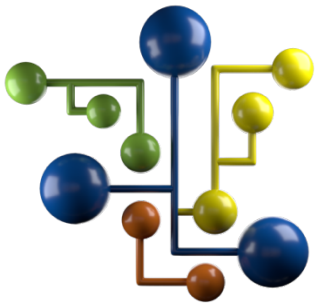

Data Science Academy

www.datascienceacademy.com.br

Big Data Real-Time Analytics com Python e Spark

Estudo de Caso
Agregação e Sumarização com MapReduce e PySpark

Definição do Problema

Neste estudo de caso, analisaremos um dataset fictício de aluguel de bicicletas em uma determinada cidade. Vamos analisar o dataset, construir algumas funções e responder as perguntas abaixo.

Leia e estude atentamente o código de cada célula (esse é o seu trabalho neste estudo de caso), faça comentários sobre sua compreensão e aproveite para resumir e revisar tudo que aprendeu durante o curso. Faz parte do seu trabalho ler e estudar o código de todas as funções e fazer alterações se julgar necessário, como por exemplo imprimir resultados intermediários com a função print().

Perguntas a serem respondidas

- Quais são as Top 5 estações com maior número de aluguel de bikes?
- Quais são as Top 5 rotas, com base na estação inicial e final, e a média de duração de cada aluguel?
- Quem aluga mais bikes, homens ou mulheres? Qual o tempo médio de aluguel de bikes?
- Qual faixa etária aluga mais bikes? Qual o tempo médio de aluguel de bikes?
- Quais são as estações com maior número de bikes alugadas/devolvidas?

Em anexo você encontra o Jupyter Notebook (que deve ser executado com PySpark) e o dataset.

Este Estudo de Caso é mais um recurso de aprendizado adicional fornecido pela Data Science Academy. Aproveite.