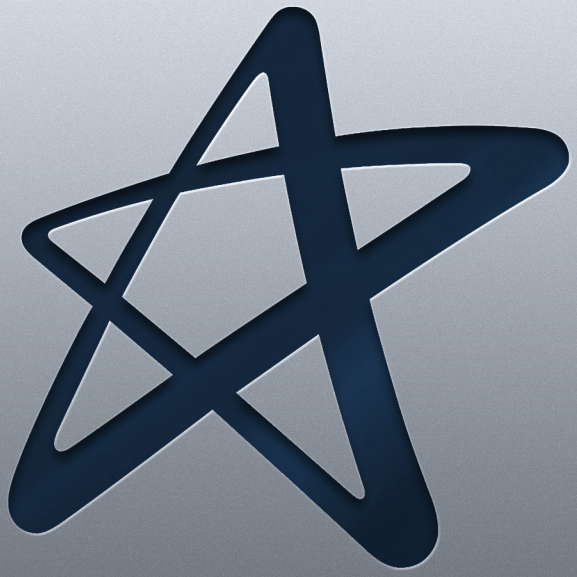
Big Data



Cruzeiro do Sul Virtual
Educação a distância

Material Teórico

Definições, Fontes de Dados e Exemplos de *Big Data*

Responsável pelo Conteúdo:
Prof. Dr. Alberto Messias

Revisão Textual:
Prof.ª Dr.ª Selma Aparecida Cesarin

UNIDADE Definições, Fontes de Dados e Exemplos de *Big Data*

- Novos Vs: Veracidade e Valor;
- Porque *Big Data* é Importante;
- Fontes de Dados em *Big Data*;
- Exemplos de *Big Data*.

OBJETIVO DE APRENDIZADO

- Aprofundar-se em outros conceitos sobre Big Data, bem como a veracidade e o valor dos Dados;
- Saber qual é a importância de Big Data, quais são suas principais fontes de dados e exemplos de 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:

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.

Novos Vs: Veracidade e Valor

Além dos três Vs inicialmente propostos para caracterizar a Tecnologia de *Big Data*, foram adicionados mais dois, o **valor** e a **veracidade** dos dados.

O **valor** está relacionado a agregar ganho e valor ao negócio da organização, enquanto a **veracidade** está relacionada à confiabilidade dos Dados em si.

Veracidade: os Dados são Confiáveis?

A veracidade ou confiabilidade dos dados possui três características importantes:

- A qualidade ou limpeza, consistência e acurácia dos dados;
- A origem ou fonte de dados ao longo do tempo e sua evolução ou linhagem;
- Como se pretende usar o conjunto de Dados, pois o nível de confiança ou qualidade devem ser aceitáveis para o negócio.

Alguns questionamentos podem ser feitos quanto à veracidade dos dados:

- De onde os dados são provenientes;
- Os dados foram gerados interna ou externamente à organização;
- Se os dados poderão ser públicos, como números de telefone ou dados comportamentais a partir dos dados agregadores;
- Se as transações que originam os dados são auditadas ou aprovadas;
- Se o dado é verdadeiro ou uma opinião;
- Se o dado foi fabricado intencionalmente;
- Se o dado bruto pode ser usado, se contém *outliers*, como aberrações ou fraudes, ou se são necessárias padronizações ou limpeza dos Dados;
- Se os métodos de governança na Organização são usados para vetar ou medir a veracidade ou classificar as dimensões de Dados; se as fontes de Dados internas se tornam externas, elas devem ser auditadas;
- Como classifica o fator de confiabilidade dos dados. Organizações sérias, consideram essa classificação como parte do processo de governança.

Em organizações que possuem dados ou transacionam dados através da *WEB*, a confiabilidade nos Dados torna-se mais crítica e também envolve aspectos de veracidade e governança de Dados.

Observe que a validação dos dados oriundos de diversos Sistemas, Mídias ou Redes Sociais, trazem uma complexidade grande para a escolha, transformação, processamento, análise e validação dos resultados e informações geradas.

Valor: Investir em *Big Data* me Dará Retorno?

Big data pode ter valor para uma Empresa? Procure exemplos.

Obter vantagem sobre a sua concorrência pode significar a identificação de uma tendência, problema ou oportunidade em apenas alguns segundos, ou até mesmo microssegundos.

Cada vez mais, os dados que são produzidos possuem uma vida útil muito curta, por isso as Organizações devem ser capazes de analisá-los quase em tempo real, se eles esperam encontrar ideias e oportunidades nesses Dados.

Observa-se que 1 em cada 3 gestores tomam decisão com base em informações que não confiam ou não tem; 56% se sentem sobrecarregados com a quantidade de dados que gerenciam; 60% acreditam que precisam melhorar a captura e entender informações rapidamente e, 83% apontam que BI & *Analytics* fazem parte de seus planos para aumentar a competitividade (fonte: *Survey KPMG)*.

Projetos com *Big Data* frequentemente não obtêm sucesso quando o V de *valor* é ignorado; também foi mostrado que as Empresas que investem em análise de Dados como um ativo para a tomada de decisão são mais bem sucedidas.

Nesse contexto, o valor é qualquer aplicação de *Big Data* que impulsione aumentos de receita (como a análise de fidelidade de clientes), identifique novas oportunidades de receita, melhore a qualidade e a satisfação do cliente (como a manutenção preditiva), economize custos, melhores resultados (por exemplo, atendimento ao paciente).

É fundamental que as Organizações que utilizam *Big Data* ganhem experiência usando a Tecnologia em um modo "experimentação", que é incentivada, ao mesmo tempo, com a identificação de um caso de teste e que impulsione o sucesso do negócio.

Já há novas referências que apontam que *Big Data* pode ter até dez Vs, nesse caso, acrescentando:

- **Variabilidade**: que possui relação com as inconsistências que poderão ser encontradas nos Dados, dada as diversas fontes de Dados e à quantidade de dimensões que os Dados poderão possuir;
- **Validade**: que possui relação com a veracidade dos Dados, nesse caso, com a correção dos Dados, uma etapa importante no dia a dia do analista de Dados;
- **Vulnerabilidade:** os Dados provenientes das fontes de *Big Data* poderão também sofrer ataques, em especial de roubo de Dados. Sua proteção é um grande desafio;

- **Volatilidade:** as Organizações devem mensurar qual seria o período em que seus Dados tornam-se irrelevantes para análises, ou seja, Dados que possuem maior volatilidade poderão ser descartados num dado período de tempo;
- **Visualização:** a visualização dos Dados provenientes das fontes de *Big Data* é um desafio para as ferramentas e os Analistas de Dados, em especial de exibição de DADOS, dadas as características de velocidade, volume e multiplicidade de variáveis de Dados. As alternativas, em grande parte, são as visualizações em *clusters*, árvores e diagramas, entre outras.

Porque *Big Data* é Importante

Soluções de *Big Data* são ideais para analisar não apenas os dados estruturados, mas os dados não estruturados e os semiestruturados a partir de uma ampla variedade de fontes de Dados.

Soluções de *Big Data* são ideais quando todos, ou a maioria, dos Dados precisam ser analisados contra uma amostragem de Dados não é tão eficaz como um conjunto maior de Dados do que para derivar análise (confrontar as análises).

Soluções de *Big Data* são ideais para análise exploratória, iterativa e quando as medidas comerciais sobre os Dados não são predeterminadas.

As Organizações são desafiadas com uma série de considerações, como, quais outras fontes de Dados estão disponíveis e como elas podem ser usadas com o que é conhecido ou como permitir novos *insights*.

Aqui estão alguns exemplos:

- Como gerenciar, usar e analisar os Dados quando não estiverem num formato comum, familiar ou prontamente utilizável (como um Banco de Dados relacional);
- Como usar fontes não tradicionais de Dados de clientes (como Dados de *Call Center* ou comentários de mídia social) para entender melhor e prever o comportamento do cliente;
- Como processar dados sensíveis ao tempo para a tomada de decisões em tempo real, como a identificação de riscos de segurança;
- Como usar o intervalo de Dados de *log* e de sensor para responder a eventos baseados em máquinas, prever o tempo de inatividade ou garantir que os contratos de nível de serviço sejam mantidos;
- Como encontrar e obter Dados de alto valor a partir das novas fontes de dados que podem se conectar a fontes tradicionais para obter *insights* aprimorados;
- Como tirar proveito das novas Tecnologias para diminuir o custo total de propriedade (*TCO*), enquanto ainda obtendo o máximo valor de seus dados.

Embora existam certamente diferenças e variações entre as Indústrias, cinco casos de uso primário surgiram em torno desses desafios:

- Grande exploração de Dados;
- Vista 360° do cliente;
- Extensões de segurança e inteligência;
- Análise de operações;
- Aumento do *Data Warehouse*.

Esses casos de uso não são mutuamente exclusivos. A identificação e o uso de informações baseadas em sensores podem se resumir e se correlacionar com outras informações tradicionais, como vendas ou preços num *Data Warehouse* aumentado ou estendendo e, em seguida, conduzir a análises, proporcionando maior visão sobre melhores campanhas de *marketing* para clientes ou melhor identificação de fraudes ou riscos de segurança.

Considere que quando se trata de resolver desafios de Gerenciamento de Informações, utilizando *Big Data*, sugere-se considerar:

- *Big Data* é o inverso do paradigma tradicional análise de dados, apropriado para a tarefa de negócio;
- Veja uma plataforma de *Big Data* complementando o que se tem atualmente para análise;
- É importante que se tenha sinergia com as soluções existentes para melhores resultados de negócios;
- É adequado para se resolver os desafios de informação que não se encaixam de forma nativa dentro da abordagem Banco de Dados relacional tradicional.

Fontes de Dados em *Big Data*

Os profissionais da Área de Gerenciamento de Dados estão relativamente confortáveis com o conceito de Dados Estruturados. Esses Dados correspondem a padrões repetidos de estrutura de Dados, como, por exemplo, a modelagem de Dados tradicional, com o formato de linhas e colunas.

A análise de *Big Data* pode conter dados semiestruturados e não estruturados, como textos em claro, imagens, vídeos, áudio ou *XML*.

A Figura 1, presente no *RedBook IBM*, ilustra as principais fontes de dados para *Big Data* vistas atualmente.

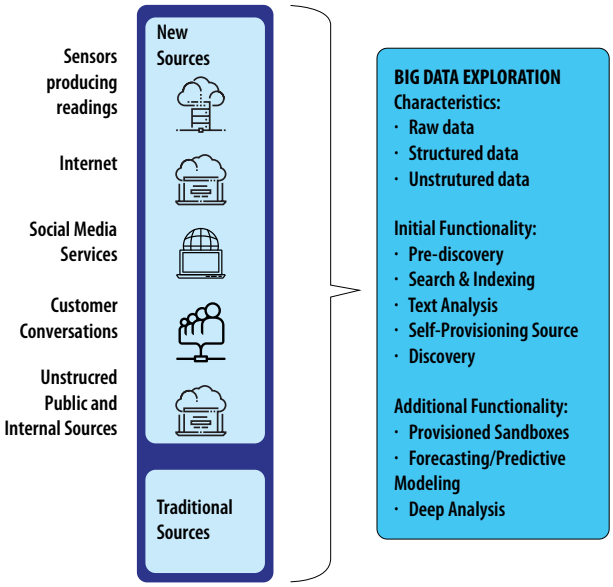


Figura 1 – Fontes de dados na exploração de dados com *Big Data*

Segue uma pequena descrição dos principais tipo de Dados encontrados.

Social Media

Blogs, *tweets*, *sites* de Redes Sociais (por exemplo, *LinkedIn* e *Facebook*), feeds de notícias, fóruns de discussão, *sites* de vídeos, e todos os outros abaixo dessa categoria, possuem *APIs* específicas para resgate de Dados e seus metadados.

Web Logs

Os *logs* de serviços *WEB* existem em vários semiestruturados formatos. Tipicamente, eles possuem informações a respeito do ambiente de execução, entradas sobre as atividades do servidor de aplicação e, essencialmente, tudo o que acontece no servidor. Tradicionalmente, um *log* contém informações transacionais de cada conexão, a origem, o início e o fim das requisições e conexão, incluindo qualquer erro que tenha ocorrido.

Os *web logs*, tradicionalmente, são utilizados para diagnósticos de erros e análises técnicas, mais recentemente, eles são utilizados em combinação com outras fontes de dados para se entender o padrão de comportamento das ações dos usuários em *sites* e para a identificação de ameaças de segurança.

Dados Gerados por Máquinas

Dados gerados por máquinas constituem uma grande variedade de dispositivos, leitores de *RFID*, sensores ópticos, de áudio, sísmicos, térmicos, químicos, dispositivos médicos ou de clima, sensores em estradas ou ruas, televisores, câmeras de vídeo, sensores corporais ou vestíveis.

Observa-se que diversos *sites* ou Sistemas disponibilizam os dados de clima em tempo real e em vários formatos, como numéricos ou textuais, como, por exemplo, **26 graus *celsius* e com nuvens esparsas** ou **September 24, 2016 at 17:56, 26 *degrees.***

***Celsius* com poucas nuvens**; note que para uma aplicação deverá existir um padrão de leitura.

GPS ou Geolocalização

Dados de geolocalização se tornaram ubíquos. Temos como origens desses dados, Sistemas de GPS em veículos, aviões, navios, *smartphones* e utilizamos esses dados para guiar nossos movimentos ou para rastrear nossos movimentos em aplicações de segurança ou emergência, ou para rastreamento de nossos *smartphones* em lojas e *shoppings* em busca de análise de comportamento.

Outro uso importante de geolocalização são os serviços de rastreabilidade logística que pode mostrar para a Empresa ou para o cliente em que se encontra o determinado produto.

Streaming Data

Os dados de fluxo são uma categoria especial de *Big Data;* ao invés de ser um formato, é um tipo especial de processamento. A transmissão é contínua de qualquer tipo de Dados e quase em tempo real. Alguns exemplos de aplicações de fluxo contínuo incluem detecção de fraude, segurança física, monitoramento de tráfego, monitoramento veicular ou monitoramento médico.

Essas aplicações utilizam protocolos específicos para enviar e receber informações; quase sempre, recebem os dados, processam-nos e os encaminham para outras aplicações.

Os tipos de origens de dados para *Big Data* não se limitam a esses colocados no texto; porém, essas categorias englobam grande parte dos exemplos de fontes de Dados em *Big Data*.

Exemplos de *Big Data*

Existem diversos exemplos de aplicabilidade da Tecnologia de *Big Data*, como, por exemplo, sistemas de recomendação de filmes, presentes no *Netflix*, sistemas de recomendação de leituras e notícias, sistemas de monitoramento para segurança física, sistemas para classificação ou criação de perfis de clientes, análise de sentimento, análise de textos de redes sociais de modo a perceber as opiniões de clientes quanto à Empresa, análise de comportamento de alunos em Ambientes Virtuais de Aprendizagem, análise de pacotes de redes em busca de pacotes

infectados ou de anomalias, *web analytics* que acompanha o perfil de cliente em *sites* de vendas, análise de séries temporais de dados de sensores, análise de dados financeiros em busca de fraudes, aplicabilidade em jogos massivos, aplicabilidade em análise de dados médicos.

Há diversas aplicabilidades da Tecnologia de *Big Data*. O quadro a seguir ilustra algumas áreas que já utilizam a análise de dados com *Big Data*.

Quadro 1 – Alice Leplante, *The Big Data Transformation*: Understanding Why Change Is Actually Good for Your Business, 2016, Estados Unidos: O'Reilly.

Industry	Big Data use cases
Automotive	Auto sensors reporting vehicle location problems
Financial services	Risk, fraud detection, portfolio analysis, new product development
Manufacturing	Quality assurance, warranty analyses
Healthcare	Patient sensors, monitoring, electronic health records, quality of care
Oil and gas	Drilling exploration sensor analyses
Retail	Consumer sentiment analyses, optimized marketing, personalized targeting, market basket analysis, intelligent forecasting, inventory management
Utilities	Smart meter analyses for network capacity, smart grid
Law enforcement	Threat analysis, social media monitoring, photo analysis, traffic optmization
Advertising	Costumer targeting, location-based advertising, personalized retargeting, churn detection/prevention

Material Complementar

Indicações para saber mais sobre os assuntos abordados nesta Unidade:

Sites

TDWI

Leitura na ual o autor coloca novos Vs em Big Data, com novas variações.

https://bit.ly/3Qs3ABd

Exame

Artigo com entrevista com um especialista e fundador da Empresa R18.

https://bit.ly/3C7L24P

DataFloq

Referência que ilustra as diversas fontes de dados em *Big Data*.

https://bit.ly/3A2KoD1

Leitura

Você realmente sabe o que é *Big Data*?

Artigo IBM sobre *Big Data*.

https://ibm.co/3w94ytS

Referências

BALLARD, Chuck *et al.* **Information Governance Principles and Practices for a Big Data Landscape**. USA: Redbook IBM 2014. Disponível em: <http://www.redbooks.ibm.com/abstracts/sg248165.html?Open>.

LEPLANTE, Alice. **The Big Data Transformation:** Understanding Why Change is Actually Good for Your Business. 2016. Estados Unidos: O'Reilly. Disponível em: <https://saas.hpe.com/en-us/asset/big-data-software/big-data-transformation-understanding-why-change-actually-good-your-business>.


Cruzeiro do Sul
Educacional