ARA0168
TÓPICOS DE BIG DATA EM PYTHON

Aula 1 – Introdução a Disciplina de Tópicos de Big Data em Python

Universidade Estácio de Sá

Prof. Simone Gama

simone.gama@estacio.br

BIG DATA EM PYTHON: Ementa Geral

- **Princípios de Big Data.**
- **Introdução ao Hadoop e Armazenamento de Dados.**
- **Princípios de Desenvolvimento com PySpark.**
- **Análise de Dados em Python com Pandas.**
- **Big Data Analytics.**

Processo de Avaliação

Avaliação Nota Final (**NF**)

Para a **Aprovação** na disciplina, o aluno deverá:

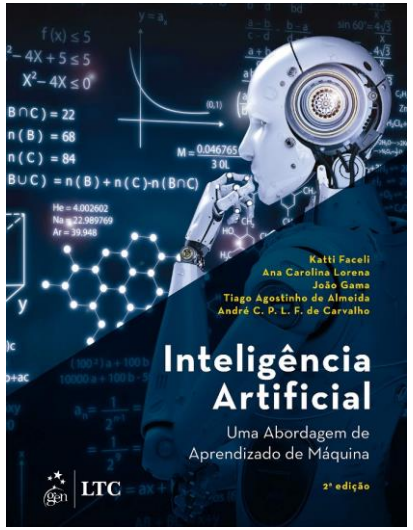
- Desenvolvimento de Trabalho a ser definido.
- **A média aritmética obtida será o grau final do aluno na disciplina, no mínimo 6,0**.
- Frequentar, no mínimo, 75% das aulas ministradas.

Informações Gerais

- **Frequências**: O aluno(a) deve ter Presença **confirmada** igual ou superior a **75% das aulas ministradas**. Presença menor que esse valor o aluno será considerado Reprovado por Falta (**RF**) ao final do semestre na disciplina.

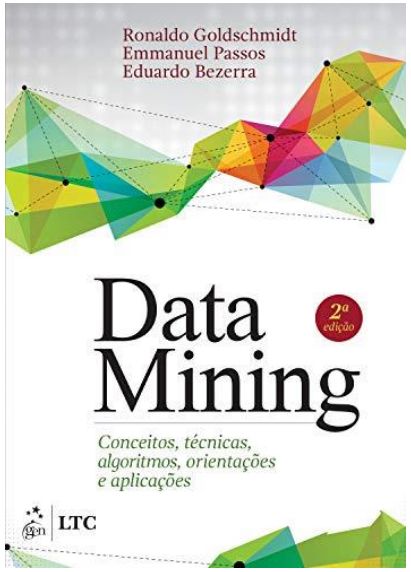
- **Exercícios e Avaliações**: Imagens e/ou trabalhos **copiados** de outros, **códigos de programação claramente semelhantes e/OU iguais** serão **desconsiderados** e em caso de Avaliações, terão as notas devidamente descontadas.

Bibliografia Básica: Big Data

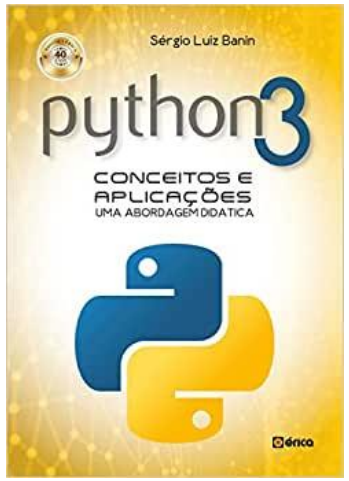


FACELI, Katti. **Inteligência Artificial Uma abordagem de aprendizado de máquina.** 2ª Ed. Rio de Janeiro: LTC, 2021. Disponível em: https://integrada.minhabiblioteca.com.br/books/9788521637509/

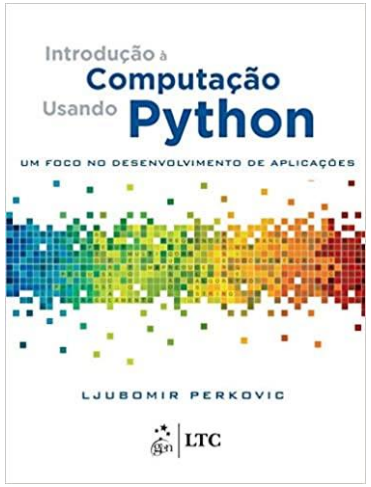
GOLDSCHMIDT, Ronaldo. **Data Mining Conceitos, técnicas, algoritmos, orientações e aplicações**. Rio de Janeiro: Elsevier, 2015. Disponível em: https://integrada.minhabiblioteca.com.br/books/9788595156395/



Bibliografia Básica: Linguagem Python



BANIN, Sérgio Luiz. **Python 3 - Conceitos e Aplicações** - Uma Abordagem Didática. São Paulo: Érica, 2018.



PERKOVIC, Ljubomir. **Introdução à Computação Usando Python - Um Foco no Desenvolvimento de Aplicações.** Rio de Janeiro: LTC, 2016.

ARA0168
TÓPICOS DE BIG DATA EM PYTHON

1.1 – Definição de Big Data

Universidade Estácio de Sá

Prof. Simone Gama

simone.gama@estacio.br

Volume de Dados

- O **Twitter** registra aproximadamente 500 milhões de tweets por dia, são cerca de 6.000 tweets por segundo;
- 42 milhões de acessos são registrados pelo **LinkedIn** apenas de dispositivos móveis, como os smartphones;
- 95 milhões de fotos e vídeos são postados diariamente no **Instagram**;
- O **Facebook** diz que ocorrem mais ou menos 4.000.000.000 de curtidas por minuto na plataforma;
- Bilhões de transações financeiras com **cartões de crédito e débito** são registradas diariamente;
- **Em 2022 haverá, mais ou menos 5,2 gigabytes de informação sobre cada pessoa do planeta**.

BIG DATA: O que é?

BIG DATA: O que é? (1)

- ***Big data*** é um termo que descreve o grande volume de dados que inunda uma empresa no dia a dia. Assim, o big data pode ser analisado em busca de *insights* que levam a melhores decisões e movimentos estratégicos de negócios.
- Em relação a esse volume total de dados, o que importa não é necessariamente a sua quantidade, **mas o que as organizações fazem com os dados**.

Fonte: Big Data: o que é, como funciona e como aplicar? - TOTVS

BIG DATA: O que é? (2)

- A ***Big Data*** são quantidades de dados que os recursos tradicionais de processamento são incapazes de tratar em tempo razoável.
- Via de regra, o Big Data é composto por uma quantidade inacreditavelmente grande de dados. Estamos falando de dezenas de centenas de *terabytes*.

Fonte: Big Data Analytics: O Que É e Por Que Aplicar Na Sua Empresa? (marketingconteudo.com)

BIG DATA: O que é? (3)

Big Data é a **<u>análise</u>** e a **<u>interpretação</u>** de grandes volumes de dados de grande variedade. Para isso são necessárias soluções específicas para Big Data que permitam a profissionais de TI trabalhar com informações não-estruturadas a uma grande velocidade.

Fonte: O que é Big Data? - Canaltech

BIG DATA: O que é?

Oracle and Gartner

A **Big Data** são dados com maior **variedade** que chegam em **volumes** crescentes e com **velocidade** cada vez maior.

Fonte:

- O que é Big Data? | Oracle Brasil
- Definition of Big Data - IT Glossary | Gartner

BIG DATA: Os 5 V's

Volume	**A quantidade de dados importa**. O Big Data processa grandes volumes de dados não estruturados de baixa densidade. Podem ser dados de valor desconhecido, como feeds de dados do Twitter, fluxos de cliques em uma página web ou em um aplicativo para dispositivos móveis, ou ainda um equipamento habilitado para sensores.
Velocidade	**Velocidade é a taxa mais rápida na qual os dados são recebidos** e talvez administrados. Normalmente, a velocidade mais alta dos dados é transmitida diretamente para a memória, em vez de ser gravada no disco.
Variedade	**Variedade refere-se aos vários tipos de dados disponíveis**. Tipos de dados tradicionais foram estruturados e se adequam perfeitamente a um **banco de dados relacional**. Com o aumento de big data, os dados vêm em novos tipos de dados não estruturados. Tipos de dados não estruturados e semiestruturados, como texto, áudio e vídeo, exigem um pré-processamento adicional para obter significado e dar suporte a **metadados**.

BIG DATA: Os 5 V's

Veracidade	Entre as milhares de informações que são geradas todos os dias, **muitas podem ser falsas e é preciso excluí-las da sua análise**. Quando se entende o que é Big Data e para que serve, descobre-se que esse processo ajuda a "filtrar" o que é real do que não é. Um dos princípios para isso é que, se várias fontes apontam para determinada informação, entende-se que aquela é a real.
Valor	O objetivo de ter acesso a tantas informações é fazer com que elas agreguem, de alguma forma, **valor para a sua empresa**. O Big Data tem justamente esse propósito: fazer uma análise precisa dessas informações e **gerar *insights*** valiosos para os gestores que irão utilizá-las.

BIG DATA: Histórico

Embora o conceito de **big data** em si **seja relativamente novo**, as origens de grandes conjuntos de dados remontam às décadas de 1960 e 1970, quando o mundo dos dados estava apenas começando, com os primeiros data centers e o desenvolvimento do banco de dados relacional.

BIG DATA: Histórico

Por volta de 2005, as pessoas começaram a perceber a quantidade de usuários de dados gerados pelo Facebook, YouTube e outros serviços online. O **Hadoop** (uma estrutura de código aberto criada especificamente para armazenar e analisar grandes conjuntos de dados) foi desenvolvido no mesmo ano. O **NoSQL** também começou a ganhar popularidade durante esse período.

BIG DATA: Histórico

O desenvolvimento de estruturas de código aberto, como o **Hadoop**, (e, mais recentemente, o **Spark**) foram essenciais para o **crescimento do big data**, porque elas tornaram o trabalho com big data mais fácil e seu armazenamento mais barato. Nos anos seguintes, o volume de big data disparou.

Usuários ainda estão gerando grandes quantidades de dados, mas não são somente humanos que estão fazendo isso.

BIG DATA: Histórico

Com o advento da **Internet das Coisas** (***IoT***), mais objetos e dispositivos estão conectados à internet, reunindo dados sobre padrões de uso do cliente e desempenho do produto.

O surgimento do *machine learning* produziu ainda mais dados.

BIG DATA: Histórico

Apesar da evolução do big data, sua utilidade ainda está no começo. A **computação em nuvem** expandiu ainda mais as possibilidades do big data.

A nuvem oferece uma **escalabilidade verdadeiramente elástica**, na qual os desenvolvedores podem simplesmente criar ***clusters ad hoc*** para testar um subconjunto de dados.

BIG DATA: Histórico

Ad hoc significa “para esta finalidade", “para isso” ou "para este efeito".
É uma expressão latina, geralmente usada para informar que determinado acontecimento tem **caráter temporário** e **que se destina para aquele fim específico**.

clusters ad hoc para testar um subconjunto de dados.

BIG DATA: Desafios

Para começar, o **big data** é grande.

Apesar de novas tecnologias terem sido desenvolvidas para o armazenamento de dados, os volumes de dados estão dobrando em tamanho a cada dois anos.

As empresas ainda se esforçam para acompanhar a evolução de seus dados e encontrar maneiras de armazená-los com eficiência.

BIG DATA: Desafios

Mas armazenar os dados não é o suficiente. Eles devem ser usados para serem **úteis**, e isso depende da curadoria.

Dados limpos ou **relevantes** para o cliente e organizados de maneira que permita uma análise significativa exigem muito trabalho.

Cientistas de dados gastam de 50 a 80 por cento de seu tempo **curando** e **preparando** dados antes de serem usados.

BIG DATA: Desafios

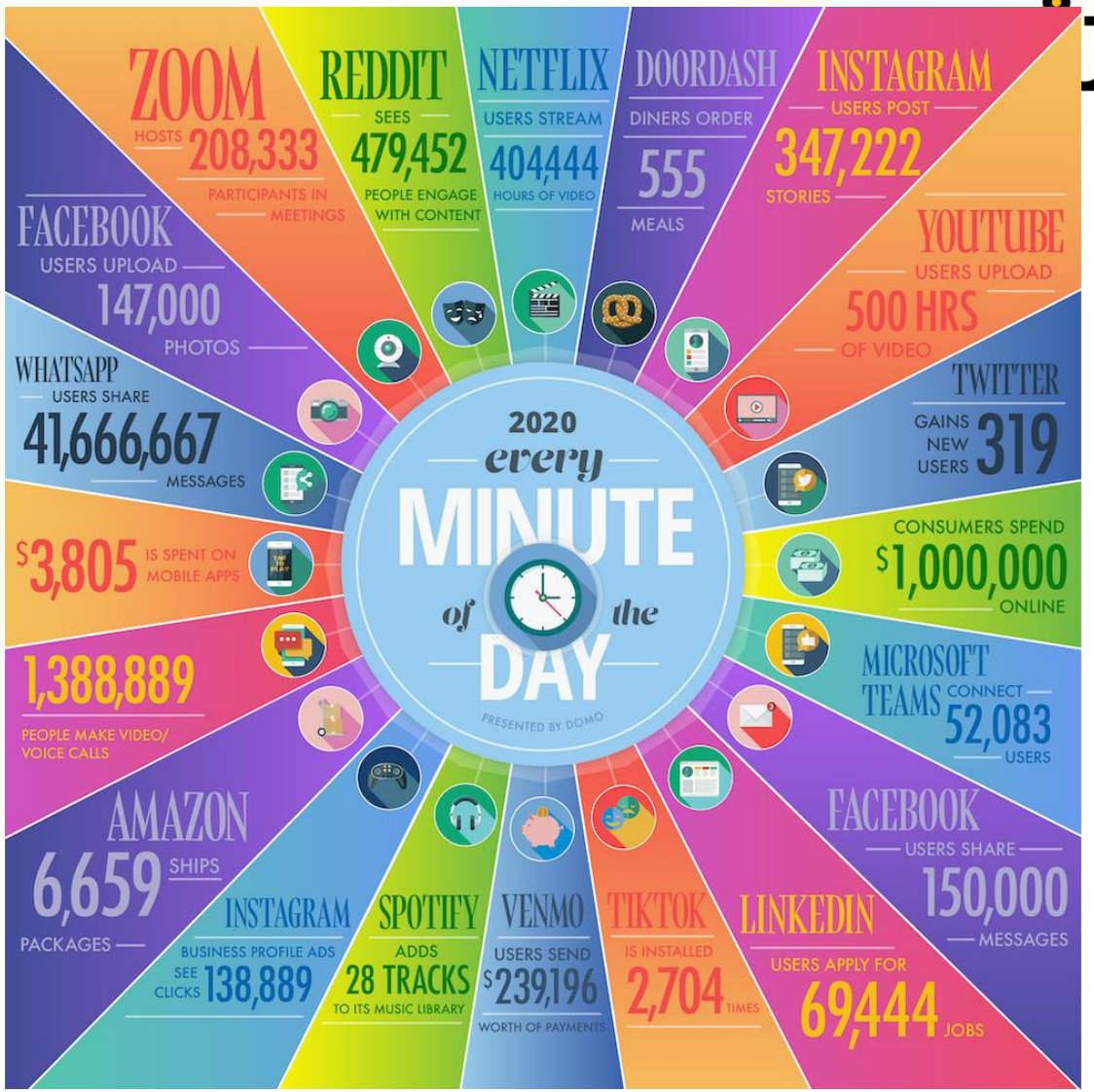
Por conta dessa problemática, surgiu o papel do ***Data Mining*** (Mineração de Dados).

Dados limpos ou **relevantes** para o cliente e organizados de maneira que permita uma análise significativa exigem muito trabalho.

Cientistas de dados gastam de 50 a 80 por cento de seu tempo **curando** e **preparando** dados antes de serem usados.

BIG DATA: Aplicações

Quantidade de dados gerados por minuto: [Data Never Sleeps 8.0 Infographic | Domo]



BIG DATA: Aplicações

Atualmente, empresas de tecnologia como o **Deezer**, a **Netflix** e a **Spotify** utilizam de *big data* para definir as preferências dos seus usuários, e fornecer para eles conteúdos mais individualizados.

Elas são uma das principais representantes atualmente no uso de Big Data no dia a dia.

BIG DATA: Aplicações

As ferramentas de propaganda do **TikTok**, do **Facebook** e do **Instagram** são baseadas em *big data*, pois correlacionam dados dos usuários das redes sociais com suas preferências de consumos e serviços.

BIG DATA: Exercício

1º Questão: Assistir ao vídeo do link abaixo e elaborar um texto resumo de no máximo 15 linhas sobre o seu entendimento do que é *Big Data,* sua importância, os 5 V's e sobre principais aplicações em *Big Data*. Enviar em formato pdf na Sala de Aula virtual da disciplina.

Vídeo: https://www.youtube.com/watch?v=32QnYF7IDyc

ARA0168
TÓPICOS DE BIG DATA EM PYTHON

1.2 – Tipos de Dados e Análises em Big Data

Universidade Estácio de Sá

Prof. Simone Gama

simone.gama@estacio.br

BIG DATA: Funcionamento

O big data fornece novas informações que abrem **novas oportunidades** e **modelos de negócios**. Os primeiros passos envolvem três ações principais:

1. Integrar
2. Gerenciar
3. Analisar

BIG DATA: Funcionamento

1. Integrar

O big data reúne dados de diversas fontes e aplicativos diferentes. Mecanismos tradicionais de integração de dados, como **extrair**, **transformar** e **carregar**, geralmente não estão aptos à tarefa. Isso requer novas estratégias e tecnologias para analisar conjuntos de big data em terabytes ou até mesmo em escala de petabytes.

Durante a integração, você precisa inserir os dados, processá-los e verificar se estão formatados e disponíveis de forma que seus analistas de negócios possam começar a utilizá-los.

BIG DATA: Funcionamento

2. Gerenciar

Big data exige armazenamento. Sua solução de armazenamento pode estar na nuvem, no local ou em ambos. Você pode armazenar seus dados da forma que desejar e trazer os requisitos de processamento desejados e os mecanismos de processo necessários para esses conjuntos de dados sob demanda.

A nuvem está gradualmente ganhando popularidade porque é compatível com as suas necessidades atuais de computação e permite que você crie recursos conforme necessário.

BIG DATA: Funcionamento

3. Analisar

Seu investimento em big data é compensado quando você analisa seus dados e age com base neles. Obtenha mais clareza com uma análise visual dos seus conjuntos de dados variados.

Explore ainda mais os dados para fazer novas descobertas. Crie modelos de dados com ***machine learning*** e **inteligência artificial**.

BIG DATA: Tipos de Dados

BIG DATA: Tipos de Dados

Existem **tipos básicos de dados** que são estudados pelos especialistas em *big data*, que em geral envolvem:

***Enterprise Data**:* Ou seja, **Dados Empresariais**. Em geral coletados pelo RH de empresas, setores de vendas, finanças e/ou logística, esses dados são atributos sobre funcionários e setores diferentes dentro de um ambiente empresarial, podem ser utilizados para otimizar processos e identificar falhas ou fraudes dentro de uma determinada seção, que visam minimizar gastos e otimizar lucros.

BIG DATA: Tipos de Dados

Existem **tipos básicos de dados** que são estudados pelos especialistas em *big data*, que em geral envolvem:

***Personal Data**:* Ou seja, **Dados pessoais**, facilmente relacionados ao conceito da Internet das coisas, são dados obtidos através de aparelhos de uso pessoal ou coletivo, tais como *smartphones*, geladeiras, televisões, etc. Esse tipo de dado mostra as preferências pessoais de um determinado indivíduo através do estudo de padrões; por meio do uso do *Personal Data* é possível desenvolver metodologias personalizadas de interação com o cliente.

BIG DATA: Tipos de Dados

Existem **tipos básicos de dados** que são estudados pelos especialistas em *big data*, que em geral envolvem:

***Social Data**:* Ou seja, **Dados sociais**. São dados coletados de redes sociais ou ambientes de interação entre usuários, geralmente **demográficos** e **comportamentais**, ou seja, ditam um padrão de um determinado grupo com as mesmas característica. O *Social Data* é muito utilizado na análise de campanhas de marketing, de maneira a oferecer um serviço ou produto mais personalizado de acordo com diferentes segmentos.

BIG DATA: Tipos de Análises

BIG DATA: Tipos de Análises

As análises que podem ser realizadas com o Big Data são:

1. Análise **preditiva**
2. Análise **prescritiva**
3. Análise **descritiva**
4. Análise **diagnóstica**

BIG DATA: Tipos de Análises

Análise **preditiva**

A análise preditiva faz uma **previsão** sobre possibilidades futuras, tendo como base os padrões encontrados nos dados analisados da empresa. É uma forma de análise avançada que verifica dados ou conteúdos para responder à pergunta: **o que é provável que aconteça no futuro**?

BIG DATA: Tipos de Análises

Análise **prescritiva**

O objetivo da **análise prescritiva** é apresentar as **possíveis consequências** que cada ação pode gerar para o negócio. Isso contribui para escolher as estratégias mais adequadas, que gerarão mais e melhores resultados para a empresa.

BIG DATA: Tipos de Análises

Análise **descritiva**

O objetivo da **análise descritiva** é trazer informações sobre questões presentes. Ou seja, esse tipo de análise colabora para **decisões que precisam ser tomadas em tempo real**. Como o nome diz, essa análise descreve os dados observados, de forma que eles “digam” o que aconteceu. Por vezes, é chamada de "**estatística descritiva**", já que é comum utilizar métodos da estatística nesse tipo.

BIG DATA: Tipos de Análises

Análise **diagnóstica**

Conhecidas as características dos dados, imediatamente vem a pergunta: **qual a causa desses dados?** A resposta a essa pergunta é justamente o objetivo da análise diagnóstica.
O objetivo da análise **diagnóstica** é analisar os resultados e desdobramentos de determinadas ações. Com isso, é possível ajustar as estratégias que estão sendo aplicadas.

ARA0168
TÓPICOS DE BIG DATA EM PYTHON

1.3 – Definições importantes em Big Data

Universidade Estácio de Sá

Prof. Simone Gama

simone.gama@estacio.br

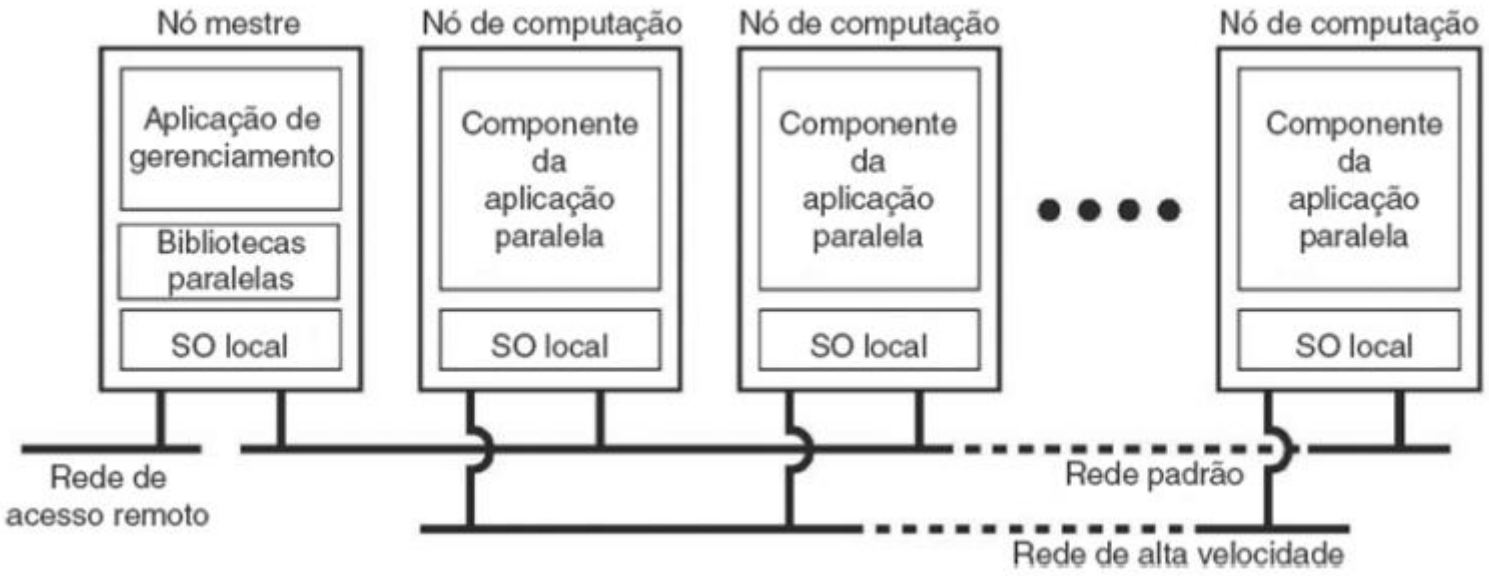
BIG DATA: Algumas definições

Cluster

- Se trata da **conexão entre dois ou mais computadores com o propósito de melhorar o desempenho dos sistemas** na execução de diferentes tarefas.
- No cluster, cada computador é denominado "**nodo**" ou "**nó**", sendo que não há limites de quantos nodos podem ser interligados.
- Com isso, os computadores passam a atuar dentro de um único sistema, trabalhando em conjunto no processamento, análise e interpretação de dados, informações e/ou realização de tarefas simultâneas.

BIG DATA: Algumas definições

Cluster

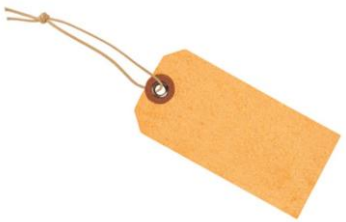


Exemplo de um Sistema Cluster (Fonte: Tanenbaum, 2007)

BIG DATA: Algumas definições

Metadados

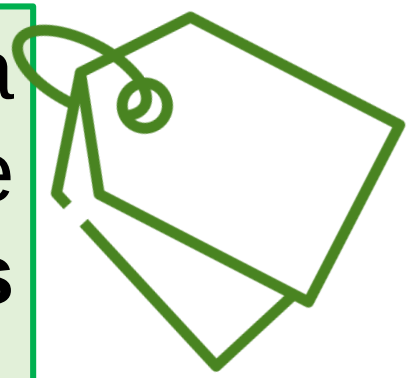
Os metadados são as estruturas responsáveis por organizar o conteúdo de forma inteligível não só pelo sistema em que se publica (*WordPress*, *Drupal*, *Joomla Blogspot*, etc), como também pelos buscadores e – ainda mais importante – pelo seu usuário. Fazem parte desse grupo os *sitemaps* e as ***tags***.



BIG DATA: Algumas definições

Metadados

Metadados existem em uma variedade de estruturas de **cabeçalhos de tabelas**, **aplicativos legados**, **arquivos de configuração**, em *IoT*, na **nuvem**, **mídia social** e **modelos de dados**.



BIG DATA: Algumas definições

Tags

As *tags* tem como objetivo **delimitar** os tópicos centrais de um conteúdo, quer ele seja um texto, uma imagem, um vídeo ou *podcast*. Este recurso apresenta ao usuário quais são os temas abordados e complementa a informação passada pelo título.

Computação Distribuída

A computação distribuída refere-se a um campo da ciência da computação que trabalha com **sistemas distribuídos**.

Um sistema distribuído inclui vários computadores que se conectam e se comunicam por meio de uma rede de computadores para atingir um objetivo comum.

Computação Distribuída

Segundo Tanenbaum (2007), é uma "**coleção de computadores independentes que se apresenta ao usuário como um sistema único e consistente**".

Computação Distribuída: Aplicações

IoT (*Internet of Things*)

1. **Objetos físicos (ou "coisas"):** Componentes eletrônicos e sensores responsáveis pela coleta de dados e aplicação de ações.
2. **Computação**: Faz o gerenciamento do ciclo de vida dos dados, desde a coleta e o armazenamento até o processamento dos dados.
3. **Protocolos de comunicação:** Viabilizam a troca dados via Internet entre os objetos físicos e outros sistemas.
4. **Serviços**: Provêm autenticação e gerenciamento de dispositivos, além de oferecer a infraestrutura.

Computação Distribuída: Aplicações

IoT (*Internet of Things*)

1. **Objetos físicos (ou "coisas")**: Componentes eletrônicos e senso de ações.
2. **Comp** dos dados, desde mento dos dados
3. **Proto** dados via Interne
4. **Serviços**: Provêm autenticação e gerenciamento de dispositivos, além de oferecer a infraestrutura.

Para tratar de aplicações de *IoT*, utiliza-se **algoritmos distribuídos** que reconhecem os dispositivos e os utilizam de forma eficiente para transmitir e receber dados.

Computação Distribuída x *Cloud Computing*

	CLOUD COMPUTING	COMPUTAÇÃO DISTRIBUÍDA
01	A computação em nuvem se refere ao fornecimento de recursos / serviços de TI sob demanda, como servidor, armazenamento, banco de dados, rede, análise, software, etc. pela Internet.	A computação distribuída se refere a resolver um problema em computadores autônomos distribuídos e eles se comunicam entre eles por meio de uma rede.
02	Na computação em nuvem simples pode-se dizer que é uma técnica de computação que entrega serviços hospedados pela internet aos seus usuários / clientes.	Na computação distribuída simples pode-se dizer que é uma técnica de computação que permite a vários computadores se comunicarem e trabalharem para resolver um único problema.
03	É classificado em 4 tipos diferentes, como nuvem pública, nuvem privada, nuvem comunitária e nuvem híbrida.	É classificado em 3 tipos diferentes, como **Sistemas de Computação Distribuídos**, **Sistemas de Informação Distribuídos** e **Sistemas Pervasivos Distribuídos**.

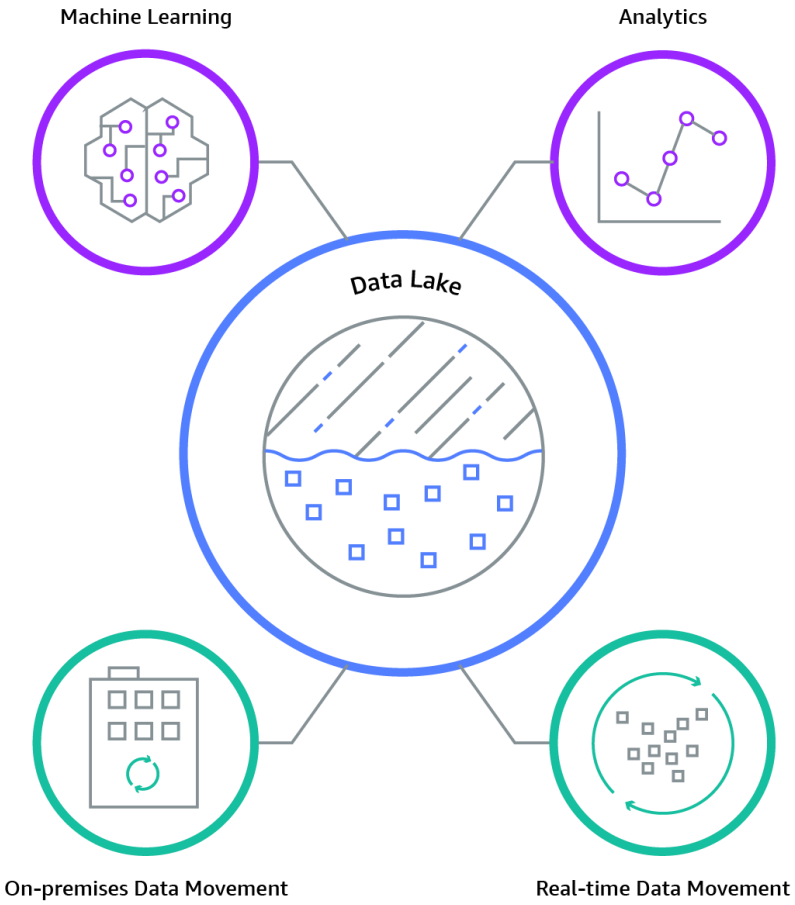
Computação Distribuída x *Cloud Computing*

	CLOUD COMPUTING	COMPUTAÇÃO DISTRIBUÍDA
04	Existem muitos benefícios da computação em nuvem, como custo-benefício, elasticidade e confiabilidade, economia de escala, acesso ao mercado global, etc.	Existem muitos benefícios da computação distribuída, como flexibilidade, confiabilidade, desempenho aprimorado, etc.
05	A computação em nuvem fornece serviços como hardware, software, recursos de rede por meio da Internet.	A computação distribuída ajuda a realizar tarefas computacionais com mais rapidez do que usar um único computador, pois leva muito tempo.

Data Lake

O **Data lake** é um repositório utilizado para armazenar todos os dados estruturados e não estruturados.

Ao armazená-los de forma não estruturada pode-se realizar diferentes tipos de análise, incluindo **processamento de big data**, **análise em tempo real** e ***machine learning***, a fim de adquirir melhores tomadas de decisões.



Data Lake

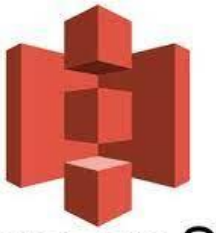
Esses dados são armazenados em **objetos** – conhecidos como ***object storage*** – que contêm *tags* de metadados e um identificador único.

Essa estrutura de entidade dos dados permite que possamos analisá-los e buscar por padrões, pois as consultas são realizadas com bastante eficiência.

Tais objetos de armazenamento podem ser consultados pelas demais aplicações de Big Data.

Data Lake

O *data lake* é recurso essencial nas plataformas de *Big Data*, pois as organizações utilizam os dados como a base para realizar **análises** e desenvolver **estratégias** que as auxiliem a potencializar seus negócios. Cada plataforma oferece uma tecnologia de *data lake*.



Amazon Simple Storage Service (S3)

Azure Data Lake

Google Cloud Storage

Oracle Cloud

Data Lake

O *data lake* é pode ser também caracterizado por **sistemas distribuídos individuais**. Alguns são amplamente conhecidos e utilizados pela comunidade científica *Big Data*.

Apache Hadoop

Apache Spark

Personal DataLake
Cardiff University

Data Lake

Esses dados são armazenados em **objetos** – conhecidos como ***object storage*** – que contêm *tags* de metadados e um identificador único.

Essa estrutura de entidade dos dados permite que possamos analisá-los e buscar por padrões, pois as consultas são realizadas com bastante eficiência.

Tais objetos de armazenamento podem ser consultados pelas demais aplicações de Big Data.

Data Lake vs Data Warehouse

Fonte: Data Lake vs Data Warehouse: Key Differences | Talend

Data lake	***Data warehouse***
Armazenamento de dados **desestruturados**, Dados **semi-estruturados** e estruturados.	Dados estruturados
Esquema definido na leitura	Esquema definido na escrita
Ciência de dados, análise preditivas, BI	BI baseado em SQL
Armazenamento de dados detalhados, brutos e também processados	Armazenamento de dados frequentemente acessados, assim como dados agregados e sumarizados
Separação entre o armazenamento e o processamento	Acoplamento entre o armazenamento e o processamento

Dúvidas?

Bibliografia

- FACELI, Katti. **Inteligência Artificial Uma abordagem de aprendizado de máquina.** 2ª Ed. Rio de Janeiro: LTC, 2021. Disponível em:

Leitura Auxiliar

- Big Data: O que é Big Data? | Oracle Brasil
- Cloud x Sistemas Distribuídos: Diferença entre computação em nuvem e computação distribuída – Acervo Lima
- Metadados e tags: Metadados: a importância do uso de tags (luisabwk.com.br)
- TANENBAUM, Andrew S.; VAN STEEN, Maarten. **Distributed Systems: Principles and Paradigms**.2.ed. Vrije Universiteit. Amsterdam, Holanda, 2007.

Observações

- Esse material / slide tem o propósito de auxiliar na didática da disciplina, não englobando, sob hipótese alguma, a ementa completa.
- O estudo do conteúdo deve ser continuado na bibliografia da disciplina e no conteúdo da Sala de Aula Virtual.