Ciência de Dados com Python

Fundamentos e aplicações

Objetivo do Curso

Neste curso o aluno compreenderá os fundamentos de Ciência de dados, exemplos de aplicações de Técnicas de Data Science nas mais diversas áreas, sendo um curso brevemente introdutório.

“Uma jornada de mil milhas começa com um único passo” - Provérbio Chinês.

Sequência

- ❑ Introdução à Ciência de Dados
- ❑ Profissional de Ciência de Dados
- ❑ Aplicações Ciência de dados
- ❑ Introdução ao Big Data
- ❑ Ecossistema Big Data
- ❑ Introdução ao Python
- ❑ Breve História do Python
- ❑ Variáveis e Tipos Embutidos
- ❑ Estruturas de Decisão
- ❑ Estruturas de repetição
- ❑ Estruturas de dados (Listas, Tuplas, Set, Dicionário)
- ❑ Funções
- ❑ Manipulação (Pandas, Numpy)

Ciência de Dados a profissão mais Sexy do Século 21 - Harvard Business Review

Harvard Business Review

Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)

Fonte: Harvard Business Review, 2012

Interesse por Ciência de Dados 2012 - 2022

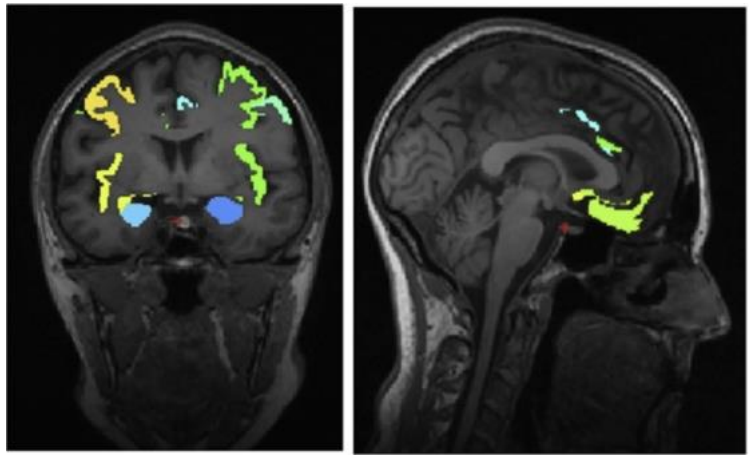
Fonte: Google Trends, 2022

Cientista de Dados – Exempls Campos de Atuação

Mercado Financeiro

Fonte: Mercado Financeiro - r7.com

Saúde



Fonte: Data Drive - How Data Science is Being Used in HealthCare, 2021

Marketing Digital - Ecommerce

Fonte: Black Friday – Bem Paraná, 2016

Segurança Pública

Fonte: South China Morning Post, 2018

“O que é Ciência ?”

“O que são Dados ?”

Introdução à Ciência de Dados

O que é Ciência ?

Ciência é conhecimento sobre a estrutura e comportamento do mundo natural e físico, com base em fatos que você pode provar, por exemplo, por experimentos.

O que são Dados ?

Uma coisa dada ou concedida; algo conhecido ou assumido como fato, e feito a base do raciocínio ou cálculo; uma suposição ou premissa a partir da qual são tiradas inferências.

Fonte: Oxford Advanced American Dictionary

O que é Ciência de Dados ?

Fonte: Dados ao Cubo – Profissão: Cientista de Dados Parte I, 2020.

"A ciência de dados é um **campo interdisciplinar** que usa **métodos científicos**, processos, algoritmos e sistemas para **extrair conhecimento** e **insights de dados** ruidosos, estruturados e não estruturados, e aplicar o conhecimento de dados em uma ampla gama de domínios de aplicação" (**Fonte:** Wikipédia - Data Science, 2022)

A era do Big Data

A era do Big Data

- Em 2020, cada pessoa gerou 1.7 megabytes por segundo – (Fonte: IBM).
- Google recebe mais de 3,5 bilhões de buscas diariamente (Fonte: Internet Live Stats).
- O mercado de análise de Big Data na área da saúde pode valer US$ 67,82 bilhões até 2025 (Fonte: Globo News Wire).
- Os usuários do WhatsApp trocam até 65 bilhões de mensagens diariamente (Fonte: *Connectiva Systems*).
- Hoje, uma pessoa levaria aproximadamente 181 milhões de anos para baixar todos os dados da internet. *(Fonte: Unicorn Insights).*
- Usando big data, a Netflix economiza US$ 1 bilhão por ano em retenção de clientes. *(Fonte: Statista, Inside Big Data)*

O que é Big Data ?

Não existe um consenso formal do que seja Big Data. De acordo com a IBM “Big data é um termo aplicado a conjuntos de dados cujo tamanho ou tipo está além da capacidade de bancos de dados relacionais tradicionais de capturar, gerenciar e processar os dados com baixa latência. O big data tem uma ou mais das características a seguir: grande volume, alta velocidade ou grande variedade (3V)” (Fonte IBM).

Fonte: IBM big-data-analytics, 2022

❑ Volume de Dados

Refere-se à enorme quantidade de dados envolvidos.

Estima-se que até 2020, existam cerca de 35 ZB de dados armazenados no mundo.

Um ZB (Zettabyte) equivale a 10^{21} bytes, ou 1 bilhão de Terabytes.

De acordo com o IDC (2011), a informação do mundo dobra a cada dois anos.

❑ Variedade

Os dados incluem não apenas dados estruturados, por exemplo, (bancos de dados comuns / estruturados), mas também oriundos de:

- Páginas Web
- Índices de pesquisa
- Arquivos de log
- Fóruns
- Mídias sociais
- E-mails
- Dados de sensores variados
- Áudio e vídeo

❑ Variedade

Os sistemas tradicionais não conseguem armazenar, processar e entender essa vasta gama de dados.

Assim, deve-se utilizar novas tecnologias, algoritmos e técnicas para realizar a analise desses dados, tanto estruturados quando não estruturados, em conjunto.

No geral, apenas 20% do volume de dados é estruturado, sendo 80% restantes, não estruturados.

❑ Velocidade dos Dados

Os dados são gerados em grande velocidade.

Definimos essa velocidade de acordo com o quão rápido os dados são resgatados, armazenados e recuperados.

Basicamente, falamos em taxa de fluxo de dados quando nos referimos à sua velocidade.

Assim, o fluxo de (geração e transmissão) de dados pode se tornar tão elevado, que os sistemas tradicionais de análise não conseguem manipulá-los.

Tipos de dados

Dados estruturados são aqueles organizados e representados com uma **estrutura rígida**, a qual foi previamente planejada para armazená-los.

0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

Exemplos. Qualquer Dados dentro de um Banco de Dados Relacional.

Dado Estruturado

Imagens, assim como **vídeos** ou **arquivos de áudio**, são também exemplos de dados não estruturados.

Não estruturado

Não se encaixam em um banco de dados, mas são estruturados por (tags). Sua estrutura pode ser classificada, mas não sempre.

JSON
XML,
RDF,
OWL

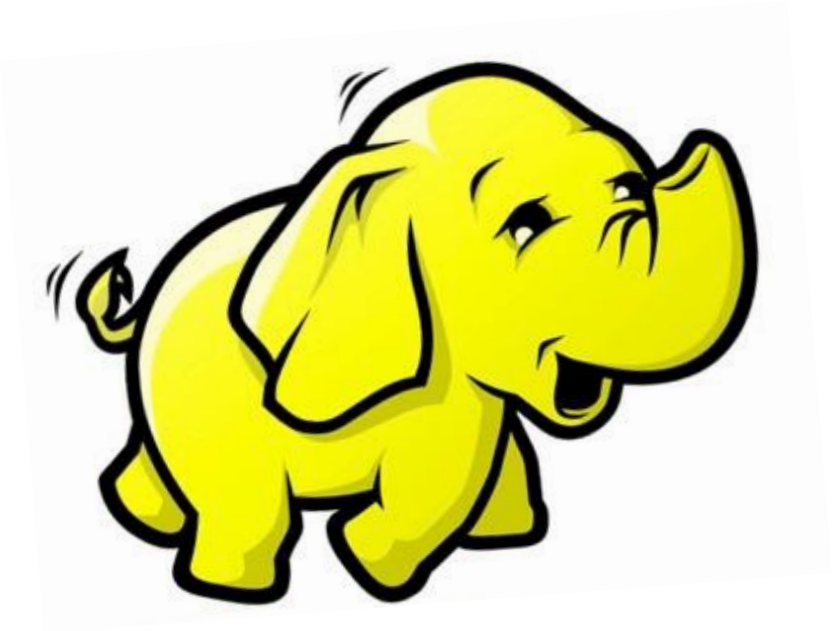
Semi Estruturado

Big Data - Frameworks

Fonte: Open Source Technology Landscape, 2021

Hadoop

Apache Hadoop é um framework livre administrado pela Apache Software Foundation construído em Java para computação distribuída, de alta escalabilidade, grande confiabilidade e tolerância a falhas.



Doug Cutting - Criador Hadoop

Fonte: Techgoondu, 2015

Hadoop

Hadoop é um **framework** livre administrado pela Apache Software Foundation construído em Java para **computação distribuída**, de alta escalabilidade, grande confiabilidade e tolerância a falhas.

- Tolerância a falhas
- Balanceamento de carga
- Comunicação entre máquinas
- Distribuição de tarefas
- Alocação de máquinas
- Ordenação dos dados
- Transferências de dados
- Escalonamento de jobs

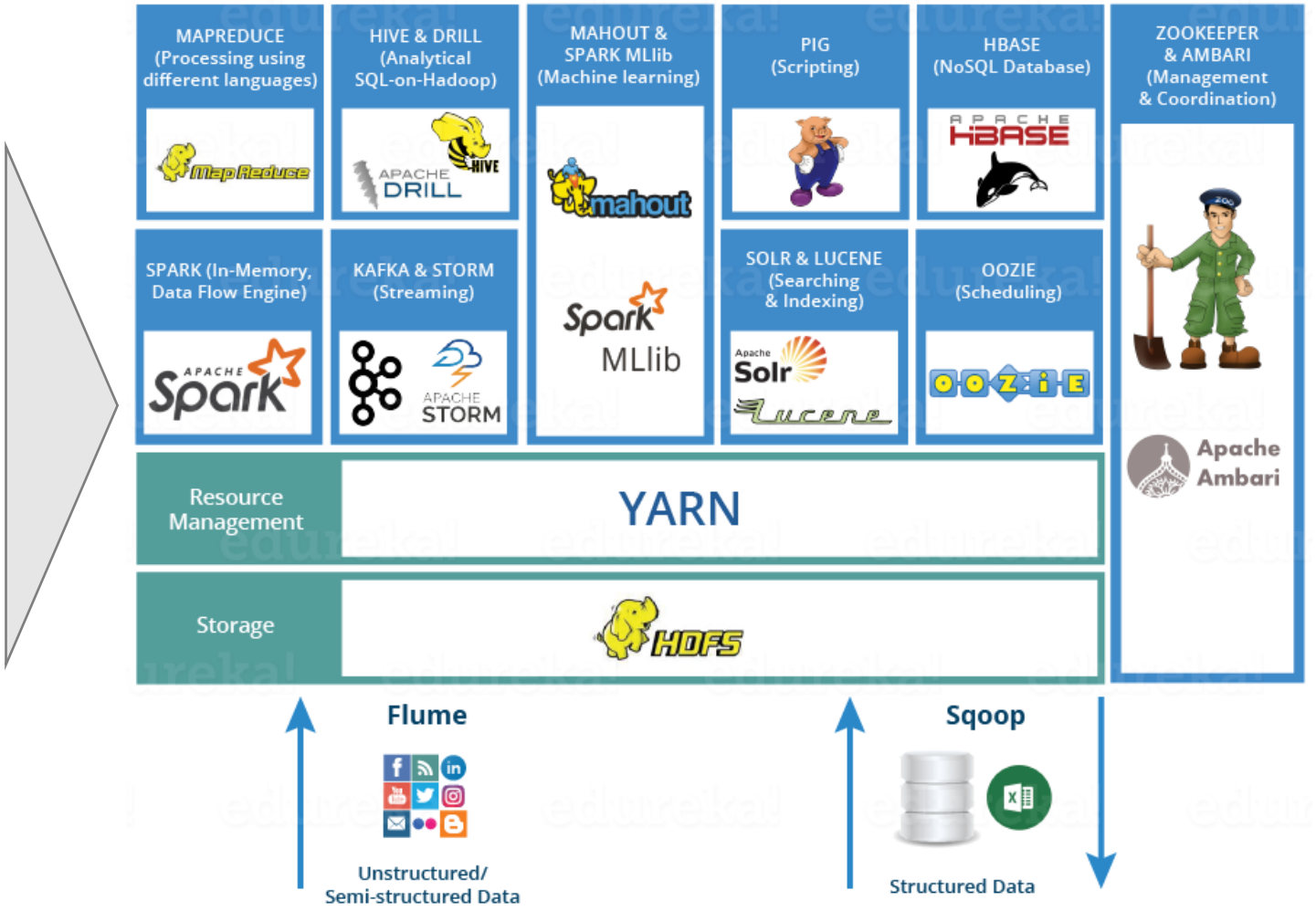
Doug Cutting - Criador Hadoop

Fonte: Techgoondu, 2015

Ecossistema Hadoop

Analogia

Ecossistema



Fonte: Edureka - What is the purpose of Zookeeper in Hadoop Ecosystem , 2018**.**

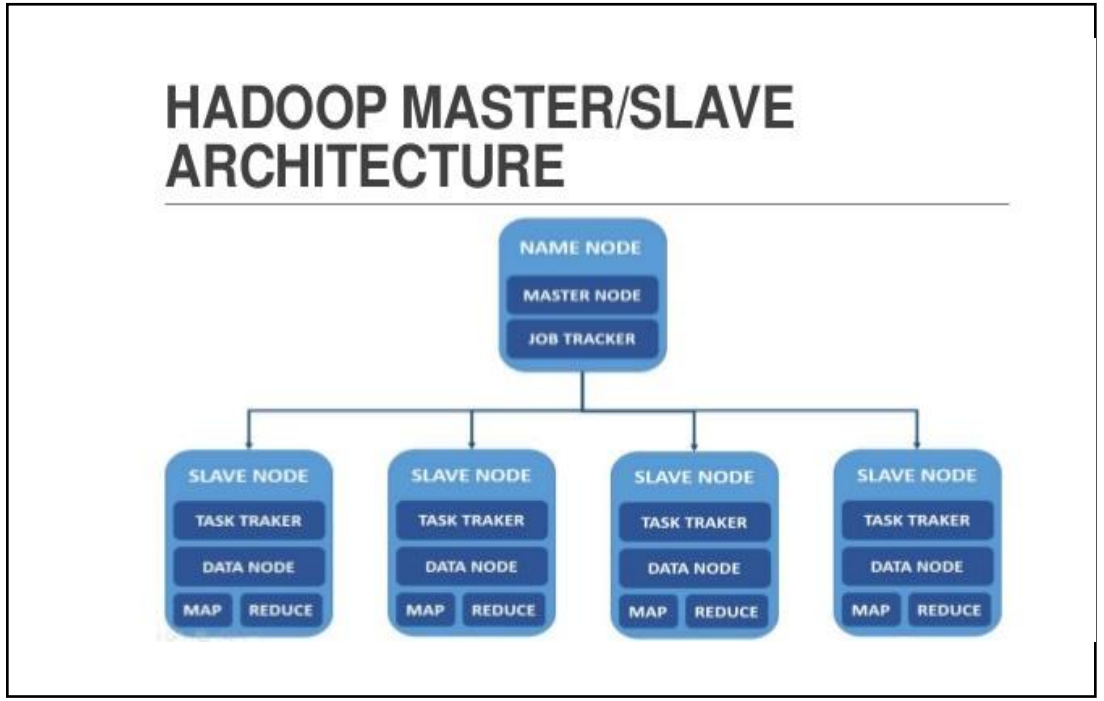
HDFS – Hadoop Distributed File System

Hadoop Distributed File System (HDFS), é responsável por gerenciar o disco das máquinas que compõem o Cluster. HDFS também serve para leitura e gravação dos dados.

HDFS - Características

- Tolerância a falhas e recuperação automática
- Portabilidade entre hardware e sistemas iguais
- Escalabilidade para armazenar grande volume de dados
- Confiabilidade, através de diversas cópias de dados

Nodes HDFS

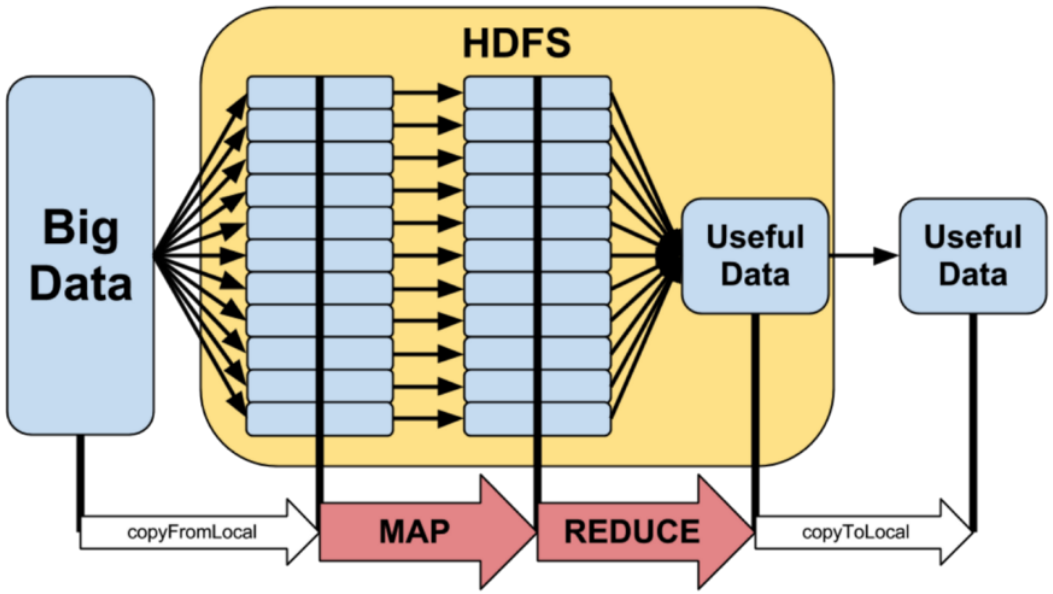


Fonte: Medium - Hadoop Cluster Setup with HDFS Architecture Demo, 2020.

Map Reduce

O MapReduce, é responsável por gerenciar o **processamento dos dados em um ambiente Cluster** (conjunto de máquinas). **Cluster** é um conjunto de computadores que tem o mesmo propósito de processamento para uma aplicação

Map Reduce



Fonte: Medium - Hadoop Cluster Setup with HDFS Architecture Demo, 2020.

- Flexibilidade – processa todos os dados independente do tipo e formato
- Confiabilidade – permite que os Jobs sejam executados em paralelo
- Acessibilidade – suporte a diversas linguagens de programação.

Perguntas ?

Referências Bibliográficas

Wikipédia – Ciência de Dados, 2022. Disponível em: https://pt.wikipedia.org/wiki/Ci%C3%AAncia_de_dados . Acesso em Junho de 2022.

ORACLE - O que é Ciência de Dados?. Disponível em: https://www.oracle.com/br/data-science/what-is-data-science/. Acesso em Junho de 2022.

Dados ao Cubo - Profissão: Cientista de Dados Parte I, 2020. Disponível em: https://dadosaocubo.com/profissao-cientista-de-dados-parte-i/. Acesso em Junho de 2022.

Medium - Big Data - An Introduction. Disponível em: https://medium.com/analytics-vidhya/big-data-an-introduction-b7bc048081c9. Acesso em Junho de 2022.

Luz .G – Medium - A Era do Big Data, 2019. Disponível em: https://medium.com/gabriel-luz/a-era-do-big-data-64ebad5859f2. Acesso em Junho de 2022.

SAP - O que é Big Data – Disponível em: https://www.sap.com/brazil/insights/what-is-big-data.html. Acesso em Junho de 2022.

Referências Bibliográficas

D. Yuki , Medium - Ciência de Dados e seus conceitos, 2021. Disponível em: https://medium.com/permalink-univesp/ci%C3%AAncia-de-dados-e-suas-aplica%C3%A7%C3%B5es-em-diversas-%C3%A1reas-8f6119e7d789. Acesso em Junho de 2022.