



Big Data

MSc. Wesley Lima

Roteiro

- Definição de Dados
- Formato dos Dados
- A era dos Dados
- Os V's de Big Data
- BI vs Big Data
- O profissional de Big Data
- Ciclo de vida dos Dados
- Oportunidades perdidas



O que são Dados?



Dados são fatos coletados e normalmente armazenados.

Informação é o dado analisado e com algum significado.



Conhecimento é a informação interpretada, entendida e aplicada para um fim.



O formato dos Dados



O dado analógico era armazenado em disco de vinil, fita de vídeo e fita cassete.

O dado digital é armazenado em forma de “zero e uns”, independente da estrutura.



Não Eletrônico na Biblioteca do Congresso Americano existem mais de 150 milhões de exemplares de livros.



Como são gerados os Dados

Dados gerados por humanos

Conteúdo gerado a partir do pensamento de uma pessoa;

Propriedade intelectual está integrada ao dado;

Interação das pessoas no mundo digital - Mídias sociais, Blogs, avaliação em e-commerce.

Dados gerados por máquinas

Dados digitais produzidos por processos de computadores, aplicação e outros mecanismos;

Arquivos de log;

Localização;

Interação máquina com máquina - IoT.



Como surgiu o Big Data?



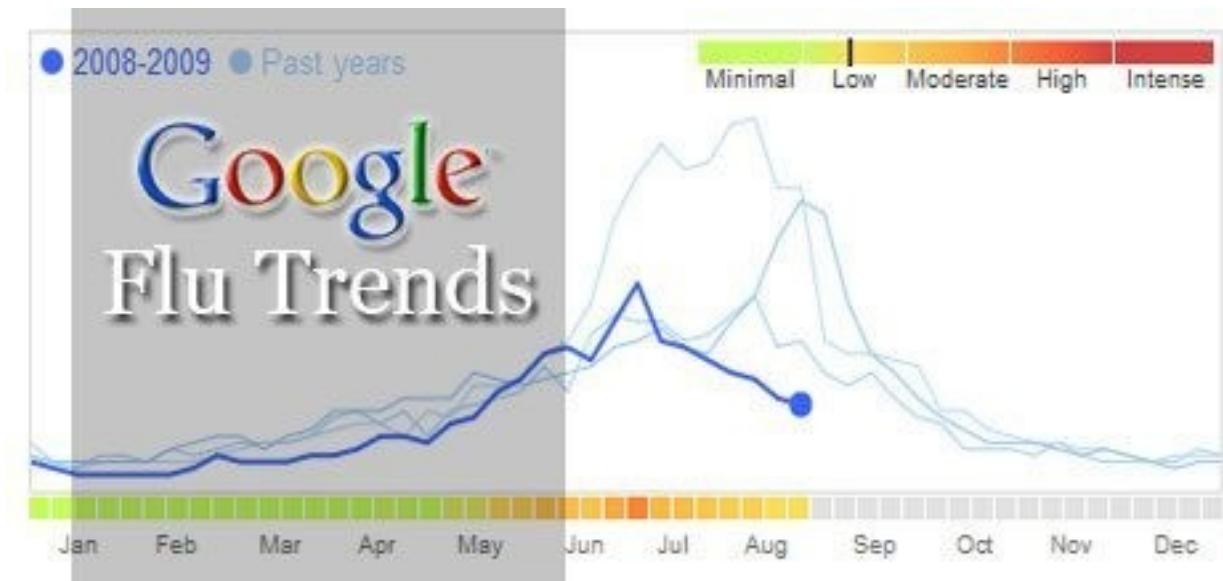
BIG DATA



Como surgiu o Big Data?

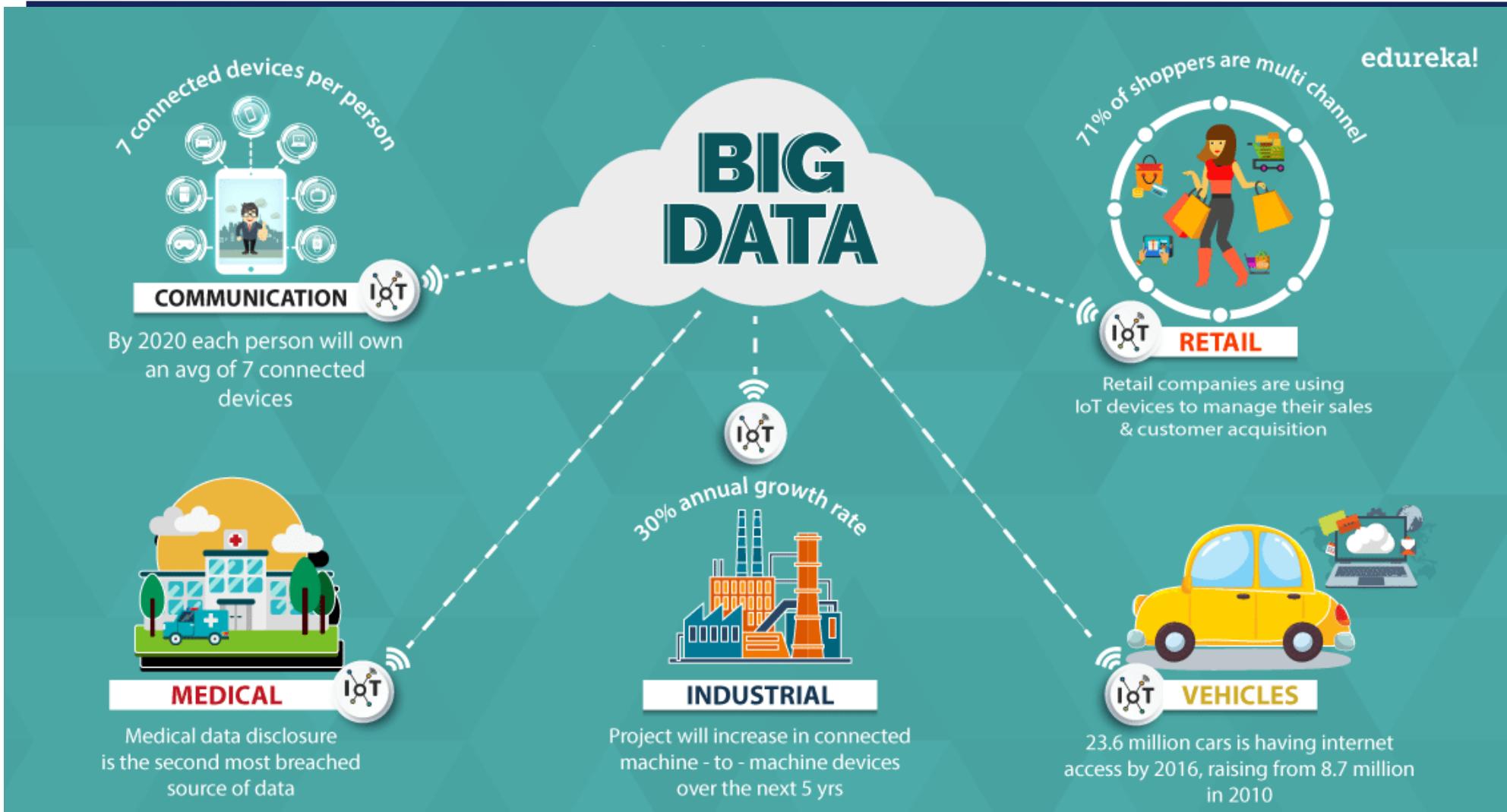
Google Flu Trends - 2009

- Surgimento da gripe H1N1;**
- Analisava os dados de busca;**
- Criou uma metodologia de previsão de surtos da gripe;**
- Criticas: não possuir uma teoria bem definida;**
- Quais as causas que levaram a metodologia a cometer falhas?**





A era dos Dados





A era dos Dados

2019 *This Is What Happens In An Internet Minute*





A era dos Dados

Principais fatores para o aumento do volume de dados

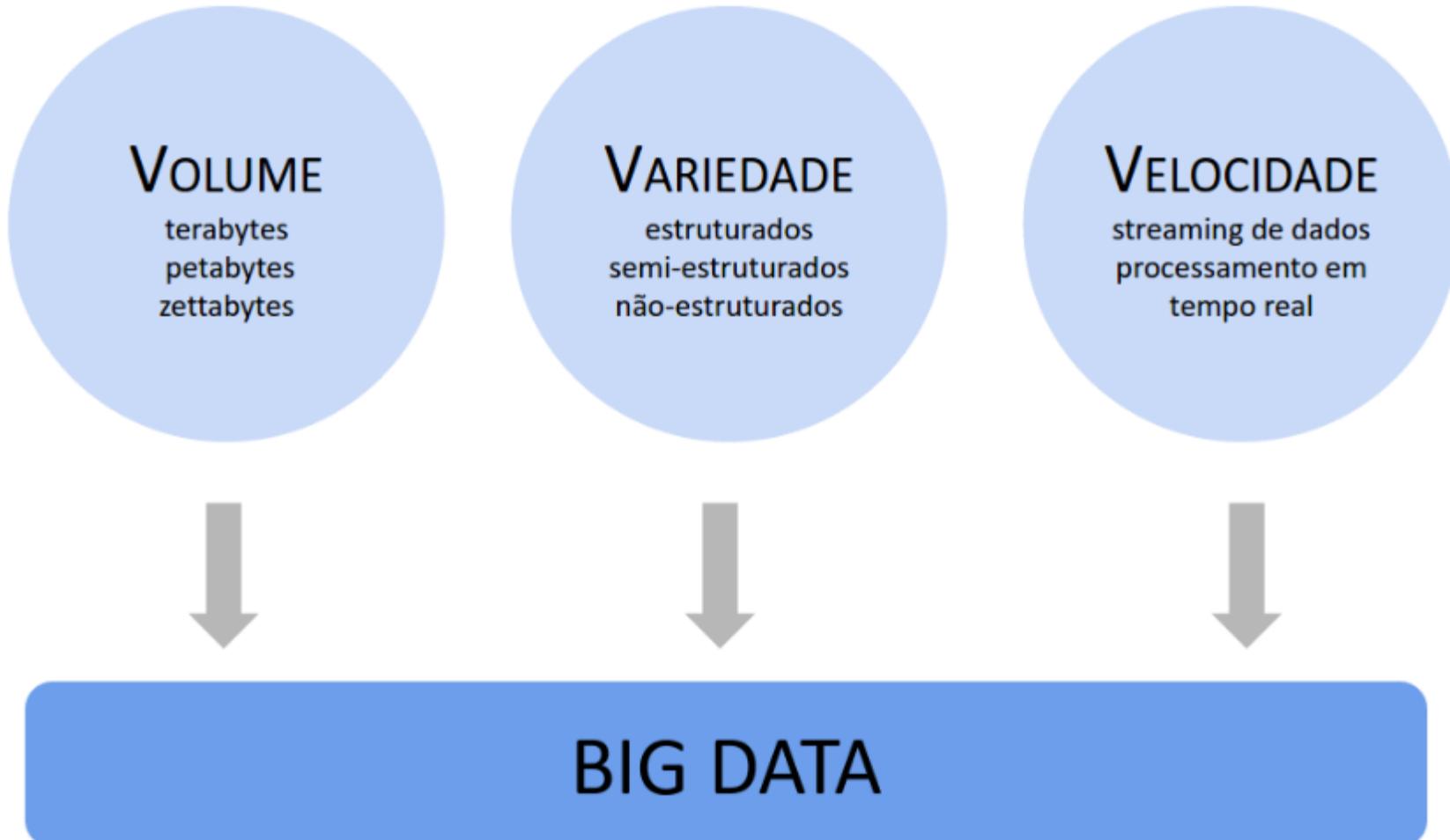
- + Uso de dispositivos móveis
- + Poder de processamento
- + Internet das Coisas



- Custo de armazenamento de dados em disco rígidos



Os 3 V's do Big Data





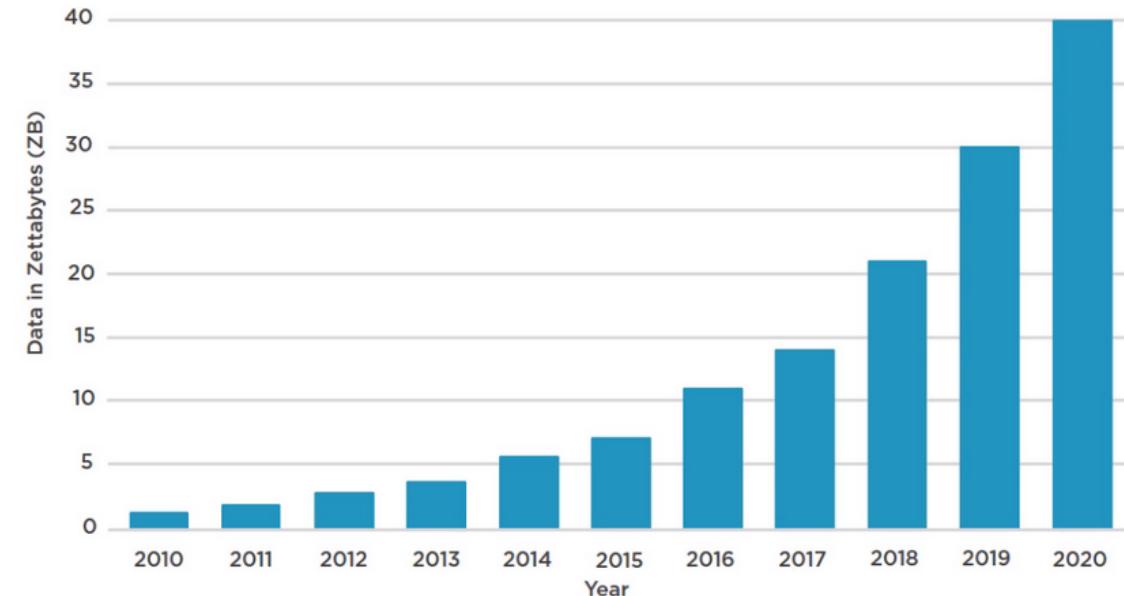
Volume

É a característica mais significativa no conceito de Big Data;

Não é apenas volume;

Qual o volume mínimo para ser considerado Big Data?

Limitação das ferramentas tradicionais para lidar com o grande volume de dados.





Variedade

Os dados são gerados nos mais diversos tipos e formatos;

Estruturado, semi - estruturado e não - estruturado;

**Dados semi - estruturados: possuem uma estrutura pré-definida, porém como menos rigor que os bancos de dados relacionais?
Ex: json e XML.**

Dados não - estruturados: vídeos, imagens, alguns formatos de texto, etc.





Velocidade

A velocidade está relacionada com a coleta, análise e utilização dos dados;

Recomendação de produto em um e-commerce;

Com o tempo os dados podem perder “Valor”.

Geração de dados.

Atualização de preços a cada 10 minutos de acordo com a demanda em tempo real.





Mais 2 V's

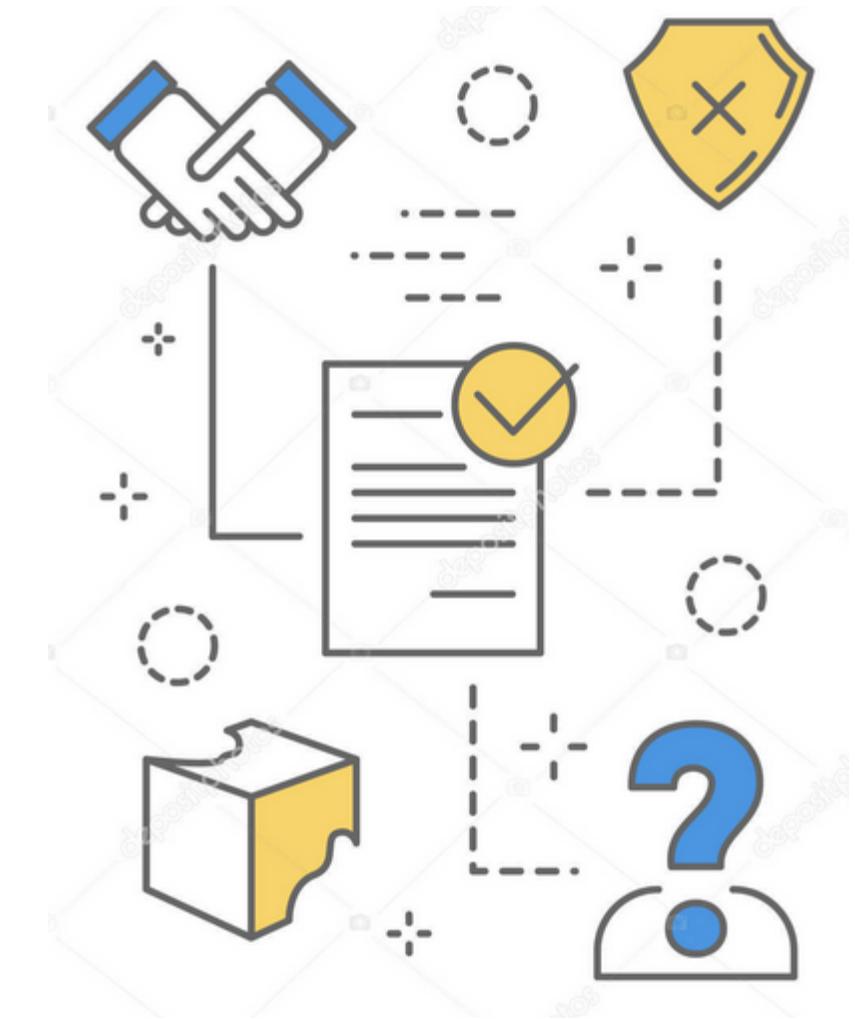
★ Veracidade:

Garantia que os dados coletados são autênticos (em relação a fonte);

No momento da análise os dados são verdadeiros;

Nem tudo postado nas redes sociais é verdadeiro;

Todo sistema pode possuir dados com erros.





Mais 2 V's

★ Valor:

Faz referência ao quanto valioso e significativo um dado pode ser para uma solução;

A análise do valor é importante para determinar quais dados serão priorizados para a solução.





Big Data

“Big Data, em geral, é definido como ativos de alto volume, velocidade e variedade de informação que exigem custo – benefício, de formas inovadoras de processamento de informações para maior visibilidade e tomada de decisão.”

Gartner Group (2012)





O que é Big Data...

Big Data é uma mudança social, cultural, é uma nova fase da revolução industrial;

Uma estratégia baseada em tecnologia que permite a coleta de insights mais profundos e relevantes dos clientes;

Um novo paradigma no qual a TI colabora com usuários empresariais e “cientistas de dados” para identificar e implementar análises que ampliam a eficiência operacional e resolvem novos problemas empresariais.



O que não é Big Data...

Só tecnologia. No nível empresarial, refere-se a explorar as amplamente melhoradas fontes de dados para ganhar insights.

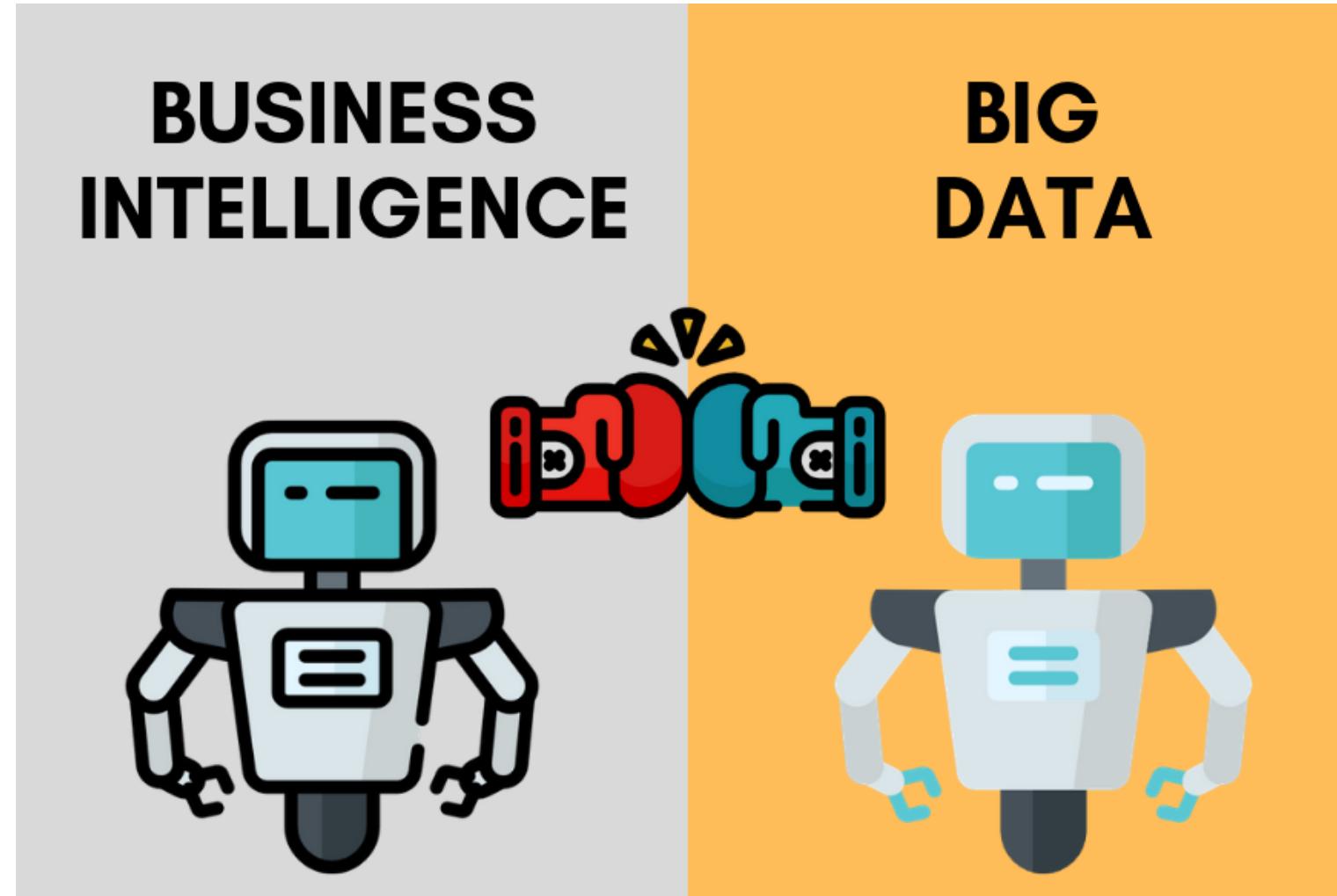
Somente volume. Também refere-se à variedade e velocidade. Mas, talvez mais importante, refere-se ao valor derivado dos dados;

Apenas MapReduce;

O fim do modelo relacional ou do data warehouse.



BI vs Big Data





BI vs Big Data

**BI depende de dados estruturados;
Não necessita de profissionais especialistas em
estatística/engenharia;
Análise descritiva;
Análise de diagnóstico.
O quê? Quanto? Quando? Onde?**

**Big Data não depende da estrutura dos dados;
Necessita de profissionais com conhecimento
multidisciplinar;
Análise preditiva;
Análise prescritiva;
Por quê? E se? O que acontecerá? Como otimizar?**





Professional de Big Data

Professional de Mercado

- Conhecimento multidisciplinar
- Gerência de Projetos
 - Liderança
- Equipe de Especialistas

Professional Idealizado

- Especialista em todas as áreas
- Foco em Conhecimento Técnico
- Trabalha sozinho
- Especialista em todas as áreas

Eles queriam o Watson??



Professional de Big Data

Um Engenheiro de Dados precisa ser bom em:

- Arquitetar sistemas distribuídos;
- Criar pipelines confiáveis;
- Combinar fontes de dados;
- Criar a arquitetura de soluções;
- Colaborar com a equipe de Data Science e construir as soluções certas para essas equipes.



Engenheiro de Dados



Profissional de Big Data

Um Cientista de Dados precisa ter:

- **Conhecimento em matemática/estatística;**
- **Conhecimentos sobre algoritmos, Inteligência Artificial e Machine Learning;**

• Familiaridade com sistemas de programação como Python, R e Scala;

• Domínio sobre linguagens de consulta e tecnologias associadas (SQL, MySQL, PostgreSQL, MongoDB, Cassandra etc.).



Cientista de Dados



Profissional de Big Data

Equipe de extração:

- Possui função crítica no projeto;
- Pode consumir até 90% do tempo e dos recursos do projeto;
- Geralmente são chamados de spiders;
- Tem a função de extrair e armazenar os dados;
- Checar se os dados extraídos são: os esperados, estão completos, íntegros e atualizados.





Ciclo de vida dos Dados





Ciclo de vida dos Dados

CICLO DE VIDA DOS DADOS		
FASE DO CICLO	ANTES DA LGPD	COM A LGPD
Coleta	Os dados pessoais são coletados indiscriminadamente.	Os dados pessoais coletados devem obedecer ao princípio da necessidade e da finalidade.
Processamento	Os dados podem ser processados sem um tratamento específico.	O processamento de dados só poderá ser realizado se o tratamento estiver enquadrado no Art. 7º da LGPD.
Análise	A análise de dados é feita para entender o mercado, conhecer o perfil das pessoas e definir estratégias para oferecer bens e serviços para o público-alvo.	A análise de dados deve levar em consideração a finalidade da coleta. Devem ser obedecidos os princípios de tratamento, com propósito legítimo, específico e explícito.
Compartilhamento	Os dados pessoais são compartilhados sem a necessidade do consentimento de seus titulares.	O compartilhamento de dados deve ser consentido pelos seus titulares.
Armazenamento	Os dados pessoais são armazenados e mantidos por tempo indeterminado.	Os dados pessoais devem ser armazenados e mantidos por prazos definidos, ou seja, até que finalidade seja alcançada ou deixem de ser necessários ou pertinentes ao alcance da finalidade.
Reutilização	Os dados pessoais são reutilizados sem a necessidade de consentimento de seus titulares.	Um novo consentimento deve ser solicitado sempre que houver mudança de finalidade.
Eliminação	Os dados pessoais são mantidos sem a obrigatoriedade de serem eliminados.	Os dados pessoais devem ser eliminados após o término de seu tratamento.



Oportunidades perdidas

Principais causas para se perder oportunidades em empresas:

1. Os dados não estão integrados	Eles já são gerados pela empresa, mas por serem armazenados em diferentes sistemas e bases, não fornecem uma visão ampla de um problema.
2. Os dados demoram para ser analisados	Gasta-se muito tempo no processo de análise dos dados, o que impede a identificação de informações em tempo hábil.
3. Os dados não estão categorizados	Os registros dos conjuntos de dados estão armazenados de diferentes maneiras, sem padronização dos campos.
4. Os dados estão obscuros	Só é possível obter informações a partir da análise de outros dados, como a identificação de padrões em streaming de vídeos, dados manuscritos, etc.



Oportunidades perdidas

Principais causas para se perder oportunidades em empresas:

5. Os dados não são usados na tomada de decisão	São os que poderiam ser utilizados no processo de apoio à tomada de decisão, mas por não serem integrantes dos dados tradicionais da empresa, são descartados.
6. Os dados não são visualizados com clareza	São situações nas quais os dados já são armazenados, porém não são analisados e apresentados de maneira efetiva para gerar percepções.
7. Os dados não são medidos	Refere-se a casos nos quais não se utilizam as métricas que os dados podem fornecer para a compreensão de um fato, até então, imperceptível.



O caso da Microsoft

Gasto com energia elétrica;

A empresa já coletava dados por meio de diversos sensores;

Cada departamento armazenava e utilizava conforme suas necessidades;

Integração, análise e visualização;

Economia de 60 milhões de dólares.





O caso da Pirelli

Entrega de produtos:

Demora na geração de relatórios;

Um relatório demorava 1 dia para ser entregue;

Estoque vs Pedido de Vendas;

Utilizando Big Data o relatório passou a ser entregue em 10 minutos;





Questões

- 1. Quais dados estruturados, semi - estruturados e não estruturados são gerados pela minha empresa ou na área que atuo?**

- 2. Como os dados gerados por humanos e por máquinas são utilizados?**

- 3. Há problemas no tempo gasto para analisar os dados?**

- 4. Existem dados que poderiam agregar valor à empresa se fossem adquiridos?**



Obrigado
Até a próxima aula...