

CIÊNCIA DE  
**DADOS**

# PROCESSAMENTO DE BIG DATA EM R E PYTHON

## FICHA DO CURSO

**Curso:** Processamento de Big Data em R e Python

**Modalidade:** EAD assíncrono

**Carga horária total:** 40 horas

**Carga horária semanal:** 4 horas

**Início da oferta:** EAD

**Fim da oferta:** EAD

**Pré-requisito:** Recomenda-se ter conhecimento básico de SQL, R e Python

**Professores:** Anderson Luiz Ara de Souza

## Objetivos

Este curso é essencial para profissionais que buscam competências avançadas no processamento de Big Data, uma habilidade crucial na era da informação. Ao completar o curso, os alunos estarão aptos a manipular eficientemente grandes volumes de dados, realizar análises complexas e tomar decisões informadas em ambientes que demandam o processamento de dados em larga escala. Além disso, o curso abrange uma variedade de ferramentas e tecnologias populares no ecossistema de Big Data, oferecendo aos alunos uma visão abrangente e prática do processamento de dados em ambientes complexos. Isso contribui para a empregabilidade dos profissionais em setores como análise de dados, ciência de dados e engenharia de dados, onde o processamento de Big Data é uma habilidade altamente valorizada.

Ao final do curso o participante será capaz de usar bancos de dados de maneira eficiente e segura para processamento de Big Data em SQL e NoSQL a partir dos ambientes da linguagem R e Python.

## Programa

O curso é composto de 3 tópicos descritos a seguir.

### Tópico 1: Introdução ao Processamento de Big Data

- **Linguagem SQL:** Capacitar os alunos a revisar e aprimorar suas habilidades em SQL, uma ferramenta essencial para a consulta eficiente de grandes conjuntos de dados.

- **Infraestrutura para Big Data:** Apresentar as principais infraestruturas utilizadas no processamento de Big Data, passando pelos tipos de bancos de dados e a infraestrutura do Apache Spark, proporcionando uma compreensão sólida dos ambientes em que o processamento ocorre.
- **APIs e processamento de dados não estruturados:** Desenvolver habilidades na manipulação de dados não estruturados, fornecendo aos alunos a capacidade de interagir com formatos como XML e JSON, comuns em ambientes de Big Data.

### Tópico 2: Processamento de Big Data com R

- **Processamento de dados tabulares:** Capacitar os alunos a realizar manipulações eficientes em grandes conjuntos de dados tabulares usando principalmente as bibliotecas *dplyr* e *data.table*.
- **Processamento de dados tabulares em Bancos de Dados:** Introduzir o uso do *dbplyr* para realizar operações SQL em bancos de dados relacionais, proporcionando uma ponte eficiente entre R e SQL.
- **Processamento de Big Data com Spark:** Explorar a integração do R com o Apache Spark por meio do *sparklyr*, permitindo aos alunos realizar análises em larga escala.
- **Outras opções de processamento:** Apresentar diferentes arquiteturas e tecnologias relevantes para o processamento de Big Data, incluindo *DuckDB*, *Parquet*, *Kafka* e *ElasticSearch*.
- **Processamento de NoSQL:** Capacitar os alunos a interagir com bancos de dados NoSQL, com foco no MongoDB, proporcionando uma compreensão abrangente do uso de bancos de dados não relacionais em cenários de Big Data.
- **Dados em XML e JSON:** Desenvolver habilidades específicas para lidar com dados em formatos não estruturados como XML e JSON no contexto de Big Data.

### Tópico 3: Processamento de Big Data com Python

- **Conexão com Bancos de Dados SQL e NoSQL:** Capacitar os alunos a conectarem-se a bancos de dados SQL e NoSQL usando Python, proporcionando flexibilidade na escolha de tecnologias.
- **Processamento de dados tabulares:** Introduzir a biblioteca *datatable* para processamento eficiente de grandes conjuntos de dados tabulares em Python.
- **Processamento com Spark:** Explorar o uso do *PySpark* para realizar processamento distribuído de Big Data em ambientes Spark usando Python.

- **Dados em XML e JSON:** Desenvolver habilidades específicas em Python para lidar com dados em formatos não estruturados como XML e JSON em contextos de Big Data.

### Conteúdo Programático

O programa do curso foi dividido em 10 módulos. Os temas de cada módulo de acordo com os tópicos do curso são apresentados na Tabela 1.

Tabela 1 - Cronograma detalhado do conteúdo dos módulos.

Tópico	Módulos	Conteúdo programado	Duração
1	1	Introdução ao Big Data	4 horas
	2	Arquiteturas para Big Data	4 horas
	3	SQL em ambientes de Big Data	4 horas
2	4	Processamento de Big Data com R	4 horas
	5	Analisando Big Data com R	4 horas
	6	R e MongoDB	4 horas
	7	Processamento de Big Data com Python	4 horas
3	8	Python e Spark	4 horas
	9	Python e MongoDB	4 horas
	10	Projeto integrador	4 horas
Encerramento do curso			

### Procedimentos didáticos e metodológicos

O conteúdo do curso será estruturado em 10 módulos. A cada módulo serão desenvolvidas as seguintes atividades pelos cursistas.

- Assistir ao conteúdo em vídeo.
- Acessar o material em texto e demais arquivos (slides, relatórios, scripts, datasets) que acompanham o módulo.

- Fazer o estudo individual com os materiais disponibilizados como: leituras complementares e acesso a vídeos adicionais.
- Realizar as atividades de avaliação do módulo no formato de quiz.

### Desempenho no curso

O desempenho no curso será determinado pela nota da atividade avaliativa aplicada aos cursistas após o último módulo. Requer-se, no mínimo, 70% de aproveitamento das 10 questões para obtenção de certificado.

### Avaliação de reação dos cursistas

Ao final do curso será realizada uma avaliação de reação dos cursistas.

### Referências bibliográficas

1. Aven, J. (2018). **Data analytics with spark using python**. Addison-Wesley Educational.
2. Badia, A. (2020). **SQL for data science: Data cleaning, wrangling and analytics with relational databases** (1st ed.). Springer Nature.
3. Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2021). **Modern data science with R** (2nd ed.). Routledge Cavendish. <https://doi.org/10.1201/9780429200717>
4. Nolan Deborah Lang, & Lang, D. T. (2013). **XML and web technologies for data sciences with R** (2014th ed.). Springer. <https://doi.org/10.1007/978-1-4614-7900-0>

5. Parsian, M. (2022). **Data algorithms with spark: Recipes and design patterns for scaling up using PySpark**. O'Reilly Media.
6. Shan, J., Goldwasser, M., Malik, U., & Johnston, B. (2022). **SQL for Data Analytics: Harness the power of SQL to extract insights from data** (3rd ed.). Packt Publishing.
7. Tandon, A., Ryza, S., Laserson, U., Owen, S., & Wills, J. (2022). **Advanced analytics with PySpark: Patterns for learning from data at scale using python and spark**. O'Reilly Media.
8. Tanimura, C. (2021). **SQL for data analysis: Advanced techniques for transforming data into insights**. O'Reilly Media.
9. Walkowiak, S. (2016). **Big Data Analytics with R**. Packt Publishing.

### Informações sobre os conteudistas

**Anderson Luiz Ara de Souza**. Graduado em Estatística (2009), Mestre em Estatística (2011), títulos obtidos pela Universidade Federal de São Carlos (UFSCar). Doutor em Estatística (2016) através dos Programas de Pós-Graduação em Estatística (PPGEst-UFSCar) e Pós-Graduação em Ciência da Computação (PPG-CC-UFSCar). Desde agosto de 2021 é Professor Adjunto da Universidade Federal do Paraná (UFPR), campus Curitiba/PR, Departamento de Estatística (DEst) do Setor de Ciências Exatas. Pesquisador Permanente do Programa de Pós-Graduação em Informática (PPGInf-UFPR) pela linha Inteligência Computacional. Pesquisador colaborador do Programa de Pós-Graduação em Métodos Numéricos (PPGMNE-UFPR). Pesquisador colaborador do Programa de Pós-Graduação em Matemática (PGMAT-UFBA) Tutor do Programa de Educação Tutorial (PET). Atua principalmente nas seguintes áreas: Aprendizado Estatístico de Máquina, Inferência Estatística e Métodos Computacionais. Tem orientado e publicado em periódicos nacionais e internacionais da área. Tem experiência no desenvolvimento de projetos multidisciplinares de pesquisa, extensão e de desenvolvimento tecnológico.

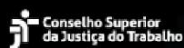
### Requisitos Técnicos

Computador com acesso à internet. Permissão para instalar programas.

# CIÊNCIA DE DADOS



APOIO



REALIZAÇÃO

