Uma abordagem lógica para base de dados multidimensionais

Alessandro Elias
aelias@c3sl.ufpr.br
Professora: Carmem Satie Hara

Disciplina CI087 - Tópicos em Banco de Dados

Universidade Federal do Paraná

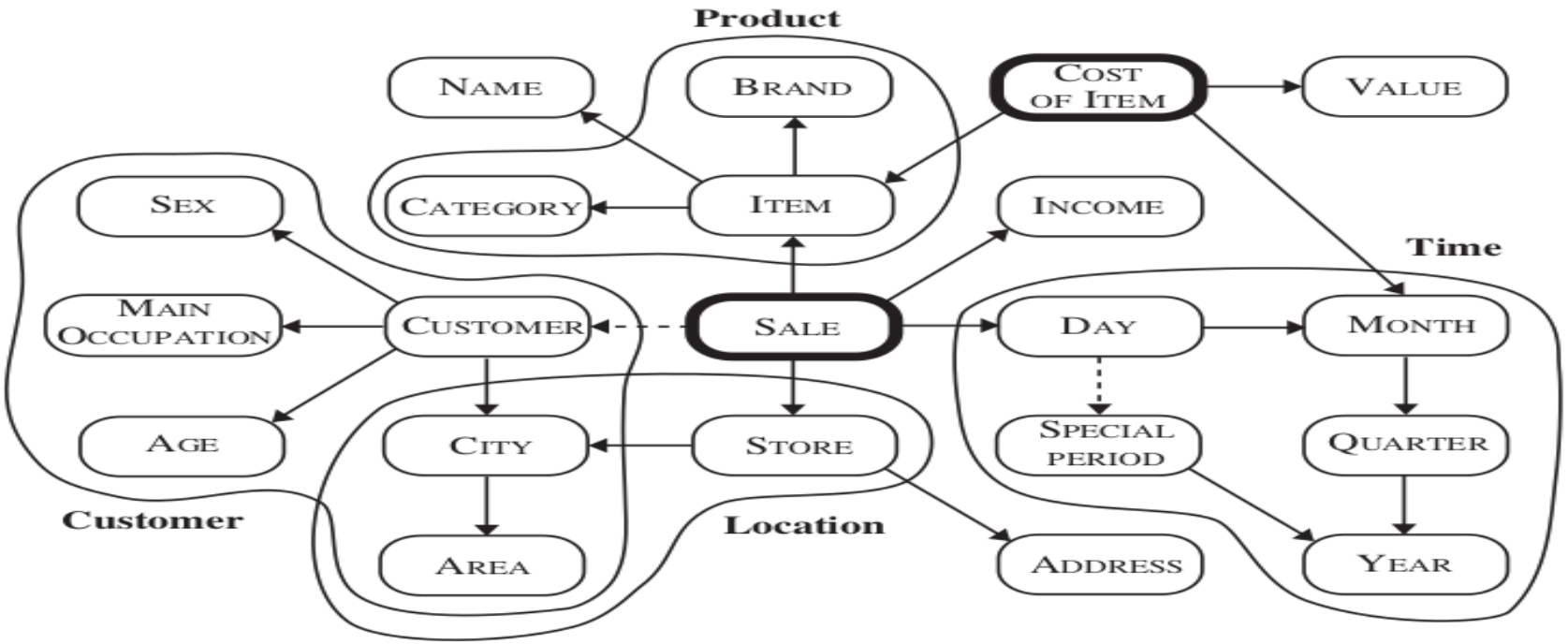
12 de Junho 2018

Agenda

1. Perguntas
2. Background
   - Arquitetura de um data warehouse
   - Taxinomia de ROLAP e MOLAP
   - Tabelas fato e dimensões
3. Artigo
   - Introdução
   - Modelando base de dados MultiDimensional
   - Design de uma base de dados MultiDimensional
     - Identificando tabelas fato e dimensões
     - Reestruturando o esquema E-R
     - A partir da E-R construir o grafo
     - Traduzir para o modelo MultiDimensional
4. Conclusão

Questões

1. Cite os passos para reestruturar o esquema entidade relacionamento, e descreva como ocorre apenas um dos passos.
2. Observe o grafo dimensional e cite apenas um conjunto de entidades que respeitam a propriedade $l_1 \preceq l_2 \preceq l_3$.

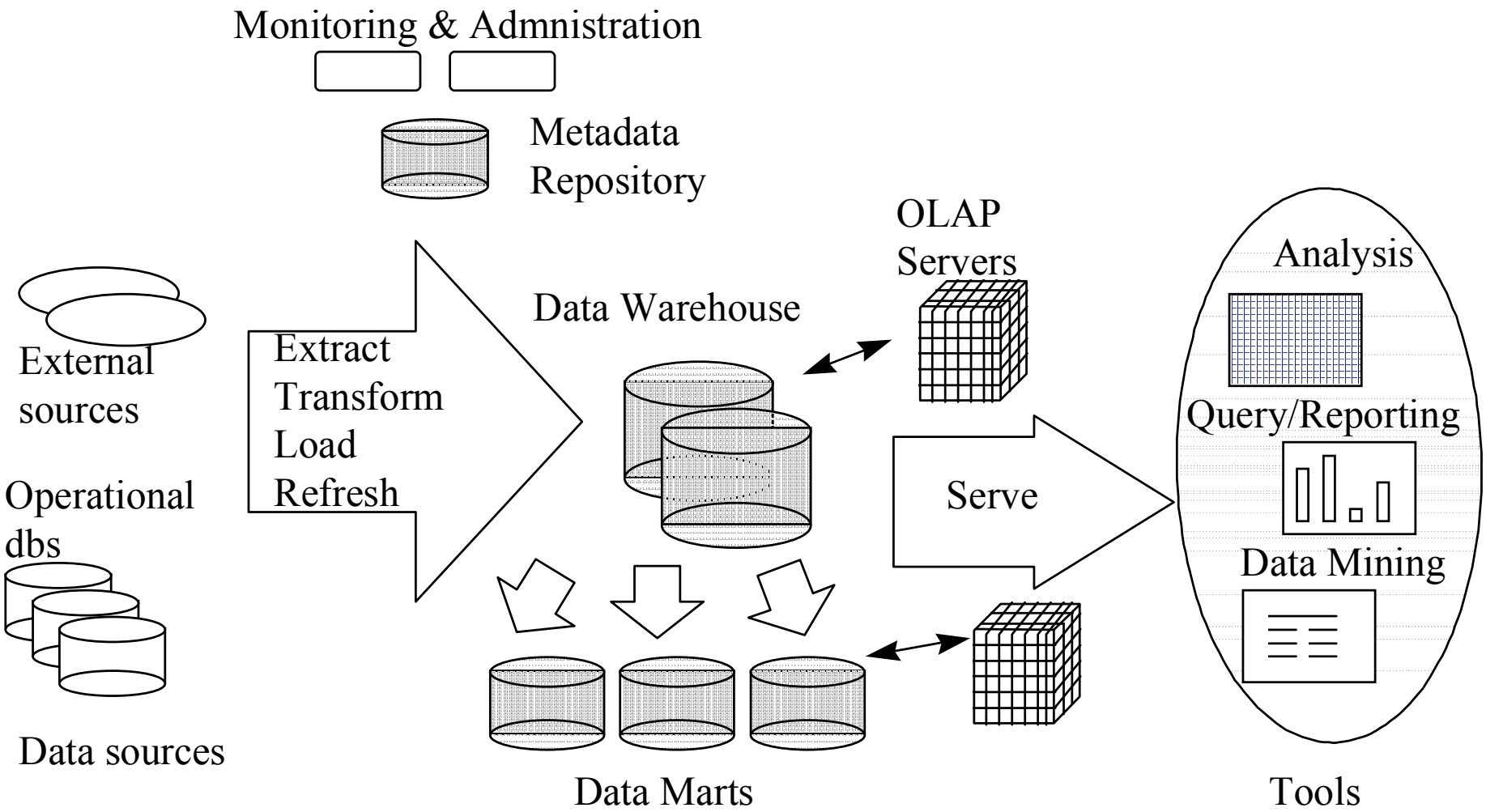


aelias@c3sl.ufpr.br

Introdução

- O que é data warehouse?
  (Uma coleção de dados integrados de toda uma empresa, orientado para tomada de decisões.)
- Onde é feito as consultas de análise para tomadas de decisão?
  (Não é feito diretamente no data warehouse, mas em data store especial, chamado de *hypercubes ou multidimensional*.)
- Para análise e tomada de decisão de uma empresa, os dados são organizados em dimensões, como categoria de produto, localização geográfica, temporal; são de maior valor (para tomada de decisões) do que saber apenas o total faturado.

Arquitetura de um data warehouse



Taxonomia de sistemas ROLAP e MOLAP

- ROLAP - Relational Online Analytical Processing
  (Armazenamento dos dados analíticos em banco de dados relacional tradicional.)
- MOLAP - Multidimensional Online Analytical Processing
  (Armazenamento dos dados em um BD multidimensional matricial.)

Funcionamento do modelo ROLAP

- Funciona direto com um banco de dados relacional tradicional. (dimensões são armazenas em tabelas relacionais)
- ROLAP não faz uso de cubos pré-calculados, logo as queries são como em qualquer banco de dados relacional.

Funcionamento do modelo MOLAP

- Base de dados MOLAP são geralmente referenciadas como apenas OLAP.
- Geralmente requerem um pré-processamento para serem carregados na base de dados.
  (conhecido como consolidação dos dados, cujo dados são chamados de cubos)
- Um cubo de dados possui as respostas de um conjunto de perguntas.
- Atualizações em OLAP podem levar tempo para serem processadas.

Representando OLAP como um cubo

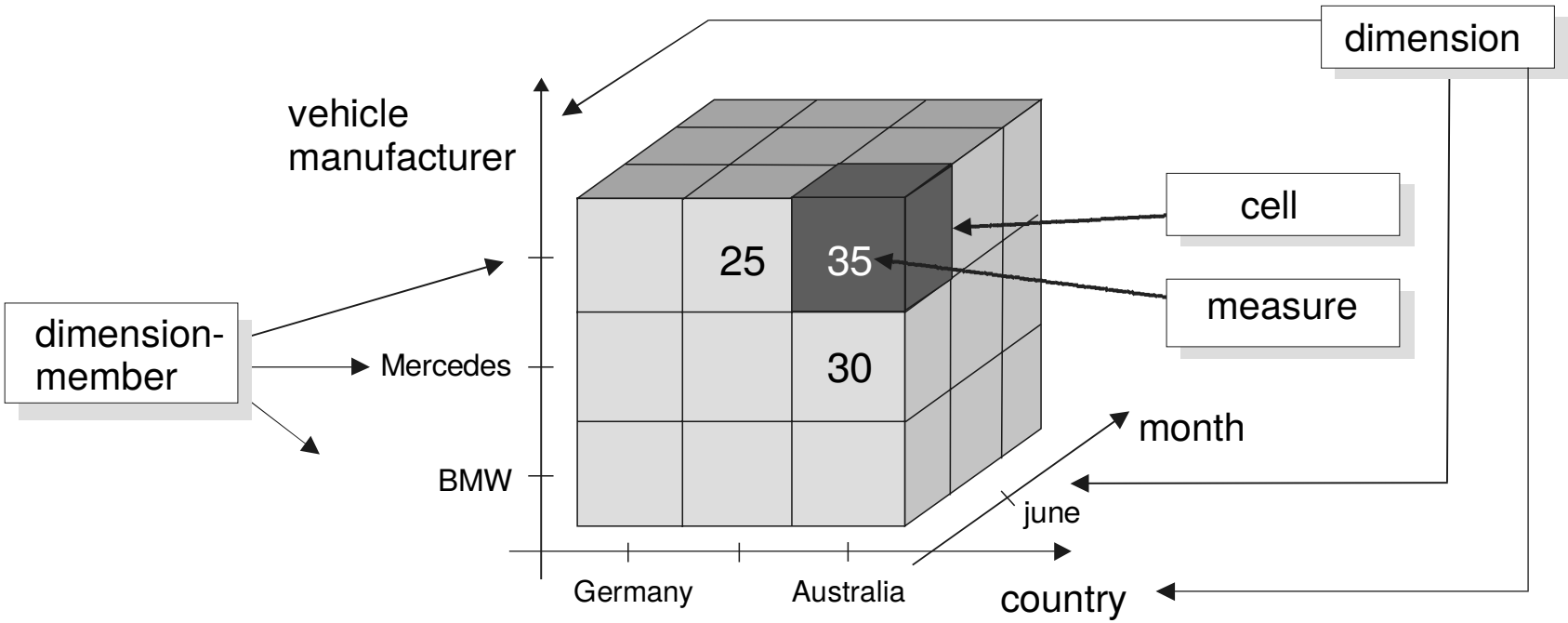
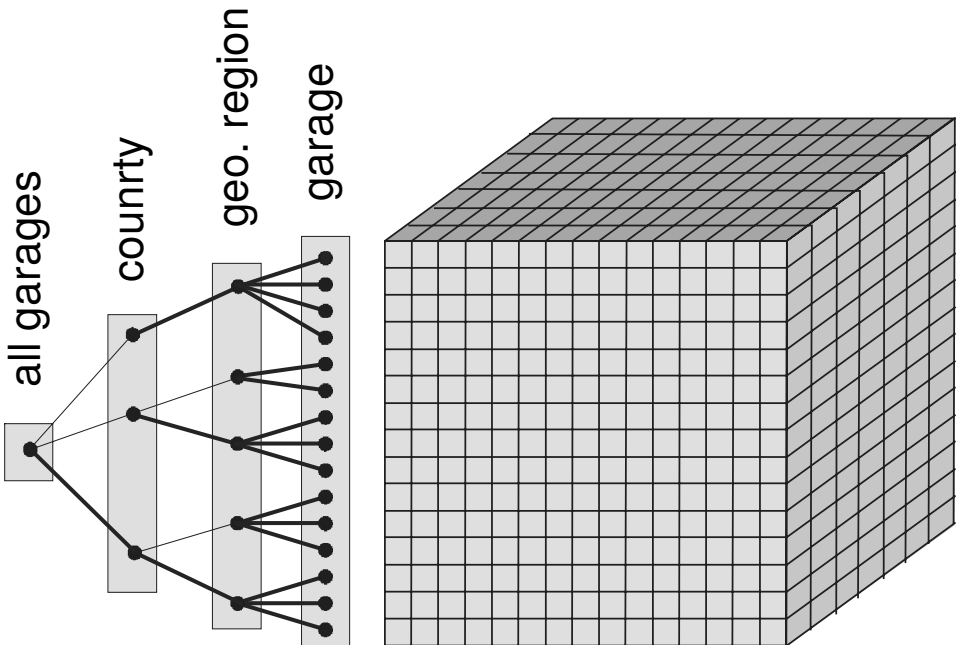


Figure: [Sapia et al., 1999]

OLAP níveis hierárquicos



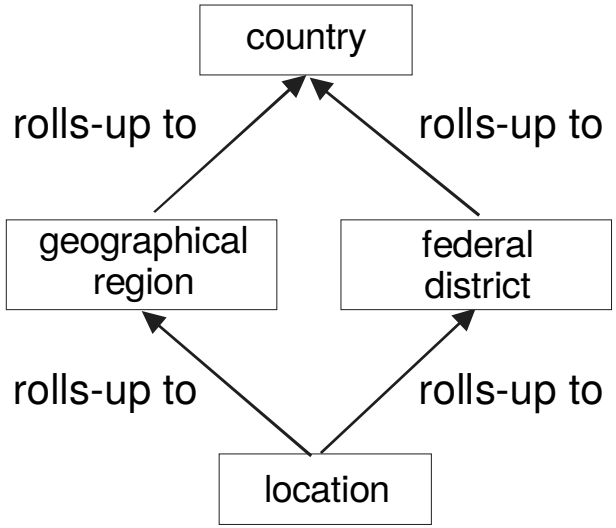


Figure: [Sapia et al., 1999]

Vantagens de sistemas ROLAP e MOLAP

ROLAP

- É considerado mais escalável quando o volume de dados é extremamente grande (milhões de tuplas)
- O tempo de executar ETL (Extract, Transform, Load) é menor, devido a habilidade de otimizar as operações ETL.
- Os dados são armazenados em BD relacionais, pode ser acessado diretamente com SQL, amplamente difundido, ou seja, não depende de uma ferramente OLAP.
- Manipula melhor dados não agregados. MOLAP tende a ser mais lento.

MOLAP

- Alta performance devido a otimização de armazenamento, indexação e cache em multi-dimensões.
- Poupa spaço de armazenamento em disco, diante das técnicas de compressão quando comparado com relacional.
- Alta automatização devido ao alto grau de agregação de dados.
- Modelos matricial provê indexação natural.
- Extração efetiva dos dados, devido a pré-estruturação das agregações.

Desvantagens de sistemas ROLAP e MOLAP

ROLAP

- O código para fazer ETL em tabelas de agregamento tem que ser customizado.
- Caso não seja feita as tabelas agregadas a performance pode degradar.
  (work around criar as tabelas agregadas; de qualquer forma é impossível criar todas as combinações de agregações das dimensões/atributos)
- BD relacionais possui seu próprio mecanismo de indexação e cache, logo não possui o mecanismos de indexação como OLAP, contudo existem alguns ROLAPs que implementam Cube Views (DB2).

MOLAP

- O processo de carregamento dos dados pode ser extremamente lento.
  (work around fazer carregamento incremental)
- Algumas metodologias introduzem redundância de dados.

Tabelas fato e dimensões

- Cada conjunto de tuplas da tabela fato representa um evento.
- Os atributos da tabela fato que são chaves estrangeiras representam dimensões.
- As dimensões devem geralmente responder uma das perguntas: who, what, where, when, how e why.

Tabelas fato e dimensões

dim_product table

product_sk	sku	description	brand	category
30	OK4012	Bananas	Freshmax	Fresh fruit
31	KA9511	Fish food	Aquatech	Pet supplies
32	AB1234	Croissant	Dealicious	Bakery

dim_store table

store_sk	state	city
1	WA	Seattle
2	CA	San Francisco
3	CA	Palo Alto

fact_sales table

date_key	product_sk	store_sk	promotion_sk	customer_sk	quantity	net_price	discount_price
140102	31	3	NULL	NULL	1	2.49	2.49
140102	69	5	19	NULL	3	14.99	9.99
140102	74	3	23	191	1	4.49	3.89
140102	33	8	NULL	235	4	0.99	0.99

dim_date table

date_key	year	month	day	weekday	is_holiday
140101	2014	jan	1	wed	yes
140102	2014	jan	2	thu	no
140103	2014	jan	3	fri	no

dim_customer table

customer_sk	name	date_of_birth
190	Alice	1979-03-29
191	Bob	1961-09-02
192	Cecil	1991-12-13

dim_promotion table

promotion_sk	name	ad_type	coupon_type
18	New Year sale	Poster	NULL
19	Aquarium deal	Direct mail	Leaflet
20	Coffee & cake bundle	In-store sign	NULL

Figure: [Kleppmann, 2014]

Operadores analitícos

- ROLL-UP
  (aumenta o nível de agregação).
- DRILL-DOWN
  (diminui o nível de agregação, ou aumenta os detalhes)
- SLICING e DICING
  (seleção e projeção)
- PIVOT
  (re-ordenação dos dados de uma visualização multidimensional)

Artigo

Uma abordagem lógica para base de dados multidimensionais.
[Cabibbo e Torlone, 1998]

Introdução

- A abordagem deste artigo visa uma abordagem alto nível, uma abordagem geral para a construção de um banco de dados multidimensional.
- As definições existentes em outros trabalhos geralmente possuem uma abordagem dependente de implementação (relacional ou proprietária multidimensional).
- Abordaremos os conceitos e como é o design de uma base de dados multidimensional.

Modelando a base de dados MultiDimensional

- O modelo de dados MultiDimensional é baseado em duas construções, com a dimensão e tabela fato.
- Cada dimensão é organizado em níveis hierárquicos, correspondendo ao domínio de dados de diferentes granularidades.
- Um nível pode ter uma descrição associado à ele.
- Dentro de uma dimensão diferentes valores podem estar associados a uma família de funções de roll-up.

Definição formal

- Fixamos dois conjuntos disjuntos enumeráveis, *nomes* e *valores*.
- Denotaremos por Γ o conjunto de *nomes*, chamados de níveis.
- Cada nível $l \in \Gamma$ é associado com o conjunto enumerável *valor*, chamado de *domínio* de l e denotado por $DOM(l)$.
- Os vários domínios são disjuntos dois à dois.

Definição de dimensão

Em $\mathcal{MD}$ uma dimensão consiste de:

- Um conjunto finito de níveis $L \subseteq \Gamma$;
- Uma ordem parcial $\preceq$ em L é $l_1 \preceq l_2$, denota que l_1 rolls up para l_2;
- Uma família de funções roll-up, incluindo a função $R - UP_{l_1}^{l_2}$ origem $DOM(l_1)$ para $DOM(l_2)$ para cada par de níveis $l_1 \preceq l_2$ - dizemos que $R - UP_{l_1}^{l_2}(o_1) = o_2$ denota o_1 rolls up para o_2.

Definição de esquema

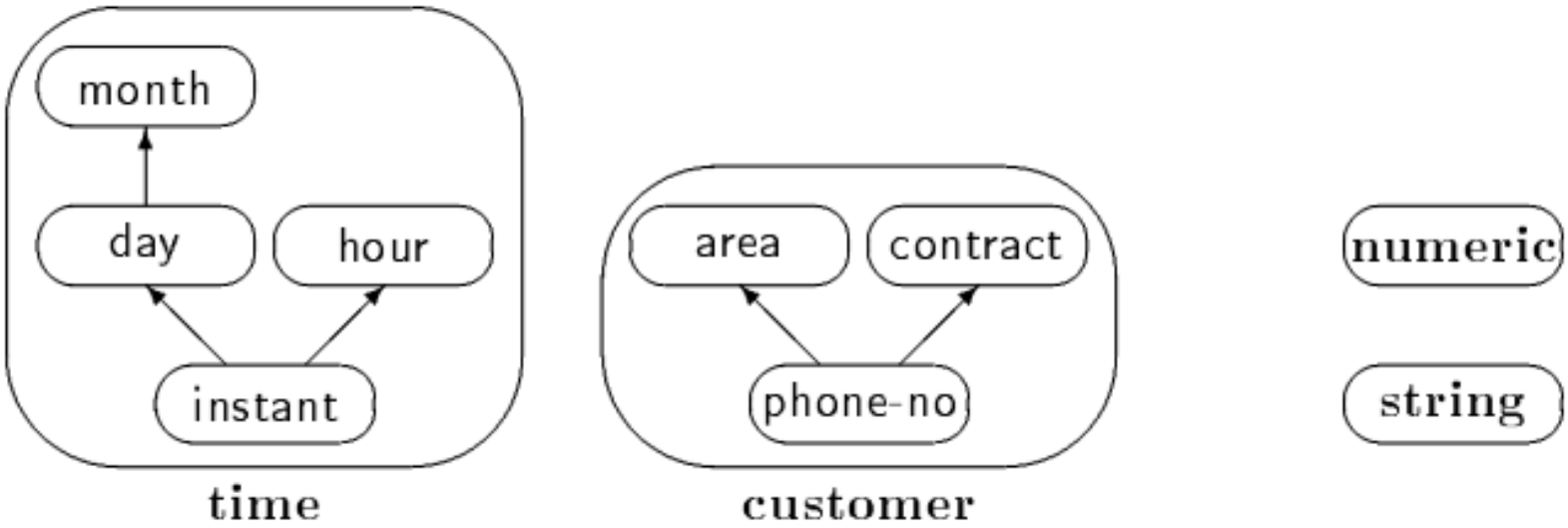
Em $\mathcal{MD}$ um esquema consiste de:

- Um conjunto finito $\mathcal{D}$ de dimensões.
- Um conjunto finito $\mathcal{F}$ de esquemas de tabelas fato da seguinte forma

$$f[A_1 : l_1\langle d_1\rangle, ..., A_n : l_n\langle d_n\rangle] : l_0\langle d_0\rangle$$

  onde f é o *nome*, e cada A_i 1 $(\leq i \leq n)$ é um *nome* distinto chamado de atributo de f, e cada l_i $(0 \leq i \leq n)$ é um nível da dimensão d_i;
- Um conjunto finito Δ de descrições de níveis na forma $\delta(l) : l'$, onde l e l' são níveis, e δ é o nome chamado de descrição de l.

Exemplo de TelCo esquema

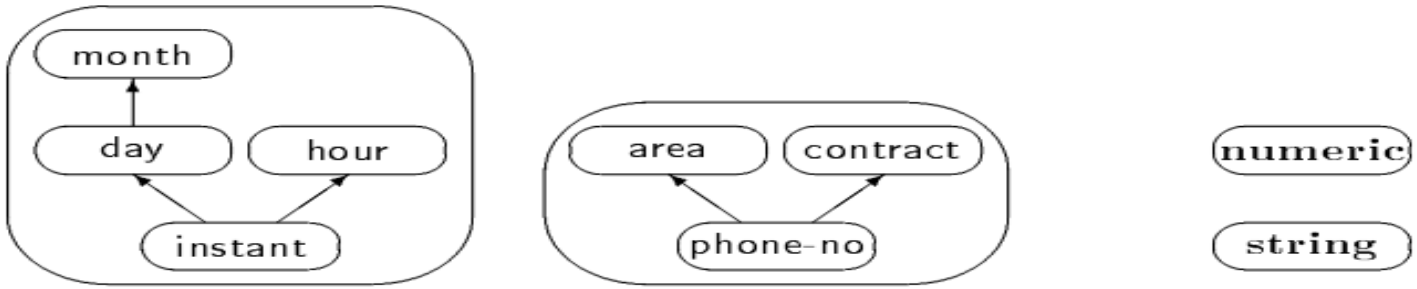


RATE [*hour* : hour, *contract* : contract, *calling-area* : area, *called-area* : area] : numeric
DURATION [*calling* : phone-no, *called* : phone-no, *start* : instant] : numeric
MONTHLY-BILL [*customer* : phone-no, *period* : month] : numeric

Owner (phone-no) : string

Instância da TelCo

- Suponha uma analise da TelCo sobre chamadas telefônicas. (estamos interessados nas dimensões **time** e **customer**)
- Dentro do domínio **instant** temos um timestamp como Jan 5, 97, 10AM:45:21.
- Observe que do timestamp rolls-up para 10AM no nível **hour** e Jan 5, 97 no nível **day**.



RATE [*hour* : hour, *contract* : contract, *calling-area* : area, *called-area* : area] : numeric
DURATION [*calling* : phone-no, *called* : phone-no, *start* : instant] : numeric
MONTHLY-BILL [*customer* : phone-no, *period* : month] : numeric

Owner (phone-no) : string

Coordenada e Instância

Seja $\mathcal{S} = (\mathcal{D}, \mathcal{F}, \Delta)$ um $\mathcal{MD}$ esquema

- Uma coordenada sobre uma f-table $f[A_1 : l_1\langle d_1\rangle, ..., A_n : l_n\langle d_n\rangle] : l_0\langle d_0\rangle$ em $\mathcal{F}$ é uma função que mapeia cada atributo A_i $(1 \leq i \leq n)$ para um elemento no $DOM(l_i)$.
- Uma instância sobre f é uma função parcial que mapeia coordenadas sobre f para elementos do $DOM(l_0)$.
- Uma instância sobre a descrição de nível $\delta(l) : l'$ em Δ é uma função parcial do $DOM(l)$ para $DOM(l')$.

Instância da TelCo esquema

hour	*contract*	*calling-area*	*called-area*	RATE
6AM	*Family*	*06*	*02*	0.44
7AM	*Family*	*06*	*02*	0.72
8AM	*Family*	*06*	*02*	1.12
		...		...
6AM	*Pro*	*06*	*055*	0.80
7AM	*Pro*	*06*	*055*	0.80
8AM	*Pro*	*06*	*055*	1.35
		...		...

MONTHLY-BILL	*Jan-97*	*Feb-97*	*Mar-97*
06-555-123	129	231	187
06-555-456	429	711	664
02-555-765	280	365	328

phone-no	*Owner*
06-555-123	*John*
06-555-456	*Ann*
02-555-765	*Mary*

Exemplo de instância da TelCo esquema

- Uma coordenada da f-table
  RATE [hour : 7AM, contract: Family, calling-area : 06, called-area: 02].
- Esta instancia associa a taxa 0.72.

Design de uma base de dados MultiDimensional.

Design de uma base de dados MultiDimensional.

- Partimos da premissa que temos um esquema E-R.
- Assumimos que este esquema descreve um primitivo data warehouse.
- Assumimos que este esquema não possui qualquer generalização hierárquica.
- Assumimos que todos os atributos são simples.
  (não são multi valorados ou atributos compostos)
- E por fim que o esquema é completamente normalizado.

Quatro passos para a construção do $\mathcal{MD}$ DB.

1. Identificar tabelas fato e dimensões.
2. Reestruturar o esquema E-R.
3. A partir da E-R construir o grafo.
4. Traduzir para o modelo $\mathcal{MD}$.

Passo 1

Identificar tabelas fato e dimensões.

E-R esquema do clássico exemplo de varejo.

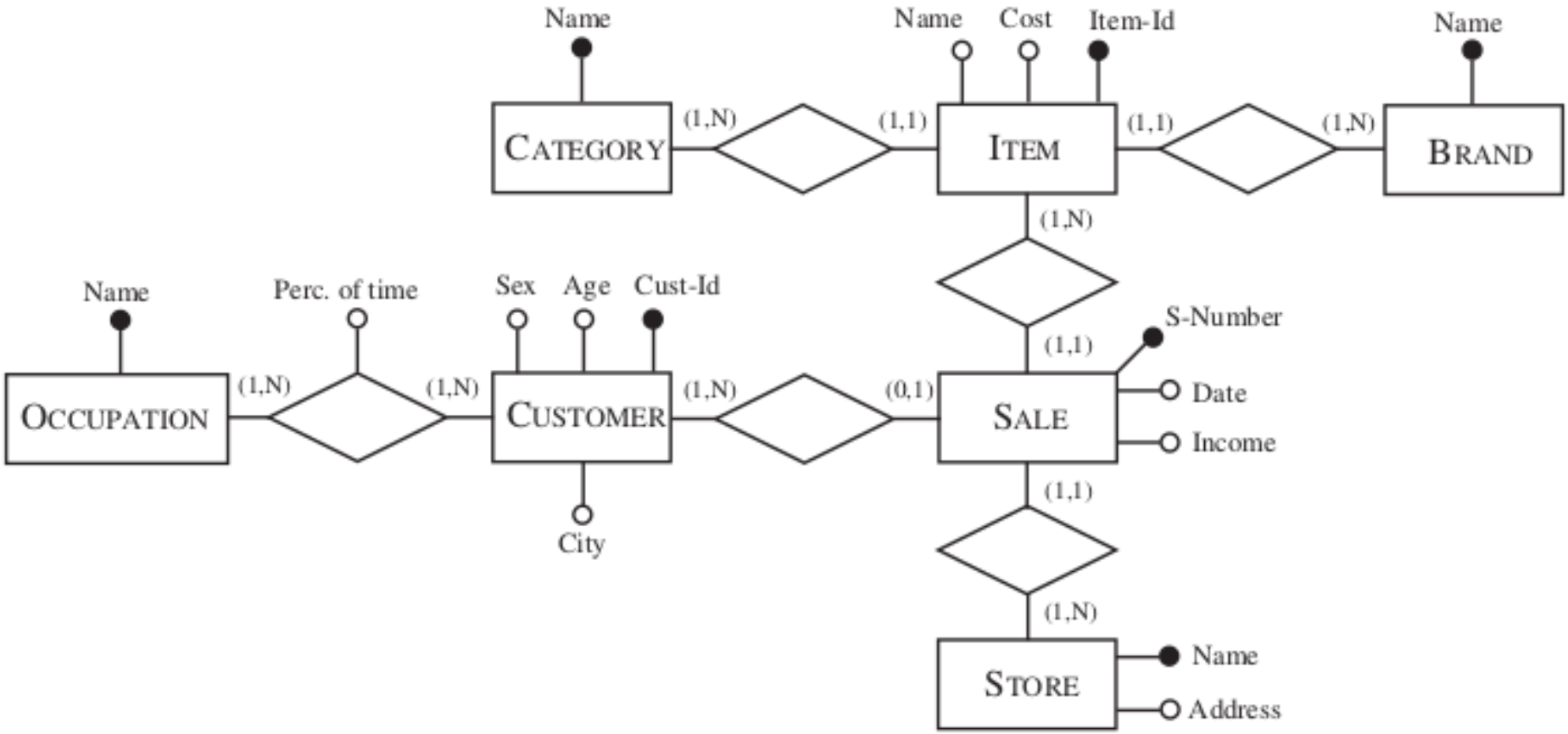


Figure: E-R esquema de uma empresa de varejo

1. Identificando tabelas fato e dimensões

- Chamamos de *fatos* dentro do conceito de E-R, entidades, relacionamentos ou atributos.
- Medidas (o que queremos mensurar na tabela fato) é uma propriedade atômica sobre um fato que queremos analisar.
- *Dimensões* é um sub-esquema do esquema E-R.
  (descreve a perspectiva no qual se quer fazer a analise)
- Suponha que estamos interessados em volume de vendas e variação do custo de um produto.
  (neste caso as entidades que nos interessa é *SALE* e o atributo *Cost* da entidade *ITEM* e o atributo *Income*)
- Queremos mensurar número de vendas.
  (número de instâncias da entidade *SALE*)

Passo 2

Reestruturando o esquema E-R

2. Reestruturando o esquema E-R

- Objetivo é criar uma nova E-R que mapeie para o modelo MultiDimensional.
- *Representando fatos como entidades*.
  (como ex. o custo de produção na E-R esquema é representado por um atributo, este será transformado em uma entidade)

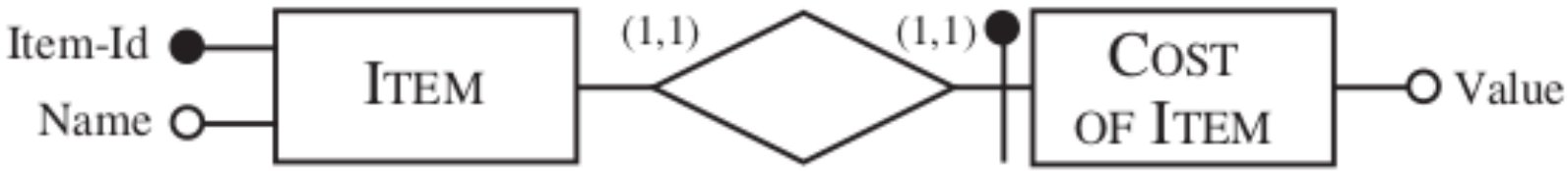


Figure: Reestruturando entidade *ITEM* da E-R esquema

2. Reestruturando o esquema E-R

- *Adicionado dimensões*.
  (em interesse de analise do fato do custo de um item, poderíamos estar interessados em uma analise temporal, logo esta informação precisa estar explícito na E-R)
- Imagine o cenário em que estamos interessados em uma analise mensal. (reestruturamos a entidade *COST OF ITEM*
  P.S.: a informação de como deve ser analise histórica vêm de meta dados da fonte.

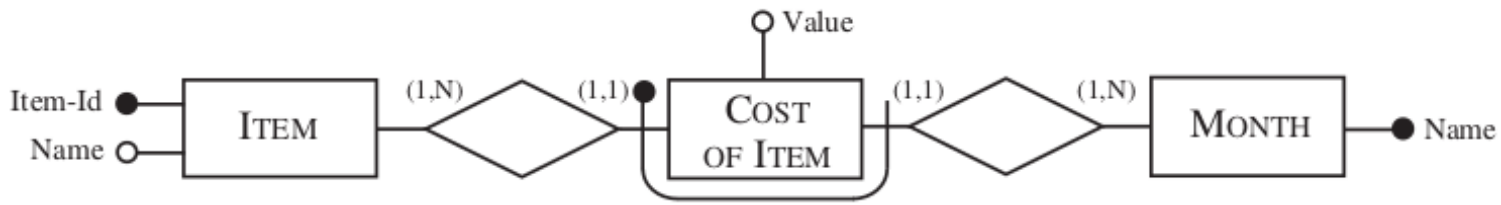


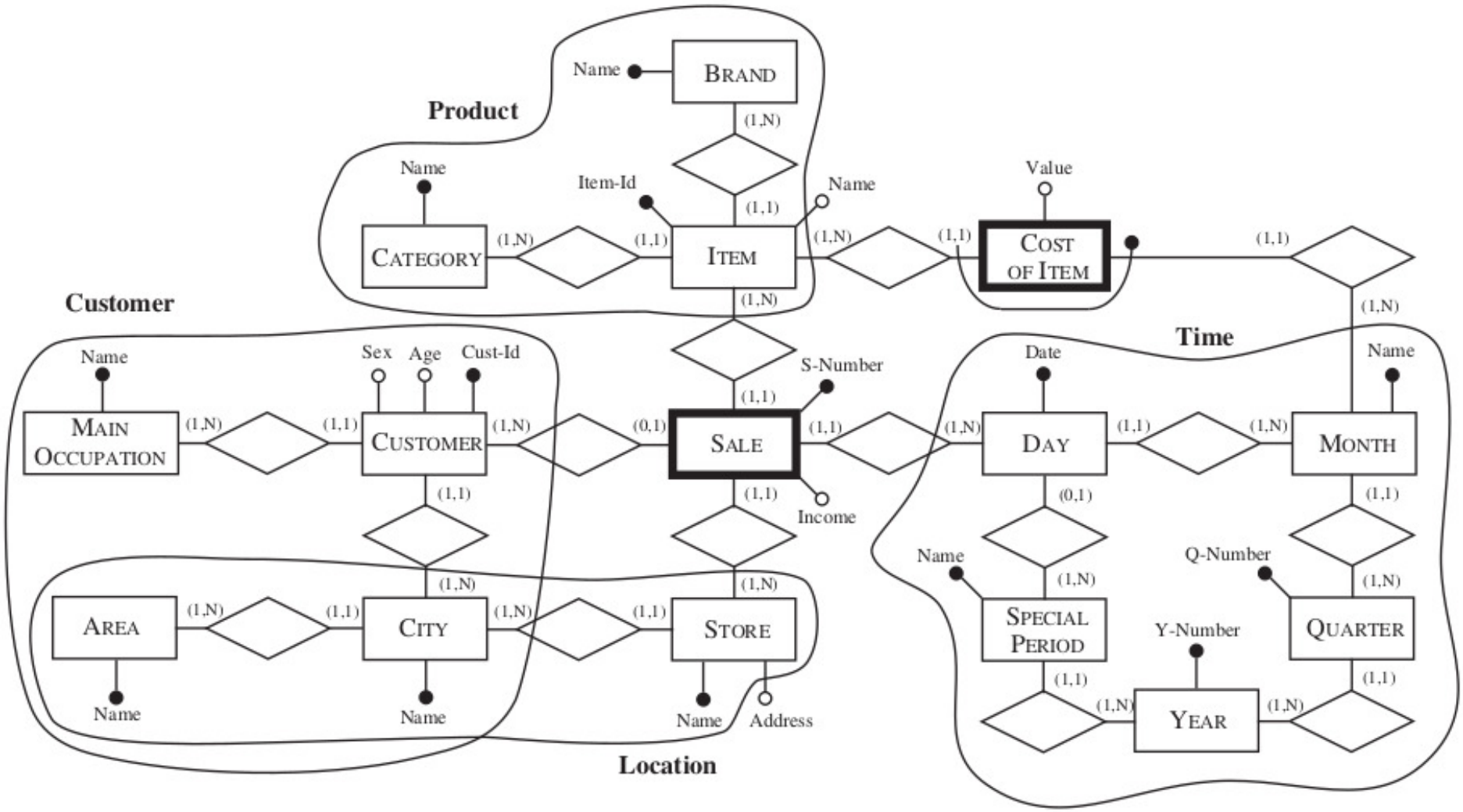
Figure: Reestruturando entidade *COST OF ITEM*

- A cada mês seria obtido os dados do custo a partir da fonte, e seria criado uma nova instância da entidade *COST OF ITEM*.

2. Reestruturando o esquema E-R

- *Refinando os níveis de cada dimensão.*
  (para cada dimensão precisamos filtrar o que é de interesse e o que não é, para analise. Como exemplo de um atributo que não é de interesse citamos o número de telefone de uma loja)
- Para atingir este objetivo, transformamos relações n:m adicionado novas entidades de modo a representar novos níveis de interesse.
- Veja que podemos construir agregações (de acordo com a entidade *CUSTOMER*) por *age*, *sex*, e *cidade*.
- Agregar por ocupação não é possível devido a cardinalidade.
  (criamos uma nova entidade, *MAIN OCCUPATION*, deste modo a cardinalidade é transformada de n:m para 1:n)

E-R depois de reestruturado



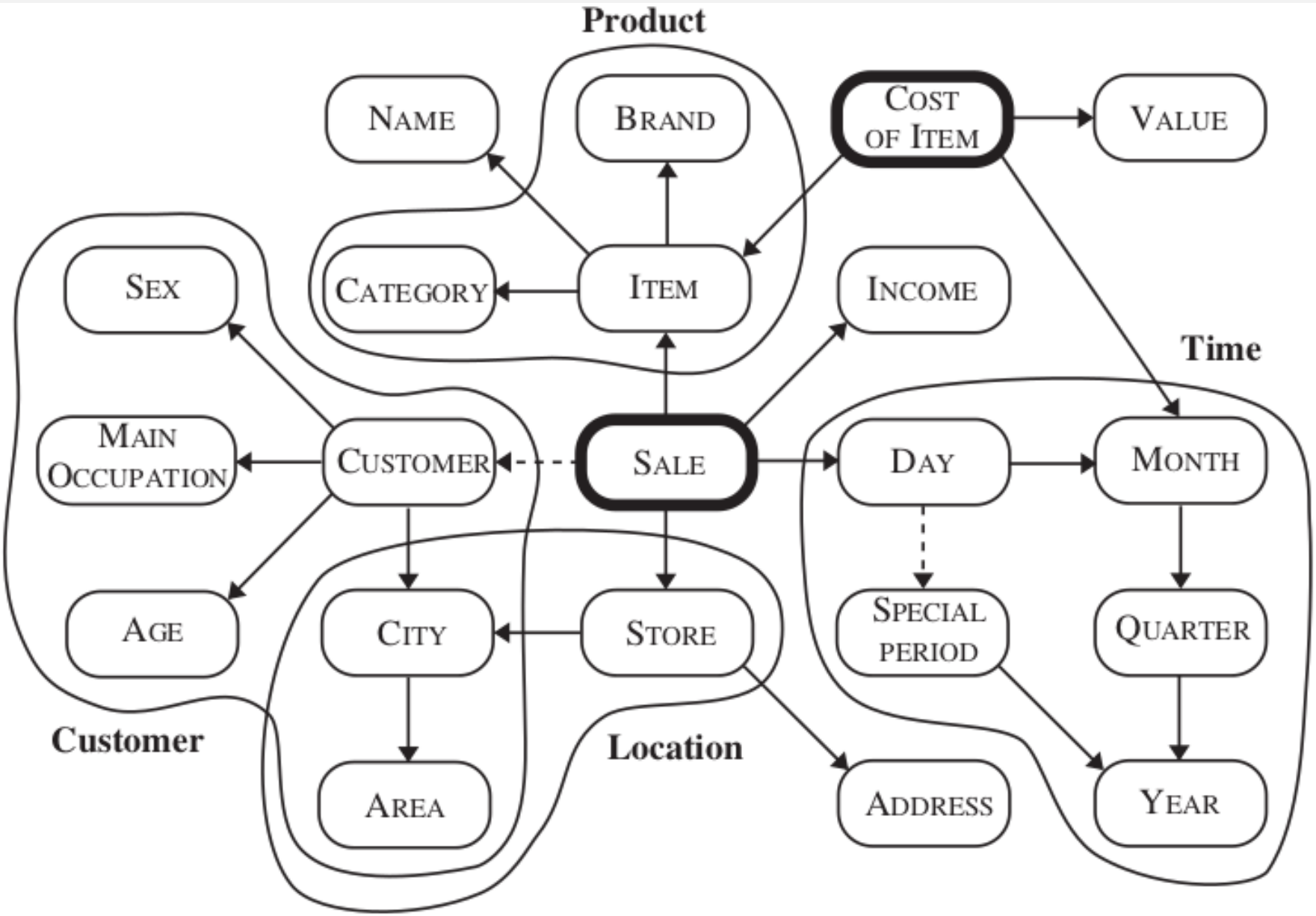
Passo 3

A partir da E-R construir o grafo.

3. A partir da E-R construir o grafo

- A partir da E-R construiremos um grafo dimensional.
- Um grafo dimensional representa fatos e dimensões do esquema E-R reestruturado.

Grafo obtido a partir da E-R reestruturada

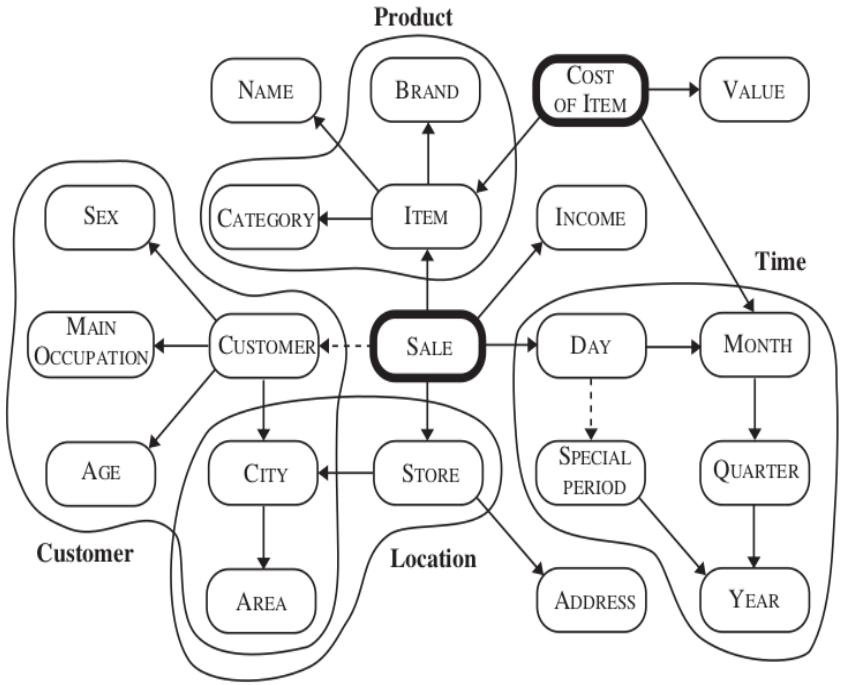


3. O que cada componente do grafo significa?

- Cada vértice do grafo representa uma entidade ou atributo.
- Se um vértice corresponde a uma entidade, então representa o domínio da chave da entidade.
- Se um vértice corresponde a um atributo, este representa o domínio do atributo.
- Um arco entre dois vértices representa uma função entre os correspondentes domínios.
- O arco pontilhado é uma função parcial.

3. O que cada componente do grafo significa?

- O vértice *ITEM* representa o domínio do atributo *Item-id*.
- De mesmo modo *MONTH* representa o domínio do atributo *Name*.
- Note que as dimensões são sub-grafos do grafo dimensional.
- Vértices dentro de uma dimensão, são níveis.



- Vértices descritivos são aqueles que estão fora de uma dimensão e um arco com calda em um vértice de nível e cabeça no vértice em questão.
- Vértices para mensurar um fato é um vértice cuja calda do arco esta em um vértice fato e cabeça no vértice em questão.
- Vértices com margem em negrito denotam fatos (este originou-se das entidades fato).

Passo 4

Traduzir para o modelo MultiDimensional.

Traduzir para o modelo MultiDimensional

- Para este objetivo precisamos definir uma função Θ que mapeie para o modelo $\mathcal{MD}$, possivelmente envolvendo agregações.
- Este mapeamento pode ser o número de instâncias na tabela fato ou uma expressão sobre o que deseja-se mensurar sobre a tabela fato.
- Uma instância da tabela fato pode ser construída da seguinte forma: para cada tupla t, temos um conjunto Φ_t de instância de uma tabela fato. (por ex. dado um item em específico e um dia, temos um conjunto de tuplas associado da venda deste dia deste item.)
- Para se obter o que deseja mensurar na tabela fato aplicamos Θ sobre Φ_t.

Traduzir para o modelo MultiDimensional

Já identificamos três medidas que gostaríamos de analisar, e suas respectivas tabelas fato podem ser representadas da seguinte forma:

- Número de itens vendidos
  $SALE[period : day, product : item, location : store] : numeric$ mapeia count($SALE$)
- Receita (revenue)
  $REVENUE[period : day, product : item, location : store] : numeric$ mapeia sum($Income(SALE)$)
- Custo de itens
  $CostOfItem[period : month, product : item] : numeric$ mapeia Value(CostOfItem)

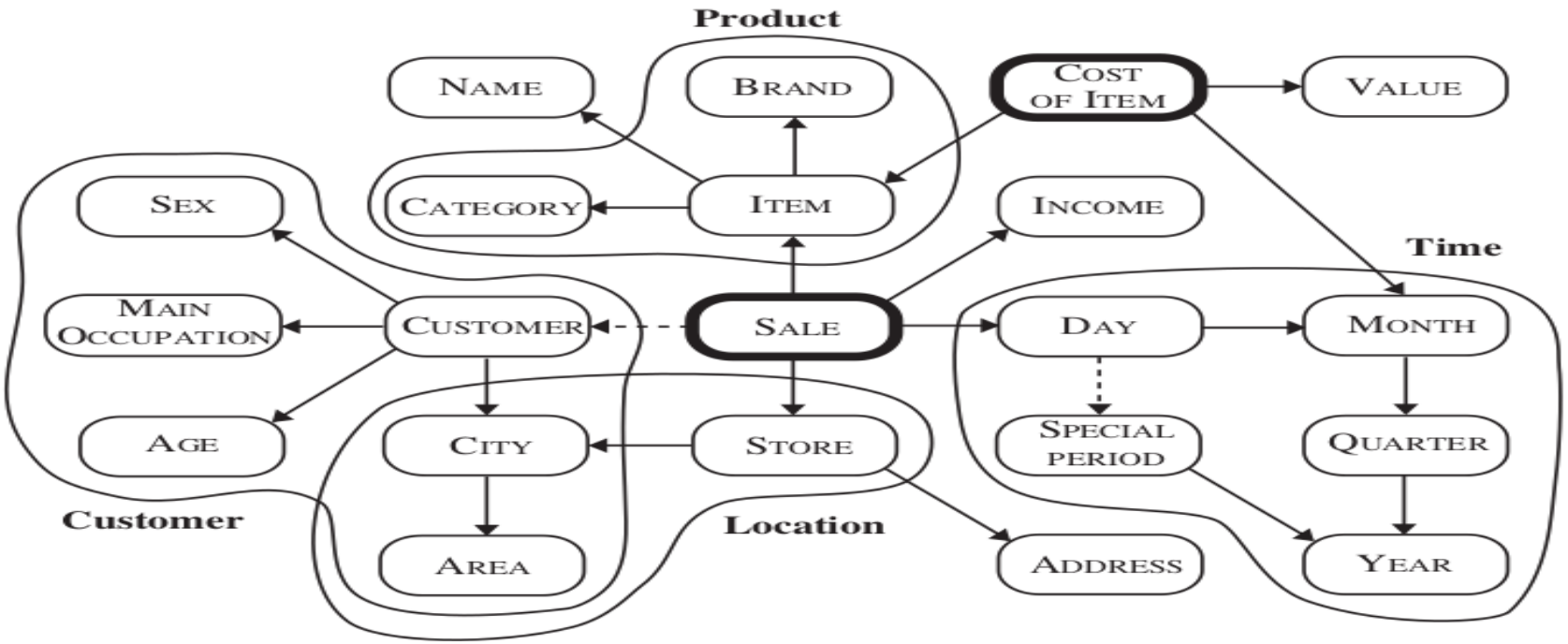
Conclusão

Abordamos neste artigo um paradigma conceitual de como construir uma base de dados MultiDimensional. Conforme dados de alguns estudos [Sapia et al., 1999][Chaudhuri e Dayal, 1997] os modelos OLAPs apresentam ótima performance em consultas, podendo chegar até 0.1% do tempo que requer para responder a mesma consulta em um modelo OLTP, contudo ainda sofre em performance em carregamento dos dados para o modelo OLAP.

Produtos que implementam OLAP

Questões

1. Cite os passos para reestruturar o esquema entidade relacionamento, e descreva como ocorre apenas um dos passos.
2. Observe o grafo dimensional e cite apenas um conjunto de entidades que respeitam a propriedade $l_1 \preceq l_2 \preceq l_3$.



aelias@c3sl.ufpr.br


Cabibbo, L. e Torlone, R. (1998). A logical approach to multidimensional databases.
Em *International Conference on Extending Database Technology*, páginas 183–197. Springer.


Chaudhuri, S. e Dayal, U. (1997).
An overview of data warehousing and olap technology.
ACM Sigmod record, 26(1):65–74.


Kleppmann, M. (2014).
Designing Data-Intensive Applications.
O'Reilly Media.


Sapia, C., Blaschka, M., Höfling, G. e Dinter, B. (1999).
Extending the e/r model for the multidimensional paradigm.
Em *Advances in Database Technologies*, páginas 105–116. Springer.