

Big Data - Introdução

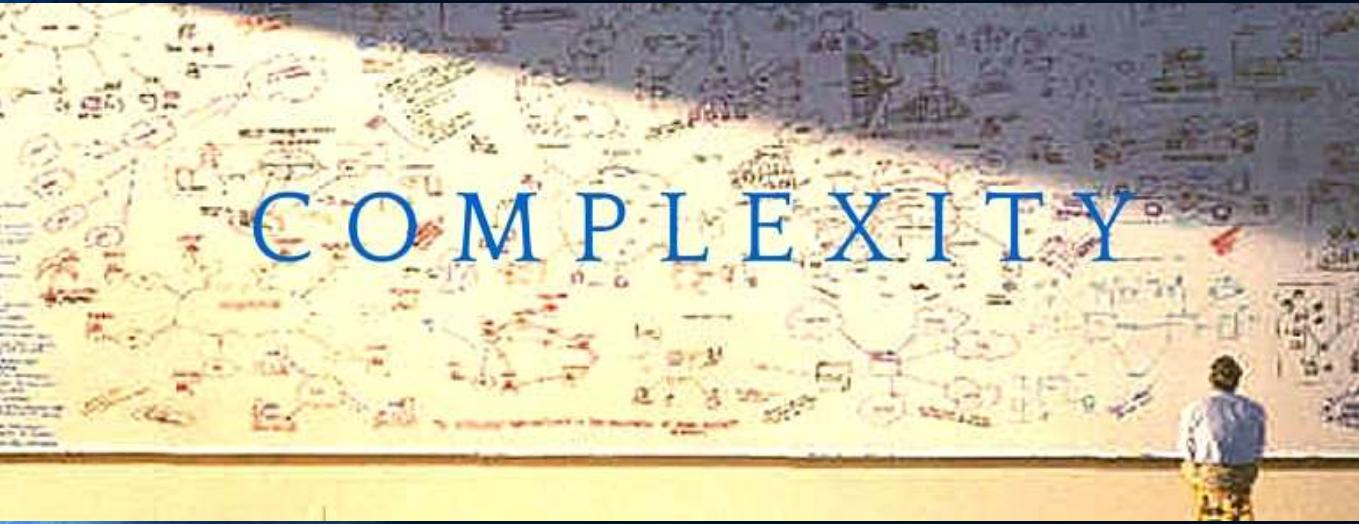
PROF. ANTONIO GUARDADO

Agenda

- 1- Conceitos
- 2 – Características do Big Data
- 3 – Big Data e Business Intelligence
- 4 – Ciência dos Dados
- 5 - Tecnologia

1 – Desafios e Motivação

- Problemas para a tomada de decisão
- Mundo dos negócios em constante mudança e alto grau de incerteza resulta em alta complexidade no processo decisório



COMPLEXITY



1 – Desafios e Motivação

Chiavenato (1997) aponta que o processo de decisão é complexo e está sujeito às características individuais do decisor quanto da circunstância em que está envolvido e da maneira como comprehende essa situação. :

1. Percepção da situação que abrange algum problema;
2. Diagnóstico e definição do problema;
3. Definição dos objetivos;
4. Busca de alternativas de solução ou de cursos de ação;
5. Escolha da alternativa mais adequada ao alcance dos objetivos;
6. Avaliação e comparação dessas alternativas;
7. Implementação da alternativa escolhida.

Modelo de Simon
Inteligência
Concepção
Seleção
Implementação

1 – Desafios e Motivação

- Quatro Estágios de Simon
- **Inteligência:** Consiste em descobrir, identificar e entender os problemas que estão ocorrendo na organização. Por que existe um problema? Onde ele está e qual seu efeito?
- **Concepção:** envolve a identificação e investigação das várias soluções possíveis para o problema.
- **Seleção:** Consiste em escolher uma das alternativas de solução.
- **Implementação:** da solução envolve fazer a alternativa escolhida funcionar e continuar a monitorar em que medida ela está funcionando.

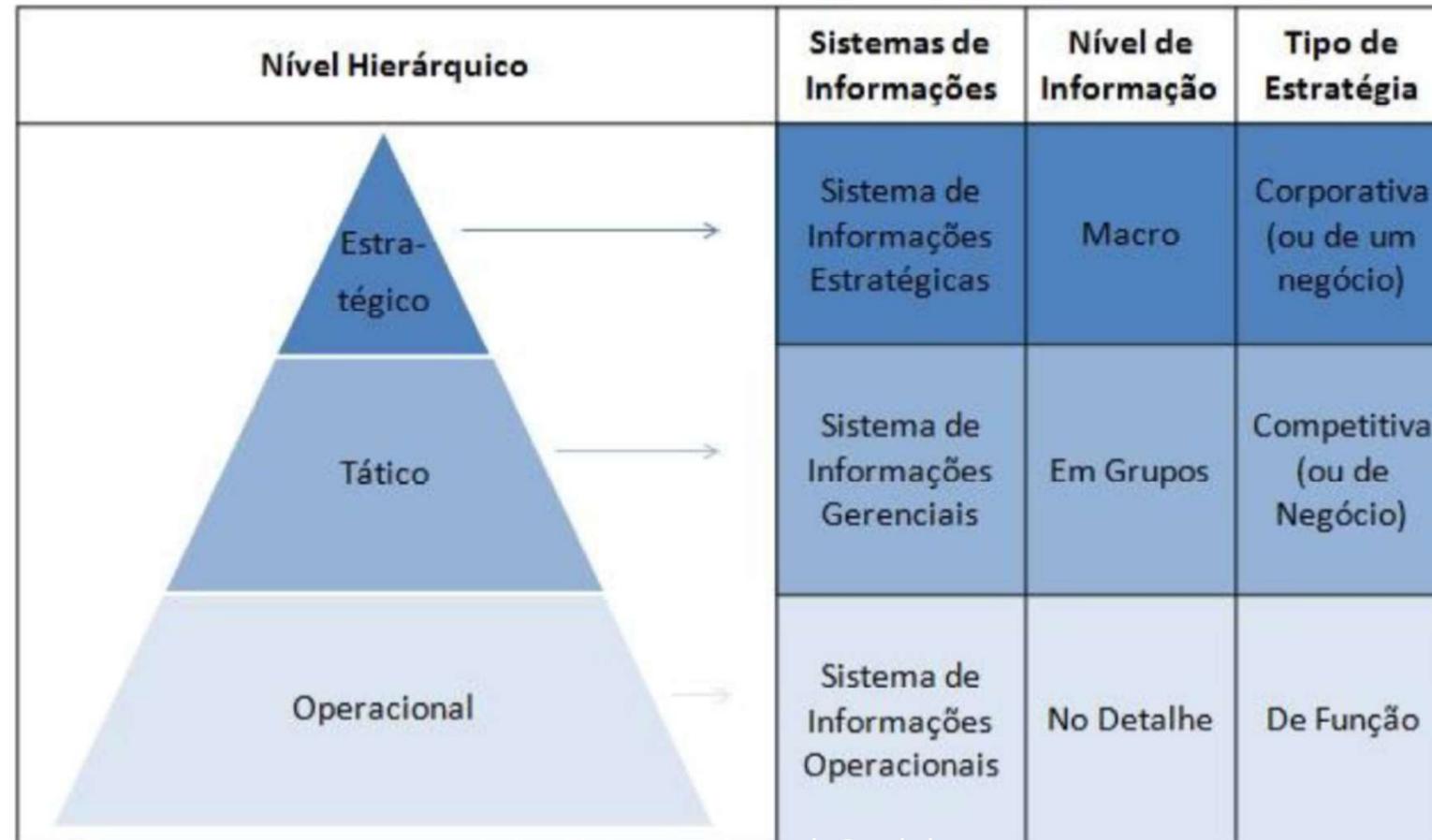
1 – Desafios e Motivação

- **Informação é a matéria-prima para todos os estágios deste processo, e o gestor ou tomador de decisão precisa coletar, selecionar e interpretar as informações, a fim de incluí-las em uma situação problema ou cenário visando a apoiar a tomada de decisão.**

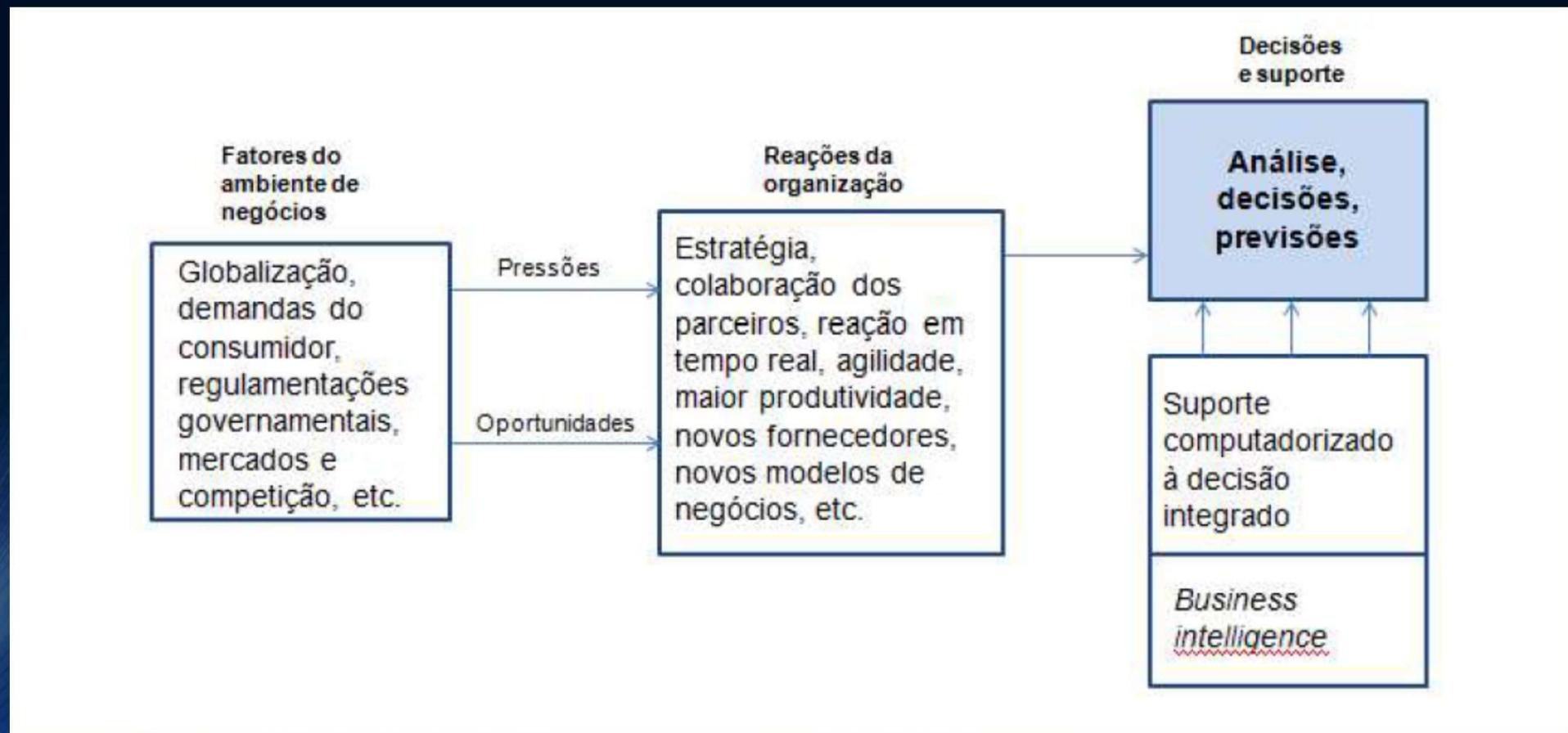
1 – Desafios e Motivação

Figura 1: Níveis hierárquicos e de informação

Fonte: DEL DUCA et. al., 2012



1 – Desafios e Motivação



1 – Desafios e Motivação

- O valor atribuído pelos gestores às informações depende dos resultados alcançados pela empresa. Os benefícios oferecidos pelas decisões acertadas, baseadas em informações valiosas representam o sucesso da empresa. O conceito de valor da informação segundo PADOVEZE está relacionado com:
 - a. A redução da incerteza no processo de tomada de decisão.
 - b. A relação do benefício gerado pela informação versus custo de produzi-la. (custo x benefício)
 - c. Aumento da qualidade da decisão.

1.2 – Conceitos - Dado

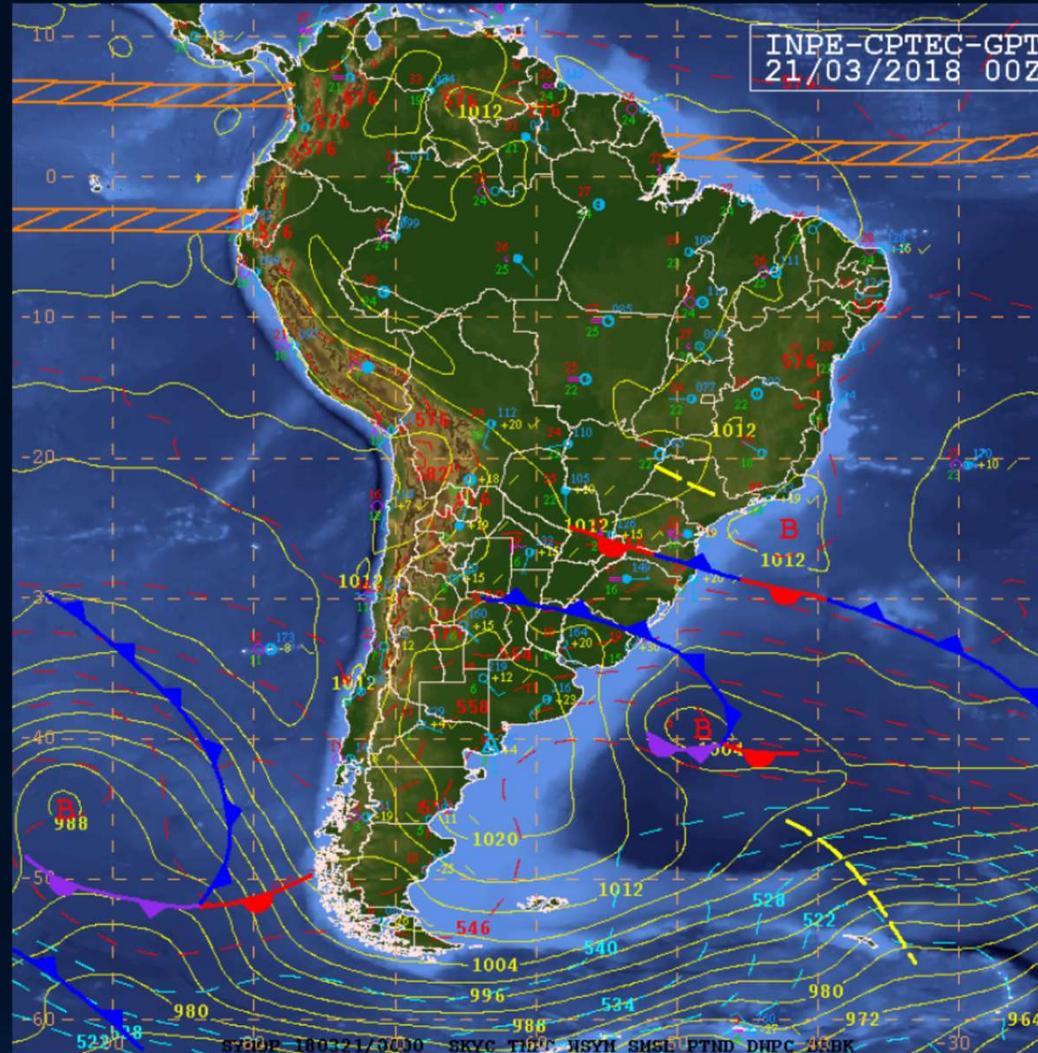
- **Dado** : qualquer elemento identificado em sua forma bruta que, por si só, não conduz a uma compreensão de determinado fato ou situação.
- No Banco de Dados é o valor de alguma característica de um objeto
 - Exemplo : Característica **Nome de Pessoa**
 - Valor : **José de Souza**
- Representados por Símbolos, Números, Marcas -> **Natureza quantificável**

1.2 – Conceitos - Informação

- Informação são dados tratados, interpretados conforme um contexto.
- O resultado do processamento de dados são as informações.
- As informações tem significado, podem contribuir no processo de tomadas decisões.
- Um conjunto de dados somente se tornará informação no momento em que for atribuído algum significado por alguém.
- Exemplo : 29 °C em São Paulo às 02:00 -> muito quente
- Natureza Qualitativa

1.2 – Conceitos – Dado x Informação

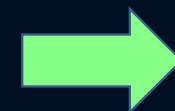
D
A
D
O
S



1.2 – Conceitos – Dado x Informação

Previsão do tempo

D Na análise da carta sinótica de superfície das 00Z do dia 21/03, nota-se um sistema frontal, com ramo A estacionário do Paraguai e SC, prosseguindo como frio sobre o Oceano Atlântico adjacente, a partir de 40°W , até um D centro de baixa pressão em oclusão no valor de 952 hPa, O posicionado em aproximadamente $60^{\circ}\text{S}/17^{\circ}\text{W}$. Outra frente S fria é observada entre o nordeste da Argentina e o norte do Uruguai se prolongando pelo Oceano Atlântico adjacente e tem associada uma baixa pressão em oclusão de 1004 hPa em $40^{\circ}\text{S}/48^{\circ}\text{W}$. Uma alta pressão pós-frontal tem valor de 1020 hPa no sul da Província de Buenos Aires na Argentina. Uma onda frontal atua no Pacífico à oeste de 80°W próximo a costa chilena. A Alta Subtropical do Atlântico Sul (ASAS) tem valor de 1024 hPa à leste de 10°W , fora do domínio da figura. A Alta Subtropical do Pacífico Sul (ASPS) tem valor de 1020 hPa em torno de $30^{\circ}\text{S}/113^{\circ}\text{W}$ (fora do domínio da figura). A Zona de Convergência Intertropical (ZCIT) atua sobre o Oceano Pacífico com banda dupla, uma em torno de $05^{\circ}\text{N}/08^{\circ}\text{N}$ e a outra em torno de $02^{\circ}\text{S}/04^{\circ}\text{S}$. No Oceano Atlântico, a ZCIT atua entre $01^{\circ}\text{N}/04^{\circ}\text{N}$.



INFORMAÇÃO



1.2 – Conceitos - Conhecimento

- Conhecimento é o conjunto de ferramentas conceituais e categorias usadas pelos seres humanos para criar, colecionar, armazenar e compartilhar a informação (Laudon & Laudon) -> **Informação trabalhada**
- Conhecimento é o ato ou efeito de abstrair idéia ou noção de alguma coisa, como por exemplo: conhecimento das leis; conhecimento de um fato (obter informação); conhecimento de um documento; conhecimento da estrutura e função de determinados sistemas.

1.3 – Conceitos – Tipos de Dados

- Dados Relacionais (Tabelas / Transação / Legacy Data)
- Texto de Dados (Web)
- Semi-estruturados de dados (XML)
- Gráfico de Dados de Redes Sociais, semanticweb (RDF)
- Data Streaming -Você só pode digitalizar os dados uma vez

1.3 - Tipos de dados

- Estruturados
- Semi-estruturados
- Não estruturados

1.3.1 - Dados Estruturados

- Dados que possuem esquema de campos fixos
- Formato bem definido
- Normalmente armazenado em BD Relacionais
- Conhecimento prévio da estrutura dos dados
- São gerados em uma ordem especificada

1.3.2 - Dados Semi-estruturados

- Possuem um fluxo lógico
- O formato pode ser bem definido, mas não necessariamente é fixo
- Não possui fácil compreensão por parte do usuário leigo
- Tem como característica marcante o uso de etiquetas e marcadores para separar elementos dos dados
- Regras complexas para manipulação dos dados
- Ex: XML , mensagens de e-mail

1.3.3 -Dados Não Estruturados

- Sem tipo predefinido;
- Não possuem estrutura uniforme (ex. Documentos, objetos);
- Pouco ou nenhum controle sobre eles;
- Dificuldade de “manipulação” para extração de informação
- Ex.: Vídeo, Áudio, Texto

2- Características

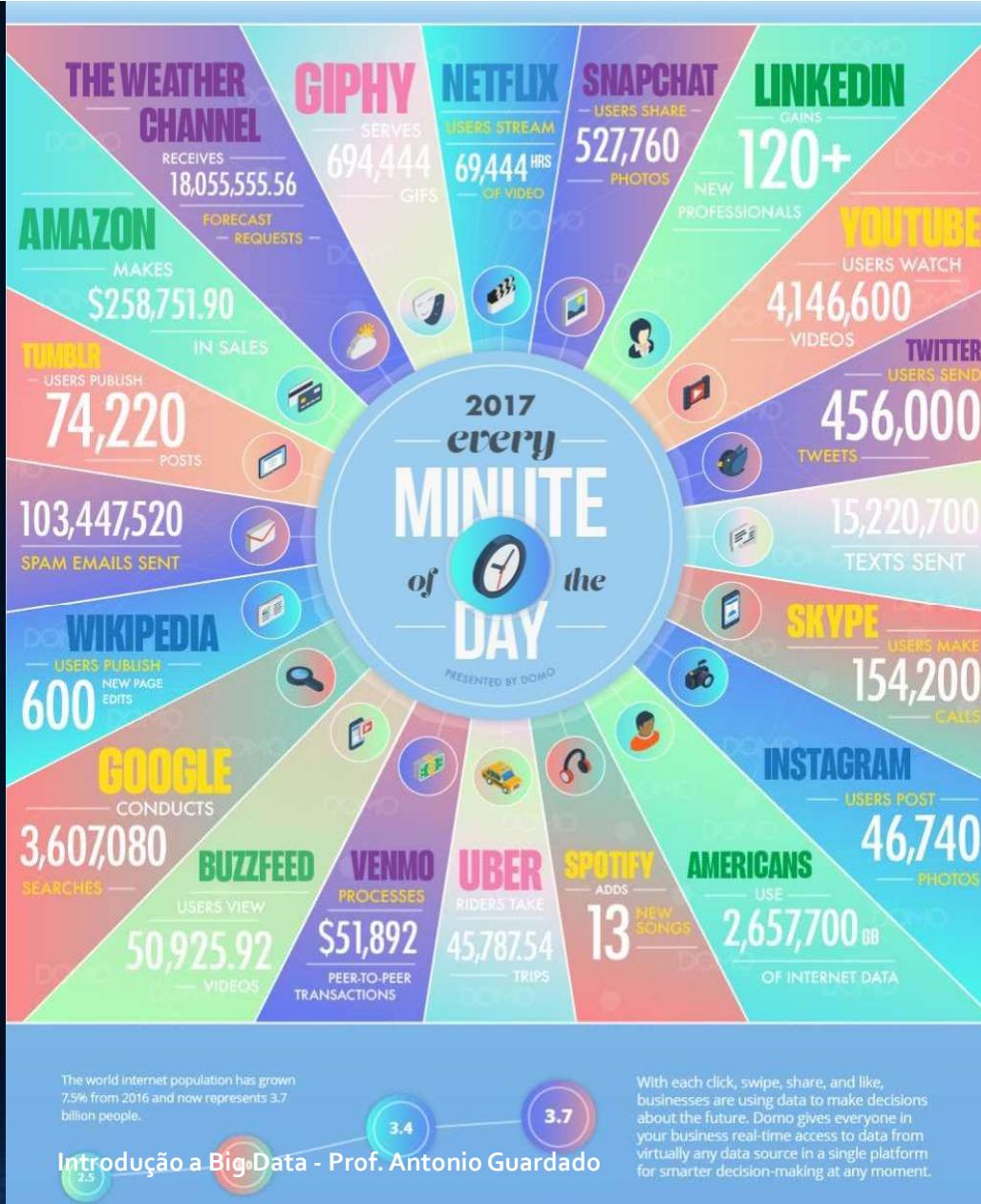
- 5Vs



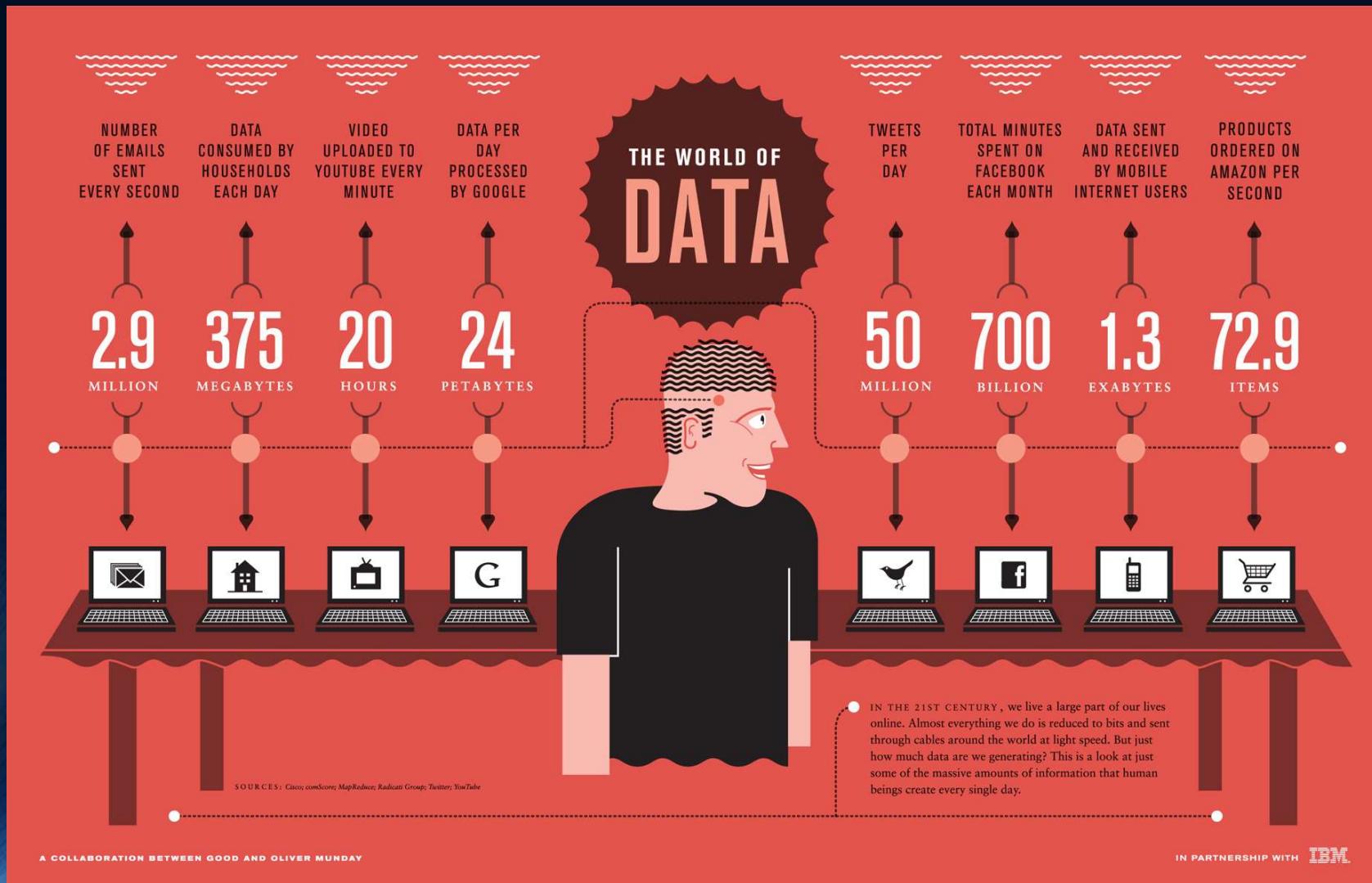
2.1- Características - Volume

- Organizações coletam dados de uma grande variedade de fontes :
 - transações comerciais
 - redes sociais
 - informações de sensores ou dados transmitidos de máquina a máquina.
- No passado, armazenar tamanha quantidade de informações teria sido um problema – mas novas tecnologias têm aliviado a carga.
- 2.5 quintilhões/dia

2.1- Características - Volume



2.1- Características - Volume



2.2 -Características - Variedade

- Os dados são gerados em todos os tipos de formatos:
 - dados estruturados
 - dados numéricos em bancos de dados tradicionais
 - documentos de texto não estruturados
 - e-mail
 - Vídeo
 - Áudio
 - dados de cotações da bolsa
 - transações financeiras.

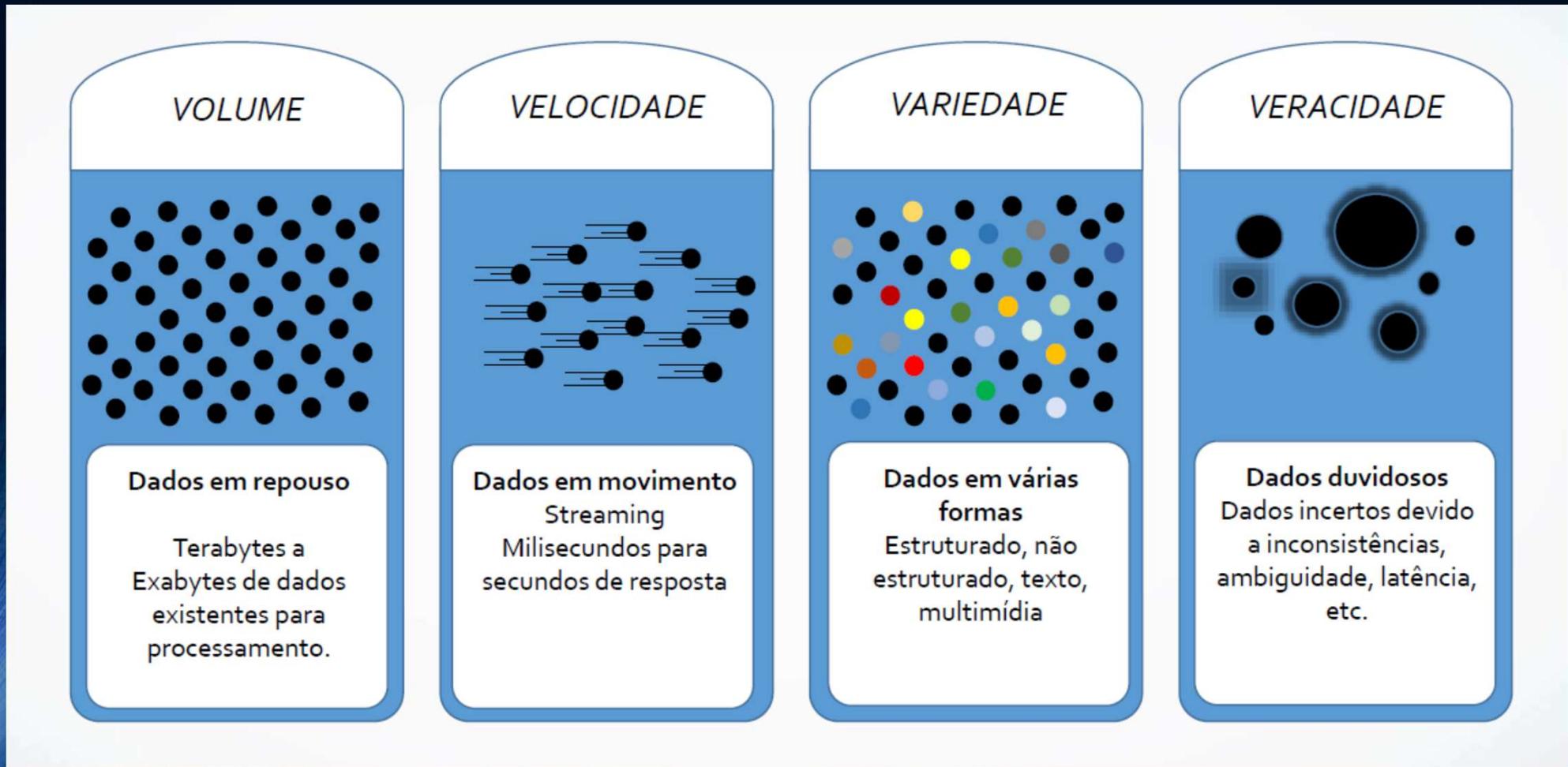
2.3 -Características-Velocidade

- Os dados fluem em uma velocidade sem precedentes e devem ser tratados em tempo hábil.
- Tags de RFID, sensores, celulares e contadores inteligentes estão impulsionando a necessidade de lidar com imensas quantidades de dados em tempo real, ou quase real.

2.4 -Características - Veracidade

- De nada adianta ter um volume massivo de dados, de fontes diferentes, em formatos diferentes se não é possível definir se estes são dados confiáveis. A qualidade do dado capturado e analisado é fundamental para o bom trabalho do Big Data.
- A qualificação da fonte, a determinação de padrões, a confiabilidade do processo de captura e também, o processo de cruzamento de novos dados com outros já existentes e sobretudo a compreensão dos dados capturados ajudam a definir a sua veracidade e consequentemente o nível de confiabilidade da informação gerada.

2.5 – Mais um V



2.5 -Características – Valor

- De nada adianta um grande volume de dados, velocidade no processamento, fontes diferentes e dados verificados se estes não possuem, **agregam valor** ou justificam o esforço do processo de consegui-los
- Se isso não reduzir custos, **melhorar os processos decisórios** ou incrementar a qualidade de serviços e produtos. Além disso, o custo da extração das informações não deve superar o valor que de fato a informação tem enquanto patrimônio da organização.

2.7 -Características – O que está gerando

Redes de Mídias Sociais (todos nós geramos)



Instrumentos Científicos
(coletando todos os tipos de dados)



Dispositivos móveis (rastreando todos os objetos o tempo todo)



Redes de Sensores de Tecnologia (medindo todos os tipos de dados)



2.7.1 – Principais Atores na Geração de dados

2.8 - De onde vem os dados do Big Data?

- Web e Redes Sociais (clicks, cookies, twitter, facebook)
- Mobilidade
- Internet das Coisas (RFID/NFC, Sensores, GPS e Telemetria)
- Biometria (Reconhecimento fácil, impressão digital, dados genéticos)
- Dados gerados por pessoas (Voz, email, SMS, etc)
- Dados gerados por governos, institutos de pesquisas e empresas

2.8.1 - Origem dos dados - WEB

- Maior fonte de Big Data utilizada na atualidade;
- Facilidade para mapear comportamento e fazer predição
- Possui conhecimento importante para tomada de decisão pelas empresas
- Gera informação objetiva e de impacto, que é difícil de se obter sem uma comunicação direta
- Possibilidade de captura de diversos tipos de eventos (Compras, visualização de produtos e vídeos, buscas etc)

2.8.2 -Origem dos dados - Texto

- Tipo mais comum e “simples” de dados
- Origina-se praticamente em todas as fontes de dados do Big Data
- Pode ser tratado como um tipo de dado “Estruturado”
- Estruturado + Muitas fontes = DIFICULDADE
- Possui ferramentas e aparato científico bem estruturado para análise
 - Processamento de linguagem natural
 - Análise sintática
 - Mineração de texto

2.8.3 - Dados de Sensores

- Peças chave da Internet das Coisas (IOT)
- Monitoramento Autônomo e Ubíquo
- Complexidade de manipulação dos sensores
- Captura muito influenciada por fatores externos (Ex. Delay)
- Dados normalmente estruturados, mas já há redes de sensores com dados não estruturados

2.8.4 - Dados de Geolocalização

- Localização e Tempo são dois atributos de grande VALOR (Ex. Google Location History)
- Possibilidades diversas para desenvolvimento de aplicações
- Muito sensível para o Big Data em Volume e Velocidade
- Binômio crítico com relação a questão de privacidade

2.8.5 -Dados de RFID e NFC

- Sofrem também efeito da privacidade
- NFC foi criada para comunicação entre objetos próximos e com pouca transmissão de dados
- Inclusão de NFC em celulares mudou a perspectiva do tráfego de dados (Ex. Pagamentos, controle de acesso)

2.8.6 -Dados de Redes Sociais

- Tão complexo que criou um novo ramo na análise de dados:
Análise Social
- Volume de dados para análise de um único indivíduo na rede
- Amplitude gerando complexidade: (Ex: Eu -> Meus amigos ->
Amigos dos meus Amigos)
- Dados crescendo indefinidamente e de forma heterogênea

2.9 –Definição de Big Data

- **Big data** é um enorme volume de **dados estruturados e não estruturados**. O volume é tão grande que é impossível processar com técnicas de banco de dados e software tradicionais.
- **Big data** é o termo utilizado para uma coleção de conjuntos de dados **tão grande e complexo** que se torna impossível processar usando ferramentas de gerenciamento de banco de dados ou aplicações de processamento de dados tradicionais.
- **Big data** são dados cuja **escala, diversidade e complexidade exigem novas arquiteturas, técnicas, algoritmos e análises** para gerenciá-los e extrair valor e conhecimento oculto deles.

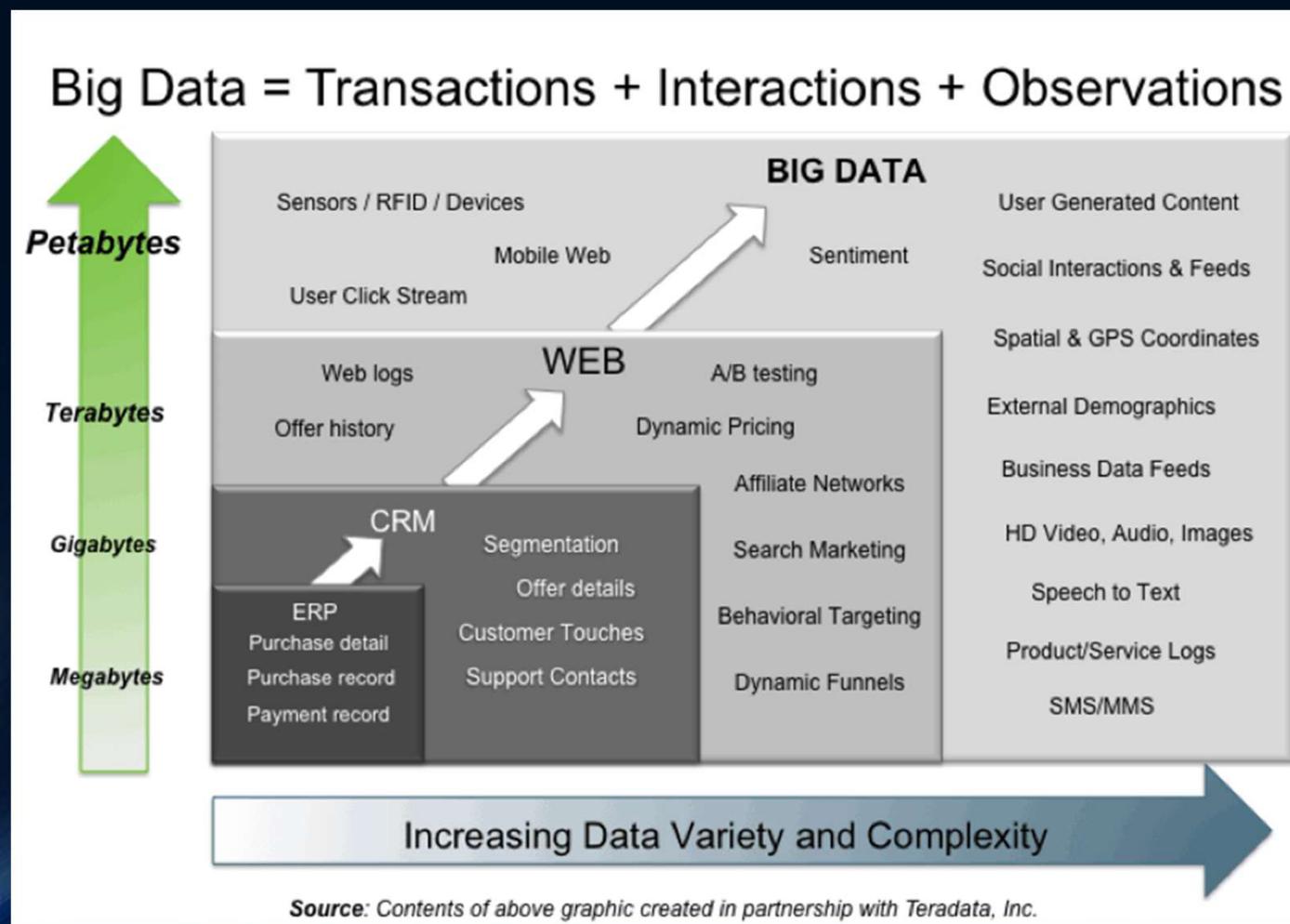
2.9.1 – Dados que compõe Big Data

- Big Data é somente dados não-estruturados (texto, imagem, etc.) ?



- Big Data = **Dados Estruturados** (BDs Relacionais principalmente) +
- **Dados não estruturados** (crescimento é exponencial !!!)

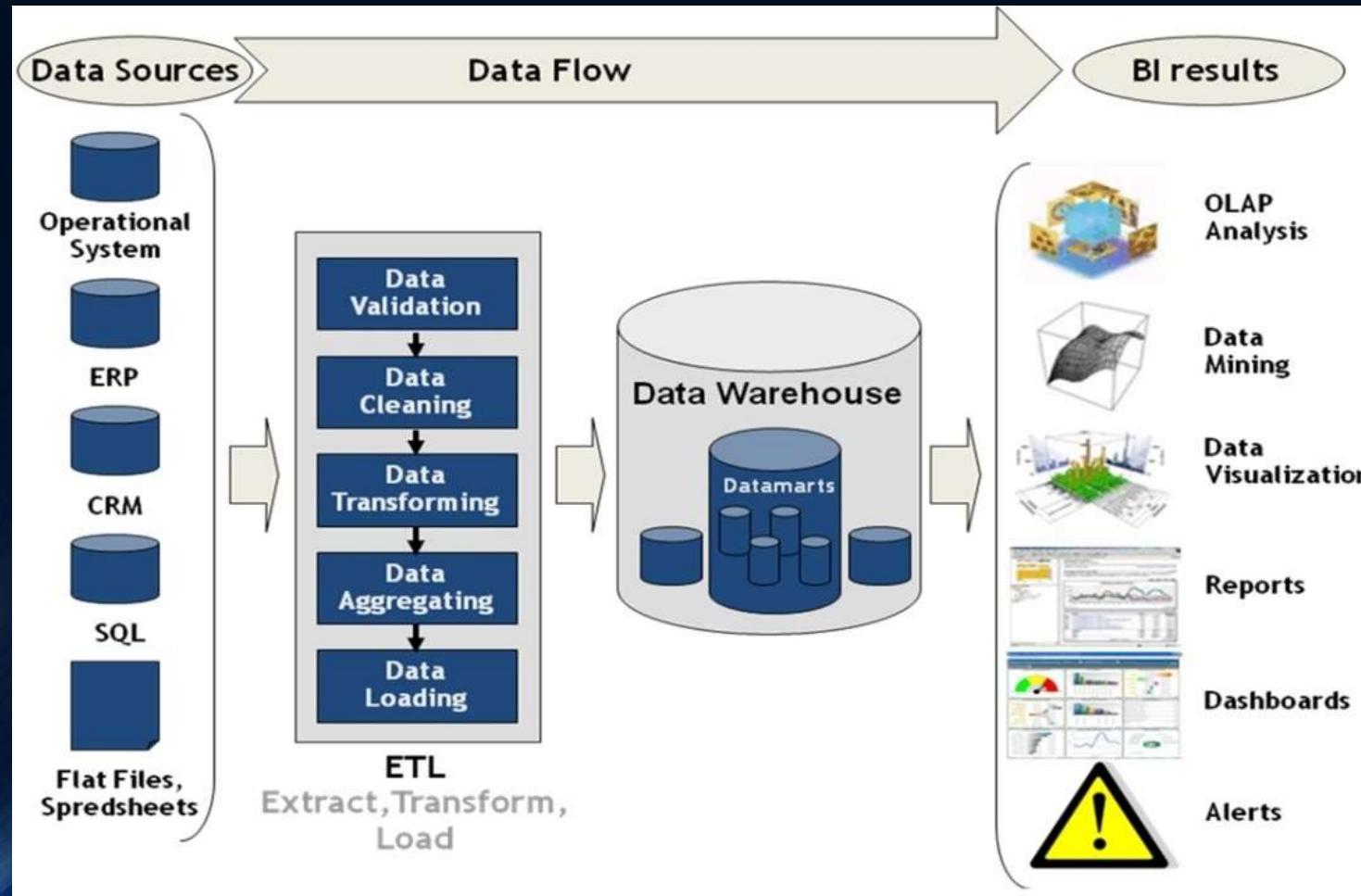
3 – Big Data e Business Intelligence



3.1 – Business Intelligence – Inteligência nos negócios

- Origem na década de 90
- - Focado na coleta, transformação e disponibilização de dados estruturados para a tomada de decisões;
- - Analisa o que já existe, definindo as melhores hipóteses;
- - Ideal para quando já se conhece as perguntas
- - Mais específico, voltado apenas para negócio
- Principais Componentes : DataMart e DataWarehouse

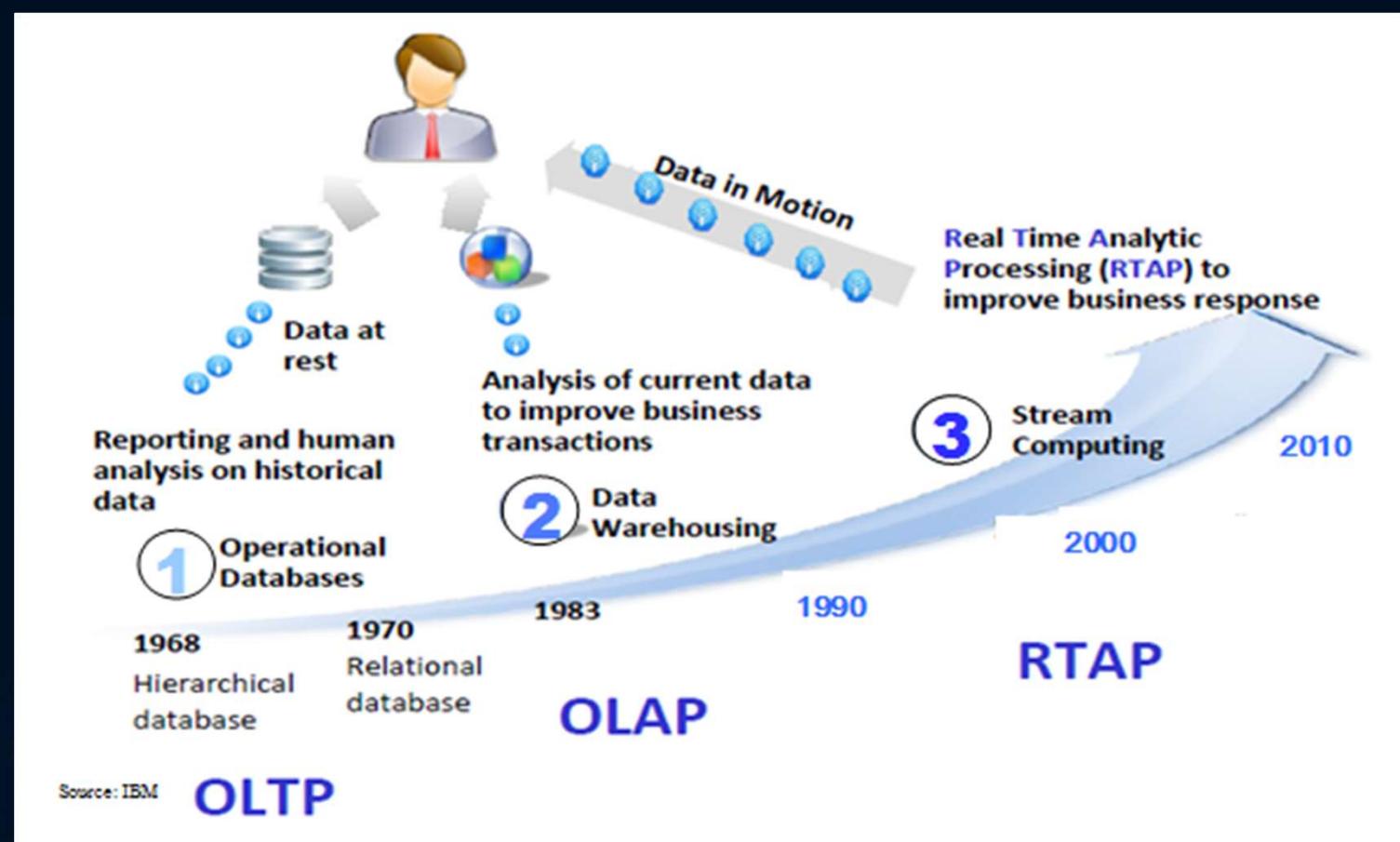
3.1.1 – BI Arquitetura



3.2- Big Data

- **Big Data (Analytics)**
- Focado no processamento de **dados estruturados e não estruturados**, bem como nas correlações e descobertas que desse processamento podem advir;
- Analisa **o que já existe e o que está por vir**, apontando novos caminhos;
- Ideal para quando se quer explorar novas possibilidades, descobrir novos padrões e explorar perguntas que ainda não haviam sido feitas;
- Mais amplo, voltado não apenas para negócios, mas para qualquer área/segmento, como saúde, entretenimento, educação.

3.3 – Evolução BI -> Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

3.3 – Evolução : O modelo mudou ...

- O modelo de geração/consumo de dados mudou

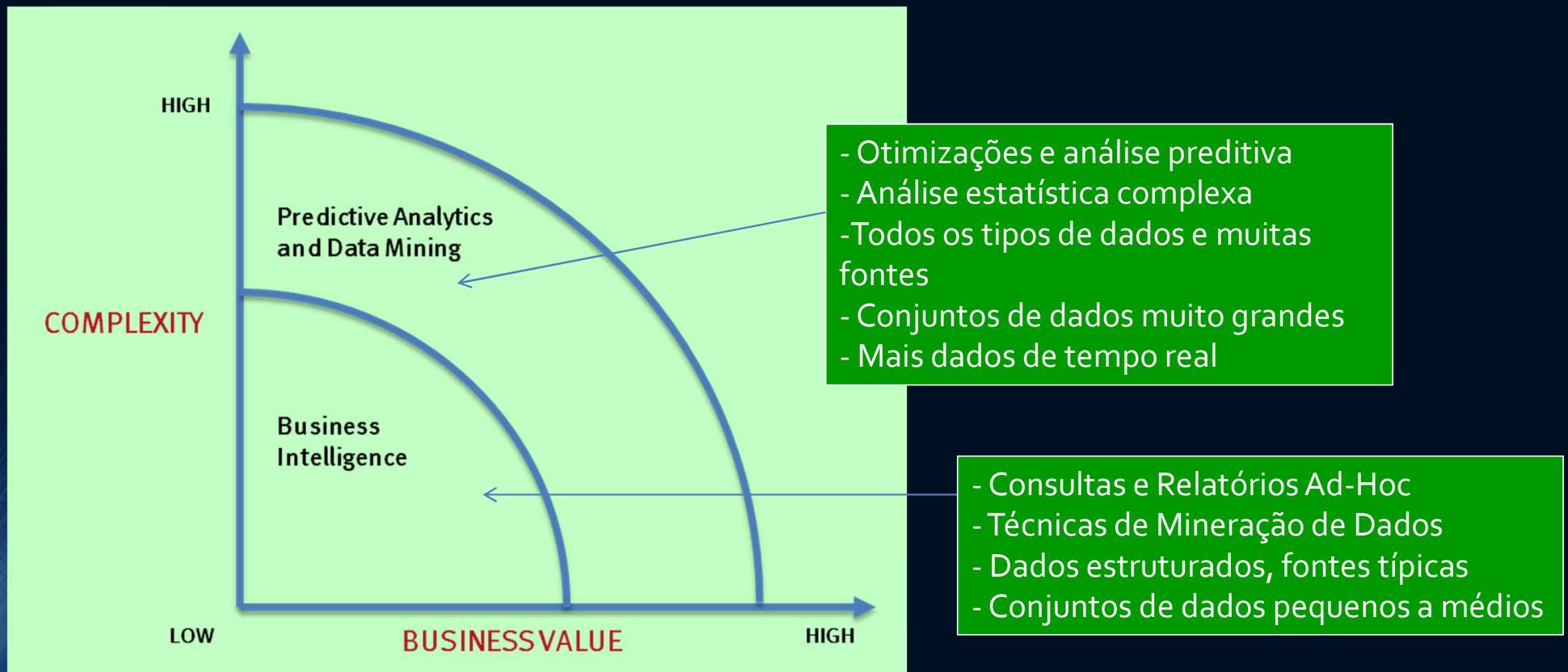
Modelo Antigo: Poucas companhias gerando dados, as outras consumindo



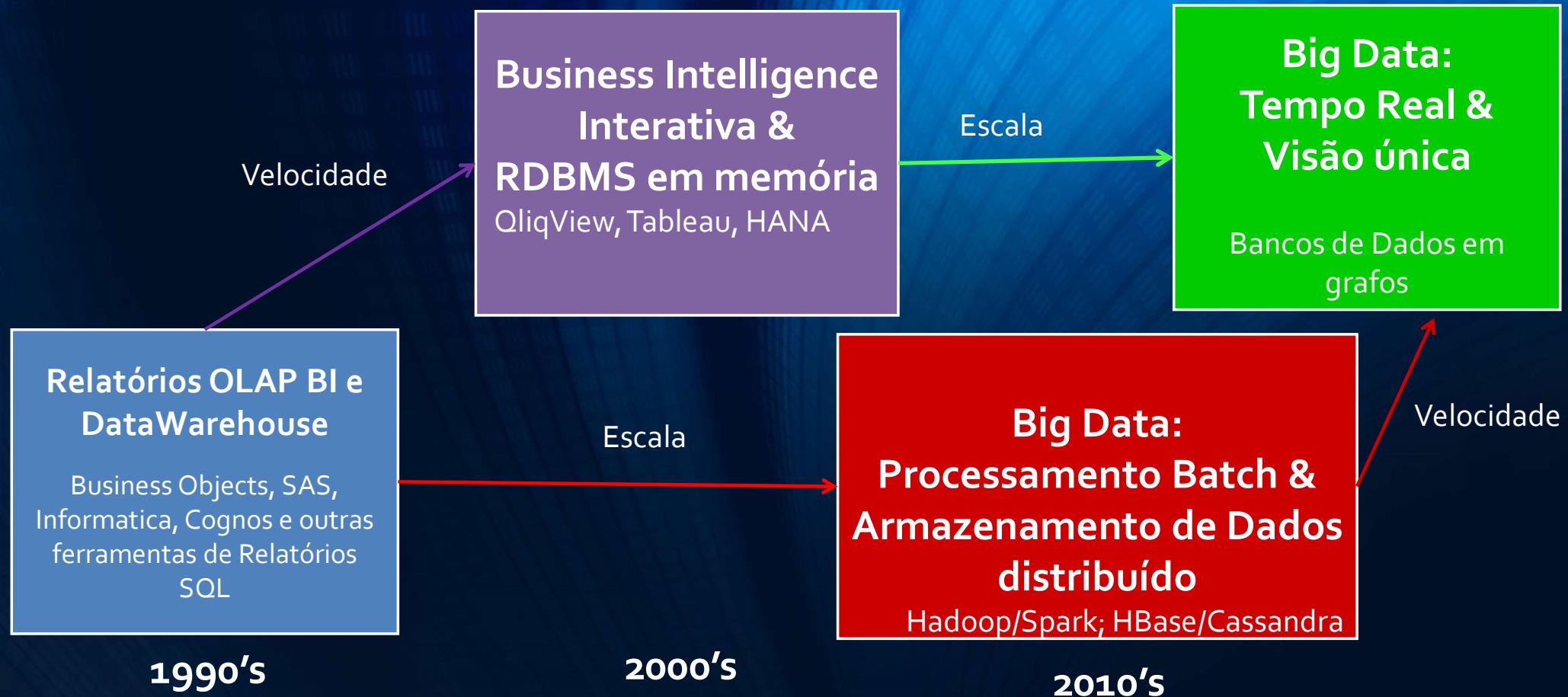
Modelo Novo: Todos nós gerando dados, todos nós consumindo



3.4 - O que está direcionando Big Data



3.5 – A evolução da Business Intelligence



3.6 - Fontes de Dados para o Big Data Analytics

- Dados extraídos de ferramentas de Inteligência de Negócios ([Business Intelligence – BI](#));
- Arquivos de log de servidores web;
- Conteúdos de mídias sociais;
- Relatórios empresariais;
- Textos de e-mails de consumidores à empresa;
- Indicadores macroeconômicos;
- Pesquisas de satisfação;
- Estatísticas de ligações celulares capturadas por sensores conectados à “[Internet das coisas](#)”;
- Bases de dados das empresas de cartão de crédito;
- [Programas de fidelidade](#);
- Reviews de produtos nos sites das empresas.

3.6.1 – Usando dados para o quê – Setor Financeiro

- Reduzir as taxas de churn
- Para se ter uma ideia do impacto da saída de clientes da base ativa do setor financeiro, nos EUA estima-se que 30% dos clientes sejam vulneráveis à migração. Sabendo desse fator crítico, muitas instituições do ramo passaram a usar a análise de dados para rastrear as manifestações emocionais dos correntistas (em mídias sociais e sites de reclamações), diagnosticando com antecedência suas insatisfações e ganhando tempo para neutralizá-las antes do fechamento de contas.

3.6.1 – Usando dados para o quê – Setor Financeiro

- Personalizar serviços
- Entender como os clientes usam cartões de crédito e para que tomam empréstimos ajuda a criar produtos que atendam assertivamente suas necessidades, ampliando o potencial de captação de novos correntistas.

3.6.1 – Usando dados para o quê – Setor Financeiro

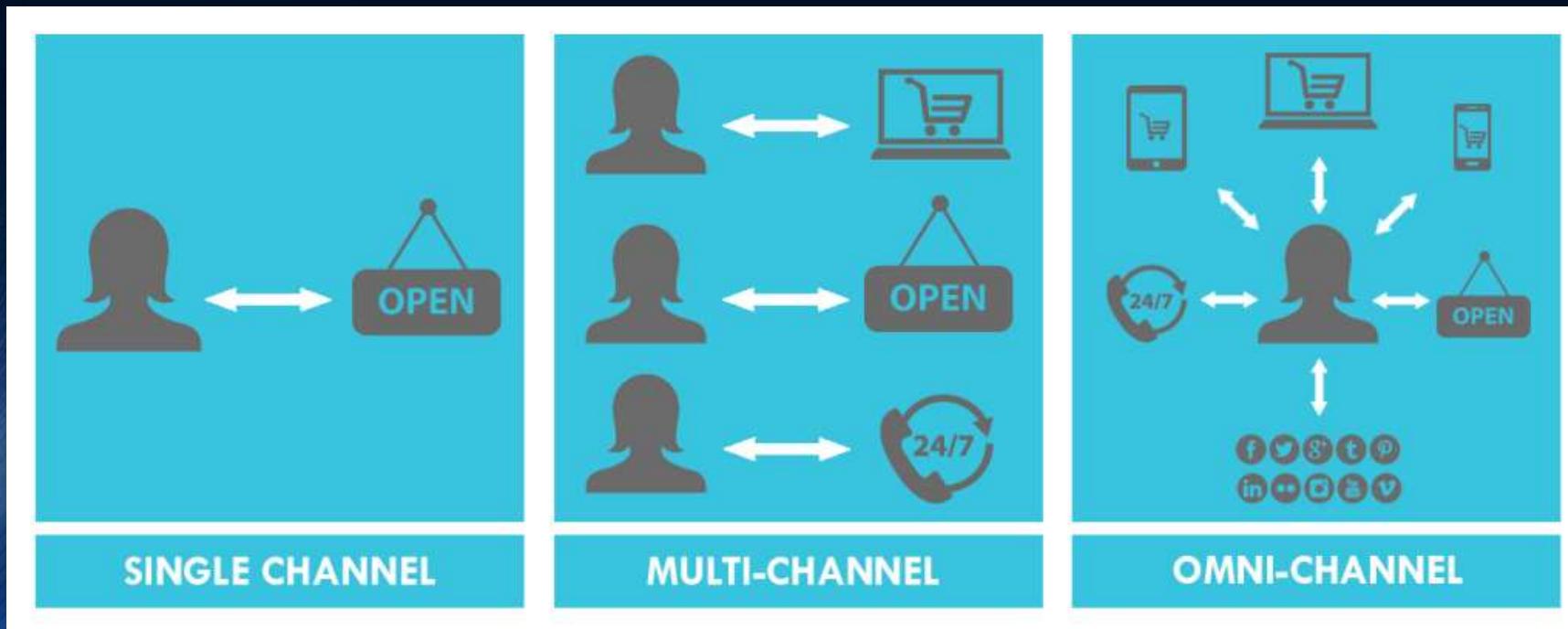
- Estreitar as relações com clientes
- Soluções em Big Data Analytics processam todas as movimentações dos correntistas (a partir do bankline, de mídias sociais, do CRM do banco, de blogs), no intuito de gerar relatórios e gráficos que revelam o valor da vida útil de cada cliente, seus anseios e expectativas em relação ao banco. Isso permite, entre outras coisas, aumentar as vendas cruzadas.

3.6.2 – Usando dados para o quê – Varejo

- Direcionar ações de marketing multicanal
- Com a disseminação de um consumidor cada vez mais omnichannel, é preciso integrar todos os canais de comunicação, o que passa pela compreensão plena do comportamento do cliente. Esta consciência de mercado pode ser alcançada por meio da maior segmentação do público-alvo, entendimento de seus hábitos e preferências de consumo e informações sociais e demográficas — tudo isso possível graças à coleta e análise dos dados de milhares de consumidores.

3.6.2- Usando dados para o quê - Varejo - OmniChannel

- É uma mistura dos mais diferentes meios de compra em uma única experiência, onde o consumidor nem percebe que está indo de canal em canal até efetivar sua compra. Resumindo, é uma união da loja física, televendas, venda de porta-em-porta, e-commerce, mobile commerce e outros.



3.6.2 – Usando dados para o quê – Varejo

- Agregar valor aos programas de fidelidade
- Entender melhor o comportamento de compra do consumidor é essencial para elaborar programas de fidelidade realmente atraentes.
- Essencial o uso de mineração de dados para revelar esse comportamento, padrões, etc.

3.6.2 – Usando dados para o quê – Varejo

- Agregar valor aos programas de fidelidade
- Case do Grupo Pão de Açúcar, que passou a utilizar, em 2015, ferramentas de análise de dados para fidelizar seus clientes. O sistema mapeia antigos consumidores que deixaram de frequentar a rede. Em seguida, realiza um levantamento eletrônico dos produtos preferidos de cada um deles.
- A descoberta desses dois fatores permite à empresa lançar cupons de descontos personalizados, oferecendo promoções especiais e distintas a cada cliente e, assim, estimulando o consumidor a retornar para a rede.
-

3.6.2 – Usando dados para o quê – Varejo

- Maximizar o ROI em marketing
- O ambiente de mercado agressivo exige aplicação de recursos com retorno “cirúrgico” aos números da empresa. Assim, cada ação de marketing deve ser acompanhada em tempo real por ferramentas de monitoramento de redes sociais. Se uma campanha não provoca o efeito esperado ou, pior, gera feedback negativo do consumidor, essa falha deve ser detectada rapidamente, a fim de que a empresa tome as medidas corretivas.
- ROI = Return on Investment – Retorno sobre Investimento
- representa a relação entre o retorno e o capital investido em um projeto

3.6.3 – Usando dados para o quê – Setor de Saúde

- Prevenção de epidemias
- Monitorar as manifestações de uma população em redes sociais — em consonância com a agregação de dados de pesquisas de campo e análises estatísticas — ajuda a visualizar antecipadamente a possibilidade de eclosão de uma epidemia, dando tempo às instituições de saúde se adequarem aos aumentos súbitos da demanda de atendimento e medicamentos.

3.6.3 – Usando dados para o quê – Setor de Saúde

- Prevenção de epidemias
- Caso do Google Flu Trends
- <https://www.youtube.com/watch?v=IEDt89eQ64o>

3.6.3 – Usando dados para o quê – Setor de Saúde

- Telemedicina e tecnologia “vestível” (wearables)
- Telemedicina é a medicina a distância, propiciada pela utilização de tecnologias de telecomunicações e análise de grandes dados no fornecimento de informações clínicas dos pacientes.
- Essas trocas de informações a distância são viabilizadas por meio de equipamentos eletrônicos de alta capacidade de processamento, como gadgets fixados no vestuário de um paciente (relógios, pulseiras ou tênis inteligentes, frutos da chamada tecnologia “vestível”).

3.6.3 – Usando dados para o quê – Setor de Saúde

Telemedicina e tecnologia “vestível” (wearables)



3.6.3 – Usando dados para o quê – Setor de Saúde

- Telemedicina e tecnologia “vestível” (wearables)
- A troca de informações permite reunir um conjunto maior de dados (Big Data na área de diagnósticos) sobre a situação clínica de um indivíduo, possibilitando um debate mais preciso na escolha do tratamento ideal a ser aplicado a um paciente e orientando com maior excelência os procedimentos ligados à promoção da saúde.
- O uso dessas tecnologias permite melhor compreensão dos profissionais de saúde sobre as patologias de seus pacientes, fornecem melhores subsídios para pesquisas e maior credibilidade aos protocolos clínicos, dentre outros benefícios.

3.6.4 – Usando dados para o quê – Setor Público

- Combater corrupção e desvio de receitas
- Desde 2007, o Ministério da Justiça vem usando sistemas de alta performance em coleta e processamento de dados, cruzando informações de milhões de contribuintes no intuito de combater a lavagem de dinheiro e outros crimes financeiros. O sucesso da iniciativa é evidenciado pelo aumento anual no montante de recursos direcionados a essa área de Inteligência.

3.6.4 – Usando dados para o quê – Setor Público

- Fortalecer a implementação de “cidades inteligentes”
- Que tal ter um sistema de monitoramento em tempo real, para que toda a população possa acompanhar o consumo de energia e as possibilidades de sobrecarga no fornecimento?
- Semáforos cuja sincronização seja alterada a depender do trânsito nas vias ?
- Zonas de maior concentração de poluição sonora e atmosférica monitoradas via sistema ?
- Já é utilizado em grandes cidades do mundo para torná-las ‘smart cities’, como Barcelona.

3.7 – Cases de Sucesso em BIG DATA



3.7.1 – BIG DATA ON BIGMAC

- McDonalds - maior rede de fast-food do planeta
- Gerenciamento de mais de 34 mil restaurantes
- Servindo mais de 69 milhões de pessoas/dia em 118 países.
- 75 hambúrgueres/segundo
- 1,7 milhão de funcionários



3.7.1 – BIG DATA ON BIGMAC



- O McDonalds coleta e combina os múltiplos dados de suas lanchonetes ao redor do mundo a fim de padronizá-los e, com isso, compreender as reações de seu público, as expectativas de cada nicho em torno de seus produtos e as alterações logísticas e de design que podem ser feitas para melhorar a cadeia de serviços.

3.7.1 – BIG DATA ON BIGMAC



- Novos sanduíches a partir de estudos de Sentiment Analysis (análises de sentimento) realizados em mídias sociais;
- Promoções em tempo real, acompanhadas de perto por cientistas de dados, que mensuraram atentamente as manifestações e reações de seu target, alterando estratégias “in real-time”;
- Logística do Drive-Thru é alterada em cada país de acordo com as reações de seus consumidores no que concerne a questões como design, tempo de espera e informações providenciadas por seus funcionários no ponto de retirada dos produtos.

3.7.2 - American Express



- A American Express conseguiu compreender que a mobilidade e os recursos digitais modificaram as expectativas de seus consumidores sobre seus serviços nos últimos anos. Seu cliente espera que a empresa o conheça profundamente, saiba dialogar com ele, entenda suas preferências e as atenda.
- Esse nível de consciência de mercado foi adquirido pela empresa por meio da implantação de um audacioso projeto de Big Data Analytics, que integra tecnologias open source, como o Hadoop, com as capacitações analíticas e operacionais da organização, ao longo de suas linhas de negócios.

3.7.2- American Express

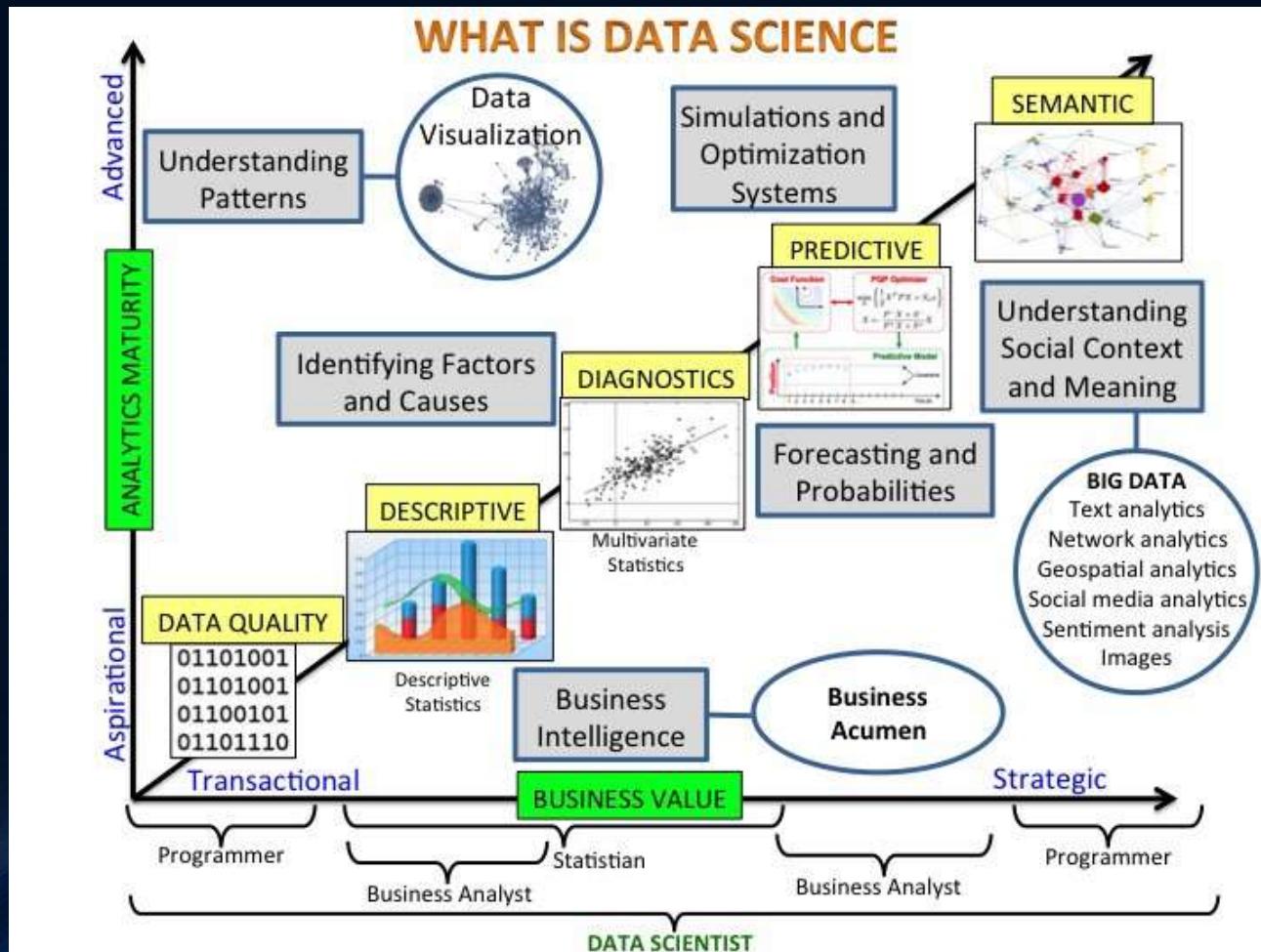


- Desenvolvimento de parcerias estratégicas e experiências “real-time” para atender seus consumidores.
- Amex Offers, que conecta seus membros a promoções personalizadas
- Parceria inovadora com o site de viagens Trip Advisor, cujo objetivo é a concessão de benefícios exclusivos e em tempo real aos clientes da operadora de cartões.

4 – Ciência dos Dados

- Área interdisciplinar voltada para o estudo e a análise de dados, estruturados ou não, que visa a extração de conhecimento ou *insights* para possíveis tomadas de decisão, de maneira similar à mineração de dados.
- Ciência de dados alia big data e machine learning, além de técnicas de outras áreas interdisciplinares como estatística, economia, engenharia e outros subcampos da computação como: banco de dados e análise de agrupamentos (*cluster analysis*).

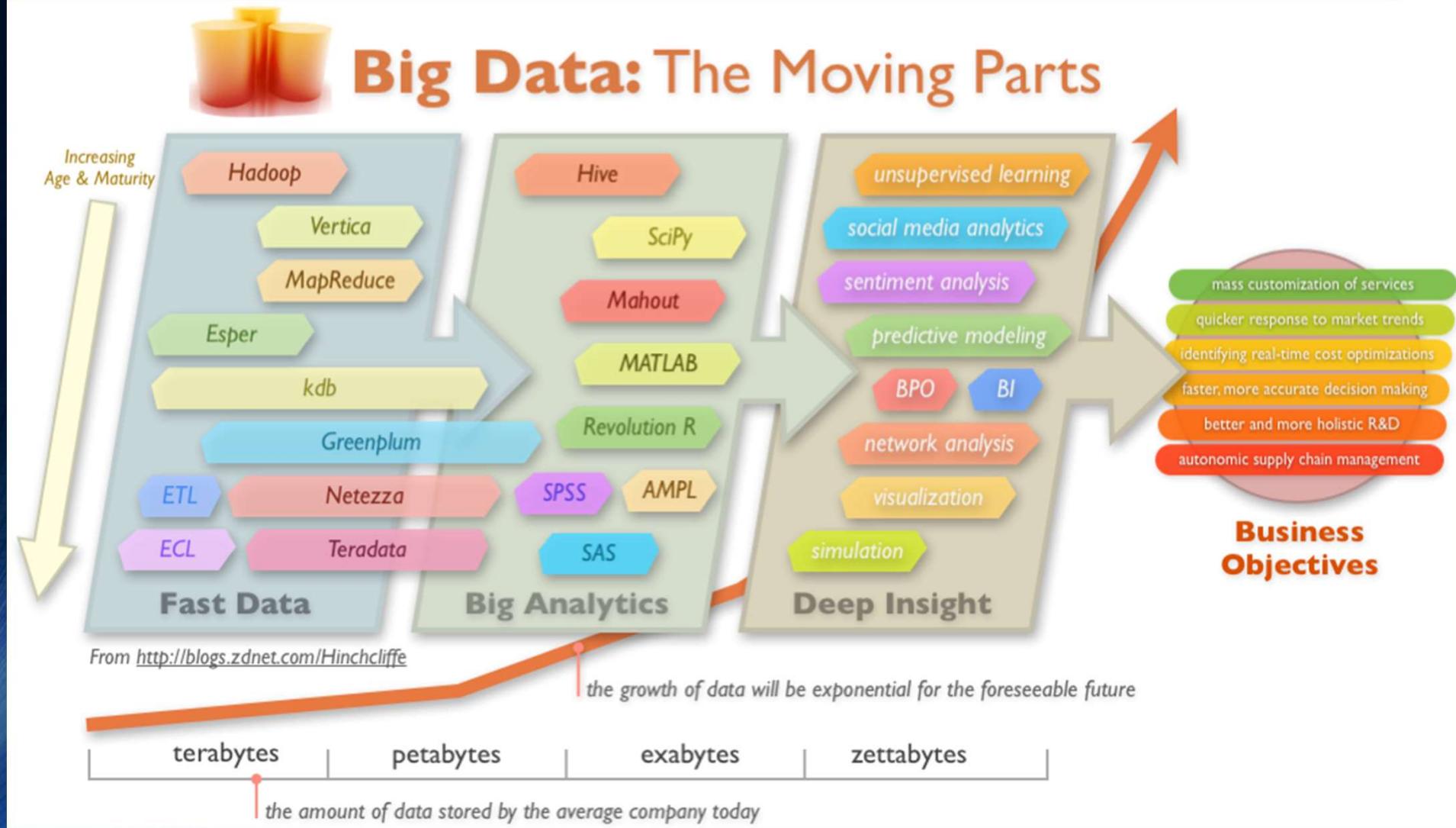
4 – Ciência dos Dados



4 – Ciência dos Dados

Área	Analista de BI	Cientista de Dados
Foco	Relatórios, KPI's, Tendências	Padrões, Correlações, Modelos Preditivos
Processo	Estático, Comparativo	Exploratório, Experimental, Visual
Fontes de Dados	Data Warehouses, Bancos Transacionais	Big Data, Dados Não-Estruturados, Bancos Transacionais e NoSQL, Dados Gerados em Tempo Real
Qualidade dos Dados na Fonte	Alta	Baixa ou Média (requer processo de limpeza e transformação)
Modelo de Dados	Esquema de dados bem definido na fonte	Esquema de dados definido no momento da consulta
Transformações nos Dados	Pouca ou nenhuma (dados já organizados na fonte)	Transformação sob demanda, necessidade de complementar os dados
Análise	Descritiva, Retrospectiva	Preditiva, Prescritiva
Responde à pergunta:	O que aconteceu?	O que pode acontecer?

5 – Tecnologia - Implantação



5 – Tecnologia Comparação

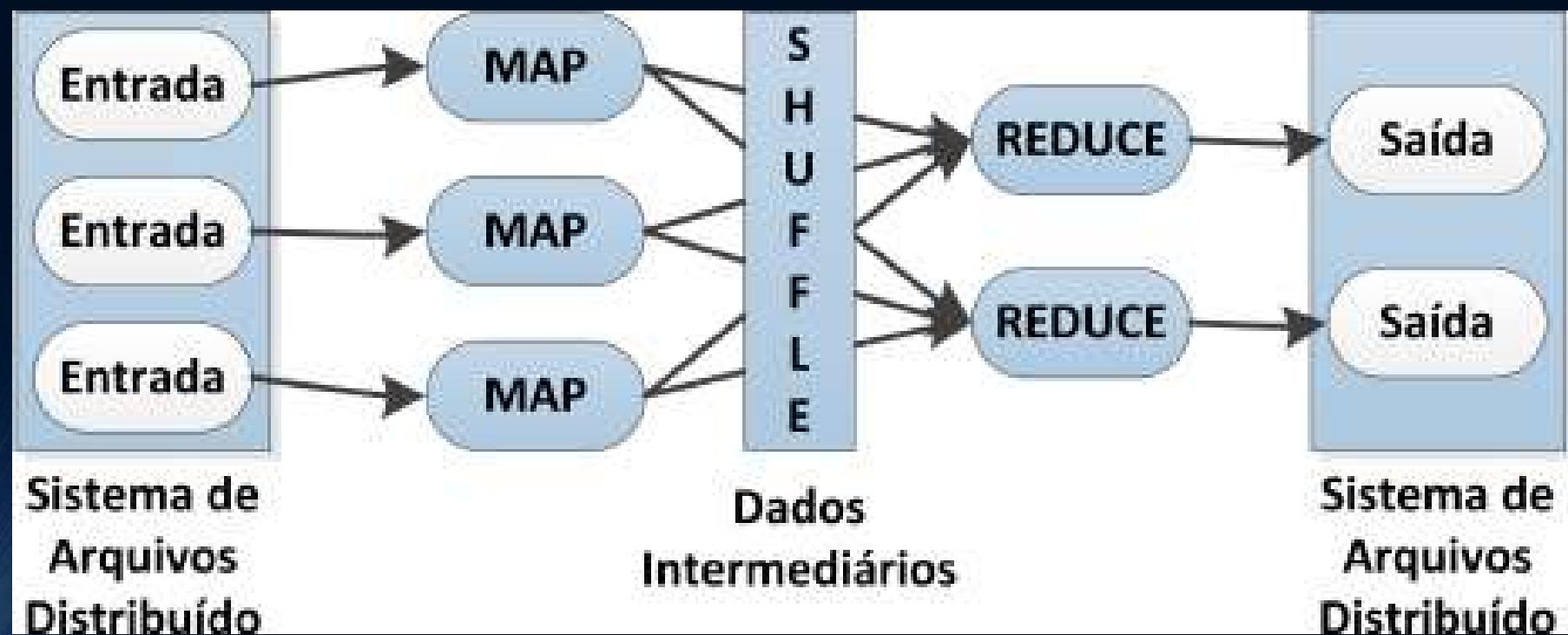
5 – Tecnologia – Map Reduce

- Modelo de programação que permite o processamento de dados massivos em um algoritmo paralelo e distribuído, geralmente em um cluster de computadores.
- A distribuição dos dados é feita no formato **chave-valor**, onde a *chave* é o identificador do registro e *valor* é o seu conteúdo.

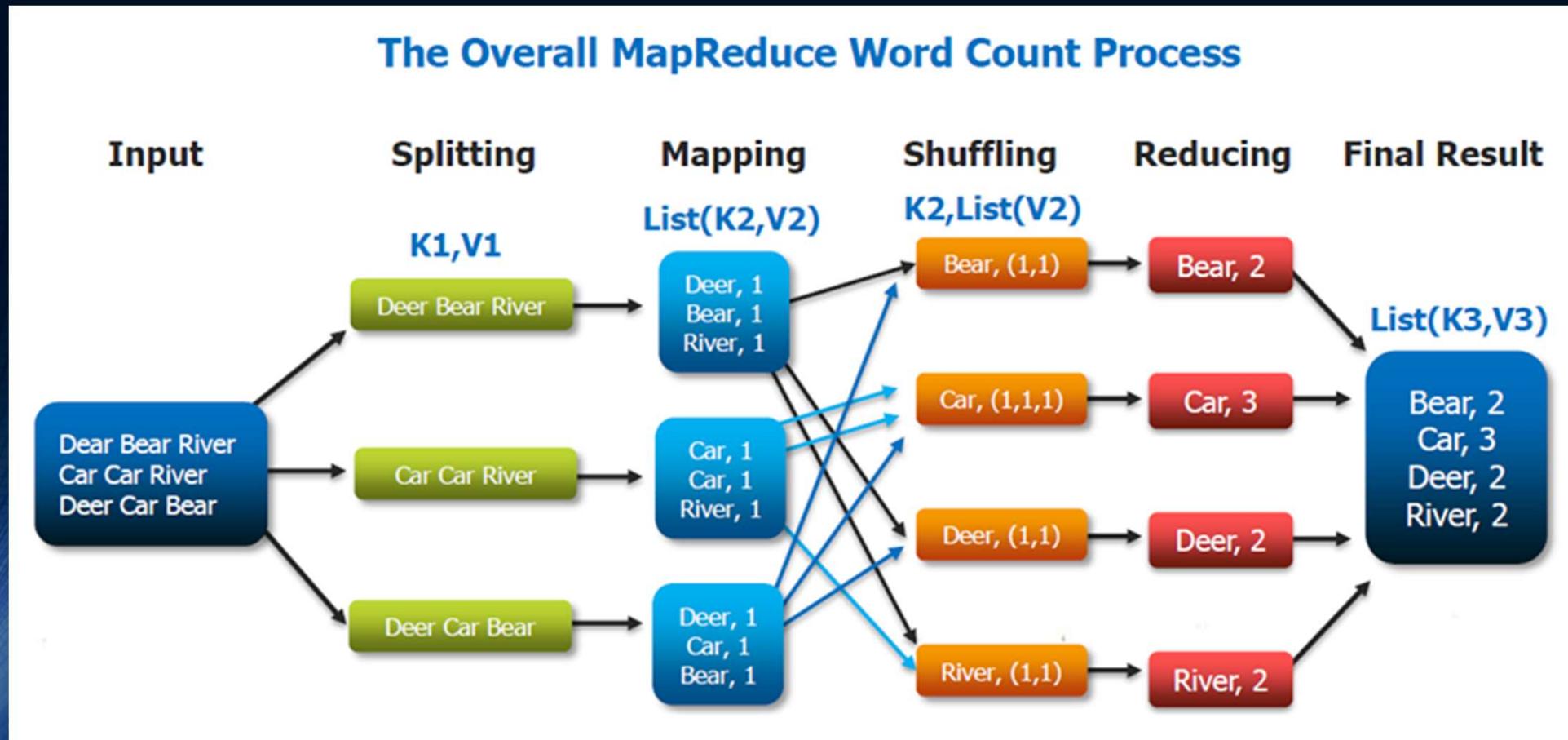
5 – Tecnologia – Map Reduce

- **Map**: Responsável por receber os dados na forma de chave/valor representando de forma lógica cada registro dos dados de entrada, podendo ser, por exemplo, uma linha em um arquivo de log ou de uma tabela. A função map retorna uma lista com zero ou mais dados chave/valor e deve ser codificada pelo desenvolvedor, através de outras ferramentas ou da API Java;
- **Shuffle**: A etapa de shuffle é responsável por organizar o retorno da função Map, atribuindo para a entrada de cada Reduce todos os valores associados a uma mesma chave. Este etapa é realizada pela biblioteca do MapReduce;
- **Reduce**: Por fim, ao receber os dados de entrada a função Reduce retorna uma lista de chave/valor contendo zero ou mais registros, semelhante ao Map, deve ser codificada pelo desenvolvedor.

5 – Tecnologia – Map Reduce - Arquitetura



5 – Tecnologia – Map Reduce - Exemplo

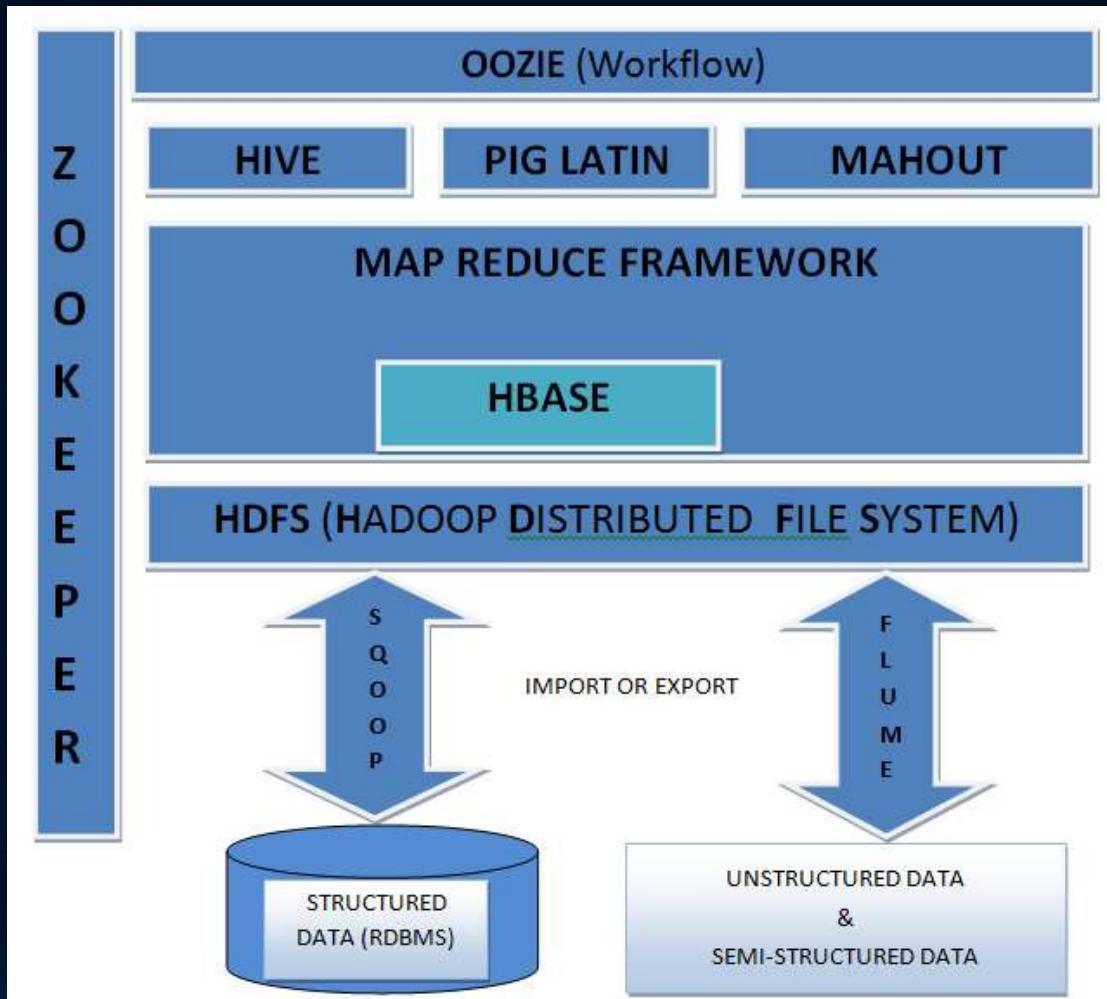


5 – Hadoop

- O Apache Hadoop é uma plataforma de software de código aberto para o armazenamento e processamento distribuído de grandes conjuntos de dados, utilizando clusters de computadores com hardware commodity.
- Os serviços do Hadoop fornecem armazenamento , processamento, acesso, governança, segurança e operações de Dados.



5 – Hadoop - Ecossistema



5 – Hadoop - Ecossistema

- **Hadoop MapReduce**: um modelo de programação e um arcabouço especializado no processamento de conjuntos de dados distribuídos em um aglomerado computacional. Abstrai toda a computação paralela em apenas duas funções: **Map** e **Reduce**.
- **Hadoop Distributed File System (HDFS)**: um sistema de **arquivos distribuído** nativo do Hadoop. Permite o armazenamento e transmissão de grandes conjuntos de dados em máquinas de baixo custo. Possui mecanismos que o caracteriza como um sistema **altamente tolerante a falhas**.