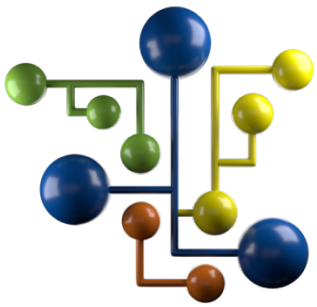


Data Science Academy

www.datascienceacademy.com.br

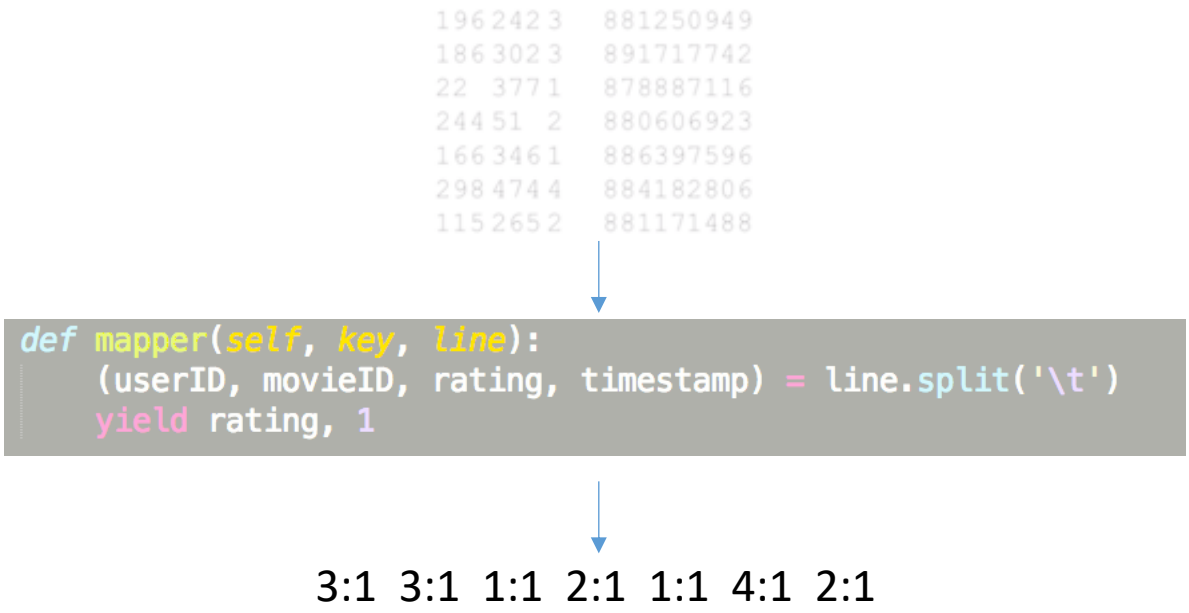
Engenharia de Dados com Hadoop e Spark

As 3 Fases do Processo de MapReduce

Considerando o Script AvaliaFilme.py, seguem as 3 fases do processo de MapReduce:

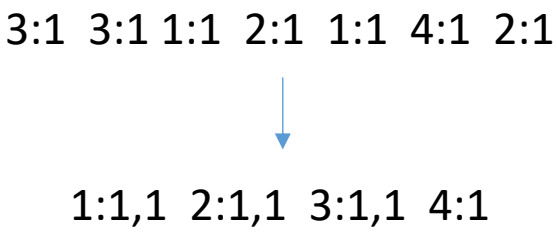
Fase 1 – Mapeamento

A palavra reservada *yield* define qual das colunas será a chave (nesse caso a coluna rating, pois queremos saber o total de filmes em cada rating, que vai de 1 a 5). Cada rating é mapeado e identificado com o valor 1, registrando a ocorrência do rating. Esse código é definido pelo Cientista de Dados.



Fase 2 – Shuffle e Sort

Essa fase é processada automaticamente pelo framework MapReduce, que então agrupa os ratings e identifica quantas ocorrências cada rating obteve ao longo do arquivo.



Fase 3 – Redução

Também definida pelo Cientista de Dados, esta fase aplica o cálculo matemático (no caso soma, com a função sum()) e retorna o resultado: total de filmes com rating 1, total de filmes com rating 2, etc....