OLAP (On-Line Analytical Processing) e Banco de Dados Multidimensionais

O que é OLAP?

Processamento de dados

Dedicado ao suporte a decisão

Visualização de dados agregrados ao longo de várias dimensões analíticas (tempo, espaço, categoria de produto, quantidade vendida, preço...)

Hierarquizadas em várias granularidades

Armazenados em BD especializadas

Modelo lógico de dados multidimensional

Data Warehouse, Data Mart ou BD multidimensionais

Exemplos de consultas OLAP

Quais foram os produtos mais vendidos no mês passado ?

A média salarial dos funcionários de informática com menos de 5 anos de experiência é maior do que a mesma para funcionários de telecomunicação?

Qual foi o total de vendas o mês passado por região de vinhos tintos importados da Europa?

Por quais semanas, quais produtos e quais cidades, a variação de venda de produtos em promoção em comparação da semana anterior sem promoção foi $\geq 15\%$?

Banco de dado operacional x data warehouse x data mart

BD operacional:

- armazena valores **correntes** e **atômicas** resultantes diretas das últimas transações
- fins **operacionais** predefinidas *ex, gerenciamento do estoque*

Data Mart:

- armazena réplicas **históricas**, não voláteis, **agregadas** ao longo de várias dimensões analíticas
- as vezes limpas, completadas e normalizadas em termos de escala e distribuição
- de dados de um único banco operacional
- fins **analíticas** abertas de escopo **departamental**

Data Warehouse:

- **integra** e padroniza dados
- de vários:
  - data marts
  - BD operacionais
  - BD de legado empacotados
  - BD semi-estruturados extraídos de páginas web
- em um único repositório coerente e limpo de dados
- fins **analíticas** abertas de escopo **organizacional**

OLTP x OLAP

Função	Automatizar operações diárias	Auxiliar tomada de decisão
Usuário humano	Cliente, Atendente, DBA	Executivo, Analista, Eng. de Conhecimento
Software cliente	Aplicativos de inventário, contabilidade, ...	Aplicativos de mineração de dados, análise matemática, ...
Modelo lógico	Relacional, orientado por aplicações	Multidimensional, orientado por assuntos
Granularidade	Única e atômica	*M*últipla e agregada
Temporalidade dos dados	Apenas valor corrente atualizada continuamente	Histórico dos valores, completado periodicamente
Consultas	Simples e predefinidas	Complexas e *ad-hoc*
Direção	Tanto ler quanto escrever	Essencialmente ler
Envolve	Acessos via índice e hash	Junções, varreduras
Registros	10	10^6
Usuários	10^3	[0-10]
Bytes	MB-GB	GB-TB
Prioridade	Disponibilidade, eficiência	Flexibilidade, interatividade
Métrica	Numero de transações	Número e tempo de cada consulta

Modelo de dados multidimensional

Cuboide:

- Espaço de dimensão N para análise de dado

Dimensão analítica:

- Atributo geralmente categórico
- Escolhido como eixo no espaço analítico N-dimensional
- Campo de uma tabela do BD relacional fonte
- *ex, tempo, local, produto, fornecedor*

Medida:

- Atributo geralmente numérico
- Escolhido como ponto no espaço analítico N-dimensional
- Agregação de valores de um campo de uma tabela do BD relacional fonte, calculada por group-by de outros campos da relação
- *ex, valor total das vendas, valor média das vendas, quantidade vendidas,*

Cuboide de dados: exemplo 4D

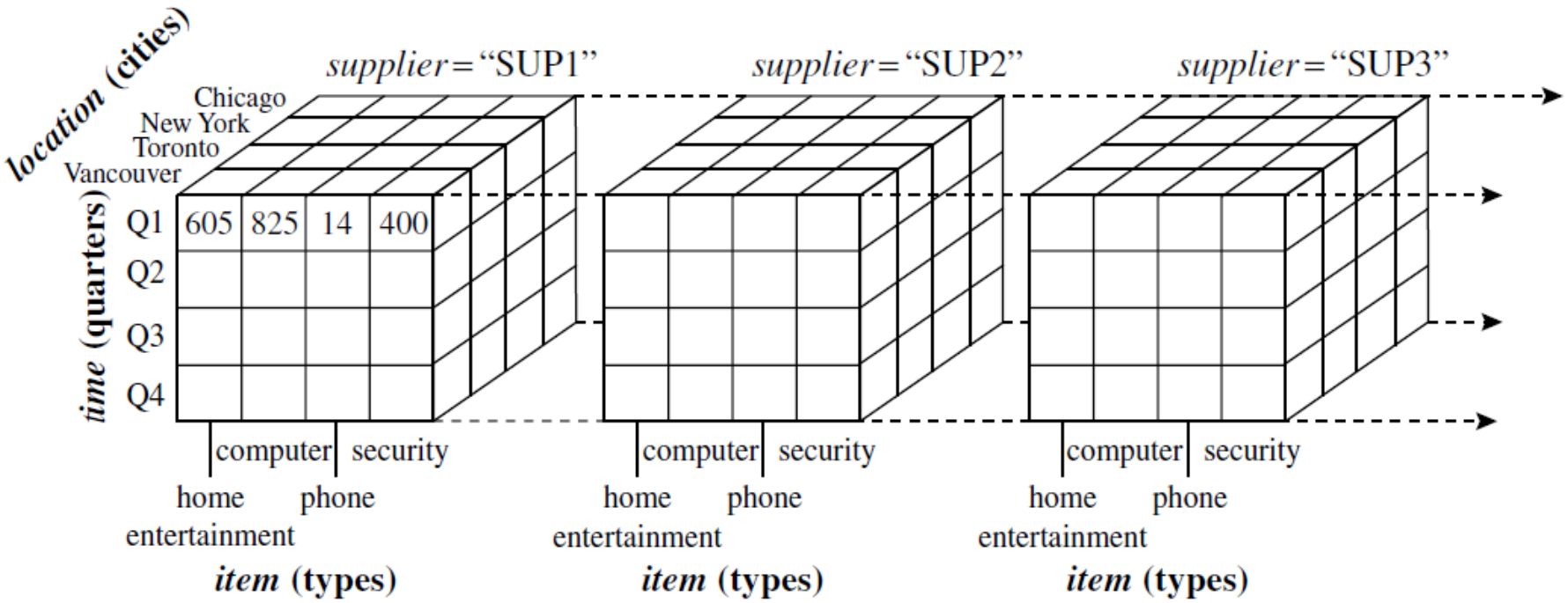
Células

Membros

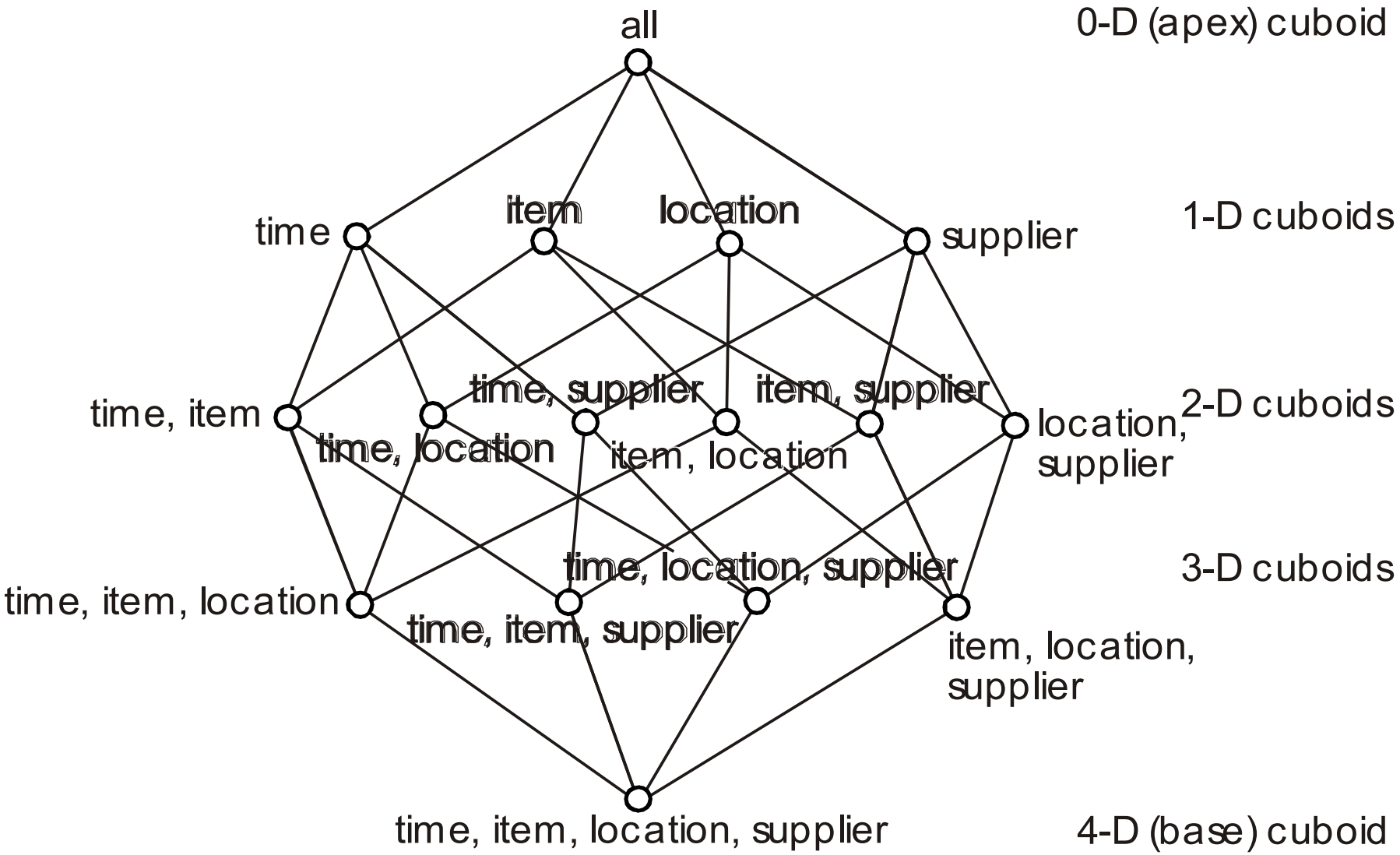
Dimensões

Location		Chicago				New York				Toronto			
Item		HE	Comp	Tel	Secu	HE	Comp	Tel	Secu	HE	Comp	Tel	Secu
Time	Supplier												
Q1	Sup1												
	Sup2												
Q2	Sup1												
	Sup2												
Q3	Sup1												
	Sup2												
Q3	Sup1												
	Sup2												
Q4	Sup1												
	Sup2												

Cuboide de dados: exemplo 4D



Reticulado de Cuboides



Tipologia e cálculo das medidas

Medida *distributiva:*

- agregada por operação distributiva sobre dados atômicos ou medidas distributivas
- *count, sum, max, min*

Medida *algébrica:*

- agregada por operações algébricas sobre dados atômicos ou medidas distributivas ou algébricas
- *avg, standev*

Medida *holística:*

- agregada por operações sem limite constante sobre o espaço necessário para armazenar os sub-agregados
- *median, mode, rank*
- em grandes data warehouses, cálculo apenas aproximativo

Hierarquias conceituais: da multidimensionalidade a multigranularidade

Hierarquia esquemática:

- implícita no esquema relacional do BD operacional fonte

Hierarquia de agrupamento:

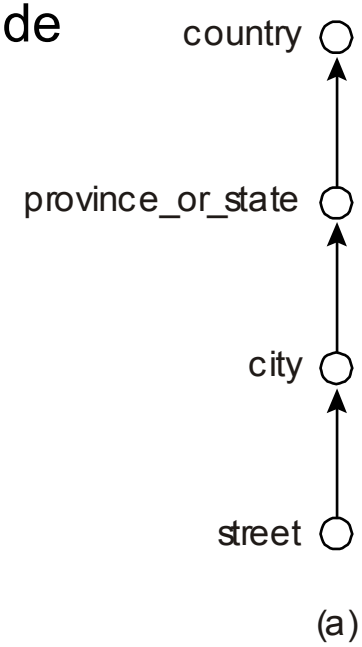
- Inexistente no esquema fonte, gerada para reduzir numerosidade

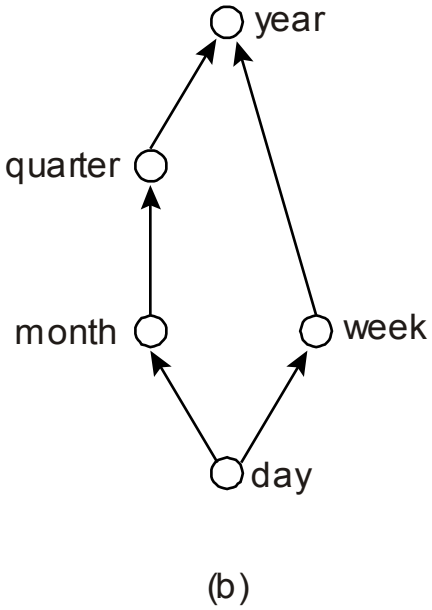
Hierarquia:

- de ordem total ou parcial
- simples ou múltipla

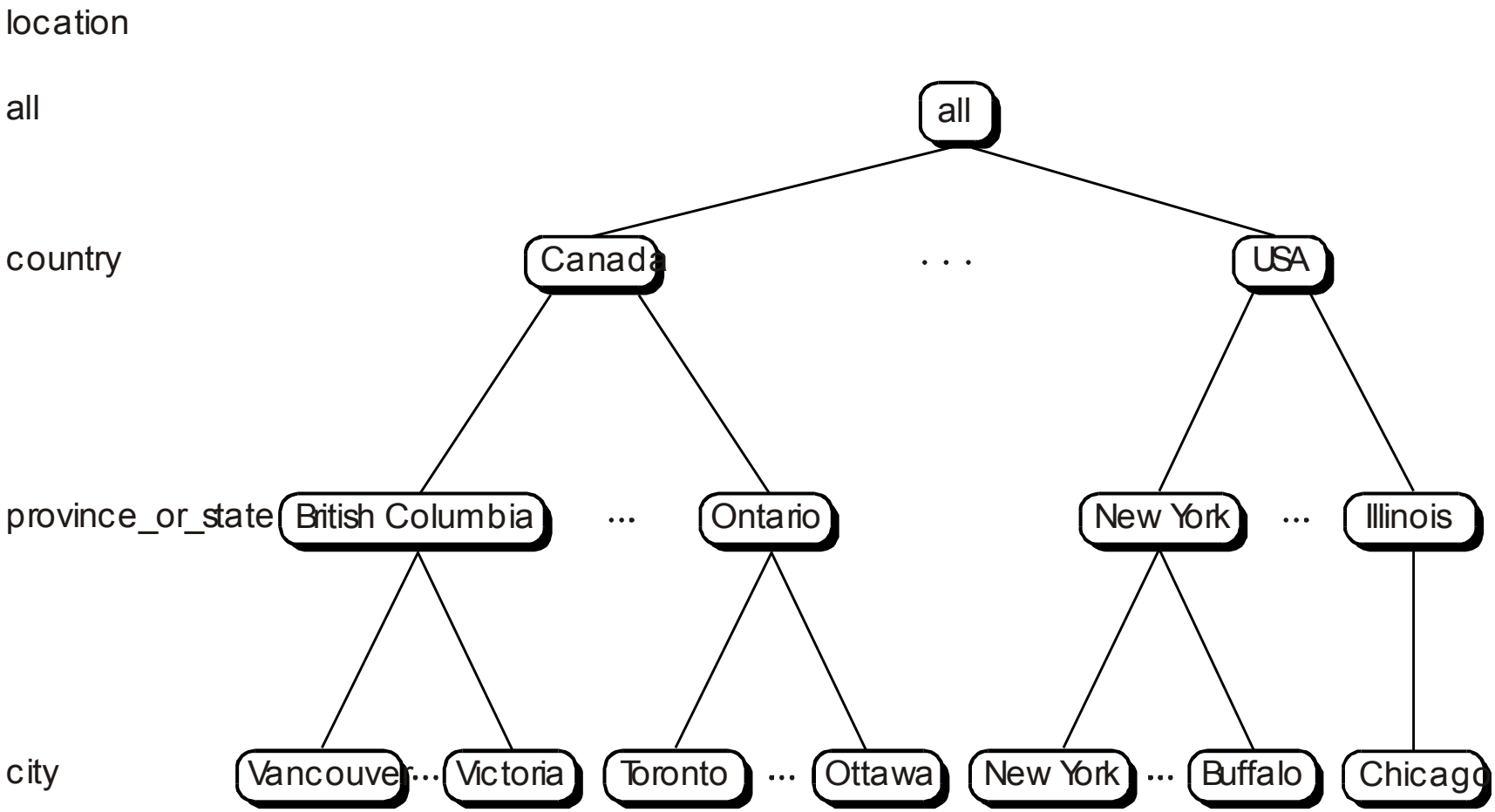
Construção de hierarquias:

- Manual via GUI
- Automática via clustering

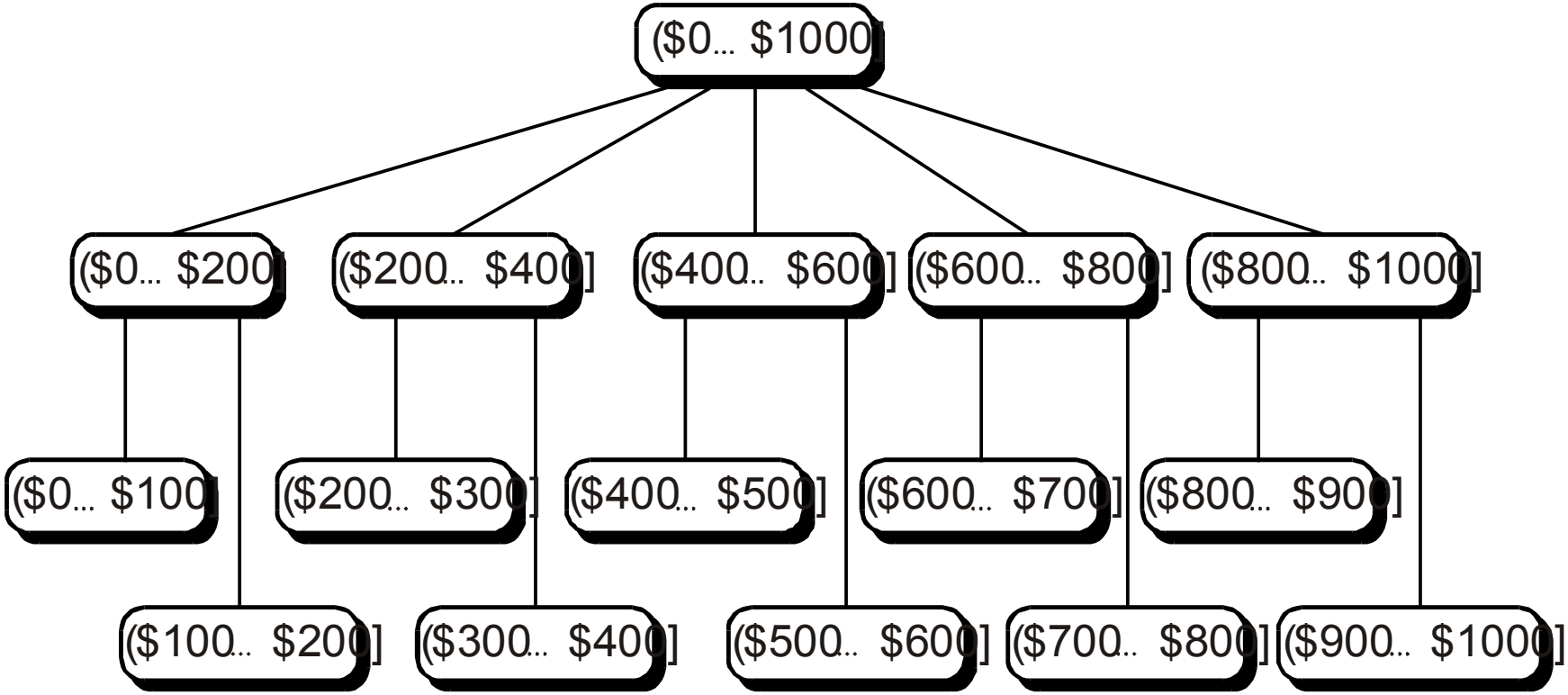




Exemplo de hierarquia conceitual esquemática



Exemplo de hierarquia conceitual de agrupamento



Operadores OLAP: navegação no espaço analítico multidimensional e multigranular

Operadores de *navegação* ao longo das *hierarquias conceituais*:

- **Roll-up**, abstrai detalhes, aplicando ao cuboide corrente um operador de agregação dado ao longo de uma dimensão dada
- *ex: região* $\rightarrow$ *pais*
- **Drill-down**, detalha o cuboide corrente desagregando ao longo de uma dimensão dada
- *ex: região* $\rightarrow$ *estado*
- **Drill-through**, detalha os valores, ao longo de uma dimensão dada, além do nível mais baixo do cuboide, por consultas SQL diretamente na fonte relacional
- **Drill-across**, detalha vários cuboides com dimensões compartilhas, por desagregação ao longo de

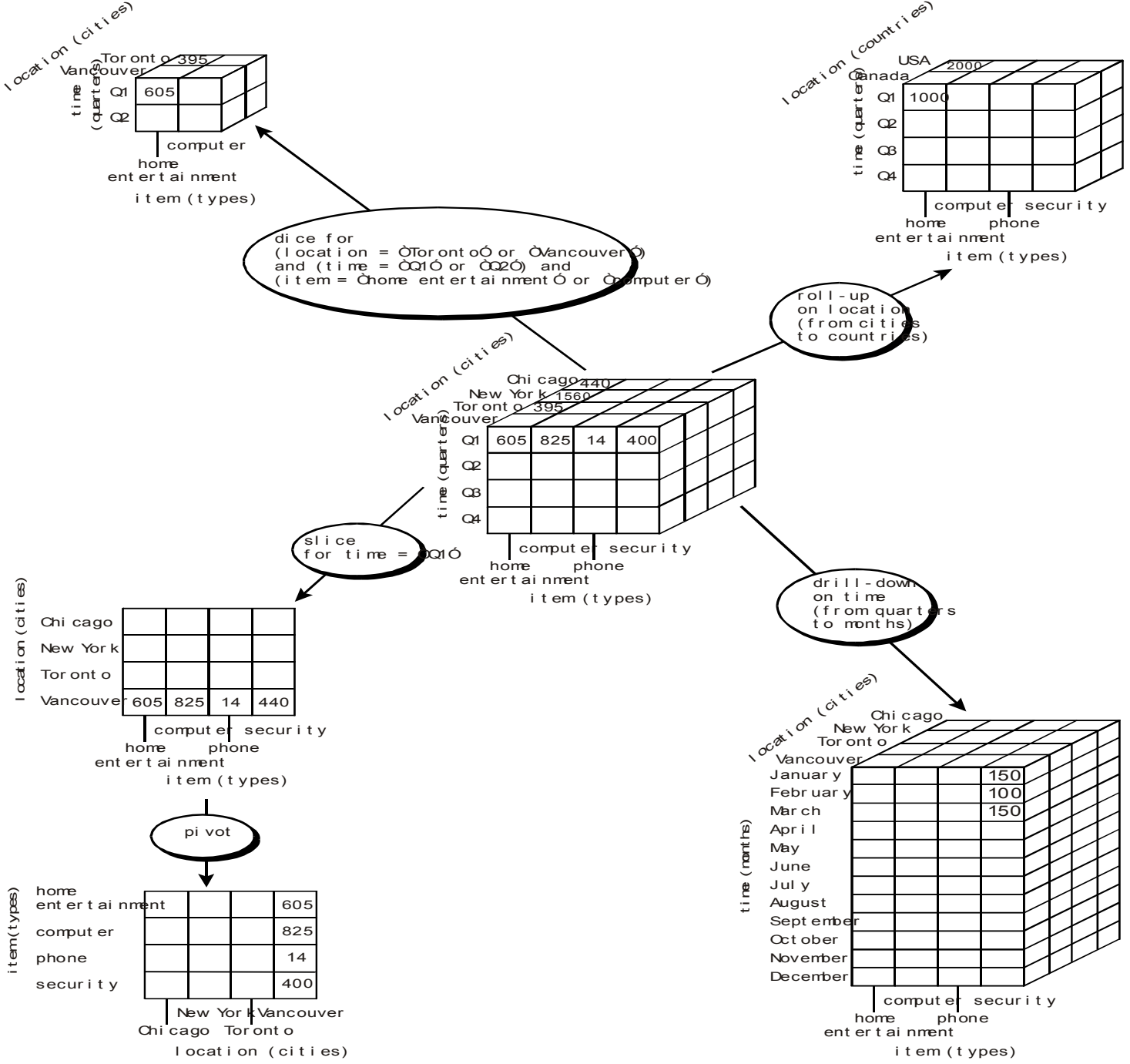
Operadores OLAP: navegação no espaço analítico multidimensional e multigranular

Operadores de *navegação* ao longo do *reticulado de cuboides*:

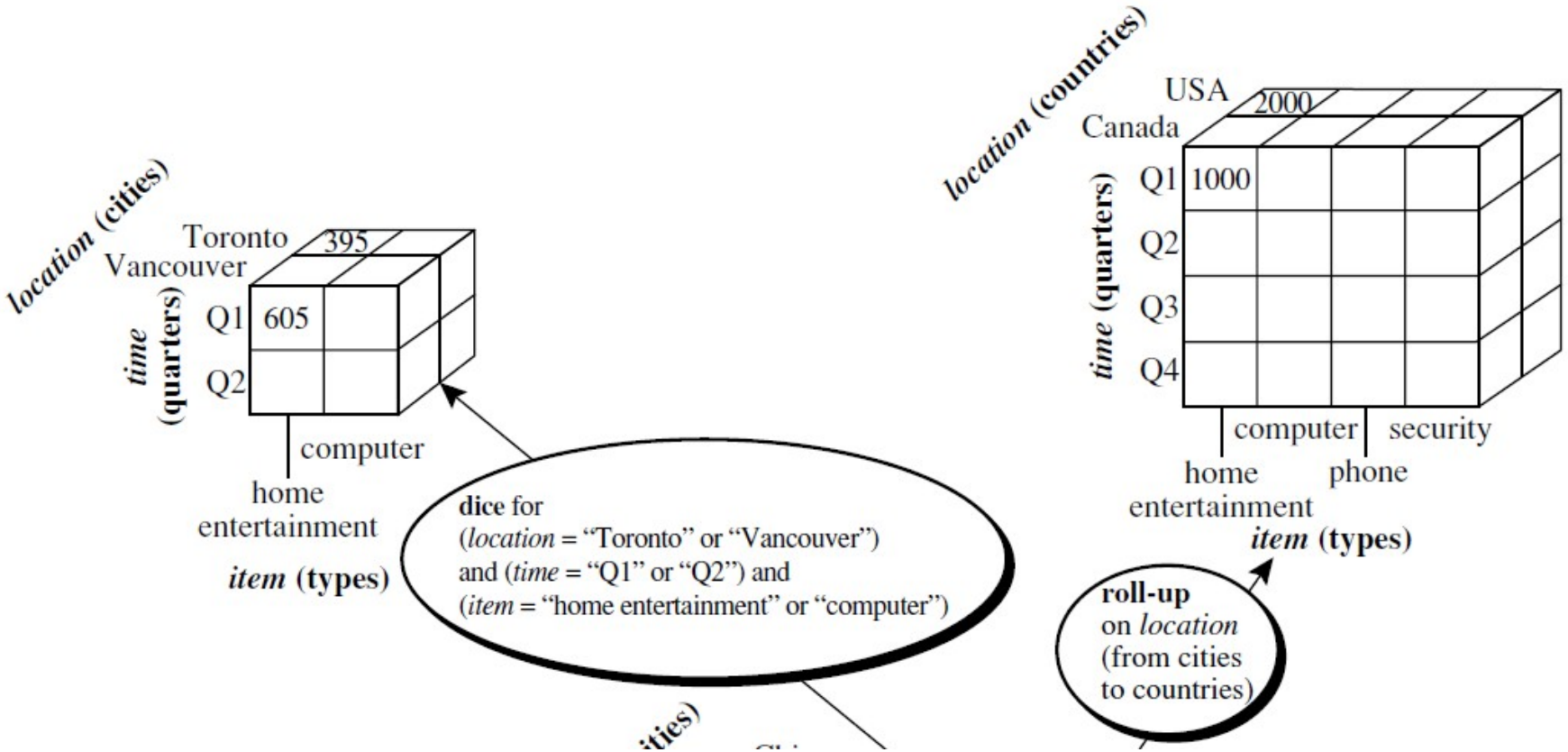
- **Slice**, extrair sub-cuboide das células verificando um restrições de valor ao longo de *uma* dimensão *(ex, time = Q1)*
- **Dice,** extrair sub-cuboide das células verificando um restrições de valor ao longo de várias dimensões *(ex, time = Q1 e item = HE)*

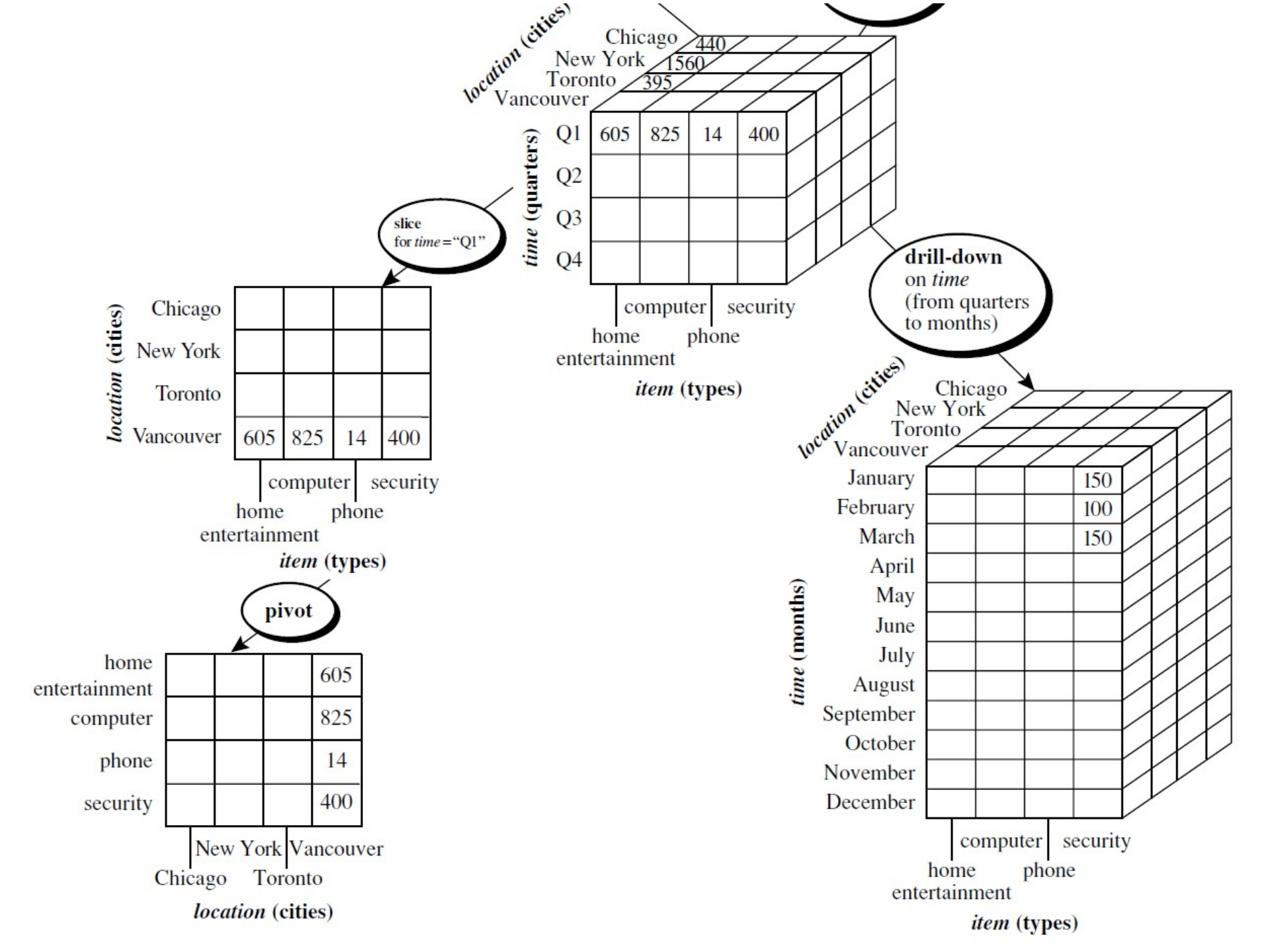
Operadores de *visualização* dos resultados:

- **Pivot,** mudar os eixos da visualização (cross-tab ou 3D grahics) do resultado de uma consultas *(ex, time na vertical no lugar da horizontal)*
- **Rank,** ordena os membros de uma dimensão de acordo com a ordem da medida corrente (ex, time retrospectivo, começando pelo mais recentes primeiro); serve também para filtragem


location (cities)
Toronto 395
Vancouver
time (quarters)
Q1 605
Q2
computer
home entertainment
item (types)
dice for
(location = "Toronto" or "Vancouver") and (time = "Q1" or "Q2") and (item = "home entertainment" or "computer")
location (countries)
USA 2000
Canada
Q1 1000
Q2
Q3
Q4
computer security
home phone entertainment
item (types)
roll-up on location (from cities to countries)
location (cities)
Chicago 440
New York 1560
Toronto 395
Vancouver
Q1 605 825 14 400
Q2
Q3
Q4
computer security
home phone entertainment
item (types)
slice for time = "Q1"
drill-down on time (from quarters to months)
location (cities)
Chicago
New York
Toronto
Vancouver 605 825 14 440
computer security
home phone entertainment
item (types)
pivot
item(types)
home entertainment 605
computer 825
phone 14
security 400
New York Vancouver
Chicago Toronto
location (cities)
location (cities)
Chicago
New York
Toronto
Vancouver
time (months)
January 150
February 100
March 150
April
May
June
July
August
September
October
November
December
computer security
home phone entertainment
item (types)

Operadores OLAP: Roll-up e dice




location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
time (quarters)
Q1
Q2
Q3
Q4
605
825
14
400
computer
security
home
entertainment
phone
item (types)
slice
for time = "Q1"
location (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
computer
security
home
entertainment
phone
item (types)
drill-down
on time
(from quarters
to months)
location (cities)
Chicago
New York
Toronto
Vancouver
time (months)
January
February
March
April
May
June
July
August
September
October
November
December
150
100
150
computer
security
home
entertainment
phone
item (types)
pivot
home
entertainment
computer
phone
security
605
825
14
400
New York
Vancouver
Chicago
Toronto
location (cities)

Modelos físicos de dados para OLAP

ROLAP (OLAP Relacional):

- Armazena dados em tabelas relacionais
- Reaproveita da tecnologia relacional, inclusive SQL
- Apenas *apresenta* dados de maneira multidimensional
- Permite acoplamento mais estreito com fontes OLTP (geralmente relacionais)
- Porém, necessita remodelagem prévio de dados em esquema especializados (estrela, floco de neve)

MOLAP (OLAP Multidimensional):

- Armazena dados em arrays de dimensões N
- Necessita desenvolvimento de novas técnicas de otimização
- Sem acesso a granularidade mínima (*i.e.,* única transações)

HOLAP (OLAP Híbrido):

- Duplica dados
- Tabelas para dados atómicos
- Arrays para agregrados
- Flexível e rápido de execução
- Custoso em memória e desenvolvimento

Modelos de dados ROLAP: *Estrela*

Uma tabela de fato com:

- uma coluna por medida agregada
- uma columa por chave de dimensão analítica

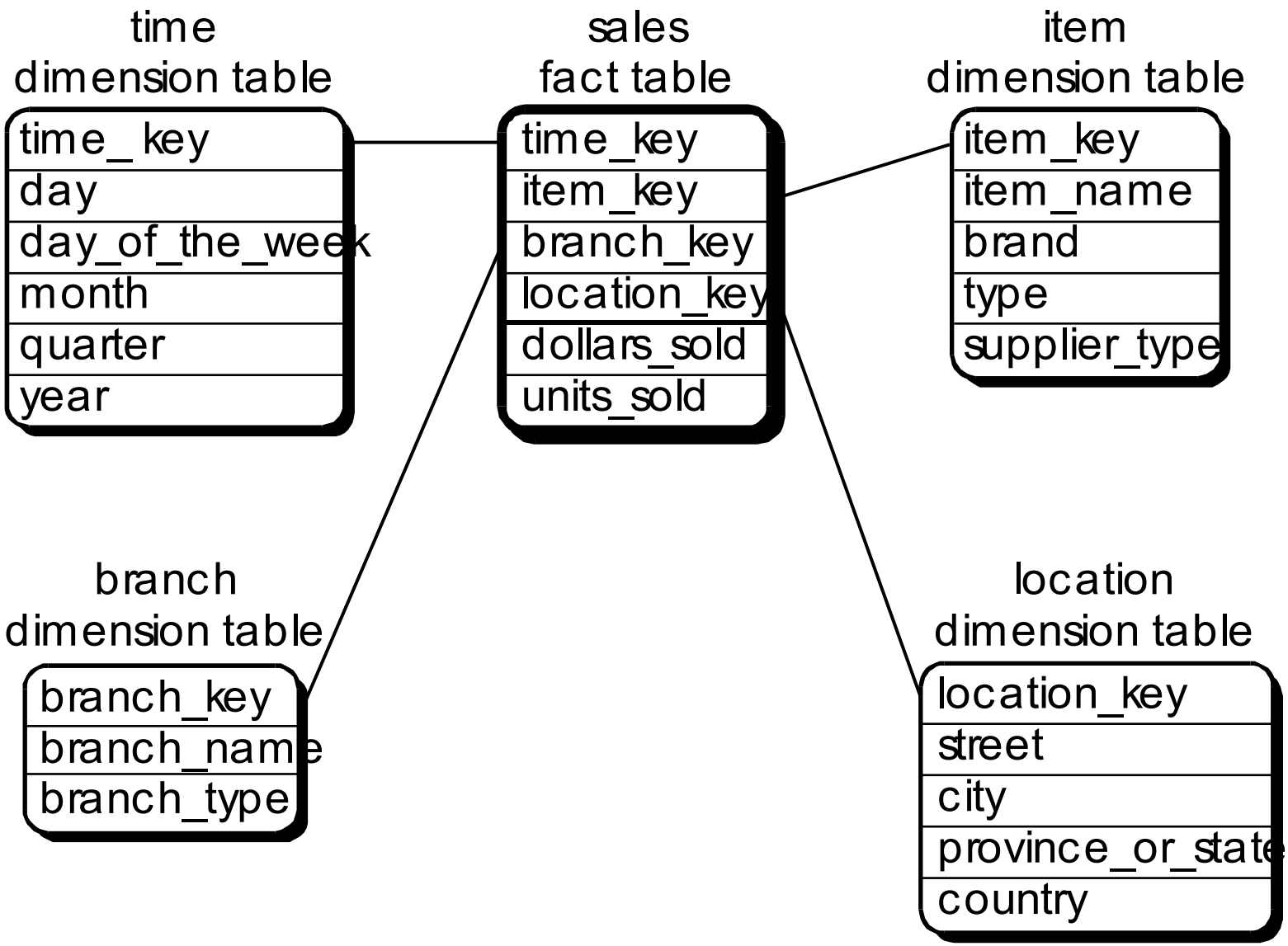
N tabelas de dimensões, uma por dimensão analítica

- uma coluna por para cada atributo descrevendo a dimensão
- geralmente um atributo por nível na hierarquia conceitual

Não normalizada:

- alguma redundância
- alguns níveis e membros aparecem em vários registros

Modelo estrela: exemplo



Modelos de dados ROLAP: *Floco de Neve*

Igual ao modelo estrela exceto pela *normalização das tabelas de dimensões*

Vantagens

- Facilita evolução das dimensões
- Reduz espaço ocupado por elas

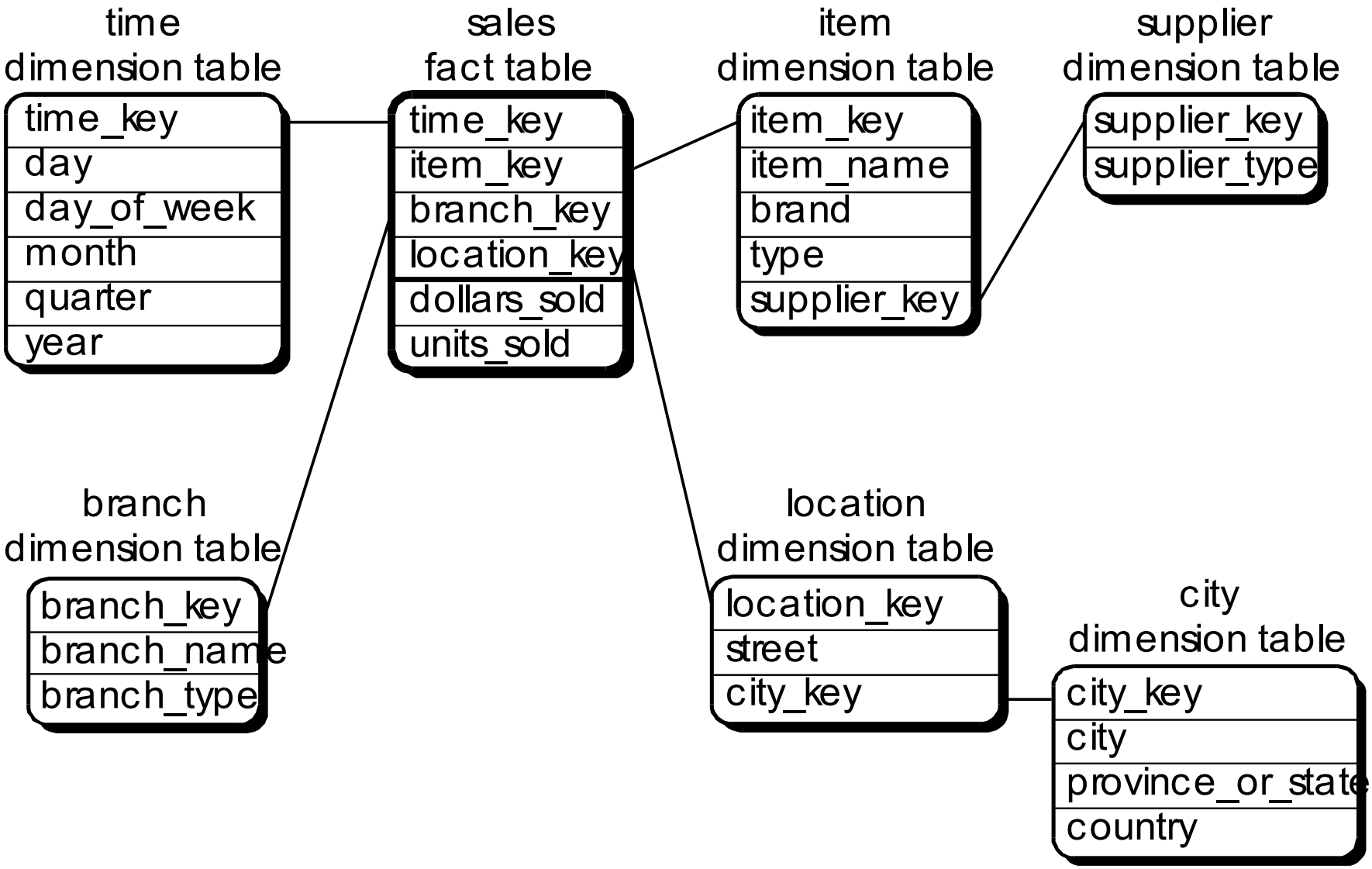
Desvantagens:

- Aumenta tempo de resposta pela necessidade de junções

Balanço:

- Espaço ganhado negligível já que espaço total do data mart é principalmente ocupado pela tabela de fato
- Modelo estrela mais popular

Modelo floco de neve: exemplo



Modelos de dados ROLAP: *Constelação*

Várias tabelas de fato: um por assunto analítico

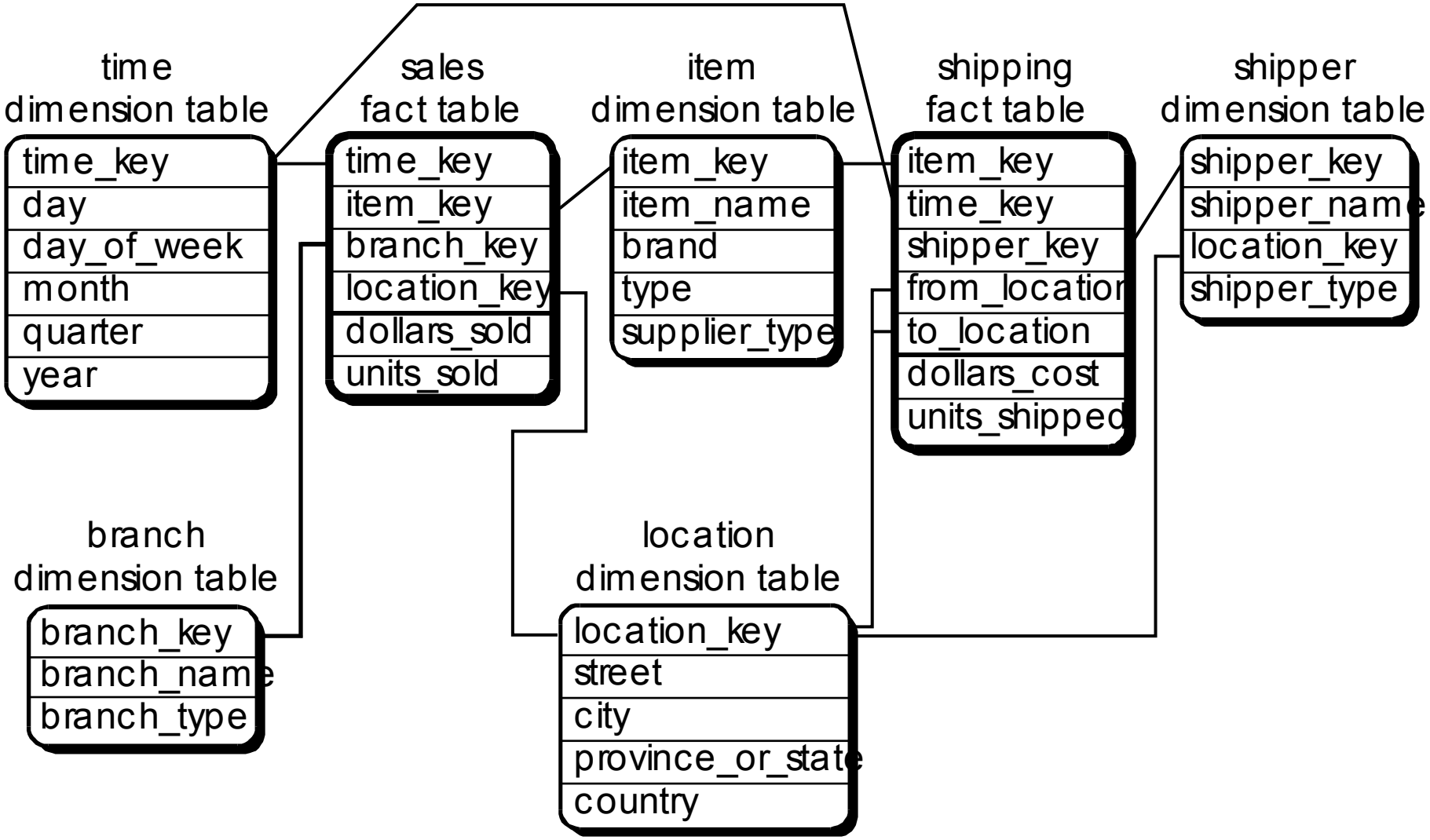
Uma tabela dimensão por dimensão analítica de algum assunto

As dimensões compartilhadas por vários assuntos não são duplicadas, mas apontadas por várias tabelas de fato

Em geral:

- data mart modelado em estrela
- data warehouse modelado em constelação
- data mart integrado em um data warehouse por:
  - uniformização das tabelas de dimensões dos vários data marts
  - ligações entre elas e as tabelas de fato

Modelo *constelação:* exemplo



Elementos de um modelo de dados lógico multidimensional

BDMD: coleção de cuboides D-dimensionais

Cuboides:

- D dimensões

(*ex,* tempo, produto, espaço)

- C **celulas** de dados quantitativos atómicos = valores das **medidas**

Dimensão:

- H **hierarquias** de N níveis de granularidade

(ex, ano/mês/dias, ano/semestre/semana)

Nível: E **membros**

(ex, {Jan, ..., Dez}, {1, ..., 31})

Cellset: subcubo resultado de uma consulta OLAP selecionando:

- um cubo A do DBMD
- d dimensões de A como analíticas
- m dimensões de A como medidas
- para cada d:

  uma hierarquia h_d

  um nivel n_d com m_d membros
- para cada m, uma função de agregação (sum, max, avg, var)

$\forall$ Π m_d celulas, cada uma contendo m dados agregados