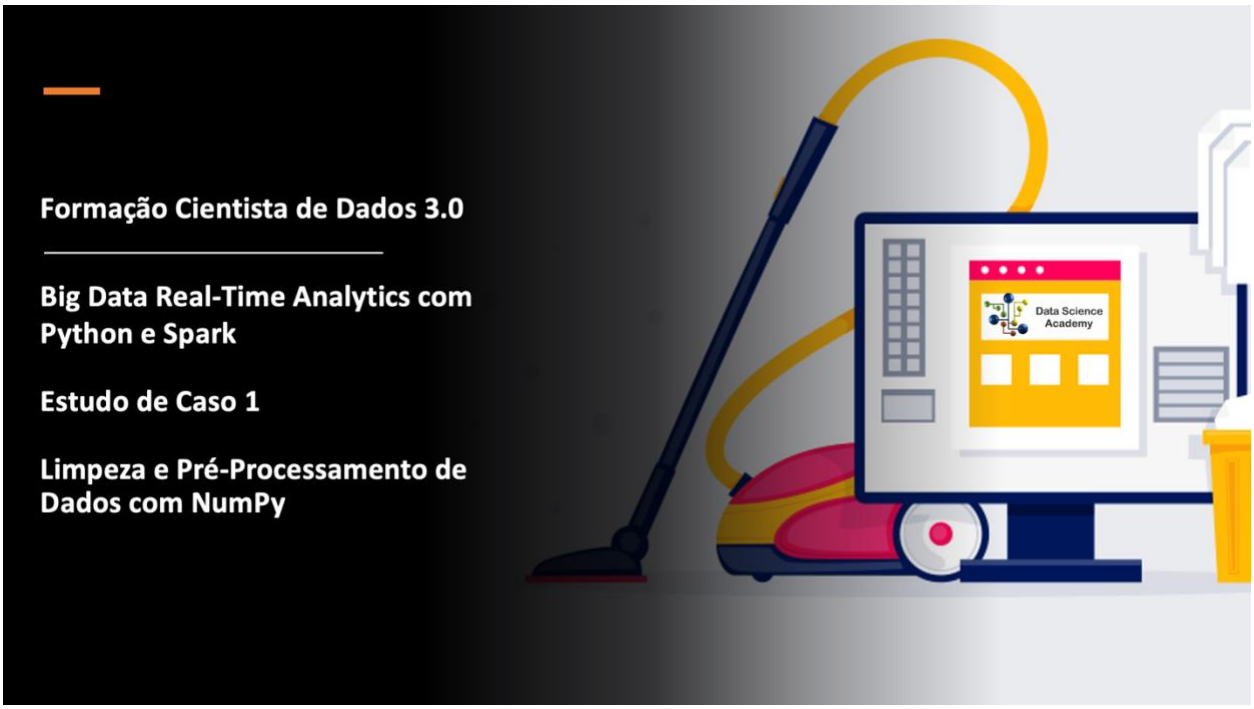
Big Data Real-Time Analytics
Com Python e Spark
Versão 3.0

Estudo de Caso 1
Definição do Problema e Fonte de Dados



Para este Estudo de Caso trabalharemos com dados reais disponíveis publicamente no link abaixo:

https://www.openintro.org/data/index.php?data=loans_full_schema

Esse conjunto de dados representa milhares de empréstimos feitos por meio da plataforma Lending Club, que é uma plataforma que permite que indivíduos emprestem para outros indivíduos.

Claro, nem todos os empréstimos são iguais. Alguém que fornece um baixo risco e que provavelmente vai pagar um empréstimo terá mais facilidade em obter um empréstimo com uma taxa de juros baixa do que alguém que parece ser mais arriscado.

E para as pessoas com alto risco de não pagar o empréstimo? Essas pessoas podem nem receber uma oferta de empréstimo, ou podem não aceitar uma oferta de empréstimo devido a uma alta taxa de juros. É importante ter em mente essa última parte, pois esse conjunto de dados representa apenas empréstimos efetivamente feitos, ou seja, não confunda esses dados com pedidos de empréstimo!

Usamos como fonte de dados o dataset disponível no link acima, mas fizemos modificações nos dados para deixá-los ainda mais problemáticos. O dataset será fornecido a você junto com os demais arquivos do capítulo.

Além disso usaremos um dataset com cotação do dólar em relação ao Euro. Extraímos uma pequena amostra de dados do site: https://finance.yahoo.com. O dataset será fornecido a você junto com os demais arquivos do capítulo.

Nosso trabalho é limpar e pré-processar os dados, deixando-os no formato ideal para um processo de análise posterior e várias decisões terão que ser tomadas no meio do caminho. Ao final do trabalho devemos salvar o dataset com os dados limpos e pré-processados.