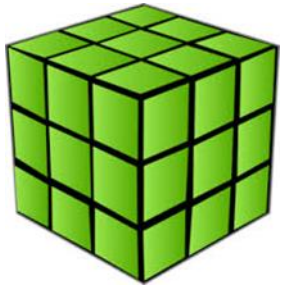
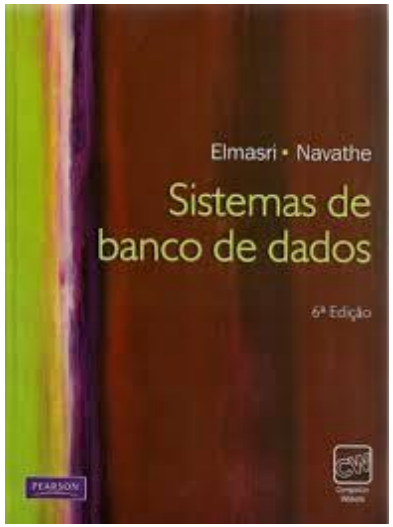
Unidade 25 – Modelagem de Dados para Data Warehouses

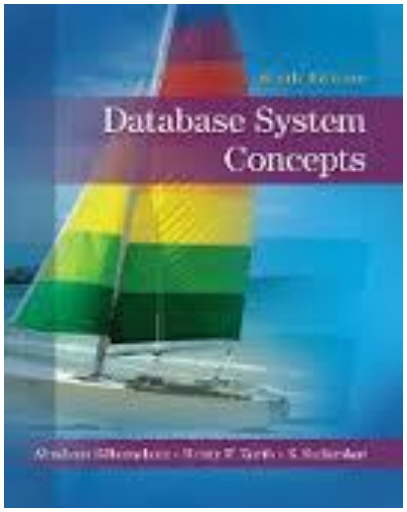


Prof. Aparecido V. de Freitas
Doutor em Engenharia
da Computação pela EPUSP

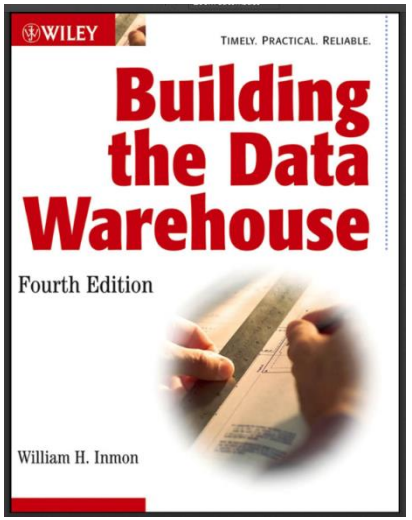
Bibliografia



Sistemas de Banco de Dados
Elmasri / Navathe 6ª edição



Sistema de Banco de Dados
Korth, Silberschatz – Sixth Editon

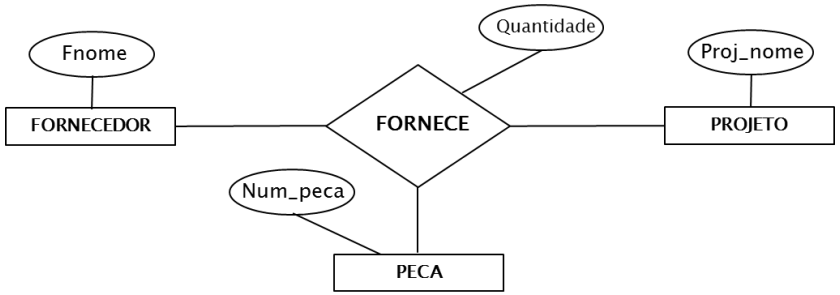


Building the Data Warehouse –
William H. Inmon – Fourth Edition

Por que não usar o Modelo Entidade Relacionamento tradicional para DW?

MER – Tradicional

- ✓ O modelo Entidade Relacionamento (**modelo conceitual**) é aplicável para aplicações **transacionais** e **não** adequado para consultas no mundo dos negócios (tomada de decisão);
- ✓ Esse modelo não foi desenhado para armazenar dados históricos. Constantemente há atualização na base de dados e informações históricas são perdidas;
- ✓ Modelos lógicos são derivados do modelo entidade relacionamento, no qual aplicam-se técnicas visando a eliminação de redundâncias e assegurando consistência de dados;
- ✓ Essas técnicas são conhecidas por **Normalização** de Banco de Dados.



Normalização de Relações

- O processo de normalização, proposto inicialmente, por Codd (1972) permite que, por meio de uma série de testes, certifique-se que uma relação satisfaz certa **FORMA NORMAL**.

- Inicialmente, Codd propôs **3** formas normais, que ele chamou de primeira, segunda e terceira formas normais.

- Todas essas formas normais estão baseadas em uma única ferramenta analítica: as **dependências funcionais** entre os atributos de uma relação.

Qual a finalidade da Normalização de Relações ?

Normalização de Relações

- Pode ser considerada um processo de se analisar os esquemas de relação com base em suas **DEPENDÊNCIAS FUNCIONAIS** e chaves primárias para conseguir as propriedades desejadas de:

  - ✓ ***Minimização da Redundância;***
  - ✓ ***Minimização das Anomalias de Inserção, Exclusão e Atualização dos dados.***

Normalização - Definição

- A forma normal de uma relação refere-se à condição de forma normal mais alta a que ela atende e, portanto, indica o grau ao qual ela foi normalizada.
- Se uma relação está na terceira formal normal, então também está na segunda e primeira formas normais.
- Se uma relação está na segunda forma normal, então também está na primeira forma normal.
- Para finalidades práticas, em geral, normaliza-se até a **terceira forma normal**.
- **2FN** e **3FN** atacam diferentes problemas, mas por motivos históricos, é comum seguir a ordem de **1FN**, **2FN** e **3FN** no processo de normalização das relações.

Resumindo ...

- Normalização é um procedimento empregado em projetos de banco de dados, visando a **minimização** das redundâncias e a diminuição da chance dos dados estarem **inconsistentes**.

Projeto DW

- Podem conter **consultas complexas** acessando uma quantidade **muito grande** de registros;
- Data warehouses têm como característica não usarem as regras de normalização;
- Assim, no projeto de Data Warehouses, deve-se quebrar o paradigma da eliminação de redundâncias, priorizando o ganho de desempenho nas consultas.

Exemplo – Dados normalizados x Dados Desnormalizados

✓ Tabelas Normalizadas

✓ Esse é um exemplo de tabelas normalizadas, evitando assim a redundância de dados.

Codigo	Nome	Endereco	Telefone
1	Jerônimo Freitas	R. Hum, 1200	31-3131-3131
2	Fernando Amaral	R. Dois, 1300	31-3232-3232

Tabela Clientes

CodigoPedido	CodigoCliente	DataPedido	DataEntrega
1	2	25/04/2012	30/04/2012
2	1	26/04/2012	02/05/2012

Tabela Pedidos

CodigoPedido	CodigoDoItem	Quantidade	ValorUnitario
1	34	2	10,5
1	21	1	7
2	76	4	8,9
2	89	7	3,45

Tabela Itens do Pedido

Modelagem do DW

- ✓ Em um **Data Warehouse**, tem-se apenas uma tabela com todos os dados importantes para a geração das informações que vão auxiliar no processo de tomada de decisão.

CodCli	NomeCli	CodPed	DataPedido	DataEntrega	Item	Qtd	VlrUnit	AnoMesPed	AnoMesEnt
1	Jerônimo Freitas	2	26/04/2012	02/05/2012	Item 1	4	8,9	201204	201205
1	Jerônimo Freitas	2	26/04/2012	02/05/2012	Item 2	7	3,45	201204	201205
2	Fernando Amaral	1	25/04/2012	30/04/2012	Item 3	2	10,5	201204	201204
2	Fernando Amaral	1	25/04/2012	30/04/2012	Item 4	1	7	201204	201204

Modelagem do DW

CodigoPedido	CodigoDoItem	Quantidade	ValorUnitario
1	34	2	10,5
1	21	1	7
2	76	4	8,9
2	89	7	3,45

- ✓ Ué, mas onde está o código do item ?
- ✓ AnoMedPed e AnoMesEnt ?

CodCli	NomeCli	CodPed	DataPedido	DataEntrega	Item	Qtd	VlrUnit	AnoMesPed	AnoMesEnt
1	Jerônimo Freitas	2	26/04/2012	02/05/2012	Item 1	4	8,9	201204	201205
1	Jerônimo Freitas	2	26/04/2012	02/05/2012	Item 2	7	3,45	201204	201205
2	Fernando Amaral	1	25/04/2012	30/04/2012	Item 3	2	10,5	201204	201204
2	Fernando Amaral	1	25/04/2012	30/04/2012	Item 4	1	7	201204	201204

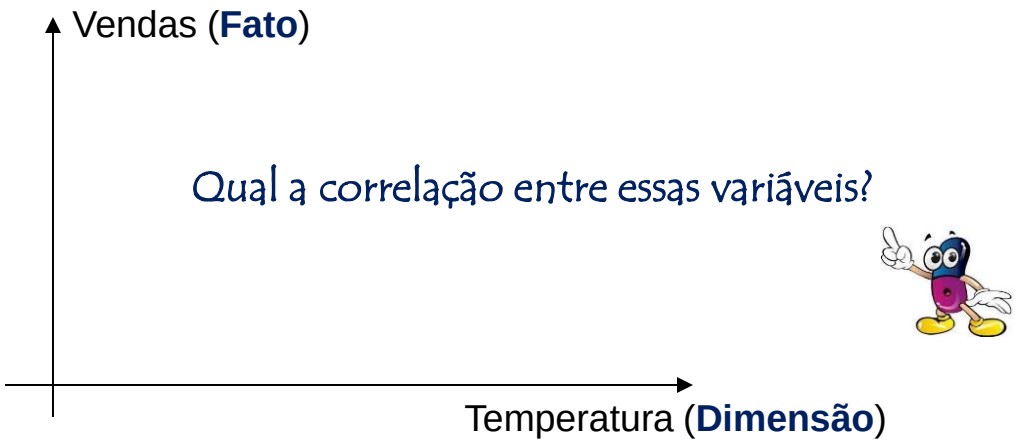
Modelagem do DW

- ✓ Só serão inseridos no Data Warehouse dados preponderantes, que vão retornar informações relevantes para o **negócio** em questão ou facilitar a busca destas informações.
- ✓ Os campos **AnoMesPed** e **AnoMesEnt**, no formato em que estão, vão facilitar o agrupamento das vendas por mês e ano sem a necessidade de usar funções do banco de dados;
- ✓ Por exemplo, o gestor que saber o **cliente que mais comprou durante o mês** ou o **item mais vendido**;
- ✓ Não tem o **código do item**, porque, o gestor quer saber o **nome do item** que mais vendeu, o código dele será necessário no ponto de venda, para facilitar a digitação do pedido.

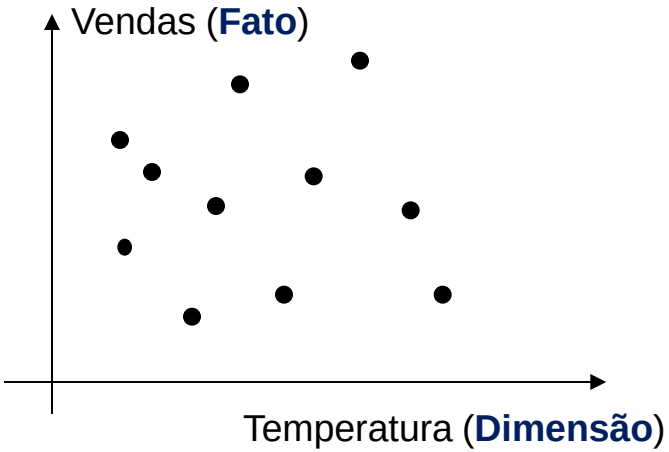
CodCli	NomeCli	CodPed	DataPedido	DataEntrega	Item	Qtd	VlrUnit	AnoMesPed	AnoMesEnt
1	Jerônimo Freitas	2	26/04/2012	02/05/2012	Item 1	4	8,9	201204	201205
1	Jerônimo Freitas	2	26/04/2012	02/05/2012	Item 2	7	3,45	201204	201205
2	Fernando Amaral	1	25/04/2012	30/04/2012	Item 3	2	10,5	201204	201204
2	Fernando Amaral	1	25/04/2012	30/04/2012	Item 4	1	7	201204	201204

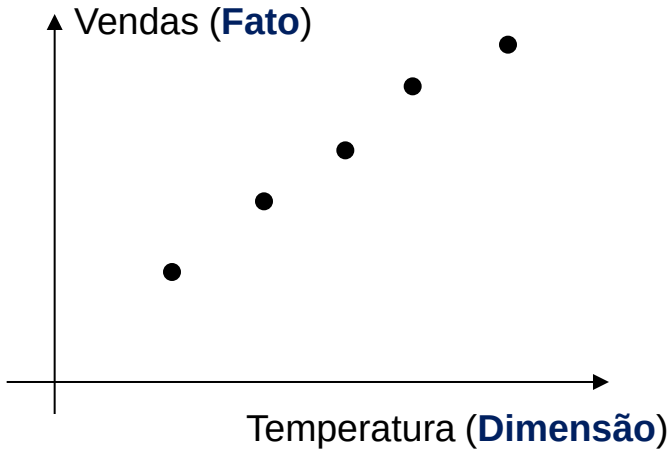
Modelagem Multidimensional

- ✓ Tem por objetivo descobrir se uma determinada variável de um objeto do problema está relacionada com outra;
- ✓ Por exemplo, deseja-se saber se a venda de um sorvete aumenta com a temperatura;
- ✓ Ou deseja-se responder a pergunta: Quando a temperatura aumenta se vende mais sorvete?
- ✓ Por meio da modelagem, tenta-se descobrir a correlação entre essas variáveis.



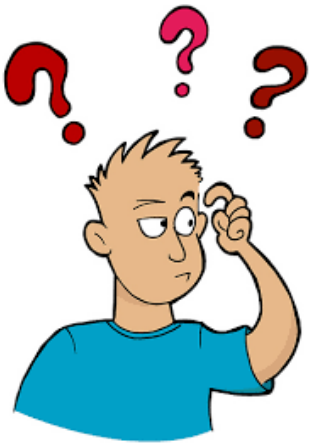
Modelagem Multidimensional – Possibilidades



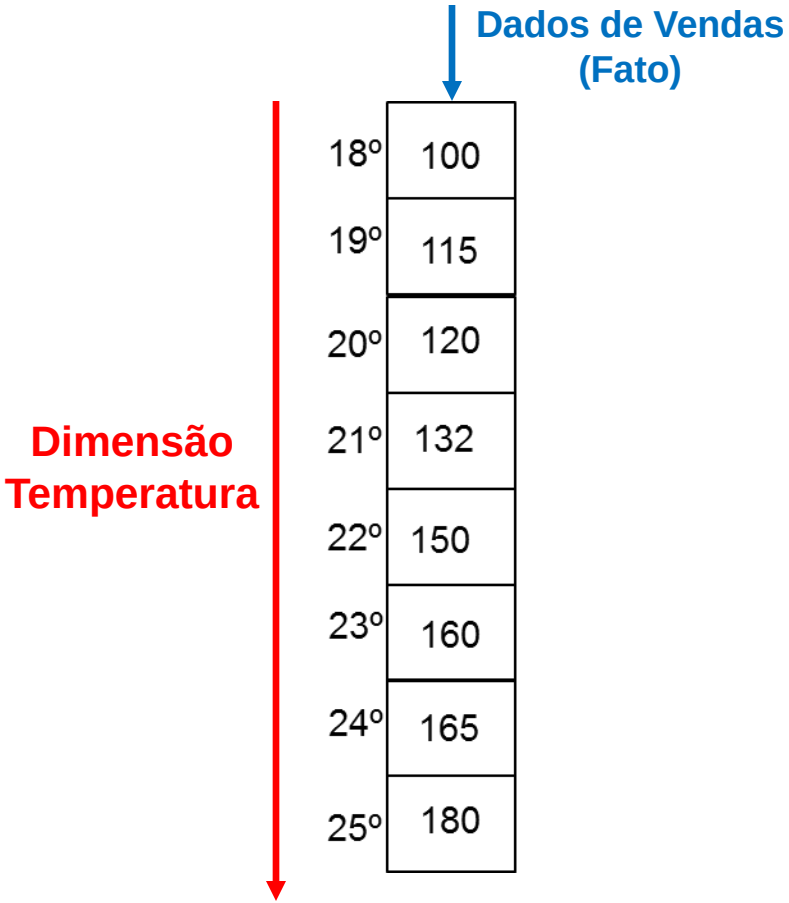


Qual a correlação entre essas dimensões?

Como se obter essa correlação entre os dados?



Visualização sob uma dimensão



Visualização sob duas dimensões

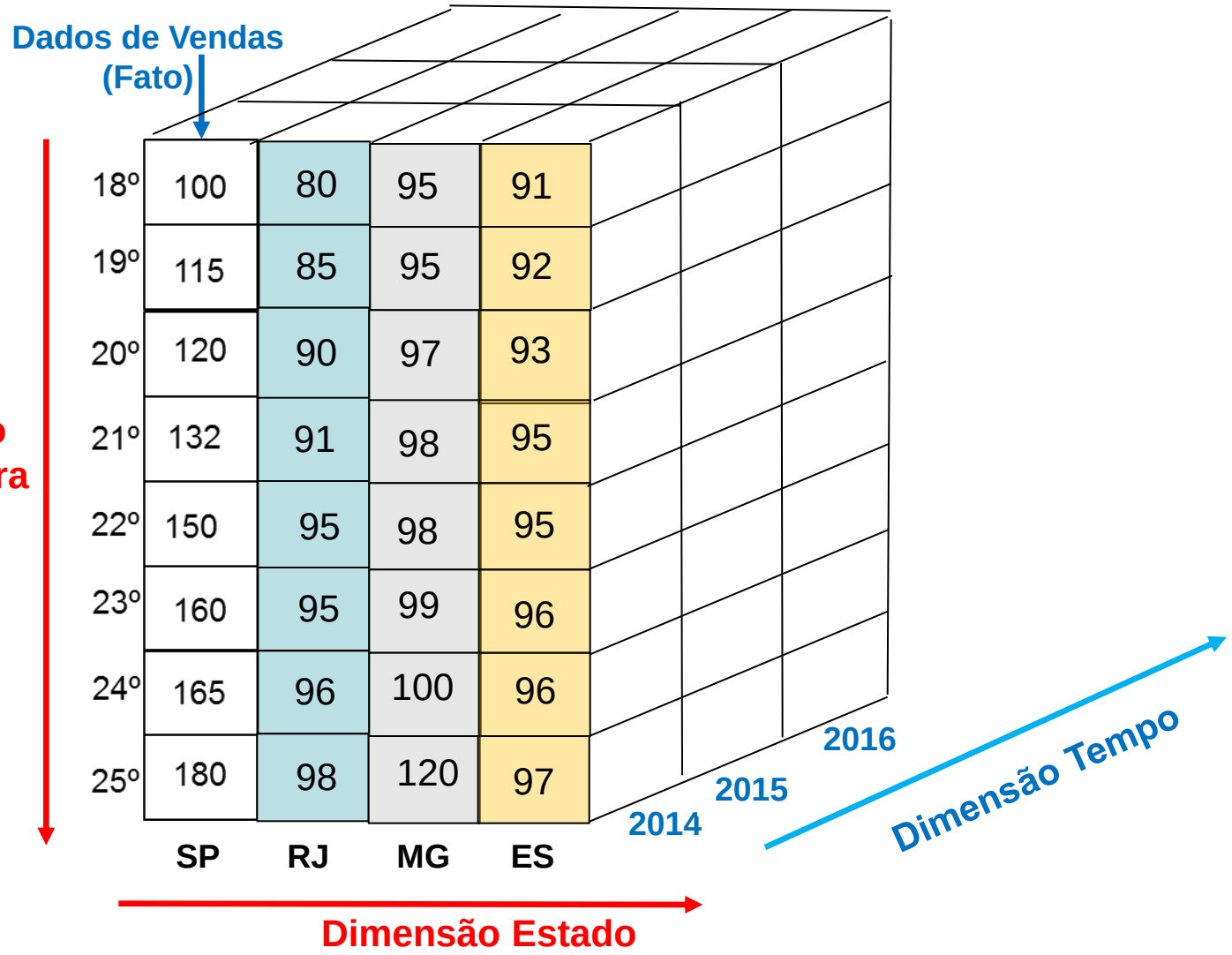
Dados de Vendas (Fato)

	SP	RJ	MG	ES
18º	100	80	95	91
19º	115	85	95	92
20º	120	90	97	93
21º	132	91	98	95
22º	150	95	98	95
23º	160	95	99	96
24º	165	96	100	96
25º	180	98	120	97

Dimensão Temperatura

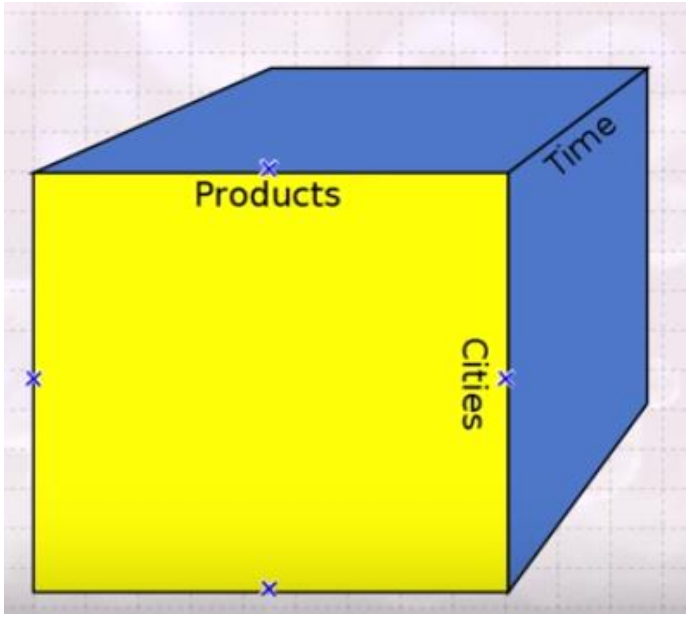
Dimensão Estado

Visualização sob três dimensões



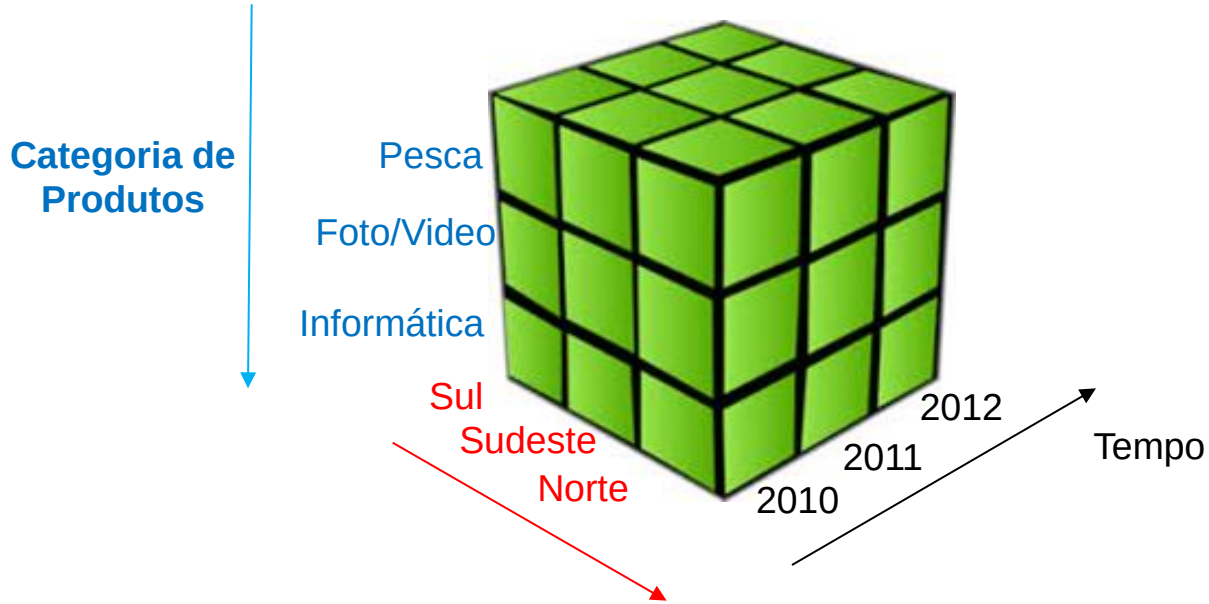
Cubo de Dados

- ✓ Por meio do cubo de dados, pode-se visualizar os dados em diferentes dimensões;
- ✓ Num cubo de dados, definem-se fatos e dimensões;
- ✓ Exemplo: O fato corresponde aos valores correspondentes às vendas. As dimensões são as perspectivas de visualização dos dados. Nesse caso, aos produtos e em que cidades esses produtos foram vendidos.



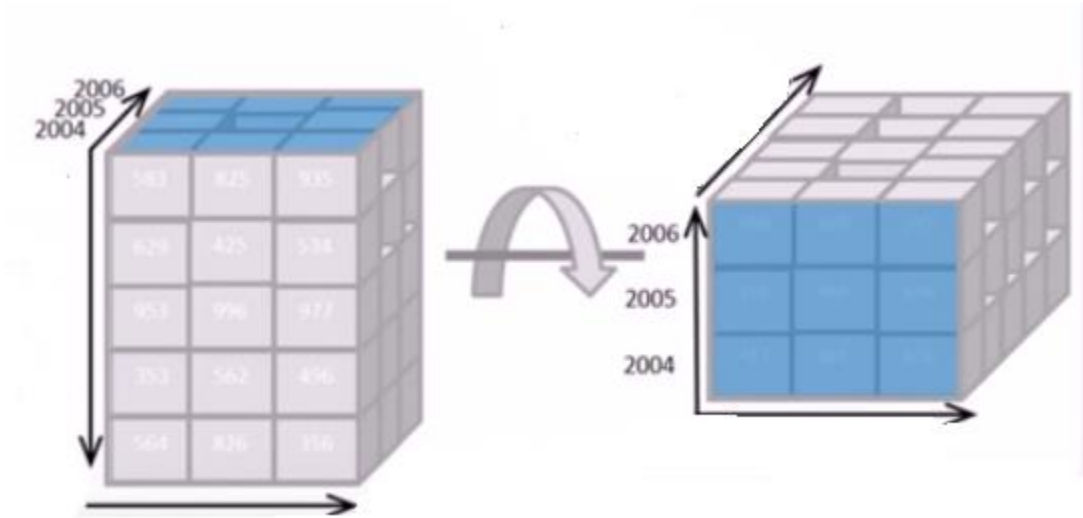
Exemplo - Cubo de Dados

- ✓ Dentro de cada célula do cubo está o objeto da Análise (Vendas realizadas – quanto foi vendido)
- ✓ O fato está dentro de cada célula do cubo.



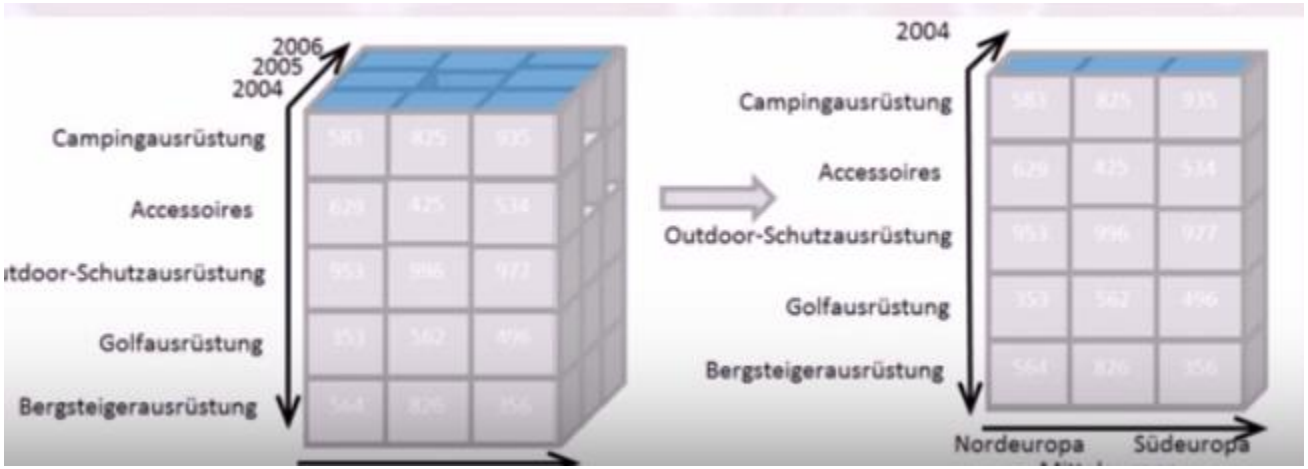
Operações – Pivoting (Rotation)

- ✓ Pode-se rotacionar o cubo e se visualizar o relacionamento entre as variáveis.



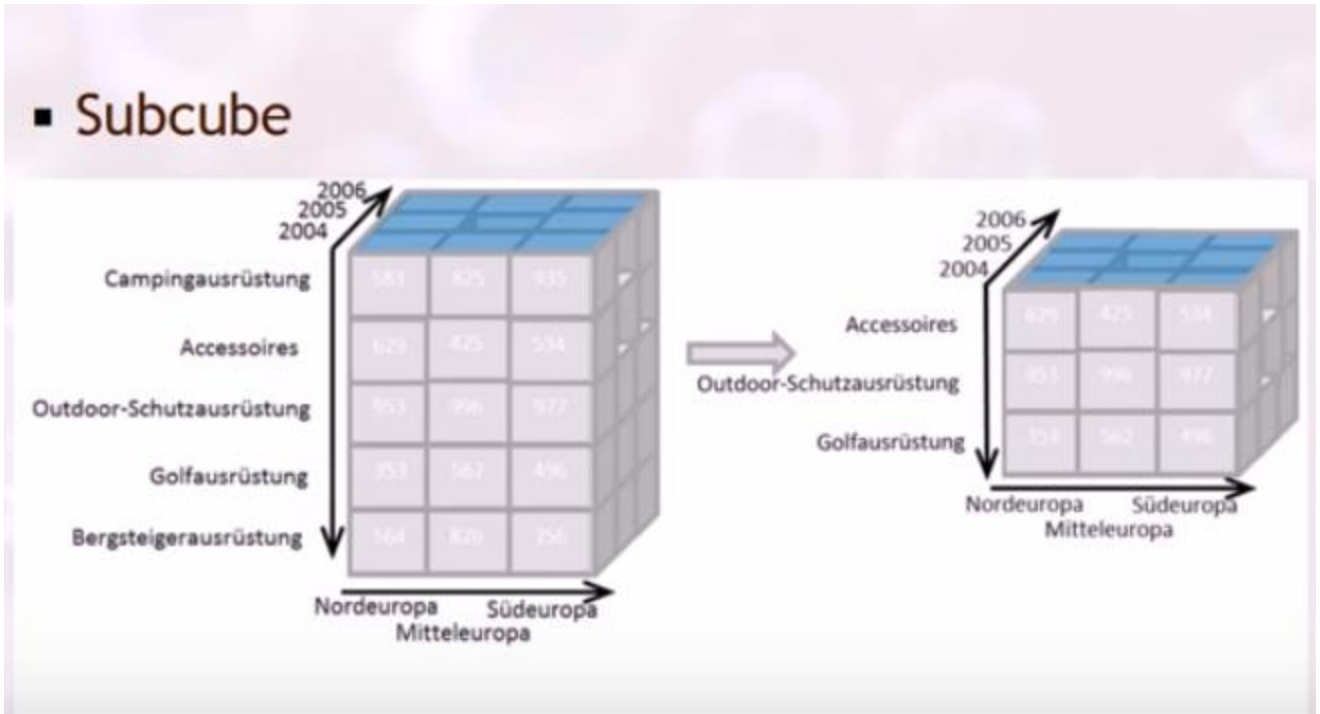
Operações – Slice (Fatia)

- ✓ Subconjunto retangular do cubo;
- ✓ Valor simples de uma dada dimensão;
- ✓ Refere-se ao cubo com menos dimensões;
- ✓ Das três dimensões, trabalha-se com duas.



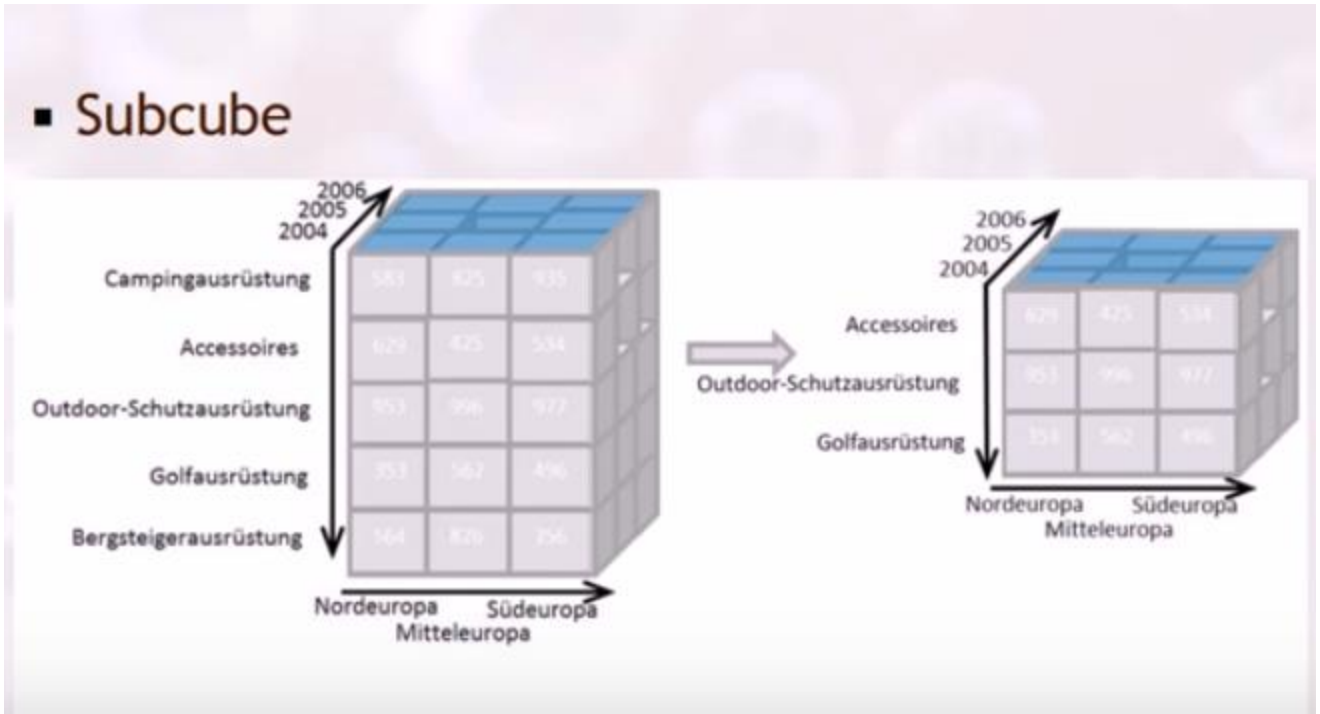
Operações – Dice (Sub-cubo)

- ✓ Trabalha-se com um subconjunto do cubo;



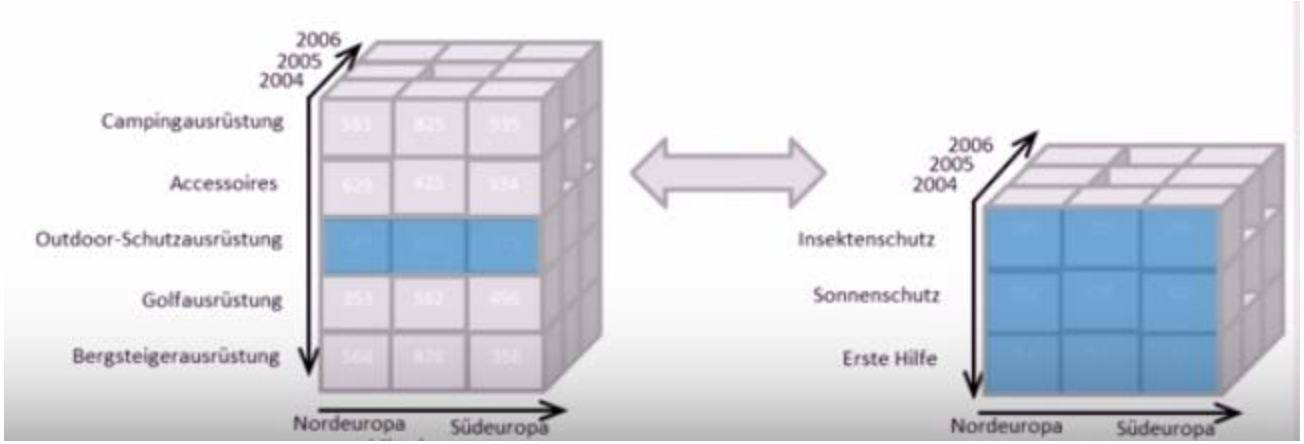
Operações – Dice (Sub-cubo)

- ✓ Trabalha-se com um subconjunto do cubo;



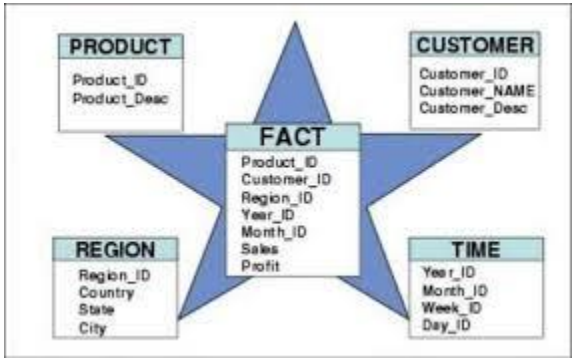
Operações – Roll-up e Drill-Down

- ✓ Roll-up: Maior granularidade (menos detalhes, grãos mais grossos). Exemplo: sumarização por região;
- ✓ Drill-down: Menor granularidade (mais detalhes, grãos mais finos). Exemplo: Detalhes por cidade.



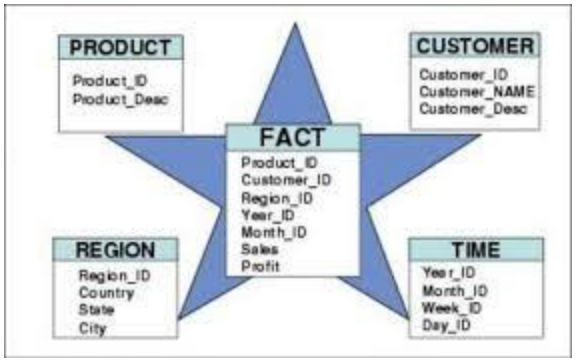
DW – Modelo Estrela (Star)

- ✓ Uma das formas de se modelar um **DW** é por meio do modelo Estrela (**Star Schema**);
- ✓ Nesse modelo, empregam-se os mesmos conceitos já utilizados na Modelagem Entidade Relacionamento (como entidades, atributos e relacionamentos);
- ✓ O modelo foi criado por Ralph Kimball (1998);
- ✓ A principal característica do modelo é a presença de dados com **redundância** para a obtenção de melhor **desempenho** das consultas.



DW – Modelo Estrela (Star)

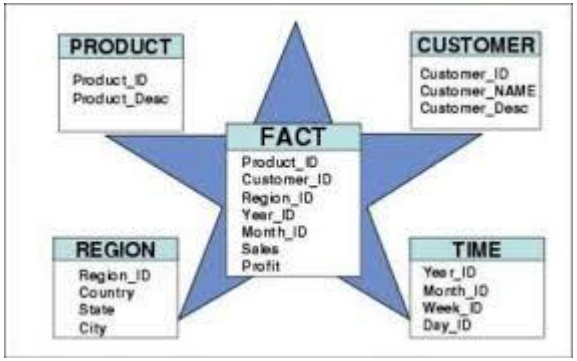
- ✓ O nome **Star Schema** foi adotado pela semelhança com uma **estrela;**
- ✓ O esquema é composto por uma tabela dominante, denominada **Tabela de Fatos**, no centro, rodeada por tabelas auxiliares denominadas **Tabela Dimensão**;
- ✓ Em um esquema estrela, estes dois tipos de tabela são combinados de forma que se tenha uma única Tabela Fato, e várias Tabelas Dimensão detalhando as informações contidas na Tabela Fato.
- ✓ Uma característica marcante deste esquema é o fato dele ser **desnormalizado**;
- ✓ A **desnormalização** auxilia na **redução** o número de **joins** utilizados nas consultas, facilitando sua escrita e melhorando seu desempenho.



DW – Modelo Estrela
Tabela Fato

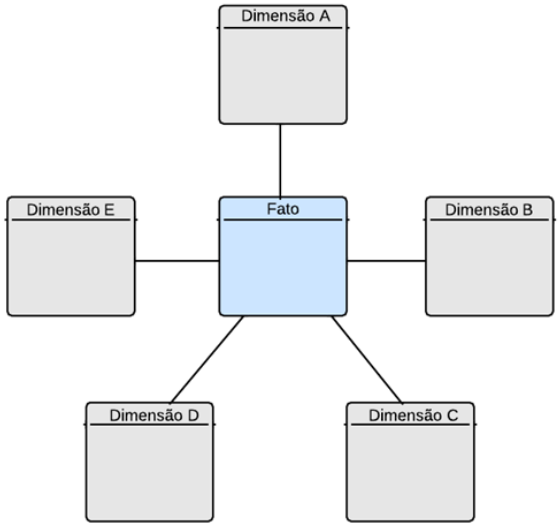
- ✓ Armazenam instâncias da realidade modelada para o negócio que podem, de alguma forma, serem medidas quantitativamente;

- ✓ Este tipo de tabela armazena uma **quantidade muito grande de informações**, sendo sua **chave primária composta por um conjunto de chaves estrangeiras** que apontam para seus respectivos **detalhamentos** nas Tabelas **Dimensão**;



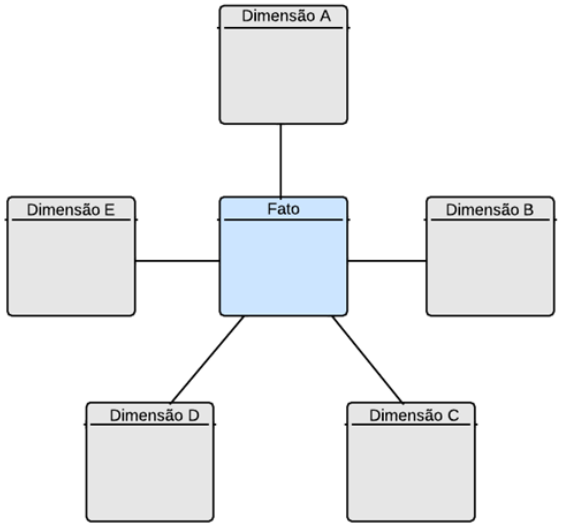
DW – Modelo Estrela
Tabela Dimensão

- ✓ **Qualificam** as informações provenientes da Tabela Fato.
- ✓ Através dela pode-se analisar os dados sob **múltiplas perspectivas** (dimensões) . Por exemplo, podemos ter Dimensões como Produto, Região e Tempo em um DW.
- ✓ As Tabelas Dimensão que compõem um esquema estrela não são normalizadas e armazenam, usualmente, **menor quantidade de dados** quando comparadas com a Tabela de Fatos.

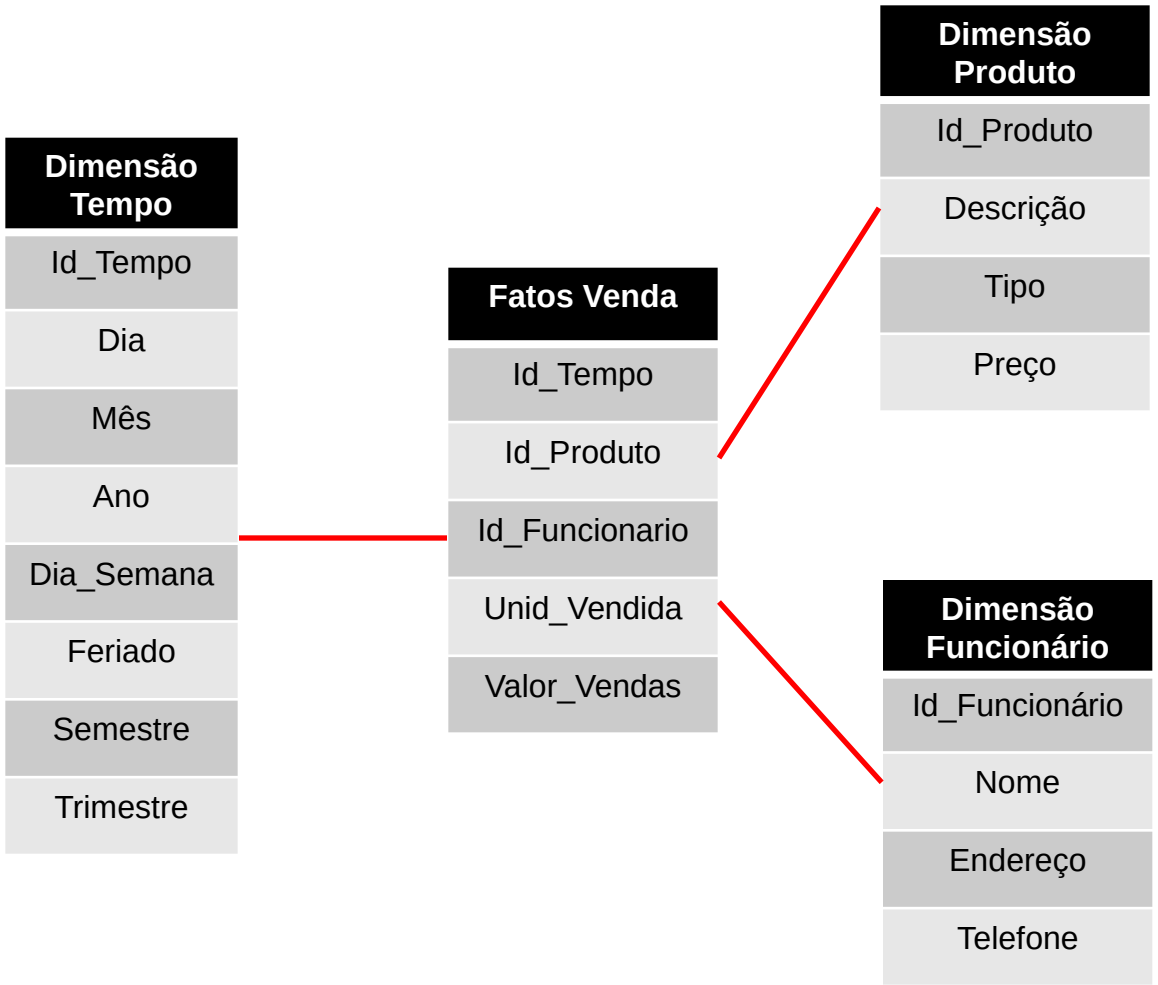


DW – Modelo Estrela

- ✓ A Tabela de Fatos armazena uma **grande quantidade de dados históricos**, obtidos a partir da intersecção de todas as dimensões da estrela;
- ✓ **Cada dimensão tem uma chave primária** que corresponde a um dos campos da chave da Tabela de Fatos;
- ✓ A dimensão **Tempo** é sempre integrante da chave primária e é na Tabela Fato onde se armazena os indicadores de desempenho (medidas) do negócio.



Esquema Estrela - Exemplo



Esquema Estrela – Exemplo

Tabela Dimensão - Produto

Id_Produto	Descrição	Tipo	Preço
101	Espaguete	Massa	10
102	Gnochi	Massa	15
103	Alcatre	Carne	30
104	Cupim	Carne	21

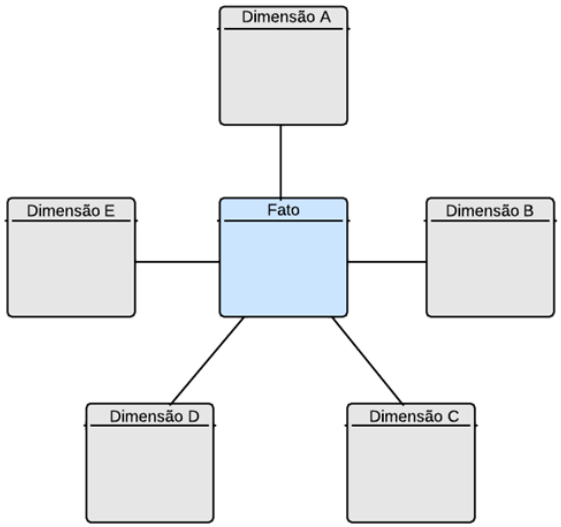
Tabela Fato

Id_Tempo	Id_Produto	Id_Funcionario	Unid. Vendidas	Valor Vendas
991020	101	200	10	500
991021	101	200	13	650
991022	101	200	15	700
991023	101	200	20	1000

DW – Modelo Estrela

- ✓ A **consulta ocorre inicialmente nas Tabelas de Dimensão** e depois nas Tabelas Fato, assegurando assim a precisão dos dados através de uma estrutura completa de chaves, onde não é preciso percorrer todas as tabelas;

- ✓ Esse procedimento, garante acesso mais eficiente e o mais alto desempenho possível;

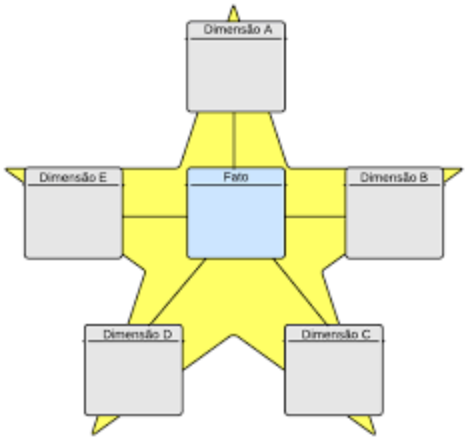


DW – Modelo Estrela - Observações

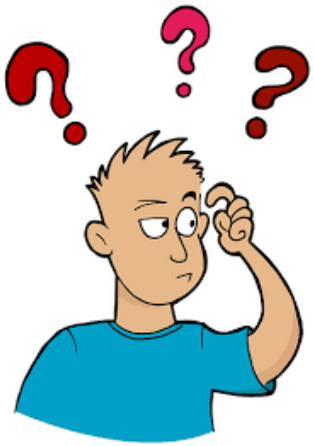
- ✓ No esquema estrela, a tabela **Fato** desempenha um papel de tabela dominante;
- ✓ O esquema é flexível para suportar a inclusão de novos elementos de dados;
- ✓ Essa flexibilidade se dá na medida em que todas as tabelas Fato e Dimensão podem ser alteradas simplesmente acrescentando-se novas colunas às mesmas.

DW – Documentação

- ✓ Estrutura de dados segundo a visão do programador;
- ✓ Estrutura de dados segunda a visão dos Analistas de Negócios;
- ✓ Fonte de dados que alimenta o DW;
- ✓ Modelagem dos Dados;
- ✓ Metadados.



Por onde iniciar a implementação de um DW?



DW – Implementação

- ✓ Esta é uma das principais questões feitas pelos consultores e Engenheiros de Software quando planejam implementar um DW;
- ✓ Há várias opções: DW Corporativo, DW's departamentais , DW's funcionais (marketing, financeiro, administrativo, etc);
- ✓ Implementação requer completa compreensão dos negócios da organização;
- ✓ Neste contexto, o Data Mart – em geral – é a opção mais interessante;
- ✓ À cada implementação de Data Mart, a equipe adquire experiência e refina implementações anteriores.

DW – Otimização

- ✓ Grande volume de dados;
- ✓ Exigências de desempenho em processos de carga e de consulta;
- ✓ Esses fatores distinguem sistemas OLTP dos sistemas de apoio a decisão.

DW – Bloco de Dados

- ✓ SGBD's armazenam dados em blocos de dados;
- ✓ Bloco representa a menor unidade de E/S para um SGBD;
- ✓ Aplicações **OLTP**, geralmente, lidam com tamanhos de blocos pequenos, em torno de 4K;
- ✓ **DW**, ao contrário, possuem tabelas com elevado número de colunas e grandes quantidades de dados;
- ✓ Para **DW**, em geral, os tamanhos de bloco são maiores, em torno de 64K.

DW – Considerações de Desempenho

- ✓ **Particionamento de Dados**: Por exemplo, tabela de vendas que contém dados históricos divididos por ano em 10 partições (2007, 2008, ... , 2016)
- ✓ **Índices B-tree**: Estruturas hierárquicas de indexação;
- ✓ **Indices Bitmap**: índice construído com mapeamento de bits nos atributos da tabela (vetores de bits).

Pentaho

- ✓ Software de código aberto (suite) desenvolvido com a Linguagem Java;
- ✓ A suite cobre as áreas de ETL (Extraction, Transformation e Load), reporting, OLAP e data mining;
- ✓ Componentes:
  - ✓ **PDI** – Pentaho Data Integration, conhecido como Kettle;
  - ✓ **PAS** – Pentaho Analysis Services, para **OLAP**;
  - ✓ **PR** – Pentaho reporting;
  - ✓ **PDM** – Pentaho Data Mining, derivado do projeto Weka para mineração de dados;
  - ✓ Pentaho **Dashboard**;
  - ✓ **PSW** – para a definição de tabelas Fato e Dimensão.