Análise de Dados com Base em Processamento Massivo em Paralelo

Introdução: Respostas dos Exercícios

Profa. Dra. Cristina Dutra de Aguiar Ciferri
André Perez
Guilherme Muzzi da Rocha
Jadson José Monteiro Oliveira
João Pedro de Carvalho Castro
Leonardo Mauro Pereira Moraes
Piero Lima Capelo

Observação:

Recomenda-se fortemente que a lista de exercícios seja respondida antes de se consultar as respostas dos exercícios.

1. OLTP (*on-line transaction processing*) diz respeito ao ambiente operacional, voltado ao processamento de transações. Isso significa que no ambiente OLTP existem muitas operações de inserção, remoção e atualização e que o objetivo de desempenho é realizar o processamento eficiente dessas operações.

   OLAP (*on-line analytical processing*) diz respeito ao ambiente informacional, voltado ao processamento de consultas analíticas. Isso significa que no ambiente OLAP existem muitas consultas e que o objetivo de desempenho é realizar o processamento eficiente dessas consultas.

2. *Data warehouse* representa o banco de dados, ou seja, é o local onde os dados são armazenados. *Data warehousing*, por sua vez, representa um ambiente, o qual é composto por *data warehouse*, *software*, *hardware* e *peopleware*.

   Data warehouse é um dos componentes de maior importância do *data warehousing*, consistindo no local onde os dados resultantes do processo de ETL (*extract*, *transform*, *load*) e modelados multidimensionalmente são armazenados.

3. Tabela com as respostas:

Ambiente Operacional	**Ambiente Informacional**
inserção/remoção/atualização	leitura (consulta)
OLTP	OLAP
interação estática/predefinida	interação dinâmica
poucos registros acessados por vez	muitos registros acessados por vez
grande número de usuários concorrentes	poucos usuários concorrentes

4. Relações entre as descrições e conceitos:

   a) Dados brutos sem significado semântico -> d) Dado.
   b) Informação interpretada, analisada, processada -> f) Conhecimento.
   c) Dados organizados, estruturados, contextualizados -> e) Informação.

5. Porque as análises propostas e o dados são consideravelmente complexos. Mesmo sendo possível usar as aplicações de bancos de dados existentes, existem diversos desafios a serem enfrentados. Esses desafios são muitas vezes extremamente custosos e, portanto, proibitivos para a produção da informação certa, na hora certa, para a pessoa certa.

   Alguns desafios que podem ser ressaltados dentro do contexto exemplificado são:

   - O dados de interesse estão espalhados em várias filiais. Consequentemente, esses dados devem ser obtidos de diferentes fontes de dados que normalmente assumem diferentes formatos e requerem processos de limpeza e tradução acurados.
   - A complexidade das consultas impacta no desempenho das mesmas. Na descrição dada, o objetivo é realizar análises para descobrir filiais que precisam ser fechadas ou remodeladas. Isso, muito provavelmente, envolveria a obtenção de inúmeros indicadores e informações a partir dos dados.
   - Tratamento dos dados temporais usualmente é incipiente, não existindo registro temporal para todas as análises possíveis de serem realizadas.

6. O dados que podem ser extraídos estão vinculados aos conjuntos de dados coletados, em sua forma bruta e sem significado semântico. Eles são: índices socioeconômicos, índices de contaminação, quantidade de testes, recuperação e óbitos para cada cidade ou região, datas referente às coletas realizadas, entre outros.

   As informações que podem ser extraídas surgem a partir da organização dos dados, estruturação e contextualização dos mesmos. Exemplos de informações que podem ser

extraídas são: regiões com maior taxa de contaminação, regiões que mais realizam a testagem de pessoas, regiões com maior taxa de óbitos, curva de contaminação ou recuperação de cada região, dentre outras.

O conhecimento provém das informações interpretadas, analisadas e processadas. Exemplos de conhecimento que pode ser extraído por meio das informações citadas supracitadas são:

- Regiões que têm características socioeconômicas parecidas, porém com curvas de contaminação diferentes, podem adotar medidas de combate à pandemia diferentes. Isto pode ser utilizado para se comparar a eficácia das diferentes medidas tomadas, por exemplo.
- Agregação das informações de índices socioeconômicos com curvas de contaminação, recuperação e óbito, possibilitando a detecção de padrões que facilitam a tomada de decisão estratégica. Essas informações agregadas podem considerar determinadas regiões com características peculiares, por exemplo.

7. Questão livre para discussão durante as tutorias. Não existe apenas uma resposta certa.