AVALIAÇÃO – CIENTISTA DE DADOS I

Orientações:

• Responder a todas as questões abaixo, detalhando as etapas, processos realizados e ferramentas utilizadas.

• Disponibilizar as respostas através de um arquivo de texto em uma plataforma de armazenamento em nuvem (Google Drive, OneDrive, WeTransfer, etc.) ou através de um link de repositório de compartilhamento de projetos de desenvolvimento (Git ou GitHub).

Importante: Todas as questões foram consultadas utilizando inteligências artificiais generativas, portanto, as respostas serão analisadas considerando aspectos de autenticidade, clareza e resolução.

Questões:

1. Você precisa integrar e manter atualizada uma base de comércio exterior de 143 países. Para isso, você utilizará uma API onde pode realizar 2500 consultas diárias. Os parâmetros de consulta são: país_origem, país_destino e ano, exemplo:

(BRA, ARG, 2019) **>>** importações e exportações feitas do Brasil para Argentina em 2019
(BRA, ARG, 2020) **>>** // em 2020
(BRA, AFG, 2019) **>>** imp/exp feitas do Brasil para Afeganistão em 2019
(BRA, AFG, 2020) **>>** // em 2020
(BRA, WLD, 2019) **>>** imp/exp feitas do Brasil para o Mundo em 2019
(BRA, WLD, 2020) **>>** // em 2020
(BRA, BRA, 2019) **>> <consulta inválida>** -> imp/exp feitas do Brasil para Brasil em 2019
(WLD, BRA 2019) **>> <consulta inválida>** -> imp/exp feitas do Mundo para Brasil em 2019

Os dados serão utilizados em um painel que responde, por exemplo, as seguintes questões:

Qual a quantidade de soja o Brasil exportou para China em 2020?

(BRA, CHN, 2020)

Qual a quantidade de soja que a China importou do Brasil em 2020?

(CHN, BRA, 2020)

Qual a quantidade de soja o Brasil exportou para o Mundo em 2020?

(BRA, WLD, 2020)

Qual a quantidade de soja o Mundo importou do Brasil em 2020?

<Questão 1.3>

Sua tarefa é criar:

1.1 - Uma estrutura de diretórios e nomenclatura de arquivos que permita armazenar as consultas de modo coerente, ex:

Haverão diretórios: BRA, ARG, AFG, ...

E arquivos: BRA.txt, ARG.txt, AFG.txt, ...

Com isso, o arquivo /BRA/ARG.txt representa a imp/exp feita do BRA para ARG

***Obs: A estrutura listada pode ser insuficiente/incorreta**

1.2 - Liste as etapas necessárias para integração de 3 anos (2019~2021) de comércio exterior, ex:

1 - Fazer request em todos paises, parceiros e anos

<opcional>

#itera sobre 143 países origem
for pais_origem in lista_paises:

#itera sobre os 143 países destino

```
    for pais_destino in lista_paises:

        #itera sobre os anos
        for ano in ['2019', '2020', '2021']:

            #consulta
            res = get(pais_origem, pais_destino, ano)
```

</opcional>

***Obs: Esse exemplo é insuficiente/incompleto**

2- Salvar arquivos

<opcional>

res.write(f'{pais_origem}/{pais_destino}.txt') *#ex: BRA/ARG.txt*

</opcional>

***Obs: Embora a estrutura listada em (1) tenha sido preservada não funcionará corretamente. Etapas adicionais (ou uma estrutura distinta) são necessárias.**

***Obs2: Note que os códigos são apenas abstrações. Não se faz necessário detalhamento: explicitar função get; garantir que objeto res possua método write e assim por diante.**

1.3 - Qual/Quais arquivo(s) da sua base de dados respondem a questão: 'Qual a quantidade de soja o Mundo importou do Brasil em 2020?'

A resposta irá depender da estrutura proposta em (1.1)

1.4 – Sua base de dados alimenta um painel que possui informações de comércio exterior entre países – e também mundo. As consultas de comércio exterior podem ser feitas com os parâmetros M (‘mês’) ou A(‘ano’). A consulta **M** representa um report dos meses disponíveis no ano até então. A consulta **A** representa o report **final/completo** das imp/exp em determinado ano. Considere que cada país escolhe uma data “aleatória” do primeiro trimestre do novo ano para realizar o report **final/completo** das suas exportações e importações anuais do ano anterior. Para o melhor entendimento, considere as seguintes consultas:

<Consultas feitas em 1 julho 2024>

(BRA, CHN, 2024, 'A') **>> retorna vazio, pois ainda não houve report final**

(BRA, CHN, 2024, 'M') **>> retorna o imp/exp nos meses disponíveis**

<Consultas feitas em 15 março 2025>

(BRA, CHN, 2024, 'A') **>> retorna report final/completo do ano de 2024**

(BRA, CHN, 2024, 'M') **>> retorna meses disponíveis (porém é incompleto)**

Vale ressaltar que o report **final/completo** pode ser diferente do mensal mesmo após o término do ano. Isso porque o país pode escolher reportar mensalmente até o penúltimo trimestre do ano e depois fazer apenas o report anual. Isso faria com que os dataset de todos os meses disponíveis em 2024 seja distinto do report anual do mesmo ano.

Com isso em mente, liste as etapas necessárias para manter as bases atualizadas de modo recorrente: considere que estamos no dia **1 de fevereiro de 2025.**

---

2. Considere os arquivos públicos da Receita Federal disponíveis em: http://200.152.38.155/CNPJ/. A página é atualizada, na média, 1 vez ao mês: os nomes dos arquivos são fixos.

Suponha que você precise automatizar o processo de baixar, empilhar e manter atualizado os arquivos de empresas (0~9). Considere que o site é instável, portanto, o arquivo baixado pode estar corrompido.

Liste as etapas necessárias para que o dataset de empresas esteja sempre atualizado.

ex:

1 - Baixar os arquivos de empresas, **<opcional>** usando requests **</opcional>**
2 - Extrair arquivos de empresas, **<opcional>** usando zipfile **</opcional>**
3 - Empilhar os arquivos, **<opcional>** usando pandas.concat **</opcional>**
4 - ...

Obs1: Não é necessário codar!

Obs2: O exemplo acima é insuficiente para garantir a criação do dataset e a atualização contínua do mesmo

Name	Last modified	Size	Description
Parent Directory		-	
Cnaes.zip	2024-06-09 19:46	22K	
Empresas0.zip	2024-06-09 20:37	326M	
Empresas1.zip	2024-06-09 20:42	74M	
Empresas2.zip	2024-06-09 20:44	75M	
Empresas3.zip	2024-06-09 20:46	81M	
Empresas4.zip	2024-06-09 20:48	86M	
Empresas5.zip	2024-06-09 20:49	93M	
Empresas6.zip	2024-06-09 20:37	90M	
Empresas7.zip	2024-06-09 20:39	95M	
Empresas8.zip	2024-06-09 20:39	95M	
Empresas9.zip	2024-06-09 20:41	91M	

---

3. Escreva uma consulta SQL que retorne, para cada funcionário, o nome do departamento em que ele trabalhou pela primeira vez (baseado na data de início de trabalho), o nome do departamento onde ele está atualmente trabalhando, e a quantidade de departamentos diferentes em que ele já trabalhou. Use as tabelas employees, departments, e employee_department_history.

Estrutura das tabelas:

- **employees**
  - employee_id (PK)
  - employee_name
- **departments**
  - department_id (PK)
  - department_name
- **employee_department_history**
  - employee_id (FK)
  - department_id (FK)
  - start_date
  - end_date ('NULL' para o departamento atual)

---

4. O arquivo exemplo1.parquet possui 7GB. O tempo de leitura e retorno de 1 linha específica são 5s - utilizando processamento paralelo (pyspark) em cluster. Contudo, você precisa criar uma API com recursos bem mais modestos, isto é, menor poder de processamento - mantendo o tempo de consulta rápido. Liste as alterações e procedimentos necessários para realizar isso.

Obs: se necesserário, considere que o arquivo exemplo1.parquet atualiza 1 vez ao mês

---

5. Você foi instruído a criar um modelo de regressão baseado em um modelo econométrico. Ao terminar a implementação você verificou que o resultado está abaixo do esperado. Liste o que poderia ser feito para melhorar o resultado.

---

6. ELT e ETL são termos comuns na rotina de trabalho de especialistas em Soluções de Tecnologia da Informação, justamente por representarem estratégias de pipelines de integração de dados em um determinado projeto. Sobre esses procedimentos, é correto afirmar que:

a) Ambos necessitam de uma etapa de cópia dos dados em um armazenamento intermediário, antes de serem transferidos para o banco de dados consolidados.
b) Modelos locais, que utilizem dados relacionais e estruturados são ideais para estratégias ELT, por apresentarem menores custos e necessidades de hardware.
c) Estratégias ETL realizam a etapa de tratamento diretamente no Banco de Dados Consolidado.
d) Estrutura de dados em nuvem são ideais para a adoção de estratégias ELT, devido a maior rapidez no carregamento dos dados e a adequada capacidade de processamento posterior.
e) Pipelines de integração do tipo ETL é indicada para casos em que a organização utilizada dados não-estruturados.