

ADAPTANDO O BENCHMARK TPC-DS PARA DATA WAREHOUSES GEOGRÁFICOS: A CARGA DE DADOS NO BANCO DE DADOS RELACIONAL

GABRIEL BREGADIOLI LIMONI¹, THIAGO LUÍS LOPES SIQUEIRA²

¹ Graduando em Tecnologia de Análise e Desenvolvimento de Sistemas, Bolsista PIBITI/CNPq/IFSP 2016-2017, IFSP, Câmpus Piracicaba, gabrielblimoni@gmail.com

² Docente, IFSP, Câmpus Piracicaba, prof.thiago@ifsp.edu.br
Área de conhecimento (Tabela CNPq): 1.03.03.03-0 Banco de Dados

Apresentado no
8º Congresso de Inovação, Ciência e Tecnologia do IFSP
06 a 09 de novembro de 2017 - Cubatão-SP, Brasil

RESUMO: O benchmark TPC-DS foi criado para aplicações de suporte à decisão tais como o data warehouse sob o banco de dados relacional. O TPC-DS, porém, não provê suporte a dados e consultas espaciais. Este artigo aborda a transformação do esquema do TPC-DS num esquema relacional de data warehouse geográfico (DWG), e detalha a criação e a avaliação de uma ferramenta que realiza a carga de dados naquele esquema. Provê-se uma alternativa aos benchmarks já existentes para DWG.

PALAVRAS-CHAVE: TPC-DS, data warehouse geográfico, modelo relacional, benchmark.

ADAPTING TPC-DS BENCHMARK FOR GEOGRAPHIC DATA WAREHOUSES: LOADING DATA INTO RELATIONAL SCHEMATA

ABSTRACT: TPC-DS is a benchmark for decision support originally created for relational databases. However, it does not support spatial data either queries. This paper describes how TPC-DS' schema was transformed into a geographic data warehouse schema and details the development and evaluation of a tool which loads data into that schema. We provide an alternative to existing benchmarks for geographic data warehouses.

KEYWORDS: TPC-DS, geographic data warehouse, relational model, benchmark.

INTRODUÇÃO

Um data warehouse (DW) é uma base de dados histórica, integrada, orientada por assunto, não volátil, variante no tempo e muito volumosa (VAISMAN; ZIMANYI, 2014). Ele armazena os dados relevantes das instituições para auxiliar na tomada de decisão. Um exemplo de consulta sobre um DW é “calcular as médias da quantidade de vendas, do preço de venda, do custo da compra completa, e o total do custo da compra em lojas, diferentes perfis de consumidores, combinações de estados da nação e lucros obtidos num determinado ano”. Já o data warehouse geográfico (DWG) é uma extensão do DW que provê suporte a dados e consultas espaciais (VAISMAN; ZIMANYI, 2014). Dados espaciais são comumente representados por geometrias, tais como pontos, linhas e polígonos. Na consulta anterior, os estados poderiam ser representados por polígonos, e o usuário poderia selecionar no mapa os estados a serem considerados.

Benchmarking é uma técnica de padronização de operações entre sistemas diferentes que visa determinar o melhor sistema para um domínio (SCABORA, 2016). Existem benchmarks para DWG (SIQUEIRA et al, 2010; NASCIMENTO et al, 2011) que se baseiam no SSB (BARATA; BERNARDINO; FURTADO; 2015). Eles foram desenvolvidos para bancos de dados relacionais, logo armazenam os dados em tabelas e as suas consultas seguem a linguagem SQL.

O benchmark TPC-DS foi criado para aplicações de suporte à decisão como DW conforme o modelo relacional (BARATA; BERNARDINO; FURTADO; 2015). Não foi identificado na literatura um benchmark para DWG baseado no TPC-DS. Este trabalho objetivou transformar o esquema do TPC-DS num esquema de DWG e criar uma ferramenta para realizar a carga de dados no esquema obtido, a fim de prover uma alternativa aos benchmarks existentes para DWG.

MATERIAL E MÉTODOS

Para a geração dos dados convencionais (numéricos e alfanuméricos) foi utilizado o software do TPC-DS. Já os dados espaciais do tipo polígono referentes a 56 estados e 3.233 condados foram obtidos a partir do Censo Estadunidense (TIGER). O desenvolvimento da ferramenta usou o paradigma de desenvolvimento de software orientado a objetos e a linguagem de programação Java 8.

Foram conduzidos testes de desempenho experimentais. O ambiente de desenvolvimento e de testes utilizado foi o Serviço de Infraestrutura para Computação Científica (SICC-IFSP), que consiste num cluster de computadores composto por cinco servidores Dell PowerEdge R720, com processadores Intel Xeon E5-2640 @ 2.50 GHz. No cluster foi criada e utilizada uma máquina virtual com as seguintes características: 16 núcleos de processamento, 16GB de memória RAM (DDR3 de barramento mesclado 1.066 MHz e 1.333 MHz), 1 TB de armazenamento em disco (SAS 15k). O sistema gerenciador de banco de dados (SGBD) relacional utilizado foi o PostgreSQL 9.5.7 junto da sua extensão para dados espaciais denominada PostGIS 2.2 (OBE, R. O.; HSU, L. S., 2015).

Para a cronometragem dos tempos de cada operação da ferramenta, foram coletados os horários inicial e final e calculada a diferença entre eles para determinar o tempo total da execução do processo. Já para a cronometragem das consultas no PostgreSQL/PostGIS foi utilizado o comando *\timing* que mostra, junto com o retorno da consulta SQL, o seu tempo de execução.

RESULTADOS E DISCUSSÃO

Um esquema de DWG foi produzido a partir do esquema do benchmark TPC-DS. Foram necessárias as seguintes adaptações para viabilizar o armazenamento de dados espaciais: adição de duas tabelas com dados espaciais referentes a estado (State) e condado (County); adição de colunas na tabela Customer_Address e Store referenciando as tabelas State e County.

Foi especificada e desenvolvida uma ferramenta para realizar a carga de dados no DWG. A carga de dados possui 8 etapas conforme segue.: (i) obtenção de parâmetros a partir de um arquivo de configuração; (ii) geração de arquivos no formato CSV contendo dados convencionais; (iii) criação das tabelas convencionais do esquema; (iv) importação dos arquivos gerados e carregamento dos seus dados nas tabelas; (v) adição de chaves primárias e estrangeiras; (vi) importação dos dados espaciais para as tabelas State e County; (vii) ajuste para esquema de DWG nas tabelas Customer_Address e Store; e (viii) criação de índices espaciais sobre as colunas das tabelas State e County que contêm os dados espaciais (denominadas state_geo e county_geo). A ferramenta foi utilizada para carregar os dados do DWG. As tabelas e índices deste ocuparam 128GB. Cada etapa de execução da ferramenta teve seu tempo de execução cronometrado. A Tabela 1 exibe os tempos de cada uma em minutos. As etapas v e iv foram as duas que tomaram mais tempo e demandaram maior processamento.

TABELA 1. Tempo de execução em minutos das etapas da carga de dados.

Etapas	Descrição	Tempo
i	Obtendo as configurações iniciais	~0
ii	Geração dos arquivos de dados convencionais	31,300
iii	Criação das tabelas	0,006
iv	Carregamento dos dados convencionais nas tabelas	84,367
v	Adição das restrições nas tabelas	141,333
vi	Importação dos dados espaciais	0,369
vii	Definição do esquema de DWG	9,383
viii	Criação de índices espaciais nas tabelas	0,310
Total		267,069

A consulta exemplificada na Introdução, pertencente ao TPC-DS, foi estendida, traduzida para SQL e executada. Nela, o predicado espacial intersecta foi introduzido e avaliado em relação a estados e condados. Na Figura 1 são exemplificados os dados espaciais, os operadores de intersecção e as janelas de consulta para condado (J_C) e para estado (J_E). Primeiramente, executou-se a consulta sobre os condados e, posteriormente sobre os estados. Essa execução consecutiva sumariza progressivamente os dados. A execução das consultas demorou 10,46 e 129,56 minutos para condado e estado, respectivamente, totalizando 140,02 minutos.

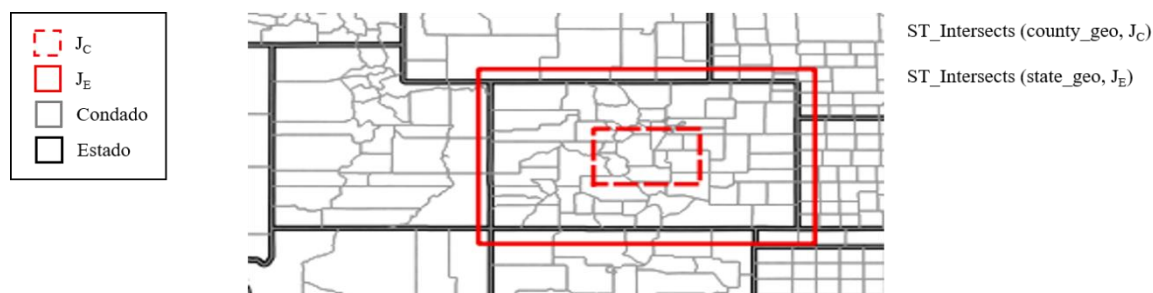


FIGURA 1. Exemplo de intersecções com estado e condado.

CONCLUSÕES

Foi estendido o esquema do benchmark TPC-DS para DWG e desenvolvida uma ferramenta que carrega os dados do esquema promovido para o SGBD relacional PostgreSQL e sua extensão espacial denominada PostGIS. A ferramenta omite do usuário tanto as operações que transformam o benchmark TPC-DS num esquema de DWG quanto as operações de carregamento dos dados. As consultas SQL do benchmark TPC-DS foram adaptadas para embutir predicados espaciais.

O tempo total de execução foi de 267,069 minutos, isto é, aproximadamente 4 horas, sendo que as operações mais demoradas foram a adição de restrições e o carregamento dos dados convencionais, respectivamente, os quais demandam um maior processamento, seja pela verificação de chaves primárias e estrangeiras ou pelo tamanho dos arquivos importados.

O principal trabalho futuro consiste na inclusão de dados do tipo linha e ponto no DWG e sua avaliação por meio de testes de desempenho experimentais. Destaca-se também o aprimoramento dos processos em que foram encontrados gargalos.

AGRADECIMENTOS

¹Bolsista PIBITI/CNPq/IFSP 2016/2017. ^{1,2}Agradecem à CIS/DIR/PRD pelo suporte técnico.

REFERÊNCIAS

- BARATA, M., BERNARDINO, J., FURTADO, P. (2015). An overview of decision support benchmarks: TPC-DS, TPC-H and SSB. In *New Contributions in Information Systems and Technologies* (pp. 619-628). Springer, Cham.
- NASCIMENTO, S. M et al. The Spatial Star Schema Benchmark. 2011. In: XII Simpósio Brasileiro de Geoinformática (GEOINFO), 2011. p. 73-84.
- OBE, R. O.; HSU, L. S. *PostGIS in Action*. 2. ed. Shelter Island, NY, USA: Manning Publications Co., 2015. ISBN 9781617291395.
- SCABORA, L. C. Avaliação do Star Schema Benchmark aplicado a bancos de dados NoSQL distribuídos e orientados a colunas. 2016. 111 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos. 2016.
- SIQUEIRA, T. L. L. et al. Benchmarking spatial data warehouses. *Data Warehousing and Knowledge Discovery*, p. 40-51, 2010.
- VAISMAN, A. A.; ZIMÁNYI, E. *Data Warehouse Systems Design and Implementation*. Springer, 2014. ISBN 978-3-642-54655-6.