Avaliação de conhecimentos

Engenharia de Dados & DataOps

Nome:_____ Data:__/__/__

Instruções gerais:

A avaliação está dividida em 3 partes:

- **Manipulação e análise de dados**
- **Programação & Tratamento de dados**
- **Inglês**

A avaliação foi dimensionada para ser resolvida em aproximadamente 120 minutos (2 horas), e servirá para conhecermos seu nível de conhecimento em Engenharia de Dados e DataOps, e não é eliminatório durante o processo. Então fique tranquilo(a), use todo seu conhecimento, justifique bem as respostas, e tente resolver tudo, mesmo se não souber.... esse ponto é bem importante na avaliação. Lembre-se de ser sucinto e direto nas respostas.

Uso de tradutores ou softwares que forneçam respostas automáticas serão consideradas como plágio e, uma vez detectado, o candidato automaticamente será eliminado do processo seletivo.

A devolução deve ocorrer digitalizada (anexada no e-mail) em até 24 horas.

Orientações:

- As questões 1 e 2 devem ser solucionadas utilizando a linguagem de programação de sua preferência. Ou seja, não serão aceitas manipulações manuais ou análises via excel.
- A questão 2 deve ser entregue através do link do github pessoal.
- Todo o código desenvolvido deve estar comentado com as explicações de cada operação desenvolvida

Boa sorte!

Queremos você em nosso time Maxxidata.

Parte 1 – Manipulação e análise de dados

A partir das tabelas *cadastro.csv* e *vendas.csv*, resolva as seguintes questões:

i. Qual a pessoa (cod_cadastro) que gastou mais? e a que gastou menos?
ii. Qual a região de procedência que gasta mais?
iii. Qual o produto que mais é vendido em quantidade?
iv. Existe alguma característica da que identifique o grupo das pessoas (top 5) que compram mais produtos em quantidade? E as que gastam mais dinheiro?
v. Você consegue ver alguma relação entre o grau de instrução e o número de filhos? Descreva a sua interpretação do caso, utilizando os dados para justificar as suas conclusões.
vi. Imagine que a empresa em questão queira desenvolver um programa de cashback de benefícios baseada no perfil dos clientes, sendo assim, o primeiro passo seria desenvolver este perfil. Proponha um critério para classificar os clientes em clientes *diamante*, *ouro* e *prata* . Justifique a sua resposta.

Parte 2 – Programação & Tratamento de Dados

Você deve realizar uma ingestão de dados da https://swapi.dev/. Esta é uma api que contém os dados dos filmes do star wars.

- swapi, swapi infos, swapi documentation
- https://swapi.dev/api/
- rotas das informações a **serem ingeridas** da swapi:
    - https://swapi.dev/api/people/?
    - https://swapi.dev/api/planets/?
    - https://swapi.dev/api/films/?

Requisitos do projeto:

1. Esta api é paginada, sendo assim, para este exercício deve-se fazer ingestão das 5 **primeiras páginas.** As bases ingeridas (people, planets, films) devem ser salvas em uma pasta chamada “raw’ em formato csv.
2. Para cada base de dados (people, planets, films) deve ser realizada os seguintes tratamentos:
    a. Padronização de strings para lower case;
    b. Remoção de caracteres especiais.
3. As bases finais, ou seja, as que foram aplicadas os tratamentos do item 2, devem ser salvas em uma pasta chamada “work” em formato csv.
4. **Não será aceito scripts feitos em jupyter notebook.**
5. Este projeto deve ser entregue através de um link do github. **Não será aceito outro formato.**

Parte 3 – Inglês

Leia o texto e responda às perguntas em inglês (a resposta deve ser redigida inglês).

What is a data lake?

A data lake is a central storage repository that holds big data from many sources in a raw, granular format. It can store structured, semi-structured, or unstructured data, which means data can be kept in a more flexible format for future use. When storing data, a data lake associates it with identifiers and metadata tags for faster retrieval.

Coined by James Dixon, CTO of Pentaho, the term "data lake" refers to the ad hoc nature of data in a data lake, as opposed to the clean and processed data stored in traditional data warehouse systems.

Data lakes are usually configured on a cluster of inexpensive and scalable commodity hardware. This allows data to be dumped in the lake in case there is a need for it later without having to worry about storage capacity. The clusters could either exist on-premises or in the cloud.

Data lakes are easily confused with data warehouses, but feature some distinct differences that can offer big benefits to the right organizations especially as big data and big data processes continue to migrate from on-premises to the cloud. They are similar in their basic purpose and objective, which make them easily confused:

- Both are storage repositories that consolidate the various data stores in an organization.
- The objective of both is to create a one-stop data store that will feed into various applications.

However, there are fundamental distinctions between the two that make them suitable for different scenarios.

- Schema-on-read vs schema-on-write — The schema of a data warehouse is defined and structured before storage (schema is applied while writing data). A data lake, in contrast, has no predefined schema, which allows it to store data in its native format. So in a data warehouse most of the data preparation usually happens before processing. In a data lake, it happens later, when the data is actually being used.
- Complex vs simple user accessibility — As data is not organized in a simplified form before storage, a data lake often needs an expert with a thorough understanding of the various kinds of data and their relationships, to read through it. A data warehouse, in contrast, is easily accessible to both tech and non-tech users due its well-defined and documented schema. Even a new member on the team can begin to use a warehouse quickly.
- Flexibility vs rigidity — With a data warehouse, not only does it take time to define the schema at first, it also takes considerable resources to modify it when requirements change in the future. However, data lakes can adapt to changes easily. Also, as the need for storage capacity increases, it is easier to scale the servers on a data lake cluster.

For more on this distinction, and to help determine which is best for your organization, see "Data Lakes vs Data Warehouses". There is also an emerging open data management architecture that combines the flexibility of a data lake with the data management capabilities of a data warehouse, known as a data lakehouse.

Questions

i) In your opinion what would be the best data storage (data lake or data warehouse) to store data from different formats (files, images, etc…) and sources (data bases, api, sensors, etc..)? Justify your answer.

ii) For a non-technical user what would be the best data storage (data lake or data warehouse) to adopt? Justify your answer.

iii) In your opinion which data storage (data lake or data warehouse) is the most sensitive to schema changes?