Atividade Técnica

Continuação...

→ Pergunta 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 (cam

A	B	C	D	E	F	G	H	I	J
ID_DATASET	ID_PONTO_VENDA	NOME_PONTO_VEI	PERFIL_PONTO_VE	DATA	ID_LINHA_PRODUT	NOME_LINHA_PROI	MARCA_LINHA_PRO	TIPO_COLETA	VALOR
1021	115	INVOLVES	ATACADAO	01/09/2020	398	BISCOITOS SORTID	INVOLVES ZERO	DISPONIBILIDADE	SIM
1022	115	INVOLVES	ATACADAO	02/09/2020	407	MARGARINA	INVOLVES ZERO	DISPONIBILIDADE	SIM
1023	115	INVOLVES	ATACADAO	03/09/2020	408	MANTEIGA	INVOLVES ZERO	DISPONIBILIDADE	SIM

O dataset enviado anexo ao email tem o nome **dataset_teste_de.csv**. Entendo que se trata do mesmo citado acima na descrição da questão, **PONTO_VENDA_UNIDADE**. Conforme o print da imagem acima, não há o campo **SELLIN** na tabela, pode ser que seja a coluna **VALOR**, mas não tenho certeza.

O dataset enviado tem apenas 40 linhas/registros, sendo que na coluna VALOR há campos 'SIM/NÃO' do tipo texto (20 linhas) e campos do tipo inteiro (min 0, max 6). O máximo da tabela é bem inferior ao solicitado na questão (20.000); portanto, mesmo que eu limpe e deixe apenas os registros de número inteiro, a tabela retorna vazia.

Pergunta2 RUN SAVE QUERY DOWNLOAD SHARE This query wi

```
SELECT *
FROM `involves-422612.involves_datasets.involves-dataset`
WHERE VALOR NOT IN ('SIM', 'NÃO') AND CAST(VALOR AS INT64) > 20000
ORDER BY NOME_PONTO_VENDA;
```

Query results SAVE RESULTS EXPLORE DATA

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

There is no data to display.

→ Pergunta 3

Pergunta3_distintos

RUN SCHEDULE MORE SAVE QUERY DOWNLOAD

```
SELECT DISTINCT (ID_PONTO_VENDA,NOME_PONTO_VENDA)
FROM `involves-422612.involves_datasets.involves-dataset`
```

Query results

SAVE RESULTS EXPLORE DATA

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	f0_._field_1	f0_._field_2
1	115	INVOLVES
2	116	INVOLVES STAGE

Conforme consulta acima, o campo NOME_PONTO_VENDA possui **apenas 2 valores distintos**, assim como ID_PONTO_VENDA. Por este motivo a consulta solicitada na questão retorna apenas a soma de VALOR para dois pontos de vendas.

Pergunta3

RUN SCHEDULE MORE SAVE QUERY DOWNLOAD SHARE

```
SELECT
  FORMAT('%02d%d', EXTRACT(MONTH FROM DATA), EXTRACT(YEAR FROM DATA)) AS ID_MES_ANO,
  NOME_PONTO_VENDA,
  SUM(CAST(VALOR AS INT64)) AS TOTAL_VALOR,
  DATA,
  EXTRACT(YEAR FROM DATA) AS ANO,
  EXTRACT(MONTH FROM DATA) AS MES,
  FROM `involves-422612.involves_datasets.involves-dataset`
WHERE VALOR NOT IN ('SIM', 'NAO')
GROUP BY NOME_PONTO_VENDA, DATA, EXTRACT(YEAR FROM DATA), EXTRACT(MONTH FROM DATA)
ORDER BY ID_MES_ANO ASC, NOME_PONTO_VENDA ASC
```

Query results

SAVE RESULTS EXPLORE DATA

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	ID_MES_ANO	NOME_PONTO_VENDA	TOTAL_VALOR	DATA	ANO	MES
1	092020	INVOLVES	22	2020-09-01	2020	9
2	092020	INVOLVES STAGE	25	2020-09-01	2020	9

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 indica 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 visitas realizadas.

→ Pergunta 4

Não foi enviada uma tabela de nome VISITAS_PONTO_VENDA, entao criei com dados aleatórios, para garantir que a consuta funcionaria.

Inserir_Visitas

RUN SAVE QUERY DOWNLOAD SHARE SCHEDULE MORE

```sql
--Criar tabela
CREATE TABLE `involves-422612.involves_datasets.involves-dataset-visitas` (
  ID_VISITA INT64 NOT NULL,
  FK_PDV INT64 NOT NULL,
  FL_VISITADO INT64 NOT NULL,
  DATA_VISITA DATE NOT NULL
);

-- Se necessario limpar a tabela
TRUNCATE TABLE `involves-422612.involves_datasets.involves-dataset-visitas`

-- Inserir dados na tablea
INSERT INTO
  `involves-422612.involves_datasets.involves-dataset-visitas` (ID_VISITA,
    FK_PDV,
    FL_VISITADO,
    DATA_VISITA)
SELECT
  id AS ID_VISITA,
  CAST(115 + FLOOR(RAND() * 2) AS INT64) AS FK_PDV,
  CAST(FLOOR(RAND() * 2) AS INT64) AS FL_VISITADO,
  DATE_ADD('2024-01-01', INTERVAL CAST(FLOOR(RAND() * 365) AS INT64) DAY) AS DATA_VISITA,
FROM
  UNNEST(GENERATE_ARRAY(1, 100)) AS id;
```

Resposta:

Pergunta4

RUN SAVE QUERY DOWNLOAD SHARE SCHEDULE MORE This que...

```sql
SELECT
  PDV_UNIDADES.NOME_PONTO_VENDA,
  COALESCE(VISITAS_COUNT.TOTAL_VISITAS, 0) AS TOTAL_VISITAS
FROM
  `involves-422612.involves_datasets.involves-dataset` AS PDV_UNIDADES
LEFT JOIN (
  SELECT
    FK_PDV,
    COUNT(ID_VISITA) AS TOTAL_VISITAS
  FROM
    `involves-422612.involves_datasets.involves-dataset-visitas`
  WHERE
    FL_VISITADO = 1
  GROUP BY
    FK_PDV) AS VISITAS_COUNT
ON
  PDV_UNIDADES.ID_PONTO_VENDA = VISITAS_COUNT.FK_PDV
WHERE
  NOME_PONTO_VENDA = 'INVOLVES'
GROUP BY
  PDV_UNIDADES.NOME_PONTO_VENDA,
  TOTAL_VISITAS;
```

Query results

SAVE RESULTS EXPLORE DATA

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	NOME_PONTO_VENDA	TOTAL_VISITAS
1	INVOLVES	22

→ **Pergunta 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 possa performar melhor. Leve em consideração que nenhum campo no banco de dados está indexado.

A indexação agiliza o encontro entre dois campos e evita a necessidade de uma consulta ser realizada linha por linha. Vários campos podem ser indexados, mas a prioridade na consulta apresentada deve ser para os campos que estão nas junções, nos filtros e nos agrupamentos. Abaixo, podemos observar que os campos FT.CICLO e FT.ID_DIM_PDV estão presentes em todas as junções e nos filtros. Portanto, são eles que precisam ser indexados com urgência, inclusive com a possibilidade de ser criado um índice composto.

Untitled query RUN SAVE DO

```
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;
```

→ Pergunta 6

Considere a instrução Python a seguir
no Python , a variável "x" conterá qual

Involves.ipynb

File Edit View Insert Runtime Tools Help Saving...

+ Code + Text

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

```
0
2
4
6
8
```

→ Pergunta 7

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

```
[11] # @title Default title text
     # Apresenta
     print ('Olá, vamos fazer a soma de dois números')

     # Pede um 1º número
     num1 = float(input("Digite o 1º número: "))

     # Pede um 2º número
     num2 = float(input("Digite o 2º número: "))

     # Calcula a soma dos dois números inseridos
     soma = num1 + num2

     # Imprime o resultado
     print("O resultado da soma é:", soma)
```

```
Olá, vamos fazer a soma de dois números
Digite o 1º número: 50
Digite o 2º número: 150
O resultado da soma é: 200.0
```

Para responder às questões 8, 9 e 10
A ETL final deve conter um job princ
10. Além disso, que tal ganhar um ponto a mais nessas questões ? Para isso, inclua o projeto criado em um repositório do Github (e importante que seja público para termos visibilidade, ok ?). Compartilhe por aqui o link para o repositório.

→ Pergunta 8

Construa uma transformação que deve usar como datasource o dataset (DATASET_TESTE_DE.csv) que contém informações de coletas de dados nos ponto de vendas. A ETL deve consultar o dataset e inserir, em uma base de dados (no bd-dimensional) as informações coletadas conforme estrutura 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

A pergunta a) não especifica o formato do ano e mês (ex. YY, YYYY-MM, MMM, etc..), deixei o padrão da função. Não encontrei o campo *perfil da linha de produto, adicionei MARCA_LINHA_PRODUTO*

Pergunta8_DIM_CALENDARIO

RUN SAVE QUERY

```
SELECT
  DATA,
  EXTRACT(YEAR
  FROM
    DATA) AS ANO,
  EXTRACT(MONTH
  FROM
    DATA) AS MES
FROM
  `involves-422612.involves_datasets.involves-dataset`
```

Pergunta8_DIM_LINHA_PRODU...

RUN SAVE Q

```
SELECT
  ID_LINHA_PRODUTO,
  NOME_LINHA_PRODUTO,
  MARCA_LINHA_PRODUTO
FROM
  `involves-422612.involves_datasets.involves-dataset`
```

Pergunta8_DIM_PDV

RUN SAVE QUERY DOWNLOAD

```
SELECT
  ID_PONTO_VENDA,
  NOME_PONTO_VENDA,
  PERFIL_PONTO_VENDA
FROM
  `involves-422612.involves_datasets.involves-dataset`
```

→ Pergunta 9

Construa uma transformação que deve usar como datasource o dataset (DATASET_TESTE_DE.csv) que contém informações de coletas de dados nos ponto 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 criadas na questão anterior e a quantidade de presenças de cada linha de produto no mês de Setembro/20, agrupado por ponto de venda no mês de Setembro/20.

Não agreguei por DATA, porque conforme descricao apenas os dados de setembro são nevessarios.

Pergunta9_FT_DISPONIBILIDADE RUN

```sql
SELECT
  ID_PONTO_VENDA,
  ID_LINHA_PRODUTO,
  COUNT(*) AS QTD_PRESENCA
FROM
  `involves-422612.involves_datasets.involves-dataset`
WHERE
  EXTRACT(YEAR
  FROM
    DATA) = 2020
  AND EXTRACT(MONTH
  FROM
    DATA) = 9
  AND VALOR = 'SIM'
GROUP BY
  ID_PONTO_VENDA,
  ID_LINHA_PRODUTO
ORDER BY
  ID_LINHA_PRODUTO,
  ID_PONTO_VENDA,
  QTD_PRESENCA DESC;
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON

Row	ID_PONTO_VENDA	ID_LINHA_PRODUTO	QTD_PRESENCA
1	115	398	2
2	115	407	1
3	116	407	2
4	115	408	1
5	115	422	1
6	116	422	2
7	115	423	2

Pergunta9_FT_DISPONIBILIDADE_AGREGA...

```sql
SELECT
  ID_PONTO_VENDA,
  COUNT(*) AS QTD_PRESENCA
FROM
  `involves-422612.involves_datasets.involves-dataset`
WHERE
  EXTRACT(YEAR
  FROM
    DATA) = 2020
  AND EXTRACT(MONTH
  FROM
    DATA) = 9
  AND VALOR = 'SIM'
GROUP BY
  ID_PONTO_VENDA
ORDER BY
  ID_PONTO_VENDA,
  QTD_PRESENCA DESC;
```

Query results

JOB INFORMATION | RESULTS | CHART

Row	ID_PONTO_VENDA	QTD_PRESENCA
1	115	7
2	116	4