



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 –

Exercícios para a 1ª Quinzena: Técnicas avançadas para Preparação de Dados em SQL

Prof. Dr. Caetano Traina Júnior

Exercícios sobre Técnicas avançadas para Preparação de Dados em SQL, Aula 1

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 Vs** da Mineração em Grandes Bases de Dados
- Ciências de Dados × Engenharia de Dados
- *Open Data* × *Big Data* × FAIR 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
- *Data warehouse* × *Data lake*
- Com referência a volumes de dados: Cardinalidade × Dimensionalidade × Resolução
- Jargão da área (procurar na internet): *Data Lake* × *Data Swamp*

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

Exercício 4) Descreva quais são as principais contribuições de cada área do conhecimento envolvidas em Ciência e Engenharia de Dados:

- Inteligência Artificial/Aprendizado de máquina
- Bases de dados
- Estatística
- Teoria da Informação
- Visualização de dados e de Informação
- Computação de alto desempenho
- Regras do negócio da aplicação

Alguns dos exercícios a seguir são propostos para utilizar o SGBD  PostgreSQL. Você pode usar outros SGBDs multimodelos (como por exemplo  ORACLE e  Microsoft SQL Server), mas não garantimos que dúvidas quanto à execução de consultas nesses sistemas serão atendidas.

É usada a base de dados [FapespCovid-19](#), a qual pode ser carregada usando o Notebook  1.3-ExploraDS-FAPESPCovid.ipynb.

Trata-se de uma base de dados “reais”, com 3 tabelas: 47.971 pacientes, 6.855.217 resultados de exames e 260.682 desfechos de internações. Os dados estão disponíveis no repositório [COVID-19 Data Sharing/BR - USP](#) em <https://repositoriodatasharingfapesp.uspdigital.usp.br/>.

Para os exercícios, podem ser carregados apenas os dados dos hospitais que disponibilizam **desfechos**: Hospital Sírio-Libanês e Hospital Beneficência Portuguesa em SP.

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

Exercício 5) Usando a base de dados [FapespCovid-19](#):

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 6) Usando a base de dados [FapespCovid-19](#):

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 7) Usando a base de dados [FapespCovid-19](#):

Considere que se pretende obter os pacientes ‘mais novos’ e ‘mais velhos’ em cada cidade. 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 8) Usando a base de dados [FapespCovid-19](#):

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 9) Usando a base de dados [FapespCovid-19](#):

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 início do atendimento correspondente aquele exame foi efetuado.

Exercício 10) Usando a base de dados [FapespCovid-19](#):

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 11) Usando a base de dados [FapespCovid-19](#):

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 12) Usando a base de dados [FapespCovid-19](#):

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

Exercício 13) Usando a base de dados [FapespCovid-19](#):

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.

Exercícios sobre Tipos de Dados não atômicos em SQL

Exercício 14) Um serviço de saúde médico (hospital, consultório, etc.) registra a anamnese dos paciente, incluindo as informações de identificação de cada paciente e seu perfil sociodemográfico, que permite a interpretação da situação clínica do paciente. São **atributos** da identificação e perfil sociodemográfico do paciente:

(cada objeto escrito em: **negrito** é um atributo; **monoespaçado** é um valor)

- **Nome completo**: sem abreviações;
- **Data de nascimento**: com a possibilidade da indicação da **Idade** na data da consulta;
- **Sexo biológico**, como **masculino** ou *feminino*, e opcionalmente **Gênero**, como *cisgênero* ou *transgênero* com respectiva **Identificação de gênero** segundo a auto-declaração do paciente;
- **Cor/etnia**, segundo a nomenclatura: Branca, Parda, Preta, Indígena, Asiática, ou outra;
- **Estado civil**, segundo a nomenclatura: Casado, Solteiro, Divorciado, Separado, Viúvo, União estável, Outro;
- **Ocupação**, com a indicação de pelo menos uma atividade produtiva que o paciente exerce no dia a dia, e **Situação** da ocupação, como **Ativo**, **Licença trabalhista**, **Aposentado**, incluindo uma descrição do **Local** e **Condições** de trabalho de cada ocupação;

- Escolaridade, segundo a nomenclatura: Alfabetizado, Analfabeto, Doutorado, Especialização, Fundamental, Graduação, MBA, Médio, Mestrado, Pós-doutorado, Técnico;
- **Religião**, é importante porque interfere na vida do paciente e na sua relação com determinadas situações e doenças;
- **Naturalidade**, o local onde o paciente nasceu, e a **Procedência** que é o local de residência atual do paciente;
- **Grau de confiabilidade** é uma estimativa do médico sobre a qualidade das informações fornecidas pelo paciente, e indicada como um valor de 1 a 5, onde 5 é o maior grau de confiança.

Defina a estrutura de dados para registrar as informações básicas da ‘anamnese’ de um paciente. A estrutura deve levar em conta todas as restrições definidas e, além disso, o Estado civil e a Escolaridade devem ser passíveis de ordenação para a geração de análises demográficas.

Exercício 15) Usando a base de dados [FapespCovid-19](#), execute a seguinte sequência de atividades:

- Usando a mesma descrição de uma estrutura para a ‘Anamnese’ do exercício anterior, defina **Ocupação** como um objeto com os quatro atributos: **nome** da ocupação, a **situação**, o **local** e **condições** de trabalho.
- Simplifique a definição da ‘Anamnese’ para ter apenas **Nome**, **DataNasc**, **Idade** e uma lista de ocupações.
- Crie uma relação **Internações** de pacientes em um hospital, que tenha como atributos:
 - O **nome** do hospital,
 - A **anamnese** com a identificação do paciente e suas informações sociodemográficas,
 - A **Data da internação**
- Dê um exemplo para inserir uma tupla com Anamnese indicando paciente com apenas uma ocupação, e outra com com Anamnese indicando paciente com duas ocupações nessa relação.

Exercício 16) Usando a base de dados [FapespCovid-19](#):

Listar os pacientes da cidade de ‘Barueri’ que tenham tipo de atendimento como sendo **internado**, indicando para cada seu identificador, sexo e ano de nascimento, mais quantas internações ele teve, listando as respectivas datas.

Exercícios sobre Textos Semi-estruturados em SQL


Exercício 17)

Considere a seguinte afirmação: “O conceito de um **Modelo de Documentos** e o conceito de um **Projeto *schema-less*** são independentes.” Explique o que é cada conceito e faça uma comparação mostrando as diferenças

Exercício 18) Usando a base de dados [FapespCovid-19](#):

Crie um novo atributo de tipo JSON na relação de **Pacientes** que armazene as informações sobre os atendimentos do paciente (**DE.TipoAtendimentos** registrados na tabela de desfecho), indicando a data de cada atendimento ou internação (**DT.Atendimento**) e de alta (**DT.Desfecho**), seu tipo, e em qual clínica ele foi internado (**DE.Clinica**).

Exercício 19) Deve ser feita a carga dos dados sobre os **Prêmios Nobel** atribuídos ao longo dos anos, que estão disponíveis em <https://github.com/jdorfman/awesome-json-datasets> como três arquivos em formato JSON.

Carregar uma tabela JSON é mais fácil se a leitura for feita primeiro como uma estrutura de texto e depois analisada usando o parser do próprio  PostgreSQL. Isso requer o dobro de espaço de armazenagem (em disco) mas é a maneira mais flexível. Além disso, deve-se considerar que a maioria das outras alternativas requerem o uso de ferramentas externas, que também vão armazenar o texto, portanto não existe muita diferença em termos de custo de memória, mas requerem o uso de duas ferramentas diferentes.

No caso dos *datasets* ‘Nobel’, cada arquivo guarda apenas um objeto JSON no formato de um *array* que tem os demais objetos (prêmios, premiados, países). Obtenha os três arquivos JSON no diretório indicado. A seguir, crie uma relação que tem apenas um atributo de tipo JSON: uma coleção de documentos, para armazenar cada coleção de prêmios, premiados e países. Por fim, faça a carga desses documentos nas respectivas relações.

Exercício 20) Usando a base de dados [Nobel](#) carregada no exercício anterior:

Mostrar os prêmios atribuídos em cada categoria, indicando para cada um, em formato de texto:

- O ano do prêmio,
- A categoria,
- Quantos foram os premiados nessa categoria nesse ano,
- Quem foram os ganhadores, um por atributo (pode assumir um número máximo=4)