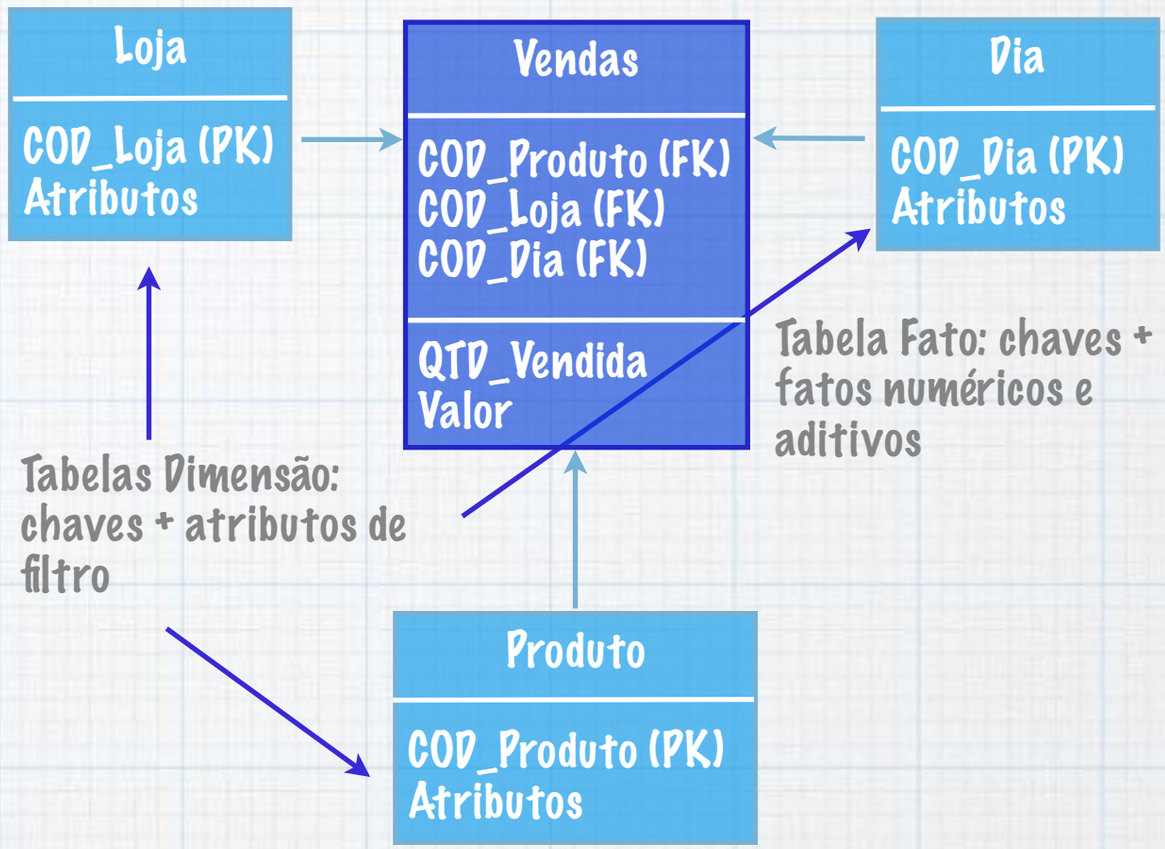
Business Intelligence
Modelagem Dimensional

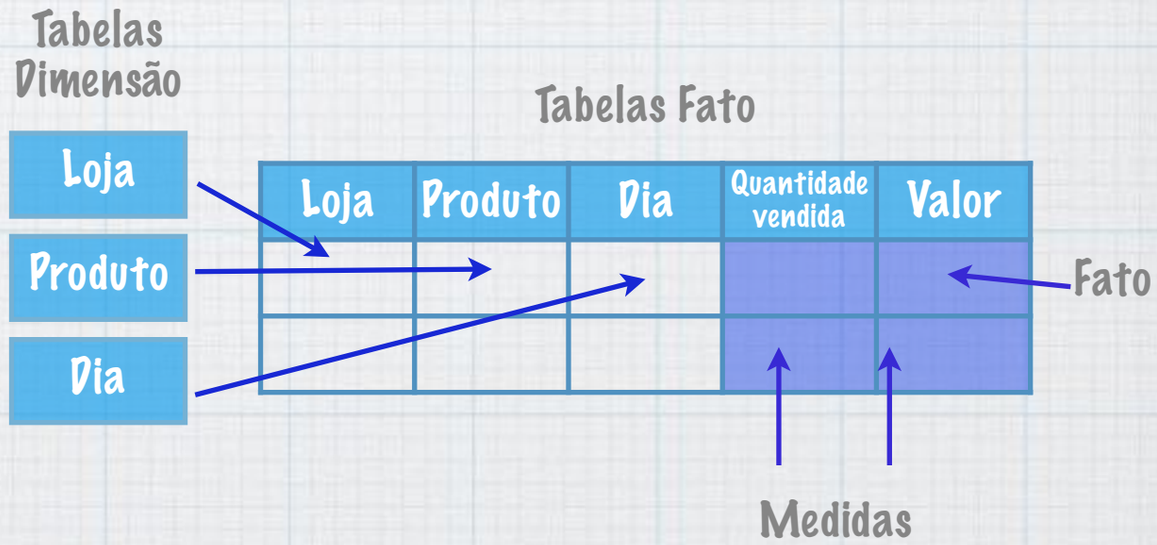
Datawarehousing e Data Mining
Prof. Sergio Bonato, 2019
asbonato@gmail.com

* **Tabelas Fato:** servem para armazenar medidas numéricas associadas aos eventos de negócio, os fatos; possuem como chave primária um campo multikey composto pelas chaves primárias das dimensões que com ela se relacionam; contém dados normalmente aditivos.

* **Tabelas Dimensão:** representam entidades de negócio e são as estruturas de entrada; tem uma relação 1:N com as tabela fato; possuem múltiplas colunas, algumas representando uma hierarquia; sempre tem chave primária; são as tabelas que fazem os filtros de pesquisa dos fatos.



Exemplo de um modelo dimensional



Composição básica de uma tabela fato

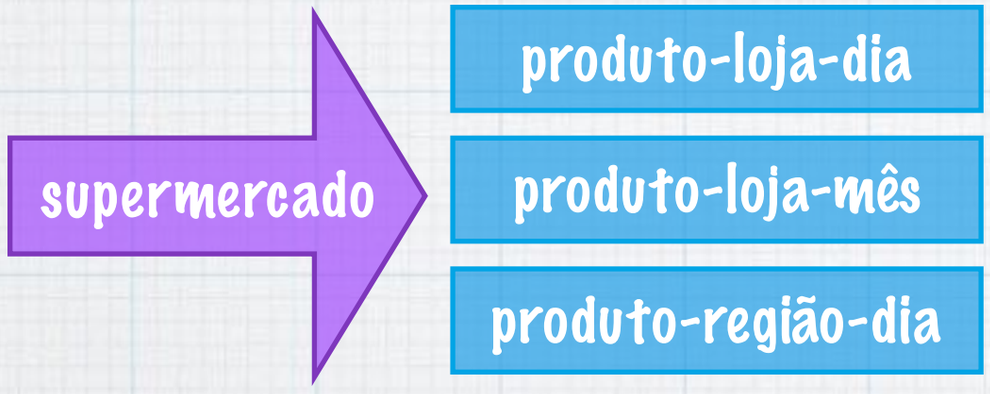
A definição da área de negócio onde vai acontecer o projeto de DW/DM deve ocorrer de acordo com as prioridades da empresa.

- Escolhe-se uma área: Marketing (clientes), Finanças, Vendas, Produção, etc.

- Definem-se os processos alvo do projeto de DW: varejo, entrega, controle de pedidos, assinaturas, etc.

  - Analisam-se os processos escolhidos, identificando entidades de dados, relacionamentos, objetos, eventos, interações.

A granularidade define, de forma combinatória, os níveis dimensionais de armazenamento de dados



deve-se partir de modelos que tenham a maior granularidade possível; assim se pode obter os outros níveis desejados.

- cada loja vende, por dia, 10 mil produtos
- temos 200 lojas
- queremos armazenar as vendas diárias por 2 anos

10000*200*365*2 = 1,46 bilhão de registros

- cada chave (produto-loja-dia) tem 5 bytes
- 3 valores numéricos com 4 bytes cada

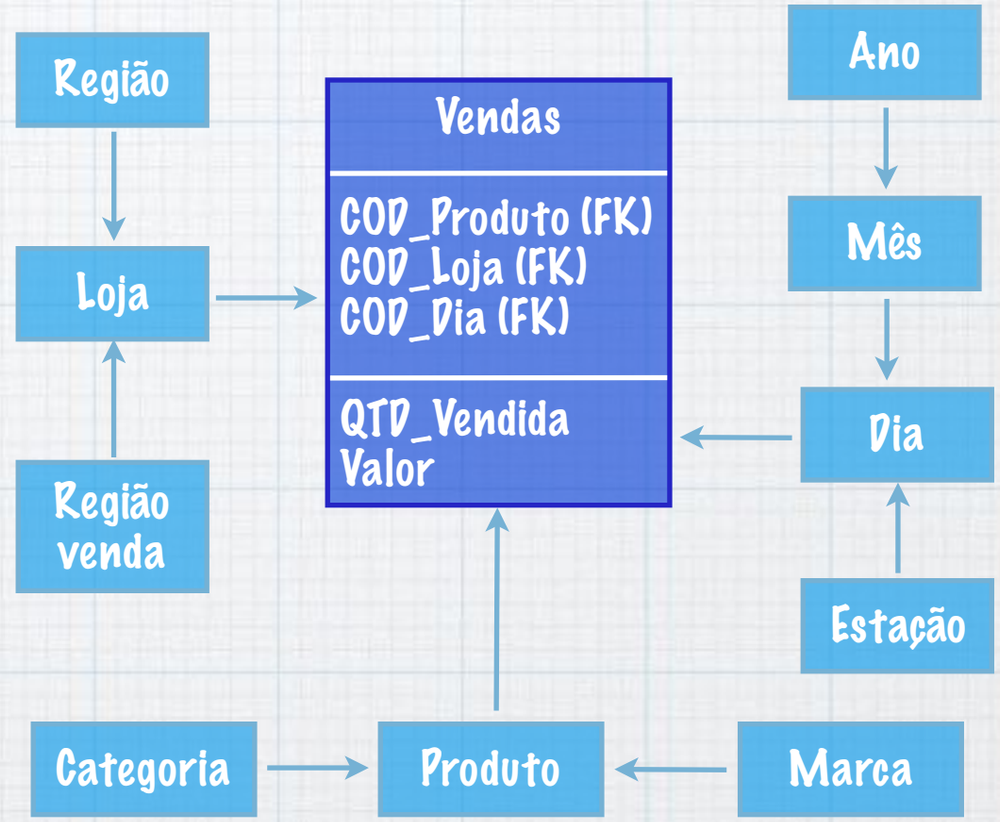
1,46 bilhão * 27 bytes = 39,4 gigabytes

cargas diárias das vendas das 200 lojas no DW

Na definição das tabelas dimensão o importante é a hierarquia das dimensões e a definição dos atributos

* As hierarquias de dimensões compõem, na forma de atributos, os registros das tabelas Dimensão.

* As várias hierarquias possibilitam diferentes caminhos à tabela Fato.



Modelo dimensional com hierarquia de dimensões

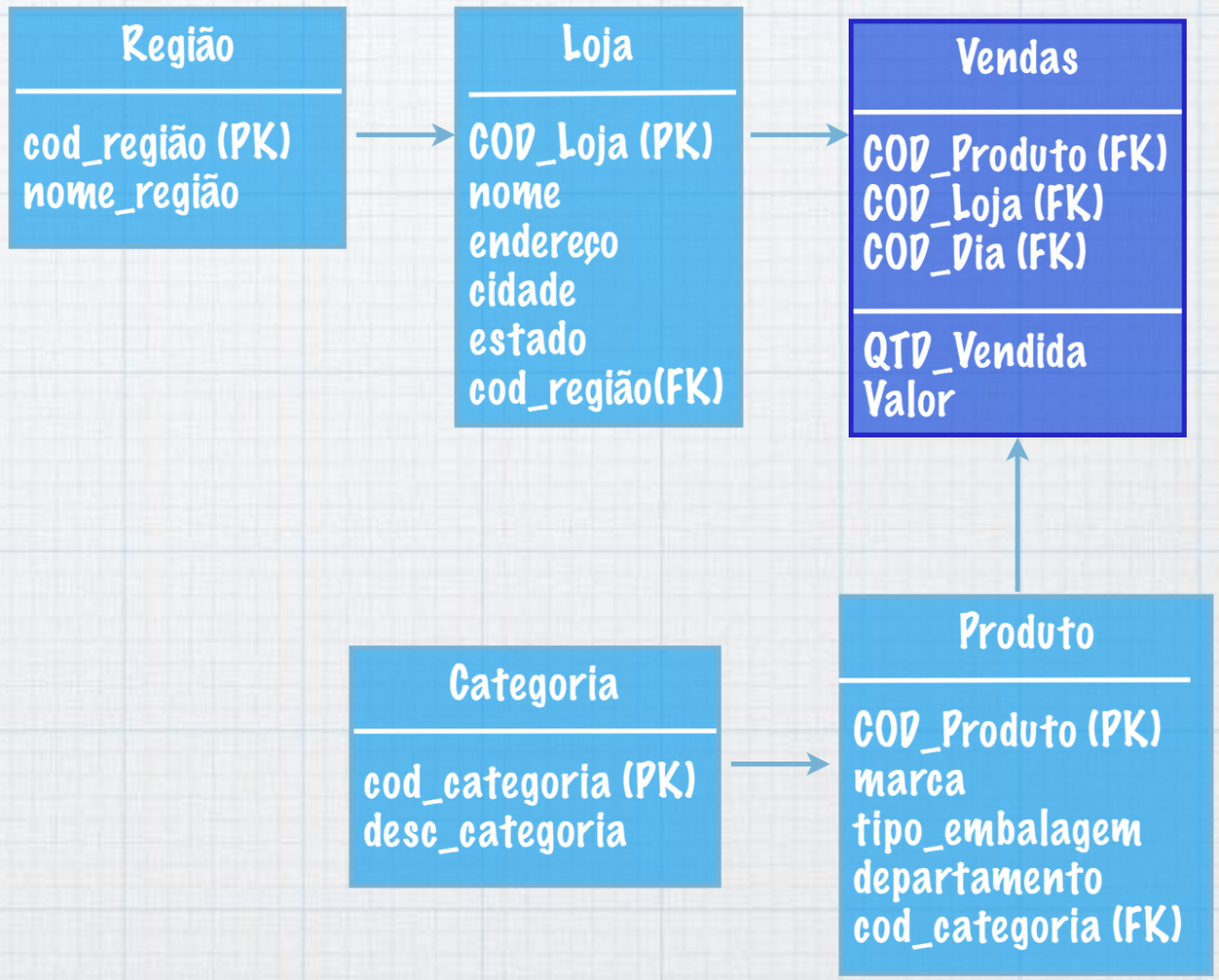
Definição das tabelas Dimensão

TdLoja (cod_loja, nome, endereço, cidade, estado, cod_região, região_venda)
TdProduto (cod_produto, marca, categoria, tipo_embalagem, departamento)
TdDia (cod_dia, mês, ano, período_fiscal, estação_ano)

Normalização das tabelas dimensão pode ser feita - snowflake schema - ou não - star schema

star schema

snowflake schema



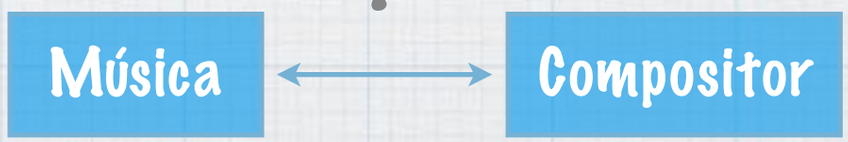
As redundâncias do star schema são compensadas pela redução no número de junções; como as tabelas dimensão são menores, o desperdício de espaço não é grande.

Os relacionamentos de atributos das tabelas dimensão podem ser classificados da seguinte forma:

* As tabelas dimensão não possuem relacionamento entre si, formando dimensões independentes. Ex: TEMPO e LOJA

* Os atributos dentro da mesma tabela possuem relacionamento hierárquico, ou seja, 1:N, formando estruturas hierárquicas de dimensões. Ex: Região->Cidade->Loja.

* Os atributos em uma dimensão possuem relacionamento N:N

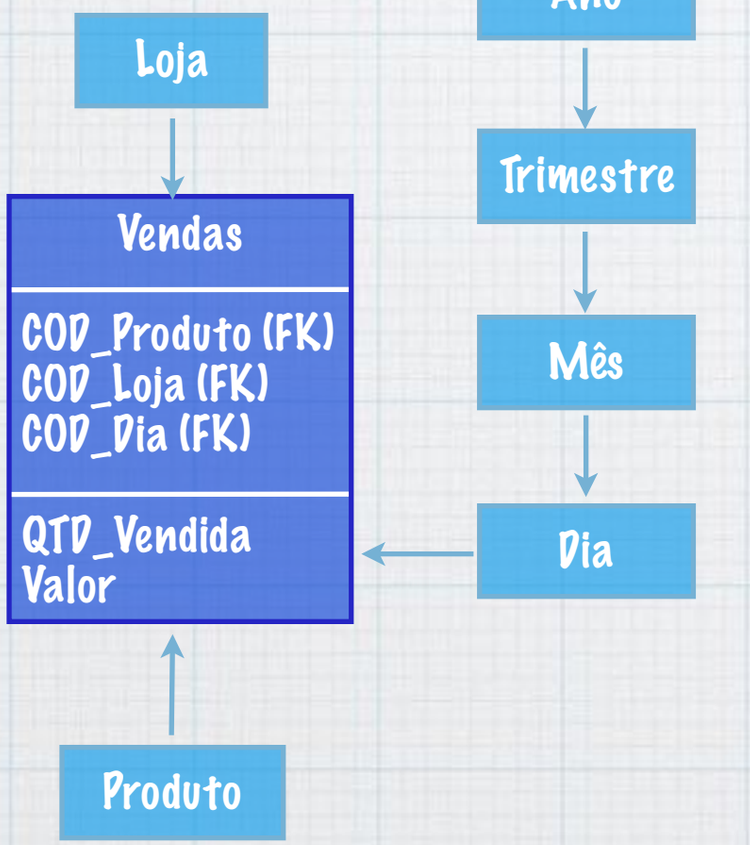


Solução:

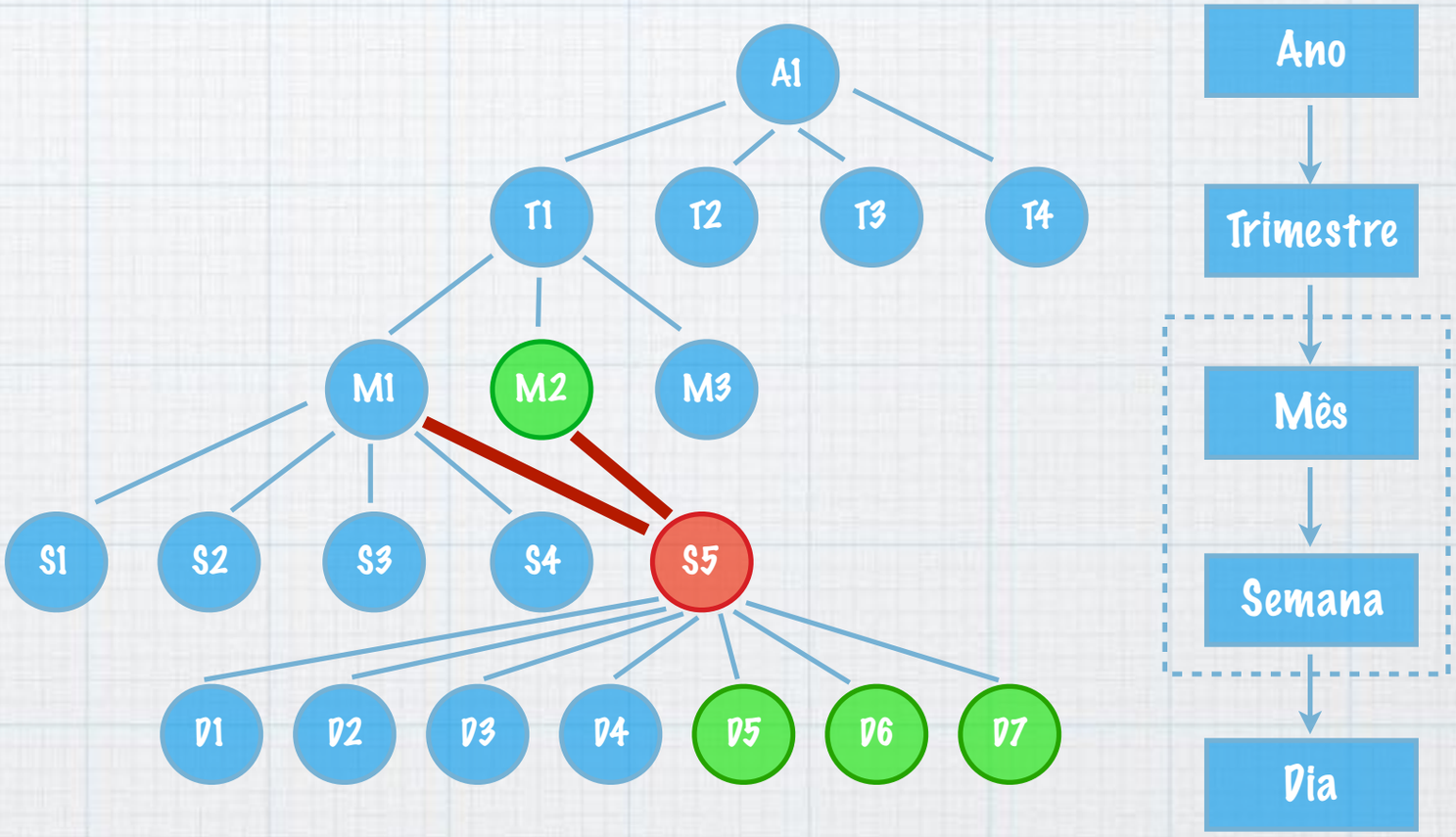
Definição dos atributos das tabelas fato, que geralmente são valores chamados de MÉTRICAS.

* Tipos de Métricas
  * Aditivas: são valores passíveis de serem somados, como valor vendido ou custo.
  * Semi-aditivas: sua soma só tem sentido em uma dimensão, como quantidade vendida, que deve ser acumulada na dimensão produto.
  * Não-aditivas: não dá para somar, como margem de contribuição de cada produto
* As tabelas fato são sempre normalizadas
* Para economizar espaço, alguns valores podem ser obtidos via transação ou por meio de drill-through

Conformidade de dimensões estabelece coerência entre as dimensões em diferentes DM ou dentro do mesmo DM



conforme

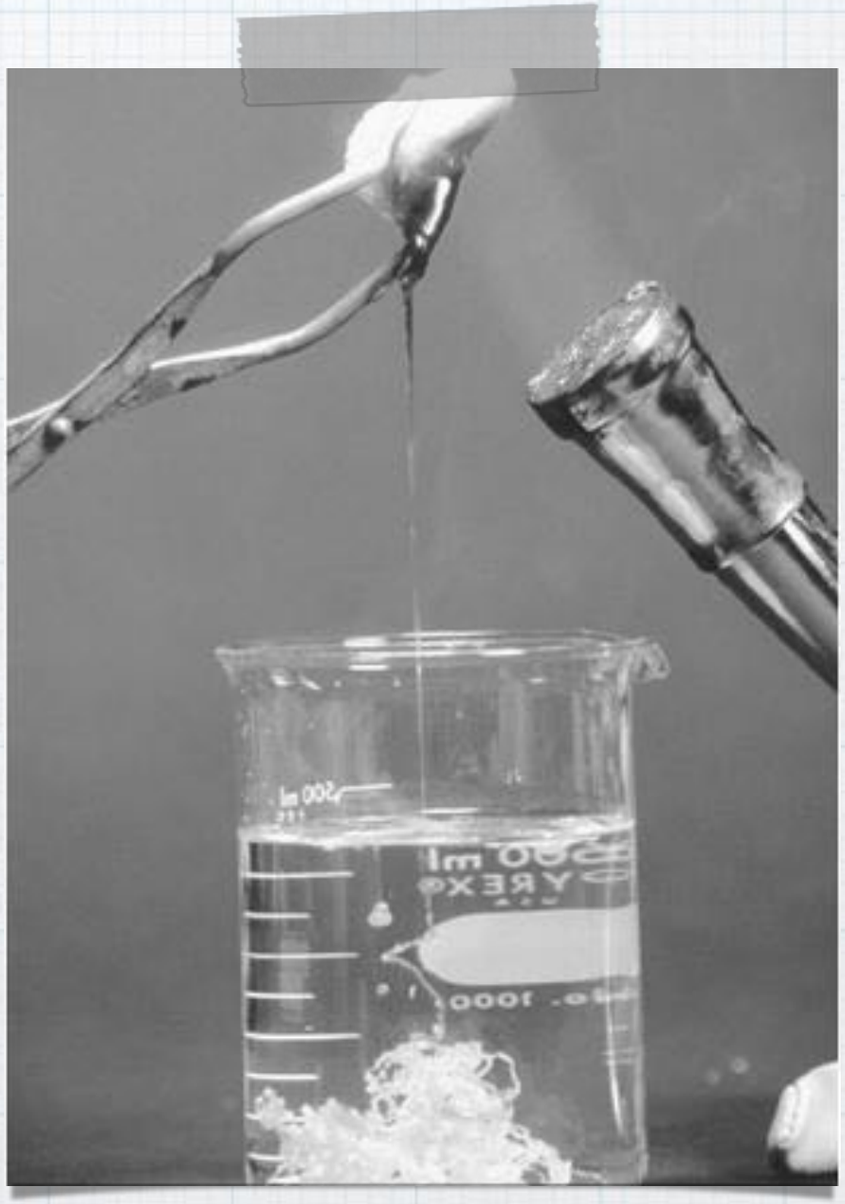


não conforme

Combinação de dimensões podem ocorrer se há grande coesão entre elas.

* Por exemplo, se certos PRODUTOS são vendidos somente em determinadas LOJAS, isso pode sugerir uma combinação de dimensões.

* Entretanto, é preciso ter cuidado com a explosão combinatória, que poderia elevar muito o número de instâncias em uma dimensão.



As dimensões especiais estão em quase todos os DW/DM. Estão relacionadas com TEMPO, ESPAÇO e o OBJETO dos sistemas.

* Tempo com granularidade DIA
  * DATA_COMPLETA: 01-01-2000
  * DIA_SEMANA: 6ª Feira
  * NÚMERO_DIA_MÊS: 01
  * NÚMERO_DIA_GERAL_CORRIDO_NO_ANO (1 - 365)
  * NÚMERO_SEMANA_MÊS (1 A 4 ou 5)
  * NÚMERO_SEMANA_GERAL_CORRIDO (1 a 52)
  * MÊS_ANO (Janeiro a Dezembro)
  * NÚMERO_MÊS_GERAL_CORRIDO (1 a 12)
  * TRIMESTRE (1 - 4)
  * PERÍODO_FISCAL (1 a 4)
  * TAG_DIA_FINAL_SEMANA (indica sáb ou dom)
  * TAG_ÚLTIMO_DIA_MÊS (indica se é o último dia do mês)
  * TAG_FERIADO (indica se é feriado)

A Dinâmica das dimensões está ligada com a estratégi de manutenção das informações quando há atualizações.

Sem histórico

Novo

Velho

N versões

Histórico completo

Atual

Anterior

2 últimas versões

Cliente (estática)

C1

TFato

C1 V1

V3

V2

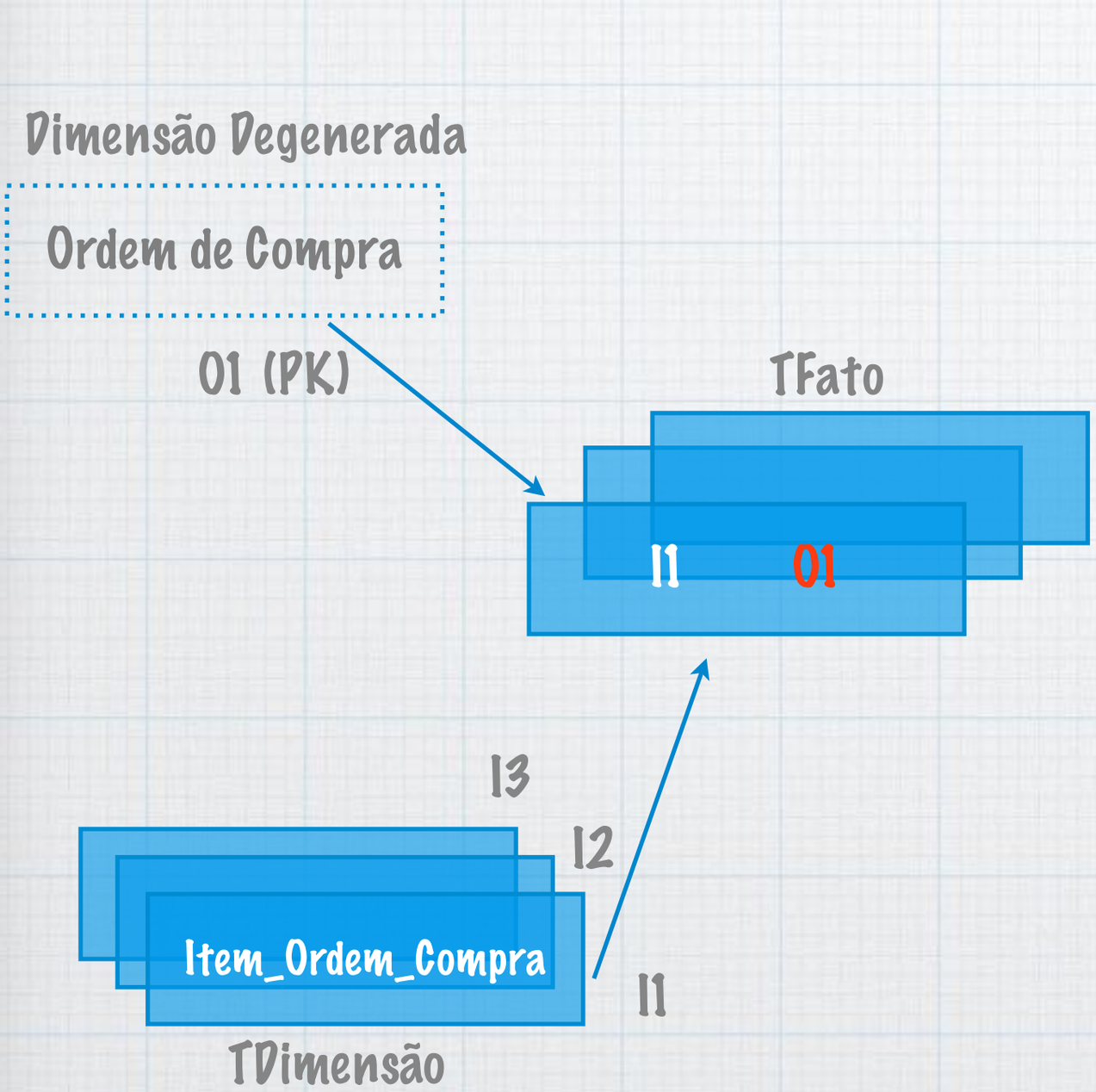
V1

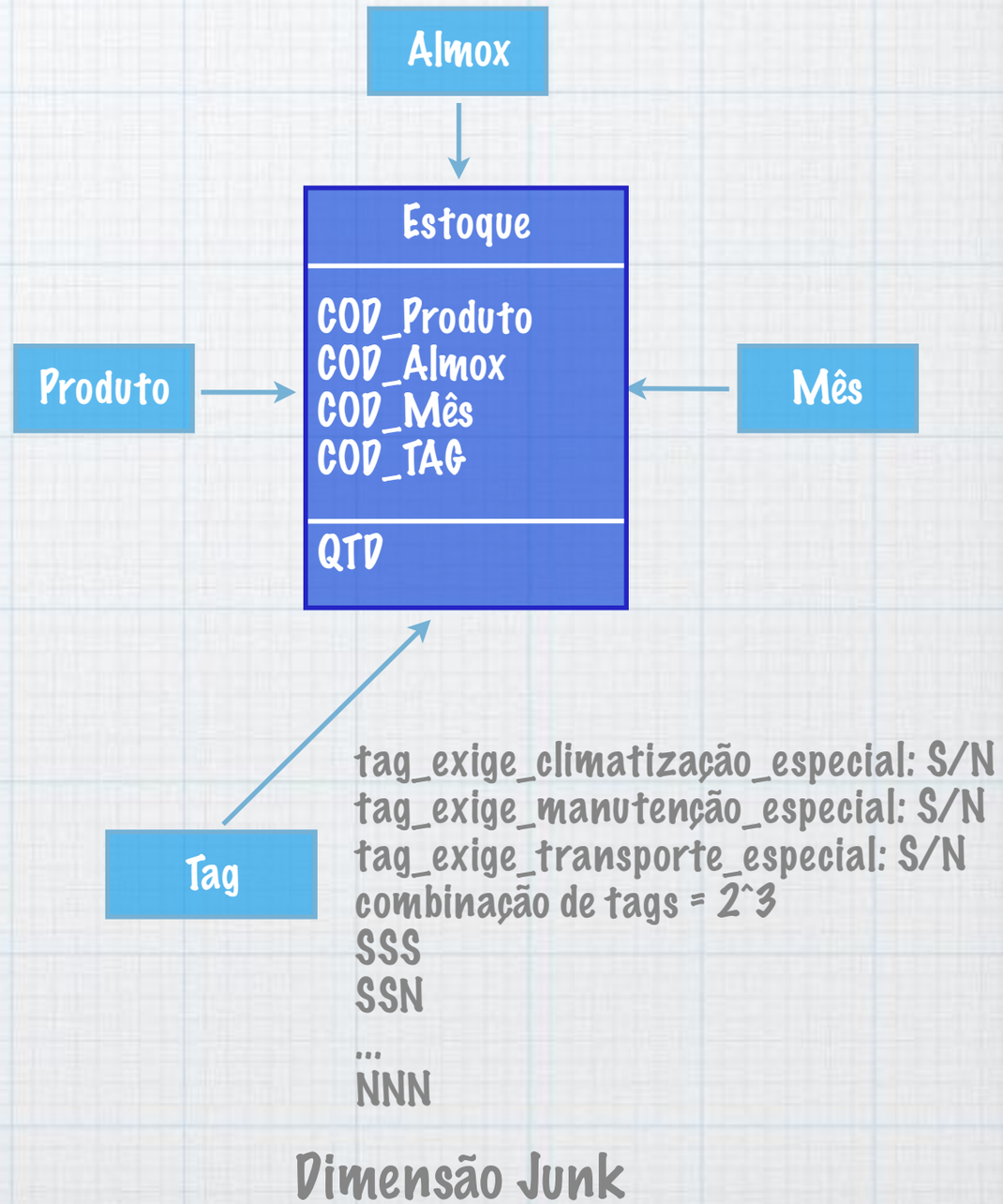
Cliente (volátil)

os dados tem data de validade - início e fim)

mantém dados estáticos e voláteis em tabelas separadas

As dimensões degeneradas existem para alinhavar outras dimensões na tabela fato, mas não tem tabela dimensão; dimensões junk são para tags, valores binários ou de pequena cardinalidade.

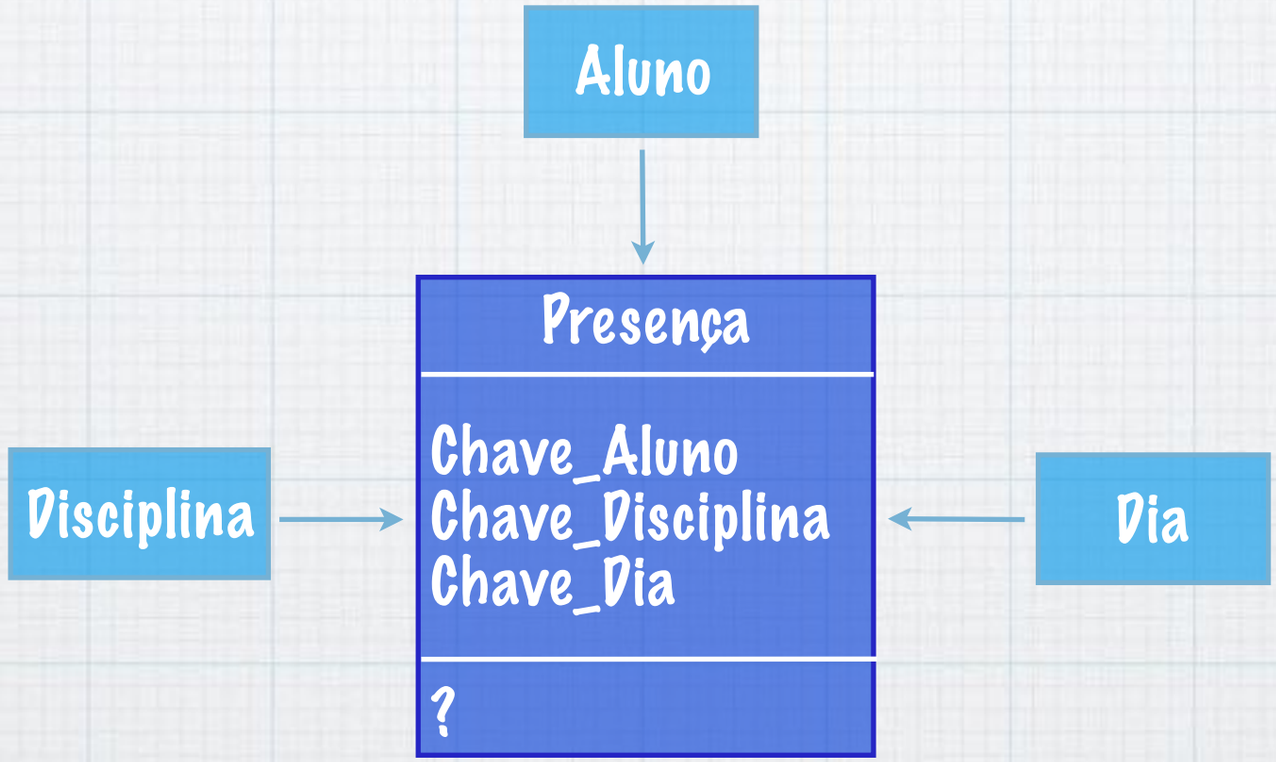




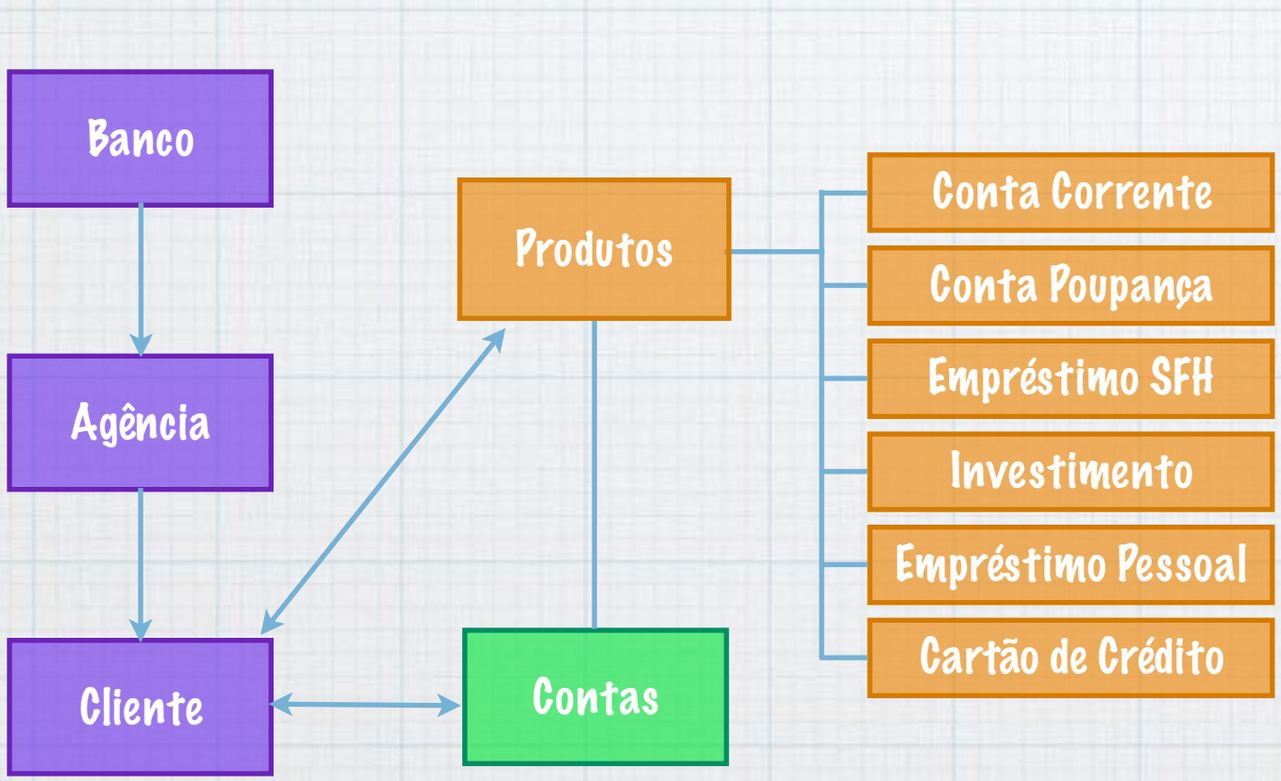
Na escolha de campos chave de dimensões e fatos deve-se preferir chaves artificiais (surrogates) às naturais, i.e., com semântica de negócio embutida, pois são mais estáveis e evitam problemas relativos a:

* Unicidade: nem sempre se pode garantir valores únicos; a esposa pode usar o CPF do marido, por exemplo.
* Ausência: algumas entidades podem não ter identificadores naturais; um cliente estrangeiro não tem CPF.
* Tamanho: as chaves naturais geralmente são maiores que as artificiais; 4 bytes tem uma faixa de 2 bilhões (2^{32}) de valores diferentes.
* Estas chaves ligam Fato e Dimensão e geralmente ficam escondidas do usuário e são produzidas durante as cargas.

Tabelas fatos sem dados ou métricas são raras e servem para relacionar tabelas dimensão envolvidas



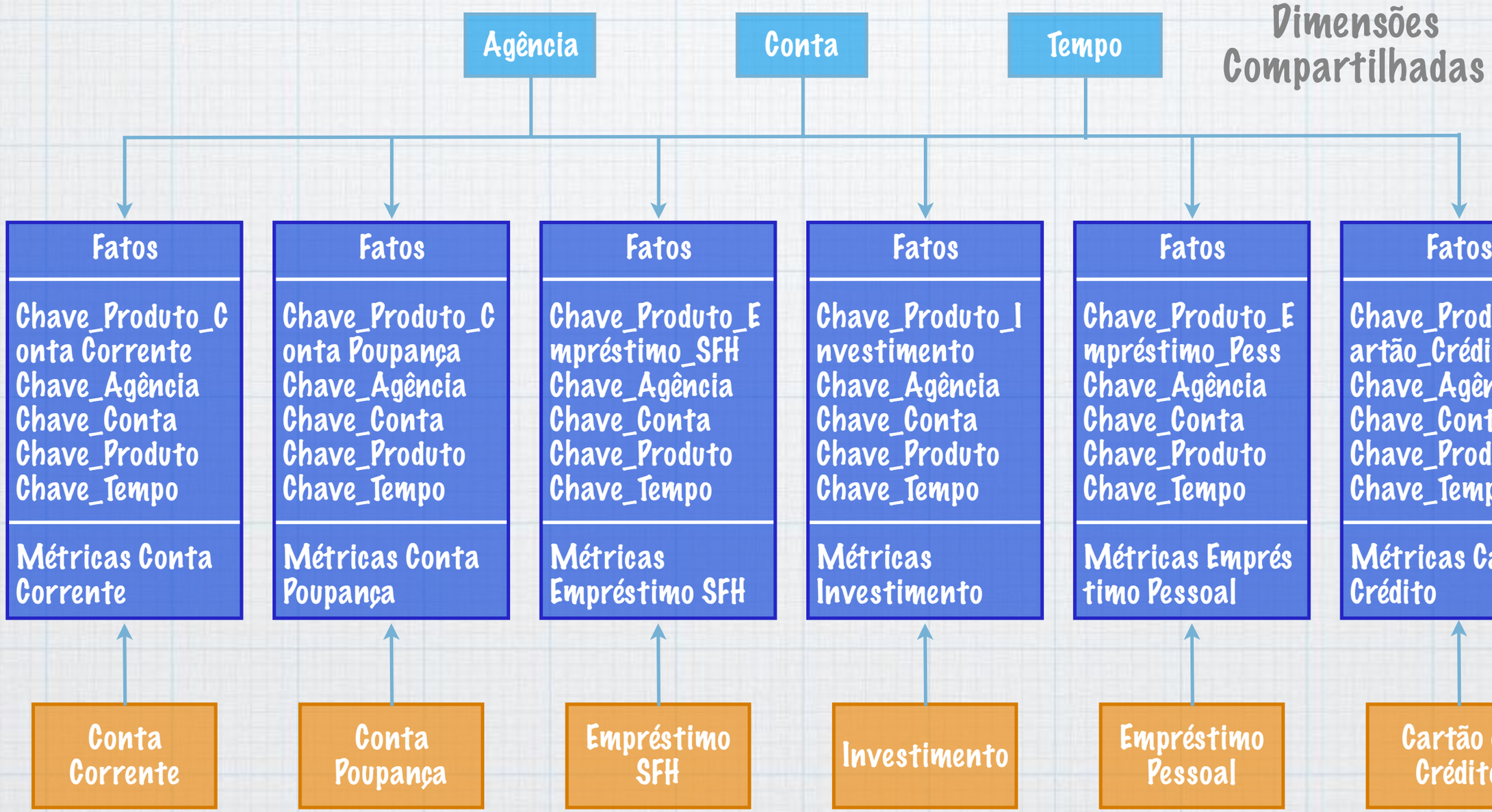
Tabelas fatos com classificação ou subtipos ocorrem quando o negócio tem vários tipos de produtos.



Modelo relacional com tipo e subtipos

Modelo dimensional para tratamento de multifatos - tabela base

O modelo base contém os produtos na sua menor granularidade, consolidados por categoria; o modelo detalhado contém as fato especializadas para cada tipo de produto.

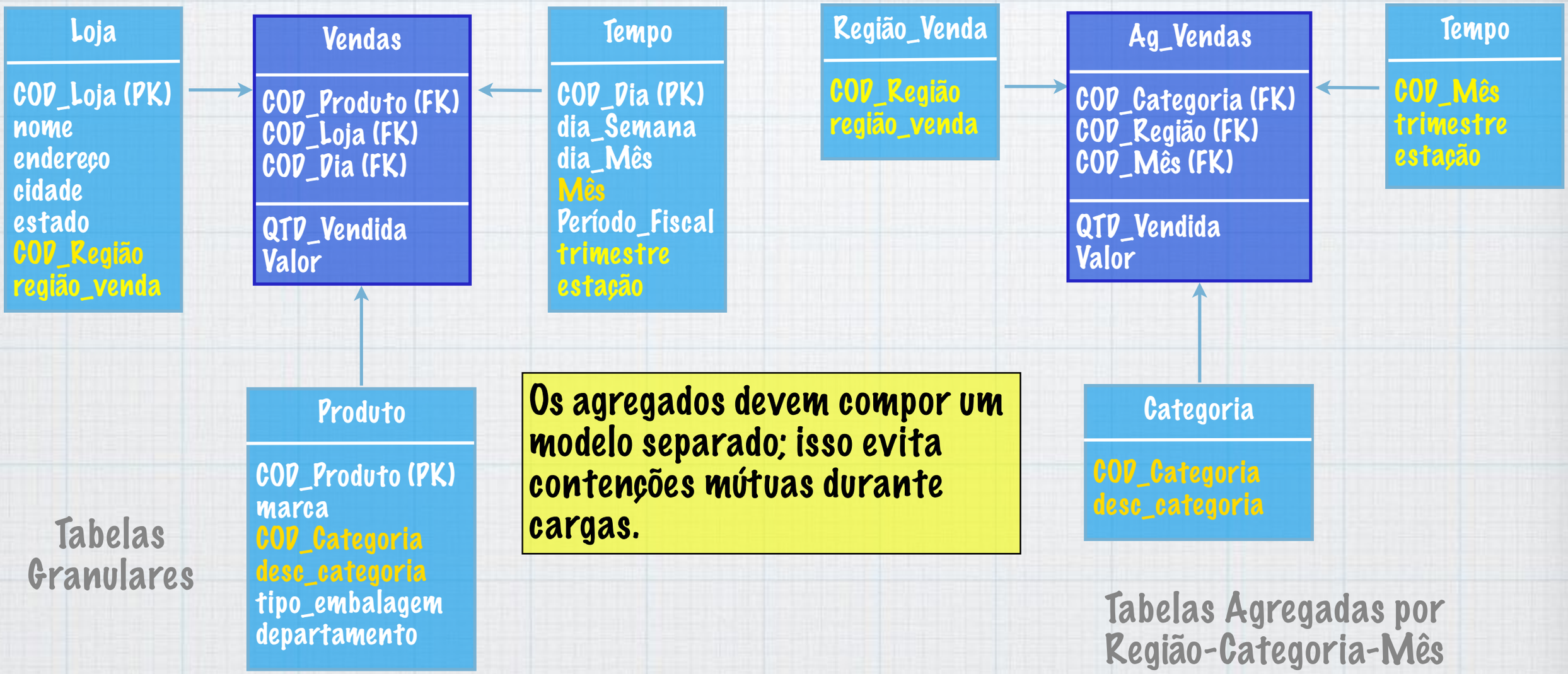


Agregados são tabelas prontas, sumariadas em várias dimensões, que facilitam a agilizam o acesso aos dados

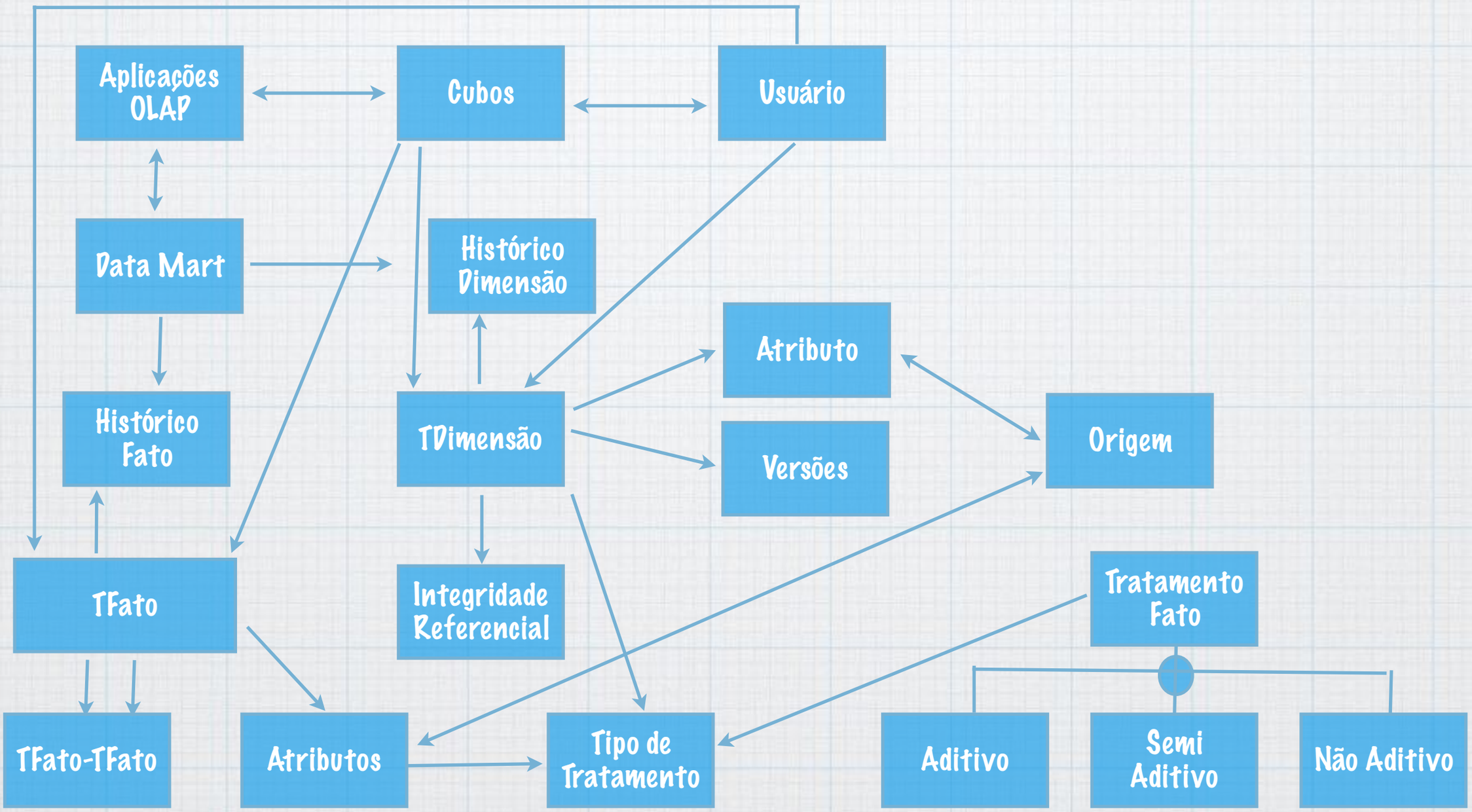
TdLoja (cod_loja, nome, endreço, cidade, estado, cod_região, região_venda)
TdProduto (cod_produto, marca, categoria, tipo_embalagem, departamento)
TdDia (cod_dia, mês, ano, período_fiscal, estação_ano)
TfVendas (cod_loja, cod_produto, cod_dia, valor_vendido_real, custo_real, lucro, qtd_vendida)

- O critério para definição de agregados está na dificuldade para obtenção de informações a partir dos dados granulares.
- Os campos em vermelho sublinhados são níveis de hierarquia:
  - REGIÃO -> LOJA: 2 níveis
  - CATEGORIA -> PRODUTO: 2 níveis
  - ANO -> MÊS -> DIA: 3 níveis
- O número de tabelas de agregados está diretamente relacionado com o número de combinações ternárias, binárias ou unárias e o volume está ligado ao número de ocorrências:
  - ternária: 2 X 2 X 3 = 12 opções. Ex: região + categoria + ano, região + categoria + mês, etc
  - binária: 16 opções (2 X 2 + 2 X 3 + 2 X 3). Ex: região + categoria, ano + loja, região + ano, etc
  - unária: 7 opções (2 + 2 + 3). Ex: agregados por loja, ou por categoria, ou por mês.
  - total: 35 combinações possíveis

Deve-se ter cuidados na definição de agregados com valores aditivos e na precisão dos valores, pois nem todos os valores são passíveis de soma e pode haver overflow nos totais se a variável não for maior.



Metadados são importantes na documentação das aplicações OLAP e do ambiente DW/DM. Exemplo de metadados armazenados em um banco relacional.



Bibliografia

* disponíveis na biblioteca
  * Fonte principal: BARBIERI, Carlos. BI - Business Intelligence: Modelagem & Tecnologia; Axcel Books, 2001
  * INMON, William H. Como construir o data warehouse. Rio de Janeiro: Campus, 1997
  * KIMBALL, Ralph; MERZ, Richard. Data Webhouse : construindo o Data Warehouse para a Web. Rio De Janeiro: Campus, 2000
* online
  * BRUZAROSCO; CASTOLDI; PACHECO; Criando data warehouse com o modelo dimensional. http://periodicos.uem.br/ojs/index.php/ActaSciTechnol/article/viewFile/3099/2225