Tópicos de Big Data em Python

Aula 04

Evandro J.R. Silva[1]

[1] Bacharelado em Ciência da Computação
Estácio Teresina

Sumário

Spark

Spark

- É um motor (*engine*) analítico unificado e de código aberto para processamento de dados de larga escala.

Spark

- É um motor (*engine*) analítico unificado e de código aberto para processamento de dados de larga escala.
- Provê uma interface para programar clusters com paralelismo de dados implícito e tolerante a falhas.

Spark

- É um motor (*engine*) analítico unificado e de código aberto para processamento de dados de larga escala.
- Provê uma interface para programar clusters com paralelismo de dados implícito e tolerante a falhas.
- Foi desenvolvido no AMPLab da Universidade da Califórnia, Berkeley. O código base foi então doado para a Apache, sua atual mantenedora, por isso que é conhecido como Apache Spark.

Spark

- O Apache Spark provê APIs em Java, Scala, Python e R, além de um motor (*engine*) otimizado que suporta execução geral da gráficos.
- Suporta também um conjunto de ferramentas de alto nível para SQL, Pandas, aprendizado de máquina e streaming de dados.
- Links úteis (documentação e guias oficiais):
  - Quick Start — exemplo rápido e simples.
  - Guia para programação RDD — o básico do Spark, incluindo API e conceitos mais antigos. RDD significa *Resilient Distributed Dataset*.
  - Spark SQL, Datasets e DataFrames — processamento de dados estruturados com consultas relacionais.
  - Streaming estruturado — processamento de streams de dados estruturados com consultas relacionais (usando Datasets e DataFrames, e API mais recente).
  - Spark Streaming — processamento de streams de dados usando DStreams (API antiga).
  - MLlib — aplicação de algoritmos de aprendizado de máquina.
  - GraphX — processamento de gráficos.
  - PySpark — biblioteca para processar dados com Spark no Python.
  - Spark SQL CLI — processamento de dados com SQL na linha de comando.

FIM