Ambiente Livre
Tecnologia

Hadoop

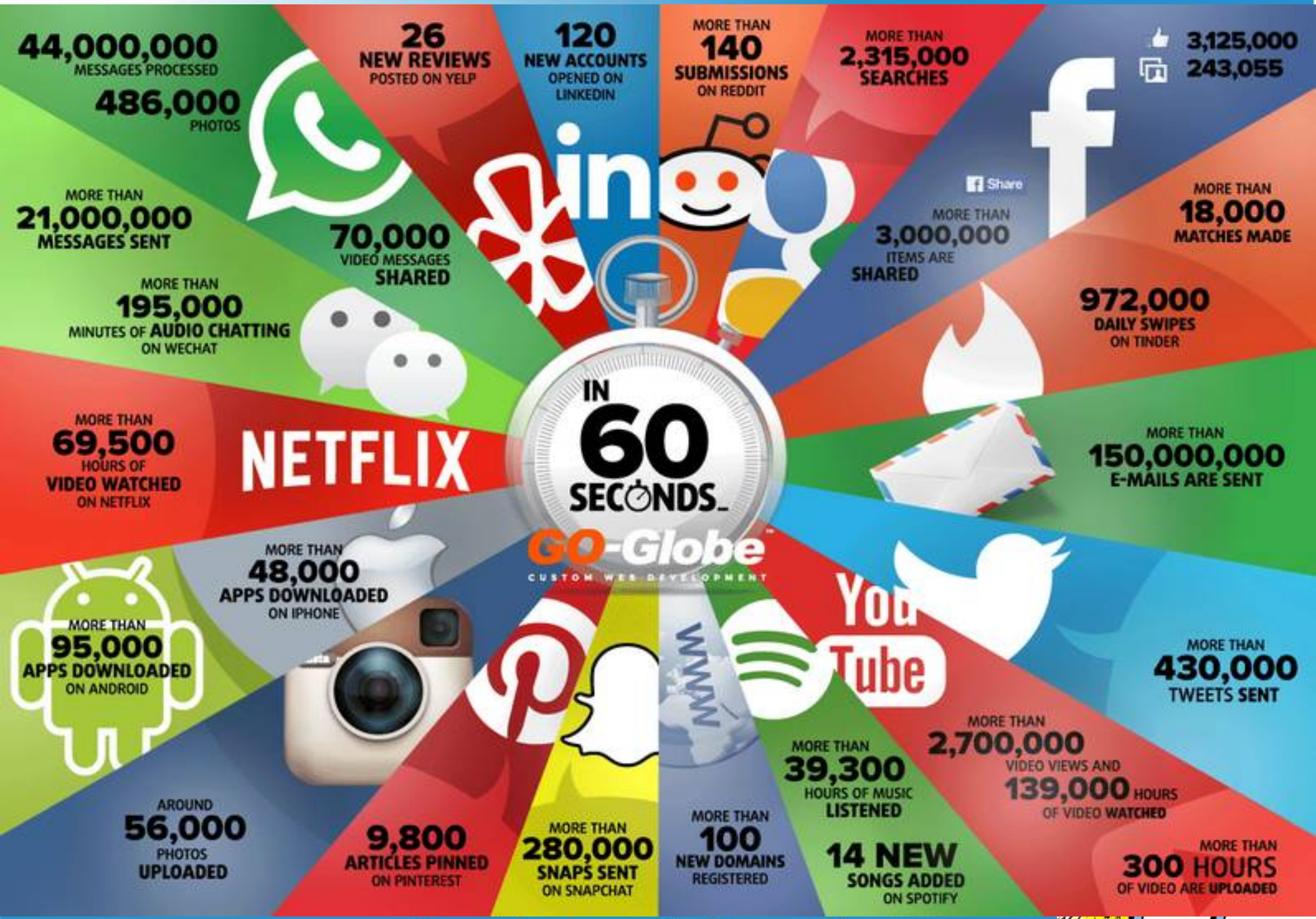
Treinamento Hadoop – Big Data Open Source - Fundamental.

Instrutor: Marcio Junior Vieira.
marcio@ambientelivre.com.br

Big Data - Muito se fala...

2005 na apresentação do Papa Bento XVI

2013 na apresentação do Papa Francisco


44,000,000
MESSAGES PROCESSED
486,000
PHOTOS
26
NEW REVIEWS
POSTED ON YELP
120
NEW ACCOUNTS
OPENED ON
LINKEDIN
MORE THAN
140
SUBMISSIONS
ON REDDIT
MORE THAN
2,315,000
SEARCHES
3,125,000
243,055
MORE THAN
21,000,000
MESSAGES SENT
70,000
VIDEO MESSAGES
SHARED
MORE THAN
195,000
MINUTES OF AUDIO CHATTING
ON WECHAT
Share
MORE THAN
3,000,000
ITEMS ARE
SHARED
MORE THAN
18,000
MATCHES MADE
972,000
DAILY SWIPES
ON TINDER
IN
60
SECONDS
GO-Globe
CUSTOM WEB DEVELOPMENT
MORE THAN
69,500
HOURS OF
VIDEO WATCHED
ON NETFLIX
NETFLIX
MORE THAN
150,000,000
E-MAILS ARE SENT
MORE THAN
48,000
APPS DOWNLOADED
ON IPHONE
MORE THAN
95,000
APPS DOWNLOADED
ON ANDROID
YouTube
MORE THAN
430,000
TWEETS SENT
MORE THAN
2,700,000
VIDEO VIEWS AND
139,000 HOURS
OF VIDEO WATCHED
MORE THAN
39,300
HOURS OF MUSIC
LISTENED
AROUND
56,000
PHOTOS
UPLOADED
9,800
ARTICLES PINNED
ON PINTEREST
MORE THAN
280,000
SNAPS SENT
ON SNAPCHAT
MORE THAN
100
NEW DOMAINS
REGISTERED
14 NEW
SONGS ADDED
ON SPOTIFY
MORE THAN
300 HOURS
OF VIDEO ARE UPLOADED

Big Data

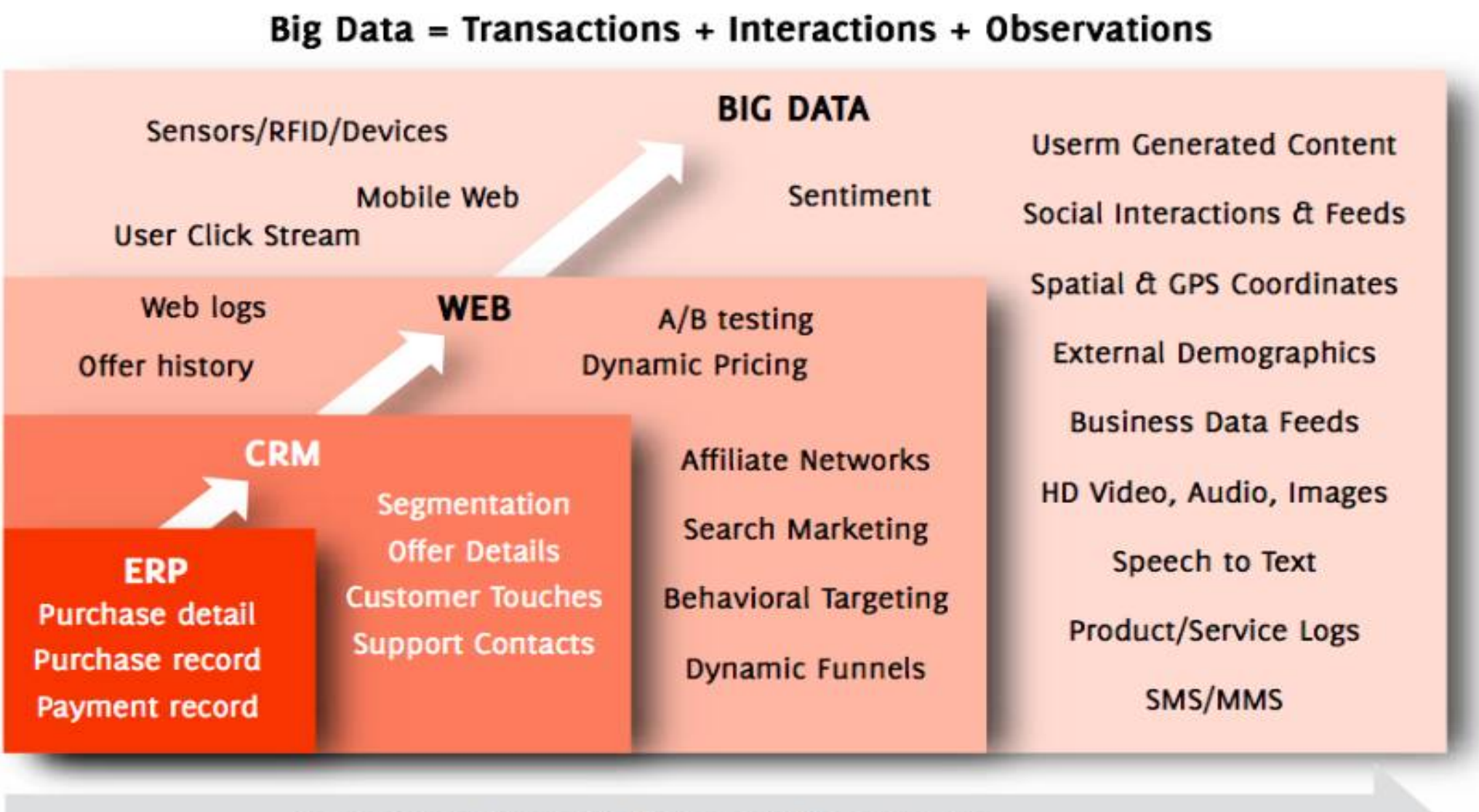
- É um novo conceito se consolidando.
- Grande armazenamento de dados e maior velocidade

O quão Big Data é Big ?

- Seus dados não cabem mais em uma única maquina.
- Quando falamos mais em Terabytes que Gigabytes
- A quantidade de dados cresce todo o mês e deve dobrar até o próximo ano.

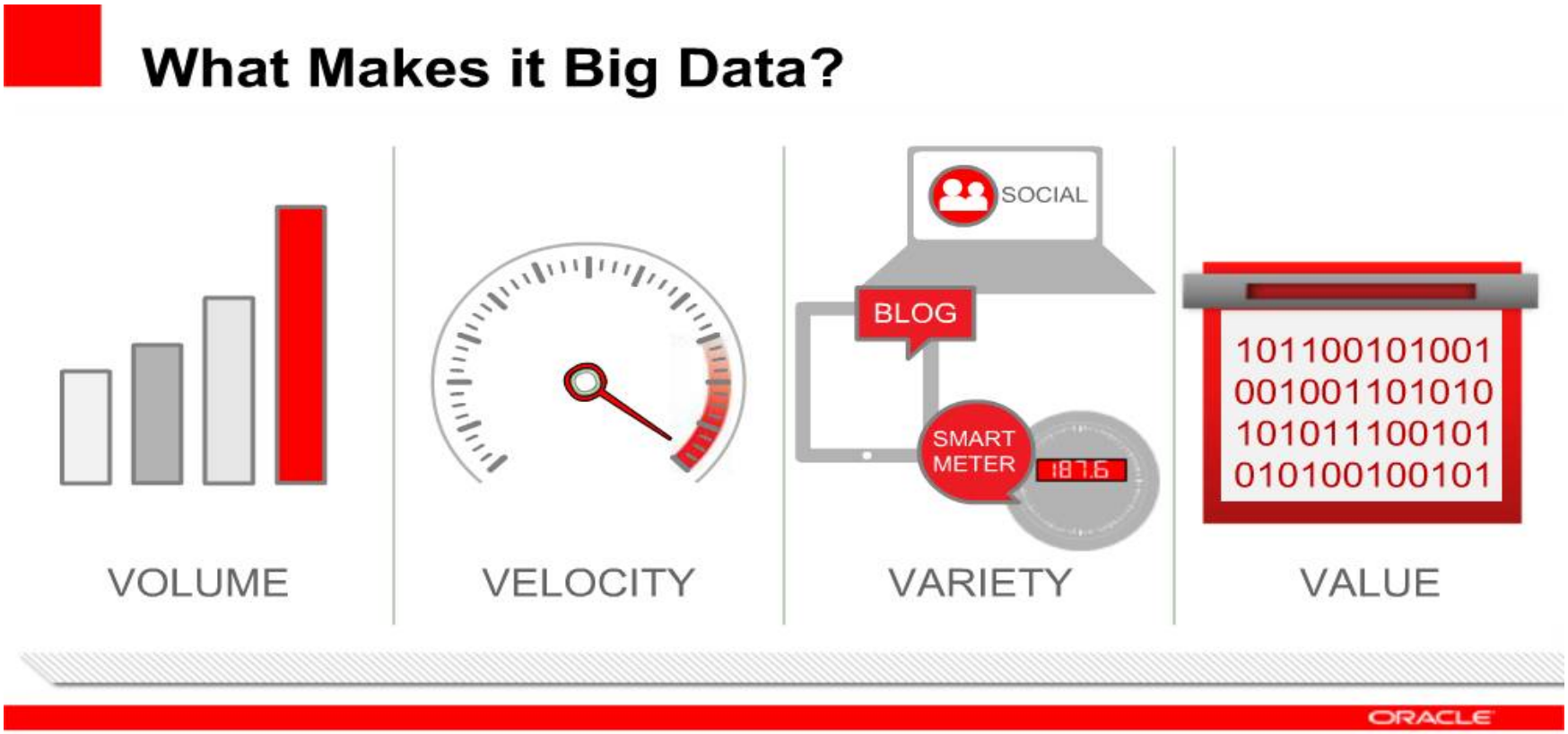
Ambiente Livre
Tecnologia

Big Data

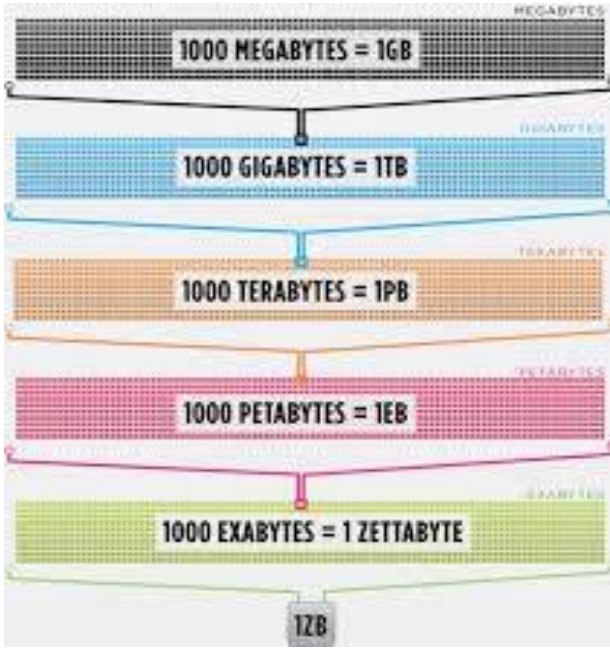


Os 4 V's

- Velocidade , Volume , Variedade e Valor



Volume



- Modelos de Persistência da ordem de Petabytes, zetabytes ou yottabyte(YB).
- Geralmente dados não estruturados.
- Um Zettabyte corresponde a 1.000.000.000.000.000.000.000 (10^{21}) ou 118059162071741130342 4 (2 elevado a 70) Bytes.

Velocidade

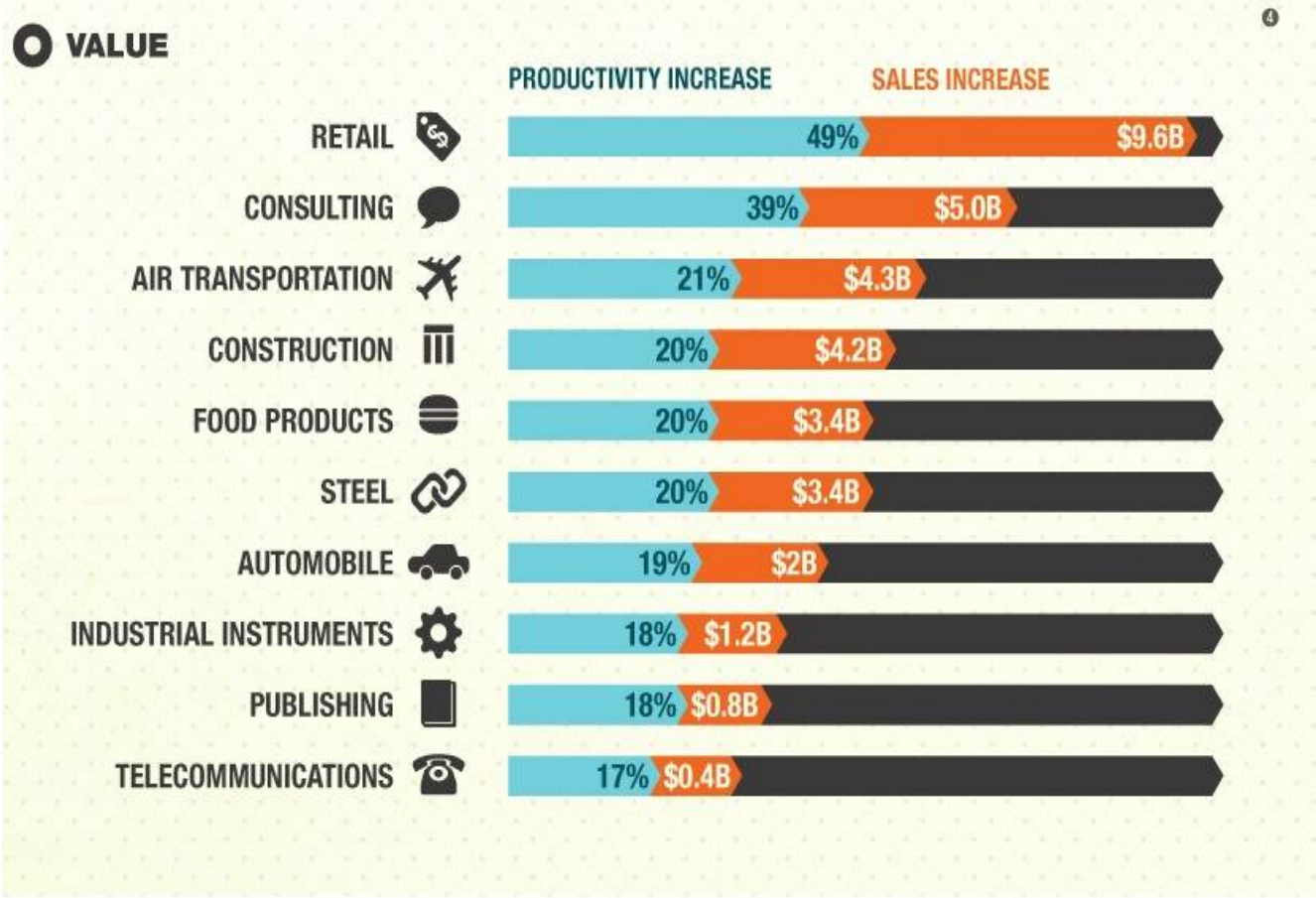
- Processamento de Dados
- Armazenamento
- Analise de Dados

Variedade

- Dados semi-estruturados
- Dados não estruturados
- Diferentes fontes
- Diferentes formatos

Valor

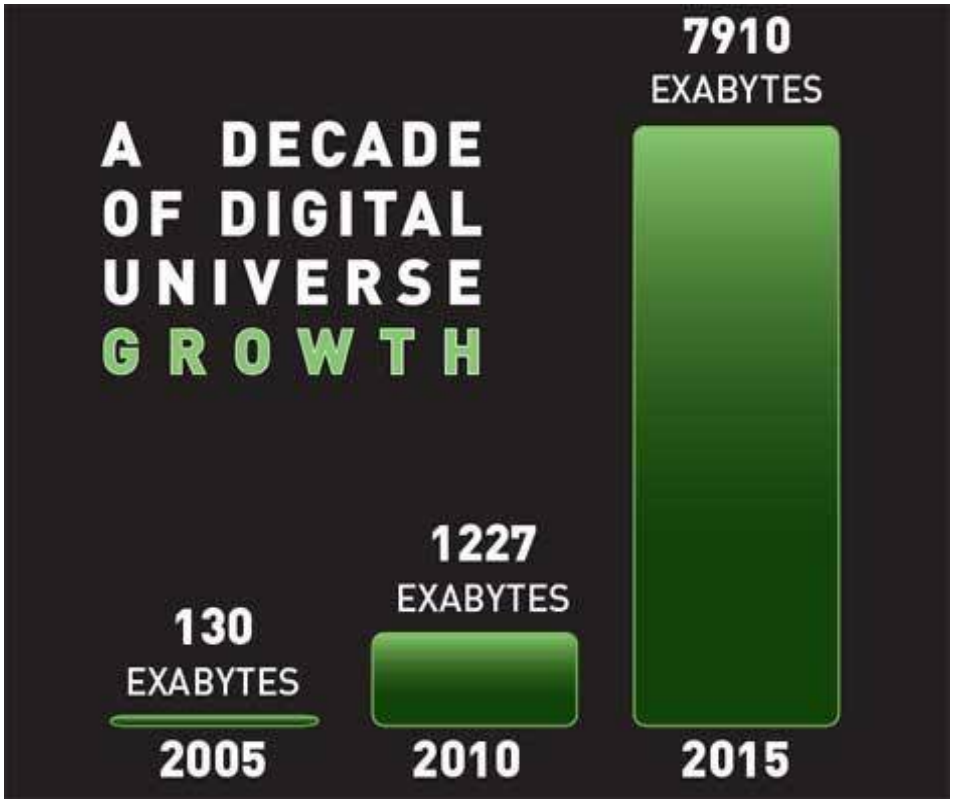
- Tomada de Decisão
- Benefícios
- Objetivo do Negócio.



O novo V: Veracidade

Veracidade refere-se a confiabilidade dos dados. Com muitas formas de grandes qualidades e precisão dos dados são menos controláveis (basta pensar em posts no Twitter com hash tags, abreviações, erros de digitação e linguagem coloquial, bem como a confiabilidade e a precisão do conteúdo), mas agora a tecnologia permite-nos trabalhar com este tipo de dados .

O momento é agora



Tomada de Decisão

- 1 em cada 3 gestores tomam decisão com base em informações que não confiam ou não tem
- 56% sentem sobrecarregados com a quantidade de dados que gerenciam
- 60% acreditam que precisam melhorar captura e entender informações rapidamente.
- 83% apontam que BI & analytics fazem parte de seus planos para aumentar a competitividade

fonte : Survey KPMG.

Onde usar Big Data ?

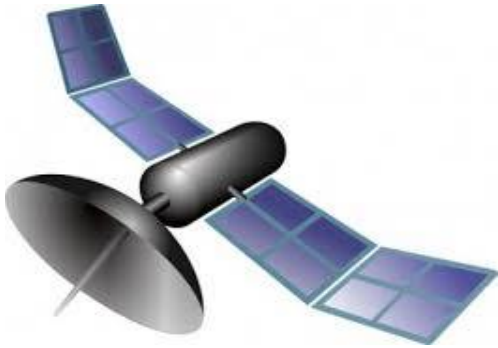
- **Sistemas de recomendação**

- **Redes Sociais**

Onde usar Big Data ?

- Analise de Risco (Crédito, Seguros , Mercado Financeiro)

- Dados Espaciais (Clima , Imagens, Trafego, Monitoramento)



- Energia Fotovoltaica (Medições , Estudos, Resultados)

Big Data X BI

- Big Data e uma evolução do BI, devem caminhar juntos
- Data Warehouses são necessários para armazenar dados estruturados

Previsão:

- BI – Casos específicos
- Big Data – Analise geral


Data Lake

Data Lake

- Fonte única
- Grande Volume
- Não Refinado
- Pode estar tratado.

Requisitos de um Data Lake

- Armazenar todos os dados
- Satisfazer relatório e rotinas de analise
- Satisfazer ad-hoc query / analises / relatórios
- Balanceamento de performance e custo

Formato Tradicional de BI

Data Mart(s)

Data Source

Ambiente Livre
Tecnologia

Arquitetura de Big Data

Data Mart(s)

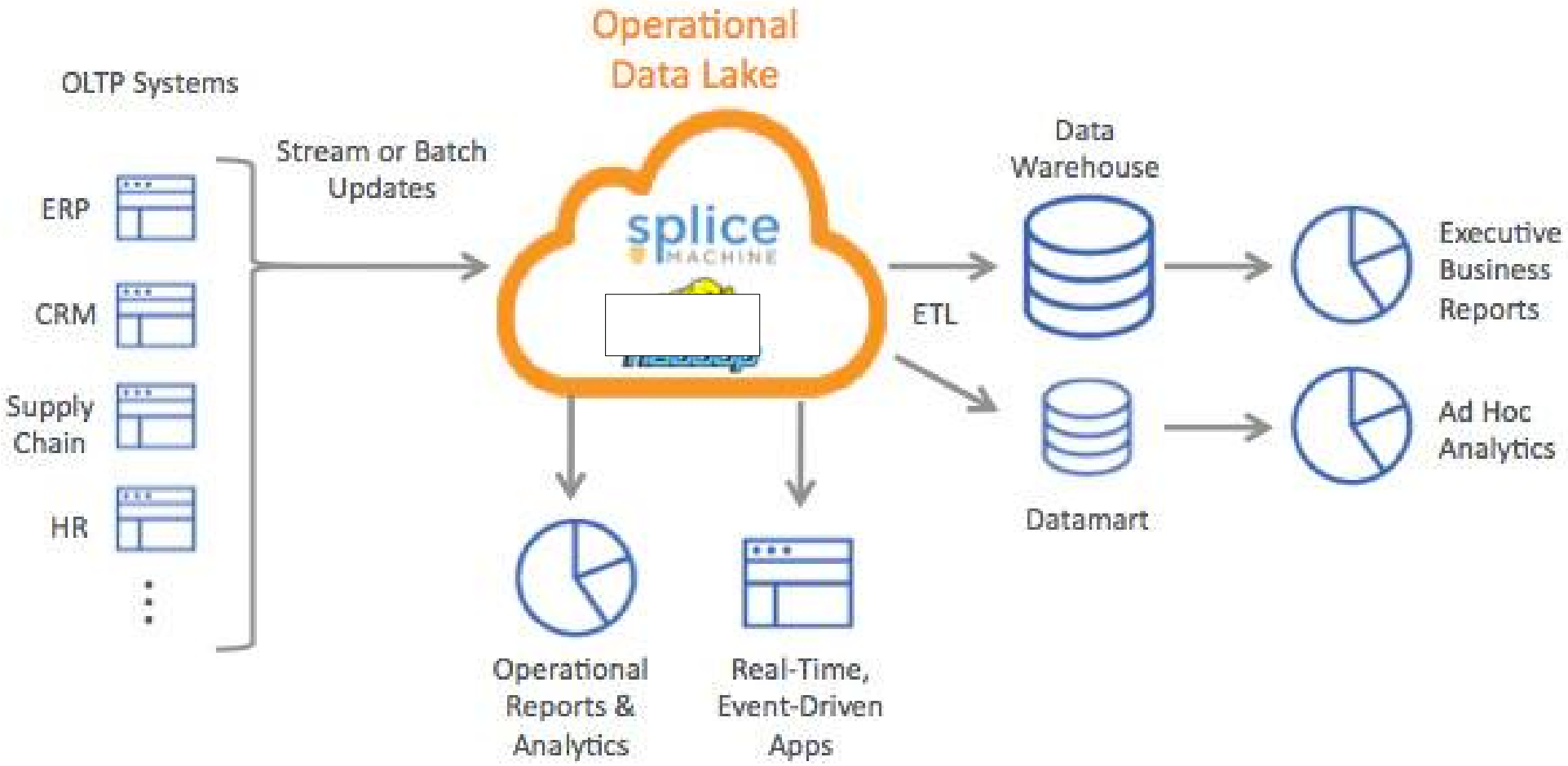
ad-hoc

Datawarehouse

Data Lake(s)

Data Source

Data Lake


Operational
Data Lake
OLTP Systems
Stream or Batch
Updates
ERP
CRM
Supply
Chain
HR
splice
MACHINE
ETL
Data
Warehouse
Executive
Business
Reports
Ad Hoc
Analytics
Datamart
Operational
Reports &
Analytics
Real-Time,
Event-Driven
Apps

Case Linkendin

- Pesquisador, percebeu que a rede social estava monótona e que as pessoas realizavam poucas interações.
- Sugeriu um algoritmo que apresentasse sugestões de amizades, conhecido como **‘People You May Know’**,
- Foi um sucesso e ajudou com que a rede se tornasse uma das mais utilizadas no mundo.
- O algoritmo utilizava informações disponibilizadas nos perfis, por exemplo, o colégio onde o usuário cursou o Ensino Médio. Comparando com os outros usuários, o algoritmo poderia sugerir pessoas que também estudaram no mesmo colégio, fazendo assim que as pessoas aumentassem seu número de conexões, proporcionando maiores interações.

Cases

A companhia Skybox tira fotos de satélite e vende a seus clientes informações em tempo real sobre a disponibilidades de vagas de estacionamento livres numa cidade em determinada hora ou quantos navios estão ancorados no mundo neste momento

O projeto Global Pulse, das Nações Unidas, vai utilizar um programa que decifra a linguagem humana na análise de mensagens de texto e posts em redes sociais para prever o aumento do desemprego, o esfriamento econômico e epidemias de doenças

A varejista americana Dollar General monitora as combinações de produtos que seus clientes põem nos carrinhos. Ganhou eficácia e ainda descobriu curiosidades: quem bebe Gatorade tem mais chances de comprar também laxante

A Sprint Nextel saltou da última para a primeira posição no ranking de satisfação dos usuários de celular nos EUA ao integrar os dados de todos os seus canais de relacionamento. De quebra, cortou pela metade os gastos com call center

No terremoto do Haiti, pesquisadores americanos perceberam antes de todo mundo a diáspora de Porto Príncipe por meio dos dados de geolocalização de 2 milhões de chips SIM de celulares, facilitando a atuação da ajuda humanitária

Um hospital no Canadá usou tecnologia da IBM e da Universidade de Ontário para monitorar em tempo real dezenas de indicadores de saúde de bebês prematuros. O cruzamento permitiu aos médicos antecipar ameaças às vidas das crianças

Em busca dos melhores lugares para instalar turbinas eólicas, a dinamarquesa Vestas Wind analisou petabytes de dados climáticos, de nível das marés, mapas de desmatamento etc. O que costumava levar semanas durou algumas horas

O Profissional “data scientist”

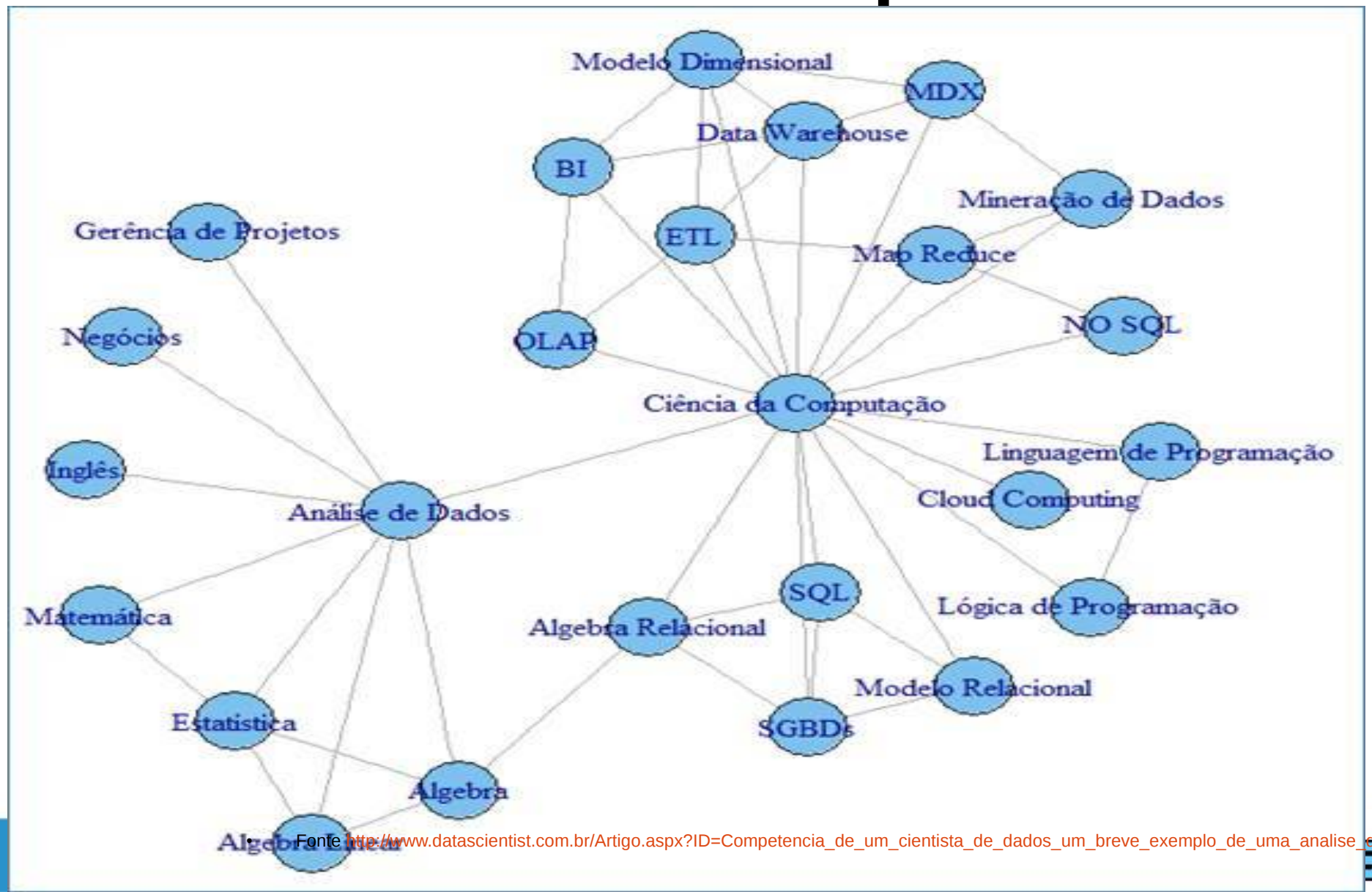
Novo profissional: **Cientista de Dados**

Cientista de dados

- **Gartner:** necessitaremos de 4,4 Milhões de especialistas até 2015 (1,9M América do Norte, 1,2M Europa Ocidental e 1,3M Ásia/Pacifico e América Latina)
- Estima-se que apenas um terço disso será preenchido. (**Gartner**)
- Brasil deverá abrir 500 mil vagas para profissionais com habilidades em Big Data
- As universidades do Brasil ainda não oferecem graduação para formação de cientistas de dados

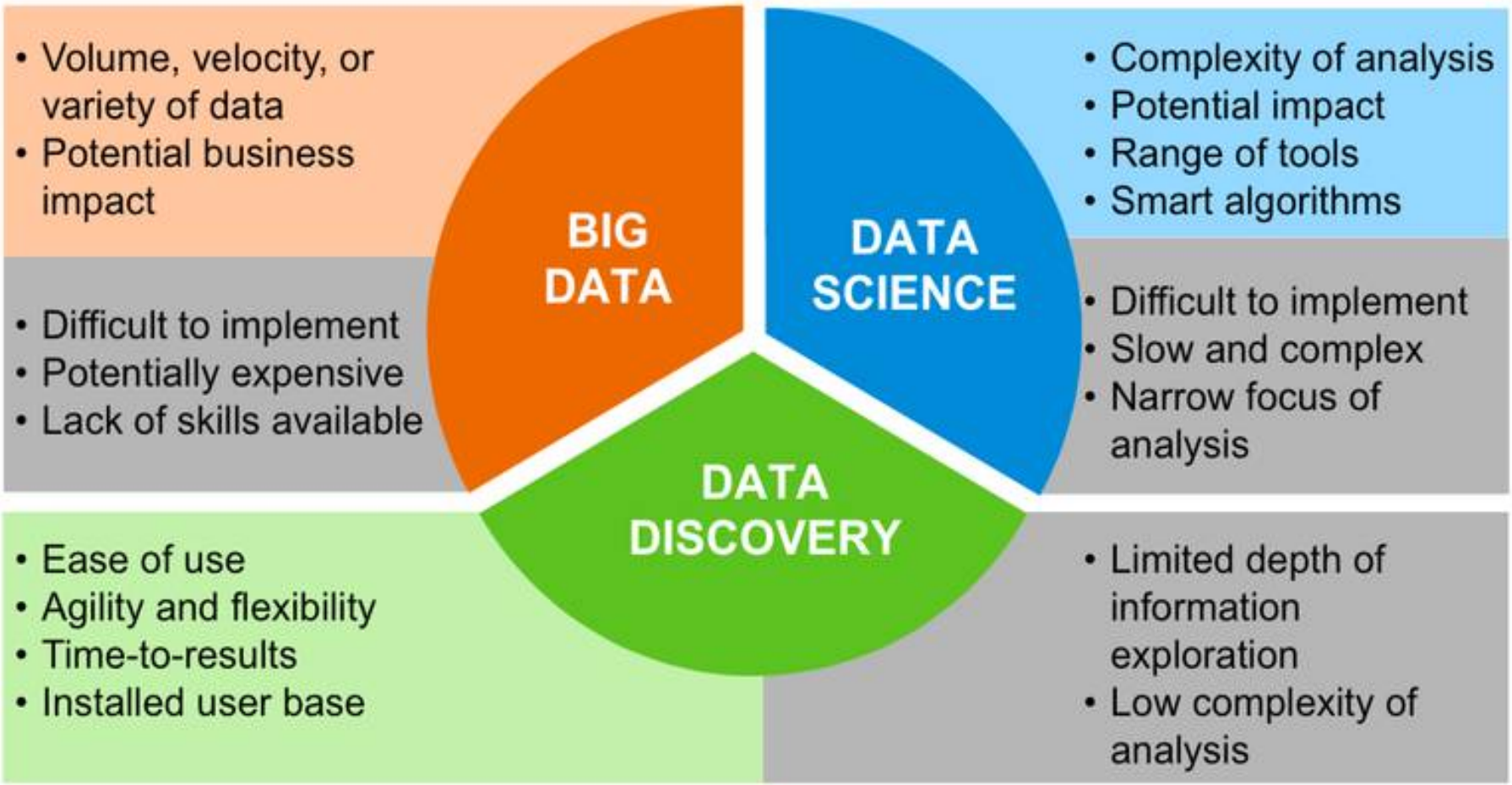
Ambiente Livre
Tecnologia

Competências

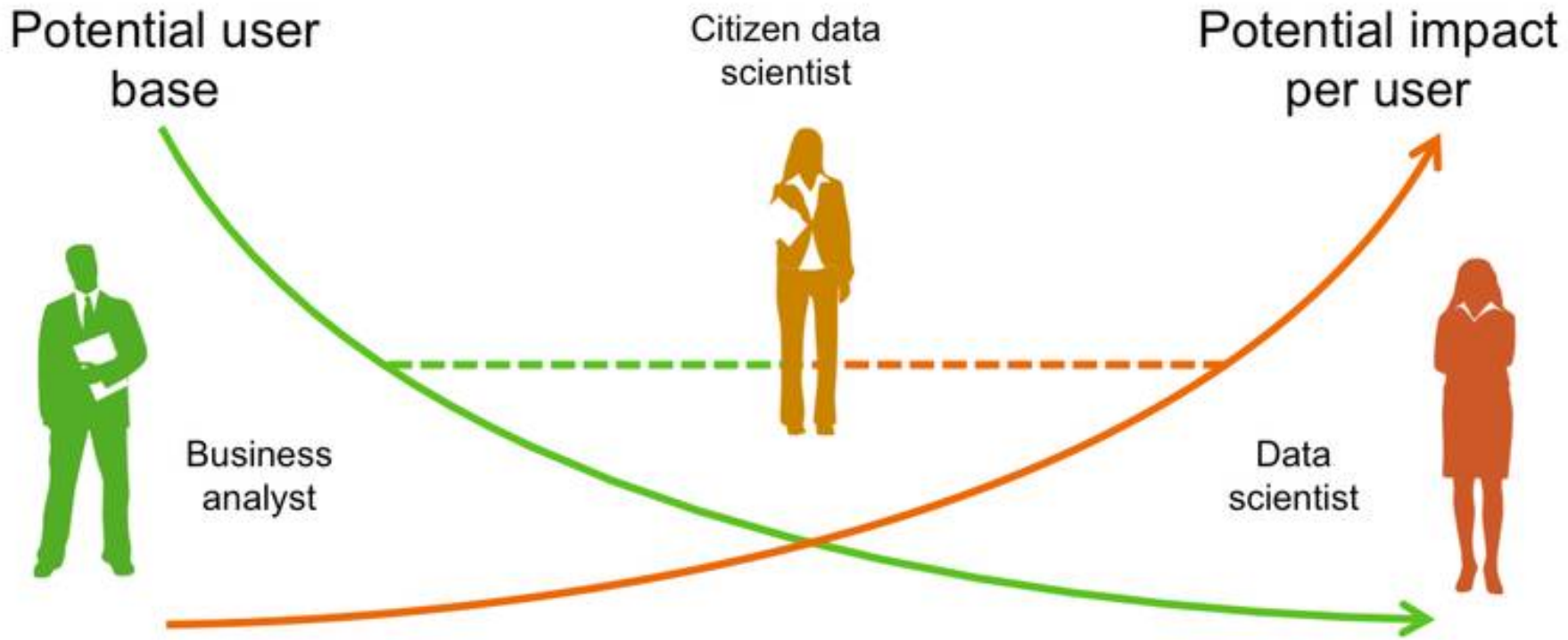


Fonte http://www.datascientist.com.br/Artigo.aspx?ID=Competencia_de_um_cientista_de_dados_um_breve_exemplo_de_uma_analise_de_redes

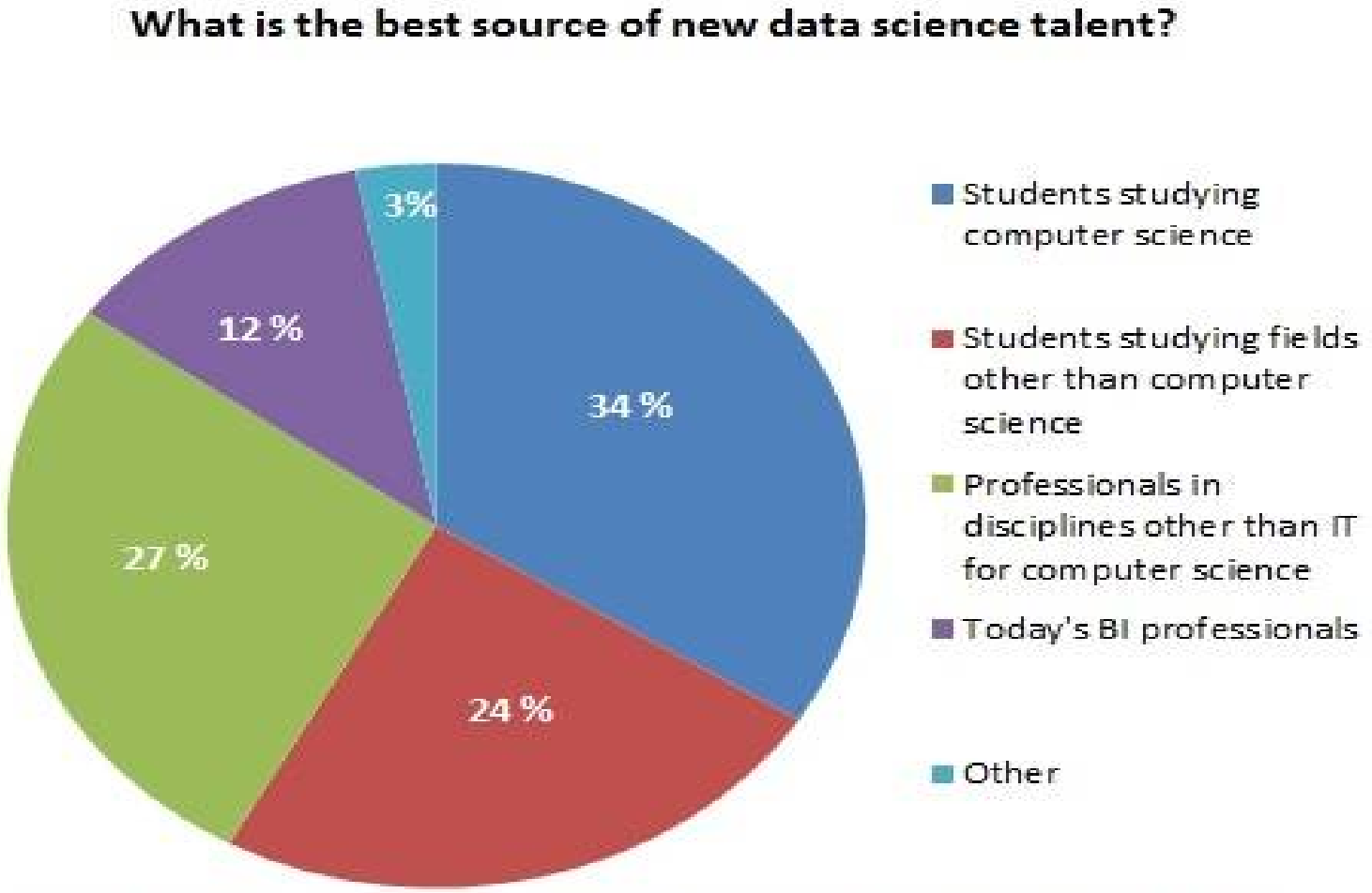
Tendências


• Volume, velocity, or variety of data
• Potential business impact
BIG DATA
• Difficult to implement
• Potentially expensive
• Lack of skills available
• Complexity of analysis
• Potential impact
• Range of tools
• Smart algorithms
DATA SCIENCE
• Difficult to implement
• Slow and complex
• Narrow focus of analysis
DATA DISCOVERY
• Ease of use
• Agility and flexibility
• Time-to-results
• Installed user base
• Limited depth of information exploration
• Low complexity of analysis

Tendências
Citizen Data Scientist

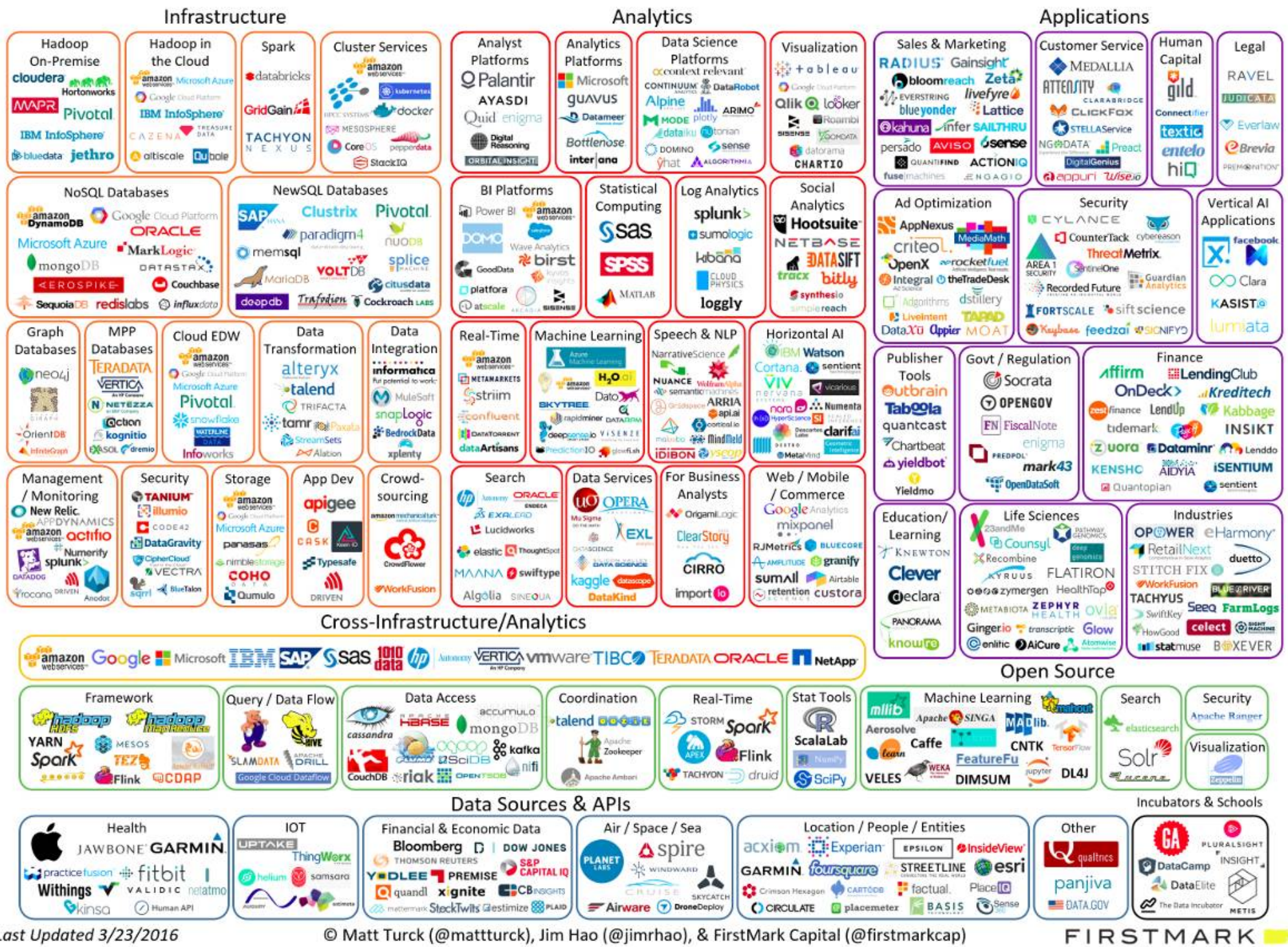


De onde ?



Source: EMC Data Science Community Survey, 497 data scientists and business professionals, Dec. 2011

Big Data Landscape 2016 (Version 3.0)



FIRSTMARK


Software Livre
Open Source

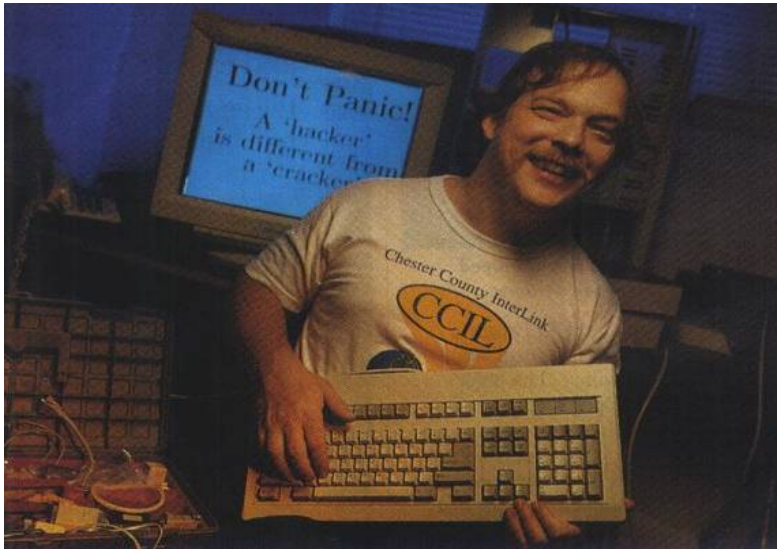
Software Livre

- **"Software livre"** se refere à liberdade dos usuários executarem, copiarem, distribuírem, estudarem, modificarem e aperfeiçoarem o software. São 4 tipos de liberdade, para os usuários do software:
- **1**. A liberdade de executar o programa, para qualquer propósito.
- **2**. A liberdade de estudar como o programa funciona, e adaptá-lo para as suas necessidades. Acesso ao código-fonte é um pré-requisito para esta liberdade.
- **3**. A liberdade de redistribuir cópias de modo que você possa ajudar ao seu próximo.
- **4**. A liberdade de aperfeiçoar o programa, e liberar os seus aperfeiçoamentos, de modo que toda a comunidade se beneficie.

GPL v3 Free Software
Free as in Freedom

Open Source

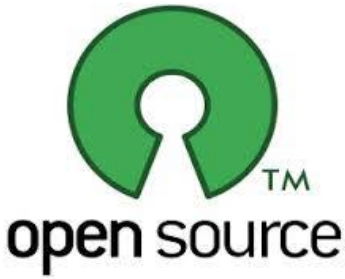
- Criado pela OSI (Open Source Initiative)
- Não refere-se a software também conhecido por software livre.
- Qualquer licença de software livre é também uma licença de código aberto (Open Source)
- Mas o contrário nem sempre é verdade
- Criado por Eric Raymond e outros fundadores da OSI.



Free Software X OSI

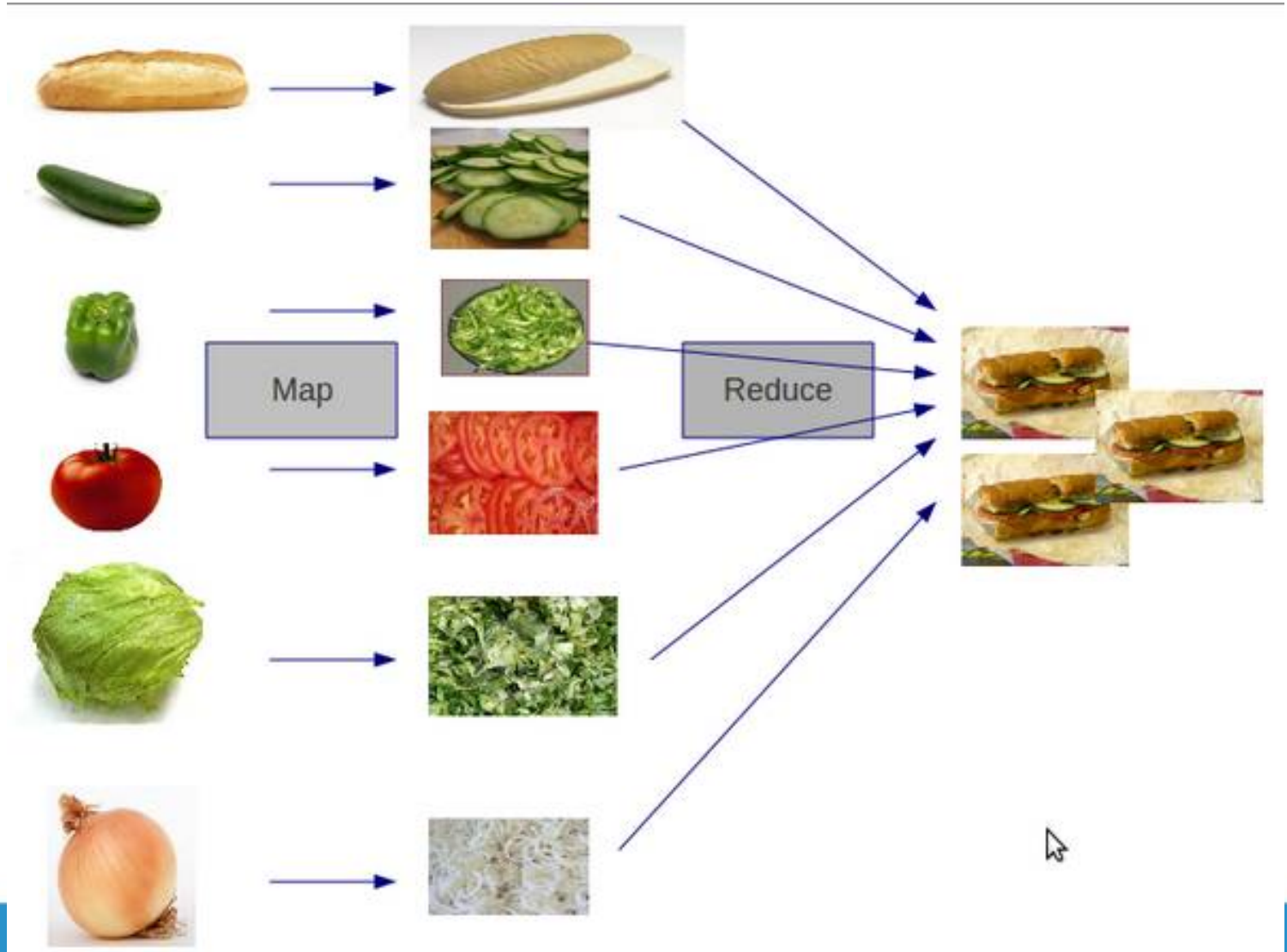
- 4 Lei da GPL
- OBRIGATORIEDADE:
  A liberdade de aperfeiçoar o programa, e liberar os seus aperfeiçoamentos, de modo que toda a comunidade se beneficie.

X



Muitos das melhores e mais conhecidas ferramentas de dados disponíveis são grandes projetos de código aberto. O mais conhecido deles é o Hadoop, o que está gerando toda uma indústria de serviços e produtos relacionados.

MapReduce by “Subway”



IoT (Internet of Things) e Big Data

- Internet das Coisas se aplica a comunicação entre objetos e entre estes e a internet, sejam eles físicos ou virtuais.

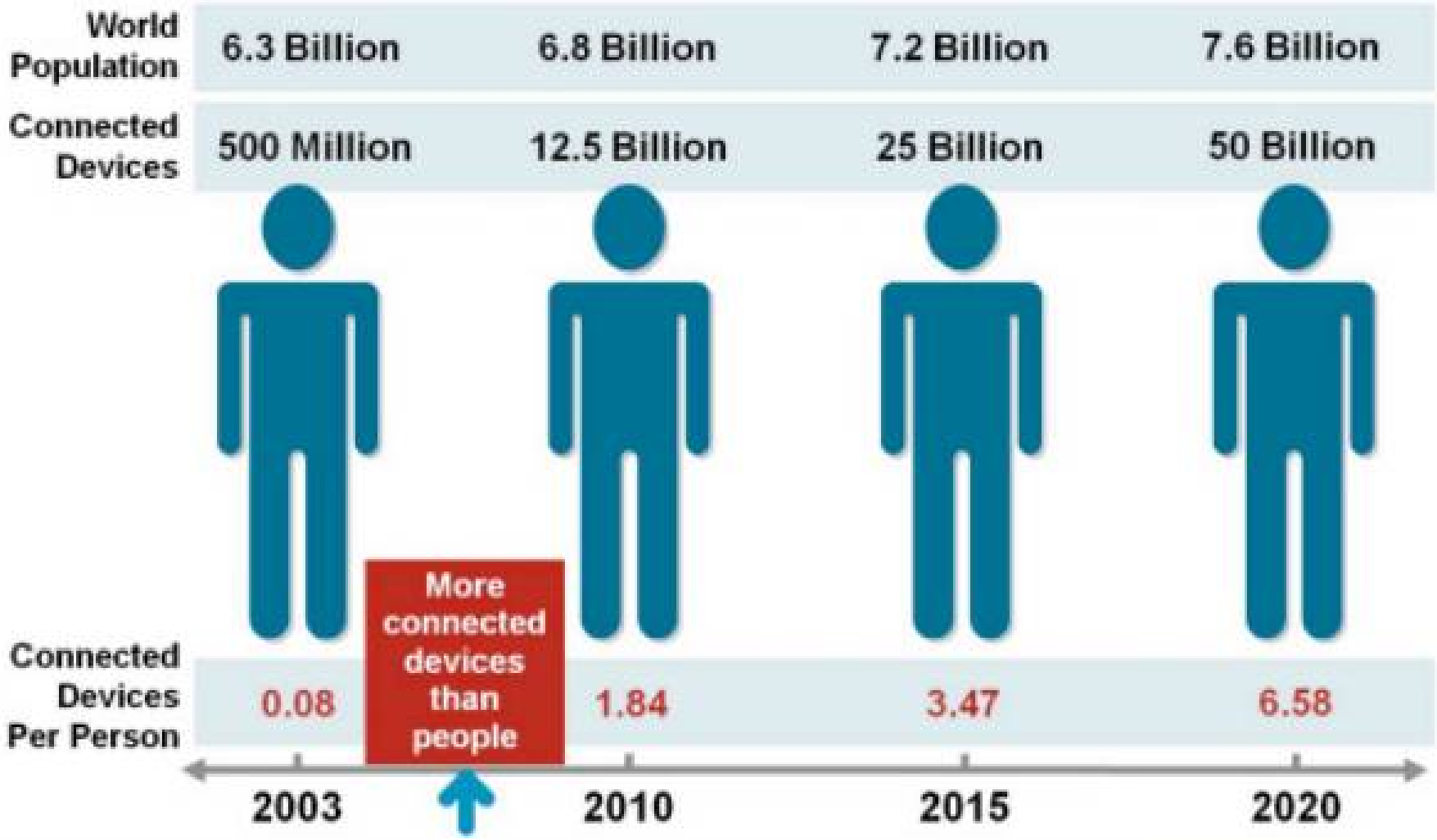
Razões para IoT

- Economia de processos;
- Maximização de lucros;
- Sustentabilidade;
- Marketing;
- Integração com redes sociais…

Elementos do IOT

- Rede de sensores sem fio (RSSF);
- Radio Frequency Identification (RFID);
- Gateway;
- Banco de dados;
- Protocolos de rede;
- Gerência de processos;
- Gerência de rede.

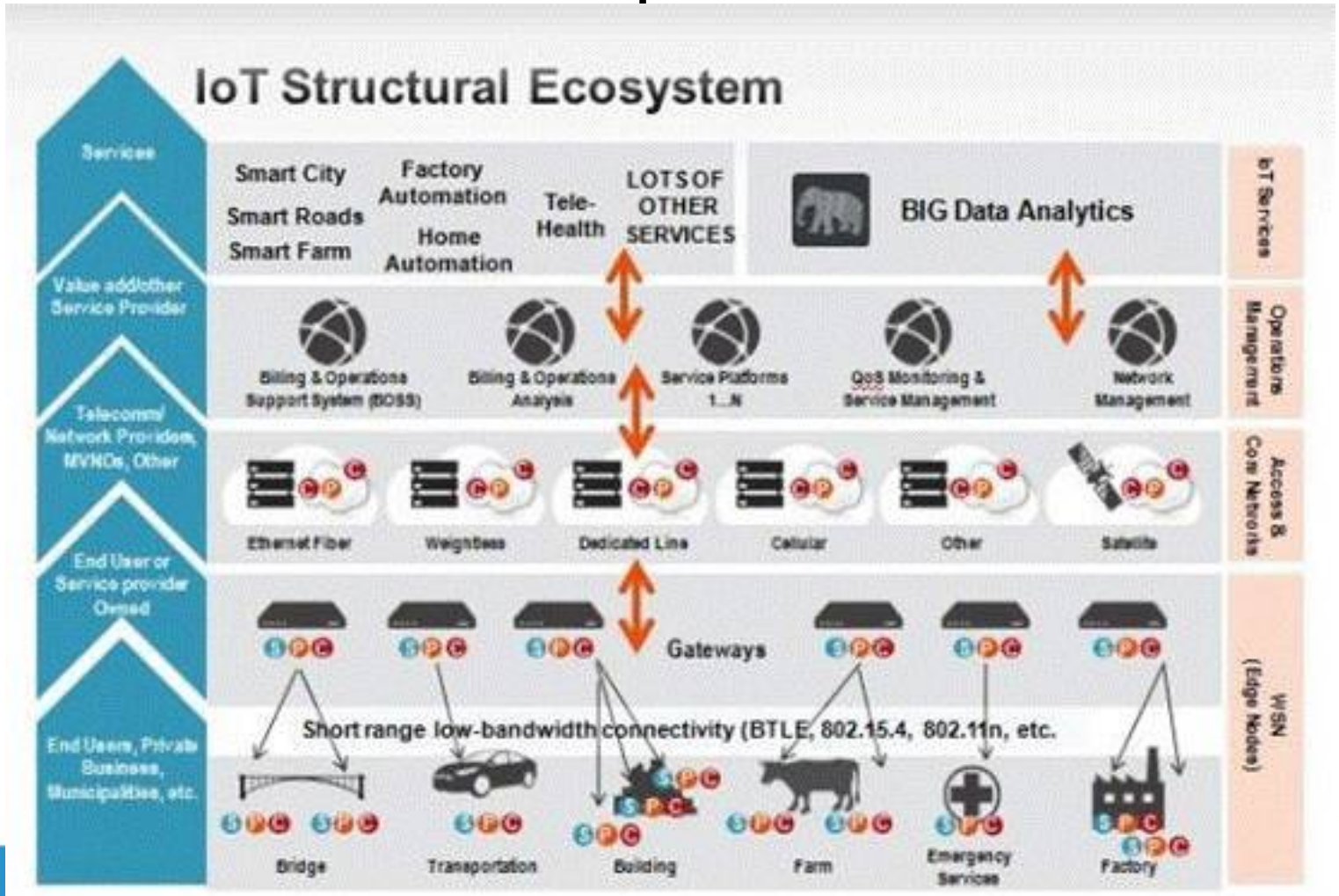
Evolução das Coisas



Ambiente Livre
Tecnologia

Impacto Econômico - IoT

- U$ 4 a 11 trilhões a partir de 2025

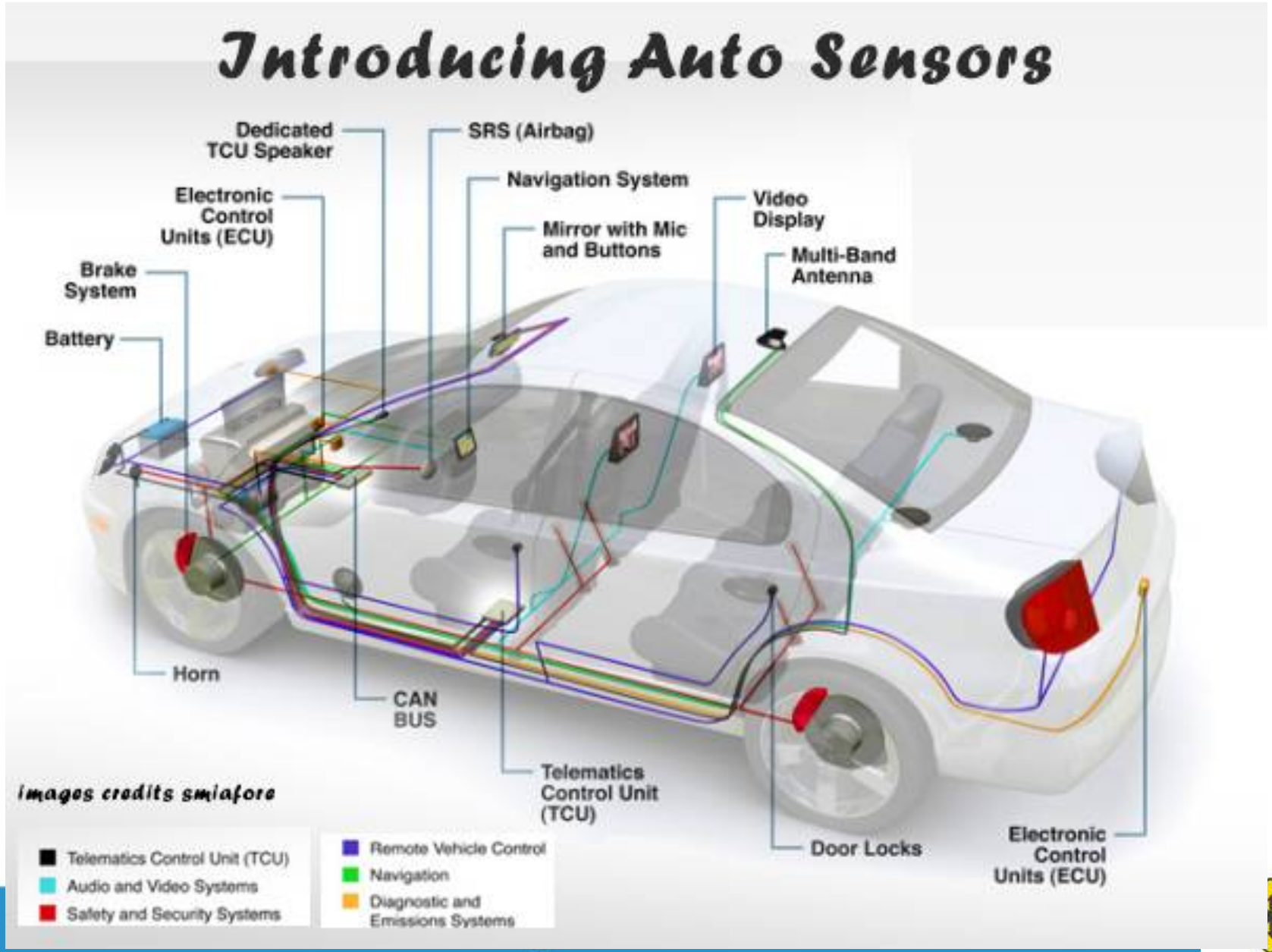


A amizade sensor Big Data

Sensores

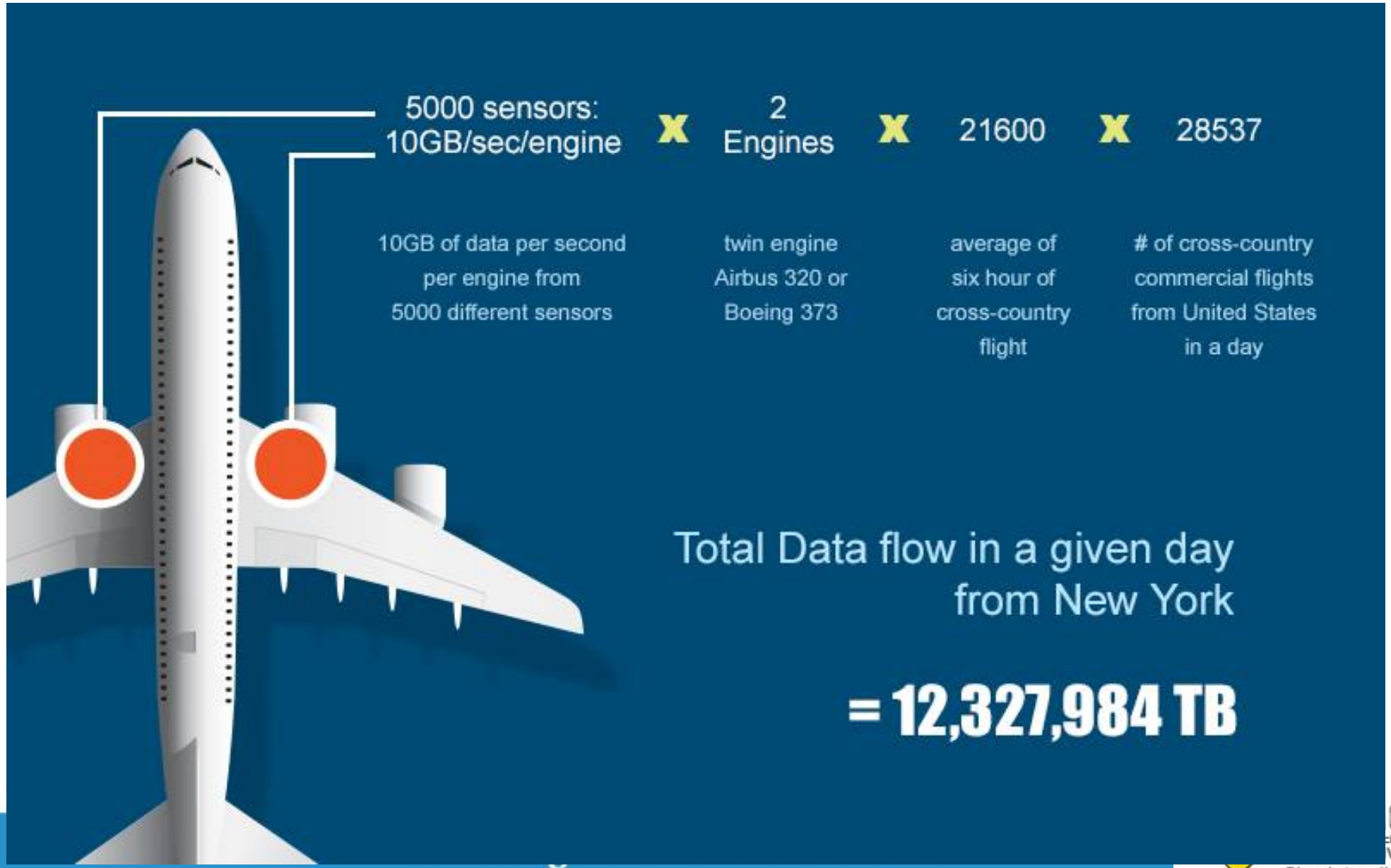
Ambiente Livre
Tecnologia

Sensores de Automóveis



Ambiente Livre
Tecnologia

Sensores de Voo



Principais desafios

- O Big Data não envolve só mudança de tecnologia, envolve adaptação de processos e treinamento relacionado à mudanç de gestão e analise de dados (MERITALK BIG DATA EXCHANGE, 2013)
- A maioria dos lideres não seba lidar com essa grande variedade e quantidade de informações, e não tem conhecimento dos benefícios que uma analise bem feita destas dados poderia tazer ao seu negocio(COMPUTERWORLD, 2012)
- Falta da cultura: a maioria das empresas não fazem um bom trabalho com as informações que já tem.
- Desafios dos Os 5 V !
- Privacidade, A identidade do usuário, mesmo preservada pode ser buscada... (Marco Civil da Internet)

Recomendações

- Comece com o problema, e não com os dados
- Compartilhe dados para receber dados
- Suporte gerencial e executivo
- Orçamento suficiente
- Melhores parceiros e fornecedores

Big Data

- “Big Data hoje é o que era a Web em 1993.
- Sabemos que será algo grande, mas não sabemos como...”

COPPEAD/UFRJ