



data

DATA WAREHOUSE (DW)

Baseado na tese de
doutorado da Professora
Cristina Dutra de Aguiar

Carlos Filipe de Castro Lemos



CONTEXTO



MERCADO DE NEGÓCIOS E SUPPLY CHAIN

Leite



Nestlé
Vigor
Itambé
Italac
Danone



Cremes
Doces
Queijos
Iogurtes
Manteiga
Whey Protein



Walmart
Carrefour
Pão de Açúcar
Extra
Markro



Produtos

Finanças

Informações



OPERACIONAL DE UMA EMPRESA

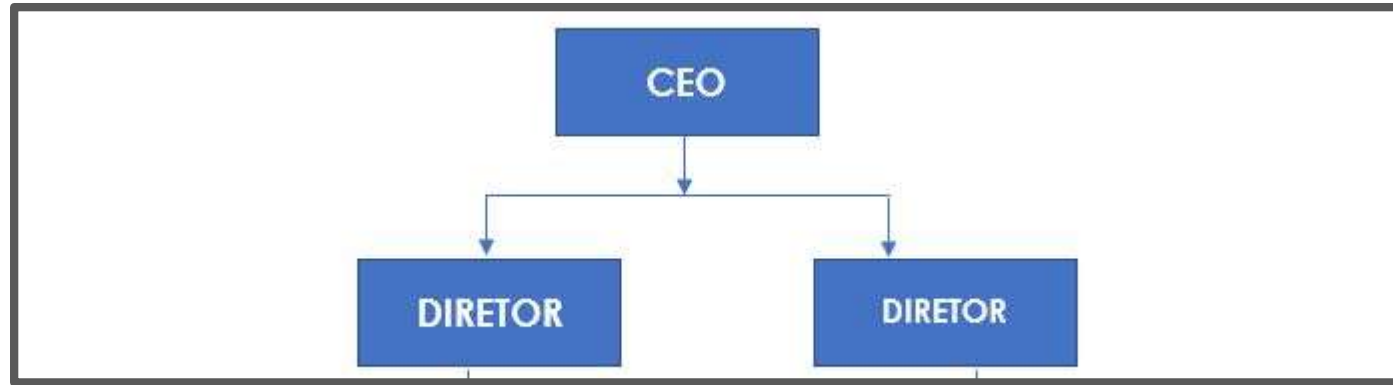


- **Vendas:** pedido de compra, nota fiscal de venda, comissões.
- **Contabilidade** registro das transações financeiras e contábeis (balanço patrimonial, demonstração de resultados, contas a pagar e a receber, previsões, etc).
- **Financeiro:** fluxo de caixa, controle de aplicações e empréstimos, relatórios de despesas e receitas.

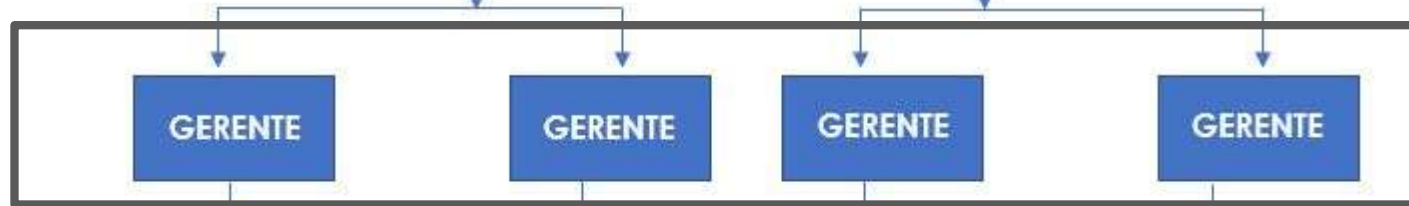


ORGANIZAÇÃO EMPRESARIAL

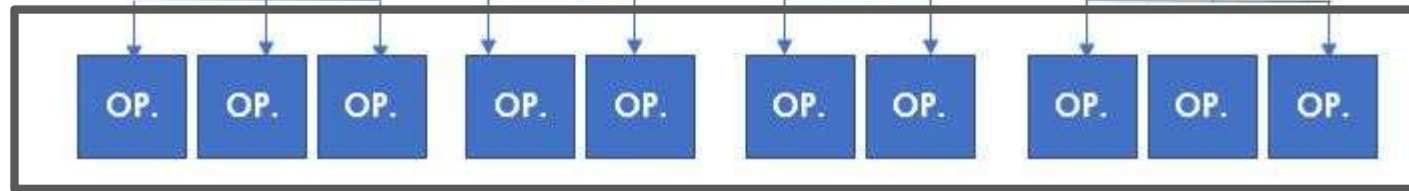
ÁREA
ESTRATÉGICA



ÁREA
GERENCIAL



ÁREA
OPERACIONAL



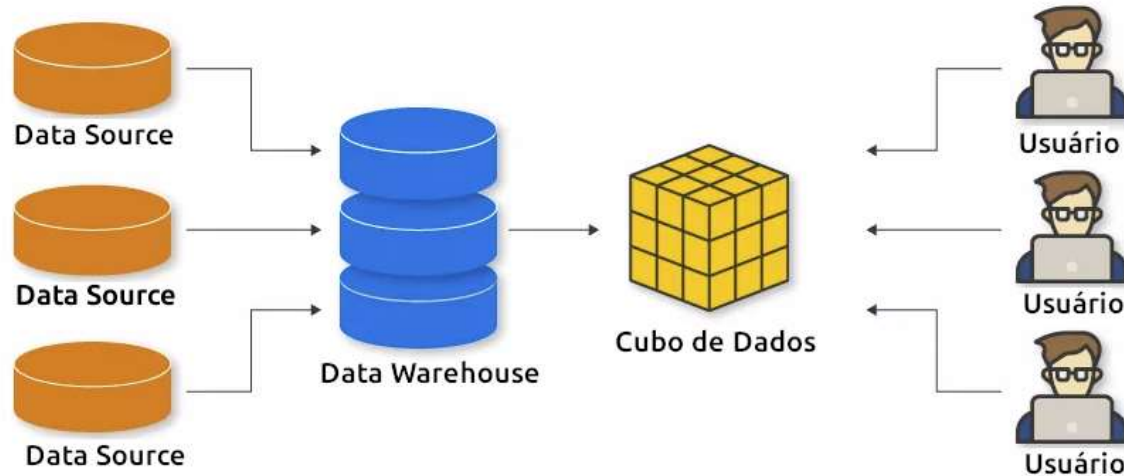


DATA WAREHOUSE



O QUE É UM DATA WAREHOUSE?

Um **Data Warehouse (DW)** é uma solução de dados que armazena dados temáticos de forma multidimensional com o objetivo de auxiliar na análise e tomada de decisões empresariais.





GRUPOS DE USUÁRIOS



CARACTERÍSTICAS: GRUPOS DE USUÁRIOS ESPECÍFICOS

Data Warehouse é direcionado a **usuários e grupos específicos** que precisam acessar, analisar e interpretar grandes conjuntos de dados para apoiar a tomada de decisões estratégicas.



Executivos e Alta Gestão: monitoração dos negócios para insights estratégicos para tomada de decisão a nível da organização. Exemplo: KPI.



Gerentes e Tomadores de Decisões Operacionais: obter informações detalhadas e planejar o futuro próximo da empresa. Exemplo: Gerente de Supply Chain preparando a cadeia de negócios para vendas de Natal.



CARACTERÍSTICAS: GRUPOS DE USUÁRIOS ESPECÍFICOS



Profissionais de Dados: grupos de suporte para fornecer informações precisas e dados estratégicos a área estratégica e gerencial. Exemplo: performance de vendas de um produto em uma região por determinado preço.




Profissionais de Finanças e Contabilidade: informações de fluxos de caixas, tributos e demonstrações financeiras. Exemplo: controle orçamentário.



Atendimento ao Cliente e Vendas: acesso a perfil do cliente, histórico de compras e nível de satisfação. Exemplos: Gerente de Marketing analisando a experiência do usuário





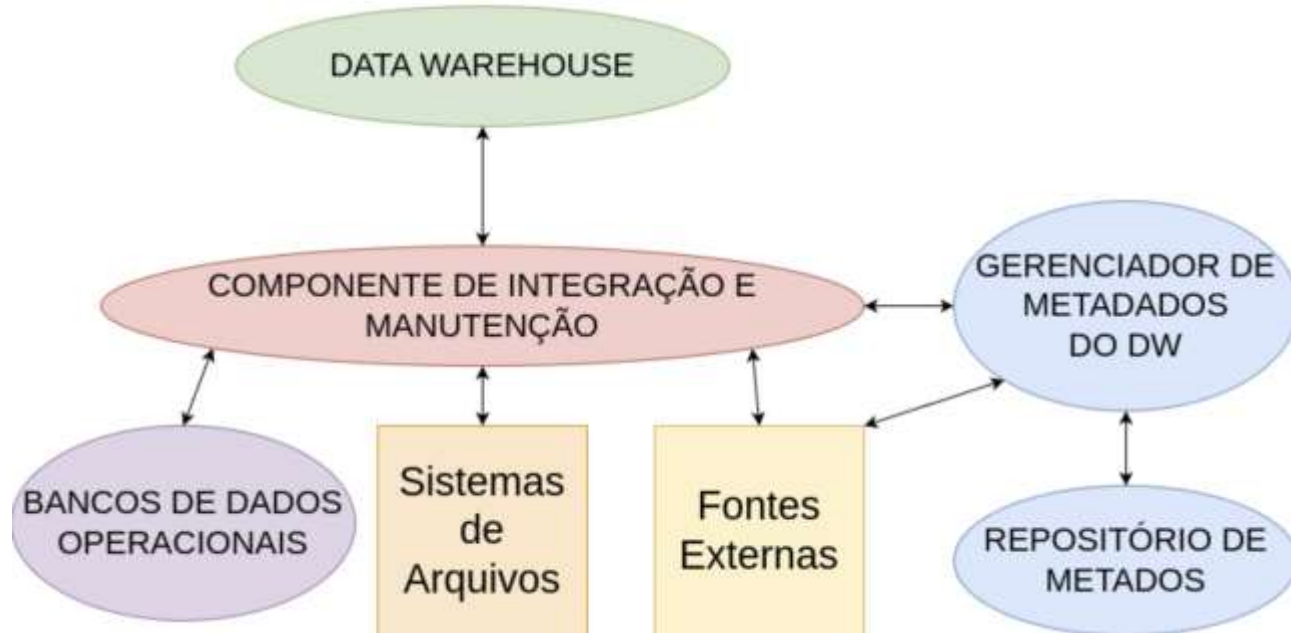
EXTRAÇÃO, TRANSFORMAÇÃO E LOADING (ETL)



COMPONENTE DE INTEGRAÇÃO E MANUTENÇÃO

- **Conceito**: trata-se de um módulo de código que é responsável pela extração, tradução, filtragem, integração e armazenamento dos dados no DW.

UPDATES TAMBÉM



EXTRAÇÃO DOS DADOS

Data Warehouse é projetado e otimizado para centralizar e organizar grandes volumes de dados. A extração dos dados envolvem múltiplas origens e variados formatos de arquivos e são voltados para dados históricos com a finalidade de auxiliar na tomadas de decisões gerenciais e estratégicas.

Soluções Empresariais



Databases



Arquivos



TRANSFORMAÇÃO DOS DADOS: CARACTERÍSTICAS

- Os dados coletados na etapa anterior serão **transformados** para que fiquem **limpos**, **consistentes** e **estruturados** para a formação do cubo de análise.
- Serão formadas as **Tabelas de Fato** e **Tabelas de Dimensão**.

Data cleaning: remover ou corrigir dados imprecisos, incompletos ou duplicados. Ex.: datas (“12/23/2023” para “23-12-2023”), preencher lacunas com valores padrão (null para “Sem dados”) ou preenchimento por imputação (utilização de média ou mediana).

Filtragem dos dados: remover ou selecionar apenas os dados relevantes para o propósito da análise temática. Ex.: filtrar dados do último ano fiscal ou dos últimos 12 meses.



TRANSFORMAÇÃO DOS DADOS: CARACTERÍSTICAS

Conversão de Tipos de Dados: é preciso garantir consistência de tipos quando se importa dados, bem como quando os dados são lidos de arquivos (sem tipo definido). Ex: números decimais em inteiros ou inteiros em decimais ou varchar (MySQL) para text (PostgreSQL).

Derivação de Dados: criação de novos atributos a partir dos existentes. Ex: classificar clientes como “VIP” de acordo com o histórico de compras ou cálculo de idade a partir da data de nascimento.

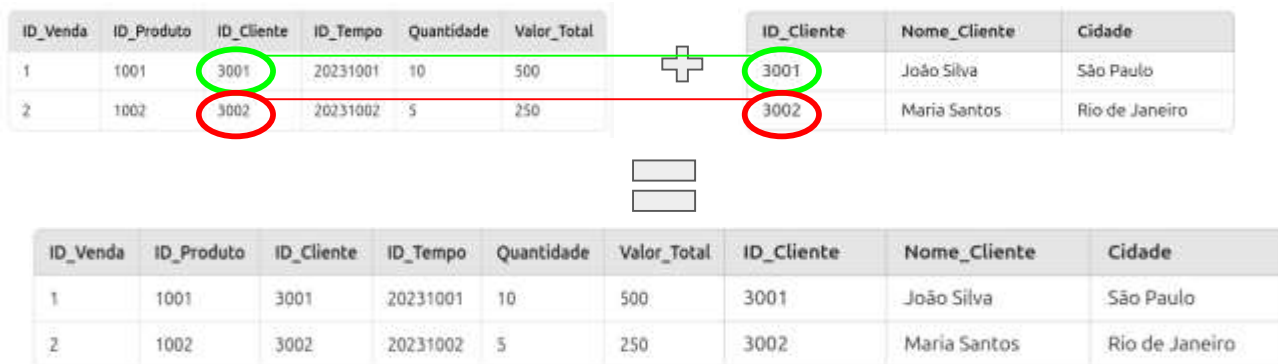
Normalização e Padronização dos Valores de Dados: os dados devem seguir o padrão temático. Ex.: utilizar capitalize em nomes próprios (pessoas, cidades, etc), normalizar quantidade de casas decimais, moeda corrente (real, dólar ou outra), normalização de categorias como gênero (masculino, feminino, homem ou mulher).



TRANSFORMAÇÃO DOS DADOS: CARACTERÍSTICAS

Cálculos e Agregações: operações matemáticas e cálculos de agregações. Ex.: contagem (quantidade de usuários, vendas e serviços), soma (valores vendidos), média (preço médio, tempo médio de atendimento), máximo (lucro), mínimo (custo).

Junções de Dados (JOIN): combinar diferentes tabelas de para criar uma visão unificada. Ex.: junção da tabela Venda com a tabela Cliente.



TRANSFORMAÇÃO DOS DADOS: CARACTERÍSTICAS

Filtragem Outliers: identificar e remover dados que não combinam com o conjunto de dados. Ex.: '-50°C' de temperatura em área equatorial ou '120%' do abastecimento de água.

Mapeamento dos Dados: a categorização de uma filial pode ser diferente da utilizada na âmbito estratégico. Ex.: categorias como 'sapatos', 'tênis' e 'botas' podem ser agrupados em 'calçados' ou substituir códigos 'P001' para descrições completas 'Produto A'.

Enriquecimento de Dados: adicionar dados externos para enriquecer os dados existentes. Ex: inserir a cotação do dólar de 1964 a 1990 em uma tabela que possui dados após 1990.

Desagregação de Dados: dados podem ser desagregados para otimizar consulta. Ex.: data '23-12-2023' em dia ('23'), mês ('12') e ano ('2023').



TRANSFORMAÇÃO DOS DADOS: CARACTERÍSTICAS

Pivoteamento dos Dados: trata-se de técnica de reorganização ou reestruturação dos dados de modo a modificação a disposição de linhas em colunas visando melhorar a visualização de dados.

Mês	Produto	Vendas
Janeiro	Produto A	100
Janeiro	Produto B	150
Fevereiro	Produto A	200
Fevereiro	Produto B	180
Março	Produto A	210
Março	Produto B	190

Os atributos *Produto* (coluna de agrupamento) e *Vendas* (coluna de valores) deixarão de existir e seus valores serão transformados em novas colunas e valores. Isto é, os valores de *Produto* (*Produto A* e *Produto B*) serão novos atributos e os valores das vendas são transformados em valores dos novos atributos.

**Resultado
Final**

Mês	Produto A	Produto B
Janeiro	100	150
Fevereiro	200	180
Março	210	190



CARACTERÍSTICAS: LOADING DOS DADOS

Depois de realizadas todas as transformações, os dados serão carregados nas **Tabelas de Fato** e **Tabelas de Dimensão**

Tabelas de Fato representam *dados históricos do modelo temático* e trazem quantidades massivas de registros (grande altura com milhões ou bilhões tuplas) e poucas colunas (pouca largura), bem como *medidas quantitativas*.

Data	Produto_ID	Loja_ID	Vendas	Custo
2023-10-01	1	101	200	100
2023-10-01	2	102	150	90
2023-10-02	1	101	180	95
2023-10-02	3	103	220	120

Tabelas de Dimensão representam *atributos descritivos* e descrevem entidade de contexto. Possuem poucos registros (baixa altura) e muitas colunas (grande largura com centenas ou milhares de colunas).

Data	Ano	Mês	Dia	Trimestre
2023-10-01	2023	10	1	4º Trimestre
2023-10-02	2023	10	2	4º Trimestre



CARACTERÍSTICAS: CABEÇALHO DE ARQUIVOS (METADADOS)

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
status	topo				L	I	S	T	A	G	E	M		D	A		F	R	O
20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
T	A		D	O	S		V	E	I	C	U	L	O	S		N	O		B
40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
R	A	S	I	L	C	O	D	I	G	O		I	D	E	N	T	I	F	I
60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
C	A	D	O	R	:		A	N	O		D	E		F	A	B	R	I	C
80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
A	C	A	O	:		Q	U	A	N	T	I	D	A	D	E		D	E	
100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119
V	E	I	C	U	L	O	S	:		E	S	T	A	D	O	:		0	N
120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139
O	M	E		D	A		C	I	D	A	D	E	:		1	M	A	R	C
140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
A		D	O		V	E	I	C	U	L	O	:		2	M	O	D	E	L
160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179
O		D	O		V	E	I	C	U	L	O	:		proxRRN				nroRegRem	
180	181																		

Pode conter campos como quantidade de registros?

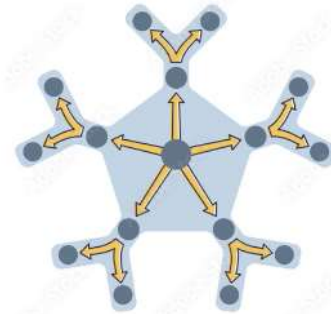
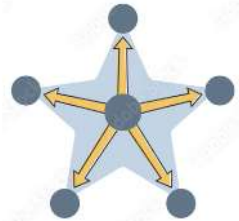
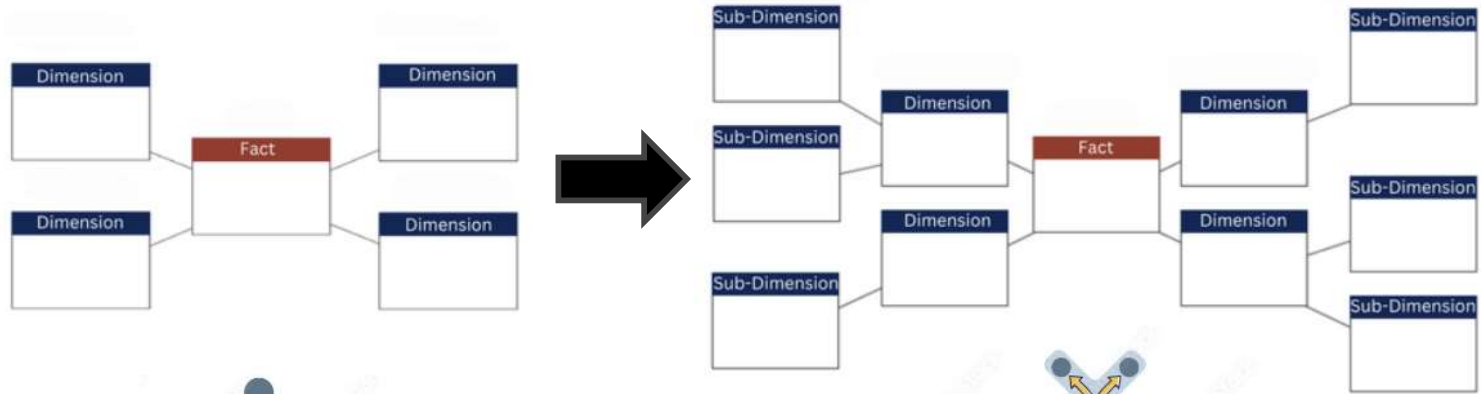




MODELAGEM DE DADOS DO DW



MODELO DE DADOS: STAR E SNOWFLAKES SCHEMA



EXEMPLO DE STAR SCHEMA: RESULTADO DO ETL

Tabela Dimensão (Tempo)

Data_ID	Ano	Mês	Dia
2023-01-01	2023	01	01
2023-02-01	2023	02	01

Tabela Dimensão (Filial)

Loja_ID	Nome_Loja	Localização
101	Loja A	São Paulo
102	Loja B	Rio de Janeiro

Tabela Fato (Vendas)

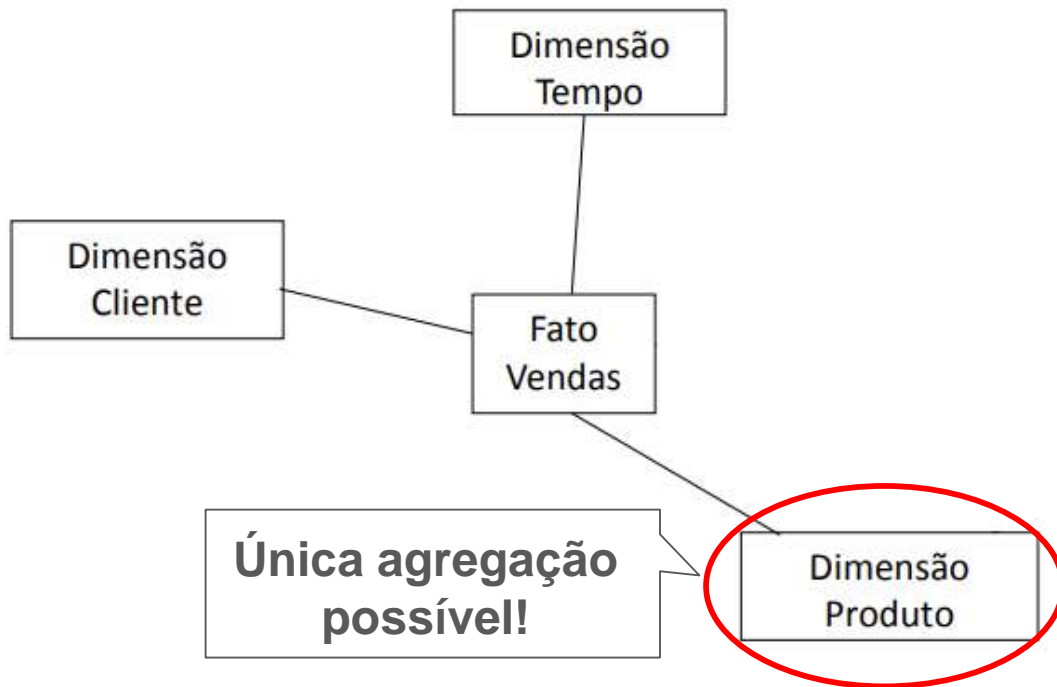
Data_ID	Produto_ID	Loja_ID	Vendas	Custo
2023-01-01	1	101	200	120
2023-01-01	2	101	150	90
2023-02-01	1	102	180	110

Tabela Dimensão (Produto)

Produto_ID	Nome_Produto	Categoria
1	Produto A	Eletrônicos
2	Produto B	Roupas



MODELO STAR SCHEMA



Observação

- Existem medidas numéricas **não-aditivas (não fazem sentido)**. Ex.: somar os dias do mês.
- Existem medidas numéricas **semi-aditivas (somam, mas resultado é equivocado)**. Ex: soma clientes que compraram os produtos P1 e P2 vendidos no mesmo dia não é uma operação válida, pois podemos contar duas vezes o mesmo cliente.



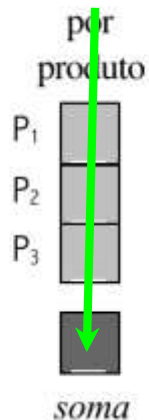
CUBO MULTIDIMENSIONAL: FATO(VENDA) DIMENSÕES (PRODUTO, FILIAL, DIA)

cubo 0-dimensional

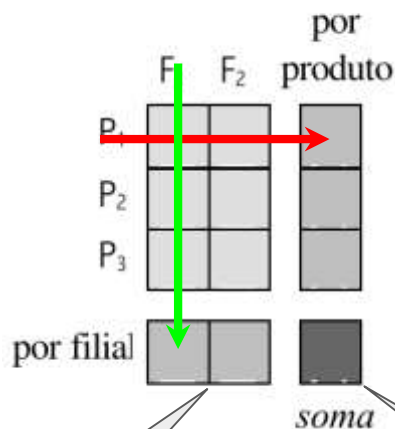


soma

cubo unidimensional



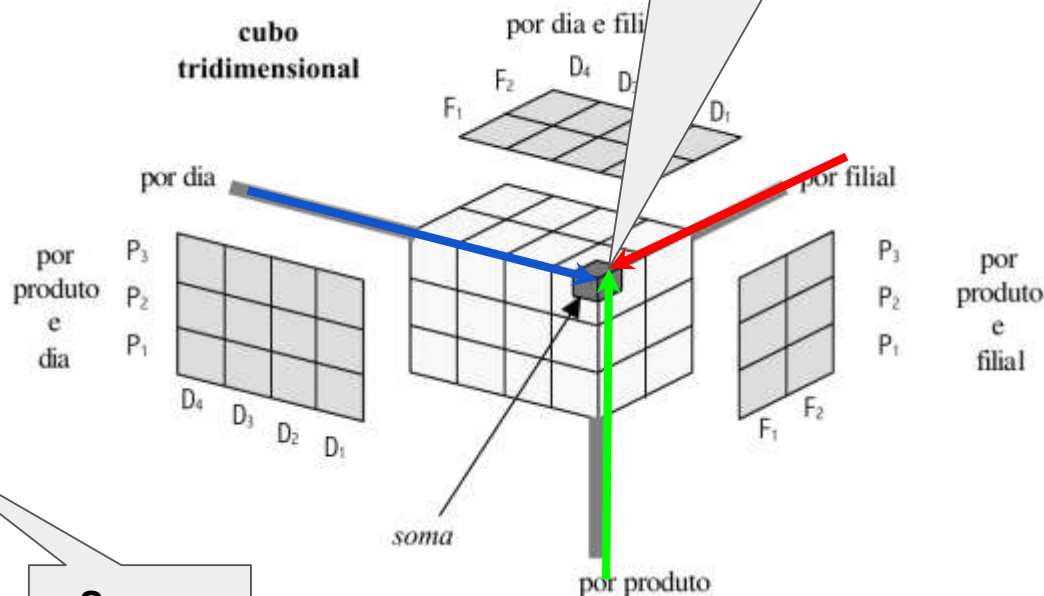
cubo bidimensional



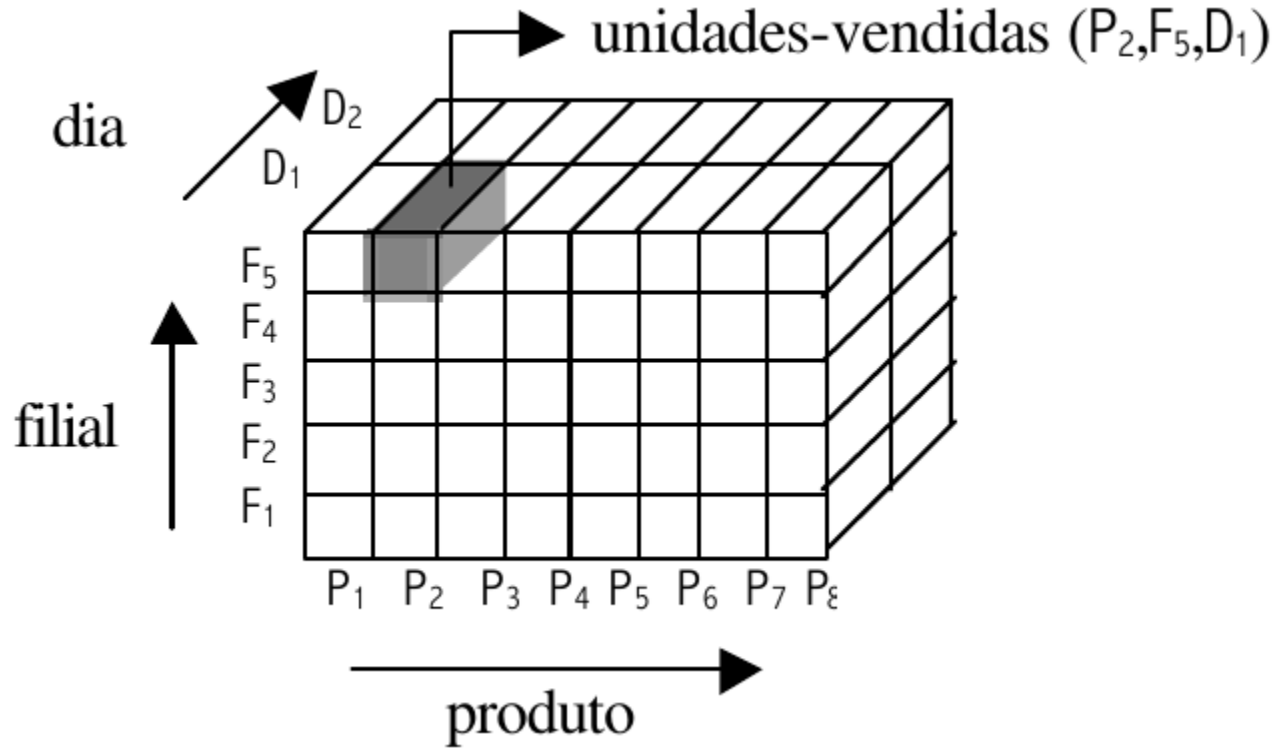
Somas
Marginais

Somas
Totais

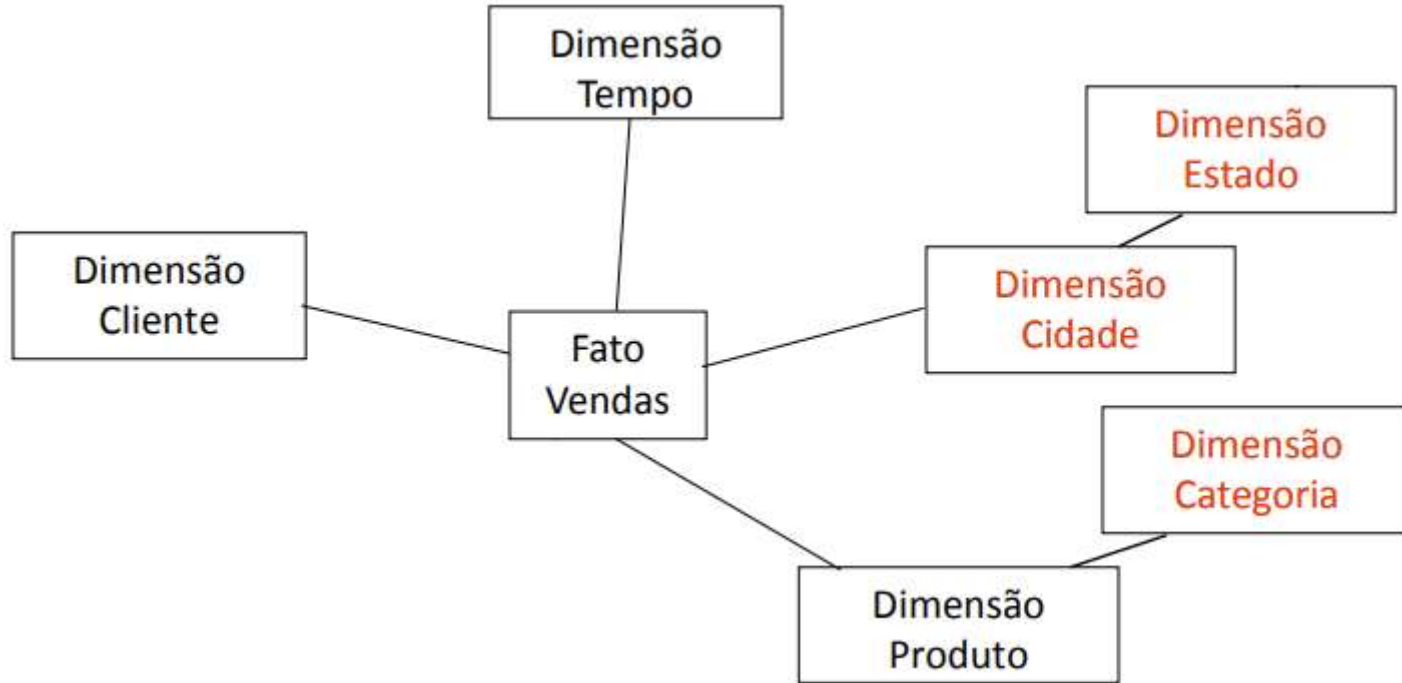
cubo tridimensional



DERIVAÇÃO EM CUBO MULTIDIMENSIONAL

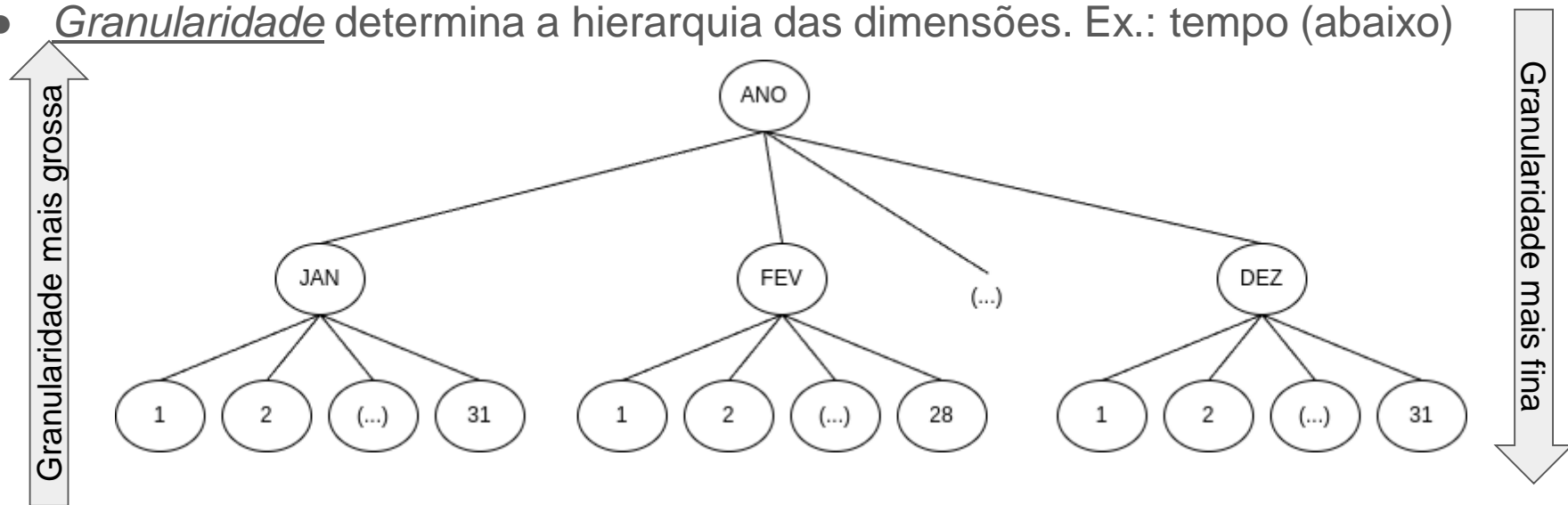


MODELO SNOWFLAKES



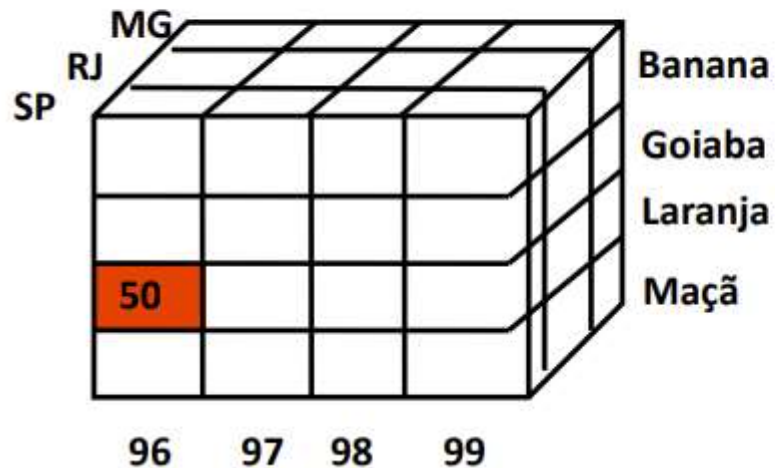
CUBO MULTIDIMENSIONAL: HIERARQUIA DAS DIMENSÕES

- As dimensões podem possuir categorias. Depende da regra de negócio e da forma como o assunto do banco de dados foi modelado. Ex.: Sapato, Tênis, Chinelo podem ser categorias de Calçados ou Cidade e Estado podem ser espécies do gênero Localidades.
- Granularidade determina a hierarquia das dimensões. Ex.: tempo (abaixo)

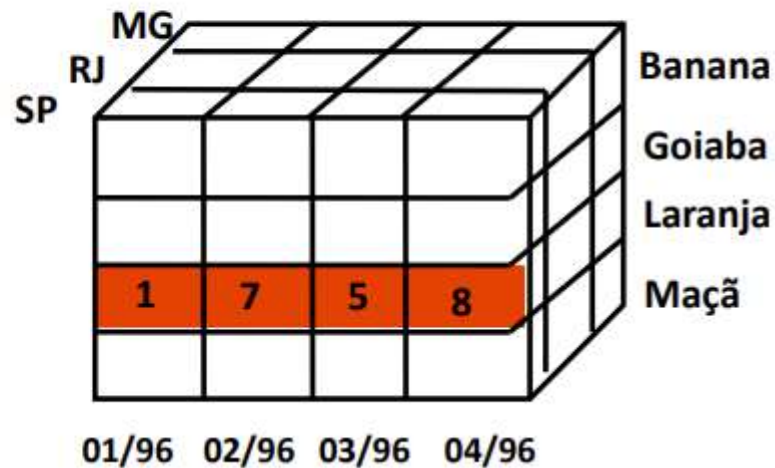


CUBO MULTIDIMENSIONAL

MENOR Granularidade



MAIOR Granularidade

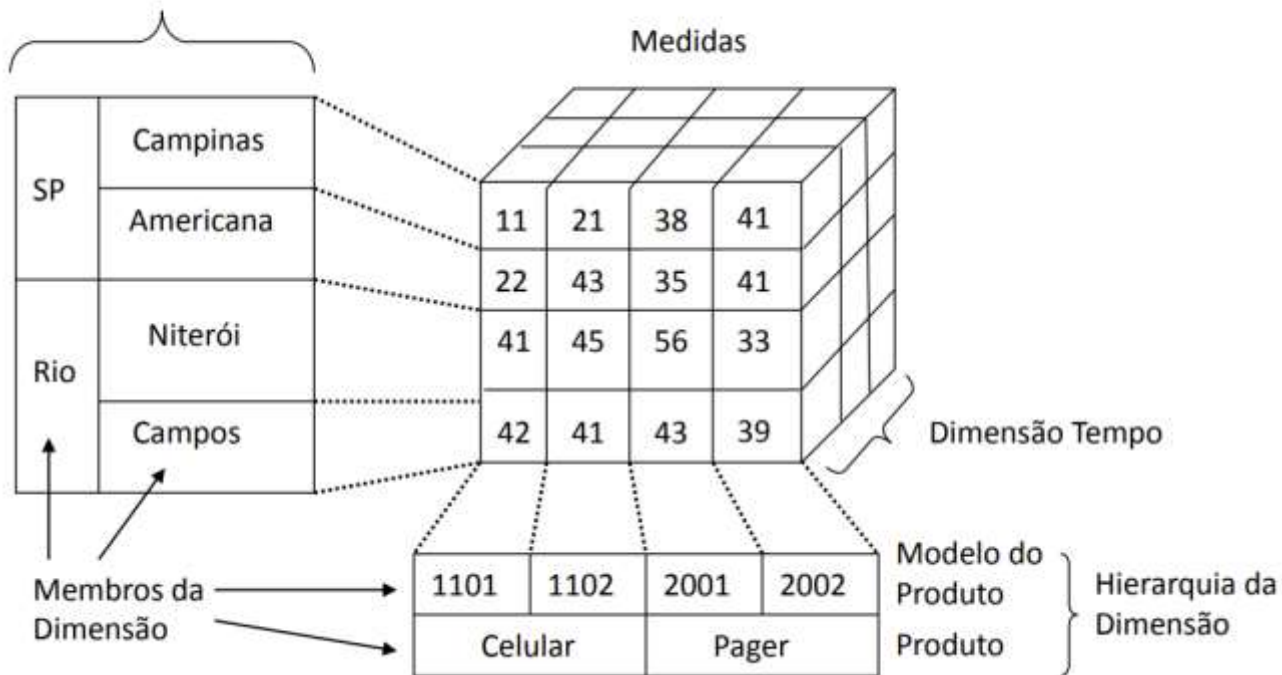


E se fosse por dia?



CUBO MULTIDIMENSIONAL

Dimensão Localização
Hierarquia da Dimensão



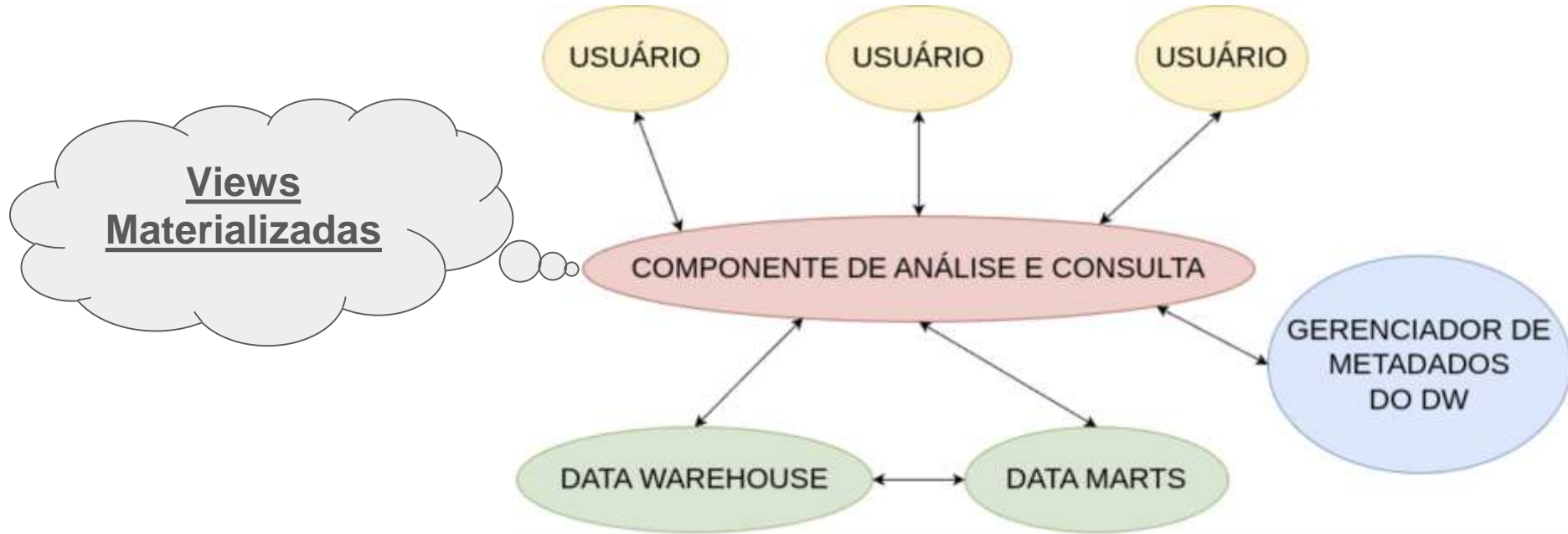


ACESSO AOS DADOS



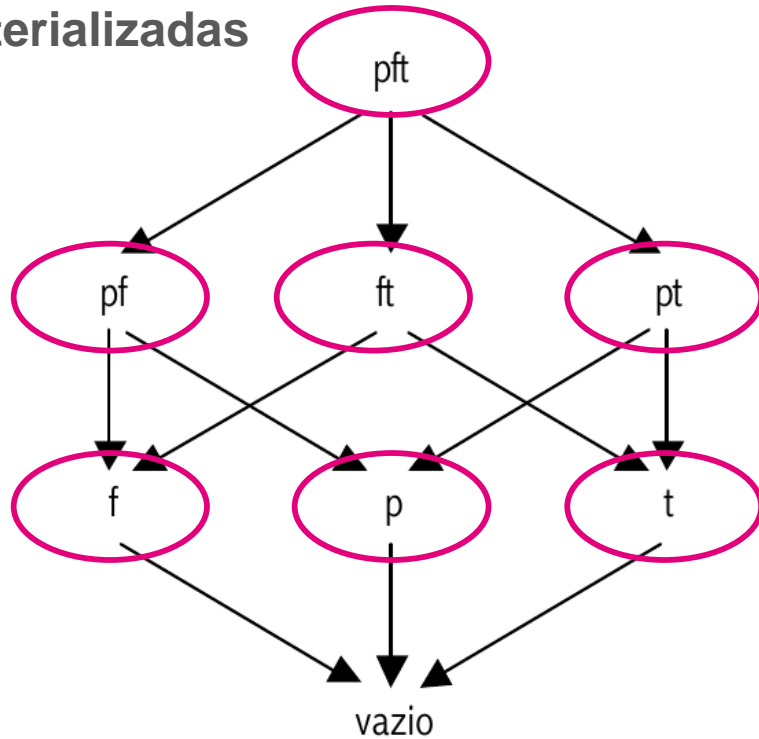
COMPONENTE DE ANÁLISE E CONSULTA

- É responsável por fazer a conexão entre o Data Warehouse e os Usuários, isto é, garantir o acesso às informações a quem precisa ter acesso às informações.



GRAFO DE DERIVAÇÃO

View
Materializadas

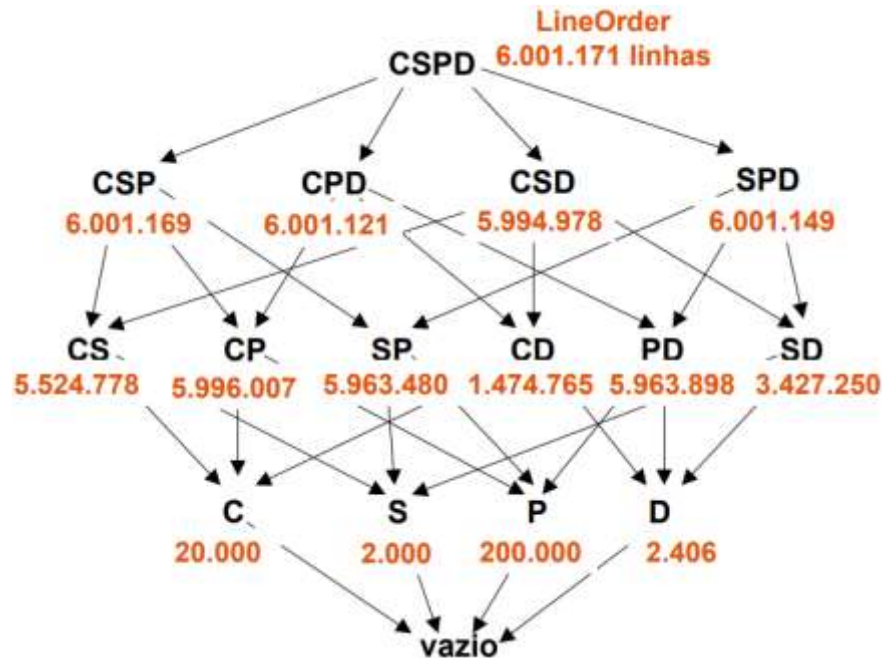


CUIDADO!

Não é um cubo. É um grafo de derivação de produto (p), filial (f) e tempo (t) em visões materializadas.



GRAFO DE DERIVAÇÃO EM NÚMEROS



Dimensões

- *Customer* (C)
- *Supplier* (S)
- *Part* (P)
- *Date* (D)



OTIMIZAÇÕES TÉCNICAS DO COMPONENTE

- **Estender a Linguagem SQL**: além das funções convencionais (sum, count, avg, max, min) é possível criar funções personalizadas. Exemplo: operador *cube* (usados nas análises dos cubos multidimensionais).
- **Estruturas de Indexação**: utilização de árvores ou outros mecanismos para acelerar a busca de informações nas views materializadas ou outras fontes.
- **Otimização de Consultas Complexas**: a criação de uma grande multiplicidade de views podem deixar calculados agregações de dados demandadas pelos usuários.
- **Execução Paralela**: a paralelização das buscas ou das execuções pode ser um fator para aumentar o *speedup* e a *eficiência*.



MUITO OBRIGADO!!



**CONTAT
O**



Linkedin: www.linkedin.com/in/cflemos



Email: filipelemos@usp.br





BIBLIOGRAFIA



- CIFERRI, Cristina Dutra de Aguiar Distribuição dos Dados em Ambiente de Data Warehousing: o sistema WebD²W e Algoritmos Voltados à Fragmentação Horizontal dos Dados. Universidade Federal de Pernambuco: Pós-Graduação em Ciência da Computação. Recife, 2002.
- <https://github.com/dpavancini/eng-de-analytics-livro> (StarSchema)
- <https://unstop.com/blog/difference-between-olap-and-oltp-in-dbms> (OLAP/OLTP)
- FONSECA, George H.G. Sistemas de Apoio à Decisão. Modelagem de Data Warehouse. Universidade Federal de Ouro Preto. Disponível em:
https://professor.ufop.br/sites/default/files/george/files/a06_modelagem_de_data_warehouse.pdf. Acesso em 13/10/2024

