Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
MBA em Inteligência Artificial e Big Data
– Curso 3: Administração de Dados Complexos em Larga Escala –

Questões da 1' Quinzena: Técnicas avançadas para Preparação de Dados em SQL

Prof. Dr. Caetano Traina Júnior

Exercícios sobre Conceitos Básicos de Mineração em Grandes Bases de Dados

---

Exercício 1) Responda o que você entende por

- Mineração de Dados
- Descoberta de Conhecimento em Bases de Dados
- Data *Warehouse*
- *Big Data*
- Escalabilidade
- Os **Big V**s da Mineração em Grandes Bases de Dados
- Ciências de Dados × Engenharia de Dados
- *Open Data* × *Big Data*

Exercício 2) Qual a diferença entre:

- Processos de Mineração de Dados × *Warehousing* de Dados
- Mineração de Dado × Descoberta de Conhecimento em Bases de Dados
- OLAP × OLTP
- Com referência a volumes de dados: Cardinalidade × Dimensionalidade × Resolução
- Jargão da área (procurar na internet): Datalake × Dataswamp

---

Exercício 3) Quais são as principais técnicas para se conseguir Escalabilidade nos processos de extração de conhecimento em grandes volumes de dados?

Exercícios sobre Dados Agregados em SQL (`CUBE` e `ROLLUP`)

Exercício 4) Mostrar o total de pacientes em cada cidade por faixa de idades (usar a década da idade como faixa: de 0 a 9 anos, de 10 a 19, etc.). Contabilizar também o total de pacientes em cada faixa (independente da cidade) e de cada cidade (independente da faixa).

Exercício 5) Mostrar o total de pacientes total, quantos foram a óbito e quantos sobreviveram em cada cidade por faixa de idades (usar a década da idade como faixa: de 0 a 9 anos, de 10 a 19, etc.).
Contabilizar também o total de pacientes em cada faixa (independente da cidade) e de cada cidade (independente da faixa).
Indicar com clareza quais são as cidades e idades conhecidas e desconhecidas (`NULLS`) e quais medidas correspondem a sub-totalizadores.

Exercícios sobre Funções de Janelamento em SQL

Exercício 6) Considere que se pretende obter os pacientes 'mais novos' e 'mais velhos' em cada cidade, na base Fapesp-Covid. Escreva um comando que responda a essa consulta:

- com uma sub-consulta usando apenas a cláusula '`GROUP BY`';
- com sub-consultas usando a construção CTE (*Common Table Expression* '`WITH queries`');
- usando '`Window functions`'.

Exercício 7) A tabela de Exames ('`ExamLabs`') reporta uma medida sobre um analito em cada tupla. Portanto, os exames que medem diversos analitos são representados em diversas tuplas. No entanto, pode-se assumir que, se foram registrados dois exames iguais no mesmo dia para o mesmo paciente, pode-se assumir como valor a ser considerado a média dos valores medidos em cada analito.

- Escreva uma consulta que mostre quais analitos podem ser medidos em exames de '**hemograma**', em cada hospital.
- Compare os nomes dos analitos entre os diferentes hospitais, e execute um processo de atualização dos nomes, corrigindo e integrando as variantes e grafias óbvias.

Exercício 8) Escreva uma consulta que associe qual é o desfecho do atendimento correspondente a cada exame, e inclua um atributo indicando a quantos dias desde o inicio do atendimento correspondente aquele exame foi efetuado.

Exercício 9) Escreva uma consulta que gere a relação de todos os exames de **colesterol** que foram efetuados, de maneira que cada tupla dessa relação inclua as medidas de todos analitos correspondentes desse exame (executar o pivotamento da relação de exames, reproduzindo o exemplo mostrado em aula). Para isso, considere que cada exame de cada paciente é realizado em um único dia, e que se houver repetição de medidas do mesmo analito, deve ser considerada a média de todas as medidas desse analito. Analitos não medidos num exame devem ficar nulos. Inclua nessa tabela o desfecho que o paciente teve para o atendimento onde esse exame foi feito.

Exercício 10) Escreva uma consulta equivalente à anterior, agora para os exames de hemograma que foram efetuados. Nessas tabelas, cada tipo de exame seguiu uma estrutura diferente. Neste caso a principal diferença para gerar as duas tabelas é que, enquanto para obter os exames de colesterol cada medida é independente, e a escolha das tuplas teve que ser feita diretamente pelo atributo 'De_Analito', os exames de hemograma são identificados por um único valor no tipo de exame (embora hospitais diferentes possam usar nomes diferentes para o mesmo exame) e portanto o atributo '`De_Exame`' pode ser usado como filtro de seleção.

Exercício 11) Considerando exames de Covid, substitua os valores do atributo '`De_Resultado`' que tenham valores numéricos para '`Positivo`' e '`negativo`' considerando o atributo '`CD_ValorReferencia`'.

Exercício 12) Faça uma consulta equivalente à de exames de hemograma, agora para exames vinculados a testes de covid, usando o resultado da consulta anterior. Inclua na relação resultante o número de dias entre dois exames que tenham resultado mudado a medida entre 'positivo' e 'negativo' para Covid.