# 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 numa base de dados Não relacional.

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.

**Requisitos**

Crie 2 objetos, uma classe e uma função. A **classe tem de ter 3 funções**:

**Classe**:

Parameters:

URI,

* + Função \_\_int\_\_ 1: fazer ping ao cluster de mongo DB:
    - Parameters: nenhum
    - Return: Mensagem se o ligou ao mongo ou não
  + Função para ler o ficheiro usando Spark desde o DBFS:
    - Parameters: LocalFilepath
    - Return: DataFrame
  + Função para escrever um um DataFrame Monogo:
    - Parameters: DataFrame, Dababase, Collection,
    - Return: Mensagem se o dataframe escreveu ou não

**Função:**

* + Parameters: Lista dos nomes das colunas, Lista dos DataTypes para aplicar às colunas de forma ordenada, DataFrame
  + Return: DataFrame com as colunas com os novos datatypes

**Ajuda**:

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

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

**Passo a Passo.**

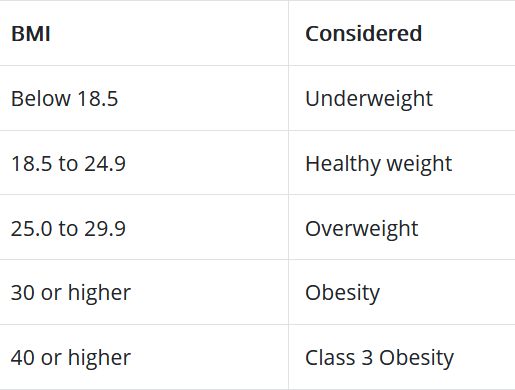
1. Crie um Notebook no Databricks
2. Trabalhe exclusivamente com estes ficheiros.
   * 1. Patients.csv
     2. Conditions.csv
     3. Devices.csv
     4. Organisations.csv
     5. Allergies.csv
3. Analise quais seriam os corretos datatypes para cada um dos ficheiros
4. Modifique os data types segundo a sua experiência.
5. Carregue os ficheiros no seu cluster no mongoDB.
6. Todo o passo anterior tem de ser executados com a classe e a função previamente criadas

**Segunda parte**

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 MongoDB faça a ligação usando o MongoDB e 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. 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:
   1. Se o estiver morto, essa será a data final para calcular a idade
   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]](#footnote-1) e máximo.
5. Faça um histograma com 100 bins (intervalos) da idade das pessoas.
   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?
   1. Faça um horizontal barplot com esta informação?
9. Identifique quantos códigos nas condições estão repetidos?
   1. Quantas descrições diferentes tem cada um dos códigos identificados?
   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.
    1. Considere que para as pessoas mortas, a data de finalização da condição é o dia da morte específico para cada um dos doentes.
    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



1. Detecte os doentes que presentam anomalias no seu peso.
   1. Use as seguintes formula com os valores calculados no ponto 17.:
   2. 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 **Atlas** para o desenvolvimento deste projeto final

tem de ser dois (2) jupyter notebooks . Não têm limites de células ou linhas de código para cada um dos Notebook. Para submeter os notebooks, terá de enviar as soluções através do GIT. na pasta projeto final terá de criar uma pasta com o seu nome e dentro colocar os 2 Notebook.

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 25 de junho de 2023 às 23h59m. Por cada dia adicional depois da data-limite será descontado 0.75 valores da nota final. Só se admitiram trabalhos até o dia 30 de junho. Ou seja, quem entregar no dia 30 até as 23h59m terá menos 3.75 valores. Os trabalhos submetidos após o 30 de junho não serão considerados e automaticamente será atribuído 0 valores no trabalho final.

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 [↑](#footnote-ref-1)