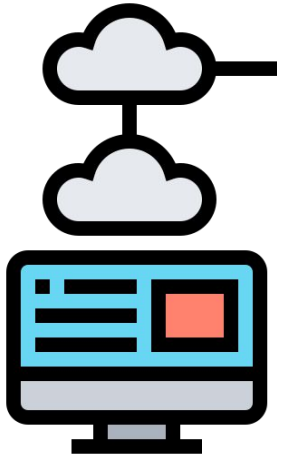
AMAZING DATAVERSE

uniesp
Centro Universitário

Wuldson Franco

- Análise e Desenvolvimento de Sistemas - UNIPÊ / Estacio SA
- MBA Engenharia e Administração de Dados - UNIESP
- Msc. Tecnologia da Informação - *Cursando*
- Engenheiro de Dados - A3Data
- Professor Membro - Artificial Inteligence e Data Analytics - AIDA
- Professor Membro - GDG Paraíba
- Professor - UNIESP

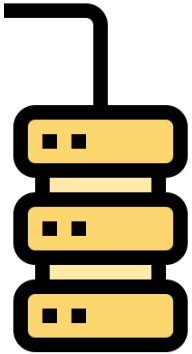
https://beacons.ai/wuldson



01 História dos Data Warehouse

02 Conceitos sobre DW

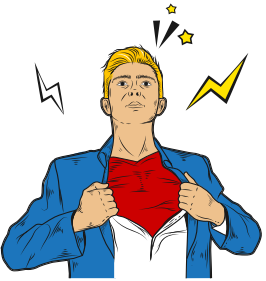
03 A importância do DW no BI



04 (Repositórios, Camadas, Níveis) do DW

05 Modelagem Dimensional Colaborativa
Fatos e Dimensões

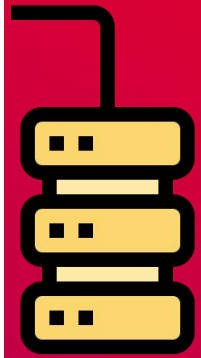
06 Práticas



uniesp
Centro Universitário

DW_BI - Práticas

- Arquitetura DW
- Modelagens - Dimensões e Fatos
- Criação do DW com Pentaho



uniesp
Centro Universitário

01 {

[História do DW]

Nos anos 1970 e 1980, os dados começaram a proliferar e as empresas precisaram encontrar uma maneira fácil para armazená-los e acessá-los.

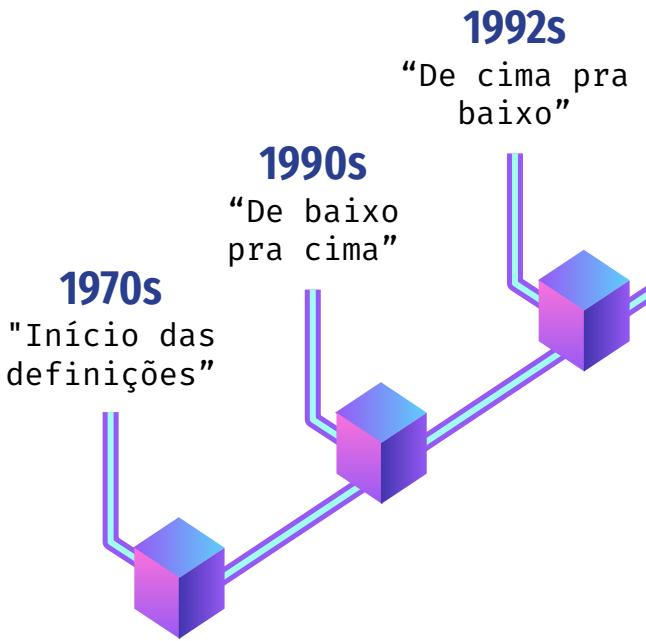
O cientista da computação Bill Inmon, que é considerado o pai do armazenamento de dados, começou a definir o conceito nos anos 1970 e é o responsável pela criação do termo **“data warehouse”.**

Em 1992, ele publicou um livro intitulado *Building the Data Warehouse*, tido como uma fonte fundamental para a tecnologia de armazenamento de dados.

02 { [Conceitos do DW]

Ele utilizava como conceito e definição de data warehouse algo como “de cima para baixo”, em que um repositório central é estabelecido primeiro e, depois, **data marts** – que contêm subconjuntos específicos de dados – são criados dentro desse repositório.

Ralph Kimball, especialista em tecnologia que publicou The Data Warehouse Toolkit em meados dos anos 1990, aplicou uma tática um pouco diferente ao conceito de data warehouse com sua abordagem “de baixo para cima”, na qual data marts individuais são desenvolvidos primeiro e, depois, integrados para criar um armazém.

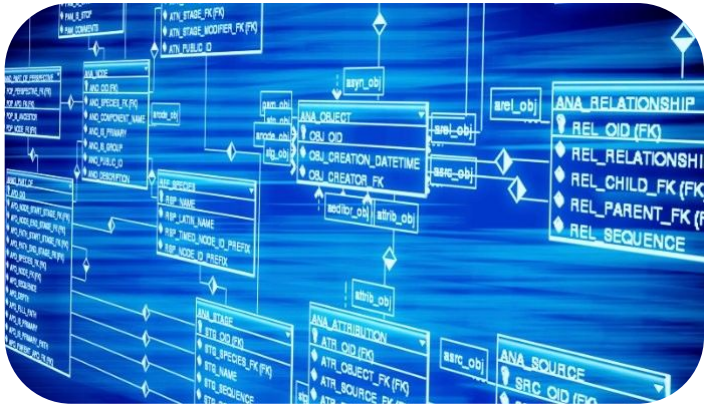


02 {

[Conceitos do DW]

- Entender as necessidades do negócio, bem como as realidades dos dados de origem subjacentes.
  - Você descobre os requisitos por meio de sessões com representantes de negócios(P.O.) para entender seus objetivos com base em ***indicadores-chave***, ***questões de negócios***, ***processos de tomada de decisão e necessidades analíticas de suporte***.

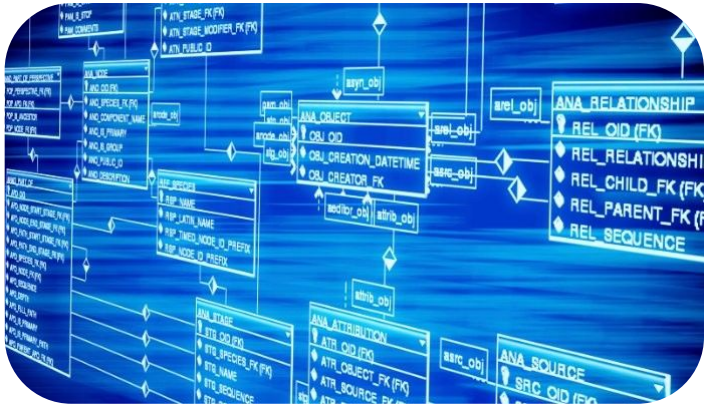
- As realidades dos dados são reveladas por meio de reuniões com ***especialistas em sistemas de origem e criação de perfis de dados de alto nível*** para avaliar as viabilidades dos dados.



02 {

[Conceitos do DW]

- Os **modelos dimensionais** devem ser projetados em colaboração com especialistas no assunto e representantes de governança de dados do negócio (*P.O de Negócio e P.O Técnico*).
- O modelador de dados está no comando, mas o modelo deve ser desenvolvido por meio de uma *série de workshops* altamente interativos com representantes comerciais.
- Esses workshops(*Negócio com Dados*) fornecem outra oportunidade para detalhar os requisitos com o **negócio**.
- Os modelos dimensionais **não** devem ser projetados isoladamente por pessoas que não entendem completamente o negócio e suas necessidades; **a colaboração é crítica!**



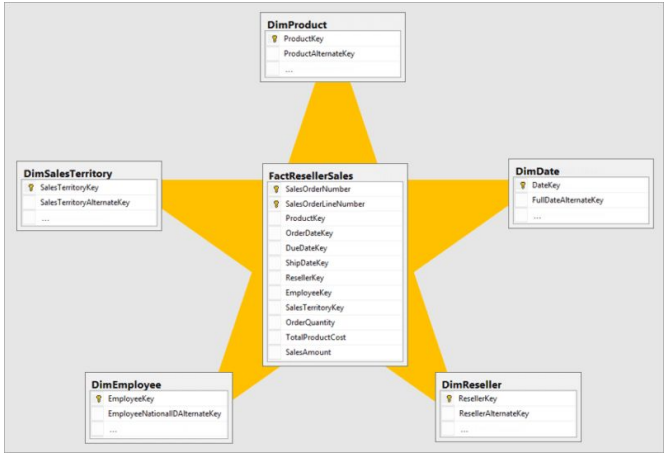
02.1 {

[Modelagem Dimensional]

- **O que é modelagem dimensional?**

  Técnica de projeto lógico normalmente usada para data warehouses que contrasta com a modelagem entidade-relacionamento.

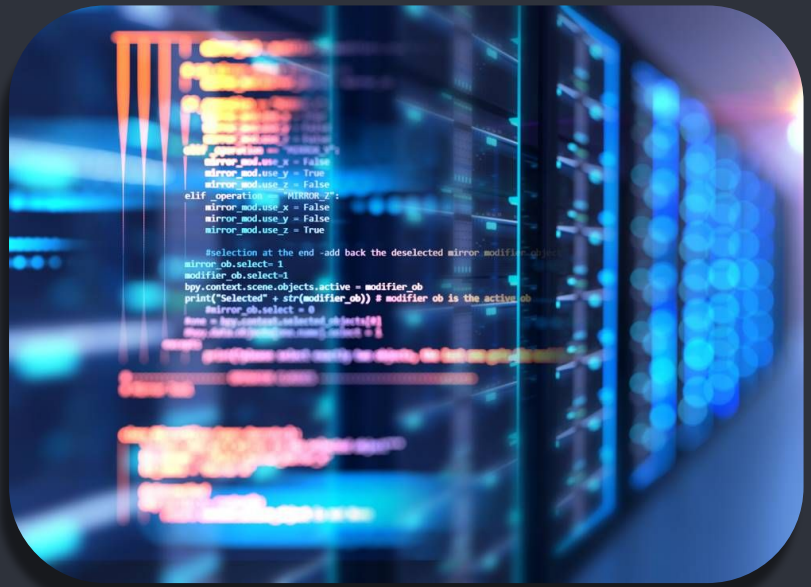
  - **Recuperação de dados mais rápida**
  - **Melhor compreensão dos processos de negócios**
  - **Flexível para mudar**



uniesp
Centro Universitário

03 {

[A Importância do Data Warehouse no Business Intelligence]



Por que alguém precisa de um Data Warehouse (DW)?

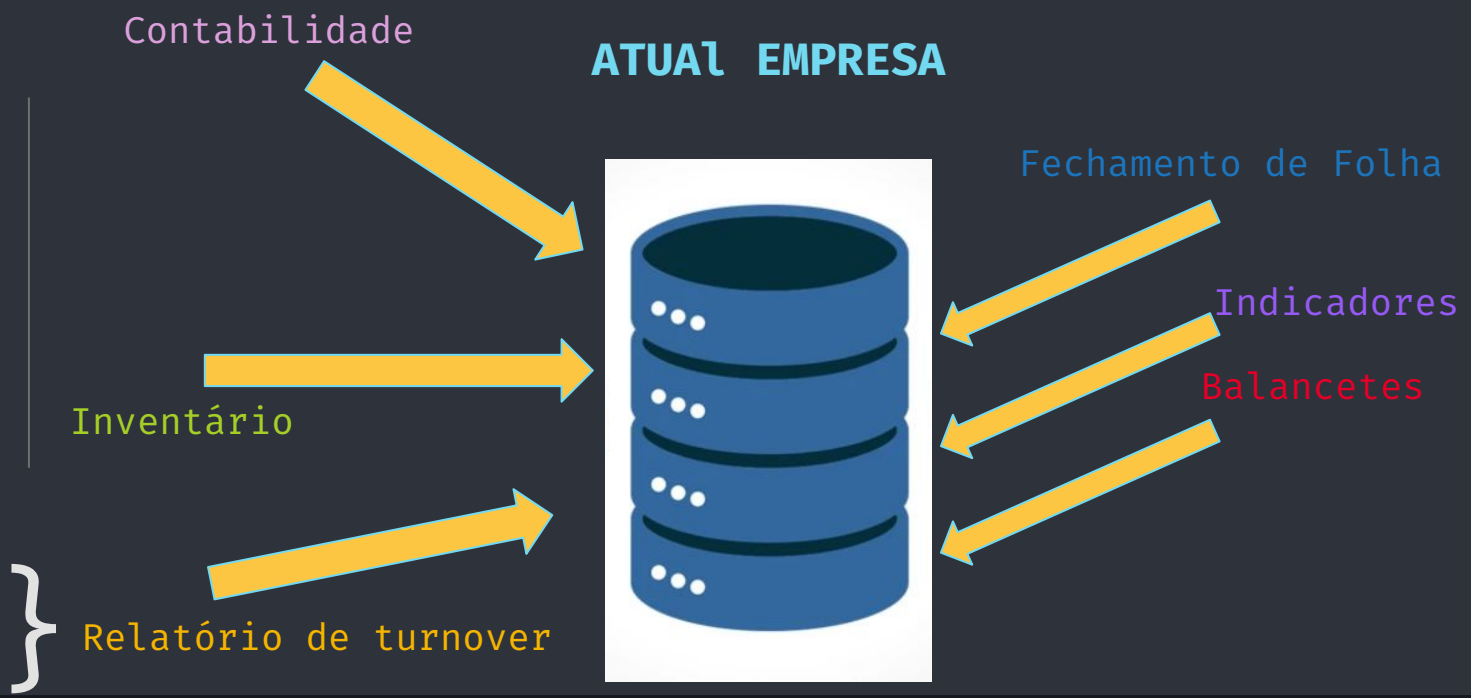
Imagine que você tem um banco SQL com 1 bilhão de linhas (por exemplo, 1 bilhão de vendas).

Pessoas que trabalharam com esses dados e informações:

== **Frontend:** Departamentos operacionais como o atendimento a clientes, que fazem a manutenção de dados (ex: inclusão, alteração e exclusão de novos produtos, clientes e vendas);

== **Backend:** Departamentos de controle, por exemplo o DP, que trabalham com informações sintetizadas (ou sumarizadas) para tomada de decisões e operação gerencial da empresa (ex: pagamento de comissões, recolhimento de impostos, etc).

}


Contabilidade
ATUAl EMPRESA
Fechamento de Folha
Indicadores
Balancetes
Inventário
Relatório de turnover

O que vai acontecer?

}

O que vai acontecer?

}

Entre as diversas abordagens, podemos avaliar essas 4 principais:

1. **Melhorar o hardware do servidor de banco:** Para converter as informações unitárias (analíticas) em totalizadas (sintéticas) é preciso poder de *processamento*. Então se melhorarmos o servidor de banco de dados, a tendência é que melhore a performance.

    a. Ótimo para pequenas e médias empresas;
    b. Não serve tanto para empresas de grande porte;

2. **Duplicar o banco de dados:** Essa base naturalmente não vai conter os dados das transações de hoje, porém em geral, relatórios gerenciais não necessitam dos dados de hoje, o que eles precisam é do fechamento de ontem ou do mês passado, por exemplo.

Vantagens:

a. Operação diária OK;
b. Solução fácil e barata de implementar comparado ao nível 3;
c. Exige gerenciamento simples.

Desvantagens:

b. Relatórios ainda depende de alguém(s) especializado que conheça bem SQL, carinha da TI, e a estrutura do banco de dados relacional, que a essa altura da sua empresa possivelmente seja complexa;
c. Cada vez que alguém gera um relatório é feito um esforço computacional que não é reaproveitado. Em última análise gera custos de desgaste de equipamentos e energia elétrica.

3. **Implantar um DW:** Para resolver as 2 desvantagens do modelo 2, podemos implementar um DW, que nada mais é que uma base de dados com as informações salvas em um formato sumarizado, como as áreas de backend necessitam em seus relatórios e interfaces.

Por exemplo: Ao invés de ter os valores de comissão espalhados em cada registro de vendas (como é na base de produção), o DW conterá uma tabela com os totais de comissão de cada vendedor.

bd_prd.tb_vendas

id	dt_venda	id_produto	id_vendedor	id_cliente	qtd
1	2014-06-01	3	10	1	5
2	2014-06-02	4	9	3	2
3	2014-06-02	3	7	5	1
4	2014-06-03	1	2	6	9

bd_dw.metas_vendas

mes_referencia	Produto	Meta	Vendas	%meta
Jan/2017	CPU Nasa 10Ghz	R$ 1.000.000,00	R$ 1.034.000,00	103.40%
Jan/2017	HD SSD 1TB	R$ 1.000.000,00	R$ 1.012.000,00	101.20%
Jan/2017	Monitor 32 pol full HD	R$ 500.000,00	R$ 537.000,00	107.40%

Tá, mas eu já implantei tudo isso na minha empresa, e ai?

4. **Múltiplos DWs:**

a) **Ampliar os servidores:** Lembra da solução 1? Se melhorarmos o servidor do DW, melhora a performance.

b) **Criar cópias do DW:** Lembra da solução 2? Também é prática comum criar várias cópias do DW e disponibilizar um DW para cada área. Então nosso RH poderia ter uma cópia e nosso comercial outra…

Tá, mas eu já implantei tudo isso na minha empresa, e ai?

4. **Múltiplos DWs:**

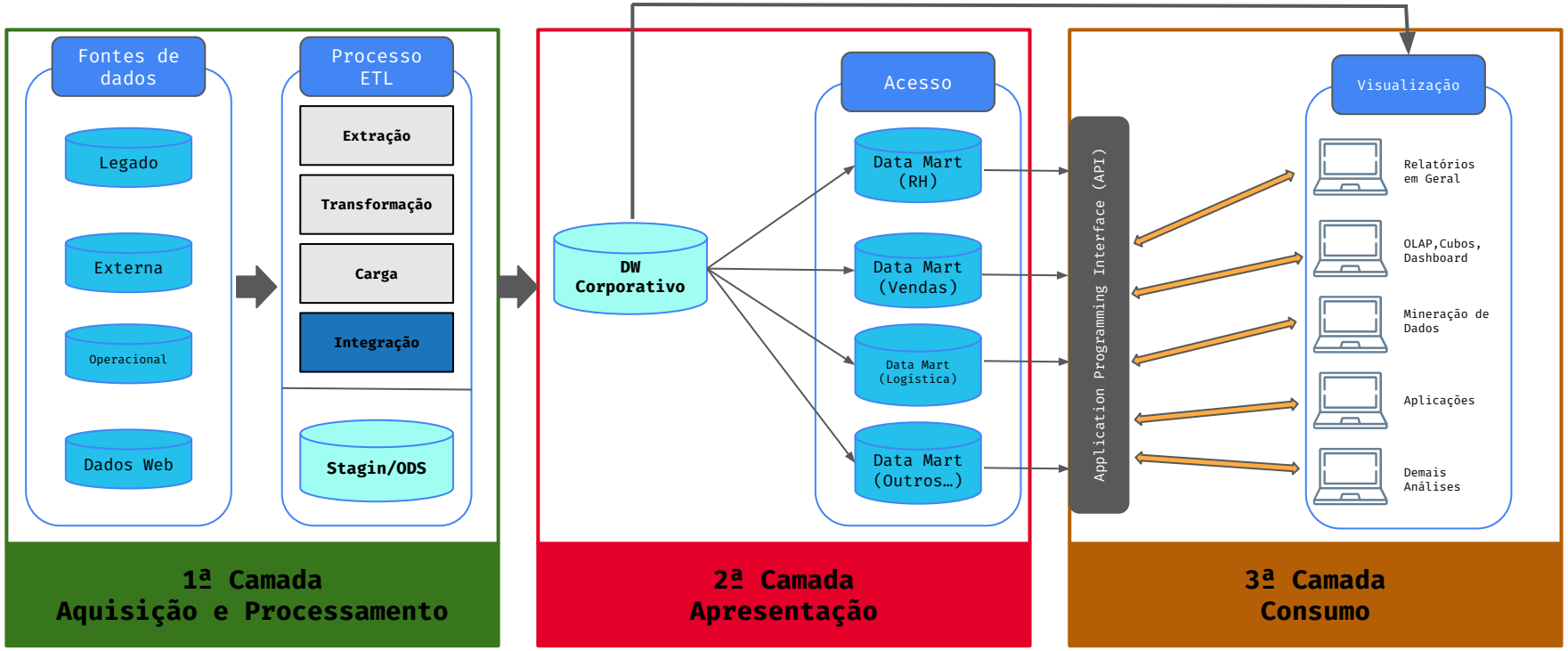
c) **Criar "sub-dws":** Se o seu DW for muito grande, ai o armazenamento ou transferência de dados pode ser um problema… *"são as dores do crescimento…"*

Então temos uma terceira alternativa, que é "dividir" o nosso DW em DWs menores.

Simplificando: você cria vários bancos de dados e distribui as tabelas de acordo com a necessidade de cada área. Por exemplo: A tabela de Vale-refeição pode não interessar ao comercial, então colocaremos ela somente no sub-DW do DP. Assim como a tabela de metas, se não interessar a contabilidade, podemos deixar só no sub-DW do comercial. O nome bonito para chamar um sub-DW no mundo de business intelligence é **"data-mart"**.

04 {

[Repositórios, Camadas, Níveis no DW]



04 {

[Repositórios, Camadas, Níveis no DW]

1ª Camada: Software de aquisição de dados (back-end), que extrai dados dos sistemas legados e fontes externas, os consolida e resume, e depois os carrega no Data Warehouse. Nesta camada está também presente o sistema operacional escolhido para uso.

2ª Camada: O próprio Data Warehouse, que contém os dados e o software associados e que são apresentados para o consumo dos usuários finais.

3ª Camada: Software cliente (front-end) ou aplicação, que permite aos usuários acessar e analisar dados a partir do DW. Geralmente usa uma camada de Middleware(utilizando API), para acesso ao DW com menor acoplamento entre as camadas.

04 {

[Repositórios, Camadas, Níveis no DW]

1ª Camada: Software de aquisição de dados (back-end), que extrai dados dos sistemas legados e fontes externas, os consolida e resume, e depois os carrega no Data Warehouse. Nesta camada está também presente o sistema operacional escolhido para uso.

2ª Camada: O próprio Data Warehouse, que contém os dados e o software associados e que são apresentados para o consumo dos usuários finais.

3ª Camada: Software cliente (front-end) ou aplicação, que permite aos usuários acessar e analisar dados a partir do DW. Geralmente usa uma camada de Middleware(utilizando API), para acesso ao DW com menor acoplamento entre as camadas.

04 {

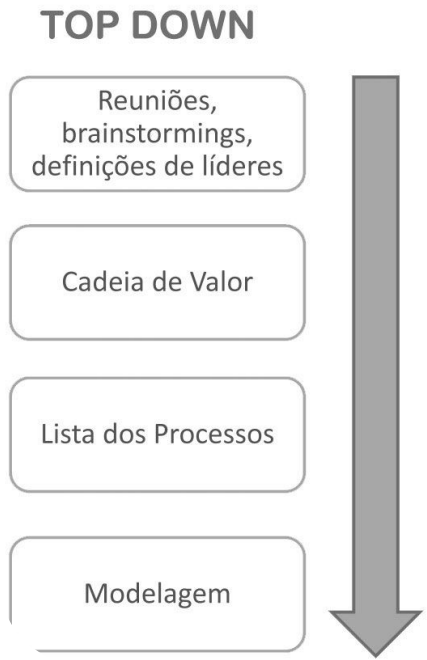
[Repositórios, Camadas, Níveis no DW]

Abordagem “top down”

Segundo Inmon:

*A implementação top down é conhecida como padrão inicial do conceito de DW. Ela requer maior planejamento e trabalho de **definições conceituais** de tecnologias antes de iniciar o projeto DW. Neste tipo de arquitetura constrói-se, em primeiro lugar, o DW e depois se extrai os dados para os Data Marts.*

The Data Warehouse Toolkit, Third Edition



04 {

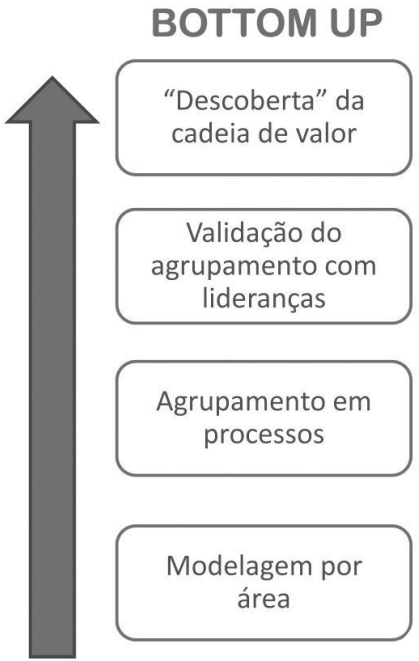
[Repositórios, Camadas, Níveis no DW]

Abordagem “bottom up”

Segundo Kimball:

Devido ao custo e tempo necessários em uma implementação top down, a implementação botton up vem se tornando mais popular. O propósito é construir um DW incremental a partir do desenvolvimento de Data Marts independentes. Ela é bastante utilizada, pois possui um retorno de investimento muito rápido.

The Data Warehouse Toolkit, Third Edition

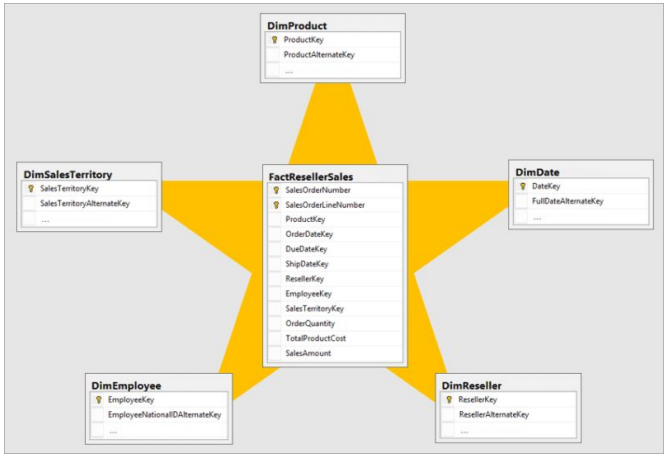


05 { [Modelagem Dimensional Colaborativa - Fatos e Dimensões]

- **O que é modelagem dimensional?**

  Técnica de projeto lógico normalmente usada para data warehouses que contrasta com a modelagem entidade-relacionamento.

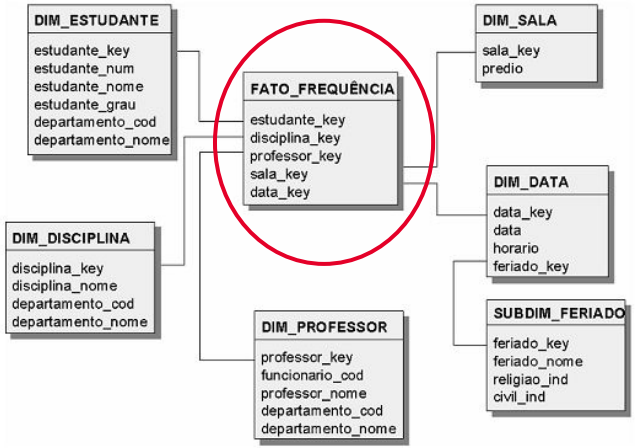
  - **Recuperação de dados mais rápida**
  - **Melhor compreensão dos processos de negócios**
  - **Flexível para mudar**



05 { [Elementos – Fatos ou Medidas de Negócio]

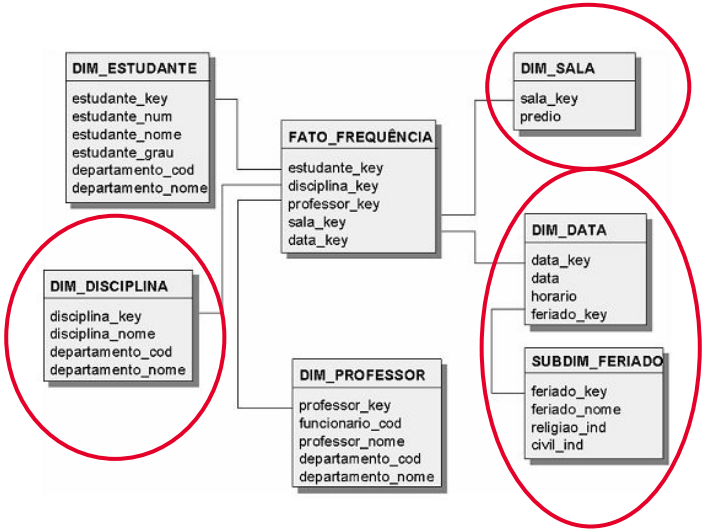
- As tabelas de fatos armazenam as informações numéricas sobre medidas de negócios e chaves estrangeiras para as tabelas dimensionais.

- Elas possuem os seguintes tipos:
  - **Fatos Aditivos:**Medidas de negócios que podem ser agregadas em todas as dimensões
  - **Fatos semi-aditivos:**Medidas de negócios que podem ser agregadas em algumas dimensões e não em outras (geralmente dimensões de data e hora).
  - **Fatos não aditivos:** Medidas de negócios que não podem ser agregadas em qualquer dimensão



05 {
[Elementos – Dimensões]

As **tabelas de dimensão** armazenam informações descritivas sobre os fatos educacionais para ajudar a compreender e analisar melhor os dados.
No exemplo apresentado na Tabela ao lado, Data, Sala e Disciplina são entidades de dimensão, dando mais informações sobre os fatos da frequência.
A quantidade total de faltas ou presença é uma medida importante a ser registrada, mas sem as **dimensões**, uma instituição não pode avaliar qual aluno(a) mais faltou ou o aluno(a) que mais esteve presente.



05 {
[Elementos – Etapas]

Etapa 1: Identificar os processos de negócios

Antes de modelar os dados, você deve encontrar os tipos de modelagem dimensional **apropriados** para seu modelo de dados. O processo de modelagem dimensional (ou qualquer modelagem de dados) começa com a identificação do **processo de negócios** que você deseja rastrear. No nosso exemplo, vamos acompanhar as vendas dos dois tipos de blusão.

Etapa 2: identificar fatos e dimensões em seu modelo de dados dimensionais

As informações em um modelo dimensional são categorizadas em dois tipos de tabelas – **Fatos e Dimensões**. A próxima etapa é identificar os fatos de negócios que você deseja medir e suas dimensões associadas. Em nosso exemplo, a venda de blusão é o fato que queremos medir. Data, localização da loja (Califórnia e Pensilvânia) e tipo de produto (blusões de náilon e blusões de poliéster) são as dimensões que nos fornecem mais informações sobre o processo de vendas.

Dimensão de Data

Chave de data	Data	Dia
10201	6/3/2018	Domingo
10202	6/4/2018	Segunda-feira

Dimensão do Produto

Código Do Produto	Coleção	Material	Cor
131620	Blusão - coleção outono	Nylon	Laranja
131571	Blusão - coleção outono	poliéster	Preto

Dimensão da Loja

Chave da Loja	Nome da loja	Cidades	Estado
151	AngAngie'sparel	Los Angeles	Califórnia
152	AngAngie'sparel	Pittsburgh	Pennsylvania

05 {
[Elementos - Etapas]

Etapa 3: identificar os atributos para dimensões
Depois de identificar as dimensões e fatos para seu processo de negócios, a próxima etapa é identificar os atributos e criar uma tabela dimensional separada para cada dimensão. Existem diferentes tipos de tabelas dimensionais para cada tipo de dados. Cada registro na tabela de dimensões deve ter uma chave exclusiva. Essa chave será usada para identificar os registros na tabela de dimensões e como chave estrangeira na tabela de fatos para referenciar a dimensão específica e juntá-la à tabela de fatos.

Etapa 4: definir a granularidade dos fatos comerciais
A granularidade refere-se ao nível de informações armazenadas em qualquer tabela. Em nosso exemplo, o valor das vendas é registrado diariamente; portanto, a granularidade, neste caso, é diária. As tabelas de fatos em um modelo dimensional devem ser consistentes com a granularidade pré-definida.

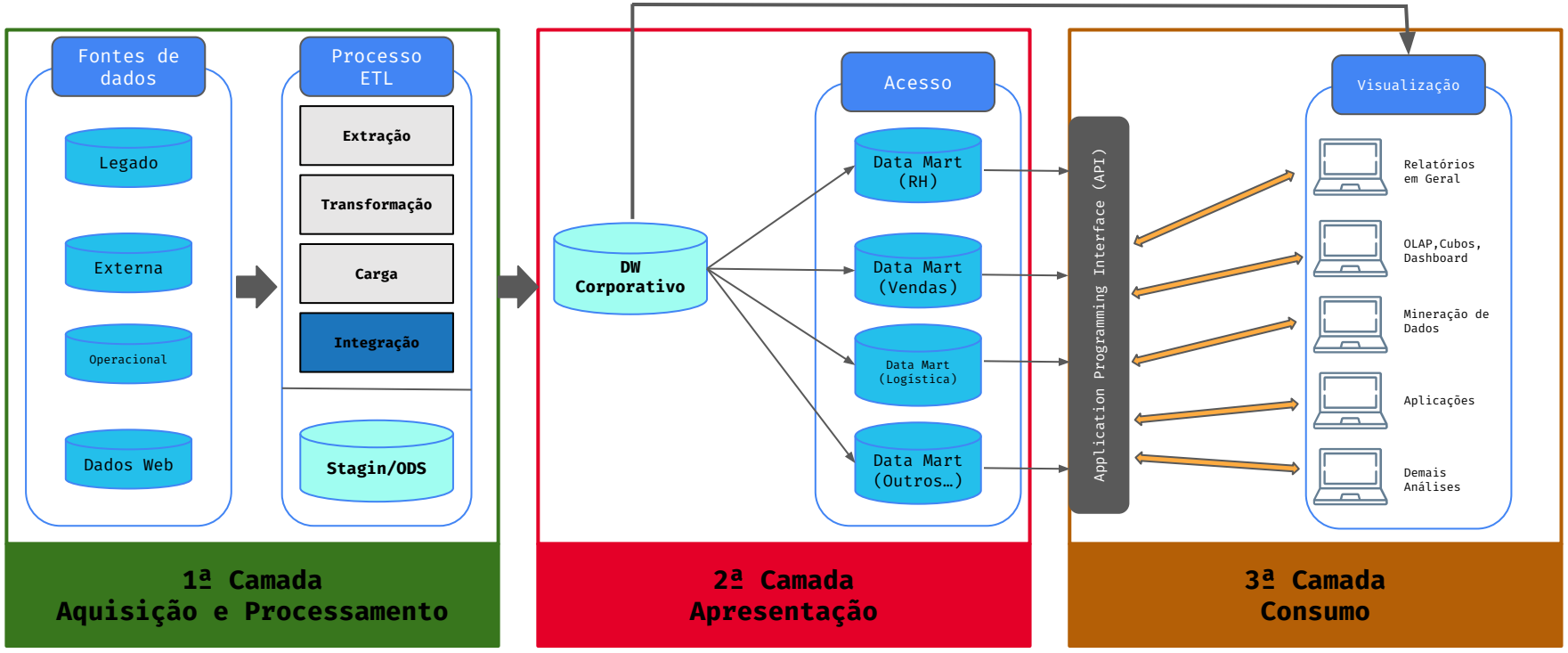
05 {

[Elementos - Etapas - CUIDADO!]

Etapa 5: armazenamento de informações históricas (dimensões que mudam lentamente)
Uma característica importante dos modelos dimensionais é que os atributos dimensionais podem ser facilmente modificados sem alterar as informações completas da transação. Por exemplo, a linha de roupas decide continuar o blusão de nylon da coleção de outono para a coleção de primavera.

Então como vamos manter o histórico?
Dimensões que mudam lentamente ao longo do tempo são chamadas de **Dimensões de Alteração Lenta.** Além disso, a tabela de dimensão de tempo em um data warehouse é gerada automaticamente e captura a hora em que ocorrem diferentes transações.

- Através de Rastreamento do dado;
- “Congelando” e definindo a carga histórica desde o processo inicial;


Fontes de dados
Legado
Externa
Operacional
Dados Web
Processo ETL
Extração
Transformação
Carga
Integração
Stagin/ODS
1ª Camada
Aquisição e Processamento
Acesso
DW Corporativo
Data Mart (RH)
Data Mart (Vendas)
Data Mart (Logística)
Data Mart (Outros…)
2ª Camada
Apresentação
Application Programming Interface (API)
Visualização
Relatórios em Geral
OLAP,Cubos, Dashboard
Mineração de Dados
Aplicações
Demais Análises
3ª Camada
Consumo

05 { [Prática]

Escolha duas área de negócio em uma empresa fictícia e crie um DW:

- Fontes:
  - PostgreSQL;
  - Relatórios em Excel;

Áreas de negócio:

- RH;
- Financeiro;
- Contabilidade;
- Vendas;
- Estoque;
- Consulta Médica;
- Setores de Internação de Paciente;

Modelagens Dimensionais

- **RH:**
  - Acompanhamento de Contratações;
- **Financeiro:**
  - Contas a Pagar;
- **Contabilidade:**
  - Entradas e Saídas, fechamento mensal;
- **Vendas:**
  - Acompanhamento de vendas por vendedor;
- **Estoque:**
  - Controle de Quantidade de Produtos;
- **Consulta Médica:**
  - Relatório de pacientes atendidos;
- **Setores de Internação de Paciente:**
  - Relatório de pacientes internados na UTI;
- **Marketing:**
  - Quantidade de acessos por campanha;

Fontes

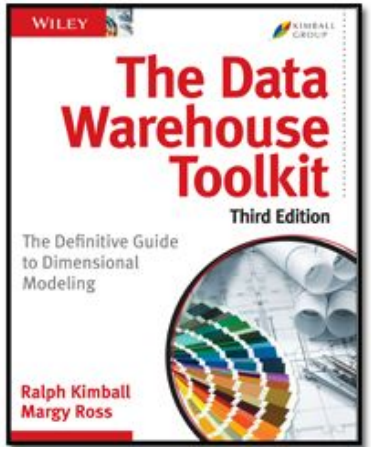
https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/

https://www.astera.com/pt/tipo/blog/guia-de-modelagem-dimensional/#elements-dimensional-model

https://dba-pro.com/o-que-e-data-warehouse/

http://www.kimballgroup.com/wp-content/uploads/2013/08/2013.09-Kimball-Dimensional-Modeling-Techniques11.pdf

Wiley, 2013



uniesp
Centro Universitário


uniesp
Centro Universitário