VisualTPCH: Uma Ferramenta para a Geração de Dados Sintéticos para Data Warehouse

Gustavo Ramos Domingues[1], Cristina Dutra de Aguiar Ciferri[1], Ricardo Rodrigues Ciferri[2]

[1]Departamento de Ciências de Computação – Universidade de São Paulo
Caixa Postal 668 – 13.560-970 – São Carlos – SP – Brasil

[2]Departamento de Computação – Universidade Federal de São Carlos
Caixa Postal 676 – 13.565-905 – São Carlos – SP – Brasil

gustavo.domingues@gmail.com, cdac@icmc.usp.br, ricardo@dc.ufscar.br

Abstract. *This paper introduces the VisualTPCH tool, which offers a graphic interface that assists the generation of synthetic data for data warehouses using as a basis the TPC-H benchmark. The tool allows the manipulation of the TPC-H schema, the generation of aggregation levels and the storage of the obtained data into different database management systems. VisualTPCH can be used to generate synthetic data aiming at different scenarios of performance tests.*

Resumo. *Este artigo apresenta a ferramenta VisualTPCH, a qual oferece uma interface gráfica que facilita a geração de dados sintéticos para data warehouses com base no benchmark TPC-H. A ferramenta permite a manipulação do esquema do TPC-H, a geração de níveis de agregação e o armazenamento dos dados gerados em diferentes sistemas gerenciadores de banco de dados. Desta forma, VisualTPCH pode ser usada para gerar dados sintéticos voltados para diferentes cenários de testes de desempenho.*

1. Introdução

Um *data warehousing* oferece suporte à tomada de decisão, provendo eficiência e flexibilidade na obtenção de informações estratégicas e garantindo a qualidade dessas informações. Ele consolida dados provenientes de fontes autônomas, heterogêneas e distribuídas em um banco de dados, o *data warehouse* (DW). Esse banco de dados é projetado de forma a refletir as requisições multidimensionais dos usuários, sendo seus dados orientados a assunto, integrados e históricos, além de comumente organizados em níveis de agregação [Chaudhuri and Dayal 1997]. O projeto do DW visa otimizar o desempenho no processamento de consultas OLAP (*On-line Analytical Processing*).

Como existe uma demanda para que consultas OLAP tenham, cada vez mais, um melhor desempenho, muitas pesquisas têm surgido visando melhorar o tempo de resposta dessas consultas. Para se testar a eficiência do processamento de consultas OLAP, existe a necessidade de se usar um volume de dados significativo nos testes de desempenho, o qual nem sempre está disponível ou é facilmente obtido. Assim, são necessárias ferramentas específicas para a geração automática de dados que podem ser usados para testes em *data warehousing*.

Os autores agradecem o apoio financeiro das seguintes agências de fomento à pesquisa do Brasil: CNPq, FAPESP, CAPES e FINEP.

O TPC (*Transaction Processing Performance Council*) é uma organização que define *benchmarks* de bancos de dados amplamente aceitos e usados na avaliação de sistemas de banco de dados. Vários *benchmarks* são definidos, dentre eles o TPC-H, o qual é voltado à tomada de decisão [Poess and Floyd 2000; TPC 2008]. Embora o TPC-H especifique como dados sintéticos para DW devem ser gerados, ele tem quatro limitações principais: (i) o aplicativo de geração de dados é de linha de comando; (ii) os dados são gerados em arquivos texto de difícil manipulação; (iii) o esquema que define os dados do DW é fixo e normalizado; e (iv) os dados gerados não são organizados em níveis de agregação. Tais limitações dificultam o uso do TPC-H e motivam o desenvolvimento de uma ferramenta flexível que supra essas limitações.

Este artigo apresenta a ferramenta VisualTPCH, a qual gera dados sintéticos para DW com base no *benchmark* TPC-H. Os diferenciais da ferramenta são:

- Oferecimento de uma interface gráfica para a geração de dados sintéticos;
- Manipulação do esquema original do TPC-H;
- Geração dos diferentes níveis de agregação de um DW;
- Geração e carga dos dados sintéticos diretamente em tabelas relacionais de diferentes sistemas gerenciadores de banco de dados (SGBD); e
- Portabilidade.

No melhor do nosso conhecimento, não existe outra ferramenta que ofereça as mesmas funcionalidades que VisualTPCH.

Este artigo está organizado da seguinte maneira. Na Seção 2 é resumida a fundamentação teórica. A ferramenta VisualTPCH é detalhada na Seção 3, em termos de objetivos, funcionalidades e interfaces. Exemplos de configurações de teste são ilustrados na Seção 4. O artigo é concluído na Seção 5.

2. Esquema do Benchmark TPC-H e Níveis de Agregação

O TPC-H define uma aplicação de *data warehousing* que armazena dados históricos relativos a pedidos e vendas de uma companhia [Poess and Floyd 2000; TPC 2008]. A modelagem multidimensional de seus dados é guiada pelo esquema ilustrado na Figura 1. *lineitem* e *partsupp* são tabelas de fatos, desde que armazenam as medidas numéricas (i.e., assuntos) de interesse. Já *customer*, *orders*, *nation*, *supplier*, *region* e *part* são tabelas de dimensão, as quais contêm as características (i.e., atributos) das dimensões. Diferentemente de um esquema estrela convencional [Kimball and Ross 2002], as tabelas de dimensão do TPC-H são normalizadas. Assim, atributos de *supplier* encontram-se divididos nas tabelas *supplier*, *nation* e *region*, enquanto que atributos de *orders* encontram-se divididos nas tabelas *customer*, *nation* e *region*.

O esquema do TPC-H ilustra somente o nível de agregação inferior de um DW. Em geral, os dados do DW são organizados na forma de um cubo [Harinarayan *et al*. 1996], ou seja, em diferentes níveis de agregação, desde um nível inferior que possui dados detalhados, até um nível superior que possui dados muito resumidos. Podem existir vários níveis intermediários, representando graus de agregação crescentes. A Figura 3 mostra graficamente esta organização do DW, usando um grafo de derivação para representar as agregações geradas a partir de *lineitem*. Nessa figura, o vértice *ops* é

o nível inferior e representa uma agregação de *orders* (*o*), *part* (*p*) e *supplier* (*s*). Já *s* é um vértice de um nível intermediário e representa uma agregação de *supplier* (*s*). O grafo da Figura 3 é completo, desde que contém todas as agregações que podem ser geradas a partir do vértice *ops*. Em contrapartida, um grafo é incompleto quando não possui todas as agregações que podem ser geradas a partir do nível inferior.

3. A Ferramenta VisualTPCH

VisualTPCH gera dados sintéticos para DW organizados em diferentes níveis de agregação, usando como base o *benchmark* TPC-H. Essa ferramenta oferece uma interface gráfica para:

- Modificar o esquema original do TPC-H para um *novo esquema*, alterando-se tabelas e atributos desse esquema. Por exemplo, pode-se gerar um esquema estrela normalizado a partir do esquema original (etiqueta **Esquema**);
- Gerar níveis de agregação para o *novo esquema*, permitindo a criação de um grafo de derivação completo ou incompleto (etiqueta **Níveis de Agregação**);
- Gerar automaticamente dados sintéticos do TPC-H para cada um dos vértices do grafo de derivação (opção **Consolidar**); e
- Armazenar os dados gerados em uma ou mais tabelas relacionais dos SGBD Oracle®, PostgreSQL® ou DB2® (opção **Consolidar**).

3.1. Etiqueta Esquema

Para a realização de testes de desempenho em diferentes cenários, pode ser necessário mudar o esquema original do TPC-H para um *novo esquema*. A etiqueta **Esquema** oferece funcionalidades voltadas a este fim. A Figura 1 ilustra a interface da etiqueta, a qual exibe inicialmente o esquema original do TPC-H. As tabelas de fatos são exibidas em linhas tracejadas porque estão inativas. O usuário pode, então, torná-las ativa.

Para cada tabela, o usuário pode: (i) selecionar quais de seus atributos irão permanecer no *novo esquema* (Figura 2a); (ii) juntá-la com outra tabela relacionada, originando apenas uma única tabela no *novo esquema* que contém os atributos das duas tabelas originais (Figura 2b); (iii) renomeá-la, se ela for uma tabela de dimensão (Figura 2c); e (iv) excluí-la, de forma que ela não exista mais no *novo esquema* (Figura 2d).

3.2. Etiqueta Níveis de Agregação

A etiqueta **Níveis de Agregação** permite que o usuário construa um grafo de derivação a partir do nível inferior do DW representado pelo *novo esquema*. A Figura 3 ilustra a interface dessa etiqueta, considerando que a tabela de fatos *lineitem* está ativa. A interface é dividida em três partes. Na parte **Dimensões**, são exibidas as dimensões que se relacionam com *lineitem* (i.e., *orders*, *part* e *supplier*), enquanto que na parte **Abreviações** são listadas as abreviações usadas para representar essas dimensões (respectivamente *o*, *p*, e *s*) no grafo de derivação. Já a parte do lado direito da interface ilustra o grafo de derivação gerado a partir de *lineitem*. Este grafo somente é desenhado quando o usuário pressiona **Desenhar Grafo**. **Definição Textual** mostra, na forma de texto, quais os vértices que compõem o grafo.

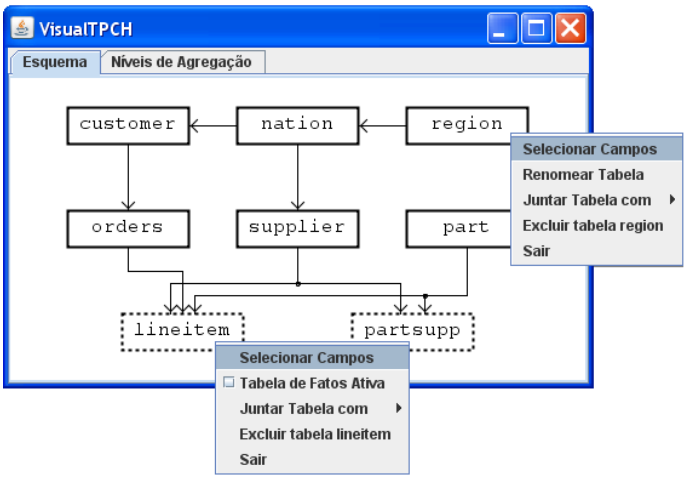

Figura 1. Interface da etiqueta esquema.

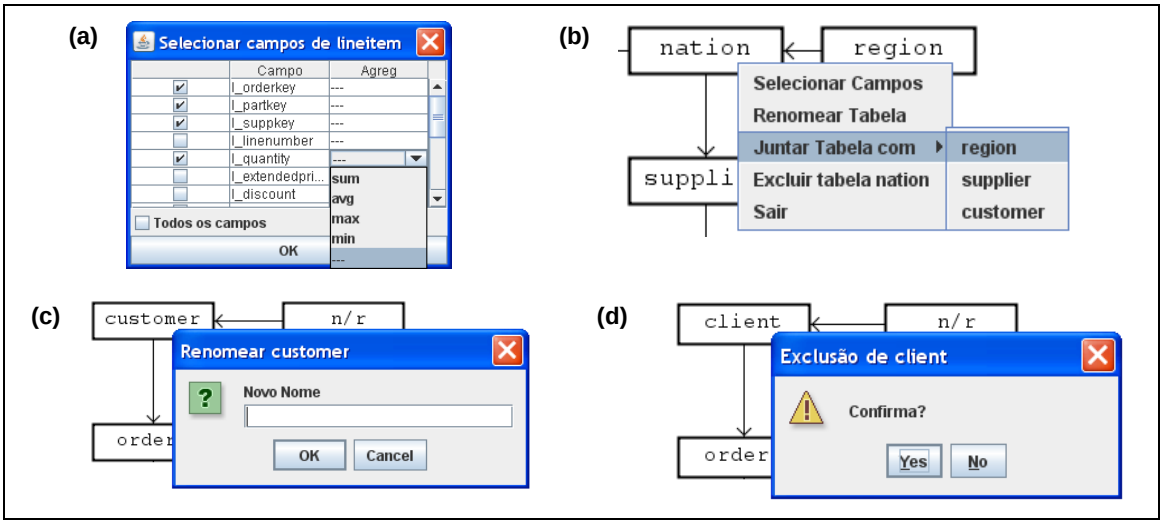

Figura 2. Funcionalidades da interface da etiqueta esquema.

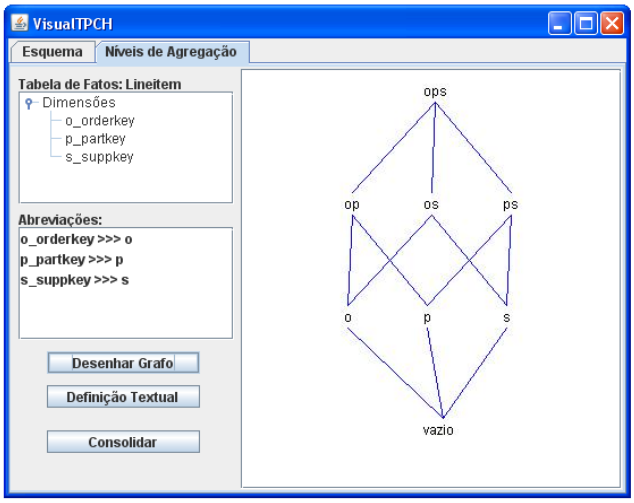

Figura 3. Interface da etiqueta níveis de agregação.

Na parte **Dimensões**, o usuário pode reposicionar as dimensões, usando recursos de arrastar e apontar, para sinalizar uma hierarquia de atributos. Na Figura 4a, a dimensão *o_orderkey* deixou de ser uma dimensão para ser um atributo da dimensão *s_suppkey*, sendo representada um pouco mais à direita que *s_suppkey*. Já na parte **Abreviações** (Figura 4b), o usuário pode definir uma nova sigla para uma abreviação. Finalmente, o usuário pode solicitar a exibição dos ancestrais diretos, dos ancestrais, dos descendentes diretos e dos descendentes de cada vértice do grafo de derivação. O usuário também pode excluir vértices, gerando assim um grafo incompleto (Figura 4c).

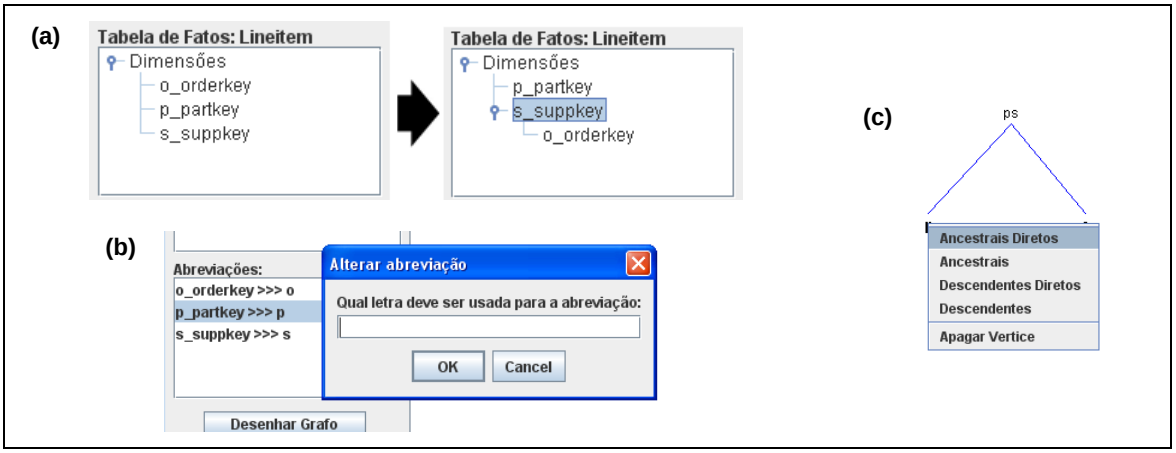

Figura 4. Funcionalidades da interface da etiqueta níveis de agregação.

3.3. Opção Consolidar

VisualTPCH realiza o armazenamento dos dados sintéticos em tabelas relacionais de diferentes SGBD, especificamente Oracle®, PostgreSQL® e DB2®. A opção **Consolidar** consiste em definir o SGBD a ser usado e estabelecer uma conexão com esse SGBD (Figura 5a). O usuário pode verificar se a conexão foi estabelecida pressionando **Testar Conectividade**. Uma conexão com o SGBD Oracle® é ilustrada na Figura 5b.

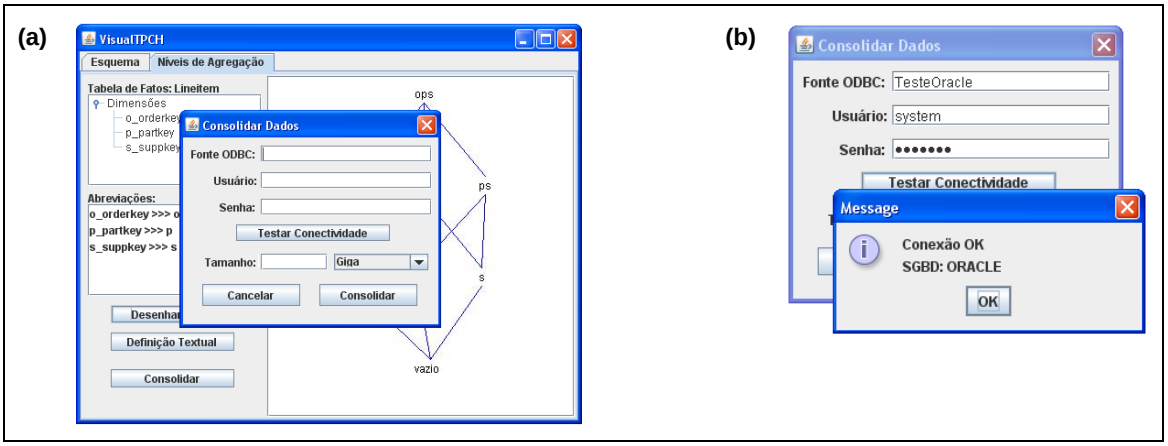

Figura 5. Funcionalidades de conexão com o SGBD.

Ao confirmar a consolidação, VisualTPCH gera e executa todos os comandos SQL (*Structured Query Language*) para a criação das tabelas e a carga dos dados. O volume de dados do nível inferior do DW é definido pelo valor do campo **Tamanho**, o

qual é informado pelo usuário. Para gerar os dados desse nível, VisualTPCH usa o aplicativo de geração de dados de linha de comando do TPC-H, porém respeitando as tabelas e os atributos selecionados pelo usuário. Os dados dos demais níveis de agregação são gerados automaticamente pela VisualTPCH com base nos dados do nível inferior.

4. Configurações de Teste

VisualTPCH visa a geração de dados sintéticos para testes de desempenho em diferentes cenários. Para exemplificar o uso da ferramenta, considere dois cenários. O **cenário 1** é formado por três tabelas de dimensão (i.e., *orders*, *supplier* e *part*) e por uma tabela de fatos (i.e., *lineitem*), além de considerar um grafo completo (Figura 3). Já o **cenário 2** consiste da tabela de fatos *partsupp* e das tabelas de dimensão *supplier* e *part*, além de considerar um grafo incompleto com os vértices *ps*, *s* e *vazio*. A Tabela 1 mostra as tabelas geradas e o número de tuplas de cada uma dessas tabelas. Para cada vértice do grafo, VisualTPCH gera uma tabela relacional de mesmo nome. Os valores constantes na Tabela 1 foram gerados para o valor do campo **Tamanho** igual a 10 MB.

Tabela 1. Número de tuplas geradas pela VisualTPCH para diferentes cenários.

Cenário 1

Tabela	Tuplas
ops	60.162
ps	7.996
os	59.036
op	60.113
s	100
p	2.000
o	15.000
vazio	1

Cenário 2

Tabela	Tuplas
os	8.000
s	100
vazio	1

5. Conclusões

Este artigo apresentou a ferramenta VisualTPCH, a qual oferece uma interface gráfica para a geração de dados sintéticos para DW. Sua implementação foi realizada usando-se a linguagem Java e o IDE Eclipse versão 3.2. A ferramenta é portátil, funcionando tanto em Linux quanto em Windows. Outros diferenciais da VisualTPCH são: a alteração do esquema do TPC-H para a geração de um novo esquema, a geração de grafos de derivação completos e incompletos a partir desse novo esquema; e a automatização da geração e da carga dos dados sintéticos desse esquema em tabelas relacionais de diferentes SGBD. Os comandos SQL gerados pela ferramenta seguem a especificação padrão e, portanto, VisualTPCH está sendo estendida para suporte para outros SGBD.

Referências

Chaudhuri, S. and Dayal, U. (1997), "An Overview of Data Warehousing and OLAP Technology", In *SIGMOD Record*, v. 26, n. 1, p. 65-74.

Harinarayan, V. *et al*. (1996) "Implementing Data Cubes Efficiently", In: *Proc. SIGMOD*, p. 205-216.

Kimball, R. and Ross, M. (2002), "The Data Warehouse Toolkit", Wiley, 2nd edition.

Poess, M. and Floyd, C. (2000), "New TPC Benchmarks for Decision Support and Web Commerce", In *SIGMOD Record*, v. 29, n. 4, p. 64-71.

TPC (2008), "Transaction Processing Performance Council", http://www.tpc.org, Maio.