Projeto Final

BIG DATA ANALYTICS.

Você foi contratado como consultor em Big Data Analytics pelo ministério de saúde dos Estados Unidos par analisar os mais recentes dados da COVID-19. Os seus 2 grandes objetivos são um procedimento de ingestão, transformação e carregamento dos dados (Extract Transformation and Load); e a outro é a análise dos dados.

Primeira Parte (ETL)

1. Como consultor em Analytics, a primeira tarefa é criar um pipeline para carregar aos dados que contêm a informação de cada doente. É o seu objetivo e criar um procedimento para receber ficheiros CSV e carregá-los diretamente no snowflake.

Para atingir dito objetivo o HHS(departamento de saúde dos Estados Unidos) , pede que o seu código seja reutilizável porque o processo que você vai criar será usado para ingerir mais dados posteriormente. Crie uma definição para cada um dos processos.

Requisitos

Crie 1 objeto. uma classe com 3 funções:

Classe:

Definição init: **__init__:** parâmetros: parâmetros de ligação ao snowflake. O resultado desta função tem indicar se a conexão foi estabelecida ou não.

Primeira definição: ler tabela desde snowflake: Tem de usar os parâmetros definidos na função __init__. Esta função tem de ter a capacidade de ler uma query especifica ou uma tabela completa. Se houver algum error na leitura da tabela tem de apresentar qual é o erro específico pelo qual

não pode ser lida a tabela ou a query.

Segunda de definição: uma defingão que permita escrever qualquer spark dataframe a Snowflake.

Se for bem sucedido o output da função tem de ser um dict com a seguinte informação:

- o tempo que tardou em criar a tabela em segundos
- schema,
- nome da tabela
- Numero de colunas
- Nome das colunas
- Número de linhas

Ex:

{

Tempo total transcorrido: 25,

“Schema”: “schema onde foi gravada a tabela(tem de ser dinâmico)”,

“Tabela”: “nome da tabela”,

“Numero de columnas”: 34

“nome das colunas”: [“A”,“B”,“C”,“D”,“E”,“F”,“G”,]

“ Numero de linhas”: 12345678

}

Se falhar a escrita, tem de se indicar qual foi o erro específico pelo qual falho a escrita da tabela.

Ajuda:

https://www.w3schools.com/python/python_classes.asp

https://www.w3schools.com/python/python_functions.asp

Segunda Parte: Uso de Pyspark

O Ministério de saúde quer que você analise a informação que tem em ***pyspark*** para que depois, se a analise for realmente prometedora, possam ser integrados mais dados, e sem importar o volume destes, possam ser analisados os novos dados. Se for preciso usar algum dataset previamente carregado no Snowflake faça a ligação usando o desenvolvido no primeiro ponto não diretamente com o ficheiro csv.

O Departamento De saúde dos Estados Unidos conta consigo.

1. Qual é quantidade de pessoas do género feminino e masculino e a sua percentagem sobre o total de doentes?
   1.1. Crie uma visualização com esta informação (gráfico de barras)
2. Identifique se existe informação de doentes com data de nascimento superior à data de morte.
3. Calcule a idade(em anos) das pessoas usando as seguintes condições:
   3.1. Se o estiver morto, essa será a data final para calcular a idade
   3.2. Se estiver vivo, considere como data final, 2020-04-05 para o cálculo da idade
4. Identifique a idade máxima, idade mínima, a média, mediana[1] e máximo.
5. Faça um histograma com 100 bins (intervalos) da idade das pessoas.

[1] Lembre que o percentil 50 é considerado a mediana. Para mais informação siga este link: https://spark.apache.org/docs/3.1.1/api/python/reference/api/pyspark.sql.functions.percentile_approx.html

5.1. Encontra alguma situação estranha com a distribuição? Comente

6. Como estão distribuídas cada umas das etnias sobre o total dos doentes?
7. Qual é raça com maior e menor número de doentes e qual é o % total sobre o total da população?
8. Quais são 15 condições mais detetadas?
   8.1. Faça um horizontal barplot com esta informação?
9. Identifique quantos códigos nas condições estão repetidos?
   9.1. Quantas descrições diferentes tem cada um dos códigos identificados?
   9.2. Proponha uma forma de unificar os códigos e a suas descrições.
   10. Calcule a duração das condições(doenças) que os doentes padecem. desde a primeira vez que foi diagnosticado.
      10.1. Considere que para as pessoas mortas, a data de finalização da condição é o dia da morte específico para cada um dos doentes.
      10.2. Calcule a média em dias e anos, se for mais de 365 dias transforme a anos.
   11. O Dr Anthony Fauci recebeu informação afirmando que o número de doenças crónicas está relacionado diretamente com estádios mais severos do covid-19. A indicação dele é que toda condição detetada que tiver mais de 1 ano será considerada como uma doença crónica.
   12. Quantas doenças/condições foram classificadas como crónicas segundo a conceito do Dr Fauci.
13. Identifique a duração mínima, máxima e média (em anos) das doenças que crónicas.
14. Qual é o nome das 10 pessoas com mais doenças crónicas.
15. Identifique qual é o código que indica o peso do doente.

16. Calcule o BMI (IMC) número
17. Cria uma classificação do BMI segundo a seguinte tabela

BMI	Considered
Below 18.5	Underweight
18.5 to 24.9	Healthy weight
25.0 to 29.9	Overweight
30 or higher	Obesity
40 or higher	Class 3 Obesity

18. Detecte os doentes que presentam anomalias no seu peso.
    18.1. Use as seguintes formula com os valores calculados no ponto 17.:
    18.2. $limite\ superior = avg(BMI) + (3 \times stdev)$
    18.3. $limite\ inferior = avg(BMI) - (3 \times stdev)$
    18.4. Crie una nova coluna, boolean, onde represente se cada uma das observações calculadas no ponto 17 fica dentro ou fora do intervalo

Condições de Entrega:

Faça uso de **Databricks** e **Snowflake** para o desenvolvimento deste projeto final

tem de entregar dois (2) notebooks. O Primeiro notebook e sobre o ETL; o segundo e sobre a análise dos dados. Não têm limites de células ou linhas de código para cada um dos Notebook.

Para submeter o trabalho final, terá de enviar as soluções pelo slack por mensagem privada,

O Nome de cada um dos ficheiros tem de ser:

1. nome_apelido_etl.ipynb **Ex**: german_mendez_etl.ipynb
2. nome_apelido_eda.Ipynb **Ex**: german_mendez_eda. ipynb

Na nota final será tida em conta a apresentação e escrita do código. Por favor, evitar mais de 79 caracteres por linha e seguir as recomendações de boas práticas. **Dica**: no databricks pode usar Ctrl + Shift + F para dar formato ao seu código

A data-limite de entrega será o dia 17 de dezembro de 2023 até às 23h59m.