Tecnólogo em Análise e Desenvolvimento de Sistemas

Ciência de Dados

UNIVERSIDADE CANDIDO MENDES

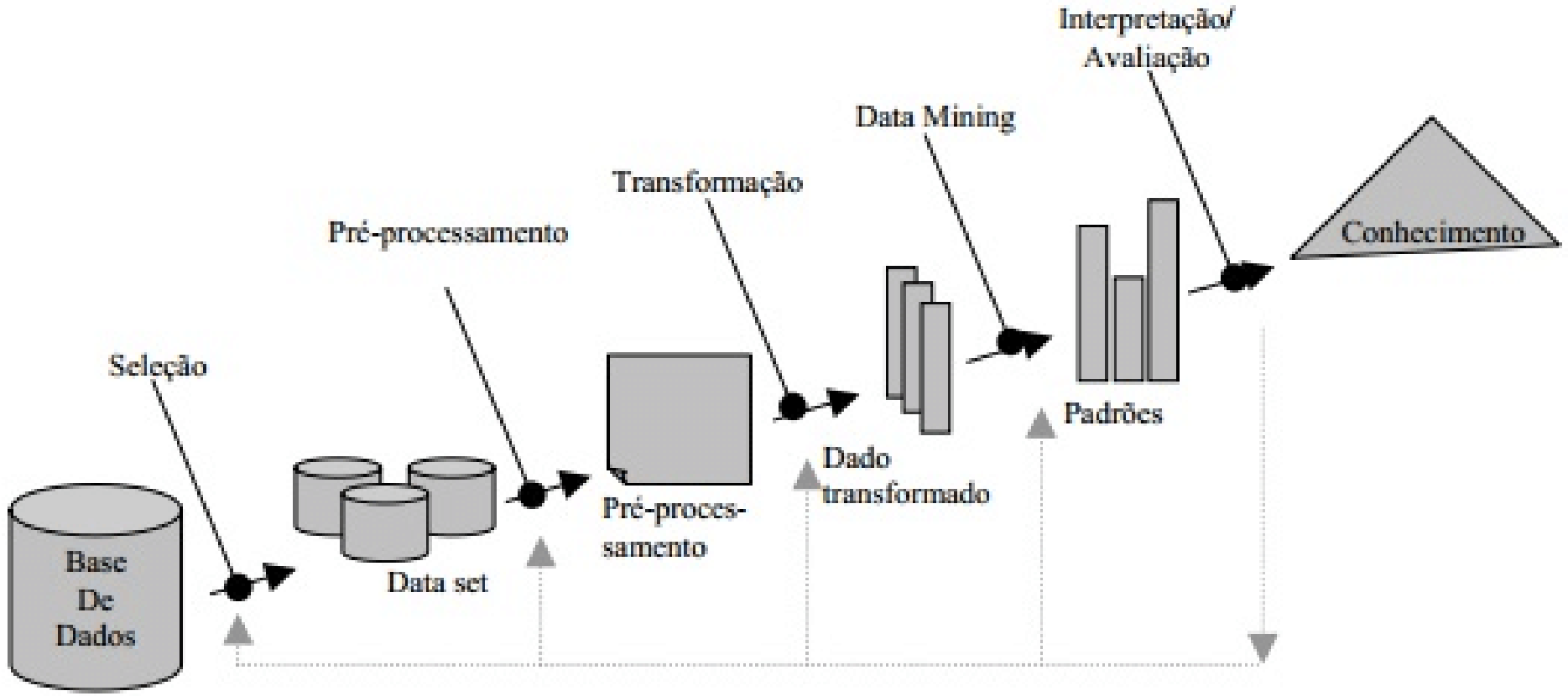
Prof. Ricardo Tavares

ricardo.tavares@ucam-campos.br

Roadmap

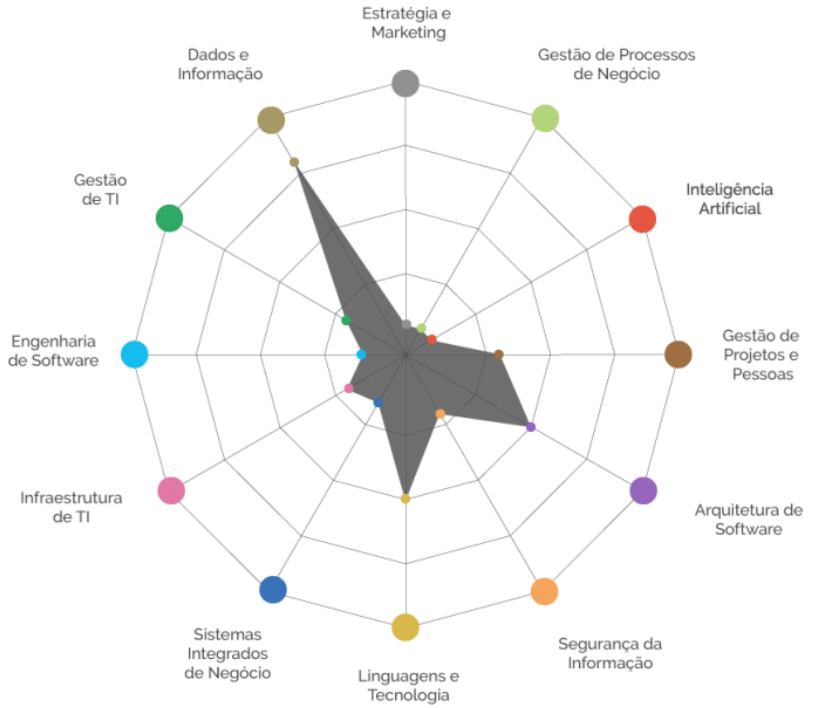
- Introdução a Big Data e Ciência de Dados:
  - Os 6 Vs do Big Data:
    - Volume, Velocidade, Variedade, Veracidade, Valor e Variabilidade;
    - Tipos de dados: estruturados, semiestruturados e não estruturados;
- Ciclo de vida dos dados:
  - DMBoK:
    - Data Managament Body of Knowledge:
    - Governança de dados;
  - Extract, Transform and Load (ETL):
    - Extração de dados, transformação e limpeza -> preparação necessária para o KDD (knowledge-discovery in databases)
    - O que são Outliers?

Roadmap


Interpretação/
Avaliação
Data Mining
Transformação
Pré-processamento
Conhecimento
Seleção
Padrões
Dado
transformado
Pré-proces-
samento
Base
De
Dados
Data set

Roadmap

- Diversos papéis em projetos de Big Data e Ciência de Dados:
  - DBA;
  - Engenheiro de Dados;
  - Analista de Dados;
  - Estatístico;
  - **Cientista de Dados;**
  - DevOps;



Roadmap

- Técnicas utilizadas em Big Data e Ciências de Dados:
  - O que é aprendizado de máquina;
  - Tipos de aprendizado de máquina;
- Ferramentas utilizadas em Big Data e Ciência de Dados:
  - Python – Utilização de Bibliotecas:
    - Pandas e Numpy;
    - Matplotlit, Seaborn e Plotly;
    - Scikit-learn e Statsmodel;
  - Anaconda;
  - Google Colaboratory;
- Estatística descritiva aplicada a projetos de Big Data:
  - Entendendo seus dados;
  - Estatística descritiva utilizando Python;

IMPORTANTE: NECESSÁRIO CONHECIMENTO BÁSICO DE PROGRAMAÇÃO EM PYTHON -> ESTUDEM

Roadmap

- Introdução ao Aprendizado de Máquina:
  - Tipos de aprendizado de máquina por tipo de problema a ser solucionado;
  - Determinando o melhor algoritmo a ser aplicado para resolução de um determinado problema;
- Aprendizados não supervisionados:
  - Problemas de agrupamentos:
    - Entendendo o k-means – teoria;
    - Aplicação do k-means – prática;
    - Utilizando Python para resolução de problemas de agrupamentos utilizando o k-means;

Roadmap

- Aprendizados supervisionados:
  - Problemas de classificação e regressão
    - Tipos de algoritmos CART (Classification and Regression Trees);
    - O que é entropia?
    - Entendendo algoritmos CART – teoria;
    - Aplicação de algoritmos CART – prática;
    - Utilizando Python para resolução de problemas de classificação e regressão utilizando algoritmos CART;

Tecnólogo em Análise e Desenvolvimento de Sistemas

Introdução a Projetos de Big Data

UNIVERSIDADE CANDIDO MENDES

Prof. Ricardo Tavares

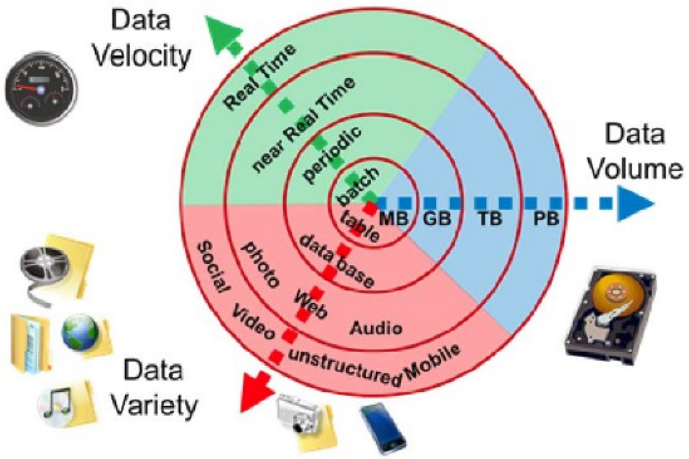
ricardo.tavares@ucam-campos.br

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - É uma área do conhecimento cujo objetivo é construir formas de **tratar, analisar e obter informações (insights)** oriundos de uma **grande quantidade de dados**;
    - Grande quantidade de dados não é só em volume!
    - Big Data é o dado que excede a capacidade de **processamento convencional** dos sistemas de bancos de dados (DUMBILL, 2012);
    - Os 5 (ou 6, 7...) V's do Big Data:
      - Volume;
      - Velocidade;
      - Variedade;
      - Veracidade;
      - Valor

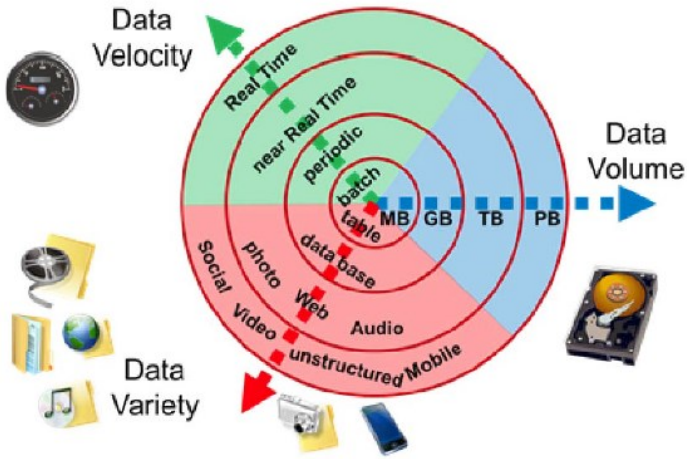
Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - Volume: Refere-se ao tamanho dos dados produzidos e à necessidade de serem armazenados;
      - Todos os dias são criados 2,5 quintilhões de bytes em forma de dados (INMOMENT, 2014);
      - Atualmente 90% de todos os dados que estão presentes no mundo foram criados nos últimos 2 anos (IBM);
      - IoT (Internet of Things): Internet das Coisas produzem dados em grande quantidade e em tempo real;
      - GPS, sensoriamento remoto, smartphones -> Dados importantes que precisam ser coletados, armazenados e interpretados;
      - Previsão é de que em 2025 serão mais de 41 bilhões de dispositivos conectados gerando aproximadamente 79 zettabytes de dados (~ 93,2 Bilhões de Terabytes);
      - Volume como fator primordial:
      - Dados de vendas de produtos em uma empresa de varejo (ERP); Banco de imagens 3D de obras de artes da Itália;



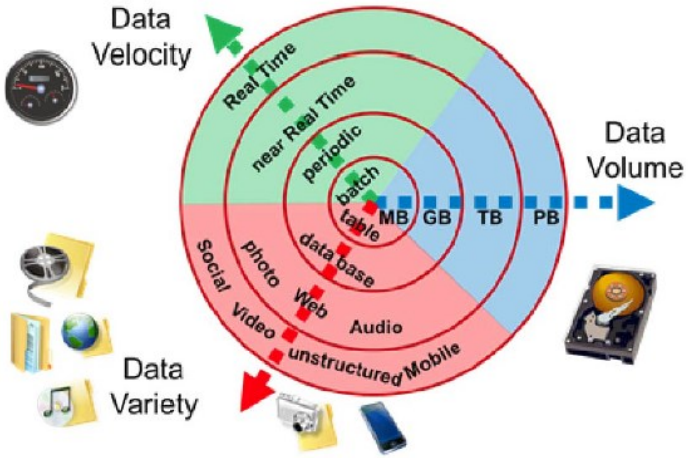
Introdução ao Big Data

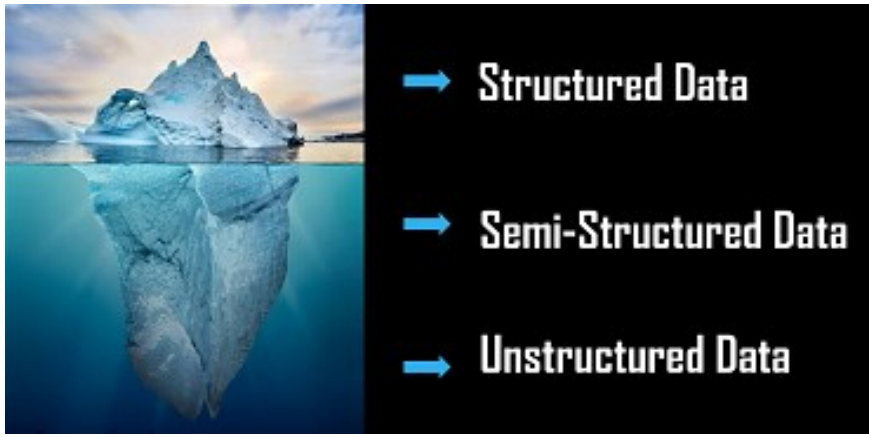
- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - Velocidade: A velocidade de produção de dados é muito alta e, muitas das vezes, sua interpretação deve ser tão veloz quanto sua geração;
      - Imagine o seguinte cenário: Se você tivesse que atravessar uma rua movimentada com os olhos vendados e a única informação que tem é uma foto de um sinal tirada a 5 minutos atrás, você se arriscaria?
      - Novamente a IoT traz um paradigma totalmente inovador:
      - A velocidade de obtenção e interpretação de dados é uma vantagem competitiva das empresas;
      - Velocidade como fator primordial:
      - Aquisição e interpretação de dados da disseminação da COVID-19; Controle de sinais de trânsito; Controle de Vendas e-commerce;



Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - Variedade: Se temos um volume enorme de dados, também obtemos a variedade dos mesmos!
      - Os dados estruturados, como planilhas, tabelas em BD, por exemplo, são apenas a ponta do Iceberg;
      - Estima-se que cerca de 80% à 90% de todos os dados no mundo estão a forma de dados não estruturados. (ICD, 2011);
      - Empresas que conseguem captar a variedade, seja de fontes ou de critérios, agregam mais valor ao negócio (GARTNER)
      - Variedade como fator primordial: As mídias sociais geram, diariamente, um grande volume de dados não estruturados, como: e-mails, fotos, vídeos, áudios; Dados que precisam ser interpretados para entender o perfil do cliente;



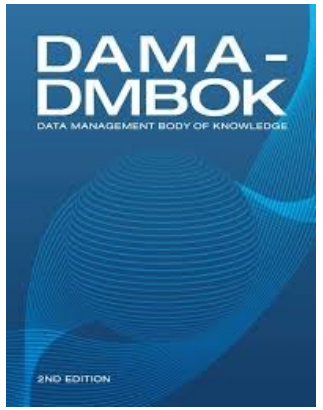


Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - Tipos de dados:
      - **Os dados estruturados** são aqueles organizados e representados com uma estrutura rígida, a qual foi previamente planejada para armazená-los, por exemplo um banco de dados, que é a representação mais típica e comum de dados estruturados.
        - Tabelas em Banco de Dados (MySQL, SQL Server, etc.) e planilhas eletrônicas;
      - **Os dados não-estruturados** são aqueles que não têm estrutura definida;
        - Posts nas Redes Sociais; Textos no Word ou Bloco de Notas; Vídeos; Áudios; Página da internet;
      - **Os dados semiestruturados** são aqueles que estão no meio termo. Possui certa estrutura, como o JSON ou XML, mas são flexíveis, pois é possível criar quantos campos de diferentes formatos quiser;
      - E os Bancos de Dados NoSQL (Mongo, Cassandra, etc.) representam dados estruturados ou não?!

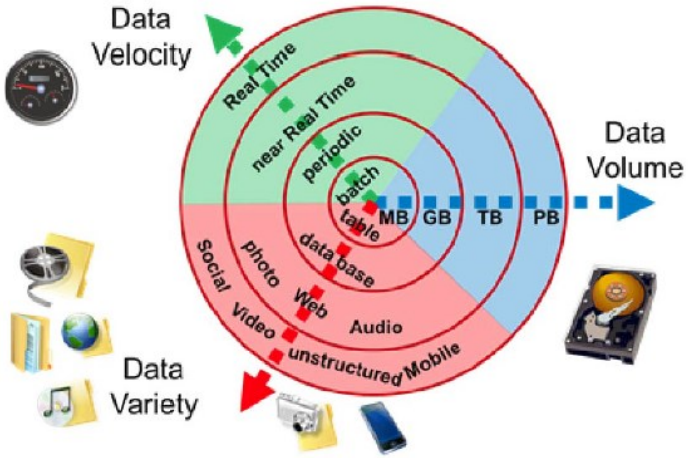
Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - Veracidade: Trata-se de planejar e projetar o saneamento do dado, provendo qualidade ao mesmo, para que este dado possa gerar informações confiáveis para suportar a tomada de decisão
    - Ótica dupla:
      - Qualidade dos dados: dados completos, sem erros ou falta de dados parciais;
      - Qualidade dos dados: Dados que representam a realidade, sem viés
    - Para colher bons frutos do processo do Big Data é necessário obter dados verídicos, de acordo com a realidade;
    - A verificação dos dados coletados para adequação e relevância ao propósito da análise é um ponto chave para se obter dados que agreguem valor ao processo. (Hurwitz et. al.);
    - Boa parte do tempo do trabalho de Data Engineer e Data Analytics é alocado para garantir a extração e limpeza dos dados → dados limpos e não enviesados.



Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - Valor:
    - O armazenamento, limpeza, transformação e análise deve gerar valor agregado que compense os custos financeiros envolvidos (TAURION, 2013);
    - “guarde, porque amanhã isso valerá muito”. Será?!! Dados passados podem ser considerados dados verídicos para o momento em que é analisado?
    - Dados são importantes para realização de Análise Preditiva. → O que quero prever? Quais são os dados importantes?



Introdução ao Big Data

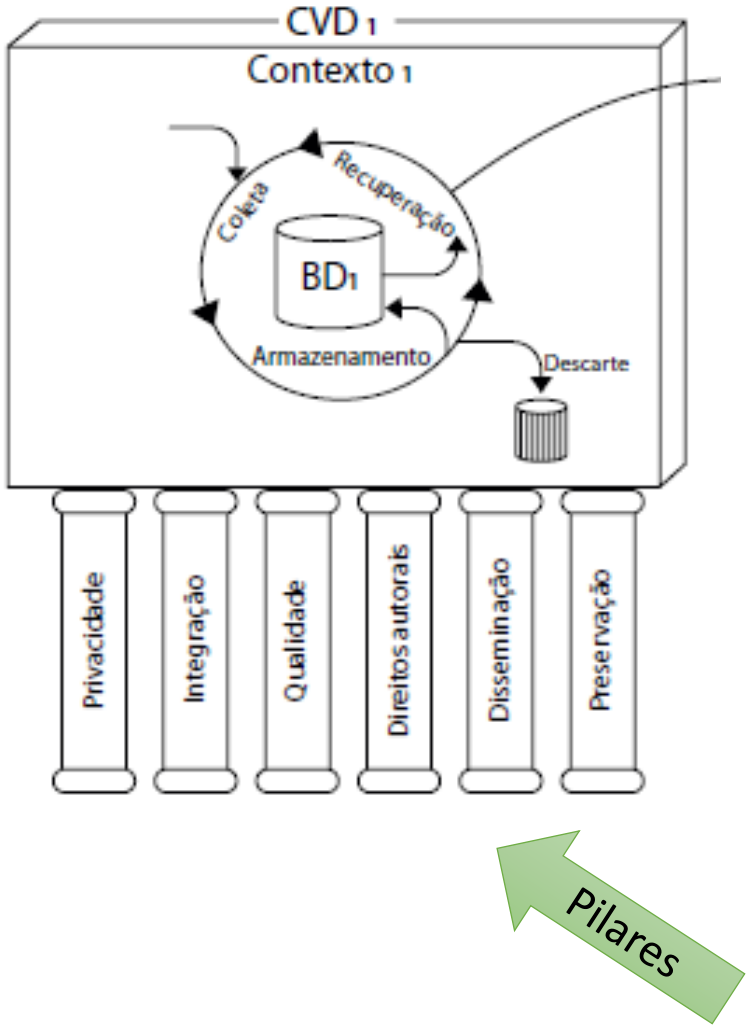
- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - Valor:
      - Quando o valor é realmente importante?!
      - Caso Nestlé: Em 2010 um vídeo associou o Kit Kat com a extração descontrolada de óleo de dendê.
      - https://www.youtube.com/watch?v=2ExNmhDLsIk
      - A empresa então montou um plano de monitoramento de Redes Sociais para entender o sentimento dos seus clientes com relação a empresa e seus produtos.

    - **O Valor: O Big Data alinhado às redes sociais elevaram a Nestlé da 16ª para a 12ª posição entre as empresas com melhor reputação do mundo no mesmo ano de 2011. Atualmente a Nestlé está entre as 10 empresas com a melhor reputação do mundo.**

- Volume: Grande volume de dados coletados de redes socias;
- Velocidade: Alta velocidade de mutação de dados, como podemos observar nas redes e mídias sociais o que é postado hoje pode ou não ser relevante para o amanhã;
- Variedade: Dados variados e não estruturados;
- Veracidade: A veracidade é obtida pela verificação em tempo real;

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Ciclo de Vida dos Dados:
    - No ciclo da sua vida, o dado pode ser extraído, exportado, importado, migrado, validado, editado, atualizado, limpo, transformado, convertido, integrado, segregado, agregado, referenciado, revisado, relatado, analisado, garimpado, salvo, recuperado, arquivado e restaurado antes de eventualmente ser eliminado (DMBOK, 2012).
    - Segundo Sant'Ana (2016), o ciclo de vida dos dados é composto por quatro fases: coleta, armazenamento, recuperação e descarte.



Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Ciclo de Vida dos Dados:
    - Pilares de apoio ao projeto de Big Data:
      - **Qualidade dos dados:** A confiança (Veracidade) no contexto dos dados está intimamente ligada à sua qualidade;
      - **Privacidade:** Princípio da confidencialidade;
        - Lei nº 13.709/18 - Lei Geral de Proteção de Dados (BRASIL, 2018): regula as atividades de tratamento de dados pessoais;
        - Todos tem o direito da privacidade e cabe a empresa que coleta os dados pessoais, seja física ou jurídica de direito público ou privado a proteção e a informação de como utilizará os dados.

Introdução ao Big Data

QUAIS DADOS SÃO COLETADOS PELO GRUPO MAGALU

Durante sua experiência em uma de nossas lojas, uso do nosso Super App ou de outros aplicativos das marcas do Grupo Magalu, podemos coletar diferentes tipos de dados pessoais, de forma automática com o objetivo de conferência, monitoramento e controle, ou fornecidas diretamente por você, como por exemplo para a realização de seu cadastro. Veja abaixo quais dados pessoais nós podemos coletar e em cada situação:

Durante o cadastro:
Nome completo;
Foto;
Número de CPF;
Endereço de e-mail;
Número de celular;
Data de nascimento;
Dados referentes aos seus endereços.

Durante o preenchimento do local de entrega e forma de pagamento:
Endereço de cobrança;
Endereço de entrega;
Dados do cartão de crédito, quando escolhido como forma de pagamento.

Durante a análise e o monitoramento de suas compras ou outras transações financeiras:
Dados cadastrais;
Tipo de produto;
Quantidade;
Valor da mercadoria (unitário);
Valor total da compra ou transação;
Natureza da transação financeira;
Informações da conta bancária e outros meios utilizados.
Filiação.
Informações de renda.

Dados que serão armazenados

COMO NÓS UTILIZAMOS OS SEUS DADOS PESSOAIS
Nós utilizamos os dados pessoais para garantir um atendimento de qualidade e uma melhor experiência na sua compra. Listamos abaixo as finalidades que poderemos utilizar seus dados pessoais:

Dados cadastrais:
Para viabilizar a prestação de diferentes serviços disponíveis em nossas lojas, nos aplicativos das marcas do Grupo Magalu ou no Super App Magalu (e-commerce, marketplace, cash-in, cashback, chargeback, pagamento de contas, recargas, transferência entre contas de pagamento, pagamentos em lojas físicas do Magazine Luiza e transações off-us).
Para realizar o atendimento de solicitações e dúvidas em nossa Central de Atendimento.
Para identificar corretamente o Usuário.
Para enviar os produtos adquiridos ou comunicações de ofertas.
Para entrar em contato com você, quando necessário. Esse contato pode contemplar diversos assuntos, como comunicação sobre promoções e ofertas, respostas a dúvidas, respostas de reclamações e solicitações, atualizações dos pedidos realizados e informações de entrega.
Para auxiliar no diagnóstico e solução de problemas técnicos.
Para desenvolver novas funcionalidades e melhorias, melhorando a sua experiência com nossos serviços disponíveis.
Para consultar suas informações nas bases de dados de agências de crédito.
Para realizar investigações e medidas de prevenção e combate a ilícitos, fraudes, crimes financeiros e crimes de lavagem de dinheiro e/ou de financiamento ao terrorismo.
Para garantir o cumprimento de obrigação legal ou regulatória ou garantir o exercício regular de direitos do Magalu. Nesses casos, podemos, inclusive, utilizar e apresentar as informações em processos judiciais e administrativos, se necessário.
Para colaborar com o cumprimento de ordem judicial, de autoridade competente ou de órgão fiscalizador.
Para viabilizar o cadastro no Magalu Pay e realizar a abertura de sua conta de pagamento.

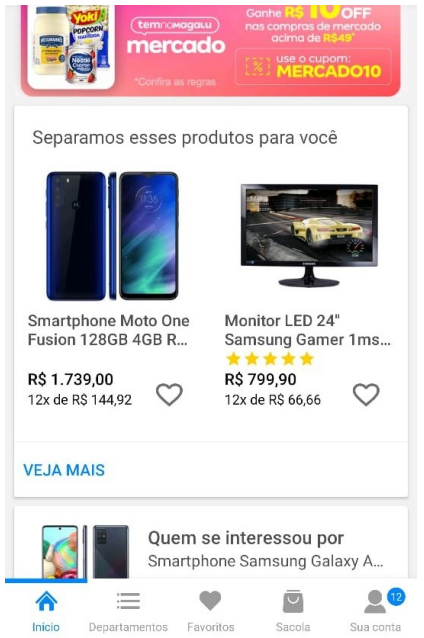
Geolocalização:
Para identificar as lojas físicas mais próximas de você.
Exibir anúncios personalizados.
Para envio de mensagens contextualizadas via push (1).
Para auxiliar nas análises que possam ser utilizadas para proteger sua conta e aumentar o nível de segurança dos seus dados cadastrais ou, ainda, prevenir possíveis fraudes.

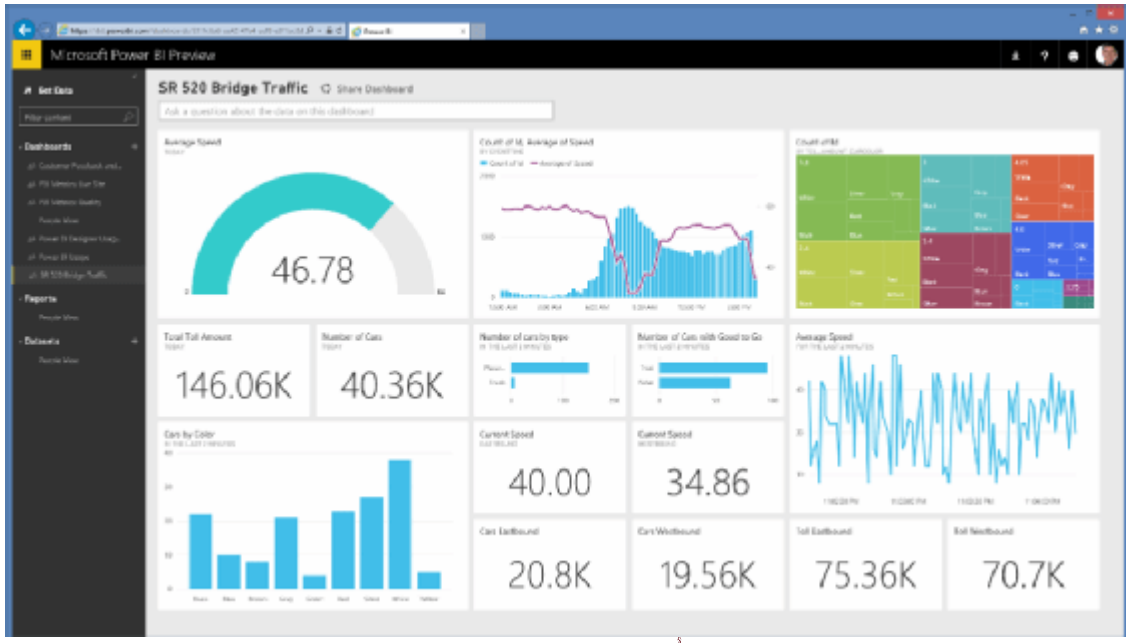
Como os dados serão utilizados

Temos ainda: Quem poderá utilizar os dados

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Ciclo de Vida dos Dados:
    - Pilares de apoio ao projeto de Big Data:
      - **Disseminação:** Deve-se haver um plano bem definido de como a informação será disseminada para os usuários;
        - Sistemas Web; APP's; BI; Planilhas analíticas;
        - Não adiantar gerar uma informação que não seja utilizada para dar valor a empresa ou usuários





Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Ciclo de Vida dos Dados:
    - Pilares de apoio ao projeto de Big Data:
      - **Direitos autorais:** Dados protegidos por direitos autorais devem ser consumidos somente através de canais permitidos ou autorizados pelo detentor dos direitos, ou outorgados formalmente;
      - Caso Napster → Compartilhamento de MP3 de forma ilegal terminou com vários processos pedindo indenização e levou ao encerramento da empresa;

Hoje o Napster opera de forma legal, através de stremming de músicas, mas nunca voltou a ter a força que tinha nos anos 2000;

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Ciclo de Vida dos Dados:
    - Pilares de apoio ao projeto de Big Data:
      - **Preservação:** Identificação dos dados necessários para construção de análise e resolução de problemas, armazenamentos e preservação para elaboração do projeto de Big Data;
        - **Os metadados são informações sobre os dados.**
        - Fotos: data, hora, localização, etc;

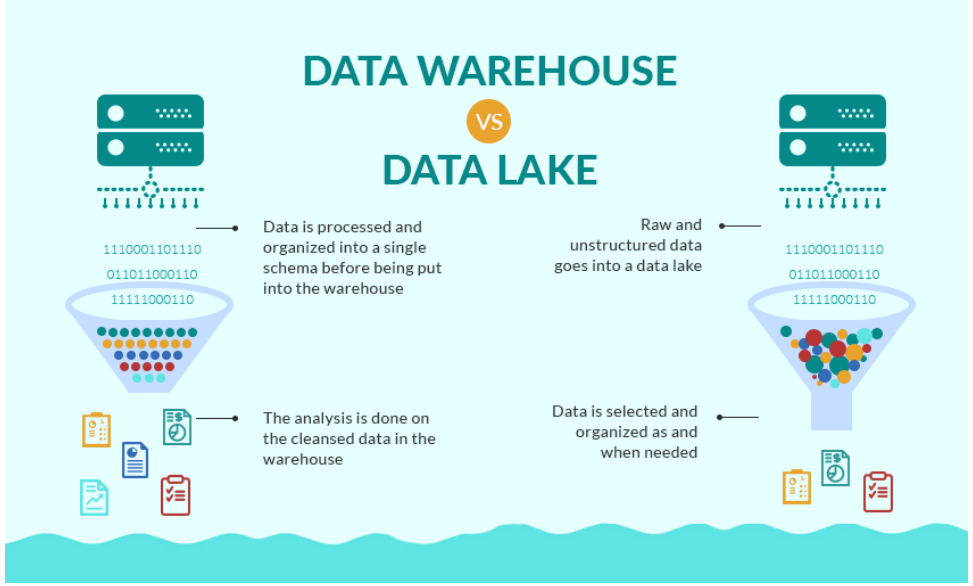
Tags especiais que o Google processa

Envie comentários

O Google é compatível com metatags no nível da página e diretivas in-line para controlar como as páginas do site aparecem na Pesquisa.

As metatags no nível da página são uma ótima maneira para os proprietários do site enviarem informações sobre os sites aos mecanismos de pesquisa. As metatags são usadas para transmitir dados a vários clientes, e cada sistema processa somente aquelas que são compatíveis, ignorando as demais. As metatags são adicionadas à seção `<head>` da página HTML e normalmente têm a seguinte aparência:

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">
    <meta name="Description" CONTENT="Author: A.N. Author, Illustrator: P. Picture, Category: Books
    <meta name="google-site-verification" content="+nxGUDJ4QpAZ5l9Bsjdi102tLVC21AIh5d1Nl23908vVuFHs
    <title>Example Books - high-quality used books for children</title>
    <meta name="robots" content="noindex,nofollow">
```



Introdução ao Big Data

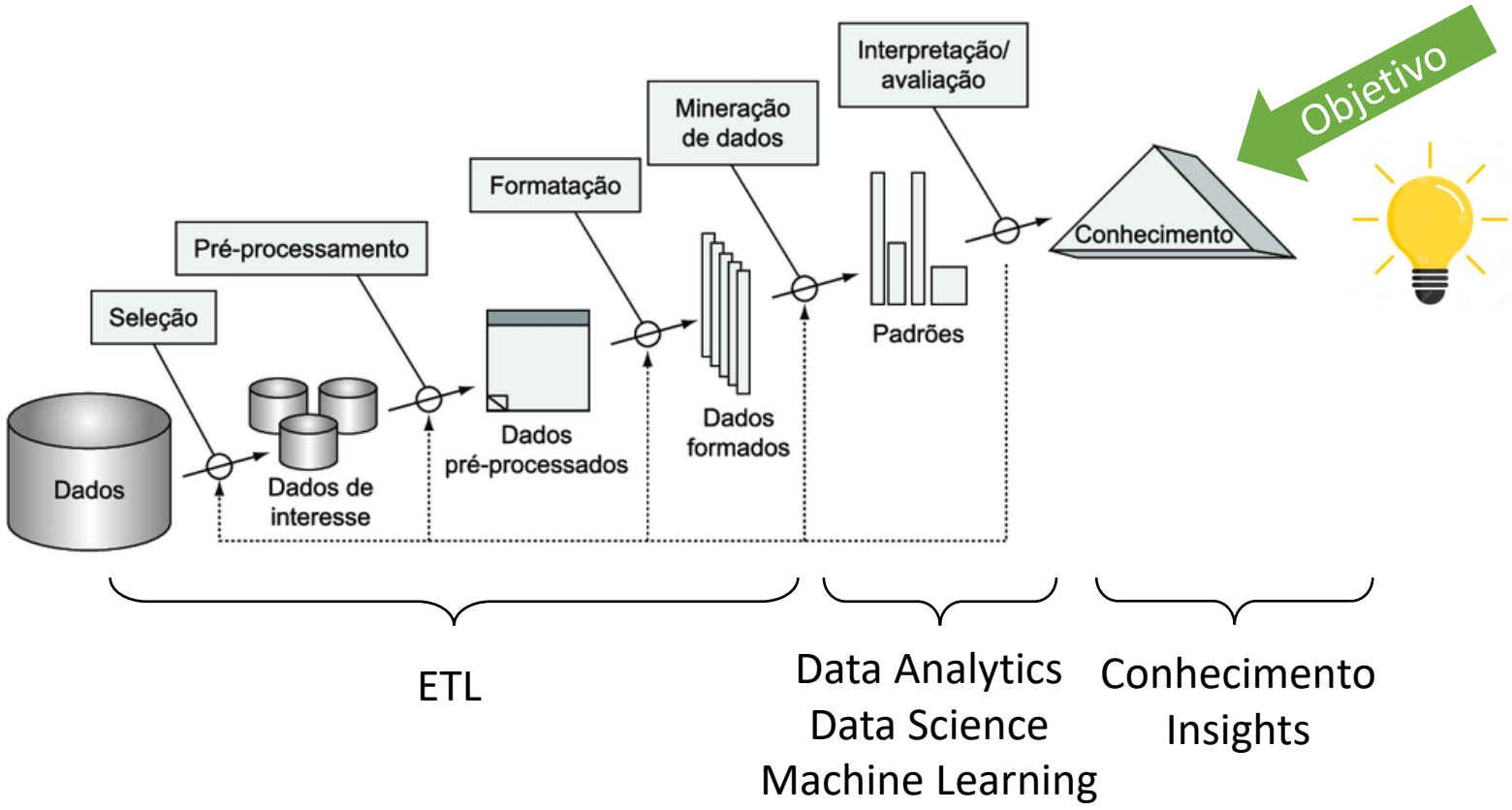
- Introdução a Big Data e Ciência de Dados:
  - Ciclo de Vida dos Dados:
    - Pilares de apoio ao projeto de Big Data:
      - **Integração:** Utilização de várias fontes de dados diferentes
      - **Enriquecimento de Dados;**
      - Exemplo:
        - Avaliação de Imóveis:
          - Dados de ofertas e vendas de imóveis;
          - Dados socioeconômicos:
            - Renda per capita;
            - IDH – Índice de Desenvolvimento Humano;
            - Taxa de Desemprego;
            - Oferta de serviços públicos;

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Ciclo de Vida dos Dados:
    - **Descarte dos dados:**
      - Deve-se implementar políticas de descarte de dados;
        - Garantia de sigilo pessoal e empresarial;
        - Garantia do copyright;
        - LGPD garante ao usuário que solicitar a exclusão de dados que a empresa coletora deverá empregar todas as técnicas e excluir todos os dados que não tenham a necessidade de cumprimento de obrigações legais, e estes devem estar descaracterizado (utilização de código ou alguma chave para identificar o usuário);

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Não confunda! Ciclo de Vida de um projeto de Big Data:



Etapas do processo KDD (FAYYAD et al., 1996)

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Ciclo de Vida em um projeto de Big Data:
    - **Seleção**: esta etapa consiste em selecionar um conjunto ou subconjunto de dados que farão parte da análise. As fontes de dados podem ser variadas (planilhas, sistemas gerenciais, data warehouses) e possuir dados com formatos diferentes (estruturados, semiestruturados e não-estruturados).
    - **Processamento**: esta etapa consiste em fazer a verificação da qualidade dos dados armazenados. A base passa por um processo de limpar, corrigir ou remover dados inconsistentes, verificar dados ausentes ou incompletos, identificar anomalias (outliers).
    - **Transformação** : esta etapa consiste em aplicar técnicas de transformação como: normalização, agregação, criação de novos atributos, redução e sintetização dos dados. Aqui os dados ficam disponíveis agrupados em um mesmo local para a aplicação dos modelos de análise.
    - **Mineração de Dados**: esta etapa consiste em construir modelos ou aplicar técnicas de mineração de dados. Essas técnicas têm por objetivo (1) verificar uma hipótese, (2) descobrir novos padrões de forma autônoma. Além disso, a descoberta pode ser dividida em: preditiva e descritiva. Esses modelos geralmente são aplicados e refeitos inúmeras vezes dependendo do objetivo do projeto.
    - **Interpretação e Avaliação**: esta etapa consiste em avaliar o desempenho do modelo, aplicando em cima de dados que não foram utilizados na fase de treinamento ou mineração. A validação pode ser feita de diversas formas, algumas delas são: utilizar medidas estatísticas, passar pela avaliação dos profissionais de negócio.

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Papéis técnicos nos processos de ciência de dados:
    - Administradores de Bancos de Dados (DBA's):
      - Responsáveis por criar e manter bancos de dados;
        - manutenção do servidor físico do banco de dados;
        - recuperação de desastres;
        - melhoria no desempenho de consultas ao banco de dados feito por aplicações da empresa;
        - controle de acesso aos dados;
        - criação de objetos (tabelas, funções, procedimentos, visualizações, etc.) no banco de dados;
        - uso dos dados da empresa para alimentar os processos de negócios.

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Papéis técnicos nos processos de ciência de dados:
    - Engenheiro de dados:
      - Responsáveis por criar a infraestrutura necessária para disponibilizar os dados necessários para consumo;
      - Esse papel, muitas das vezes, são realizados pelos próprios DBA's, já que estes possuem o conhecimento da infra e da estrutura dos dados armazenados;
      - Primeiro filtro de dados;
      - Responsável pela criação de Data Warehouse e Data Lakes;
      - Responsável pela migração de dados on-premisses para cloud;

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Papéis técnicos nos processos de ciência de dados:
    - Cientista de dados:
      - Extrair conhecimento através dos dados brutos;
        - Conhecimento em programação e softwares utilizados para manipular dados;
        - Conhecimento matemático e estatístico;
          - Pode ocorrer 02 ou mais profissionais trabalharem em conjunto: Cientista de dados + estatístico + matemático;
        - Realiza a limpeza, tratamento, transformação, carga e análise de dados;
          - Pode trabalhar em conjunto com um Analista de Dados;
        - Conhecimento do negócio;
          - Pode trabalhar em conjunto com um Analista de Negócios;
        - Responsável pela construção ou utilização de algoritmos de Aprendizado de Máquina;
          - Geralmente, a construção de algoritmos de Machine Learning está migrando para um Engenheiro de Machine Learning;
            - Conhecimento avançado em cálculo, álgebra, programação matemática;

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Papéis técnicos nos processos de ciência de dados:

DBA's e Data Engineers	Data Analyst	Data Scientist
• Modelagem, construção e manutenção de Banco de Dados; • Bancos de dados relacionais e NoSQL; • Construção e manutenção de DW e DL; • Linguagem: SQL • Cria e mantém infraestruturas com Hadoop, MapReduce, Spark, Hive, Pig;	• ETL; • Conhecimento de BD; • Criação de scripts para ETL e análise estatística; • Criação e manutenção de relatórios analíticos e visualização de dados; • Linguagem: SQL, Python, Java, R; • Ferramentas: Excel, Power BI, Tableu, etc;	• ETL; • Estatística e Análise de padrões; • Visualização de dados; • Criação e manutenção de modelos de Machine learning e Deep Learning; • Utiliza frameworks: Spark, Hive, Pig, Hadoop; • Co-responsável pela realização de deploy de modelos;

UNIVERSIDADE
CANDIDO
MENDES

EAD