



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 1ª 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 Vs** 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).

**Resposta:**

Dificuldade: 

```
SELECT CD_Municipio, 10*ROUND((2021-AA_Nascimento)/10) AS FaixaIdade,
       Count(*)
FROM Pacientes
GROUP BY ROLLUP (CD_Municipio, FaixaIdade)
ORDER BY CD_Municipio, FaixaIdade;
```

### 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.

**Resposta:**

Dificuldade: 

```
SELECT CASE WHEN GROUPING( 10*ROUND((2021-P.AA_Nascimento)/10))=1
            THEN '---' ELSE (10*ROUND((2021-P.AA_Nascimento)/10))::TEXT
            END FaixaIdade,
       CASE WHEN GROUPING(CD_Municipio)=1 THEN '---' ELSE CD_Municipio
            END Municipio,
       COUNT(*) Total,
       Count(*) FILTER (WHERE FD.DE_FDesfecho~*'[oó]bito') Obitos,
       Count(*) FILTER (WHERE FD.DE_FDesfecho!~*'[oó]bito') Sobreviventes
FROM Pacientes P LEFT JOIN
    (SELECT ID_Paciente, Dt_Fatendimento, De_FDesfecho
     FROM (SELECT ID_Paciente,
                  MAX(Dt_Atendimento) OVER(Partition BY ID_Paciente)
                  AS Dt_Fatendimento,
                  MAX(De_desfecho) OVER(Partition BY ID_Paciente)
                  AS De_FDesfecho
          FROM Desfechos
        ) As temp
     GROUP BY ID_Paciente, Dt_Fatendimento, De_FDesfecho
    ) AS FD -- Desfecho Final
ON P.id_paciente=FD.ID_paciente
WHERE FD.DE_FDesfecho IS NOT NULL
GROUP BY ROLLUP (CD_Municipio, 10*ROUND((2021-P.AA_Nascimento)/10))
ORDER BY CD_Municipio NULLS LAST, FaixaIdade NULLS LAST;
```

## 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**’;

**Resposta:**

Dificuldade: 

```
SELECT P.ID_Paciente, P.CD_Municipio, P.AA_Nascimento,
       MM.AAN_Min, MM.AAN_Max
FROM Pacientes P, (
    SELECT P.CD_Municipio, Min(P.AA_Nascimento) AAN_Min,
           Max(P.AA_Nascimento) AAN_Max
    FROM Pacientes P
    GROUP BY CD_Municipio) MM
WHERE (P.CD_Municipio=MM.CD_Municipio AND P.AA_Nascimento=MM.AAN_Min) OR
      (P.CD_Municipio=MM.CD_Municipio AND P.AA_Nascimento=MM.AAN_Max)
ORDER BY 2,3;
```

- com sub-consultas usando a construção CTE (*Common Table Expression* ‘[WITH queries](#)’);

**Resposta:**

Dificuldade: 

```
WITH MM AS (SELECT P.CD_Municipio,
                   Min(P.AA_Nascimento) AAN_Min, Max(P.AA_Nascimento) AAN_Max
            FROM Pacientes P
            GROUP BY CD_Municipio)
SELECT P.ID_Paciente, P.CD_Municipio, P.AA_Nascimento, MM.AAN_Min, MM.AAN_Max
FROM Pacientes P, MM
WHERE (P.CD_Municipio=MM.CD_Municipio AND P.AA_Nascimento=MM.AAN_Min) OR
      (P.CD_Municipio=MM.CD_Municipio AND P.AA_Nascimento=MM.AAN_Max)
ORDER BY 2,3;
```

- usando ‘[Window functions](#)’.

**Resposta:**

Dificuldade: 

```
SELECT * FROM (
    SELECT P.ID_Paciente, P.CD_Municipio, P.AA_Nascimento,
           Min(AA_Nascimento) OVER(Partition by CD_Municipio) AS AAN_Min,
           Max(AA_Nascimento) OVER(Partition by CD_Municipio) AS AAN_Max
    FROM Pacientes P) AS MM
WHERE (AA_Nascimento=AAN_Min OR
      AA_Nascimento=AAN_Max) AND
      CD_Municipio IS NOT NULL
ORDER BY 2,3;
```

Notas:

1. A cláusula `ORDER BY` e a listagem dos atributos `AAN_Min` e `AAN_Max` é opcional, colocadas aqui apenas para conferir o resultado.
2. A condição `AND CD_Municipio IS NOT NULL` do terceiro item evita colocar no resultado tuplas onde a cidade é desconhecida, que são naturalmente eliminadas nas demais opções por causa da comparação `P.CD_Municipio=MM.CD_Municipio`. Portanto também pode ser considerada opcional.

### 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.

#### Resposta:

Dificuldade: 

Existem muitas respostas possíveis. Uma consulta básica pode ser:

```
SELECT DE_Exame, DE_Analito, DE_Hospital, Count(*), Count(DE_Analito)
FROM Examlabs
WHERE DE_Exame~*'hemograma'
GROUP BY DE_Exame, DE_Analito, DE_Hospital
ORDER BY 2, 1, 3;
```

- 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.

#### Resposta:

Dificuldade: 

Analisando a resposta da consulta anterior, percebe-se que os nomes dos exames e dos analitos têm grafias diferentes nos diversos hospitais, sendo que a acentuação dos analitos é provavelmente a mais óbvia. Mas também existem variações na maneira como são requisitados, como por exemplo 'Hemograma' no 'BPSP' e 'Hemograma completo' no 'Hemograma completo, sangue total' no 'HSL'. O comando seguinte pode corrigir alguns problemas localizados de acentuação nos analitos e correção de nomes dos exames:

```
SELECT regexp_replace(lower(DE_Exame), '( completo|sangue total|,)', '', 'g')
deexame,
    regexp_replace(
    regexp_replace(
    regexp_replace(
    regexp_replace(
    regexp_replace(lower(DE.analito), '[ó]filos', 'ófilos')
        , '[ó]citos', 'ócitos')
        , '[ó]crito', 'ócrito')
        , 'plaquet.rio m.dio', 'plaquetário médio')
        , 'fracao', 'fração') deanalito,
    DE_Hospital, Count(*)
FROM ExamLabs
WHERE De_Exame ~*'hemograma'
GROUP BY 2, 1, 3
ORDER BY 2, 1, 3;
```

### 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 início do atendimento correspondente aquele exame foi efetuado.

**Resposta:**

Dificuldade: 

```
SELECT E.DT_Coleta - FIRST_Value(E.DT_Coleta)
      OVER (PARTITION BY D.id_paciente
            ORDER BY E.DT_Coleta, E.DE_Exame, E.de_analito) Separacao,
D.de_desfecho Desfecho,
E.*
FROM ExamLabs E JOIN Desfechos D
  on (E.id_paciente, E.id_atendimento) = (D.ID_Paciente, D.id_atendimento)
```

### 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.

**Resposta:**

Dificuldade: 

```
WITH Colest AS (
  SELECT P.id_paciente, E.id_atendimento, E.dt_coleta,
    Max(E.de_resultado||' '||cd.unidade) FILTER(WHERE
      E.DE_Exame ~*'colesterol - fra[cç][aã]o ldl|ldl - colesterol'
    ) AS LDL,
    Max(E.de_resultado||' '||cd.unidade) FILTER(WHERE
      E.DE_Exame ~*'Colesterol - fra[cç][aã]o HDL|hdl - colesterol'
    ) AS HDL,
    Max(E.de_resultado||' '||cd.unidade) FILTER(WHERE
      E.DE_Exame ~*'Colesterol - fra[cç][aã]o vldl|vldl - colesterol'
    ) AS VLDL,
    Max(E.de_resultado||' '||cd.unidade) FILTER(WHERE
      E.DE_Exame ~*'Colesterol n[aã]o-hdl, soro|n[aã]o - hdl - colesterol'
    ) AS nao_HDL
  FROM ExamLabs E JOIN Pacientes P on E.id_paciente = P.ID_Paciente
  WHERE E.DE_Exame ~*'colest'
  GROUP BY P.id_paciente, E.id_atendimento, E.dt_coleta)
SELECT C.id_paciente, C.dt_coleta,
  C.LDL, C.HDL, C.VLDL, C.Nao_HDL, D.de_desfecho
FROM Colest C JOIN Desfechos D
  on (C.id_paciente, C.id_atendimento) = (D.ID_Paciente, D.id_atendimento)
WHERE HDL IS NOT NULL OR LDL IS NOT NULL;
```

### 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.

Resposta:

Dificuldade:

NOTA:

1. Os analitos a serem considerados são os obtidos pelo exercício 5, possivelmente agrupando só os analitos e desconsiderando os hospitais e nomes de exames (mas mantendo a condição de seleção na cláusula `WHERE`).
2. Usar os analitos com nos nomes corrigidos ou os originais é uma opção de cada um.
3. A indicação `WHERE COALESCE ...` foi usada aqui para garantir que ao menos alguns analitos fundamentais estejam presentes no exame. Essa cláusula não foi solicitada no exercício.


WITH Hemograma AS (

```
SELECT P.id_paciente, E.id_atendimento, E.dt_coleta,
       Max(E.de_resultado||' '||cd.unidade)
       FILTER(WHERE E.DE_Analito ~*'basofilos') AS Basofilos,
       Max(E.de_resultado||' '||cd.unidade)
       FILTER(WHERE E.DE_Analito ~*'bastonetes') AS Bastonetes,
       Max(E.de_resultado||' '||cd.unidade)
       FILTER(WHERE E.DE_Analito ~*'blastos') AS Blastos,
       Max(E.de_resultado||' '||cd.unidade)
       FILTER(WHERE E.DE_Analito ~*'chcm') AS CHCM,
       Max(E.de_resultado||' '||cd.unidade)
       FILTER(WHERE E.DE_Analito ~*'Eosinofilos') AS Eosinofilos,
       ... os demais analitos ...
       Max(E.de_resultado||' '||cd.unidade)
       FILTER(WHERE E.DE_Analito ~*'Volume plaquetário médio') AS VolPlaq
FROM ExamLabs E JOIN Pacientes P on E.id_paciente = P.ID_Paciente
WHERE E.DE_Exame ~*'hemograma'
GROUP BY P.id_paciente, E.id_atendimento, E.dt_coleta)
SELECT C.id_paciente, C.dt_coleta,
       C.Basofilos, C.Bastonetes, C.Blastos, C.CHCM, C.Eosinofilos, ...
       C.VolPlaq, D.de_desfecho
FROM Hemograma C JOIN Desfechos D
on (C.id_paciente, C.id_atendimento) = (D.ID_Paciente, D.id_atendimento)
WHERE COALESCE (Basofilos, Blastos, Eosinofilos) IS NOT NULL;
```

### 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'.

**Resposta:**

Dificuldade: 

```
SELECT de_analito, de_resultado::VARCHAR(20), cd_valorreferencia,
       de_resultado !~ '^[^d.,+-]' TemNum, -- tem medida numérica?
CASE WHEN de_resultado !~ '^[^d.,+-]' THEN
  CASE WHEN ((CASE WHEN DE_Resultado !~ '^[^d.,+-]' THEN
    Regexp_Replace(Regexp_Replace(DE_Resultado, '^[^d.,+-]', '\1'),
      ',,', '.'):REAL ELSE 0.0 END)
    >
    (CASE WHEN cd_valorreferencia ~ '.*[^\d.,+-]' THEN Regexp_Replace(
      (Regexp_Replace(cd_valorreferencia, '.*([^\d.,+-])', '\1'),
        ',,', '.'):Real ELSE 0.0 END))
  THEN 'Positivo' ELSE 'Negativo' END -- Resultado numérico e ref.
WHEN de_resultado ~* 'posit|^detec' THEN 'Posit'
WHEN de_resultado ~* 'negat|n.o detec' THEN 'Negat'
ELSE 'desconhecido' END REsultCovid, -- Resultado do exame de covid
CASE WHEN DE_Resultado !~ '^[^d.,+-]' THEN Regexp_Replace(
  Regexp_Replace(DE_Resultado, '^[^d.,+-]', '\1'),
    ',,', '.'):REAL ELSE 0.0 END Medida, -- Medida numérica
CASE WHEN cd_valorreferencia ~ '.*[^\d.,+-]' THEN Regexp_Replace(
  Regexp_Replace(cd_valorreferencia, '.*([^\d.,+-])',
    '\1'), ',,', '.'):Real ELSE 0.0 END Limite -- Limite numérico
FROM examLabs
WHERE de_exame ~* 'covid';
```

NOTA: Neste exercício, a maior dificuldade é extrair os números dos campos alfabéticos, o que foi feito de uma maneira simplificada usando padrões em expressões regulares.

## 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.

Resposta:

Dificuldade: 

```
WITH Temp AS (SELECT id_paciente, ExamLabs.DT_Coleta,
    de_resultado !~ '^[^d.,+-]' TemNum, -- tem medida numérica?
    CASE WHEN de_resultado !~ '^[^d.,+-]' THEN
        CASE WHEN ((CASE WHEN DE.Resultado! ~ '^[^d.,+-]' THEN
            Regexp_Replace(Regexp_Replace(DE.Resultado, '^[^d.,+-]', '\1'),
                ',, ', '.')::REAL ELSE 0.0 END)
            >
            (CASE WHEN cd_valorreferencia ~ '.*[^d.,+-]' THEN Regexp_Replace(
                Regexp_Replace(cd_valorreferencia, '.*([^\d.,+-])', '\1'),
                    ',, ', '.')::Real ELSE 0.0 END))
        THEN 'Positivo' ELSE 'Negativo' END -- Resultado numérico
        WHEN de_resultado ~* 'posit|^detec' THEN 'Posit'
        WHEN de_resultado ~* 'negat|n.o detec' THEN 'Negat'
        ELSE 'desconhecido' END RESultCovid, -- Resultado do exame de covid
    RoW_Number(*) OVER(
        PARTITION BY ID_paciente ORDER BY DT_Coleta) seqExame,
    DT_Coleta - lag(DT_Coleta) OVER (
        PARTITION BY id_paciente ORDER BY DT_Coleta) Separacao
FROM examLabs
WHERE de_exame ~* 'covid')
SELECT *, CASE WHEN (RESultCovid != lag(RESultCovid) OVER (
    PARTITION BY id_paciente ORDER BY DT_Coleta))
    THEN Separacao ELSE Null END Mudou
FROM Temp;
```

NOTA: Este exercício basicamente usa a solução do anterior, calculando a tabela de exames de Covid como a tabela-visão [Temp](#), e acrescentando o atributo 'Mudou'.