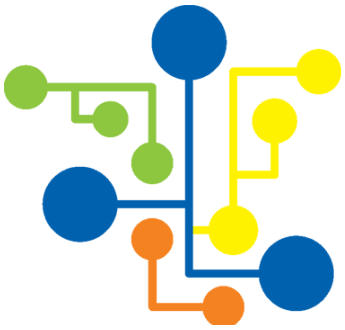

Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Mini-Projeto 3
Design de um Job MapReduce para Gastos Totais por Cliente

Você já parou para pensar quantas vendas são realizadas por dia em grandes empresas como, por exemplo, Amazon ou WalMart? Empresas que faturam bilhões vendendo os mais variados produtos para um grande número de clientes.

E se você fosse contratado para um projeto em uma dessas empresas e seu primeiro trabalho fosse calcular o total de vendas por cliente? Tarefa aparentemente simples. Sua primeira abordagem talvez fosse buscar o banco de dados transacional com as informações de vendas, cruzar os dados com o cadastro de clientes e obter o valor total gasto por cliente. Mas quantos clientes uma empresa como a Amazon possui? E se a solicitação fosse para gerar o total gasto por cliente nos últimos 5 anos, de modo a criar uma campanha personalizada para os clientes que tiveram os maiores gastos ao longo dos anos? Após alguma pesquisa, você poderia obter um dataset no seguinte formato:

Código do cliente	Valor gasto em uma única compra
128	899.90
1029	349.12
128	45.76

Sua pesquisa identificou que todos os registros dos últimos 5 anos geram um dataset com apenas duas colunas, mas 200 milhões de registros. Definitivamente esse não é um trabalho para um banco de dados relacional. Você precisa de uma ferramenta que possa rapidamente processar os dados e retornar apenas um valor total por cliente. Você então decide criar um job de mapeamento e redução. Com poucas linhas de código e usando linguagem Python, você consegue gerar o resultado esperado. Mas ainda tem um problema. Como processar esse job da forma mais rápida possível? Spark/Hadoop é a solução ideal.

Esse é um exemplo claro de projeto de Big Data. Um grande volume de dados e tudo que você precisa é extrair uma simples informação, que poderá fazer toda diferença na estratégia da empresa.

Seu trabalho agora é criar o Job de MapReduce e executá-lo com Spark e Hadoop.