

**Projeto em ciência de
dados com soluções
para processamento
paralelo e distribuído
de dados**



PÓS-GRADUAÇÃO

Processos com R Programming e Data Lakes

Bloco 1

Marcelo Tavares de Lima






► Objetivos

- Apresentar procedimentos para tratamentos de processos e dados com o R *Programming*.
- Apresentar conceitos fundamentais de *Data Lake*.
- Apresentar aplicações de *Data Lake* com R *programming*.



► Processos com o R *Programming*

- O programa R é um software para tratamento e análise de dados, gratuito e de código aberto (*open source*).
 - É resultante de um projeto elaborado por diversos colaboradores ao redor do mundo, composto por implementações para análises estatísticas, desde as mais simples até a mais complexas, além de poder manipular grandes volumes de dados, segundo Siqueira e Tibúrcio (2011).
- 



► Processos com o R *Programming*

- O programa R é composto por vários pacotes, também compostos por funções, sendo compostas por linhas de comando.
- A linguagem de programação R, conforme definem Oliveira, Guerra e McDonnell (2018, p. 10), pode ser entendida como “um conjunto de pacotes e ferramentas estatísticas, munido de funções que facilitam sua utilização, desde a criação de simples rotinas até análises de dados complexas”.



► Processos com o R *Programming*

- A interface do programa R é muito simples e se baseia, basicamente, em linhas de comando. No entanto, para executar suas linhas de programação, recomenda-se a utilização de um editor de texto ou de uma *Integrated Development Environment* (IDE), ou seja, uma interface amigável para facilitar a sua.



► Processos com o R *Programming*

- “Inúmeras organizações, atualmente, têm a necessidade de realizar o processamento e análise de uma grande quantidade de dados em tempo computacional hábil”. (HÖLBIG; MAZZONETTO; PAVAN, 2017, p. 27). v



► Processos com o R *Programming*

- Por isso, ao longo dos anos, muitas empresas passaram a utilizar o programa R como um de seus recursos computacionais por conta de suas características, conforme algumas citadas anteriormente.
- É possível citar algumas dessas empresas, conforme apresentado no Quadro 1.

► Processos com o R *Programming*

Quadro 1 – Empresas que utilizam programação R em suas atividades

Empresa	Finalidade do uso do R.
Google	Para descobrir o retorno sobre o investimento em publicidade.
Facebook	Análise de atualizações de status do Facebook.
Microsoft	Serviço Xbox, estatístico com estrutura do Azure Machine Learning.
Ford Motor	Para análise estatística e decisão orientada por dados.
John Deere	Modelagem de séries temporais e análise geoespacial.
Lloyds	Desenvolvimento de gráficos de movimento para fornecer análises para investidores.

Fonte: Shinde, Oza e Kamat (2017).



► Processos com o R *Programming*

- A IDE mais popular para o programa R é o RStudio, disponibilizada em diversas versões para os vários sistemas operacionais existentes.
- Além disso, existem versões gratuitas e pagas, que se diferenciam em suas funcionalidades e facilitam a utilização de programação R.



► Processos com o R *Programming*

- Também é possível utilizar programação em linguagem R em ambientes paralelos, que trabalham em *clusters*, *grids*, processadores multicore e *Graphics Processing Unit (GPU)*, ou unidade de processamento gráfico.
- Existem diversos pacotes elaborados para ambientes com arquitetura paralela, descritos com maiores detalhes em seu material de estudo e resumidas a seguir.



► Processos com o R *Programming*

- Rmpi.
- snow.
- snowfall.
- foreach (exemplo na LF).
- doMC.
- Etc.



► Processos com o R *Programming*

- Voltando a falar de *Big Data*, em se tratando de linguagem R, Hölbig, Mazzonetto e Pavan (2017) declaram que há uma dificuldade encontrada nesse ambiente de programação.
- Os autores afirmam que tal dificuldade está ligada com a necessidade de manter os dados na memória no ambiente R.
- “Mesmo para os computadores modernos com grande capacidade de armazenamento, isto pode se apresentar como um desafio significativo”. (HÖLBIG; MAZZONETTO; PAVAN, 2017, p. 34).



► Processos com o R *Programming*

- Os principais pacotes em linguagem R capazes de manipular grandes quantidades de dados e fazer gerenciamento de seu uso em memória são denominados “RevoScaleR”, “parallel” e “RHadoop”.
- Mais informações podem ser encontradas em Hölbig, Mazzonetto e Pavan (2017).



► Processos com o R *Programming*

- Para arquiteturas paralelas e sistemas distribuídos também existem alguns pacotes em linguagem R que são apropriados para isso.
- Exemplos: os pacotes sparklyr, dyplr, entre outros.

PÓS-GRADUAÇÃO

Processos com R Programming e Data Lakes

Bloco 2

Marcelo Tavares de Lima





► Conceitos fundamentais de *Data Lake*

- No ano de 2010 foi apresentado um novo conceito, o de Data Lake ou Data Hubs, introduzido por James Dixon.
- Esse conceito, inicialmente, foi menosprezado por ter sido entendido como um rótulo de marketing para um produto que fosse compatível com o Hadoop, segundo Miloslavskaya e Tolstoy (2016).



► Conceitos fundamentais de *Data Lake*

- *Data Lake*, em sua essência, é uma estratégia de armazenamento de dados, semelhante ao um *Data Warehouse*.
- “O Data Lake pode armazenar dados estruturados e não estruturados em seu formato bruto e são projetados para o consumo de dados, apoiando a descoberta de novas perguntas”. (DATAPREV, 2019, p. 11).



► Conceitos fundamentais de *Data Lake*

- Os *Data Lake*, geralmente, incluem um banco de dados semântico, um modelo conceitual que utiliza os mesmos padrões e tecnologias usados para criar hyperlinks da Internet e adicionar uma camada de contexto sobre os dados, o que define o significado dos dados e suas inter-relações com outros dados.



► Conceitos fundamentais de *Data Lake*

- As estratégias de *Data Lake* podem combinar abordagens de banco de dados SQL e NoSQL e processamento analítico on-line (OLAP) e recursos de processamento de transações on-line (OLTP), segundo Miloslavskaya e Tolstoy (2016).



► Conceitos fundamentais de *Data Lake*

Ao contrário de um *Data Warehouse* hierárquico, com armazenamento de arquivos ou pastas, o *Data Lake* utiliza uma arquitetura simples, em que cada elemento de dados tem um identificador exclusivo e um conjunto de *tags* de metadados estendidas.



► Conceitos fundamentais de *Data Lake*

- Ele não requer um esquema rígido ou manipulação de dados de todas as formas e tamanhos, mas requer manter a ordem de chegada dos dados.
- Pode ser imaginado como um grande conjunto de dados que traz registros históricos acumulados e novos (estruturados, não estruturados e semiestruturados), em tempo quase real, em um único local, no qual o esquema e os requisitos de dados não são definidos até que sejam consultados, segundo Miloslavskaya e Tolstoy (2016).



► Conceitos fundamentais de *Data Lake*

- A utilização de linguagem R para tratamento de dados armazenados em *Data Lake* pode ser realizada com o software Spark. Para sua utilização em ambiente R, é necessário utilizar o pacote “SparkR.”



► Conceitos fundamentais de *Data Lake*

- Existe uma interface para R que utiliza Apache Spark para manipulação de grandes volumes de dados. Para sua utilização, basta acessar o repositório do R (CRAN) e fazer o download e instalação.



► Conceitos fundamentais de *Data Lake*

- Funciona como se fosse um pacote qualquer do ambiente R, o que significa que, após sua instalação, é necessário ativar o pacote com o comando “require()” ou “library()”.

PÓS-GRADUAÇÃO

Teoria em prática


Bloco 3

Marcelo Tavares de Lima






► Teoria em prática

- Considere que a empresa que você trabalha está passando por uma série de reformas, inclusive na área de Tecnologia de Informação (TI).
 - Você é usuário de alguns sistemas da empresa e, por conta dessas mudanças, muitos sistemas de dados serão substituídos por sistemas mais modernos e mais atuais.
- 




► Teoria em prática

- Para que você possa continuar utilizando os sistemas de forma satisfatória, precisará realizar alguns treinamentos relacionados ao manuseio dos novos sistemas existentes.
 - Para isso, você precisa priorizar os treinamentos em sistemas que mais utiliza e que são prioritários para o bom andamento de seu trabalho.
- 



► Teoria em prática

- Em um desses sistemas, você precisará utilizar a linguagem R para ter acesso aos dados.
 - Como você daria início ao estudo dessa linguagem de programação?
 - Pense, supondo que você tem pouco conhecimento sobre o assunto e precisa utilizar sistemas que lidam com grandes massas de dados.
- 



► Teoria em prática

- Primeiramente, você pode verificar se existirá algum treinamento sobre os sistemas que utiliza.
- Caso positivo, deve procurar se inscrever nesse treinamento.
- Deve verificar também o que é necessário para que possa participar do treinamento.
- Enfim, verificar como estão suas habilidades com a linguagem R.

Dica do professor

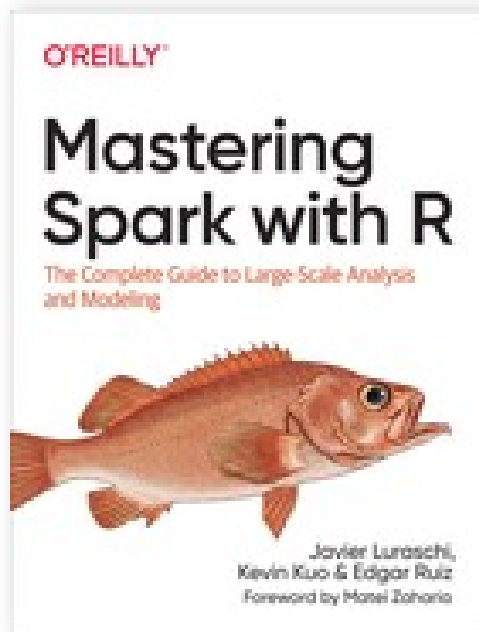
Bloco 4

Marcelo Tavares de Lima



► Indicação de livro

Figura 1 - Livro



Mastering Spark with R

★★★★★ 1 REVIEW

by Edgar Ruiz, Kevin Kuo, Javier Luraschi

Publisher: O'Reilly Media, Inc.

Release Date: October 2019

ISBN: 9781492046370

Topic: Spark



Fonte: <https://therinspark.com/>. Acesso em: 05 fev. 2020.



► Referências

DATAPREV. **Administração de dados:** padrões para construção de modelos de dados. [s.l.]. 2019. Disponível em: https://portal.dataprev.gov.br/sites/default/files/arquivos/instrumentos_normativos/n_ad_001_05.pdf. Acesso em: 04 fev. 2020.

HÖLBIG, C. A.; MAZZONETTO, A.; PAVAN, W. **Computação paralela com a linguagem R:** técnicas, ferramentas e aplicações. Minicurso. 17ª Escola Regional de Alto Desempenho do Estado do Rio Grande do Sul. Anais, p. 25-42. Ijuí: RS, 2017. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/erad/2017/003.pdf>. Acesso: 04 fev. 2020.

MILOSLAVSKAYA, N.; TOLSTOY, A. Big data, fast data and data lake concepts. **Procedia Engineering**, 2016. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050916316957>. Acesso em: 04 fev. 2020.



► Referências

OLIVEIRA, P. F.; GUERRA, S.; McDONNELL, R. **Ciência de dados com R: introdução**. Brasília: IBPAD. 2018. Disponível em: <https://www.ibpad.com.br/o-que-fazemos/publicacoes/introducao-ciencia-de-dados-com-r#download>. Acesso em: 04 fev. 2020.

SIQUEIRA, A.; L., TIBÚRCIO, J. D. **Estatística na área da saúde: conceitos, metodologia, aplicações e prática computacional**. Belo Horizonte: Coopmed, 2011.

SHINDE, P. P.; OZA, K .S.; KAMAT, R. K. **Big Data predictive analysis: using R analytical tools**. International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, 2017. Disponível em: <https://ieeexplore.ieee.org/document/8058297>. Acesso em: 04 fev. 2020.

