

Big Data



Cruzeiro do Sul Virtual
Educação a distância

Material Teórico



Introdução à *Big Data*

Responsável pelo Conteúdo:

Prof. Dr. Alberto Messias

Revisão Textual:

Prof.^a Dr.^a Selma Aparecida Cesarin

UNIDADE

Introdução à *Big Data*



- Conceitos de *Big Data*;
- Os Vs em *Big Data*;
- A Velocidade na Qual os Dados Estão Fluindo.



OBJETIVO DE APRENDIZADO

- Compreender as características iniciais sobre a tecnologia de *Big Data*;
- Entender e diferenciar o que é *big data* de *data warehouse*;
- Aprofundar-se nos conceitos de volume, variedade e velocidades dos dados em *Big Data*.



Orientações de estudo

Para que o conteúdo desta Disciplina seja bem aproveitado e haja maior aplicabilidade na sua formação acadêmica e atuação profissional, siga algumas recomendações básicas:

Determine um horário fixo para estudar.

Mantenha o foco! Evite se distrair com as redes sociais.

Procure manter contato com seus colegas e tutores para trocar ideias! Isso amplia a aprendizagem.

Seja original! Nunca plágie trabalhos.

Aproveite as indicações de Material Complementar.

Conserve seu material e local de estudos sempre organizados.

Não se esqueça de se alimentar e de se manter hidratado.

Assim:

- ✓ Organize seus estudos de maneira que passem a fazer parte da sua rotina. Por exemplo, você poderá determinar um dia e horário fixos como seu “momento do estudo”;
- ✓ Procure se alimentar e se hidratar quando for estudar; lembre-se de que uma alimentação saudável pode proporcionar melhor aproveitamento do estudo;
- ✓ No material de cada Unidade, há leituras indicadas e, entre elas, artigos científicos, livros, vídeos e sites para aprofundar os conhecimentos adquiridos ao longo da Unidade. Além disso, você também encontrará sugestões de conteúdo extra no item **Material Complementar**, que ampliarão sua interpretação e auxiliarão no pleno entendimento dos temas abordados;
- ✓ Após o contato com o conteúdo proposto, participe dos debates mediados em fóruns de discussão, pois irão auxiliar a verificar o quanto você absorveu de conhecimento, além de propiciar o contato com seus colegas e tutores, o que se apresenta como rico espaço de troca de ideias e de aprendizagem.

Conceitos de *Big Data*

Mayer-Schonb e Cukier (2014) mencionam que *Big Data* refere-se a grandes Conjuntos de Dados que são difíceis de armazenar, pesquisar, visualizar e analisar como, por exemplo, uma Empresa aérea que coleta 10 terabytes de dados de sensores durante 30 minutos de voo do avião.

Nathan e Warren (2015) mencionam que *Big Data* é usado comercialmente para analisar grandes quantidades de Dados para se tomar decisões, baseando-se na análise do comportamento e de preferências do consumidor.

Segundo Smith (2012), *Big Data* refere-se ao processamento e à análise de repositórios de Dados extremamente grandes e que não seriam possíveis processar ou analisar com ferramentas convencionais de análise de Dados.

Big Data requer grande poder computacional para processar eficientemente grandes quantidades de dados em intervalos de tempo toleráveis.

Essa Tecnologia envolve *Massive Parallel Processing Databases (MPP)*, grids de Mineração de Dados, Sistemas de Arquivos Distribuídos, Plataformas de Computação em Nuvem, Redes de Comunicação e Sistemas de Armazenamento Escaláveis.

Os Vs em *Big Data*

Conforme se observa em Zikopoulos e Eaton (2011), a IBM caracteriza e define a tecnologia de *Big Data* por três características juntas: **volume**, **velocidade** e **variedade** de Dados ou, simplesmente, **V3**, conforme se observa na Figura a seguir.

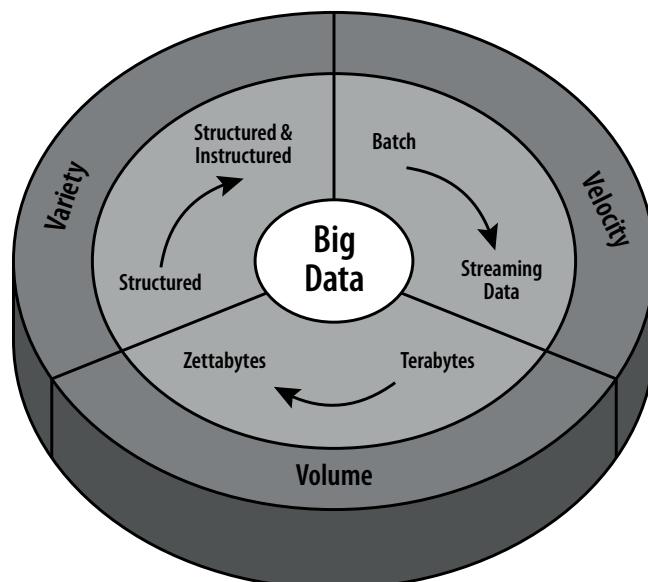


Figura 1

Segue a definição e a exemplificação para cada um dos termos.

Volume de dados

Por que o **volume** é tão grande? É importante que vocês percebam e se façam essa pergunta e procurem exemplos dessas grandes massas de Dados geradas.

O grande volume de Dados armazenados está explodindo. No ano de 2000, eram 800.000 *petabytes* (PB) de Dados que foram armazenados no mundo.

Observa-se que, porém, grande parte dos Dados que estão sendo criados não são analisados. Espera-se chegar em 35 *zettabytes* (ZB) em 2020

Outros exemplos interessantes são:

- O *Twitter* gera mais de 7 *terabytes* (TB) de dados todos os dias;
- O *Facebook* gera 10 TB.

Vê-se, ainda, que algumas Empresas geram *terabytes* de dados a cada hora de cada dia do ano. Já não é inédito que Empresas individuais possuam *clusters* de armazenamento com *petabytes* de Dados.

Por que o volume é tão grande?

Com a diminuição dos custos de produção de dados, o volume de dados da WWW tem crescido rapidamente, dispositivos móveis e ubíquos e sensores sem fio estão gerando Dados a cada minuto ou segundo.

A troca de Dados em massa está ocorrendo a cada segundo entre os milhares de milhões de serviços de *Internet*, aplicações científicas, vigilância por vídeo, registros médicos, dados operacionais da Empresa, manufatura discreta, *e-commerce*... são todas fontes de dados grandes.

Em 2011, a *International Data Corporation* alegou que “A informação do mundo está dobrando a cada dois anos”. Naquele ano, o mundo gerou um escalonamento 1,8 ZB (10^{21} B) de Dados, um aumento de 0,6 ZB ano a ano.

Por fim, prevê-se que, em 2020, o mundo vai gerar até 35 ZB de dados, e isso coloca desafios significativos para armazenamento.

O volume de dados disponíveis para as Organizações, hoje, está em ascensão, enquanto o percentual de dados que eles podem analisar está em declínio, conforme se pode observar na Figura II.



Figura 2

De acordo com o Gráfico exibido na Figura 1, observa-se que as Organizações tem disponíveis ou podem armazenar uma grande quantidade de Dados que não podem processar ou analisar.

Variedade nos Dados

Um dos desafios da Tecnologia de *Big Data* é a variedade de Dados, com a explosão dos sensores e dispositivos inteligentes, as tecnologias de colaboração social, os Dados em uma Empresa tornaram-se complexos: inclui não só os Dados tradicionais ou transacionais, mas os dados “crus”, semiestruturados e não estruturados, mas também, páginas *web*, arquivos de *log* (incluindo cliques), índices de pesquisa, mídia social, fóruns, *e-mails*, documentos, dados de sensores de sistemas ativos e passivos, *Internet das Coisas*, vídeos, áudios e Dados não relacionais.

Os Sistemas Tradicionais têm dificuldades para armazenar, processar e compreender esses Dados, ou seja, os dados gerados não são relevantes com a Tecnologia Tradicional; é preciso perceber novas oportunidades, utilizar novas Tecnologias, algoritmos e técnicas para a análise dos dados estruturados em conjunto com os dados não estruturados.

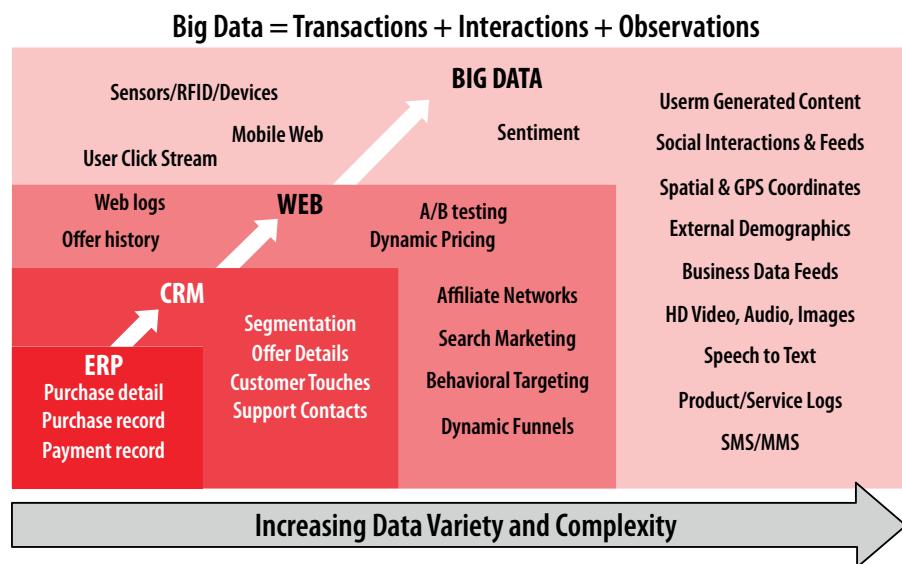


Figura 3

Os Dados Estruturados representam apenas 20% dos Dados, enquanto que os dados não estruturados representam 80%. O sucesso da Organização vai depender de sua capacidade de perceber os *insights* relacionados aos vários tipos de dados de que ela dispõe, o que inclui tanto tradicionais, quanto não tradicionais.

Velocidade nos dados



O quão rápido é rápido? Reflitem quanto a exemplos (questionamento).

Assim como o volume e a variedade de Dados que coletamos e armazenamos mudou, a velocidade a que é gerado também cresceu, a velocidade que normalmente se considera está relacionada ao quanto rapidamente os dados são resgatados, armazenados e suas taxas associadas de recuperação; porém, com *Big Data*, em vez de tratar a ideia de velocidade às taxas de crescimento, associadas aos seus repositórios de dados, sugere-se aplicar essa definição para Dados em movimento.

A Velocidade na Qual os Dados Estão Fluindo

As Empresas de hoje estão lidando com *petabytes* de dados, em vez de *terabytes*. O aumento em sensores e outros fluxos de informação têm levado a um fluxo constante de dados e a um ritmo que se tornou impossível para os Sistemas Tradicionais manusearem.

Além das características citadas, podemos acrescentar a complexidade dos tipos de dados que são produzidos nos Sistemas de Informação e nas diversas Mídias, Redes Sociais e Computação Móvel.

É crescente a análise de dados não estruturados.

Material Complementar

Indicações para saber mais sobre os assuntos abordados nesta Unidade:



Livros

Análise da informação para tomada de decisão, desafios e soluções

Leitura do item 2.3 sobre *Big Data*, do livro: TARAPANOFF, K., com o título **Análise da informação para tomada de decisão, desafios e soluções**. Curitiba: Intersaberes, 2015, presente na Biblioteca Virtual Universitária Pearson;



Leitura

O que é big data?

Leitura e vídeo com introdução à Big Data

<https://goo.gl/nbF6QV>

Volume, variedade, velocidade, veracidade e valor: Os cinco Vs do Big Data

Leitura do artigo da **Computer World** que ilustra os Vs de Big Data

<https://bit.ly/3snARXb>

Volume, variedade, velocidade, veracidade e valor: Os cinco Vs do Big Data

Autor apresenta as definições de Big Data e as aplicações dessa tecnologia

<https://bit.ly/3sdnfhk>

Referências

SOUZA, A. M. C. **Uma nova arquitetura para Internet das Coisas com análise e reconhecimento de padrões e processamento com Big Data.** 2015. Tese de Doutorado. Universidade de São Paulo. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/3/3142/tde-20062016-105809/>>. Acesso em: 07/03/2017.

ZIKOPOULOS, P.; EATON, C. **Understanding big data: Analytics for enterprise class hadoop and streaming data.** McGraw-Hill Osborne Media, 2011.



Cruzeiro do Sul
Educacional