

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



Definindo um projeto de Big Data e as bases e arquivos a considerar. Arquivos de dados com Python

Bloco 1

Marcelo Tavares de Lima





► Objetivos

- Apresentar conceitos fundamentais de *Big Data*.
- Descrever sobre tratamento de dados em linguagem Python.
- Apresentar aplicações de *Big Data* em linguagem Python.



► Conceitos fundamentais de *Big Data*

- *Big Data*, no sentido literal da palavra, significa grande volume de dados. No entanto, o sentido prático e real do termo não se limita somente a isso, pois é muito mais.
- Além de se referir a grandes volumes de dados, também se refere à grande variedade de dados manipulados, das diversas fontes disponíveis, e também à velocidade em que os dados são tratados.



► Conceitos fundamentais de *Big Data*

- Não existe um consenso quando se trata do conceito básico de *Big Data*.
- Para exemplificar, Taurion (2013, [s.p.]) apresenta a definição dada pela *McKinsey Global Institute* como: “o termo *Big Data* refere-se a este conjunto de dados cujo crescimento é exponencial e cuja dimensão está além da habilidade das ferramentas típicas de capturar, gerenciar e analisar dados”.



► Conceitos fundamentais de *Big Data*

- Uma analogia feita por Taurion (2013), com respeito a *Big Data* e medicina, afirma que *Big Data* é um microscópio que permitiu que se vissem coisas que já existiam, como bactérias e vírus, mas não eram conhecidas. Essa é a descoberta do conhecimento por meio dos dados.



► Conceitos fundamentais de *Big Data*

- O espaço ocupado por esses novos dados, que passaram a ser vistos e enxergados como fonte de informação para os negócios, é muito maior, e sua produção em massa requer espaços amplos para seu armazenamento.



► Conceitos fundamentais de *Big Data*

- Por exemplo, o uso de imagens e vídeos divulgados nas redes sociais requer muito espaço em memória do que um simples dado numérico.
- Um vídeo em alta definição ocupa muito mais espaço para armazenamento em comparação a uma página de texto, e assim por diante.



► Conceitos fundamentais de *Big Data*

- Amaral (2016, [s.p.]) apresenta uma definição para *Big Data*: “*Big Data* é o fenômeno em que dados são produzidos em vários formatos e armazenados por uma grande quantidade de dispositivos e equipamentos”.



► Conceitos fundamentais de *Big Data*

- As causas do fenômeno *Big Data*, basicamente, são associadas aos investimentos feitos em tecnologia, como, por exemplo, investimentos em unidade central de processamento (CPU), memórias e unidades de armazenamento, dentre outros equipamentos, tornando-os cada vez mais baratos.

► Tratamento de dados com Python

- Segundo Santos (2018), a linguagem Python foi criada em 1989, pelo pesquisador Guido Van Rossum, do *National Research Institute for Mathematics and Computer Science in Amsterdam* (CWI).
- O nome Python foi dado à linguagem por conta de um seriado de comédia que existia na época, cujo nome era *Tropa Monty Python*.



► Tratamento de dados com Python

- O ambiente de desenvolvimento da linguagem Python ou IDE (*Integrated Development Environment*), assim como para a linguagem R, é muito utilizado por conta de uma série de facilidades que trazem para o uso da linguagem.



► Tratamento de dados com Python

- Em número também são vários, pois além de versões distintas, diferentes IDEs e diferentes plataformas exigem algumas condições para que possam ser instalados.
- Também existem versões gratuitas e versões pagas. A interface (IDE) que será utilizada para o desenvolvimento de linguagem de programação Python será a Anaconda-Spyder e a Jupiter Notebook, que pode ser encontrada facilmente na Internet.



► Tratamento de dados com Python

Santos (2018) afirma que:

O mercado de trabalho para programadores demanda Python como uma das principais linguagens de programação, isso porque o seu uso para programação nas áreas de ciência de dados, análise de dados e inteligência artificial como um todo faz uso principalmente dessa linguagem. (SANTOS, 2018, [s. p.])



► Tratamento de dados com Python

Segundo Santos (2018), dentre os tipos de linguagens existentes, classificadas como compiladas e interpretadas, a linguagem Python é considerada uma linguagem interpretada simples.



► Tratamento de dados com Python

É considerada uma linguagem interpretada simples, pois para sua compilação precisa de um interpretador interno à máquina onde é inserida.

A função do interpretador é traduzir a linguagem Python para a linguagem de máquina.

O código fonte da linguagem Python é convertido para *bytecode*, que tem formato binário e instruções para o interpretador.

Definindo um projeto de Big Data e as bases e arquivos a considerar. Arquivos de dados com Python

Bloco 2

Marcelo Tavares de Lima





► Tratamento de dados com Python

O tratamento de grandes volumes de dados, por meio da linguagem Python, pode ser feito pela plataforma Hadoop, de computação distribuída, com alta escalabilidade, de grande confiabilidade e muito tolerante a falhas.



► Tratamento de dados com Python

- De acordo com o site do Apache Hadoop, o Hadoop foi projetado para ser dimensionado de maneira fácil para qualquer quantidade de máquinas, com a ajuda do poder computacional e de armazenamento.
- A Hadoop foi criada por Doug Cutting e Mike Cafarella, no ano de 2005, segundo Madhavan (2015). O nome Hadoop era o nome do elefante de brinquedo do filho de Doug Cutting.



► Tratamento de dados com Python

O MapReduce é um conjunto de bibliotecas que permite realizar processamento em paralelo, de grandes quantidades de dados, usando todo o hardware disponível em um *cluster* Hadoop. Divide o processamento em duas etapas principais:

- 1) Mapeamento e validação dos dados (chamado MAP).
- 2) Os dados validados na etapa MAP são tratados gerando como resultados os valores finais do processo.



► Tratamento de dados com Python

- A implementação do MapReduce é feita pelo Apache Hadoop, que é *open source* e fornece estrutura para executar as aplicações de um modelo de MapReduce, segundo Araújo e Montini (2016). Também “permite o armazenamento confiável e distribuído de dados, utilizando dispositivos de armazenamento baratos (*commodity hardware*)”. (ARAÚJO; MONTINI, 2016, p. 92).



► Tratamento de dados com Python

- Os dados a serem analisados são armazenados de forma distribuída através do sistema de arquivos distribuídos *Hadoop Distributed File System* (HDFS), o qual é uma implementação de código aberto (*open source*) da *Google File System* (ARAÚJO; MONTINI, 2016).



► Tratamento de dados com Python

- A plataforma Hadoop pode ser obtida no site da empresa. (HADOOP, 2019)
- Para executar códigos em linguagem Python, pode-se utilizar a interface de programação de aplicação (API) de *streaming* do Hadoop, que auxilia no uso de programação que possua uma entrada e saída padrão, tal como um programa MapReduce.

PÓS-GRADUAÇÃO

Teoria em prática


Bloco 3

Marcelo Tavares de Lima





► Teoria em prática

- Na empresa em que você trabalha são produzidos muitos dados sobre uma série de medidas e registros.
 - No departamento em que você trabalha, por exemplo, o departamento de marketing, você e seus pares produzem dados sobre campanhas e sobre intenção de consumo de clientes de vários ramos de produtos.
- 



► Teoria em prática

- A grande massa de dados produzida, traz para vocês um grande desafio não só para armazenamento, mas também para tratamento e análise.
- Você recebeu um treinamento recente sobre linguagem Python, o que fez com que tivesse a ideia de utilizar esse recurso computacional para realizar o tratamento e a análise dessa grande massa de dados que sua empresa produz. Você acha que será possível implementar sua ideia? Como faria?

Dica do professor

Bloco 4

Marcelo Tavares de Lima



► Indicação de livro

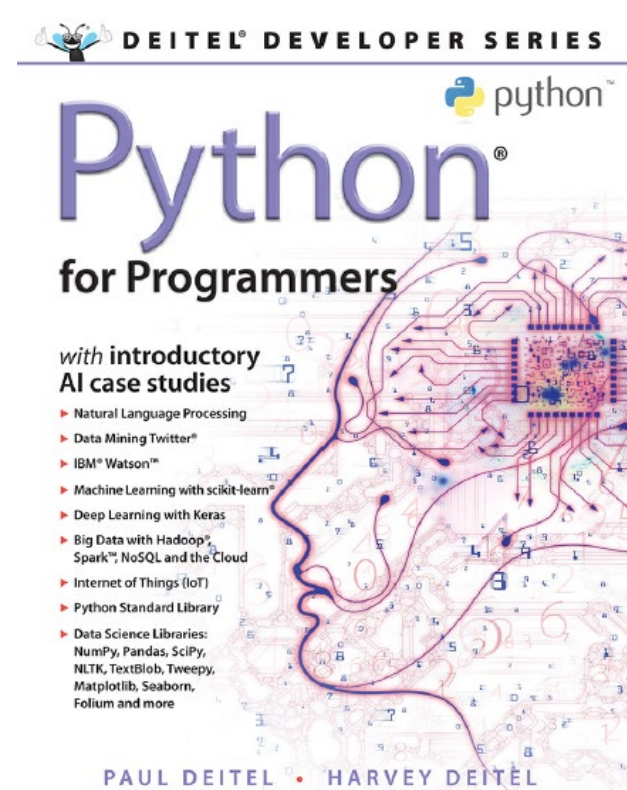
Título: Python for Programmers.

Autores: Paul Deitel, Harvey Deitel.

Editora: Pearson Education, Inc.

Ano de publicação: 2019.

Figura 1 - Livro



Fonte:

<https://www.oreilly.com/library/view/python-for-programmers/9780135231364/>.

Acesso em: 04 fev. 2020.



► Referências

AMARAL, F. **Introdução a ciência de dados**: mineração de dados e Big Data. Rio de Janeiro: Alta Books, 2016. KINDLE.

ARAÚJO, A. C.; MONTINI, A. A. Técnicas de Big Data e projeção de risco de mercado utilizando dados em alta frequência. **Future Studies Research Journal**. São Paulo, v. 8, n.3, p. 83-108, Disponível em: <https://revistafuture.org/FSRJ/article/view/219/375>. Acesso em: 04 fev. 2020.

HADOOP. **Apache™ Hadoop®**. 2019. Disponível em: <https://hadoop.apache.org/>. Acesso em: 04 fev. 2020.

MADHAVAN, S. **Mastering Python for Data Science**. Brimingham, UK: Packt Publishing, 2015. Disponível em: <http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1058787&lang=pt-br&site=ehost-live>. Acesso em: 04 fev. 2020.

SANTOS, R. F. V. C. **Python**: guia prático do básico ao avançado. Série cientista de dados. 2018. E-BOOK KINDLE.

TAURION, C. **Big Data**. Rio de Janeiro: Brasport, 2013. EPUB. Disponível em: <https://bv4.digitalpages.com.br/#/legacy/epub/160676>. Acesso em: 04 fev. 2020.

