INTRODUÇÃO À CIENCIA DE DADOS

Ciência de Dados e Big Data

- Qual a relação entre Ciência de Dados e Big Data?
- R- **Ciência de Dados** compreende o conjunto de técnicas, ferramentas e procedimentos para análise de dados e **Big Data** é a nossa matéria prima, é onde será aplicada a Ciência de Dados.

O que são Dados?

São fatos individuais, estatísticas ou medidas que são Coletados por observação/medição. Em termos práticos, tudo que se pode medir, observar, coletar podem ser chamados de dados.

Tecnicamente falando, os dados são um conjunto de valores de variáveis qualitativas ou quantitativas sobre eventos, pessoas ou objetos.

O que são Dados?

- Dados x Informação
- Dados: Consistem em fatos brutos (não trabalhados)
- Informação: Coleção de fatos organizados de modo que adquirem um valor adicional.

De Onde Vem os Dados?

- Usamos smartphones desde que acordamos e em tempos pré-determinados durante o dia.
- A tecnologia está em constante evolução. Há vinte anos atrás não existiam soluções capazes de facilitar nossas ações diárias.
  - Smartphone nos acorda com a música favorita.
  - Nossos compromissos são notificados com antecedência.
  - Documentos podem ser buscados facilmente acessando a internet em um serviço de computação em nuvem para armazenamento de dados.
  - Solicitar serviço de transporte de passageiros por meio de um aplicativo.
  - Etc...
- Você é capaz de imaginar sua rotina diária sem os recursos tecnológicos existentes? Seja para lazer, viagens, compras ou trabalho, a tecnologia nos proporciona facilidades que antes eram inimagináveis.

De Onde Vem os Dados ?

PROCEDIMENTOS MÉDICOS; E-COMMERCE; COMPARAÇÃO DE PREÇOS DE PASSAGENS; DEFINIÇÃO DE TRAJETO POR AUXÍLIO DE GPS; SERVIÇOS DE STREAMING DE FILMES; SÉRIES E MÚSICAS; PEDIDOS ON-LINE DE SERVIÇOS ALIMENTÍCIOS; INTERNET BANKING; SENSORES E SISTEMAS DE MONITORAMENTO; COMPARTILHAMENTO DE MOMENTOS EM REDES SOCIAIS; BUSCA E CANDIDATURA DE VAGAS DE TRABALHO ON-LINE,...

De Onde Vem os Dados ?

Qual a semelhança entre os serviços apresentados anteriormente?

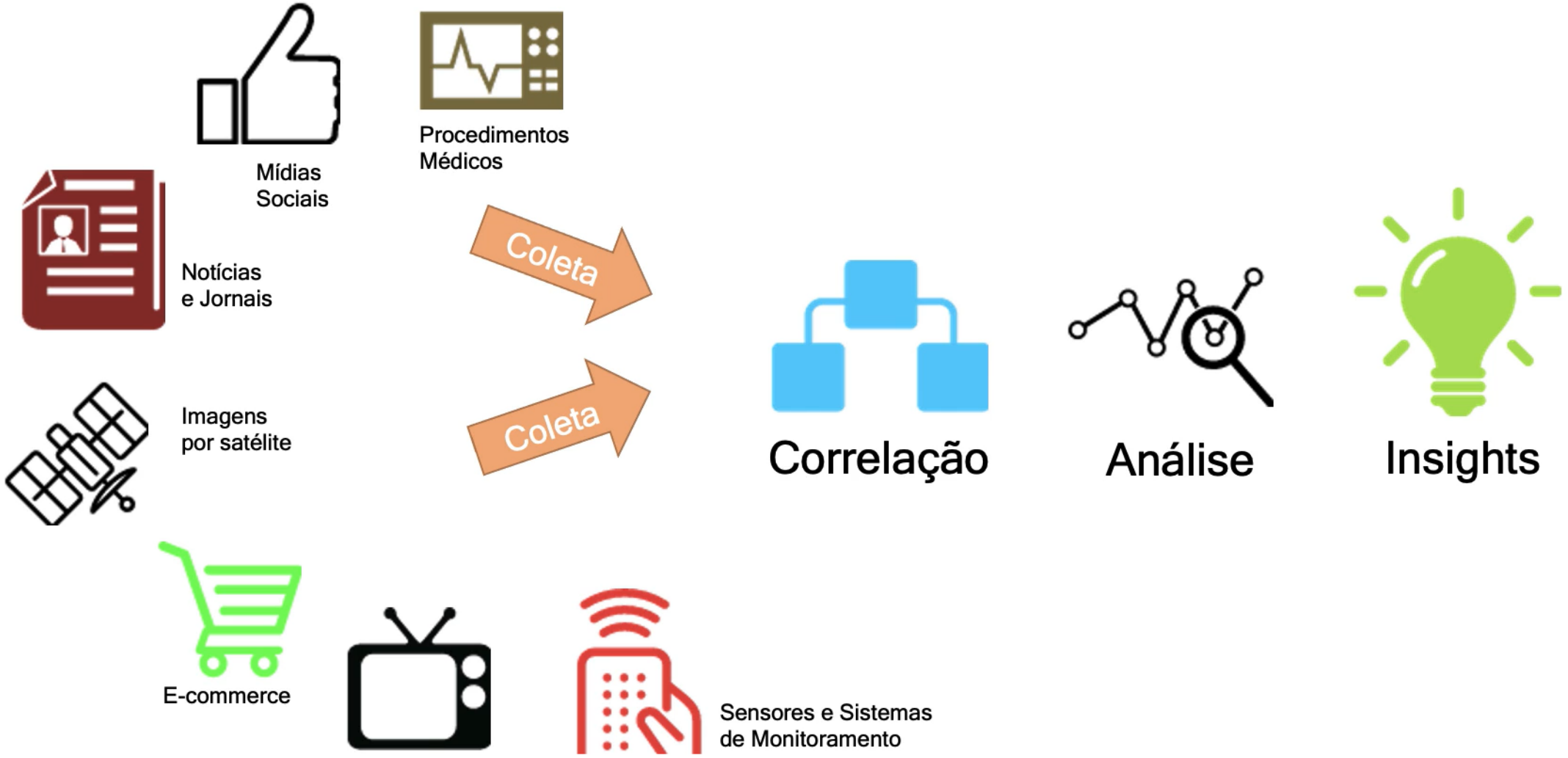
R- A quantidade de dados que eles geram.

- Os avanços em **hardware**, **software**, **tecnologias** e **infraestrutura de redes** foram responsáveis para que chegássemos à “era dos dados”.
- Um estudo feito pela revista Science apontou que, em 1996, somente 0.8% dos dados eram armazenados em formato digital, enquanto em 2007 a quantidade de dados digitais já era de 94%.

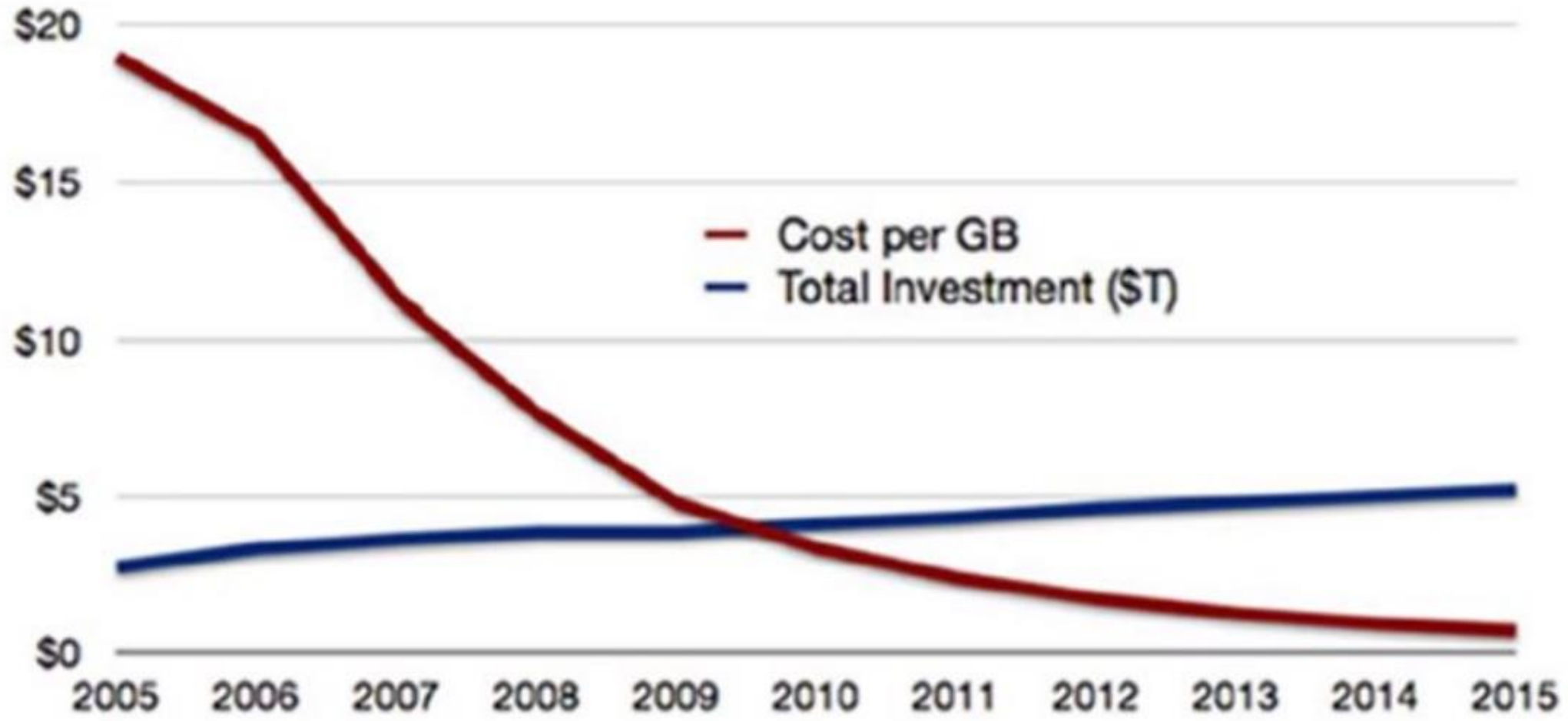
De Onde Vem os Dados ?

- Principais fatores para o aumento no volume de dados:

Big Data e Data Science



Custo de Armazenamento de Dados



Custo de Armazenamento de Dados

- Custo de armazenamento de 1 megabytes em 1990 era de aproximadamente U$ 12.000, a média atual é de apenas U$ 0.03.
- Aumento do poder de processamento.
- Lei de Moore.
  - A capacidade de processamento dos computadores dobraria aproximadamente a cada 18 meses.
- Com o volume de dados crescendo e novas tecnologias habilitadoras para a geração desses dados, empresas de diversos segmentos passaram a perceber o **potencial** dos diferentes tipos de dados.
  - Aperfeiçoar processos.
  - Aumentar a produtividade.
  - Melhorar o processo de tomada de decisão.
  - Desenvolver novos produtos ou serviços.
- Logo, surgiram soluções que utilizam uma série de dados para inúmeros propósitos.

Soluções de Big Data

➢ Na indústria varejista, que adotam etiquetas de identificação por radiofrequência (RFID).

➢ Na agricultura, utilização de redes de sensores, que coletam fluxos de dados em tempo real para fornecer suporte às ações referentes ao processo de plantação, cultivo e colheita.

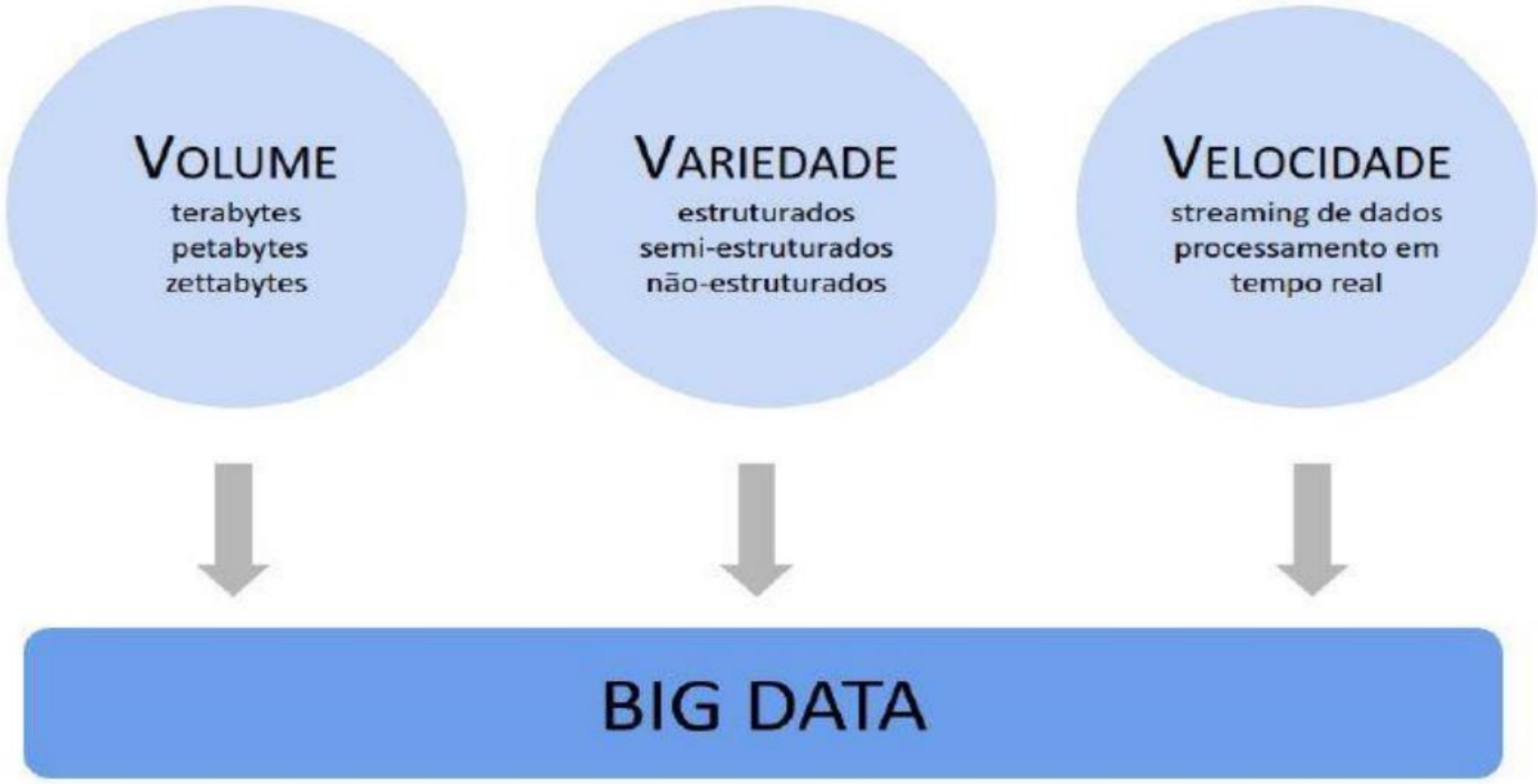
➢ Mesmos havendo tantos dados, um estudo do EMC apontou que, em 2012, de todos os 643 exabytes de dados existentes no mundo digital, somente 3% foram utilizados.

Os Vs de Big Data

Além do próprio nome Big Data, “grande quantidade de dados”, dizer uma de suas principais características, existem outras como os Vs de Big Data.

- Os 3 Vs de Big Data está relacionado com as suas características.
- Volume.
- Variedade.
- Velocidade.

Os Vs de Big Data



Os Vs de Big Data: Volume

- O atributo **volume** é a característica mais significativa no conceito de Big Data. Ele faz referência à **dimensão** sem precedentes do volume de dados.
- 90% dos dados foram criados nos últimos dois anos.
- Origem para tanto dados:
  - A cada segundo, cerca de 40.000 buscas são realizadas no Google.
  - A empresa Walmart manipula mais de 1 milhão de transações dos clientes por hora.
- Uma dúvida frequente relacionada ao volume de dados é a identificação de **quando** um determinado conjunto de dados pode ser considerado **Big Data**.
- É preciso ter uma quantidade de **petabytes de dados** para ter uma solução de Big Data?
  - A resposta é não.
- O que define se o atributo volume necessita de uma tecnologia de Big Data é **limitação das ferramentas tradicionais** para lidar com volumes de dados.

Os Vs de Big Data: Volume

Os Vs de Big Data: Variedade

O banco de dados relacional é o modelo de armazenamento de dados mais usado nos últimos 40 anos pelas corporações.

- Dados rígidos, bem estruturados.
- Tamanho e os tipos de dados bem definidos.

➢ Embora seja muito eficiente e aplicado a diversos cenários, devido às características acima, o banco de dados relacional se torna uma limitação para Big Data, uma vez que esse termo inclui dados semiestruturados e não estruturados.

Os Vs de Big Data: Variedade

Dados semiestruturados são aqueles que possuem uma estrutura pré-definida, porém não com o mesmo rigor dos dados relacionais.

- Arquivos no formato JSON (JavaScript Object Notation).
- XML (eXtensible Markup Language).

➢ **Dados não estruturados** incluem os vídeos, imagens, e alguns formatos de textos. Considerando todos os dados disponíveis globalmente, apenas 20% são considerados dados estruturados.

Os Vs de Big Data: Variedade

Formato dos Dados

Fonte dos Dados	Estruturado	Não-estruturado
Interno 	• Resultados de pesquisas • Registros de vendas • Medidas de controle de processos • Bancos de dados de sistemas internos (ERP, CRM)	• E-mails, cartas, mensagens de texto • Legendas de vídeos • Comentários de clientes • Mensagens de voz • Imagens / ilustrações • Avaliação de funcionários
Externo 	• Likes do Facebook, retweets • Horário de publicação de posts, tweets, updates • Pontuação em sites de classificação	• Conteúdo publicado em redes sociais • Comentários em foruns online • Imagens • Vídeos de câmeras de segurança

Os Vs de Big Data: Variedade

- Na área governamental, com a utilização de tecnologias para rastrear os perfis dos eleitores na campanha do presidente dos Estados Unidos, Barack Obama;
- No setor financeiro, com soluções na área de análise de risco e detecção de fraude;
- Na área de transporte e automação, com o monitoramento de tráfego e rastreamento de carga;
- No setor de varejo, com a possibilidade de gerar ofertas baseadas na análise de vendas e no perfil do consumidor;

Os Vs de Big Data: Variedade

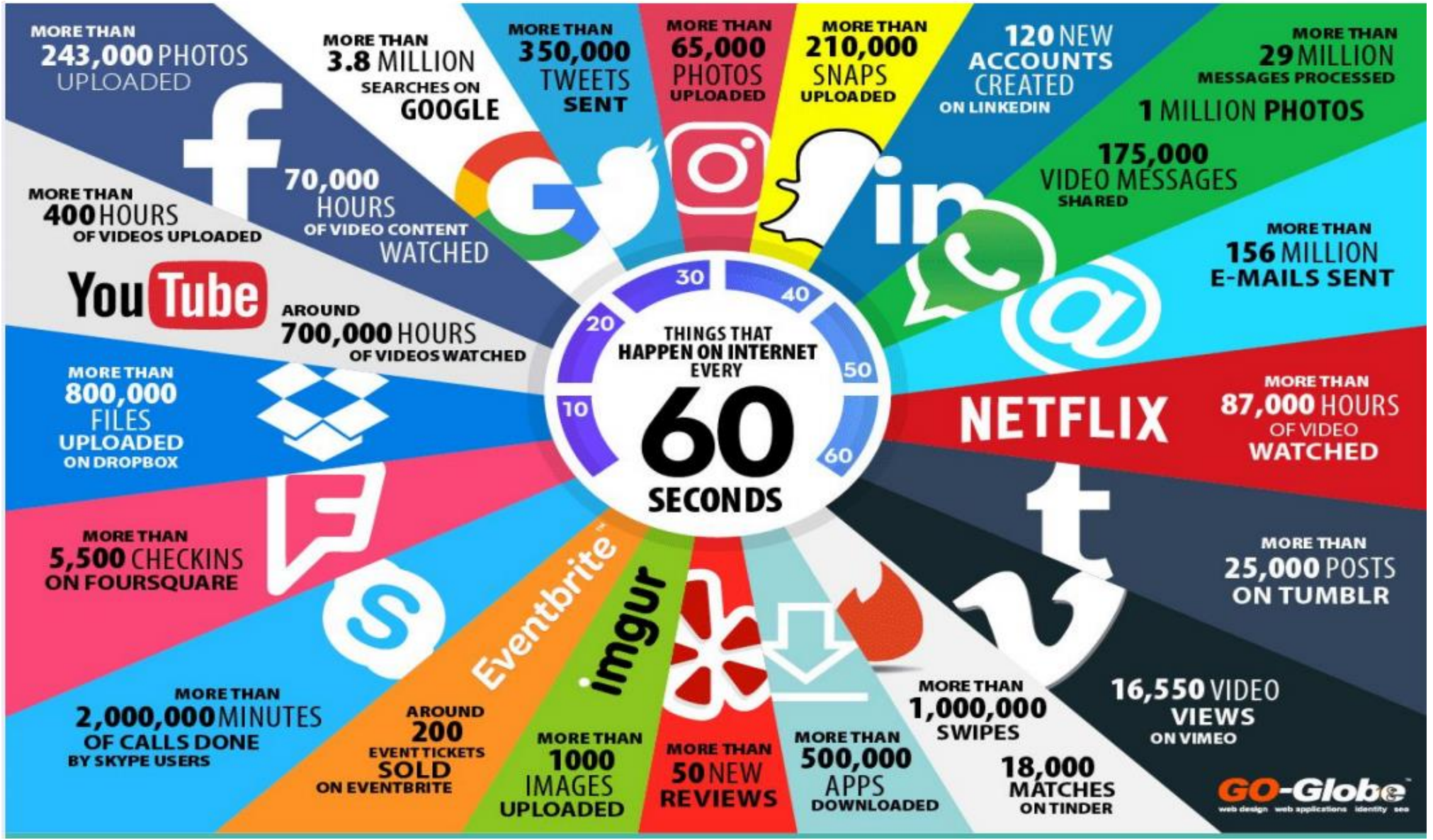
- Nas diversas possibilidades na área de marketing, por meio da análise de redes sociais;
- Na área de seguros, com a possibilidade de ofertas de planos baseados no comportamento do segurado.

Os Vs de Big Data: Velocidade

Outra característica de Big Data é a velocidade com que os dados são coletados, analisados e utilizados.

- Imagine um e-commerce que faz recomendações de produtos a um cliente depois de uma semana dele ter comprado um produto. Se fosse feito no mesmo instante teria um impacto bem maior provavelmente.
- Além da análise dos dados, outro fator de velocidade deve ser levado em consideração e a rapidez com que os dados são gerados.
- Em apenas 1 minuto são gerados:
  - 2 milhões de pesquisas no google.
  - 6 milhões de páginas são visitadas no facebook.
  - 1.3 milhão de vídeos são vistos no youtube.

Os Vs de Big Data: Velocidade



Big Data: definições importantes

Big Data não é somente um grande volume de dados armazenado. Envolve variedade e velocidade dos dados, que necessita de estratégias inovadoras capaz de extrair informações valiosas de uma massa de dados.

➢ Exige quebra de paradigmas. Novos tamanhos de dados, novas velocidades, novas tecnologias e novos métodos de análise de dados.

➢ Mudança de estratégias e tecnologias a todo momento.

Big Data: definições importantes

➢ Existem outras características além dos 3 Vs apresentados. O atributo **valor**, que consiste em quão valioso e significativo um dado é para uma solução. O atributo **veracidade**, que consiste na confiabilidade dos dados.

➢ Por ser característico de Big Data ter uma grande quantidade e variedade de dados, é comum a existência de dados inconsistentes.