01.RDD, DataFrame e Dataset

segunda-feira, 25 de setembro de 2023 10:02

RDD - Resilient Distributed Datasets(RDD)

- Estrutura básica de baixo nível
- Dados imutáveis, distribuídos pelo cluster
- Em memória
- Pode ser persistido em disco
- Tolerante a falha
- Operações sobre um RDD criam um novo RDD
- Otimização difícil pelo Spark

DataFrames e Dataset

- Semelhantes a uma estrutura de tabela de banco de dados
- Compatível com objetos Dataframe do Python e R
- Dataset tem uma particularidade: Disponível em java e scala