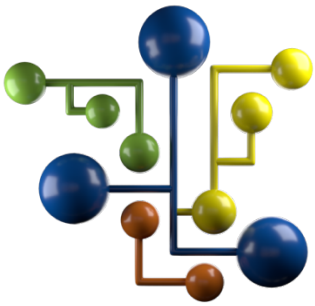


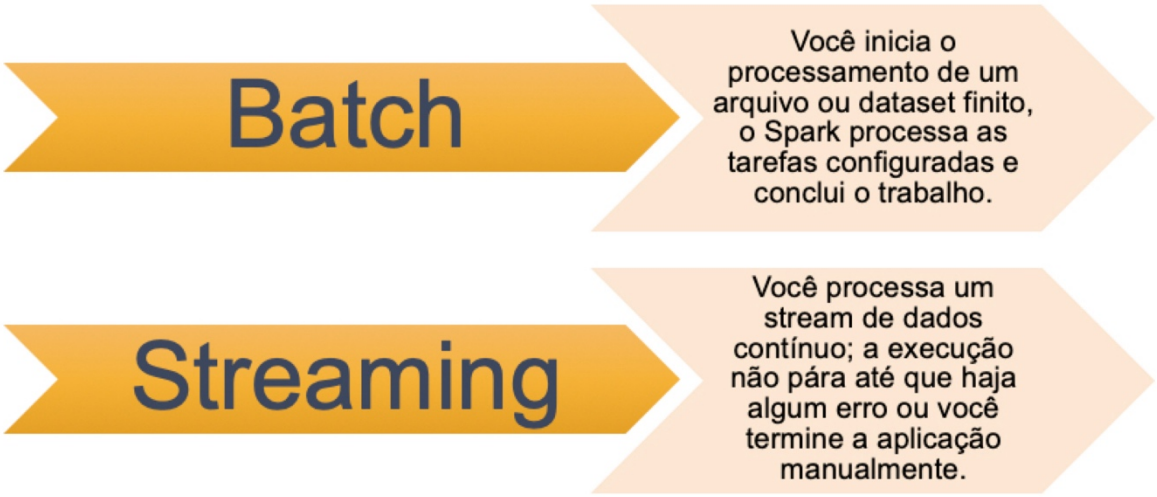
Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Processamento em Batch x Processamento de Stream

Existem basicamente 2 modos de processamento de dados:



Usamos Processamento em Batch no Apache Spark para:

- Análise exploratória de dados
- Gerar Data Warehouses sobre grandes conjuntos de dados, estilo OLAP
- Treinar um modelo de aprendizado de máquina sobre grandes conjuntos de dados
- Outras tarefas analíticas que antes eram feitas com Hadoop MapReduce

Usamos Processamento de Stream no Apache Spark para:

- Monitoramento de serviços
- Processamento de eventos em tempo real para alimentação de dashboards
- Processamento dados de cliques e eventos em web sites
- Processamento de dados de sensores de Internet das Coisas
- Processamento de dados vindos de serviços como: Twitter, Kafka, Flume, AWS Kinesis