

# ME315 - Trabalho Final

Leitura/Manipulação de Dados em Python

Benilton Carvalho, Tatiana Benaglia

2º semestre de 2020

## Introdução

A coleta de dados em épocas mais recentes teve seu custo reduzido e volume ampliado significativamente. Redes sociais combinadas com dispositivos móveis rastreiam bilhões de pessoas no mundo, com informações das mais diversas origens: rotinas de passeio, amizades, históricos de compras, históricos de localização, fotografias geo-referenciadas, etc (FYI: com o escândalo da Cambridge Analytica, surgiram evidências de que o Facebook tenha registros de histórico de chamadas e SMS, para mensagem e ligações ocorridas fora do aplicativo Facebook, em telefones Android: <https://www.theverge.com/2018/3/25/17160944/facebook-call-history-sms-data-collection-android>). Empresas, por meio de programas de fidelidade, aprendem hábitos de compras, propensão de compras de certos produtos dependendo de horários, produtos que são comprados juntos e informações financeiras referentes a pagamentos. Desta maneira, o volume das bases de dados cresce vertiginosamente, de modo que é necessária a mudança de paradigma na estatística moderna: os dados não mais podem trafegar até o analista de dados; a análise deve ser transportada até os dados.

## Objetivos

Ao fim desta atividade, usando o Python, você deve ser capaz de:

- Importar um arquivo volumoso por partes;
- Calcular estatísticas suficientes para métrica de interesse em cada uma das partes importadas;
- Manter em memória apenas o conjunto de estatísticas suficientes e utilizar a memória remanescente para execução de cálculos;
- Combinar as estatísticas suficientes de modo a compor a métrica de interesse.

## Tarefa

Utilizando o conjunto de dados `flights.csv.zip`, vocês irão ler 100.000 observações por vez e determinar o percentual de vôos por Cia. Aérea que apresentou atraso na chegada (`ARRIVAL_DELAY`) superior a 10 minutos. As companhias a serem utilizadas são: AA, DL, UA e US. A estatística de interesse deve ser calculada para cada um dos dias de 2015. Os resultados para cada Cia. Aérea devem ser apresentados em um gráfico.

Para baixar o arquivo `flights.csv.zip`, utilize o link abaixo:

[https://drive.google.com/file/d/1NsjokM5mUQr\\_99l5AOECWJ-iOromJAbI/view?usp=sharing](https://drive.google.com/file/d/1NsjokM5mUQr_99l5AOECWJ-iOromJAbI/view?usp=sharing)

## Instruções

1. Quais são as estatísticas suficientes para a determinação do percentual de vôos atrasados na chegada (`ARRIVAL_DELAY > 10`)?

2. Utilize a função `read_csv` do `pandas` para importar o arquivo `flights.csv.zip` por partes e calcule o percentual de atraso para cada cia. aérea/mês/dia. Retorne o resultado abaixo em um `DataFrame`.
  - a. Configure o tamanho do lote (`chunk`) para 100 mil registros;
  - b. Configure o argumento `usecols` de forma que ele leia, diretamente do arquivo, apenas as colunas de interesse (`AIRLINE`, `DAY`, `MONTH` e `ARRIVAL_DELAY`);
  - c. Para cada lote de registros:
    - i. Filtre o conjunto de dados de forma que contenha apenas observações das seguintes Cias. Aéreas: AA, DL, UA e US;
    - ii. Remova observações que tenham valores faltantes em campos de interesse;
    - iii. Agrupe o conjunto de dados resultante de acordo com: dia, mês e cia. aérea;
    - iv. Para cada grupo em iii., determine as estatísticas suficientes apontadas no item 1;
  - d. Combine as estatísticas suficientes para compor a métrica final de interesse (percentual de atraso por dia/mês/cia aérea);
  - e. Retorne as informações em um `DataFrame` contendo as seguintes colunas:
    - i. `AIRLINE`: sigla da companhia aérea;
    - ii. `MONTH` e `DAY`: mês e dia, respectivamente;
    - iii. `PERC`: percentual de atraso para aquela cia. aérea/mês/dia, apresentado como um número real no intervalo  $[0, 1]$ .
3. Usando a biblioteca `plotnine`, produza um conjunto de gráficos de linhas (usando o `ggplot`) para representar o percentual de atrasos para cada Cia. Aérea, dia e mês de 2015, em que:
  - a. Cada mês deve estar representado em um gráfico, ou seja, use `facet_wrap` nos meses;
  - b. O eixo x representa os dias do mês e o eixo y o percentual de atraso para aquele mês e dia;
  - c. As cores das linhas representam as Cia. Aéreas;
  - d. Lembre-se de adicionar um título para o seu gráfico e especificar os eixos e legenda.