Projeto
(Trabalho de Grupo)
Big Data

O trabalho prático é obrigatório para a obtenção de aprovação na unidade curricular. No caso de não entrega durante o prazo previsto os alunos serão admitidos a exame.

Objetivo: Familiarização com os conceitos de big data

Entrega: Os trabalhos devem ser inseridos na plataforma de e-Learning em data a anunciar pelo docente.

Realização do trabalho: Os trabalhos devem ser entregues em formato notebook (devidamente documentados).

Neste exercício vamos recorrer a um dataset `ghcnd_daily.tar.gz` (https://www.dropbox.com/s/oq36w90hm9ltgvc/global_climate_data.zip?dl=0) que regista dados meteorológicos diários de milhares de estações ao longo de várias décadas.
Comece por descompactar o ficheiro (quando descompactado ocupa cerca de 4 GB, mas expandirá para cerca de 12 GB de RAM, o que significa que a maioria dos computadores (que geralmente têm, no máximo, 16 GB de RAM) não consegue importar este conjunto de dados para pandas e manipulá-lo diretamente).
Cada registo no ficheiro contém um mês de atividade (temperaturas máximas). `Value1` é o valor do primeiro dia do mês. `Value31` é o valor do último dia do mês (missing = -9999).
A versão completa do dataset (temperaturas máximas, temperaturas mínimas, precipitação, etc) pode ser encontrada no ficheiro `ghcnd_daily_30gb.tar.gz` que consta do ficheiro de download (quando descompactado ocupa mais de 30GB).
O ficheiro `ghcnd-stations.txt` inclui o código da estação meteorológica (que pode depois ser encontrado no ficheiro ghcnd_daily.csv na coluna id), bem como o nome e a localização da estação.
O ficheiro `readme.txt` descreve todos os campos dos vários datasets e dos ficheiros de texto.
The journal article describing GHCN-Daily is: Menne, M.J., I. Durre, R.S. Vose, B.E. Gleason, and T.G. Houston, 2012: An overview of the Global Historical Climatology

Network-Daily Database. Journal of Atmospheric and Oceanic Technology, 29, 897-910, doi:10.1175/JTECH-D-11-00103.1. To acknowledge the specific version of the dataset used, please cite: Menne, M.J., I. Durre, B. Korzeniewski, S. McNeal, K. Thomas, X. Yin, S. Anthony, R. Ray, R.S. Vose, B.E.Gleason, and T.G. Houston, 2012: Global Historical Climatology Network - Daily (GHCN-Daily), Version 3. [indicate subset used following decimal, e.g. Version 3.12]. NOAA National Climatic Data Center. http://doi.org/10.7289/V5D21VHZ [access date].

1. Guarde o dataset numa base de dados denominada ghcnd_daily.db.
2. Escreva uma query que permita selecionar apenas os 200 primeiros registos da base de dados.
3. Escreva uma query que selecione todos os registos dos campos *id, year, month, element, value1, value2, value3, value4, value5, value6, value7, value8, value9, value10, value11, value12, value13, value14, value15, value16, value17, value18, value19, value20, value21, value22, value23, value24, value25, value26, value27, value28, value29, value30, value31* das cinco estações meteorológicas portuguesas (Horta; Funchal; Lisboa; Castelo Branco e Faro) que existem no dataset.
4. Substitua os ids presentes no dataframe pelo correspondente nome da estação meteorológica.
5. Aplique as funções de *data wrangling* e de *compression* dos dados que achar adequadas.
6. Determine a percentagem de *null values* em cada uma das variáveis.
7. Crie uma nova coluna denominada *count_nan* que registe para cada observação o número de dias com temperaturas máximas nulas (*NaN*).
8. Crie um novo dataframe que considere apenas os registos com um valor total de *NaN* inferior a 7.
9. Reccorra agora a chunking para determinar a percentagem de null values em cada uma das variáveis na totalidade dos registos do dataframe.
10. Determine a temperatura média para cada observação (estação/ano/mês) do dataframe. Guarde os valores obtidos numa nova coluna denominada daily_avg_temp. Não se esqueça de considerar apenas as colunas cujo nome começa por value.
11. Agrupe os dados por nome de estação e ano aplicando a função mean. Mostre apenas os resultados para a coluna daily_avg_temp.