Utilizando o R para Big Data, uma introdução prática ao SparkR

Daniel dos Santos

20 de outubro de 2021

Universidade Federal Fluminense

Sumário

Agenda

Agenda

- Conhecer o Spark e o SparkR;
- Instalar o SparkR e suas dependências;
- Aplicar conceitos, funções e ferramentas em um banco de dados;
- Discussão final.

Início: 17:00

Conclusão: 18:30

Quem vos fala?

- Daniel dos Santos;
- Estatística - UFF;
- Analista de dados júnior em uma empresa de comércio eletrônico;
- Coleciona histórias em quadrinhos e figuras de ação.

Big Data

Desafios do Big Data

1. Grande volume de dados;

Desafios do Big Data

1. Grande volume de dados;
2. Dados em tempo real (*streaming*);

Desafios do Big Data

1. Grande volume de dados;
2. Dados em tempo real (*streaming*);
3. Variedade nas fontes de dados;

Desafios do Big Data

1. Grande volume de dados;
2. Dados em tempo real (*streaming*);
3. Variedade nas fontes de dados;
4. Trazer valor aos dados.

Desafios do Big Data

1. Grande volume de dados; **Volume**
2. Dados em tempo real (*streaming*); **Velocidade**
3. Variedade nas fontes de dados; **Variedade**
4. Trazer valor aos dados. **Valor**

Sobre o Spark

Sobre o Spark

1. Utiliza **MapReduce** com o objetivo de melhorar e facilitar análise de dados de grande dimensão;

Sobre o Spark

1. Utiliza **MapReduce** com o objetivo de melhorar e facilitar análise de dados de grande dimensão;
2. Desenvolvido em 2009 em Berkley, Universidade da Califórnia;

Sobre o Spark

1. Utiliza **MapReduce** com o objetivo de melhorar e facilitar análise de dados de grande dimensão;
2. Desenvolvido em 2009 em Berkley, Universidade da Califórnia;
3. Doado à Fundação Apache em 2010;

Sobre o Spark

1. Utiliza **MapReduce** com o objetivo de melhorar e facilitar análise de dados de grande dimensão;
2. Desenvolvido em 2009 em Berkley, Universidade da Califórnia;
3. Doado à Fundação Apache em 2010;
4. Utilizado em diversas instuições como, Yahoo!, IBM, Huwaei, Alibaba, Tencent, etc.

MapReduce

MapReduce

MapReduce é um paradigma computacional que visa processar informações em duas etapas, nomeadas de **Map** e **Reduce**.

- **Map**: Processar um conjunto de valores e transformá-los em valores intermediários;

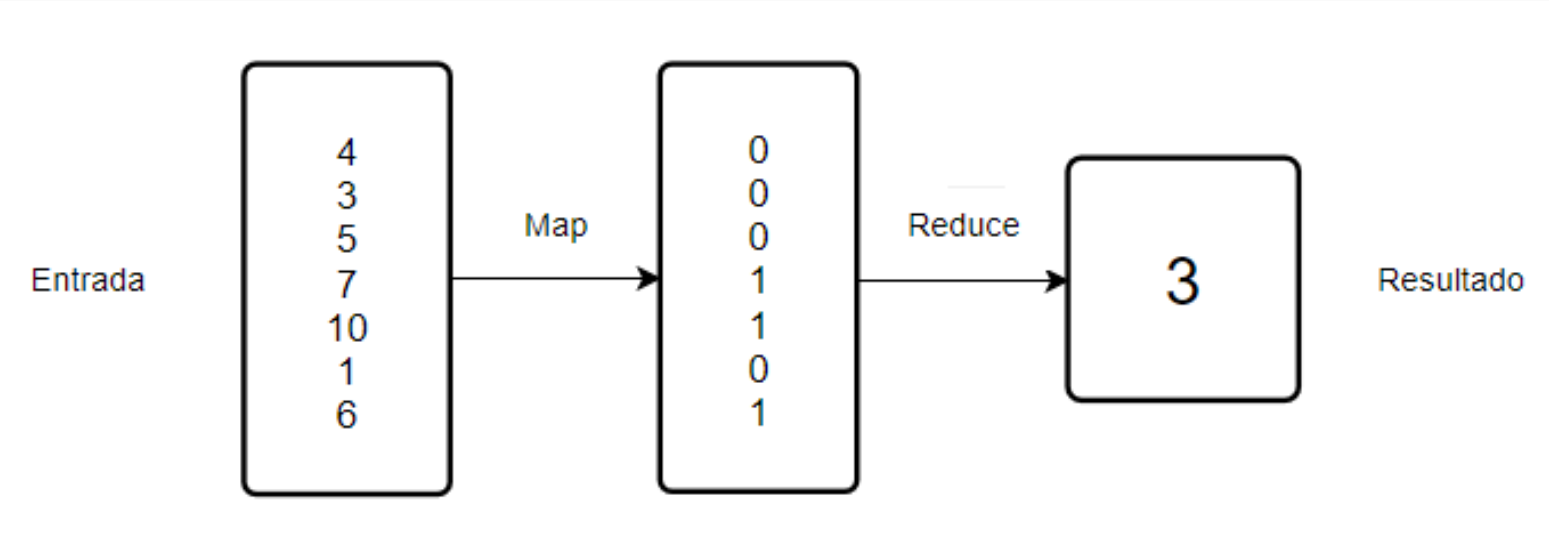
MapReduce

MapReduce é um paradigma computacional que visa processar informações em duas etapas, nomeadas de **Map** e **Reduce**.

- **Map**: Processar um conjunto de valores e transformá-los em valores intermediários;
- **Reduce**: Resumir os valores transformados, de forma que gerem o resultado esperado.

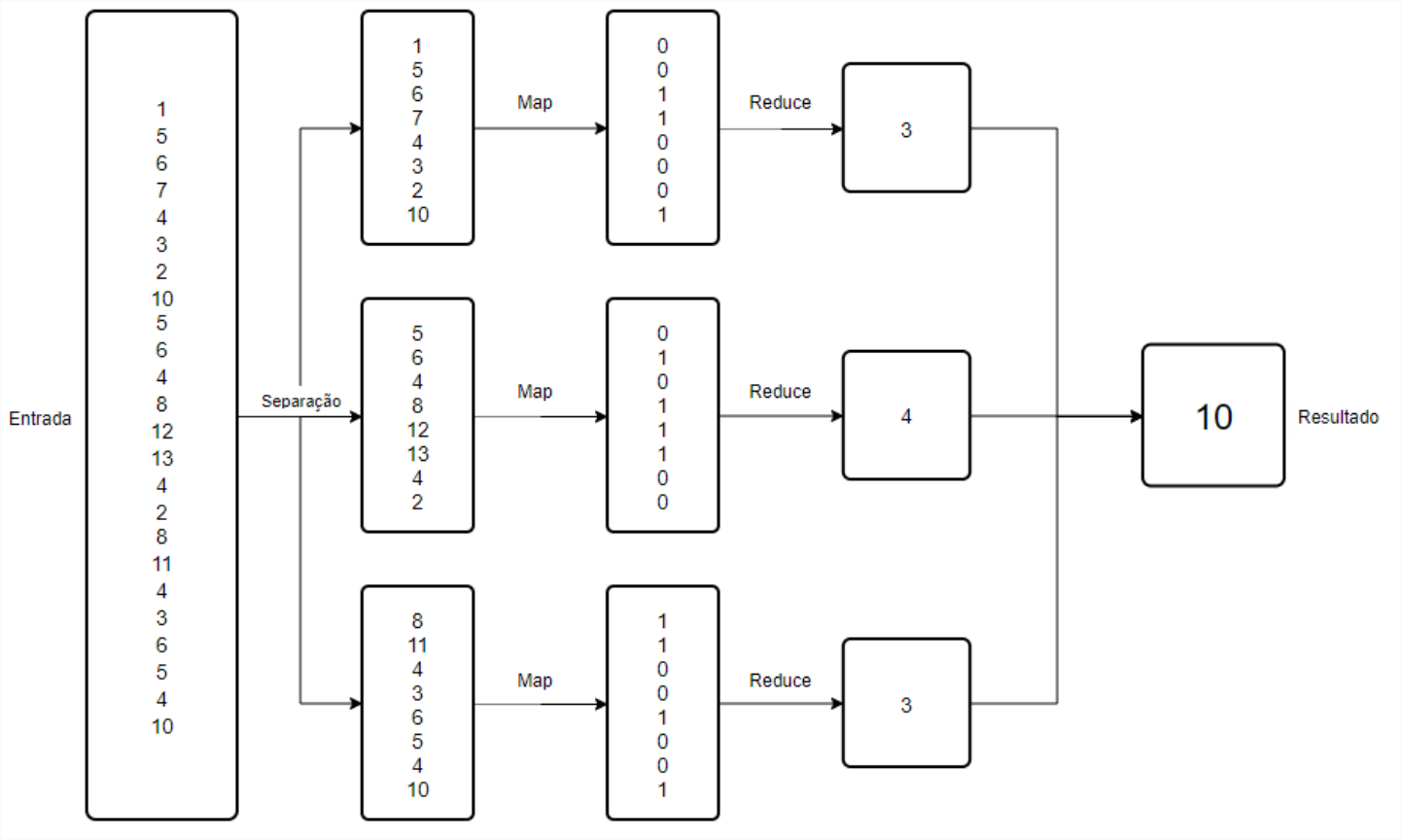
MapReduce

Suponha-se que deseja-se calcular a quantidade de valores maiores do que 5 em um conjunto de dados, o diagrama baixo exemplifica a utilização do **MapReduce**.



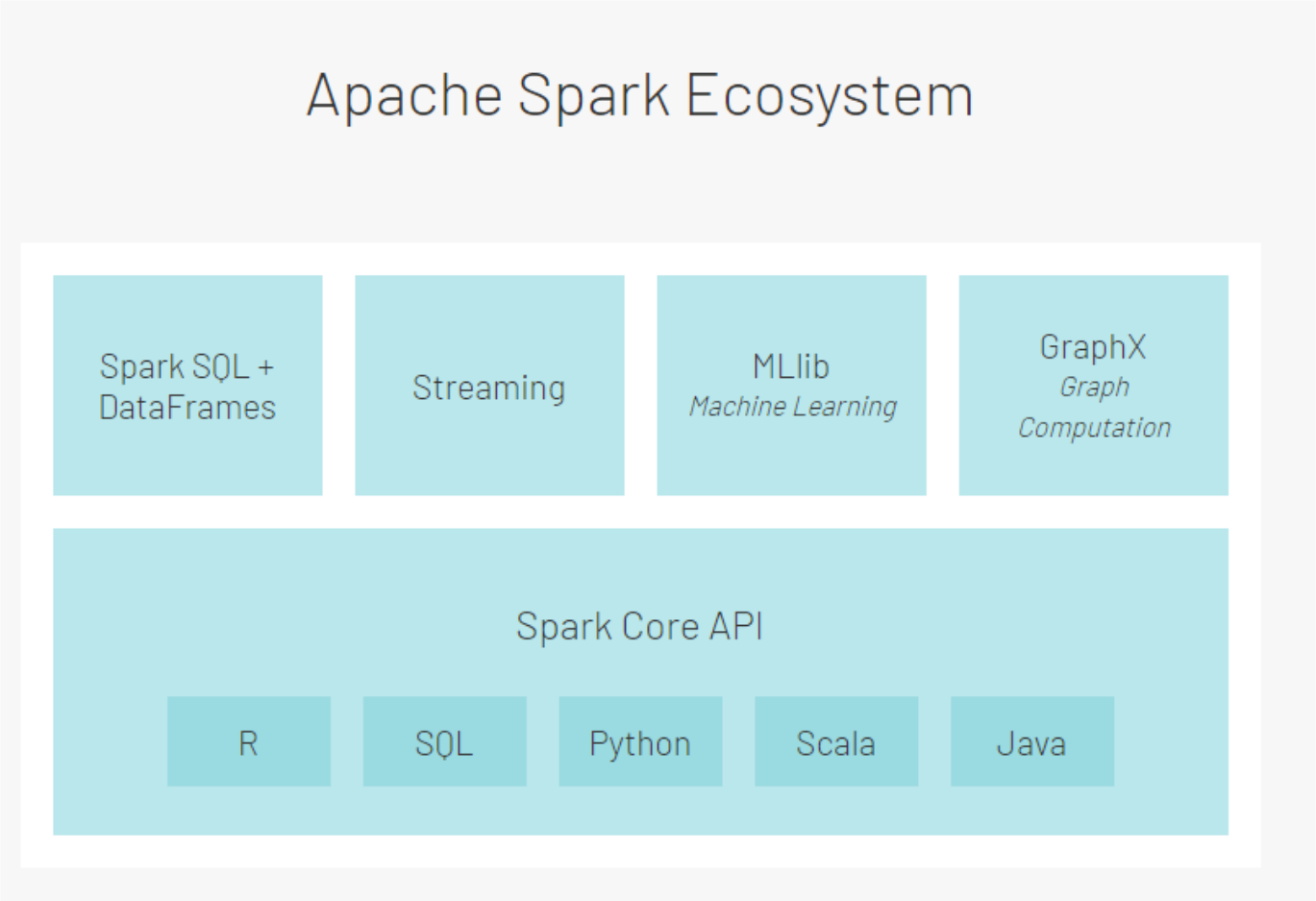
MapReduce Paralelizado

Este paradigma é facilmente escalável, já que o processamento dos dados pode ser feito em paralelo, assim o **Hadoop MapReduce** foi capaz de paralelizar uma série de ações.



Ecossistema Spark

Ecossistema Spark



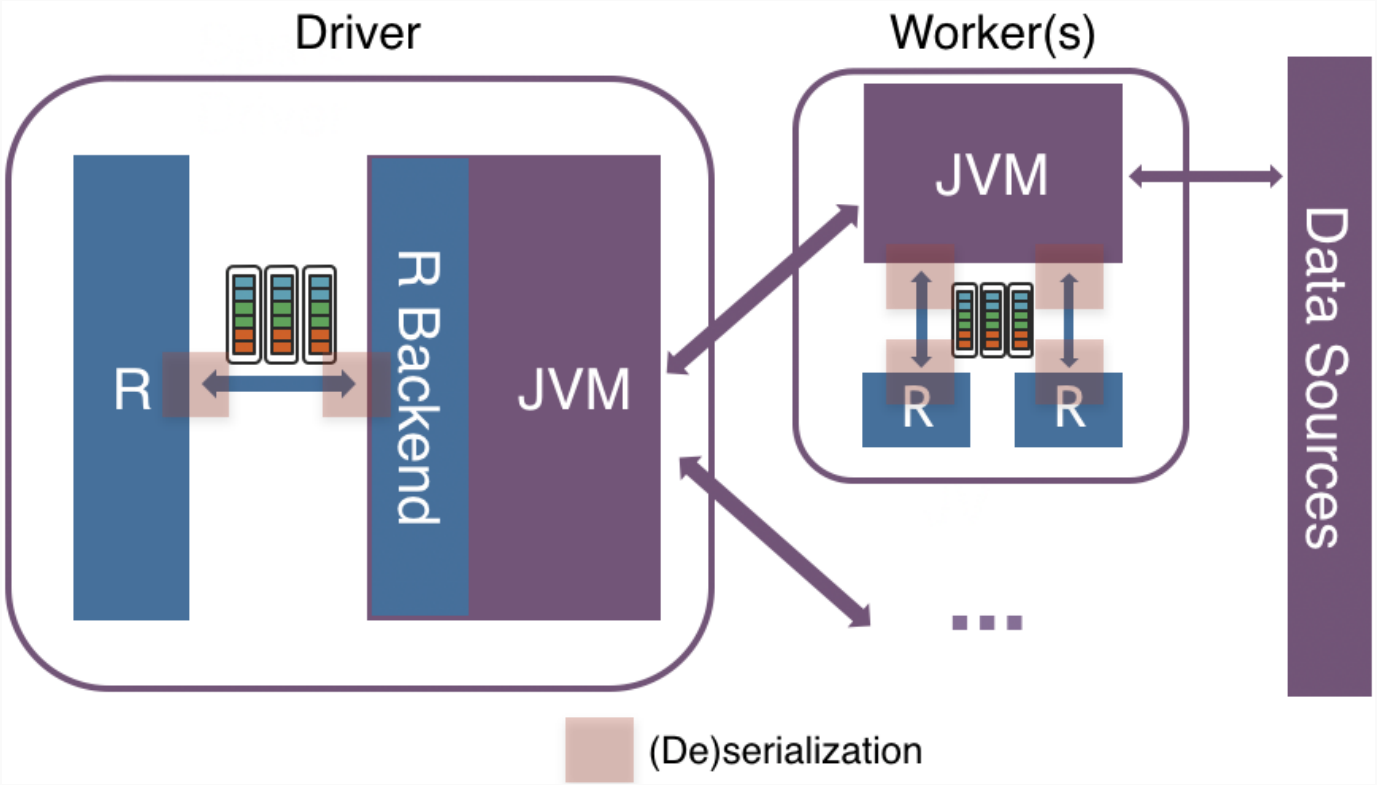
SparkR

SparkR

Através do pacote **SparkR** é possível trazer diversas ferramentas do Spark para dentro da liguagem R.

SparkR

O **SparkR** faz com que o **R** se comunique com o Spark através de uma **API**.



Mão na massa

Instalação

1. **Java**: Download;
2. **Spark versão 3.1.2 com Hadoop 2.7**: Download;

OkCupid

Informações de perfil de usuários do site de relacionamentos **OkCupid** no ano de 2012.

Logo OkCupid

OkCupid

- Manipulação de dados;
- Análise descritiva;
- Análise gráfica.

Dúvidas?

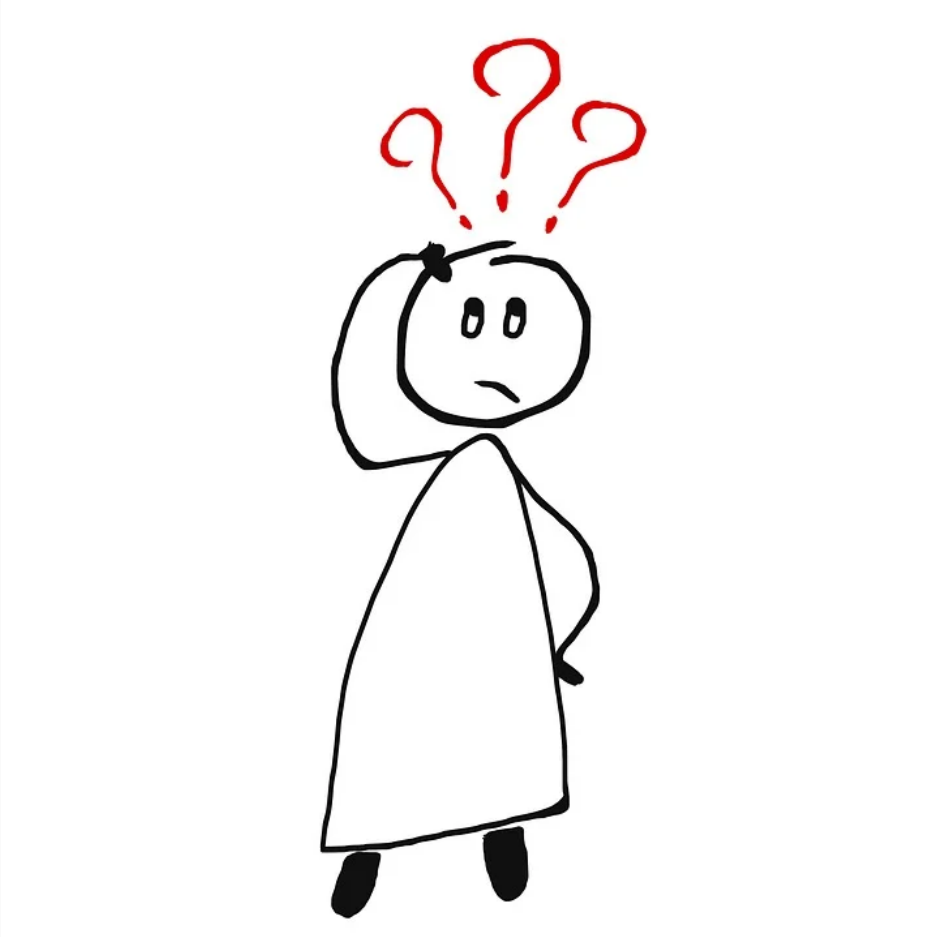


Image de uso livre, Pixbay

Onde me encontrar?

Minicurso:

`https://github.com/Daniel-EST/sparkr_semest_2021`

TCC: `https://github.com/Daniel-EST/spark-tcc`

Discord: NielINSIDER#1997

Email: dd_santos@id.uff.br