Teste Técnico: Pessoa Engenheira de Dados Involves

Nome: **Thiago Bragança Soares de Oliveira**
Empresa: **Involves**

1 Questões

1. Descreva com suas palavras os principais conceitos abaixo:

   (a) O que é um Data Warehouse?

   (b) Quais características possuem as tabelas do tipo Fato e Dimensão?

   (c) O que é ETL?

   (d) Quais são as principais atribuições de um Engenheiro de Dados?

   (e) O que é Trade Marketing?

2. Crie uma query, considerando o SGBD MySQL, para exibir todos os dados de uma tabela de Pontos de Venda (tabela origem PONTO_VENDA_UNIDADE) e restringir apenas os pontos de venda que possuem sell in maior que 20.000 (campo SELLIN) e ainda ordená-los por nome do ponto de venda (campo NOME_PDV).

3. Considerando a tabela de origem da questão anterior, crie uma query que some o valor de sell in de acordo com cada ponto de venda e agrupe os resultados por mês (campo MES) e ano (campo ANO). Ordene os registros por um período cronológico de forma crescente e por nome do ponto de venda.

4. Considerando a tabela de origem da questão 2 e uma segunda tabela VISITAS_PONTO_VENDA, crie uma query que calcule a quantidade de visitas do ponto de venda de nome INVOLVES, sabendo-se que a tabela de visitas possui um campo que identifica se o ponto de venda foi visitado ou não chamado FL_VISITADO (Se 1 = Ponto de venda visitado / Se 0 = Ponto de venda não visitado). O campo chave que liga as duas tabelas é ID_PDV (na tabela PONTO_VENDA_UNIDADE) e FK_PDV (na tabela VISITAS_PONTO_VENDA). A query deve mostrar apenas as informações de nome do ponto de venda e quantidade de vistias realizadas.

5. Considerando a query abaixo, a pessoa engenheira de dados identificou que a performance da query está muito abaixo do esperado. Imaginando que um dos problemas possa estar relacionado aos índices das tabelas do banco de dados, a pessoa resolveu criar os índices nas tabelas. Liste quais possíveis campos devem ser indexados nas tabelas do banco de dados para que a query criada possa performar melhor. Leve em consideração que nenhum campo no banco de dados está indexado.

```sql
    select
        FT.CICLO,
        FT.ID_DIM_PDV,
        FT.ID_BLOCO_ITEM,
        SUM(FT.QTD_PONTO_EXTRA),
        SUM(FTPI.TOTAL_NOTA_ITEM)
    from FT_DOMINANCIA_PONTO_EXTRA_COMPLIANCE FT
```

```
    inner join TABREF_PAINEL_LOJAS_LP TPLL
        on FT.ID_DIM_PDV = TPLL.ID_DIM_PDV
           and FT.CICLO = TPLL.CICLO
    inner join FT_PERFECTSTORE_ITEM FTPI
        on FT.CICLO = FTPI.CICLO
           and FT.ID_DIM_PDV = FTPI.ID_DIM_PDV
           and FT.ID_BLOCO_ITEM = FTPI.ID_BLOCO_ITEM
           and FT.SEMANA_LP = FTPI.SEMANA_LP
    where
        FT.CICLO = 202009
    and FT.ID_DIM_PDV = 223459792
    group by FT.CICLO,
        FT.ID_DIM_PDV;

```

6. Considere a instrução Python a seguir:

```
x = [print(i) for i in range(10) if i % 2 == 0]

```

   Após a execução dessa instrução no Python, a variável "x" conterá qual valor?

7. Faça um script em Python que peça dois números e imprima a soma.

   Para responder às questões 8, 9 e 10 utilize a ferramenta Pentaho Data Integration (PDI) na versão de sua preferência. A ETL final deve conter um job principal que, por sua vez, deve conter as transformações criadas nas questões 8, 9 e 10. Além disso, que tal ganhar um ponto a mais nessas questões? Para isso, inclua o projeto criado em um repositório do Github (é importante que seja público para termos visibilidade, ok?). Compartilhe por aqui o link para o repositório.

   Seguem as questões:

8. Construa uma transformação que deve usar como datasource o dataset (DATASET_TESTE_DE.csv) que contém informações de coletas de dados nos pontos de vendas. A ETL deve consultar o dataset e inserir, em uma base de dados (modelo dimensional), as informações coletadas, conforme as tabelas abaixo:

   (a) Dimensão Calendário (DIM_CALENDARIO): Deve conter data, mês e ano da coleta.
   (b) Dimensão Ponto de Venda (DIM_PDV): Deve conter o id, nome e perfil do ponto de venda.
   (c) Dimensão Linha de Produto (DIM_LINHA_PRODUTO): Deve conter o id, nome e perfil da linha de produto.

9. Construa uma transformação que deve usar como datasource o dataset (DATASET_TESTE_DE.csv) que contém informações de coletas de dados nos pontos de vendas. A transformação deve consultar o dataset e inserir, em uma base de dados (modelo dimensional), as informações coletadas, conforme as tabelas abaixo:

   (a) Fato Disponibilidade (FT_DISPONIBILIDADE): Deve conter os ids de ligação das tabelas de dimensões criadas na questão anterior e a quantidade de presenças de cada linha de produto no mês de Setembro/20.
   (b) Fato Disponibilidade Agregada (FT_DISPONIBILIDADE_AGREGADA): Deve conter os ids de ligação das tabelas de dimensões (Dimensão Calendário e Ponto de Venda) e a quantidade de presença de linhas de produto agrupadas por ponto de venda no mês de Setembro/20.

Obs: Os dados de "Disponibilidade" estão categorizados na coluna TIPO_COLETA com o valor "Disponibilidade". A presença é contada sempre que no campo VALOR aparecer o valor "SIM".

10. Construa uma transformação que deve usar como datasource o dataset (DATASET_TESTE_DE.csv) que contém informações de coletas de dados nos pontos de vendas. A transformação deve consultar o dataset e inserir, em uma base de dados (modelo dimensional), as informações coletadas, conforme as tabelas abaixo:

    (a) Fato Ponto Extra (FT_PONTO_EXTRA): Deve conter os ids de ligação das tabelas de dimensões criadas na questão anterior e a soma de pontos extras de cada linha de produto do mês de Setembro/20.

    (b) Fato Ponto Extra Agregada (FT_PONTO_EXTRA_AGREGADA): Deve conter os ids de ligação das tabelas de dimensões (Dimensão Calendário e Ponto de Venda) e a soma de pontos extras de linhas de produto agrupadas por ponto de venda no mês de Setembro/20.

    Obs: Os dados de "Ponto Extra" estão categorizados na coluna TIPO_COLETA com o valor "Ponto Extra".

2 Respostas

1. (a) Um Data Warehouse trata-se de um repositório central de dados de uma organização, isto significa que, as diversas fontes de dados das diferentes áreas de uma corporação são unificadas. Isso garante que se tenha uma versão única da "verdade", ou seja, certifica-se que o mesmo conceito, para um determinado indicador ou assunto de negócio, terá o mesmo significado para as diferentes áreas sejam elas financeira, produção, contabilidade etc. Um Data Warehouse é orientado ao negócio e, portanto, a sua função é auxiliar na tomada de decisões. Para que isso seja possível, este repositório central armazena os dados históricos da organização e a sua modelagem é baseada no que chamamos de modelo dimensional. Um Data Warehouse armazena dados estruturados em tabelas chamadas dimensões e fatos e utiliza-se de SGBDs relacionais para armazenamento dos dados.

   (b) As tabelas Fato amarzenam, como o próprio nome diz, o fato em si, o que ocorreu. Todas as medições do negócio que se deseja acompanhar são armazenadas na Fato. Como exemplo podemos citar uma Fato Venda. Essa Fato seria responsável por armazenar as métricas referentes ao acontecimento venda. Essas métricas poderiam ser, quantidade de produtos vendidos, lucro, desconto, total em valor etc. Já as dimensões são responsáveis por qualificar o fato, ou seja, são informações descritivas sobre o que ocorreu de modo que tenhamos várias visões sobre aquele acontecimento. Seguindo essa linha de raciocínio poderíamos ter uma dimensão de produto, que atrelada à fato, nos informaria que produto foi vendido naquela determinada transação, um dimensão de tempo, nos dando uma visão do espaço temporal em que aquela venda foi efetuada, dimensão cliente, dando-nos características sobre aquele cliente em específico e por aí poderíamos continuar definindo quais seriam nossas dimensões. A tabela fato está sempre ligada a duas ou mais dimensões.

   (c) ETL é um dos tipos de integração de dados que podemos aplicar em um processo. O termo é um acrônimo para **Extract, Transform, Load**, ou seja, é o processo em que se extrai os dados de uma ou mais fontes, sejam elas bancos OLTPs, fonte externas à organização, como redes sociais, APIs, entre outras, ou até mesmo arquivos de textos, planilhas etc. Uma vez extraídos, os dados podem ser transformados de acordo com alguma regra de negócio, podem ser, também, normalizados para que os dados de diversas fontes possam ser integrados. Nesta etapa também podemos calcular as métricas derivativas, isto é, métricas que são originadas de outras métricas já presentes nas fontes. O último passo é então, uma vez extraído e trabalhado o dado, carregá-lo em um data warehouse ou em algum outro banco.

   (d) Um engenheiro de dados é o profissional responsável pela coleta de dados de fontes externas ou internas à organização, fluxo de dados, armazanemaento e transforamção de dados, ou seja, ele deve garantir toda a infraestrutura de dados de forma a propiciar que os analistas e cientistas de dados possam exercer o trabalho deles. Esta infraestrutura compreende os data warehouses, data lakes etc. Portanto, este profissional tem a função de assegurar que os dados estarão disponíveis e com qualidade.

   (e) O Trade Marketing é uma estratégia de marketing B2B, ou seja, de empresa para empresa, que visa estabelecer uma relação de ganho mútuo entre a marca e os canais de distribuição como os atacados, distribuidoras etc. O setor de Trade Marketing atua nos pontos de venda dos canais de distribuição a fim de tentar incentivar a compra de seu produto pelo cliente seja por meio de um melhor posicionamento dos produtos nas gôndolas do mercado, seja por promoções ou outros meios. Resumidamente, portanto, o trade marketing visa o melhor posicionamento da marca nos pontos de venda e determina em quais destes é interessante que o produto da marca esteja presente.

2. Segue a query abaixo

```
1    SELECT *
```

```
    FROM PONTO_VENDA_UNIDADE
    WHERE SELLIN > 20000
    ORDER BY NOME_PDV ASC;

```

3. Segue a query abaixo

```
    SELECT
        ANO
    ,   MES
    ,   NOME_PDV
    ,   SUM(SELLIN) AS TOTAL_SELLIN
    FROM PONTO_VENDA_UNIDADE
    GROUP BY
        ANO
    ,   MES
    ,   NOME_PDV
    ORDER BY
        ANO ASC
    ,   MES ASC
    ,   NOME_PDV ASC;

```

4. Segue a query abaixo

```
    SELECT
        P.NOME_PDV
    ,   SUM(V.FL_VISITADO) AS QTD_VISITAS
    FROM PONTO_VENDA_UNIDADE AS P
    LEFT JOIN VISITAS_PONTO_VENDA AS V ON(
        P.ID_PDV = V.FK_PDV
    )
    WHERE P.NOME_PDV = "INVOLVES"
    GROUP BY P.NOME_PDV;

```

5. Os campos que poderiam ser indexados a fim de otimizar o tempo de resposta da consulta são aqueles que são utilizados para realizar o join entre as tabelas e os utilizados na condição *WHERE*. Sendo assim, seriam os seguintes campos:

   - FT.ID_DIM_PDV;
   - TPLL.ID_DIM_PDV;
   - FT.CICLO;
   - TPLL.CICLO;
   - FTPI.CICLO;
   - FTPI.ID_DIM_PDV;
   - FT.ID_BLOCO_ITEM;
   - FTPI.ID_BLOCO_ITEM;
   - FT.SEMANA_LP;
   - FTPI.SEMANA_LP.

6. Após executar o comando abaixo no python, a variável x conterá uma lista com cinco posições cujos valores são *None*, ou seja, há ausência de valor. Isso é ocasionado pelo fato de que a instrução não armazena nenhum valor na lista e apenas imprime na tela os valores múltiplos de 2 que se encontram dentro do intervalo de 0 a 9.

```
x = [print(i) for i in range(10) if i % 2 == 0]

```

7. Segue abaixo um script python que pede dois números ao usuário e imprime na tela a soma dos números escolhidos.

```

print("Bem vindo! \nO programa em questão é uma calculadora que executa apenas a soma
entre dois números", end="\n\n")

num1 = float(input("Insira o primeiro número: "))
num2 = float(input("Insira o segunda número: "))

print(f"Os números escolhidos foram {num1} e {num2}")
print(f"A soma entre os dois números é tal que: \n\n{num1} + {num2} = {num1 + num2}")

```

Obs: Foi elaborado, também, um script mais completo que simula uma calculadora simples. O script acima, bem como o mais completo, se encontra no meu repositório git. Para acessá-lo clique aqui!

2.1 Questões 8, 9 e 10

As transformações e jobs desenvolvidos para as questões 8, 9 e 10 se encontram no meu repositório git. Para acessá-lo clique aqui!