INTEGRAÇÃO DE DADOS I: MODELAGEM DIMENSIONAL

BIBLIOGRAFIA

- BARBIERI, Carlos. BI2-Business intelligence: modelagem & qualidade. Rio de Janeiro: Elsevier: Campus, 2011.
- KIMBALL, Ralph et al. The data warehouse lifecycle toolkit: Practical techniques for building data warehouse and business intelligence systems. 2nd ed. Indianopolis: Wiley, 2008.
- material da Profa. Viviane Dias

ETAPAS DE UM DATA WAREHOUSE

- Levantamento de requisitos
- Modelagem Multidimensional
- ETL (Extract Transform and Load)
- Visualização do resultados

LEVANTAMENTO DE REQUISITOS

- de acordo com o que o cliente precisa analisar

- fonte de dados operacional

MODELAGEM

Passo	Perguntas a serem feitas para o usuário	Elementos a serem definidos no modelo
1	O que estamos avaliando?	Fatos ou métricas (sempre um valor numérico)
2	Como serão avaliados ou analisados?	Dimensões de negócios relacionadas às métricas
3	Qual o nível mais baixo de detalhe das informações?	Granularidade das informações relacionadas as métricas
4	Como se espera agrupar ou sumariar as informações?	Hierarquia de agrupamento das informações em cada dimensão.

MODELAGEM DE DADOS

- Modelagem de dados
  - Peter Chen
  - James Martin
- Novas necessidades:

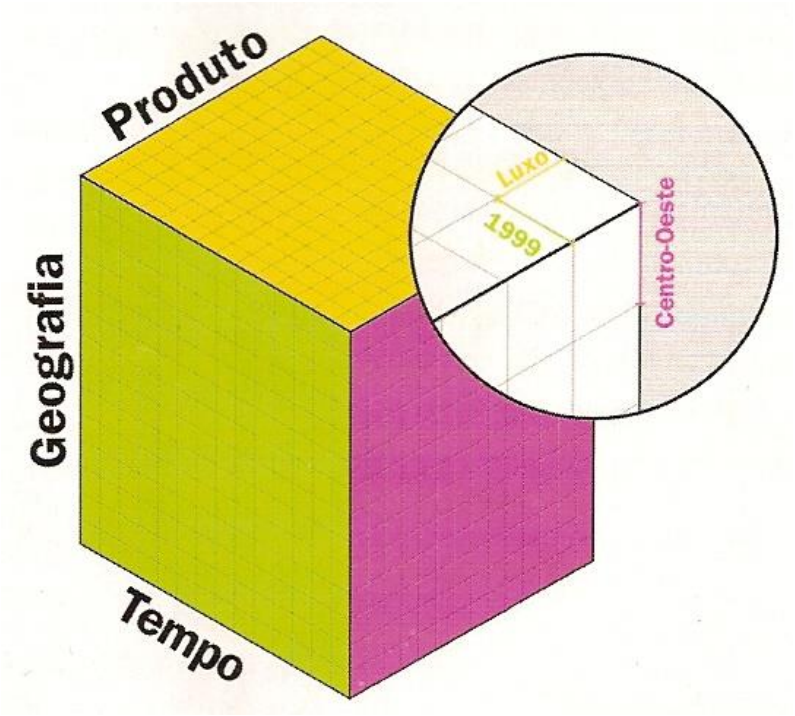
Competitividade

MODELAGEM DIMENSIONAL DE DADOS

- O que significa dimensional?
  - A estrutura dimensional modifica a ordem de distribuição de campos por entre as tabelas;
  - Permitindo uma formatação estrutural mais voltada para os muitos **pontos de entradas específicos** (as chamadas **dimensões**)
  - E menos para os dados granulares em si (os chamados **fatos**).

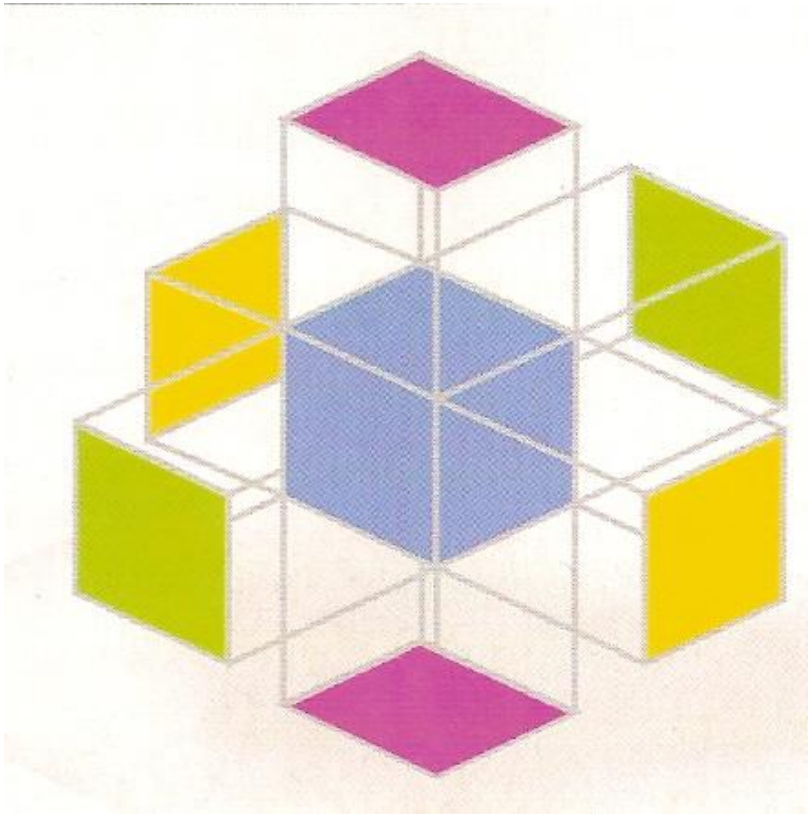
MODELAGEM MULTIDIMENSIONAL

- A representação dos dados em um DW é estruturada como um **cubo**, transmitindo a ideia de múltiplas dimensões.



MODELAGEM MULTIDIMENSIONAL

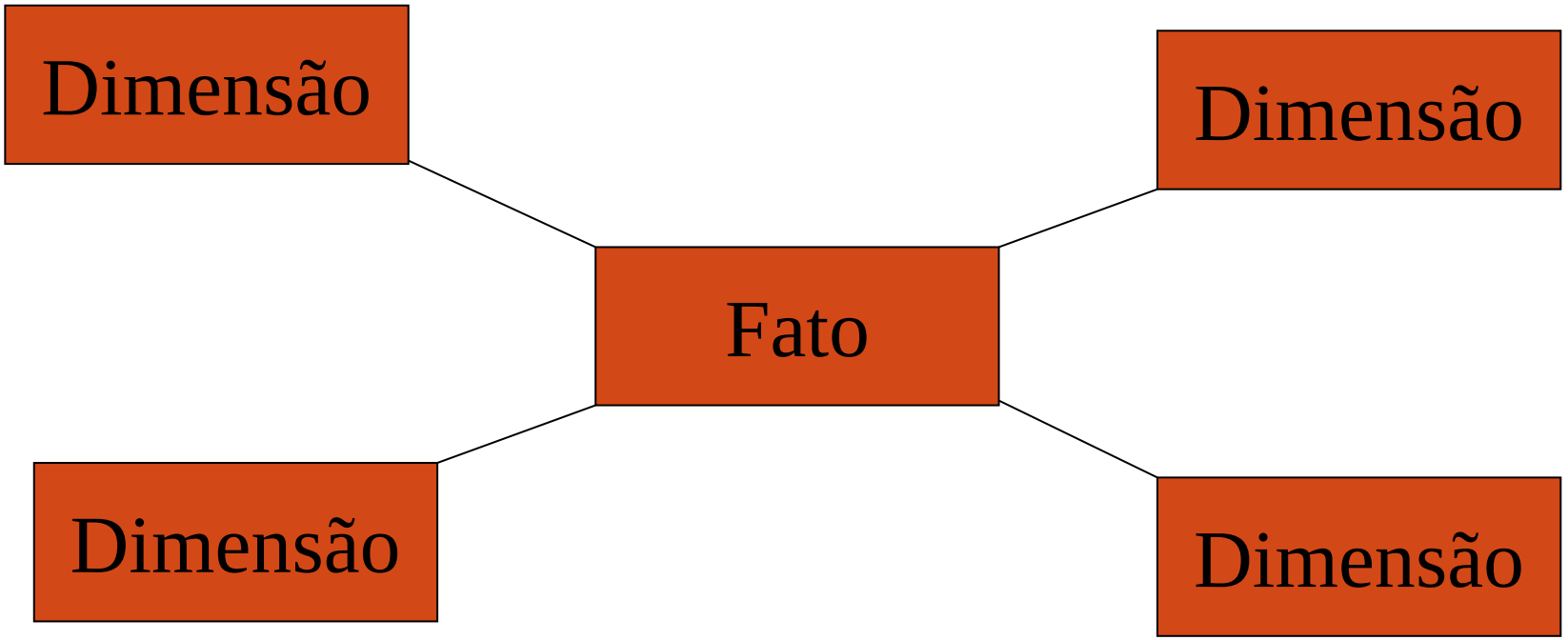
- A inclusão de dados no DW passa a ideia de **crescimento** na largura, comprimento e profundidade de **cubo**.



MODELAGEM MULTIDIMENSIONAL

- Construção:
  - começa pela definição da tabela denominada **Fato**;
  - em seguida, são definidos seus elementos relacionados, que são tabelas denominadas **Dimensões**;
  - Na interseção das dimensões, são obtidas as medidas, que são as medições numéricas da tabela **Fato**.

EXEMPLO DE MODELAGEM MULTIDIMENSIONAL



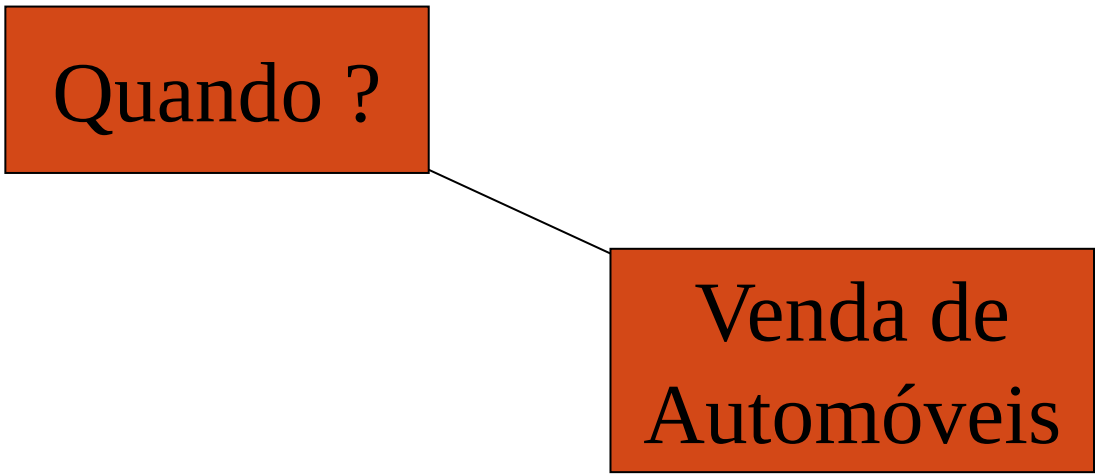
No **centro**, temos a tabela **Fato** e, nas **pontas**, **Dimensões**, ou seja, os elementos que participam de um Fato

Star Schema – Esquema Estrela

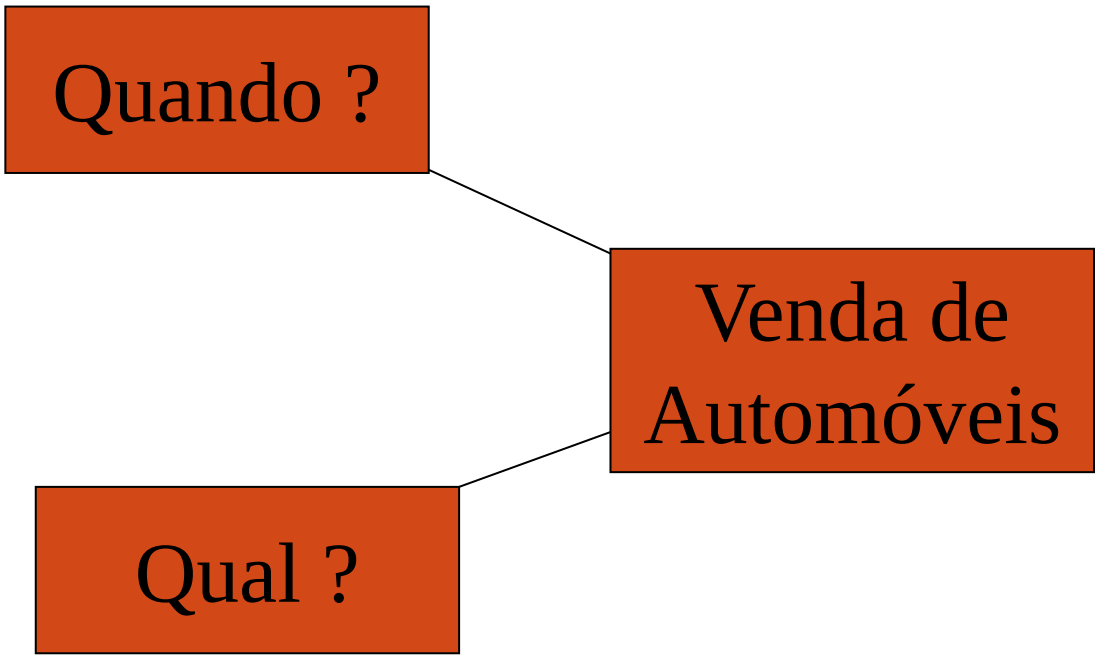
EXEMPLO DE MODELAGEM MULTIDIMENSIONAL

Venda de Automóveis

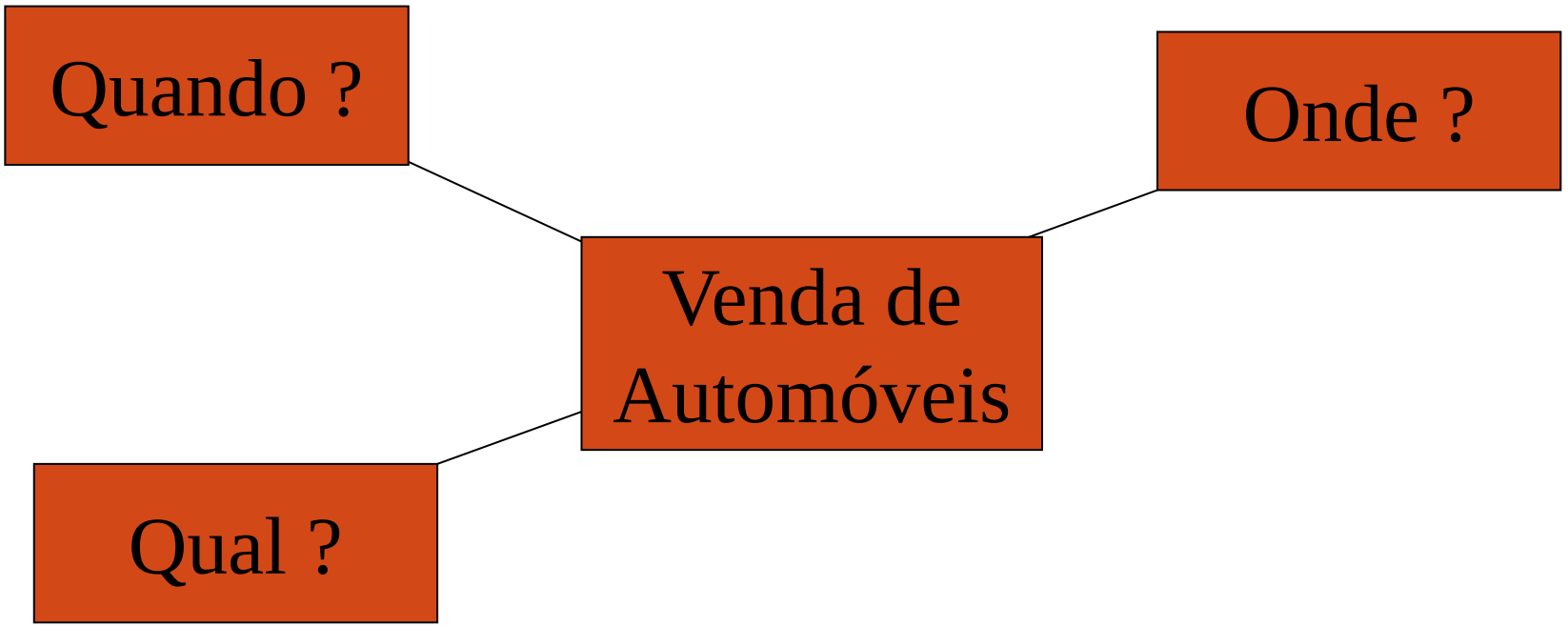
EXEMPLO DE MODELAGEM MULTIDIMENSIONAL



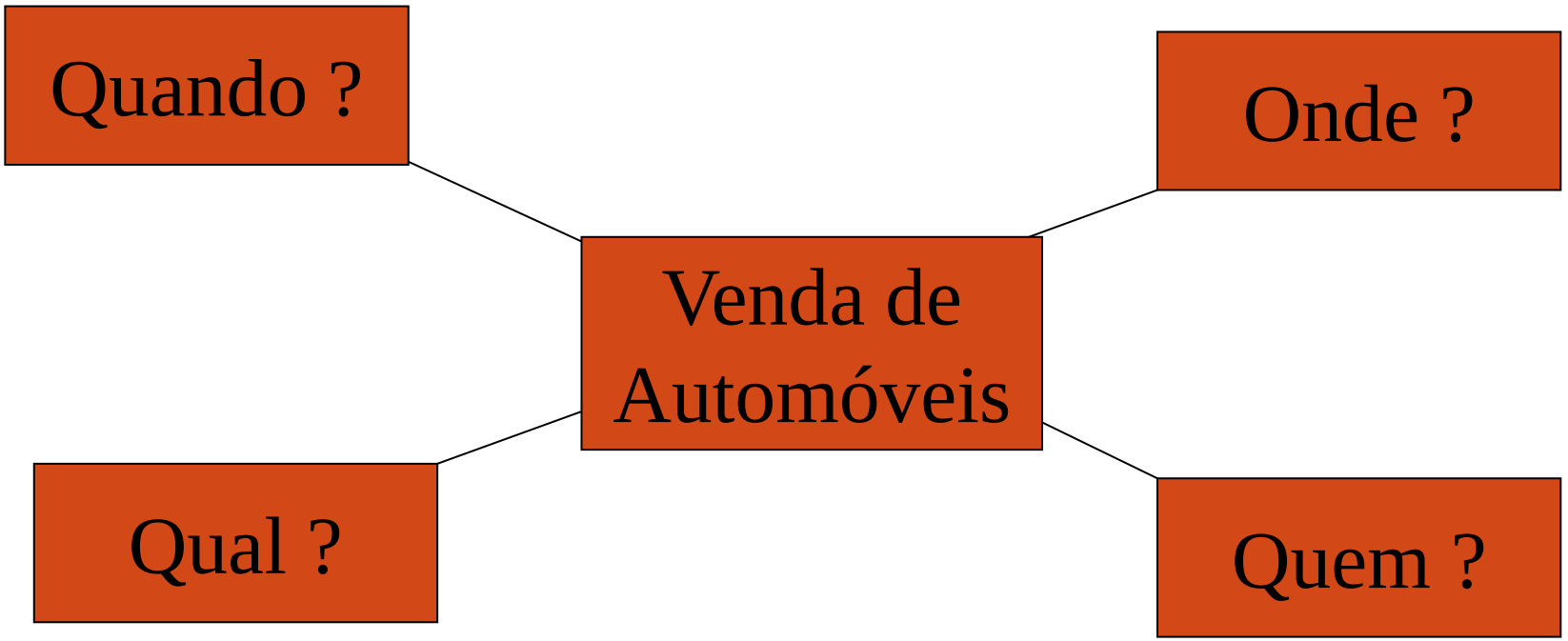
EXEMPLO DE MODELAGEM MULTIDIMENSIONAL



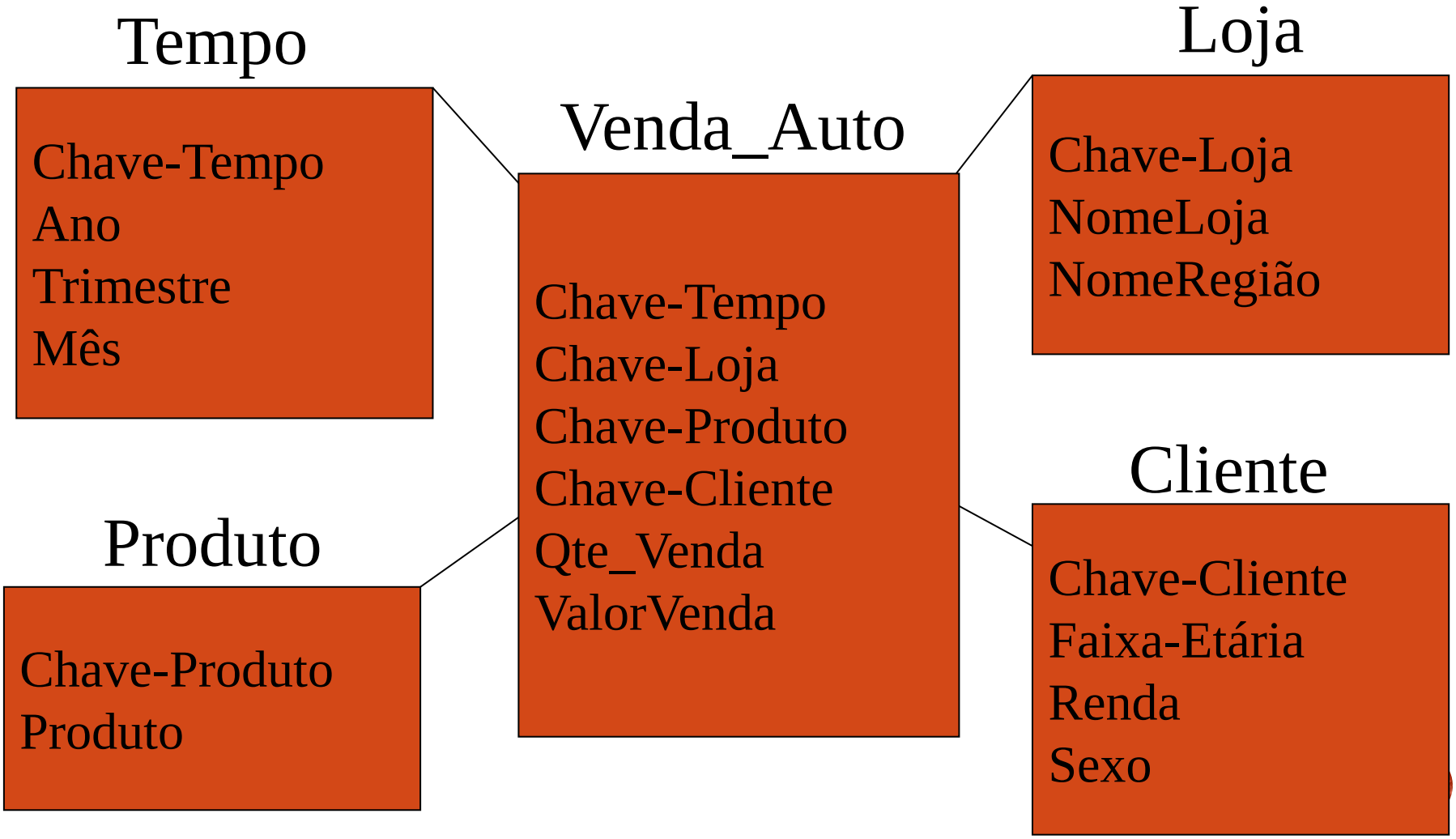
EXEMPLO DE MODELAGEM MULTIDIMENSIONAL



EXEMPLO DE MODELAGEM MULTIDIMENSIONAL



EXEMPLO DE MODELAGEM MULTIDIMENSIONAL



MODELAGEM: DEFININDO FATOS E MÉTRICAS

- O que queremos avaliar no DW/Data Mart?
  - **Fatos:**
    - números que serão medidos e analisados através das diferentes dimensões de negócios
      - números que o usuário lida no seu dia a dia
  - Definida a área de negócios, responde-se a pergunta:
    - **o que estamos avaliando?**

MODELAGEM: EXEMPLO

Área comercial de uma rede de lojas de varejo, tomando por base as possíveis necessidades de informação de um Gerente Comercial.

MODELAGEM: DEFININDO FATOS E MÉTRICAS

- **Cenário**:
  - o Gerente Comercial de uma rede de lojas quer avaliar:
    - quantidade de itens vendidos, valor da venda, custo de cada um dos itens e a margem alcançada
  - Onde estão estes valores/dados?
    - são originários de sistemas transacionais onde é mantida cada uma destas métricas
    - **Obs.: nem sempre as métricas são originadas de um só sistema.**

MODELAGEM: DEFININDO FATOS E MÉTRICAS

- De acordo com o **Cenário**:

  - As métricas ou fatos que o usuário deseja avaliar são:

    - Valor da venda – realizado
    - Valor da venda – previsto
    - Quantidade de itens – realizada
    - Quantidade de itens – prevista
    - Preço médio de venda
    - Custo médio
    - Margem de venda
    - % de variação entre o realizado e o planejado

MODELAGEM: DEFININDO FATOS E MÉTRICAS

- Algumas métricas poderão ser calculadas durante o processo de extração, transformação e carga e serão armazenadas no DW, já calculadas, ou então;

- poderão ser calculadas diretamente, durante a consulta (pelas ferramentas OLAP).

MODELAGEM: DEFININDO FATOS E MÉTRICAS

- Definir as dimensões relacionadas às métricas

- Função:
  - reunir os atributos que serão utilizados para qualificar as consultas
  - seus valores serão utilizados para agrupar as métricas (fatos)

- Como identificar as dimensões?
  - perguntar ao usuário: **Como as métricas serão analisadas?**
    - ou seja, a partir de quais dimensões de negócio os fatos serão avaliados?

MODELAGEM: DEFININDO FATOS E MÉTRICAS

- **Por exemplo**: cada uma das métricas precisa ser analisada ao longo do tempo.
  - isso significa analisar a quantidade de itens vendidos por mês, ou talvez por dia;
  - pode-se querer comparar períodos de vendas:
    - analisando a quantidade de itens vendidos no último mês em comparação com o mesmo mês do ano anterior.

MODELAGEM: DEFININDO DIMENSÕES

- **Dica**: Converse com o cliente/usuário
  - dê exemplos/sugestões como se fosse um relatório de resultados;
  - deixe que ele informe as suas necessidades.

MODELAGEM: DEFININDO DIMENSÕES

- Considerando o exemplo, as dimensões identificadas são:
  - **Tempo**:
    - indica os períodos de tempo para a análise;
  - **Produto**:
    - indica quais produtos estão relacionados às métricas;
  - **Geografia**:
    - indica a região geográfica onde se encontram as lojas que efetuam as vendas.

MODELAGEM: DEFININDO DIMENSÕES: PRODUTO

- Conteúdo da dimensão Produto

Chave do produto	Item	Linha	Categoria
1	Lápis n° 2 – Faver Carel	Lápis	Escolar/Escritório
2	Caneta Clic azul - fina	Caneta	Escolar/Escritório
3	Caneta Clic vermelha - fina	Caneta	Escolar/Escritório
...	...	...	...
36	Caderno espiral 96 folhas - Clip	Caderno	Escolar/Escritório
37	Cartucho preto – Impressora Vulcan 482	Suprimento	Informática
38	Caixa presente 20x28	Embalagem	Escolar/Escritório
39	CD regravável - Tent	Suprimento	Informática
40	Bloco recibo Jordel	Impresso	Escolar/Escritório
41	Teclado computador Micel	Hardware	Informática
42	Personal Firewall Protect v3.4	Software	Informática
...	...	...	...

MODELAGEM: DEFININDO DIMENSÕES: TEMPO

- Conteúdo da dimensão Tempo

Chave do Tempo	Data	Mês	Semestre	Ano
1	01/01/2003	Janeiro 2003	1º. Semestre 2003	2003
2	02/01/2003	Janeiro 2003	1º. Semestre 2003	2003
3	03/01/2003	Janeiro 2003	1º. Semestre 2003	2003
...	...	...	...	...
50	19/02/2003	Fevereiro 2003	1º. Semestre 2003	2003
51	20/02/2003	Fevereiro 2003	1º. Semestre 2003	2003
...	...	...	...	...
214	02/08/2003	Agosto 2003	2º. Semestre 2003	2003
215	03/08/2003	Agosto 2003	2º. Semestre 2003	2003
216	04/08/2003	Agosto 2003	2º. Semestre 2003	2003
...	...	...	...	...
520	03/06/2004	Junho 2004	1º. Semestre 2004	2004
421	04/06/2004	Junho 2004	1º. Semestre 2004	2004
...	...	...	...	...

MODELAGEM: DEFININDO DIMENSÕES: GEOGRAFIA

- Conteúdo da dimensão Geografia

Chave da geografia	Loja	Cidade	Estado	Região
1	Shopping Anália Franco	São Paulo	SP	Sudeste
2	Leblon	Rio de Janeiro	RJ	Sudeste
3	Shopping Morumbi	São Paulo	SP	Sudeste
4	Guabiraba – Loja 1	Recife	PE	Nordeste
5	São José	Recife	PE	Nordeste
6	Guabiraba – Loja 2	Recife	PE	Nordeste
...	...	...	...	...
12	Central	Curitiba	PR	Sul
13	Shopping Mueller	Curitiba	PR	Sul
14	Cristo Rei	Curitiba	PR	Sul
15	Barra Shopping	Rio de Janeiro	RJ	Sudeste
		...		...

MODELAGEM: DEFININDO DIMENSÕES

- Verificar se cada métrica se relaciona com todas as dimensões definidas:
  - **cada métrica pode ser analisada ao longo de cada dimensão?**
    - **Exemplo**: Faz sentido analisar o valor das vendas por produto? E por Loja? E ao longo do tempo?

MODELAGEM: DEFININDO GRANULIDADE

- É importante saber qual o nível de detalhe, ou granularidade, mais baixo que será avaliado

- Dimensão Tempo:
  - Podemos questionar o usuário da seguinte forma: **Qual o nível de detalhe desejado?**
  - Faz sentido avaliar a métrica quantidade vendida por dia?

MODELAGEM: DEFININDO GRANULIDADE

- Para cada uma das métricas definidas vamos identificar qual o nível mais baixo de detalhe será armazenado no DW.

- Se para a dimensão **Tempo** o nível mais baixo for **dia**, então todas as métricas deverão ser obtidas com valores por dia.

MODELAGEM: DEFININDO GRANULIDADE

- **Estudo de Caso**:
  - Nível de granularidade mais baixo:
    - Dia → dimensão Tempo
    - Item de Produto → dimensão Produto
    - Loja → dimensão Geografia
  - O processo de ETL deve trazer os dados para o DW de acordo com a granularidade definida nas dimensões

MODELAGEM: DEFININDO GRANULIDADE

- **Estudo de Caso-Exemplo:**
  - A métrica valor da venda: deve ser o valor de venda realizado para cada item de produto em cada dia e em cada loja.

MODELAGEM: DEFININDO GRANULIDADE

- Informações necessárias para se preparar o DW

Tempo (Dia)	Produto (Item)	Geografia (Loja)	Valor da venda (R$)	Quantidade de itens	Preço médio de venda (R$)	...
05/01/2004	Lápis n° 2 – Faver Carel	Loja 04	78,00	65	1,20	...
05/01/2004	Lápis n° 2 – Faver Carel	Loja 06	150,00	125	1,20	...
05/01/2004	Caneta Clic azul - fina	Loja 04	117,60	84	1,40	...
05/01/2004	Caneta Clic vermelha - fina	Loja 04	39,20	28	1,40	...
...	...	...	...	...	...	...
23/03/2004	Caneta Clic azul - fina	Loja 06	123,00	82	1,50	...
23/03/2004	Bloco recibo Jordel	Loja 12	132,50	53	2,50	...
...	...	...	...	...	...	...

MODELAGEM: DEFININDO A HIERARQUIA DE AGRUPAMENTO DE INFORMAÇÕES

- Os dados estarão armazenados no DW de acordo com o nível de detalhes estabelecido pelo usuário;

- Porém, o usuário deseja informações como:
  - **Qual o total de canetas vendidas nas lojas de São Paulo no último semestre?**

MODELAGEM: DEFININDO A HIERARQUIA DE AGRUPAMENTO DE INFORMAÇÕES

- **Essa pergunta indica**:
  - deveremos nos preocupar com o agrupamento ou sumarização das informações no DW.

- Portanto, deve-se definir quais as possibilidades de **agrupamento** das informações que o usuário deseja
  - especificando a hierarquia desses agrupamentos – em cada dimensão

MODELAGEM: DEFININDO A HIERARQUIA DE AGRUPAMENTO DE INFORMAÇÕES

- **Estudo de Caso:**
  - Hierarquia natural – **dimensão tempo**
    - meses normalmente são agrupados em bimestres ou trimestres
      - que por sua vez são agrupados em semestres e em anos.

MODELAGEM: DEFININDO A HIERARQUIA DE AGRUPAMENTO DE INFORMAÇÕES

- Estudo de Caso:

- Volte a perguntar ao usuário
  - é importante saber o que o usuário necessita – já que algumas regras de negócios requerem agrupamentos temporais diferentes (até para o tempo)

MODELAGEM: DEFININDO A HIERARQUIA DE AGRUPAMENTO DE INFORMAÇÕES

- Estudo de Caso - considerar:
  - Dimensão tempo:
    dia→mês→semestre→ano
  - Dimensão produto: item de produto→linha de produto→categoria
  - Dimensão Geografia:
    loja→cidade→estado → região

MODELAGEM: NORMALIZAÇÃO

- Impulso:
  - Aplicar as regras para normalizar;
  - Se normalizarmos as tabelas dimensão, o BD levará mais tempo para recuperar as linhas;
  - Custo (processamento) muito alto;
  - Por ser uma base de consultas e de grande volume, devemos nos preocupar em favorecer o tempo de resposta aos usuários, mantendo as informações de forma redundante.

MODELAGEM: NORMALIZAÇÃO

- Aspecto que difere a modelagem de um DW/Data Mart da modelagem das bases operacionais / transacionais.

MODELAGEM: NORMALIZAÇÃO

- Tabelas dimensão não normalizadas
  - Star Schema

- É possível normalizar as dimensões
  - Snowflake Schema

MODELAGEM: ESQUEMA FÍSICO

- cada dimensão corresponde a uma tabela física na base de dados.

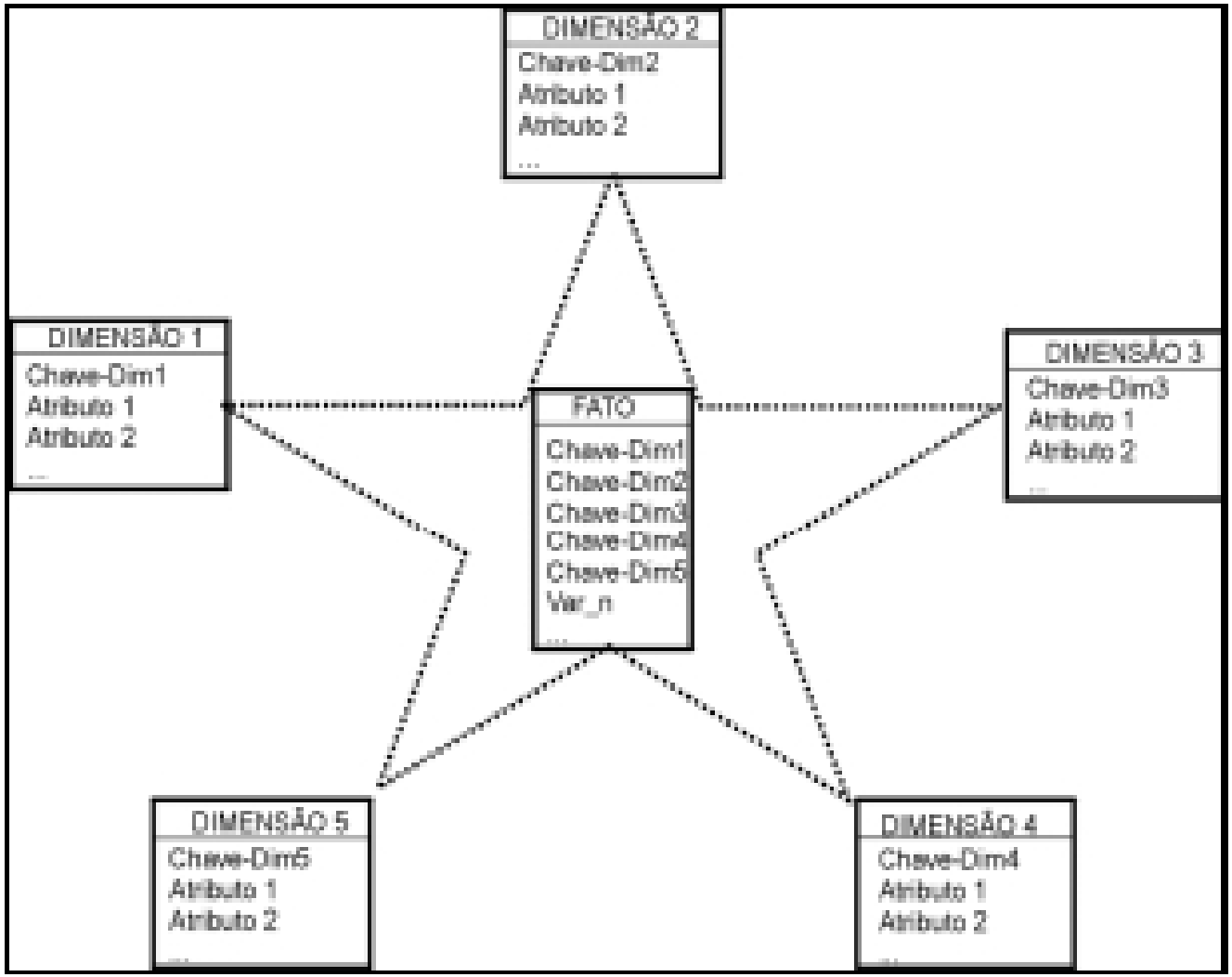
MODELAGEM: STAR SCHEMA

- Definir:
  - Tabela que conterá as métricas, ou valores, a serem analisados pelos usuários, através das informações representadas nas dimensões
- Tabela Fato: Quais informações serão analisadas
- Tabela Dimensão: Como serão analisadas

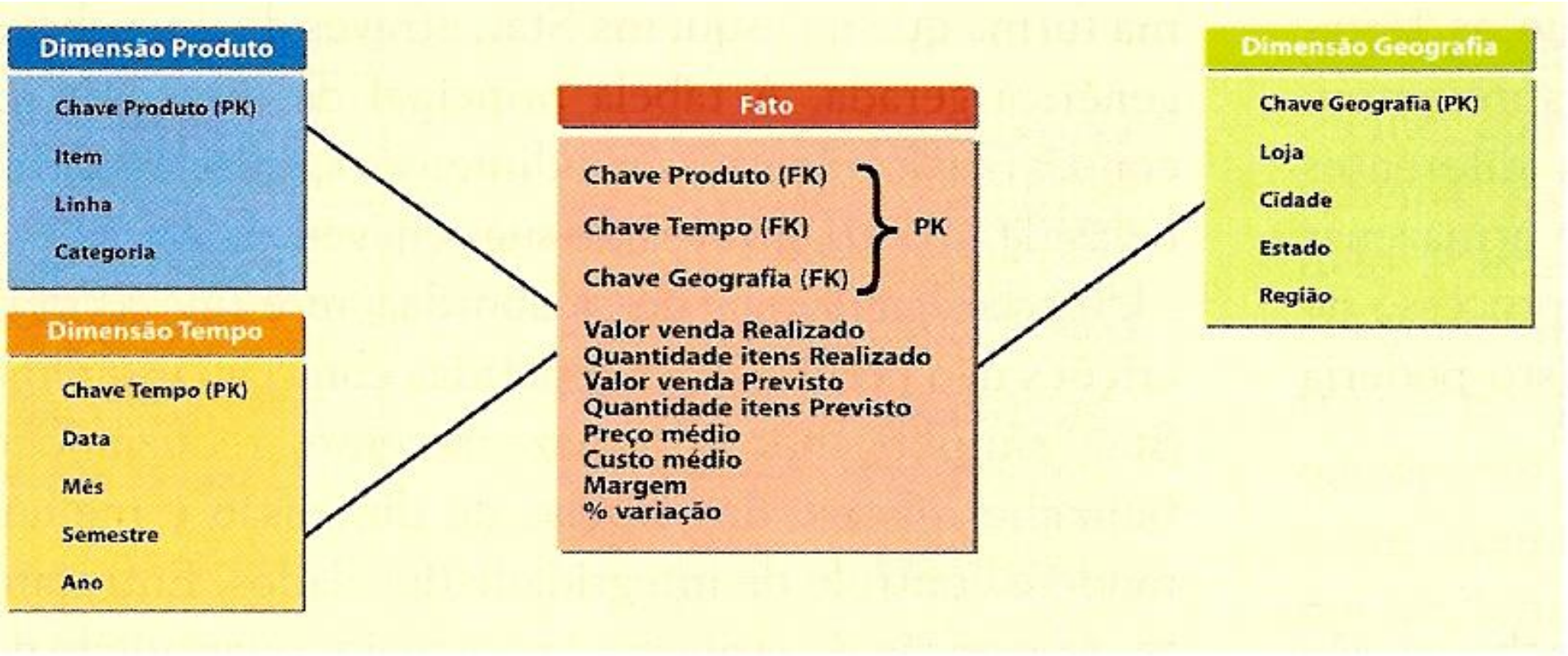
MODELAGEM: STAR SCHEMA

- Tabela Fato
  - Contém atributos chave e métricas ou fatos numéricos;
  - Ligada às tabelas dimensão através das chaves;

MODELAGEM: STAR SCHEMA



MODELAGEM: STAR SCHEMA



MODELAGEM: STAR SCHEMA

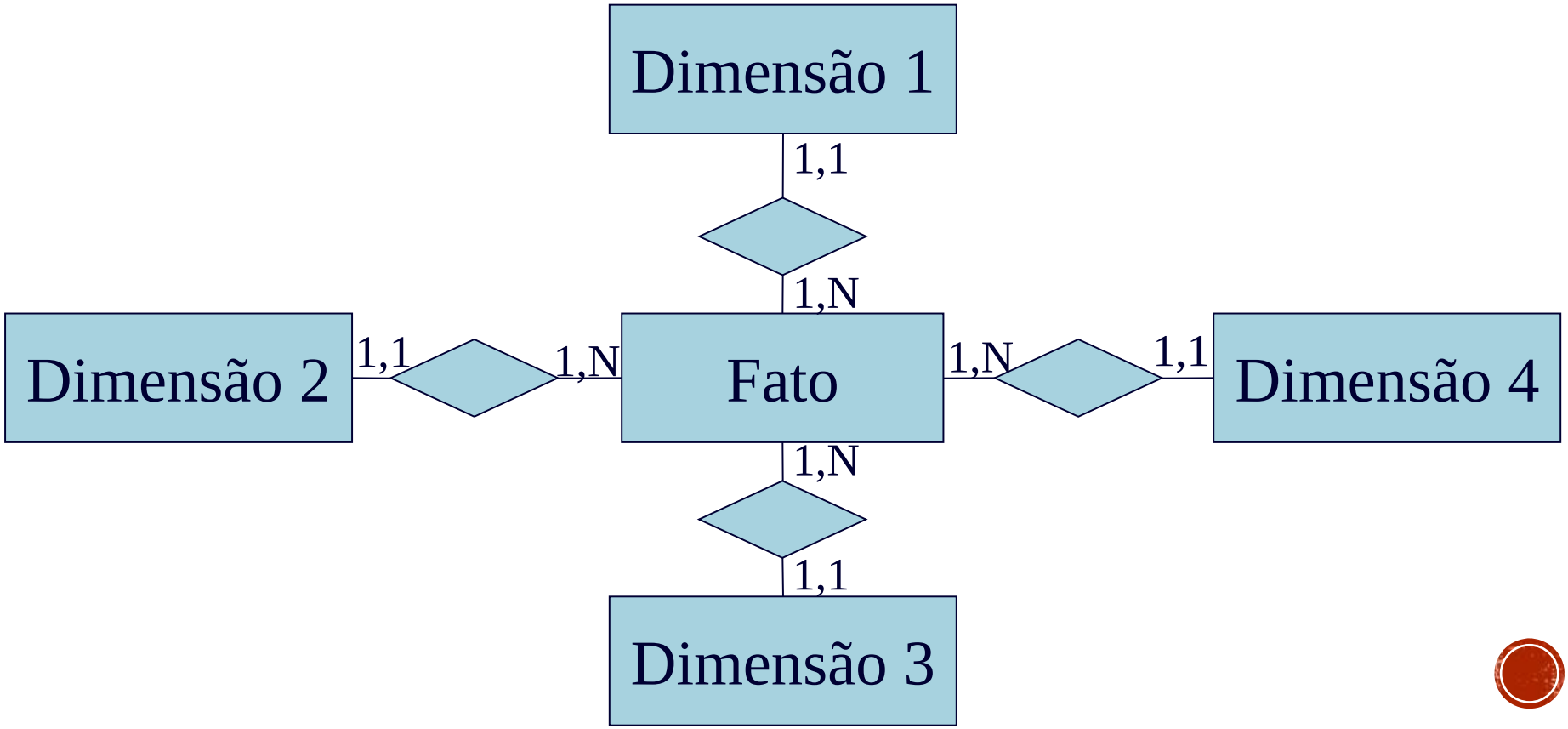
- Tabela Fato:
  - Cada linha representa como foi a venda (ou um conjunto de vendas);
  - Em uma determinada data;
  - De um determinado item;
  - E em uma determinada loja;
  - Armazena qual o valor total das vendas;
  - Quantos destes itens foram vendidos;
  - Qual o preço médio das vendas, o custo médio do item vendido;
  - A margem obtida com as vendas.

MODELAGEM: STAR SCHEMA

- Tabela Fato:
  - Todas estas vendas são somadas e apresentadas em apenas uma linha da tabela;
  - As outras métricas indicam o que estava previsto para ser vendido deste item, nesta data, nesta loja;
  - A métrica **%variação** indica qual o percentual de variação entre o previsto e o realizado.

MODELAGEM: STAR SCHEMA

- A estrutura básica deste modelo pode ser representada por um diagrama entidade relacionamento



MODELAGEM: SNOWFLAKE SCHEMA

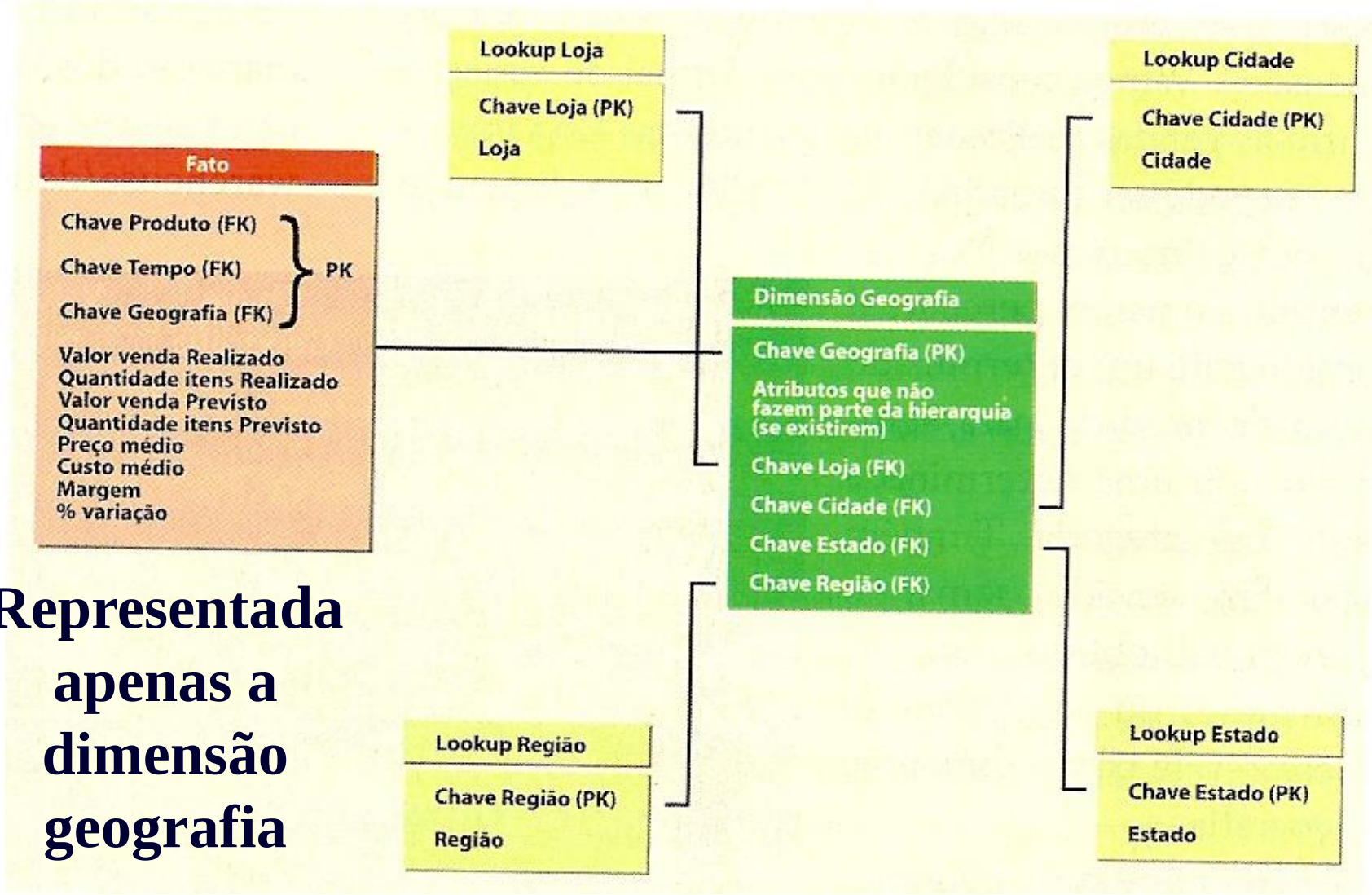
- Emprega uma combinação de normalização da base de dados;
- Para manter a integridade e reduzir os dados armazenados de forma redundante;
- As dimensões são normalizadas em subdimensões,
- Sendo que cada nível da hierarquia fica em uma subdimensão;

MODELAGEM: SNOWFLAKE SCHEMA

- A tabela principal da dimensão tem uma chave para cada nível hierárquico representado na subdimensão e não mais uma única chave, como Star;

- Possui duas variações (diferem na disposição das tabelas que representam dimensões):
  - Snowflake Lookup
  - Showflake Chain

MODELAGEM: SNOWFLAKE LOOKUP



Representada apenas a dimensão geografia

MODELAGEM: SNOWFLAKE LOOKUP

- Emprega tabelas adicionais para nomes e descrições de atributos;
- Todas ligadas a uma tabela principal da dimensão;
- É possível reduzir o tamanho da tabela dimensão eliminado a redundância;
- As tabelas adicionais atuam como tabelas lookup para a chave ou valores codificados da tabela principal da dimensão;
- Que está ligada a uma única tabela fato;

MODELAGEM: SNOWFLAKE LOOKUP

- Vantagem
  - simplificação do armazenamento, reduzindo o tamanho relativo das tabelas de dimensão;
  - melhora do controle de integridade dos dados.

- Desvantagem
  - acontece um número maior de joins, comparando com o esquema Star;
    - porque precisa buscar as descrições nas tabelas adicionais.

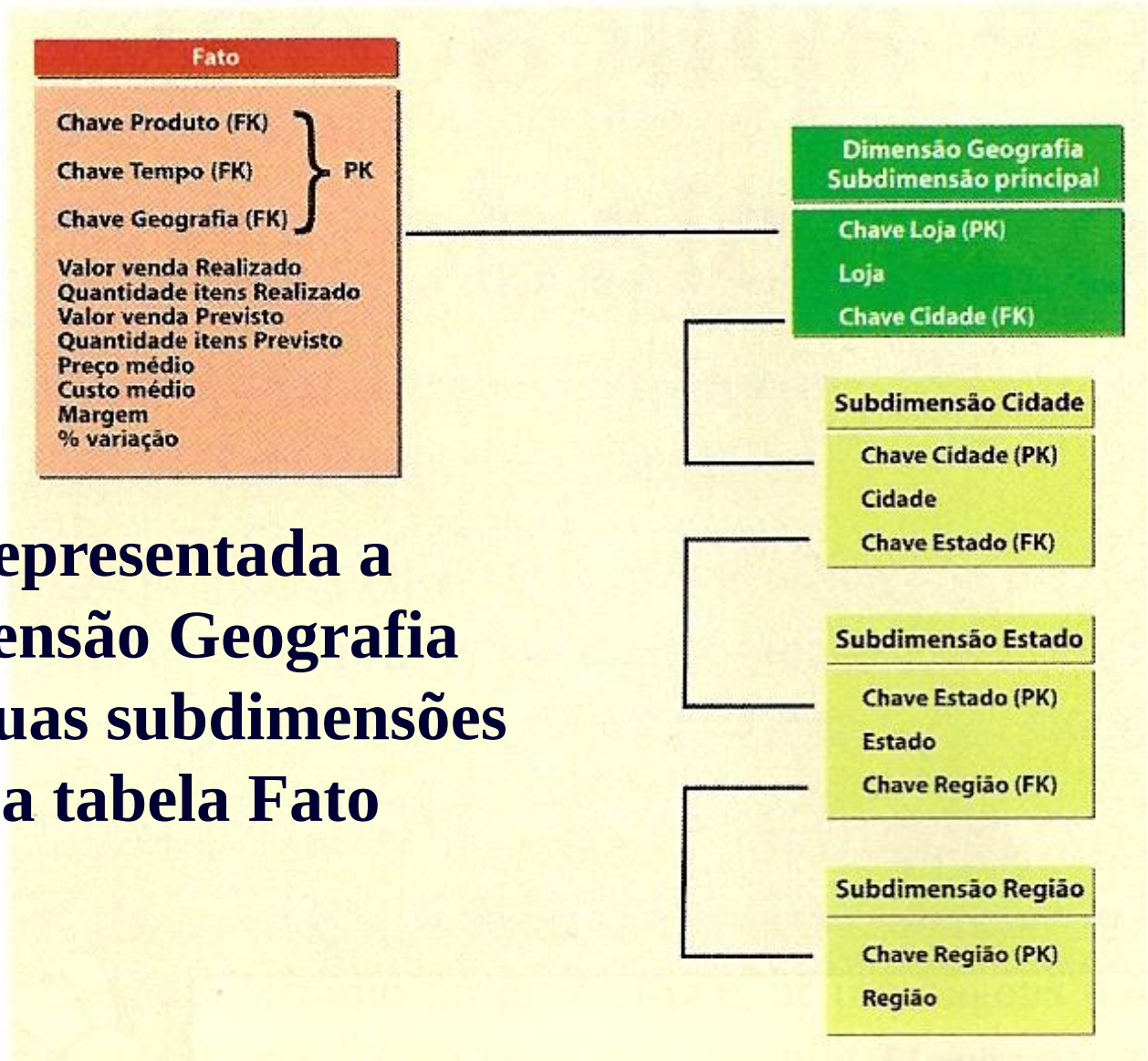
MODELAGEM: SNOWFLAKE LOOKUP

- Desvantagem
  - manutenção da base de dados requer um custo alto, pois o número de tabelas físicas distintas torna-se maior.

MODELAGEM: SNOWFLAKE CHAIN

- Também possui subdimensões particionadas pelos níveis hierárquicos da dimensão;
- A tabela principal da dimensão representa o nível mais baixo (mais detalhado) da hierarquia;
- As subdimensões estão encadeadas;
- A tabela Fato fica ligada à subdimensão de mais baixa granularidade (chamada de principal ou raiz)
  - exemplo: subdimensão - Loja

MODELAGEM: SNOWFLAKE CHAIN



Representada a dimensão Geografia com suas subdimensões e a tabela Fato

MODELAGEM: SNOWFLAKE CHAIN

- Esta hierarquia é sempre 1:N;
- Cada tabela da subdimensão contém sua chave primária e suas descrições associadas;
- Contém também a chave para o próximo nível da hierarquia da dimensão;
- Até chegarmos ao nível mais alto (menos detalhado da hierarquia);

MODELAGEM: SNOWFLAKE CHAIN

- Não é recomendada quando os relatórios necessitam, frequentemente, de vários níveis de agregação da informação;
  - já que são necessários vários passos na cadeia para se chegar ao resultado.

QUAL O MELHOR ESQUEMA?

- Depende:
  - do projetista;
  - da ferramenta OLAP;
    - algumas funcionam melhor com o Star Schema outras com Snowflake;
  - existem aquelas que independem, podem ser utilizadas com qualquer opção de modelagem.

- Star Schema tem sido mais utilizado.

Dúvidas, Perguntas ou Sugestões?