

Benchmark TPC

Uma introdução completa aos benchmarks TPC e sua importância para avaliação de desempenho em sistemas de banco de dados

Eduardo Ogasawara

eduardo.ogasawara@cefet-rj.br

<https://eic.cefet-rj.br/~eogasawara>

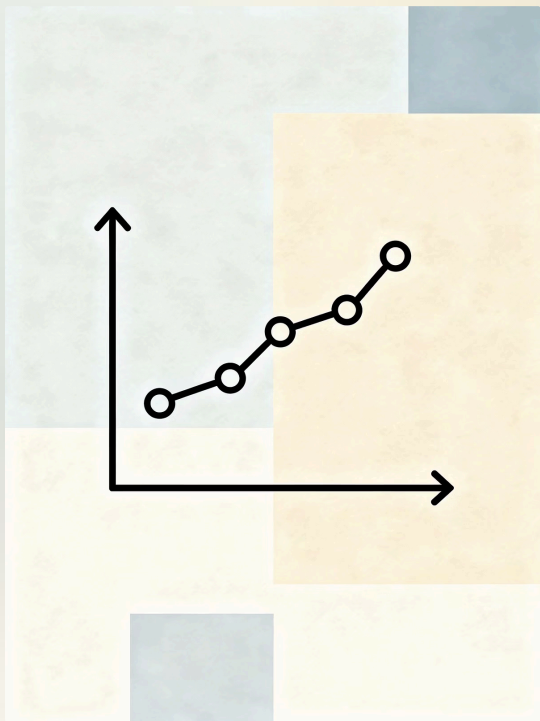
O que são Benchmarks?

Definição

Um benchmark é uma carga padronizada utilizada no processo de avaliação e medição de desempenho de sistemas computacionais. Ele permite comparações objetivas entre diferentes configurações de hardware e software.

Atributos Essenciais

<div>Relevante</div> <div>Representa cargas de trabalho reais do mundo empresarial</div>	<div>Entendível</div> <div>Facilmente compreensível por desenvolvedores e gestores</div>	<div>Métricas bem definidas</div> <div>Resultados claros e mensuráveis</div>
<div>Escalável</div> <div>Adaptável a diferentes tamanhos de sistemas</div>	<div>Alta cobertura</div> <div>Testa diversos aspectos do sistema</div>	<div>Bem aceito</div> <div>Reconhecido pela indústria</div>



Benefícios dos Benchmarks TPC



Padronização de Avaliação de Desempenho

Define critérios e métricas padronizadas para avaliação de desempenho de sistemas de bancos de dados



Direcionamento de Projeto e Otimização

Fornecer metas quantitativas que orientam decisões de projeto, otimização e avaliação de sistemas de bancos de dados



Comparação Controlada de Desempenho

Permite a comparação objetiva de desempenho entre sistemas sob cargas padronizadas e regras de execução bem definidas



Evolução de Arquiteturas e Mecanismos

Estimula a melhoria de arquiteturas, mecanismos de execução e estratégias de otimização a partir de cargas padronizadas

Conselho de Desempenho TPC

O Transaction Processing Performance Council (TPC) é uma **organização sem fins lucrativos** fundada para definir benchmarks de processamento de transações e bancos de dados, além de divulgar dados de desempenho objetivos e verificáveis para a indústria.

Missão Principal

Define regras rigorosas e testes padronizados para avaliação de desempenho de sistemas de bancos de dados

Composição

Composta por representantes de empresas líderes do setor de tecnologia da informação

Objetivo

Provê dados confiáveis para testes de desempenho relevantes em diversos cenários de uso empresarial

Tipos de Benchmarks TPC



TPC-C

Benchmark OLTP (Online Transaction Processing)

- Simula ambientes de processamento transacional
- Focado em operações de alta concorrência
- Ideal para sistemas de comércio e varejo



TPC-H

Benchmark OLAP (Online Analytical Processing)

- Avalia consultas analíticas complexas
- Orientado para tomada de decisões
- Voltado para business intelligence

OLTP

Benchmark TPC-C

O TPC-C é o benchmark padrão da indústria para comparação de desempenho OLTP em diversos ambientes de hardware e software. Ele foi projetado para representar com precisão o ambiente computacional encontrado em aplicações comerciais típicas.



Ambiente Completo

Simula um sistema com grande quantidade de usuários executando transações simultaneamente contra um banco de dados



Transações Variadas

Executa transações concorrentes de complexidade e tipos diferentes, refletindo cenários reais



Base Robusta

Banco de dados composto por nove tabelas inter-relacionadas com milhões de registros



Métrica Principal: O resultado é medido em transações por minuto (tpmC), permitindo comparações objetivas de desempenho entre diferentes sistemas

Transações do TPC-C

O benchmark TPC-C utiliza cinco tipos de transações que simulam atividades típicas de um ambiente de distribuição e comércio atacadista:

New-Order 45% das transações Insere uma nova ordem de compra de um cliente no sistema	Payment 43% das transações Atualiza o saldo do cliente refletindo um pagamento realizado	Delivery 4% das transações Processa e atende lotes de ordens de compra pendentes
Order-Status 4% das transações Recupera o status das ordens de compra mais recentes dos clientes		Stock-Level 4% das transações Monitora e verifica os níveis de estoque de produtos

Modelo de Dados TPC-C

Warehouse (W armazéns)

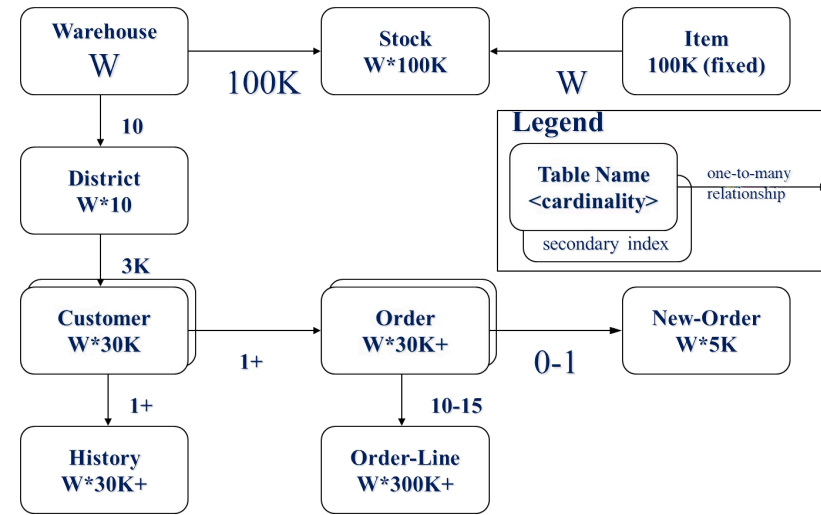
Representa os centros de distribuição

Stock (W × 100K itens)

Estoque por armazém

Item (100K itens fixos)

Catálogo de produtos



O modelo de dados é escalável, permitindo ajustar o tamanho do benchmark através do número de armazéns (W). Cada armazém mantém relacionamentos complexos com clientes, pedidos e itens em estoque.

Processo de Execução TPC-C

Seleção de Transação

O sistema seleciona uma transação do menu baseado nas porcentagens definidas

- New-Order: 45%
- Payment: 43%
- Order-Status: 4%
- Delivery: 4%
- Stock-Level: 4%

Think Time

Simulação do tempo de análise do usuário



Menu Response Time

Medição do tempo de resposta da tela de entrada

Keying Time

Simulação do tempo de digitação do usuário

Transaction RT

Medição do tempo de resposta da transação

📄 **Tempos definidos pela especificação do benchmark TPC-C:** Os tempos de digitação (*keying time*), espera (*think time*) e resposta fazem parte da carga padronizada do benchmark e não refletem parâmetros internos do SGBD.

Teste ACID do TPC-C

O benchmark TPC-C não apenas mede desempenho, mas também verifica rigorosamente que as propriedades ACID fundamentais de bancos de dados transacionais estão sendo respeitadas:



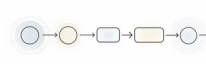
Atomicidade

Confirma que todas as mudanças de uma transação são completamente confirmadas ou totalmente abortadas, sem estados intermediários



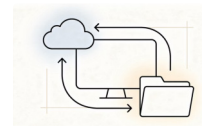
Consistência

Verifica que o banco de dados permanece em um estado válido antes e depois de cada transação



Isolamento

Assegura que transações concorrentes não interferem umas com as outras durante a execução



Durabilidade

Demonstra a capacidade de recuperação do sistema em diversos cenários de falha

O TPC-C testa especificamente a durabilidade através de simulações de faltas de energia, problemas de memória e falhas em mídias de armazenamento

OLAP

Benchmark TPC-H

O TPC-H é um benchmark de suporte à decisão que examina grandes volumes de dados e executa consultas de alta complexidade. Ele foi projetado para avaliar o desempenho de sistemas de banco de dados utilizados em aplicações de **Business Intelligence** e **Data Warehousing**.

Consultas ad-hoc complexas

Orientadas a negócio e representando análises reais

Grande volume de dados

Examina extensas porções da base para insights profundos

Atualizações Concorrentes Controladas

Simula ambientes analíticos onde atualizações ocorrem de forma limitada e controlada durante a execução das consultas

Respostas críticas

Fornece informações essenciais para decisões estratégicas

Características das Consultas TPC-H

01	02	03
Alto grau de complexidade Consultas envolvem múltiplas junções, agregações e subconsultas aninhadas	Diversidade de Estratégias de Acesso Exploram diferentes estratégias de acesso aos dados, incluindo varreduras completas, uso de índices e operações de junção complexas	Natureza ad-hoc Não seguem padrões previsíveis, simulando consultas exploratórias reais
04	05	06
Processamento em Larga Escala Processam porções significativas da base de dados para responder consultas analíticas complexas	Diversidade entre consultas Cada consulta é única e testa diferentes aspectos do sistema	Parâmetros dinâmicos Possuem parâmetros que mudam durante a execução, testando adaptabilidade

Tipos de Consultas OLAP

O TPC-H inclui 22 consultas que cobrem cenários analíticos típicos de negócios. Estas consultas são agrupadas em cinco categorias principais:



Preço e Promoção

Análises de estratégias de precificação, impacto de descontos e efetividade de campanhas promocionais



Oferta e Procura

Gerenciamento de estoques, previsão de demanda e otimização da cadeia de suprimentos



Lucros e Rendimentos

Análise de margens de lucro, rentabilidade por produto e performance financeira



Satisfação do Cliente

Estudos de comportamento, retenção e satisfação de clientes ao longo do tempo



Gerenciamento de Remessa

Otimização de logística, análise de prazos de entrega e custos de transporte

Cenário de Execução TPC-H

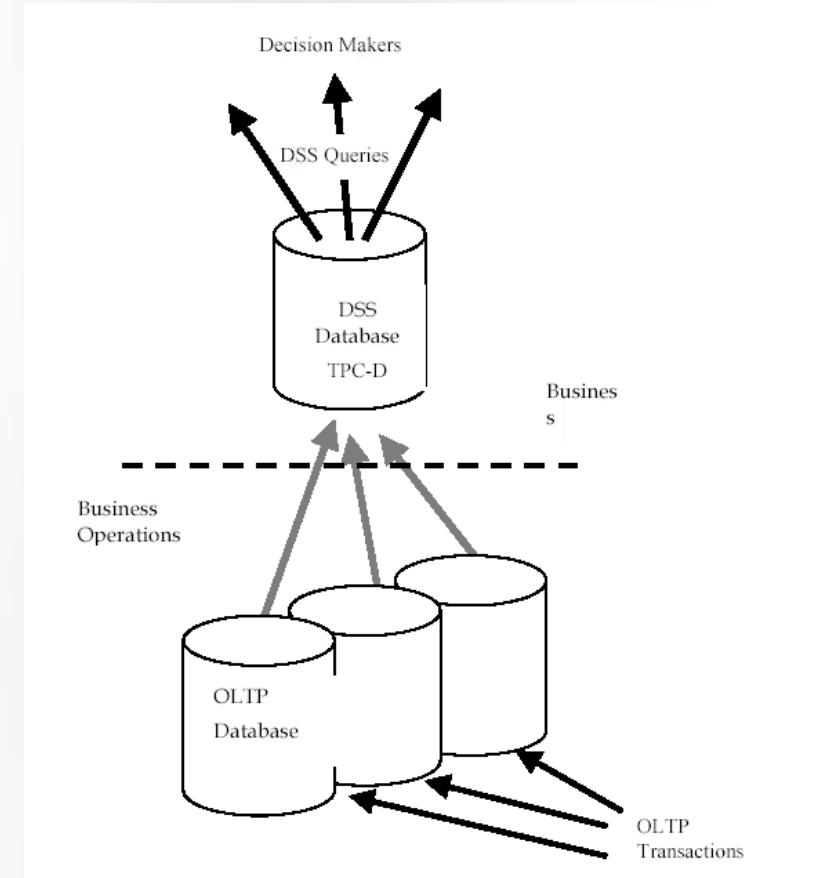
TPC-H vs TPC-C

Enquanto o TPC-C foca em transações rápidas e concorrentes, o TPC-H é projetado para consultas analíticas de longa duração que examinam grandes volumes de dados históricos.

📄 **Importante:** O sistema deve ser capaz de executar qualquer carga de trabalho OLAP, não apenas as consultas específicas do TPC-H

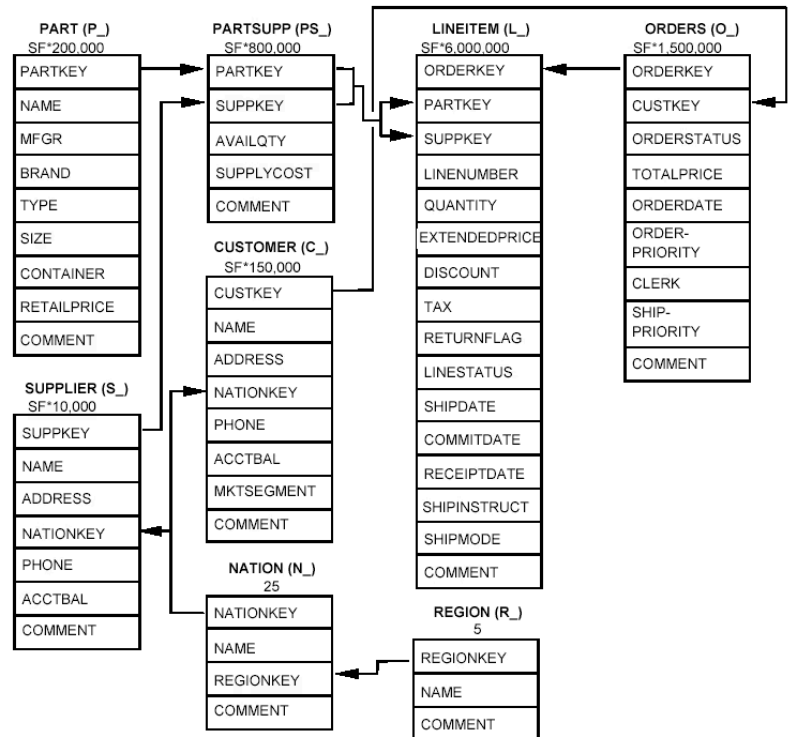
Neutralidade da Base de Dados

A base de dados deve ser construída conforme a especificação do TPC-H, sem otimizações específicas que favoreçam exclusivamente o benchmark.



Esquema do TPC-H

O esquema de banco de dados do TPC-H é baseado em um modelo de negócios de distribuição e vendas. Ele consiste em **oito tabelas principais** que representam entidades como clientes, pedidos, fornecedores, produtos e entregas.



Relacionamentos Complexos

As tabelas possuem múltiplas chaves estrangeiras e relacionamentos que simulam um ambiente empresarial real

Dados Escaláveis

O tamanho do banco pode variar de 1GB a vários terabytes através do fator de escala

Distribuição Estatística Controlada

Os dados seguem distribuições estatísticas definidas pela especificação do TPC-H para representar padrões de negócios de forma reproduzível

Executando o TPC-H



Preparação da Base

Utilize o **DBGen** (Database Generator) para gerar os dados do benchmark

- Define o fator de escala desejado
- Cria arquivos .tbl com dados formatados
- Gera tabelas conforme especificação



Preparação das Consultas

Use o **QGen** (Query Generator) para criar as 22 consultas parametrizadas

- Gera parâmetros aleatórios
- Cria variações das consultas
- Produz scripts executáveis



Ferramentas Oficiais do Benchmark

Utilize o DBGen e o QGen, conforme especificação do TPC-H, para geração de dados e consultas padronizadas.

Dicas para Carga de Dados

Conformidade com a Especificação

A estrutura das tabelas deve seguir rigorosamente a especificação do TPC-H, pois alterações na ordem ou definição das colunas comprometem a validade dos resultados.

Opções de Carga

Opção 1: Funções Nativas

Utilize funções específicas do SGBD que permitem a carga direta de arquivos com separadores (delimitadores)

- Mais rápida e eficiente
- Aproveita otimizações do banco
- Recomendada para grandes volumes

Opção 2: Programa Customizado

Desenvolva um programa (Java, .NET, Python) que leia os arquivos .tbl e faça a carga conectando-se à base

- Maior controle sobre o processo
- Permite transformações de dados
- Útil para debugging

Carga via SGBD: Exemplos Práticos

Cada sistema de gerenciamento de banco de dados possui seu próprio mecanismo otimizado para carga em massa. Aqui estão exemplos de comandos para os principais SGBDs:

PostgreSQL

```
-- Adicionar coluna temporária
alter table supplier add stub char(1);

-- Carregar dados
COPY supplier FROM '/data/tpch/supplier.tbl'
USING DELIMITERS '|' WITH NULL AS 'NULL';

-- Remover coluna temporária
alter table supplier drop column stub;
```

Observação: o comando COPY requer acesso do servidor ao arquivo; alternativamente, pode-se usar \copy via cliente.

Referência: <http://www.postgresql.org/docs/8.2/interactive/sql-copy.html>

MySQL

```
-- Carregar dados diretamente
load data infile "/var/lib/mysql-files/region.tbl"
IGNORE INTO TABLE region
fields terminated by '|';
```

Comando simples e direto para carga de arquivos delimitados

- ❑ **Boas práticas de carga em massa:** Para cargas volumosas, recomenda-se desabilitar índices e restrições temporariamente e reativá-los após a carga, conforme suportado pelo SGBD, a fim de reduzir o custo de inserção.

Referências



Elmasri & Navathe

Fundamentals of Database Systems

Pearson, 2016

Referência abrangente sobre fundamentos de sistemas de bancos de dados, cobrindo aspectos teóricos e práticos.

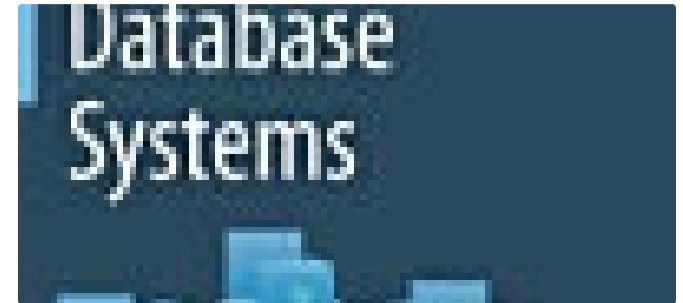


Korth, Sudarshan & Silberschatz

Database System Concepts

McGraw-Hill, 2019

Texto fundamental que serviu como base para a maioria dos exemplos apresentados nesta apresentação.



Özsu & Valduriez

Principles of Distributed Database Systems

Springer Nature, 2019

Obra especializada em sistemas de bancos de dados distribuídos, essencial para compreensão avançada.

❏ **Nota:** Os conceitos e exemplos apresentados baseiam-se principalmente na literatura clássica de sistemas de bancos de dados, em especial *Database System Concepts* e *Fundamentals of Database Systems*.