto de armazéns de dados (data warehouse), a área intermediária na qual os dados coletados pelo processo de ETL são armazenados antes de serem processados e transportados para o seu destino é chamada de

a) cubo OLAP.

b) dicionário de dados.

c) staging.

d) data vault.

e) data mart.

HORA DE PRATICAR!

22. IBADE - Técnico em Informática (Pref Jaru)/2019

Todo dado é relevante. Baseado nessa premissa algumas empresas acumulam e mantém grandes quantidades de dados que, organizados e analisados, fornecem informações relevantes para os processos de decisão. A esses “depósitos” de dados chamamos Data:

a) Check.

b) Mart.

c) Pool.

d) Mining.

e) Warehouse.

23. CEBRASPE (CESPE) - Assistente Judiciário (TJ AM)/Programador/2019

Com relação a arquitetura e tecnologias de sistemas de informação, julgue o próximo item.

Data warehouse, o principal dispositivo de armazenamento de um computador, é formado pelo processador, pela entrada e pela saída de dados.

24. CEBRASPE (CESPE) - Assistente Judiciário (TJ AM)/Suporte ao Usuário de Informática/2019

A respeito de data warehouse e data mining, julgue o item que se segue.

Chamados de data mart, os servidores de apresentação de data warehouse permitem consultas.

25. DECEx - Curso de Formação de Oficiais do Quadro Complementar (EsFCEx)/Informática/2019/CA CFO-QC 2020

Nas palavras de LAUDON e LAUDON, em sistemas de informação, existem quatro tipos de sistemas que apoiam os diferentes níveis e tipos de decisão. Os Sistemas de Informações Gerenciais (SIG) fornecem resumos e relatórios de rotina com dados no nível de transação para a gerência de nível operacional e médio. Sistemas de Apoio à Decisão (SAD) fornecem ferramentas ou modelos analíticos para analisar grandes quantidades de dados, além de consultas interativas de apoio para gerentes de nível médio que enfrentam situações de decisões semiestruturadas. Sistemas de Apoio ao Executivo (SAE) são sistemas que fornecem à gerência sênior, envolvida em decisões não estruturadas, informações externas e resumos de alto nível. Sistemas de Apoio à Decisão em Grupo (SADG) são sistemas especializados que oferecem um ambiente eletrônico no qual gerentes e equipes podem coletivamente tomar decisões e formular soluções.

Considerando os conceitos do universo dos Sistemas de Apoio à Decisão (SAD), avalie as seguintes asserções e a relação proposta entre elas.

I. DATA WAREHOUSE é uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo, voláteis, para dar suporte ao processo de tomada de decisão. É um repositório de grande volume de dados tratados, objetivando levar informação a partir dos dados.

OBTIDOS POR MEIO DE

II. Ambientes heterogêneos, geralmente de bancos transacionais, utilizando técnica de ETL para extrair, transformar e carregar, com objetivo de processamento analítico, de modo a permitir a criatividade das pessoas envolvidas, também denominado de OLAP.

A respeito dessas asserções, assinale a alternativa correta.

a) As asserções I e II são proposições verdadeiras, e a II complementa a I.

b) As asserções I e II são proposições verdadeiras, mas a II não complementa a I.

c) A asserção I é uma proposição verdadeira, e a II é uma proposição falsa.

d) A asserção I é uma proposição falsa, e a II é uma proposição verdadeira.

e) As asserções I e II são proposições falsas.

26. Ano: 2018 Banca: FCC Órgão: DPE-AM Prova: Analista em Gestão Especializado de Defensoria - Analista de Banco de Dados

Uma das características fundamentais de um ambiente de data warehouse está em

a) servir como substituto aos bancos de dados operacionais de uma empresa, na eventualidade da ocorrência de problemas com tais bancos de dados.

b) ser de utilização exclusiva da área de aplicações financeiras das empresas.

c) proporcionar um ambiente que permita realizar análise dos negócios de uma empresa com base nos dados por ela armazenados.

d) ser de uso prioritário de funcionários responsáveis pela área de telemarketing das empresas.

e) armazenar apenas os dados mais atuais (máximo de 3 meses de criação), independentemente da área de atuação de cada empresa.

27. Ano: 2017 Banca: CESPE Órgão: SEDF Cargo: Analista de gestão educacional – Especialidade: tecnologia da informação Questão: 118

Com relação a data mining e data warehouse, julgue os itens que se seguem.

[118] Comparados aos bancos de dados transacionais, os data warehouses são mais voláteis porque, para que se mantenham consistentes, são atualizados em tempo real a cada atualização que ocorrer em qualquer uma das bases originais de dados que o componham.

28. Ano: 2018 Banca: CESGRANRIO Órgão: TRANSPETRO Cargo: ANALISTA DE PROCESSO DE NEGÓCIO Questão: 22

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

29. Ano: 2016 Banca: FCC Órgão: TRT-20 Cargo: Técnico de TI – Q. 38

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) tem como vantagem a criação de legamarts ou DMs legados que facilitam e agilizam futuras integrações.

(B) é conhecida como top down.

(C) permite um retorno de investimento apenas em longo prazo, ou seja, um slower pay back.

(D) tem como objetivo a construção de um sistema OLAP incremental a partir de DMs independentes.

(E) não garante padronização dos metadados, podendo criar inconsistências de dados entre os DMs.

30. ANO: 2015 BANCA: CESPE ÓRGÃO: MEC PROVA: TÉCNICO DE NÍVEL SUPERIOR - ADMINISTRADOR DE DADOS

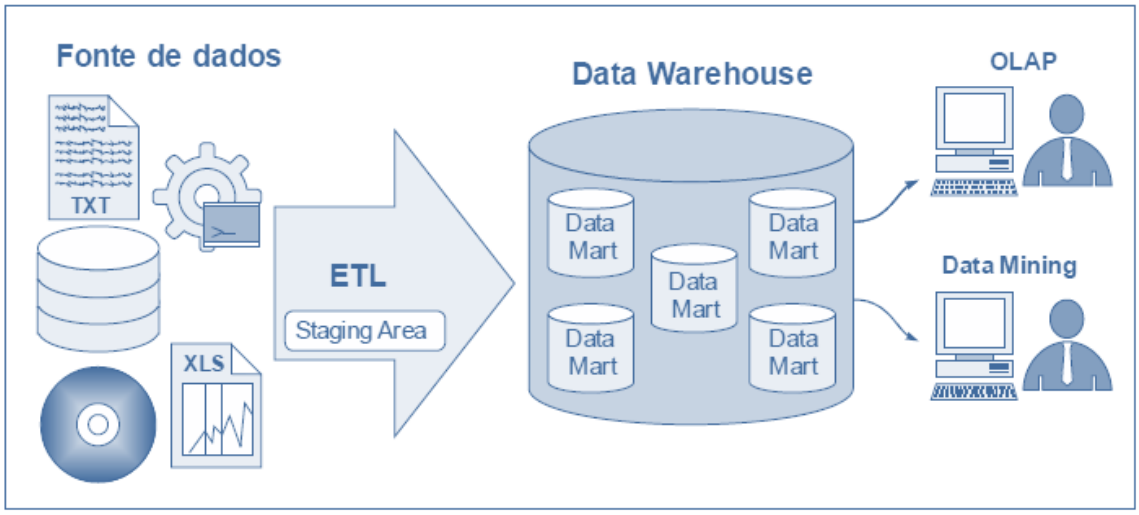
No que se refere a bancos de dados transacionais (OLTP) e a banco de dados analíticos (OLAP), julgue os itens que se seguem.

[1] Para melhor manter o controle sobre identificadores de registro de ambientes de data warehouse (armazém de dados), em geral recomenda-se a geração de chaves substitutas (surrogate keys). Assim, cada junção entre as tabelas de dimensão e tabelas fato em um ambiente de data warehouse deve se basear nessas chaves substitutas, e não nas chaves naturais existentes.

HORA DE PRATICAR!

31. ANO: 2015 BANCA: FCC ÓRGÃO: TRT - 4ª REGIÃO (RS) PROVA: ANALISTA JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

Considere a arquitetura geral de um sistema de BI - Business Intelligence mostrada na figura abaixo.



Nesta arquitetura

A Data Mining se refere ao processo que, na construção do Data Warehouse, é utilizado para composição de análises e relatórios, armazenando dados descritivos e qualificando a respectiva métrica associada.

B Data Marts representam áreas de armazenamento intermediário criadas a partir do processo de ETL. Auxiliam na transição dos dados das fontes OLTP para o destino final no Data Warehouse.

C OLAP é um subconjunto de informações extraído do Data Warehouse que pode ser identificado por assuntos ou departamentos específicos. Utiliza uma modelagem multidimensional conhecida como modelo estrela.

D os dados armazenados no Data Warehouse são integrados na base única mantendo as convenções de nomes, valores de variáveis e outros atributos físicos de dados como foram obtidos das bases de dados originais.

E o Data Warehouse não é volátil, permite apenas a carga inicial dos dados e consultas a estes dados. Além disso, os dados nele armazenados são precisos em relação ao tempo, não podendo ser atualizados.

32. ANO: 2015 BANCA: FCC ÓRGÃO: TRT - 3ª REGIÃO (MG) PROVA: TÉCNICO JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

Um técnico de TI precisa utilizar um subconjunto de dados de um Data Warehouse direcionado à área administrativa de um Tribunal. Esses dados serão armazenados em um banco de dado modelado multidimensionalmente, que será criado capturando-se dados diretamente de sistemas transacionais, buscando as informações relevantes para os processos de negócio da área administrativa. Esse banco de dados será um

A Big Data.

B Data Mart.

C OLAP.

D MOLAP.

E Data Mining.

33. ANO: 2014 BANCA: FCC ÓRGÃO: TCE-RS PROVA: AUDITOR PÚBLICO EXTERNO - TÉCNICO EM PROCESSAMENTO DE DADOS

A granularidade de dados é uma questão crítica no projeto de um Data Warehouse (DW), pois afeta o volume de dados que reside no DW e, ao mesmo tempo, afeta o tipo de consulta que pode ser atendida. Considere:

I. Quanto mais detalhe existir, mais baixo será o nível de granularidade. Quanto menos detalhe existir, mais alto será o nível de granularidade.

II. Quando há um nível de granularidade muito alto, o espaço em disco e o número de índices necessários se tornam bem menores, mas há uma correspondente diminuição da possibilidade de utilização dos dados para atender a consultas detalhadas.

É correto afirmar que a afirmativa I

A é equivalente a: quanto menos detalhes há nos dados, menor é a granularidade, consequentemente, quanto mais detalhes existem, maior é a granularidade.

B e a afirmativa II estão corretas e coerentes em relação ao nível de granularidade, espaço em disco e tipos de consultas em um DW.

C está correta. A afirmativa II está incorreta, pois apresenta incoerência em relação ao nível de granularidade, espaço em disco e tipos de consultas em um DW.

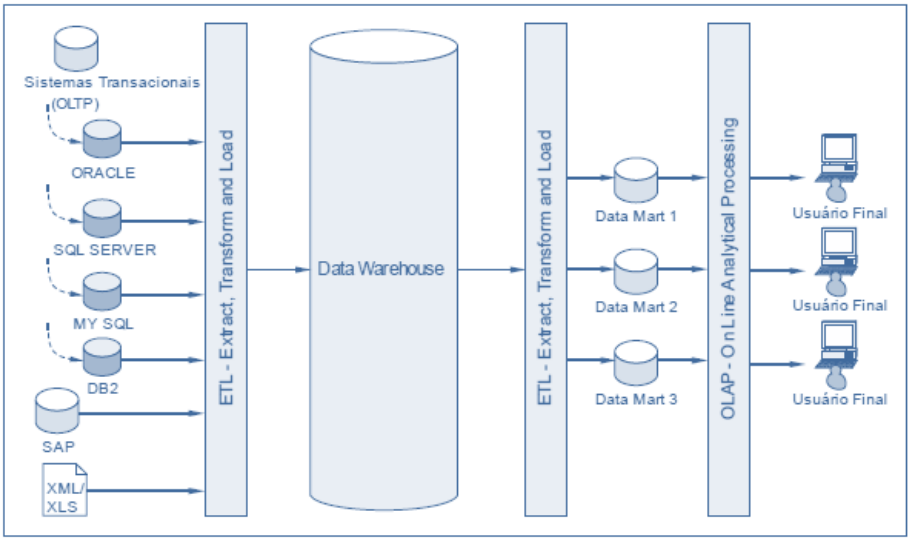
D e a afirmativa II estão incorretas. Ambas apresentam incoerência em relação ao nível de granularidade, espaço em disco e tipos de consultas em um DW.

E está incorreta. A afirmativa II está correta, pois é coerente em relação ao nível de granularidade, espaço em disco e tipos de consultas em um DW.

HORA DE PRATICAR!

34. ANO: 2015 BANCA: FCC ÓRGÃO: CNMP PROVA: ANALISTA DO CNMP - DESENVOLVIMENTO DE SISTEMAS

Considere que a equipe de Analistas de Desenvolvimento de Sistemas do CNMP está projetando a arquitetura para o Data Warehouse (DW) da instituição, conforme mostra a figura abaixo:



correto afirmar que esta arquitetura

A é bottom-up, pois primeiro a equipe cria um DW e depois parte para a segmentação, ou seja, divide o DW em áreas menores gerando pequenos bancos orientados por assuntos aos departamentos.

B é bottom-up. Permite um rápido desenvolvimento, pois a construção dos Data Marts é altamente direcionada. Normalmente um Data Mart pode ser colocado em produção em um período de 2 a 3 meses.

C é top-down. A partir do DW são extraídos os dados e metadados para os Data Marts. Nos Data Marts as informações estão em maior nível de sumarização e, normalmente, não apresentam o nível histórico encontrado no DW.

D é top-down, pois possui um retorno de investimento muito rápido ou um faster pay back. O propósito desta arquitetura é a construção de um DW incremental a partir de Data Marts independentes.

E é bottom-up. Garante a existência de um único conjunto de aplicações para ETL, ou seja, extração, limpeza e integração dos dados, embora os processos de manutenção e monitoração fiquem descentralizados.

35. ANO: 2010 BANCA: FCC ÓRGÃO: TRT - 22ª REGIÃO (PI) PROVA: ANALISTA JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

No âmbito dos DWs, uma outra concepção do ODS (Staging Area) está sendo estabelecida por alguns autores. Trata-se de

A OLAP.

B Drill throught.

C ETL.

D Data Mining.

E Dynamic Data Storage.

MODELAGEM DIMENSIONAL

36. Analista Legislativo (ALAP)/Atividade de Tecnologia da Informação/Desenvolvedor de Banco de Dados/2020

Duas definições de estruturas de dados estão determinadas para um projeto de datamart de uma loja de varejo: uma delas (tabela A) contém a data da venda, a identificação do produto vendido, a quantidade vendida do produto no dia e o valor total das vendas do produto no dia; a outra (tabela B) contém a identificação do produto, nome do produto, marca, modelo, unidade de medida de peso, largura, altura e profundidade da embalagem.

Considerando os conceitos de modelagem multidimensional de data warehouse, as tabelas A e B são, respectivamente:

a) Query e Réplica

b) Fato e Dimensão

c) Dimensão e Réplica

d) Fato e ETL

e) ETL e Query

HORA DE PRATICAR!

37. Analista de Tecnologia da Informação (EBSERH HC-UFU)/2020

Em um banco de dados de um data warehouse baseado em modelagem multidimensional, encontrou-se três tabelas: Venda, Vendedor e Produto, entre outras. Um subconjunto da estrutura referente a este trecho do modelo é o seguinte:

As tabelas **Venda**, **Vendedor** e **Produto** são classificadas, respectivamente, como:

a) fato, fato e dimensão.

b) fato, dimensão e dimensão.

c) dimensão, fato e fato.

d) dimensão, fato e dimensão.

e) dimensão, dimensão e fato.

HORA DE PRATICAR!

38. (Ministério da Economia – Especialista em Ciência de Dados - 2020) Julgue os itens a seguir, relativos a conceitos de modelagem dimensional.

Em um processo de modelagem dimensional, a operação de merge/purge agrega informações das dimensões para diminuir a tabela de fatos.

HORA DE PRATICAR!

39. CEBRASPE (CESPE) - Profissional de Tecnologia da Informação (ME)/Atividades Técnicas de Complexidade Gerencial, de Tecnologia da Informação e de Engenharia Sênior/Desenvolvimento de Software/2020

No que se refere a conceitos de modelagem de dados relacional e dimensional, julgue o item a seguir.

Na modelagem dimensional, a tabela fatos armazena as dimensões e os detalhes dos valores descritivos do armazém de dados.

HORA DE PRATICAR!

40. Ano: 2019 Banca: CESPE Órgão: SEFAZ-RS Prova: Auditor Assunto: Modelagem dimensional

Com relação aos modelos de dados multidimensionais, assinale a opção correta.

A A principal característica da tabela de fatos é a ausência de dados redundantes, o que melhora o desempenho nas consultas.

B Esses modelos são cubos de dados, sendo cada cubo representado por uma única tupla com vários atributos.

C Esses modelos proporcionam visões hierárquicas, ou seja, exibição roll-up ou drill-down.

D Os modelos de dados multidimensionais dão ênfase à coleta e às transações de dados.

E Esses modelos não utilizam processos de transferência de dados, mas sim acessos nativos do próprio SGBD utilizado.

HORA DE PRATICAR!

41. FCC - Auditor Fiscal (SEFAZ-BA)/Tecnologia da Informação/2019

Suponha que uma Auditora Fiscal da área de TI tenha proposto a seguinte modelagem multidimensional para a SEFAZ−BA:

Fato central: Controle de Receitas e Despesas

A partir do Fato Controle de Receitas e Despesas:

- Dimensão Tempo
- Dimensão Receitas
  - Dentro da dimensão Receitas: Dimensão Receitas de Impostos
  - Dentro da dimensão Receitas: Dimensão Receitas de Taxas
- Dimensão Despesas
  - Dentro da dimensão Despesas: Dimensão Tipo de Despesa
- Dimensão Cidade
  - Dentro da dimensão Cidade: Dimensão NF-e

A modelagem multidimensional proposta

a) é o resultado da decomposição de mais de uma dimensão que possui hierarquias entre seus membros, caracterizando o modelo snowflake, a partir de um fato central.

b) tem como característica um fato central, a partir do qual estão dispostas as dimensões que dele participam, em um formato simétrico, característico do modelo star.

c) parte de um elemento central, denominado pivot, a partir do qual são realizadas operações OLAP como roll up, em que busca-se aumentar o nível de detalhe ou diminuir a granularidade da consulta.

d) possui um fato central, a partir do qual estão dispostas as dimensões que dele participam e seus membros, sob uma única estrutura hierárquica, facilitando a inclusão de dados por digitação nas tabelas do DW.

e) não é um modelo normalizado, por isso evita a redundância de valores textuais em cada uma das tabelas, representadas pelas dimensões denominadas dimension tables.

42. VUNESP - Programador (CM Piracicaba)/2019

No modelo dimensional, composto por tabelas fato e tabelas dimensão,

a) as tabelas fato não admitem chaves estrangeiras.

b) as tabelas dimensão comportam apenas atributos multivalorados.

c) o relacionamento de cada tabela dimensão para a tabela fato é de “um para muitos”.

d) nas tabelas dimensão há apenas atributos numéricos.

e) não há atributos numéricos nas tabelas fato.

43. CEBRASPE (CESPE) - Analista Judiciário (TJ AM)/Analista de Sistemas/2019

A respeito de bancos de dados relacionais, julgue o item a seguir.

O esquema multidimensional estrela de data warehouse é composto por uma tabela de fatos associada com uma única tabela para cada dimensão.

44. IDECAN - Professor de Ensino Básico, Técnico e Tecnológico (IF Baiano)/Informática/2019

A modelagem de data warehouses pode ser feita seguindo diferentes esquemas. Sobre esse tópico, analise as afirmativas:

I. No esquema estrela, os dados são organizados em uma tabela de dimensão e muitas tabelas de fatos.

II. O esquema floco de neve é uma variação do esquema estrela, onde algumas tabelas de fatos são normalizadas, dividindo, assim, os dados em tabelas adicionais.

III. Quando várias tabelas de fatos compartilham tabelas de dimensão, temos o chamado esquema de constelação de fatos ou galáxia, pois podem ser considerados como coleções de estrelas.

a) se somente as afirmativas I e II estiverem corretas.

b) se somente as afirmativas II e III estiverem corretas.

c) se somente a afirmativa I estiver correta

d) se somente a afirmativa II estiver correta.

e) se somente a afirmativa III estiver correta.

HORA DE PRATICAR!

45. FUNDATEC - Auditor Fiscal da Receita Municipal (Pref POA)/2019/"Sem Edição"

A questão baseia-se na Figura 7, que mostra uma modelagem multidimensional, elaborada no Microsoft Access 365 (MS Access 365).

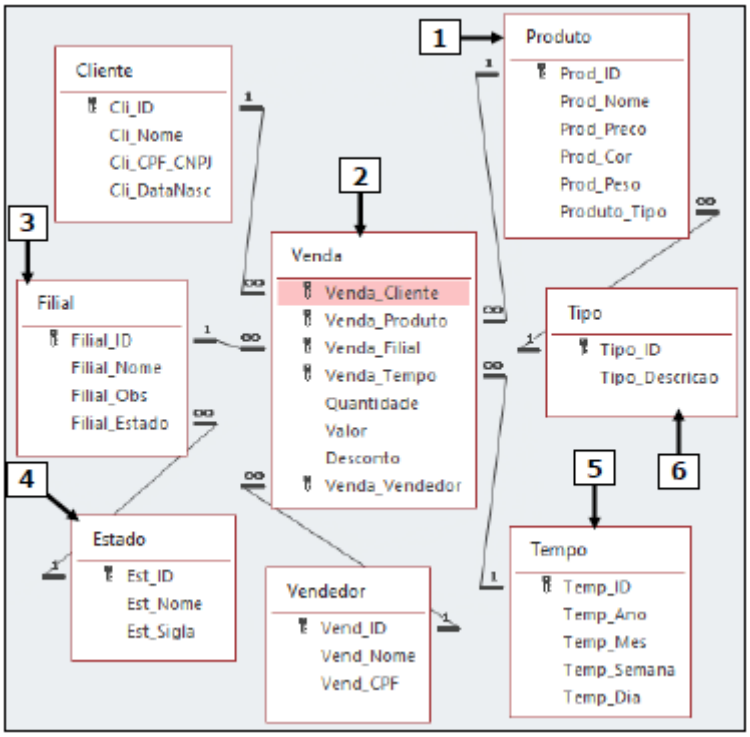


Figura 7 – Modelagem multidimensional

Após observar a Figura 7, analise as seguintes assertivas:

I. A tabela fato, dessa modelagem, é "Venda", apontada pela seta nº 2.

II. As tabelas "Produto", "Filial", "Estado", "Tempo" e "Tipo", apontados, respectivamente pelas setas nº 1, 3, 4, 5 e 6, são tabelas "Dimensão".

III. O esquema multidimensional exibido na Figura 7 é chamado de esquema "Estrela".

Quais estão corretas?

a) Apenas I.

b) Apenas III.

c) Apenas I e II.

d) Apenas II e III.

e) I, II e III.

46. CEBRASPE (CESPE) - Auditor Municipal de Controle Interno (CGM João Pessoa)/Tecnologia da Informação/Desenvolvimento de Sistemas/2018

Com relação à modelagem dimensional e à otimização de bases de dados para business intelligence, julgue o item subsequente.

O modelo snowflake acrescenta graus de normalização às tabelas de dimensões, eliminando redundâncias; em termos de eficiência na obtenção de informações, seu desempenho é melhor que o do modelo estrela, o qual, apesar de possuir um único fato, possui tamanho maior que o do snowflake, considerando-se a desnormalização das tabelas de dimensões.

47. CEBRASPE (CESPE) - Auditor Municipal de Controle Interno (CGM João Pessoa)/Tecnologia da Informação/Desenvolvimento de Sistemas/2018

Com relação à modelagem dimensional e à otimização de bases de dados para business intelligence, julgue o item subsequente.

Na modelagem multidimensional utilizada em data warehouses para se prover melhor desempenho, a tabela fato central deve relacionar-se às suas dimensões por meio da chave primária oriunda da fonte de dados original. O valor dessa chave deve ser idêntico ao da fonte, para que tenha valor semântico e garanta que o histórico das transações seja mantido.

48. Ano: 2018 Banca: CESPE Órgão: TCM-BA Cargo: Auditor de Contas Questão: 11

Acerca de modelagem dimensional, assinale a opção correta.

A As granularidades fundamentais para classificar todas as tabelas fato de um modelo dimensional são: transacional, snapshot periódico e snapshot acumulado.

B Os fatos e dimensões não são tabelas do banco de dados, pois, no modelo dimensional, são componentes do cubo de um data warehouse.

C No modelo estrela, as dimensões são normalizadas para tornar mais ágeis as consultas analíticas.

D O modelo floco de neve (SnowFlake) aumenta o espaço de armazenamento dos dados dimensionais, pois acrescenta várias tabelas ao modelo, todavia torna mais simples a navegação por software que utilizarão o banco de dados.

E Os códigos e as descrições associadas, usadas como nomes de colunas em relatórios e como filtros em consultas, não devem

ser gravados em tabelas dimensionais.

49. Ano: 2018 Banca: Cesgranrio Órgão: Petrobras Cargo: Analista de Processo de Negócio Questão: 44

Ao construir um modelo de dados para um data warehouse de sua empresa, um desenvolvedor viu-se às voltas com três tabelas relacionais: venda, cliente e vendedor. Ao fazer uma transformação para o modelo estrela, ele deve organizar:

(A) venda, como tabela fato; cliente e vendedor, como tabelas dimensão

(B) cliente e vendedor, como tabelas fato; venda, como tabela dimensão

(C) cliente, como tabela fato; venda e vendedor, como tabelas dimensão

(D) vendedor e venda, como tabelas fato; cliente, como tabela dimensão

(E) vendedor, como tabela fato; cliente e venda, como tabelas dimensão

50. Ano: 2017 Banca: CESPE Órgão: SEDF Cargo: Analista de gestão educacional – Especialidade: tecnologia da informação

Com relação aos conceitos de modelagem multidimensional de dados para inteligência computacional, julgue os seguintes itens.

[104] Diferentemente da estrutura relacional, a estrutura multidimensional oferece baixa redundância de dados e suporte a normalização até a segunda forma normal.

[106] Ao se modelar uma tabela-fato, deve-se considerar que a chave primária é composta e que a dimensão tempo sempre será parte integrante dessa chave.

51. BANCA:CESPE | CEBRASPE - ANO: 2016 - CONCURSO:FUNPRESP – CARGO 8: ESPECIALISTA - ÁREA: TECNOLOGIA DA INFORMAÇÃO (TI)

Acerca dos modelos de dados relacional e dimensional em engenharia de software, julgue os itens que se seguem.

63 Na modelagem dimensional, as tabelas dimensão estão menos sujeitas ao processo de desnormalização que as tabelas fato.

64 Em um modelo de dados relacional, a integridade referencial assegura que os valores dos campos presentes na chave estrangeira apareçam na chave primária da mesma tabela, a fim de garantir a integridade dos dados.

52. ANO: 2015 BANCA: CESPE ÓRGÃO: MEC PROVA: TÉCNICO DE NÍVEL SUPERIOR - ANALISTA DE SISTEMAS

Com relação aos passos do processo de projeto de bancos de dados e de modelagem de dados relacional e dimensional, julgue os itens subsequentes.

[1] Na modelagem dimensional, implementada em sistemas de data warehouse, o esquema snowflake caracteriza-se por possuir diversas tabelas de fatos e de dimensões, sendo estas últimas organizadas hierarquicamente na terceira forma normal (3FN).

53. Ano: 2018 Banca: FCC Órgão: SABESP Cargo: Analista de Gestão Área: Tecnologia da Informação Questão: 42

Um Analista está trabalhando em um Data Warehouse – DW que utiliza no centro do modelo uma única tabela que armazena as métricas e as chaves para as tabelas ao seu redor (que descrevem os dados que estão na tabela central) às quais está ligada. O esquema de modelagem utilizado pelo DW, a denominação da tabela central e a denominação das tabelas periféricas são, respectivamente,

(A) floco de neve, base, granulares.

(B) estrela, fato, dimensões.

(C) constelação, fato, granulares.

(D) atomic, base, branches.

(E) anel, base, dimensões.

54. BANCA: FCC ANO: 2017 ÓRGÃO: DPE-RS PROVA: ANALISTA – BANCO DE DADOS

[42] Um dos modelos mais utilizados no projeto e implementação de um data warehouse é o modelo dimensional ou multidimensional. Em um modelo dimensional (composto por uma tabela fato e várias tabelas dimensão),

a) as tabelas dimensão devem conter apenas atributos do tipo literal.

b) a tabela fato tem uma cardinalidade de mapeamento de um para um com cada tabela dimensão.

c) a tabela fato deve conter atributos numéricos, visando proporcionar dados para uma análise de atividades da empresa.

d) há um número teórico mínimo de 3 e máximo de 15 tabelas dimensão.

e) as tabelas dimensão comportam um número máximo teórico de atributos.

55. BANCA: FCC ANO: 2017 ÓRGÃO: TRT - 24ª REGIÃO (MS) PROVA: ANALISTA JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

[42] Uma das formas de apresentação de um banco de dados multidimensional é através do modelo estrela. No centro de um modelo estrela encontra-se a tabela de

a) dimensão e, ao seu redor, as tabelas de fatos.

b) dimensão, cuja chave primária deve ser composta.

c) núcleo e, ao seu redor, as tabelas de nível.

d) fatos, cuja chave primária deve ser simples.

e) fatos e, ao seu redor, as tabelas de dimensões.

56. BANCA: FCC ANO: 2016 ÓRGÃO: TRT - 20ª REGIÃO (SE) PROVA: TÉCNICO JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

[37] 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.

57. BANCA: FCC ANO: 2016 ÓRGÃO: PREFEITURA DE TERESINA - PI PROVA: TÉCNICO DE NÍVEL SUPERIOR - ANALISTA DE SISTEMAS

[57] Em um Star Schema de um Data Warehouse – DW, a tabela Dimensão possui característica

a) descritiva dentro do DW. Ela qualifica as informações provenientes da tabela Fato; A tabela Fato possui característica quantitativa dentro do DW. A partir dela são extraídas as métricas que são cruzadas com os dados das Dimensões. Dimensões são ligadas entre si e qualquer uma delas se liga diretamente a tabela Fato. Os dados devem ser normalizados.

b) quantitativa dentro do DW. Ela quantifica as informações provenientes da tabela Fato; A tabela Fato possui característica descritiva dentro do DW. A partir dela são extraídas as nomenclaturas que são quantificadas com os dados das Dimensões. Dimensões são ligadas entre si. Os dados devem ser desnormalizados.

c) quantitativa dentro do DW. Ela quantifica as informações provenientes da tabela Fato; A tabela Fato possui característica descritiva dentro do DW. A partir dela são extraídas as nomenclaturas que são quantificadas com os dados das Dimensões. Dimensões não são ligadas entre si. Os dados devem ser normalizados.

d) quantitativa dentro do DW. Ela quantifica as informações provenientes da tabela Fato; A tabela Fato possui característica descritiva dentro do DW. A partir dela são extraídas as nomenclaturas que são quantificadas com os dados das Dimensões. Dimensões são ligadas entre si. Os dados devem ser normalizados.

e) descritiva dentro do DW. Ela qualifica as informações provenientes da tabela Fato; A tabela Fato possui característica quantitativa dentro do DW. A partir dela são extraídas as métricas que são cruzadas com os dados das Dimensões. Dimensões são ligadas diretamente a tabela Fato. Outra característica marcante é que os dados são desnormalizados.

58. BANCA: FCC ANO: 2016 ÓRGÃO: PREFEITURA DE TERESINA - PI PROVA: ANALISTA TECNOLÓGICO - ANALISTA DE SUPORTE TÉCNICO

[35] O modelo dimensional utilizado na modelagem de data warehouse tem como característica:

a) Todas as tabelas dimensão de um mesmo modelo devem possuir o mesmo número de atributos.

b) A tabela fato possui pelo menos 4 atributos numéricos, além das chaves estrangeiras.

c) Poder ter quantas tabelas dimensionais, quantas forem necessárias para representar o negócio sob análise.

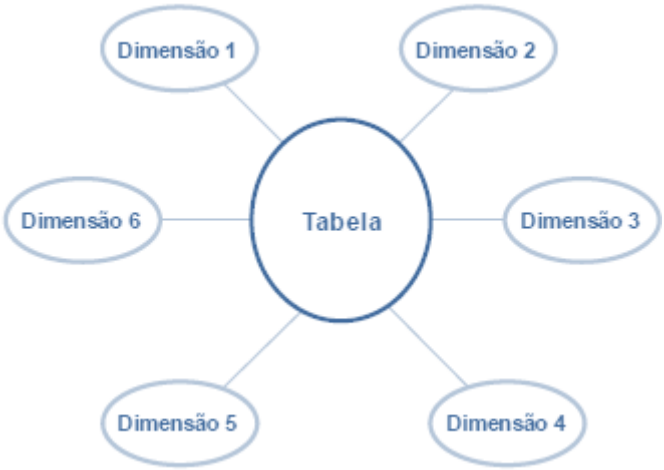
d) As tabelas dimensão não necessitam ter atributos que sirvam como chave primária.

e) A cardinalidade de relacionamento da tabela fato para as tabelas dimensão é de um para um.

HORA DE PRATICAR!

59. BANCA: FCC ANO: 2016 ÓRGÃO: ELETROBRAS-ELETROSUL PROVA: INFORMÁTICA

[57] Considere a figura abaixo que ilustra um modelo multidimensional na forma de modelo relacional em esquema estrela. Há uma tabela central que armazena as transações que são analisadas e ao seu redor há as tabelas look up, denominadas dimensões.



De acordo com o modelo estrela da figura e sua relação com um Data Warehouse, é correto afirmar:

a) Uma das candidatas à chave primária da tabela central, denominada star table, seria uma chave composta pelas chaves primárias de todas as dimensões.

b) A tabela fato armazena os indicadores que serão analisados e as chaves que caracterizam a transação. Cada dimensão registra uma entidade que caracteriza a transação e os seus atributos.

c) As dimensões devem conter todos os atributos associados à sua chave primária. Por causa disso, o modelo multidimensional estrela está na 3ª Forma Normal.

d) O modelo estrela é derivado do modelo snowflake, ou seja, é o resultado da aplicação da 1ª Forma Normal sobre as entidades dimensão.

e) Um Data Waherouse, por permitir a inclusão de dados por digitação, necessita da aplicação de normalização para garantir a unicidade de valores.

HORA DE PRATICAR!

60. BANCA: FCC ANO: 2016 ÓRGÃO: TRT - 23ª REGIÃO (MT) PROVA: ANALISTA JUDICIÁRIO – TECNOLOGIA DA INFORMAÇÃO

[34] Na abordagem Star Schema, usada para modelar data warehouses, os fatos são representados na tabela de fatos, que normalmente

a) é única em um diagrama e ocupa a posição central.

b) está ligada com cardinalidade n:m às tabelas de dimensão.

c) está ligada às tabelas de dimensão, que se relacionam entre si com cardinalidade 1:n.

d) tem chave primária formada independente das chaves estrangeiras das tabelas de dimensão.

e) está ligada a outras tabelas de fatos em um layout em forma de estrela.

HORA DE PRATICAR!

61. ANO: 2014 BANCA: FCC ÓRGÃO: TJ-AP PROVA: ANALISTA JUDICIÁRIO - BANCO DE DADOS - DBA

Os sistemas de Data Warehouse utilizam-se de um modelo de dados diferente dos bancos de dados tradicionais, que proporciona ganhos de desempenho nas consultas. Esse modelo é conhecido como modelagem

A dinâmica.

B dimensional.

C fixa.

D online.

E transacional.

HORA DE PRATICAR!

62. ANO: 2015 BANCA: FCC ÓRGÃO: TRT - 3ª REGIÃO (MG) PROVA: TÉCNICO JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

A modelagem multidimensional é utilizada especialmente para sumarizar e reestruturar dados e apresentá-los em visões que suportem a análise dos valores desses dados. Um modelo multidimensional é formado por dimensões, e por uma coleção de itens de dados composta de dados de medidas e de contexto, denominada

A schema.

B pivot.

C slice.

D fato.

E versão.

63. ANO: 2015 BANCA: FCC ÓRGÃO: TCM-GO PROVA: AUDITOR DE CONTROLE EXTERNO - INFORMÁTICA

Quando o modelo de dados multidimensionais começa a ser definido, elementos básicos de representação precisam ter sido estabelecidos, de modo a se criar um padrão de modelagem. Considere um modelo em que as dimensões e fatos são representados em tabelas, podendo haver múltiplas dimensões e múltiplas tabelas de fatos.

Ao modelar cada tabela _____I_____ devem ser considerados os seguintes pontos:

– A chave primária é composta, sendo um elemento da chave para cada dimensão;

– Cada elemento chave para a dimensão deve ser representado e descrito na tabela _____II_____ correspondente (para efetuar a junção);

– A dimensão tempo é sempre representada como parte da chave primária.

Deve haver uma tabela _____III_____ para cada dimensão do modelo, contendo

– Uma chave artificial (ou gerada) genérica;

– Uma coluna de descrição genérica para a dimensão;

– Colunas que permitam _____IV_____;

– Um indicador nível que indica o nível da hierarquia a que se refere a linha da tabela.

As lacunas de I a IV são corretas, e respectivamente, preenchidas com:

A dimensão – de fatos – de tempo – efetuar os filtros.

B dimensão – de fatos – de fatos – a junção com as tabelas de fatos.

C de fatos – de tempo – dimensão – sinalizar a presença de fatos para o período de tempo indicado na linha.

D de fatos – dimensão – dimensão – efetuar os filtros.

E de tempo – dimensão – de fatos – a junção com as tabelas de dimensão.

64. ANO: 2013 BANCA: FCC ÓRGÃO: MPE-MA PROVA: ANALISTA JUDICIÁRIO - BANCO DE DADOS

Na modelagem de um data warehouse, pode ser feito o snowflaking, que significa

A criptografar as tabelas fato e dimensão.

B normalizar as tabelas dimensão.

C excluir atributos do tipo binário.

D indexar as tabelas dimensão por todos seus atributos.

E duplicar a tabela fato.

HORA DE PRATICAR!

65. ANO: 2013 BANCA: FCC ÓRGÃO: MPE-MA PROVA: ANALISTA JUDICIÁRIO - BANCO DE DADOS

Na modelagem dimensional de um data warehouse voltado para vendas, o tipo de tabela fato que inclui pares de produtos adquiridos em uma mesma compra recebe a denominação de

A cesta de mercado.

B tabela de degeneração.

C data mart.

D outrigger.

E pacote de integralização.

HORA DE PRATICAR!

66. ANO: 2009 BANCA: FCC ÓRGÃO: TRT - 15ª REGIÃO (CAMPINAS-SP) PROVA: ANALISTA JUDICIÁRIO - TECNOLOGIA DA INFORMAÇÃO

No contexto OLAP:

I. As visões materializadas agregadas a partir de uma tabela de fatos podem ser identificadas exclusivamente pelo nível de agregação para cada dimensão.

II. Quando aplicada a configuração star schema as tabelas de fatos e as de dimensão são idênticas quanto à totalidade dos atributos que contêm e também quanto ao grau de granularidade.

III. O esquema snow flake é uma variação do star schema.

Está correto o que consta em

A I, somente.

B I e III, somente.

C II e III, somente.

D III, somente.

E I, II e III.

GABARITO

1. C
2. B
3. D
4. A
5. C
6. D
7. D
8. C
9. E
10. C
11. E
12. A
13. E
14. E
15. A
16. B
17. C
18. C
19. A
20. E
21. C
22. E
23. ERRADO
24. C
25. D
26. C
27. E
28. B
29. E
30. C
31. E
32. B
33. B
34. C
35. E
36. B
37. B
38. ERRADO
39. ERRADO
40. C
41. A
42. C
43. CERTO
44. E
45. C
46. ERRADO
47. ERRADO
48. A
49. A
50. ERRADO CERTO
51. ERRADO CERTO(?)
52. E
53. B
54. C
55. E
56. D
57. E
58. C
59. B
60. A
61. B
62. D
63. D
64. B
65. A
66. B

CONSIDERAÇÕES FINAIS

Chegamos ao final da nossa aula introdutória de *Business Intelligence* ou inteligência de negócios!

Outros assuntos dentro deste tema podem estar presentes nas próximas aulas.

Até a próxima!

Thiago Cavalcanti

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.