

Leitura Complementar

Disciplina: Linguagens de programação
para ciência de dados (*Python com Spark*)

Autor da leitura complementar: Danilo Rodrigues Pereira



Prezado aluno, selecionamos as referências abaixo visando o aprofundamento das temáticas estudadas na disciplina e a complementação dos seus estudos. Para conferir as indicações, acesse a nossa biblioteca virtual: <https://biblioteca-virtual.com/> e boa leitura!

► Tema 01 – Introdução: por que *Python*, *Spark* e *Hadoop*? Preparação do ambiente em *Spark* e *Python*

Performance evaluation of cloud-based log file analysis with Apache Hadoop and Apache Spark

Os arquivos de *log* são gerados em muitos formatos diferentes por uma infinidade de dispositivos e softwares. A análise adequada desses arquivos pode levar a informações úteis sobre vários aspectos de cada sistema. A computação em nuvem (*cloud*) parece ser adequada para esse tipo de análise, pois é capaz de gerenciar a alta taxa de produção, o tamanho grande e a diversidade de arquivos de *log*. Este artigo sugerido como leitura complementar mostra a investigação da análise de arquivos de *log* com as estruturas computacionais em nuvem *Apache*TM, *Hadoop*[®] e *Apache Spark*TM. Desenvolvemos aplicativos realistas de análise de arquivos de *log* em ambas estruturas e realizamos consultas do tipo SQL em arquivos de *log* reais do Apache Web Server. Várias experiências foram realizadas com parâmetros diferentes, a fim de estudar e comparar o desempenho das duas estruturas. *Para realizar a leitura, acesse a plataforma EBSCO Host disponível na Biblioteca Virtual da Kroton, na opção artigos, e busque pelo título do artigo.*

MAVRIDIS, I.; KARATZA, H. Performance evaluation of cloud-based log file analysis with Apache Hadoop and Apache Spark. **Journal of Systems & Software**, [s. l.], v. 125, p. 133–151, 2017.

Introdução à computação usando *Python* – um foco no desenvolvimento de aplicações

O livro sugerido apresenta a linguagem de programação *Python*, como configurar o ambiente de desenvolvimento *Python* e os principais conceitos da ciência da computação, aprender programação e aprender desenvolvimento de aplicações em geral. Traz também exercícios resolvidos para fixação do aprendizado. É sugerida especialmente a leitura do Capítulo 1 – Introdução à Ciência da Computação e do Capítulo 12 – Bancos de Dados e Processamento de Dados. *Para realizar a leitura, acesse a plataforma Minha Biblioteca, disponível na Biblioteca Virtual da Kroton.*

PERKOVIC, L. **Introdução à Computação Usando Python – Um Foco no Desenvolvimento de Aplicações**. Rio de Janeiro: LTC — Livros Técnicos e Científicos Editora Ltda, 2016.

Tema 02 – Manipulação de dados com *Python*

***Python* 3 – Conceitos e aplicações – Uma abordagem didática**

O livro sugerido apresenta a linguagem de programação *Python* 3 e suas aplicações, assim como contempla um estudo sobre algoritmos e lógica de programação. Traz também exercícios resolvidos para fixação do aprendizado. É sugerida especialmente a leitura do Capítulo 1 – *Python*: uma Linguagem de Programação, Capítulo 2 – Objetos e comandos de entrada e saída em *Python* e Capítulo 7 – Arquivos. *Para realizar a leitura, acesse a plataforma Minha Biblioteca, disponível na Biblioteca Virtual da Kroton.*

BANIN, S. L. **Python 3 - Conceitos e Aplicações - Uma abordagem didática**, São Paulo: Editora Érica, 2018.

Use of Python in data manipulation and interfacing spreadsheets (Excel)

O artigo sugerido como leitura complementar enfoca a importância do uso do *Python* na manipulação de dados em planilhas Excel. Mostra também alguns dos benefícios do processamento feito pelo *Python* sobre o uso de planilhas em comparação ao uso de macros. *Para realizar a leitura, acesse a plataforma EBSCO Host disponível na Biblioteca Virtual da Kroton, na opção artigos, e busque pelo título do artigo.*

CHAN, K, B. Use of Python in data manipulation and interfacing spreadsheets (Excel). **Python Papers Monograph**, [s. l.], v. 2, p. 1-5, 2010.

Tema 03 – Organização e visualização de dados


Matplotlib: a 2D graphics environment

Nesse artigo, o autor fala sobre sua experiência em recodificar um programa de computador desenvolvido em *Matlab* para o *Python*. Ele menciona a crescente complexidade do programa original, que incorporou visualizações tridimensionais de imagens médicas, análises espectrais e de séries temporais, e a estrutura de dados necessária para representar dados de sujeitos humanos. Ele fala sobre o processo de recodificação e como o uso foi capaz de atender às suas necessidades com o *Python*. *Para realizar a leitura, acesse a plataforma EBSCO Host disponível na Biblioteca Virtual da Kroton, na opção artigos, e busque pelo título do artigo.*

HUNTER, J. D. Matplotlib: A 2D Graphics Environment. **Computing in Science & Engineering**, [s. l.], v. 9, n. 3, p. 90-95, 2007.

Python: An Ecosystem for Scientific Computing

À medida que a relação entre pesquisa e computação evolui, novas ferramentas são necessárias não apenas para tratar problemas



numéricos, mas também para resolver problemas que envolvem grandes conjuntos de dados em diferentes formatos, novos algoritmos e sistemas computacionais. O *Python* pode ajudar a desenvolver essas ferramentas de pesquisa computacional, oferecendo um equilíbrio de clareza e flexibilidade sem sacrificar o desempenho. O artigo sugerido como leitura complementar exibe a visão geral da linguagem *Python* e algumas bibliotecas aplicadas na área da computação científica. *Para realizar a leitura, acesse a plataforma EBSCO Host disponível na Biblioteca Virtual da Kroton, na opção artigos, e busque pelo título do artigo.*

PÉREZ, F.; GRANGER, B. E.; HUNTER, J. D. Python: An Ecosystem for Scientific Computing. **Computing in Science & Engineering**, [s. l.], v. 13, n. 2, p. 13–21, 2011.

Tema 04 – Análise estatística dos dados

Probabilidade e estatística para engenharia e ciências

O uso de modelos probabilísticos e métodos estatísticos para a análise de dados tem se tornado uma prática comum em praticamente todas as disciplinas científicas. O livro sugerido para leitura complementar abrangente contém uma introdução sobre os modelos e métodos com maior probabilidade de serem encontrados e usados por estudantes de engenharia e ciências de dados. O livro contém exercícios para fixação do conteúdo. É sugerida especialmente a leitura do Capítulo 1 – Visão geral e estatística descritiva, Capítulo 2 – Variáveis aleatórias discretas e distribuições de probabilidade e Capítulo 14 – Testes de qualidade do ajuste e análise de dados categorizados. *Para realizar a leitura, acesse a plataforma Minha Biblioteca, disponível na Biblioteca Virtual da Kroton.*

DEVORE, J. L. **Probabilidade e estatística para engenharia e ciências**, Tradução da 9ª edição norte-americana. São Paulo: Editora Cengage, 2018.

Probabilidade e estatística

A Estatística é um dos mais populares ramos da Matemática devido ao tratamento científico de dados amostrais que permite aplicar o

raciocínio indutivo ou realizar análises exploratórias de dados, a fim de inferir algum conhecimento sobre a população objeto de estudo. Pesquisadores das áreas das Ciências Exatas, Sociais e Aplicadas e da Saúde utilizam a Estatística para tarefas como calcular estimativas, computar índices diversos, apresentar dados de forma sistemática por meio de gráficos e/ou tabelas, fazer previsões e testar hipóteses. O livro contém exercícios para fixação do conteúdo. É sugerida especialmente a leitura do Capítulo 1 – Amostragem e capítulo 7 – Regressão e Correlação. *Para realizar a leitura, acesse a plataforma Minha Biblioteca, disponível na Biblioteca Virtual da Kroton.*

LOESCH, C. **Probabilidade e Estatística**. Rio de Janeiro: LTC — Livros Técnicos e Científicos Editora Ltda., 2015.

► Tema 05 – *Machine Learning* em Python

Machine Learning Made Easy: A Review of “Scikit-learn” Package in Python Programming Language

O artigo sugerido como leitura complementar apresenta uma visão geral do aprendizado de máquina (*machine learning*) e da biblioteca *Scikit-learn*, escrita em *Python*, destacando as principais diferenças de *machine learning* em comparação com as inferências estatísticas mais familiares à comunidade de estatísticas educacionais e comportamentais. O objetivo deste artigo é fornecermos aos pesquisadores um senso geral sobre como o pacote *Scikit-learn* pode ser usado e explicamos por que vale a pena aprender. Incentivamos os leitores interessados a conferir mais detalhes e tutoriais no site do *Scikit-learn*. *Para realizar a leitura, acesse a plataforma EBSCO Host disponível na Biblioteca Virtual da Kroton.*

HAO, J.; HO, T. K. Machine Learning Made Easy: A Review of “Scikit-learn” Package in Python Programming Language. **Journal of Educational and Behavioral Statistics**, [s. l.], v. 44, n. 3, p. 348–361, 2019.

Introdução a **Big Data** e Internet das Coisas (IOT)

No aprendizado de máquina (*machine learning*), temos uma junção entre recursos computacionais, inteligência artificial, dados, dentre outros elementos. Esses sistemas devem estar aptos não só a memorizar dados, mas também a observá-los e explorá-los para que suas habilidades evoluam por meio da prática e, consequentemente, da construção do conhecimento. É sugerida especialmente a leitura do capítulo Aprendizado de máquina (*machine learning*). Para realizar a leitura, acesse a plataforma Minha Biblioteca, disponível na Biblioteca Virtual da Kroton.

MORAIS, I. S.; GONÇALVES, F, P. D.; LEDUR, L, C.; JUNIOR, C., R. **Introdução a Big Data e Internet das Coisas (IoT)**. Porto Alegre: SAGAH, 2018.

► Tema 06 – Processando **Big Data** com **Spark**

Apache Spark: A Unified Engine for Big Data Processing

O artigo sugerido como leitura complementar discute a estrutura de computação *Apache Spark*, que unifica as cargas de trabalho de *streaming*, *batch* e *big data* interativas para desbloquear novos aplicativos. Os tópicos incluem o uso do *Spark* de um modelo de programação RDD, o uso do *Spark* em diversas aplicações, como processamento em lote e processamento de imagens, e o custo adicional do *Spark* em outros sistemas especializados devido à tolerância a falhas. Para realizar a leitura, acesse a plataforma EBSCO Host disponível na Biblioteca Virtual da Kroton.

ZAHARIA, M. Apache Spark: A Unified Engine for Big Data Processing. **Communications of the ACM**, [s. l.], v. 59, n. 11, p. 56–65, 2016.

Introdução a *Big Data* e Internet das Coisas (IOT)


Atualmente, o tema *Big Data* desperta o interesse em todas as pessoas que têm algum envolvimento com atividades relacionadas à gestão da informação. Porém, quem não tem afinidade com a tecnologia da informação (TI), pode encontrar dificuldades para entender do que se trata. Com o aumento significativo da quantidade de dados gerados pela internet e com o surgimento das mídias sociais, é necessário gerenciar e armazenar as informações de maneira organizada. Esses dados podem ser classificados em estruturados, não estruturados e semi-estruturados com base no seu gerenciamento e armazenamento. É sugerida especialmente a leitura dos capítulos *Big Data* e Internet das Coisas (IoT), Introdução à Ciência de dados e *Big Data*, Aplicações da ciência de dados e Descoberta de conhecimento com *Big Data*. Para realizar a leitura, acesse a plataforma Minha Biblioteca, disponível na Biblioteca Virtual da Kroton.

MORAIS, I. S.; GONÇALVES, F, P. D.; LEDUR, L, C.; JUNIOR, C., R. **Introdução a Big Data e Internet das Coisas (IoT)**. Porto Alegre: SAGAH, 2018.

► Tema 07 – *Real Time Analytics* com *Python* e *Spark*

Real-Time Big Data Stream Processing Using GPU with Spark Over Hadoop Ecosystem

O artigo sugerido como leitura complementar propõe um sistema de processamento de fluxo de *Big Data* em tempo real para analisar os dados e tomar decisões imediatas usando *Apache Spark* e Hadoop. Os serviços existentes (sites sociais, redes, aplicativos da web etc.) e várias tecnologias de sensores estão produzindo gigabytes de dados dentro de alguns milissegundos continuamente. Além disso, nesta era tecnológica, vários dispositivos e objetos conectados à internet (Internet das Coisas – IoT) esse número chegaria a 50 bilhões em 2020. Com esse número,



pode-se imaginar a quantidade de geração de dados em alta velocidade. Essa enorme quantidade de dados de alta velocidade é denominada *Big Data*. Para realizar a leitura, acesse a plataforma EBSCO Host disponível na Biblioteca Virtual da Kroton.

RATHORE, M. M. Real-Time Big Data Stream Processing Using GPU with Spark Over Hadoop Ecosystem. **International Journal of Parallel Programming**, [s. l.], v. 46, n. 3, p. 630–646, 2018.

An experimental survey on big data frameworks

Recentemente, são geradas quantidades cada vez maiores de dados a partir de uma variedade de fontes. As tecnologias de processamento de dados existentes não são adequadas para lidar com as enormes quantidades de dados gerados. No entanto, muitos trabalhos de pesquisa concentram-se no *Big Data*. As estruturas propostas recentemente para aplicativos de *Big Data* ajudam a armazenar, analisar e processar os dados. Neste artigo, os autores discutem os desafios do *Big Data* e pesquisam as estruturas existentes de *Big Data*. Também apresentam uma avaliação experimental e um estudo comparativo das estruturas mais populares de *Big Data* com várias cargas de trabalho representativas em lote e interativas. Esta pesquisa é concluída com uma apresentação das melhores práticas relacionadas ao uso das estruturas estudadas em vários domínios de aplicativos, como aprendizado de máquina, processamento de gráficos e aplicativos do mundo real. Para realizar a leitura, acesse a plataforma ScienceDirect – Journals & Books e pesquise pelo título do artigo.

INOUBLI, W.; ARIDHI, S.; MEZNI, H.; MADDOURI, M.; NGUIFO, E. M. An experimental survey on big data frameworks. **Future Generation Computer Systems**, 2018.

The background features abstract geometric shapes. In the top right, there's a yellow triangle with a green gradient. On the left, a blue triangle points upwards. In the bottom left, a large yellow circle overlaps a smaller blue circle. The bottom right corner is filled with a solid blue shape.

Bons estudos!