Análise de Dados com Base em Processamento Massivo em Paralelo

Lista de Exercícios: Introdução

Profa. Dra. Cristina Dutra de Aguiar

Observação:

Esta lista contém exercícios classificados como essenciais e complementares. A indicação da classificação de cada exercício é feita junto de sua definição. A resposta de cada exercício encontra-se destacada na cor azul. Recomenda-se fortemente que a lista de exercícios seja respondida antes de se consultar as respostas dos exercícios.

1. (Essencial) Qual a diferença entre OLTP e OLAP?

   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. (Essencial) Liste os principais aspectos pelos quais o ambiente de *data warehousing* se difere do conceito de *data warehouse*.

   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.

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

3. (Essencial) Na aula, foram contextualizados dois ambientes: o ambiente operacional e o ambiente informacional. Esses ambientes são distintos entre si e cada um deles tem características importantes que os definem. Descreva cada um desses ambientes, destacando as suas principais características.

   No ambiente operacional, os tipos de operação mais frequentes são de inserção, remoção e atualização dos dados, o que é uma característica do OLTP (*on-line transaction processing*). As interações com os usuários são usualmente estáticas e predefinidas. As aplicações do ambiente operacional vislumbram o processamento eficiente das operações de inserção, remoção e atualização dos dados. Nesse ambiente, a quantidade de usuários que usam cada aplicação simultaneamente é muito grande.

   No ambiente informacional, o tipo de operação mais frequente é a consulta aos dados, ou seja, a leitura dos dados, o que é uma característica do OLAP (*on-line analytical processing*). As interações realizadas com os usuários são usualmente dinâmicas, desde que os usuários podem consultar os dados de acordo com diferentes perspectivas de análise. As aplicações do ambiente informacional vislumbram o processamento eficiente das consultas. Essas consultas são caracterizadas por acessar inúmeros registros, desde que usualmente realizam análises massivas dos dados. Poucos usuários interagem com o ambiente informacional simultaneamente. Esses usuários geralmente são executivos, analistas e gerentes, ou seja, usuários voltados à tomada de decisão estratégica.

4. (Essencial) O reitor de uma universidade precisa tomar uma decisão acerca da distribuição de recursos financeiros entre os institutos da universidade. Em reunião com os pró-reitores da instituição, foi definido que aqueles institutos com maior quantidade de publicações científicas em revistas de alto fator de impacto devem ser priorizados. Sendo assim, o reitor decidiu utilizar a ferramenta OLAP da universidade para emitir um relatório contendo a quantidade de publicações por instituto, por revista e por trimestre. O relatório emitido pelo reitor pode ser classificado em qual categoria: dado, informação ou conhecimento? Justifique sua resposta detalhando o porquê do relatório não ter sido classificado nas outras duas categorias.

   O relatório emitido pelo reitor representa uma informação, uma vez que os dados presentes neste relatório mostram a quantidade (ou seja número) de publicações agrupados por diferentes perspectivas (por instituto, por revista e por trimestre). O relatório não pode ser classificado como um dado bruto, uma vez que houve um processamento para sua geração. Além disso, o relatório também não pode ser classificado como conhecimento, visto que ainda não foi devidamente interpretado pelo reitor e ainda não foi obtido nenhum direcionamento, conclusão ou tomada de decisão a partir do mesmo.

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

5. (Essencial) Considere uma empresa de supermercados que possui várias filiais. Cada filial possui um sistema diferente para contabilizar os produtos vendidos e as promoções realizadas. Um executivo dessa empresa deseja fazer uma análise para descobrir filiais que precisam ser fechadas ou remodeladas. Por que não é ideal que esse executivo realize essa análise sobre os sistemas existentes da empresa?

   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.
   - O tratamento dos dados temporais usualmente é incipiente, não existindo registro temporal para todas as análises possíveis de serem realizadas.

6. (Complementar) Considere a seguinte situação:

   > Uma empresa brasileira especializada em análise de dados busca estudar e entender as maiores consequências que a pandemia gerada pelo novo coronavírus ocasionou nas cidades brasileiras. O principal objetivo é determinar quais diferentes medidas podem ser tomadas, levando em consideração as diversas características socioeconômicas que cada cidade ou região do país pode ter. Para dar início ao estudo, a empresa catalogou conjuntos de dados que contêm índices socioeconômicos e dados referentes aos índices de contaminação, quantidade de testes, recuperação e óbitos para cada cidade.

   Com base nessa descrição, cite quais dados podem ser extraídos, quais informações podem ser obtidas e quais conhecimentos podem ser construídos.

   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.

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

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. (Complementar) Pesquise uma situação real na qual seria interessante aplicar consultas analíticas. Descreva o problema e elabore três perguntas para exemplificar possíveis análises. A partir dessas perguntas, descreva três exemplos de conhecimento que podem ser gerados por meio dessas análises.

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

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.