UNIVERSIDADE FEDERAL DE ITAJUBÁ

Banco de Dados II

COM 231

Big Data
Aula 14

Vanessa Cristina Oliveira de Souza

O que é Big Data?

O que é Big Data?

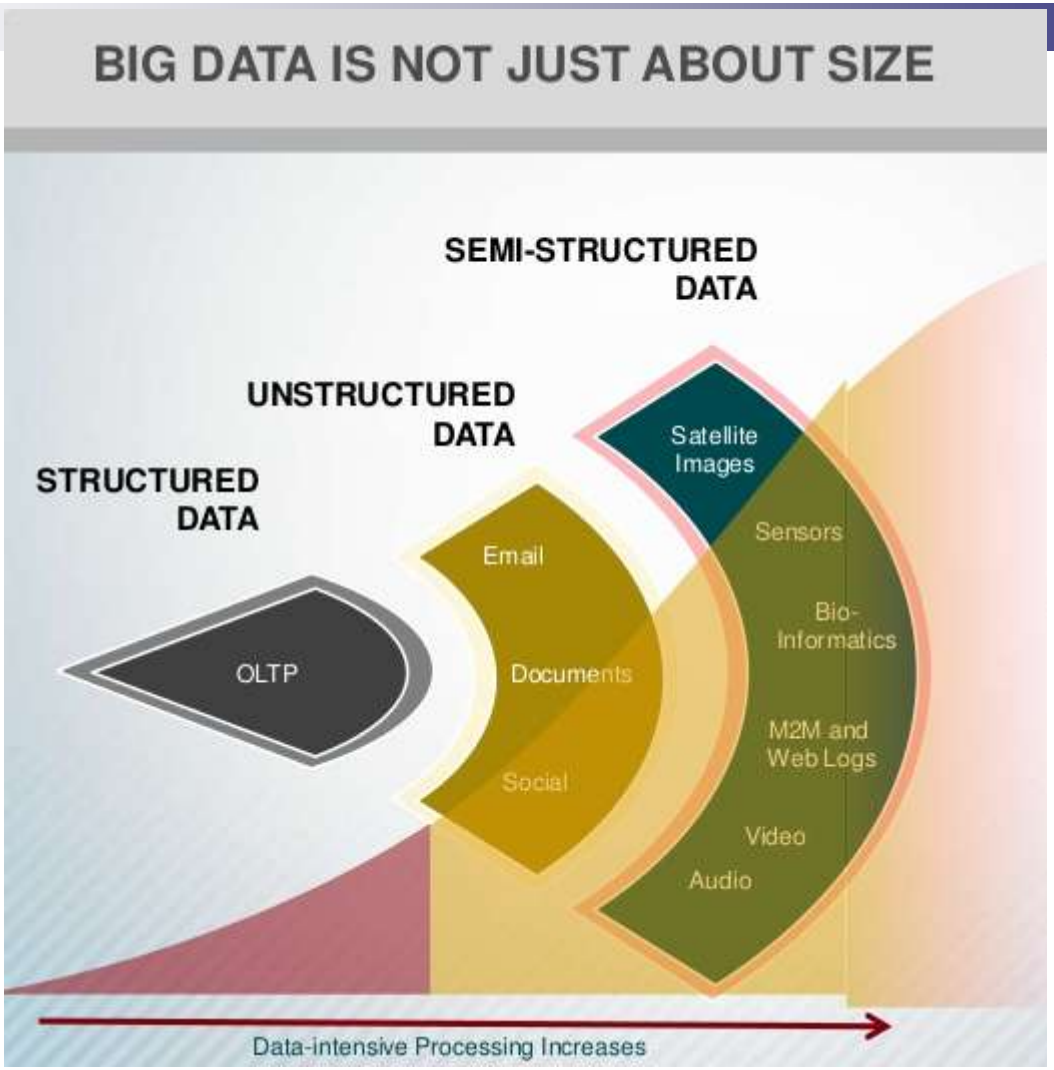
O que é Big Data?

COM 1,4 BI DE USUÁRIOS, FACEBOOK É A MAIOR "NAÇÃO" DO MUNDO

Country	Population	# of Facebook Users	% of Population Using Facebook
United Kingdom	63,742,977	36,000,000	56.48%
United States	318,892,103	180,000,000	56.45%
Canada	34,834,841	19,600,000	56.27%
Argentina	43,024,374	24,000,000	55.78%
Malaysia	30,073,353	16,000,000	53.20%
Colombia	46,245,297	22,000,000	47.57%
Turkey	81,619,392	38,000,000	46.56%
Brazil	202,656,788	92,000,000	45.40%
France	66,259,012	30,000,000	45.28%
Mexico	120,286,655	54,000,000	44.89%

O que é Big Data?

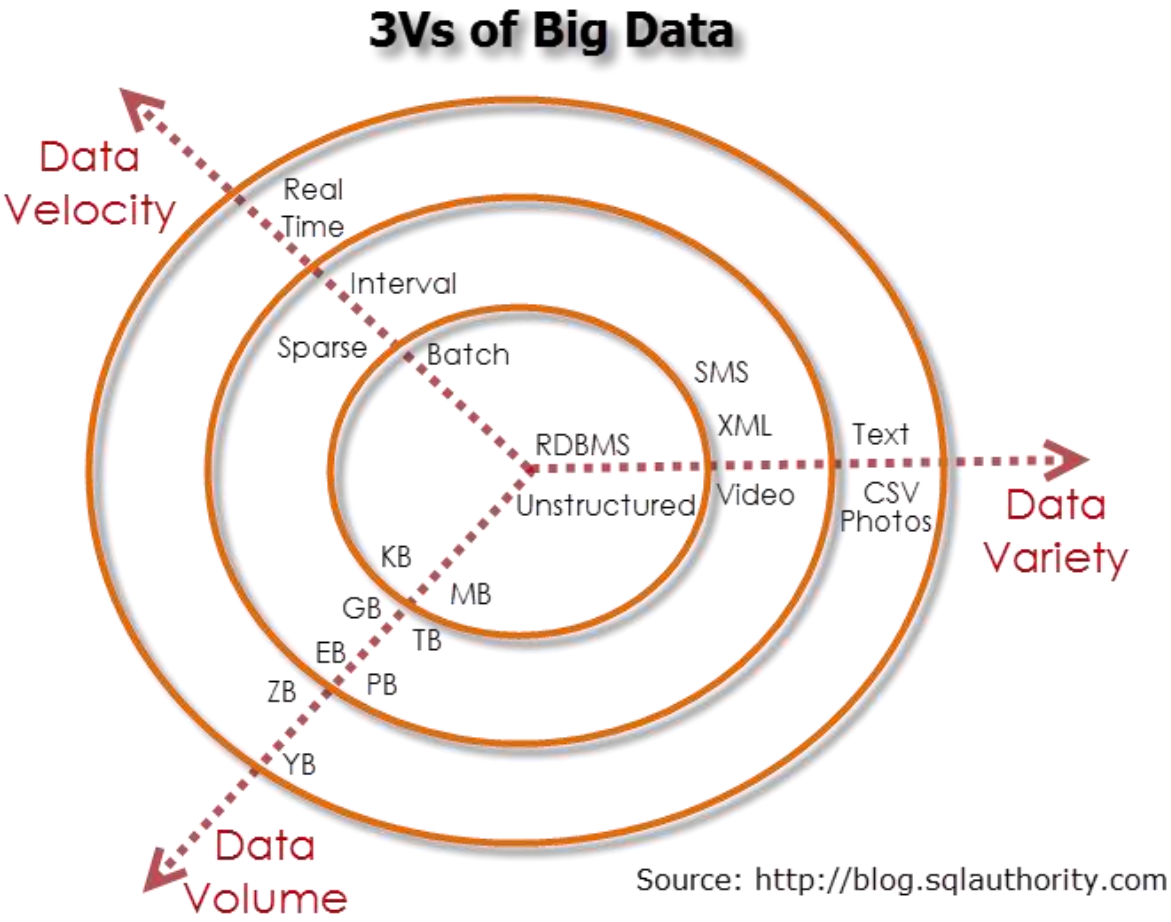
O que é Big Data?



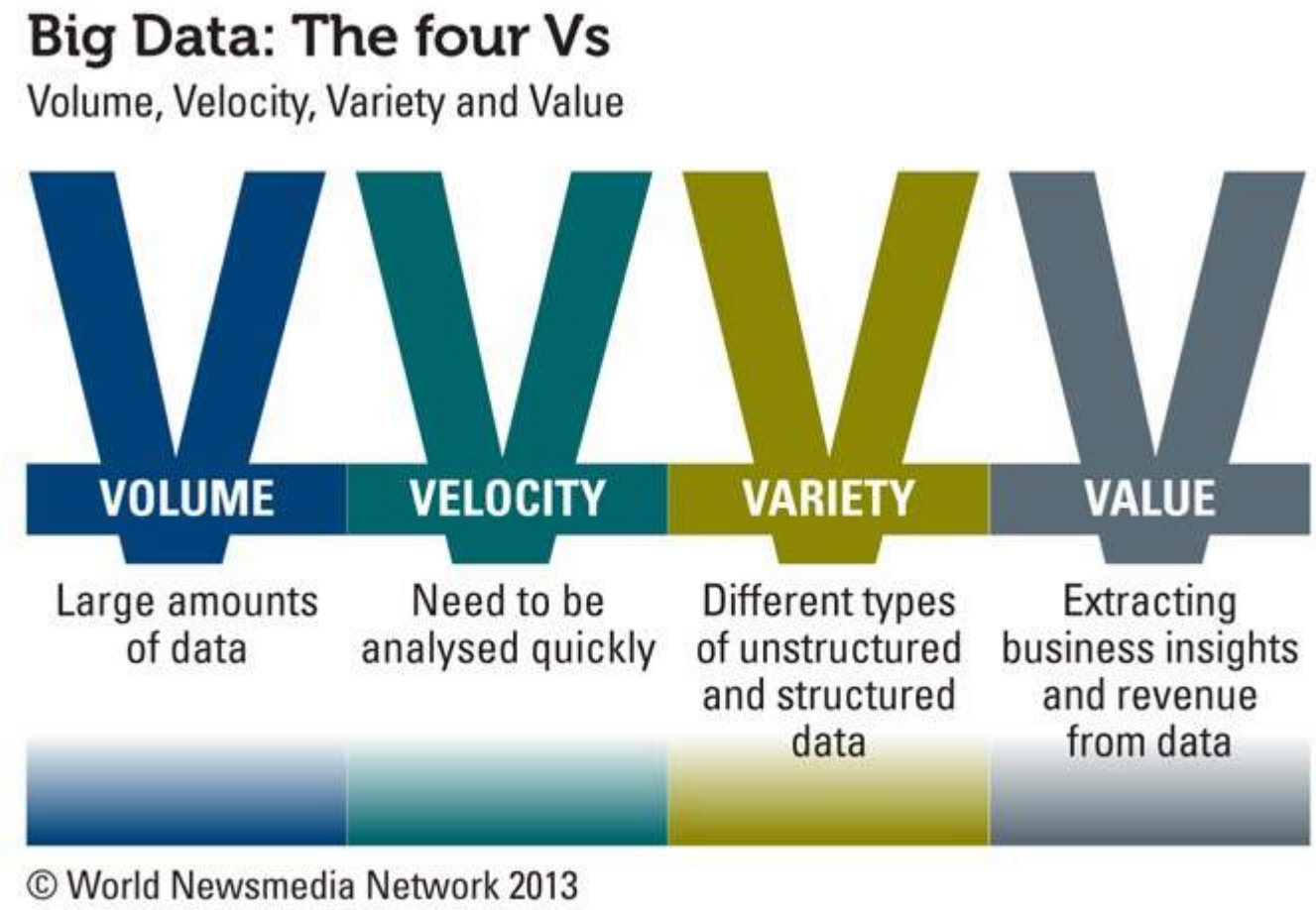
Every 60 seconds on Facebook: 510 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded.

(Source: The Social Skinny) The Implication: Again, there are a lot of engaged and active users, but also a huge amount of information competing for their attention, so quality and strategy on your part matter.

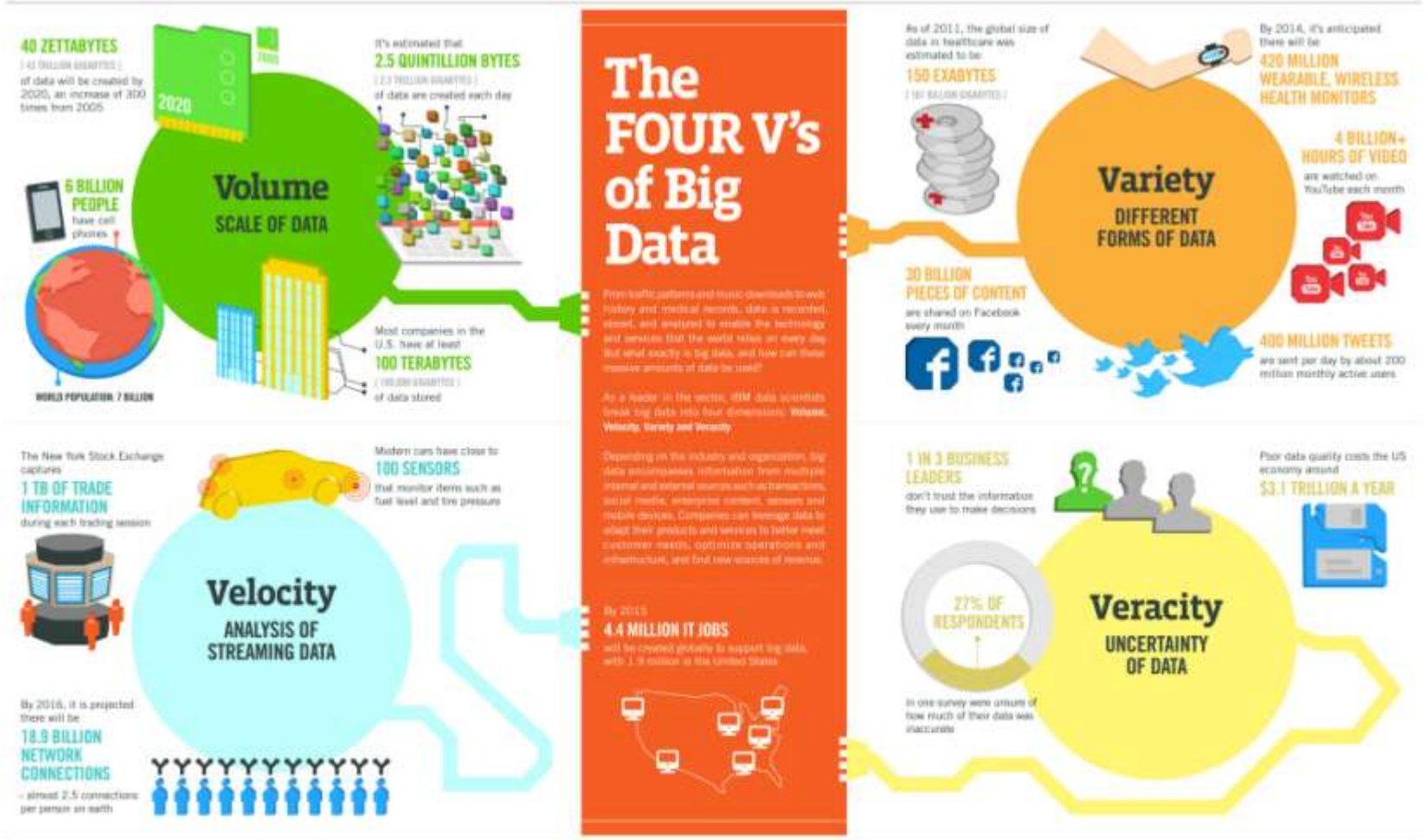
O que é Big Data?



O que é Big Data?



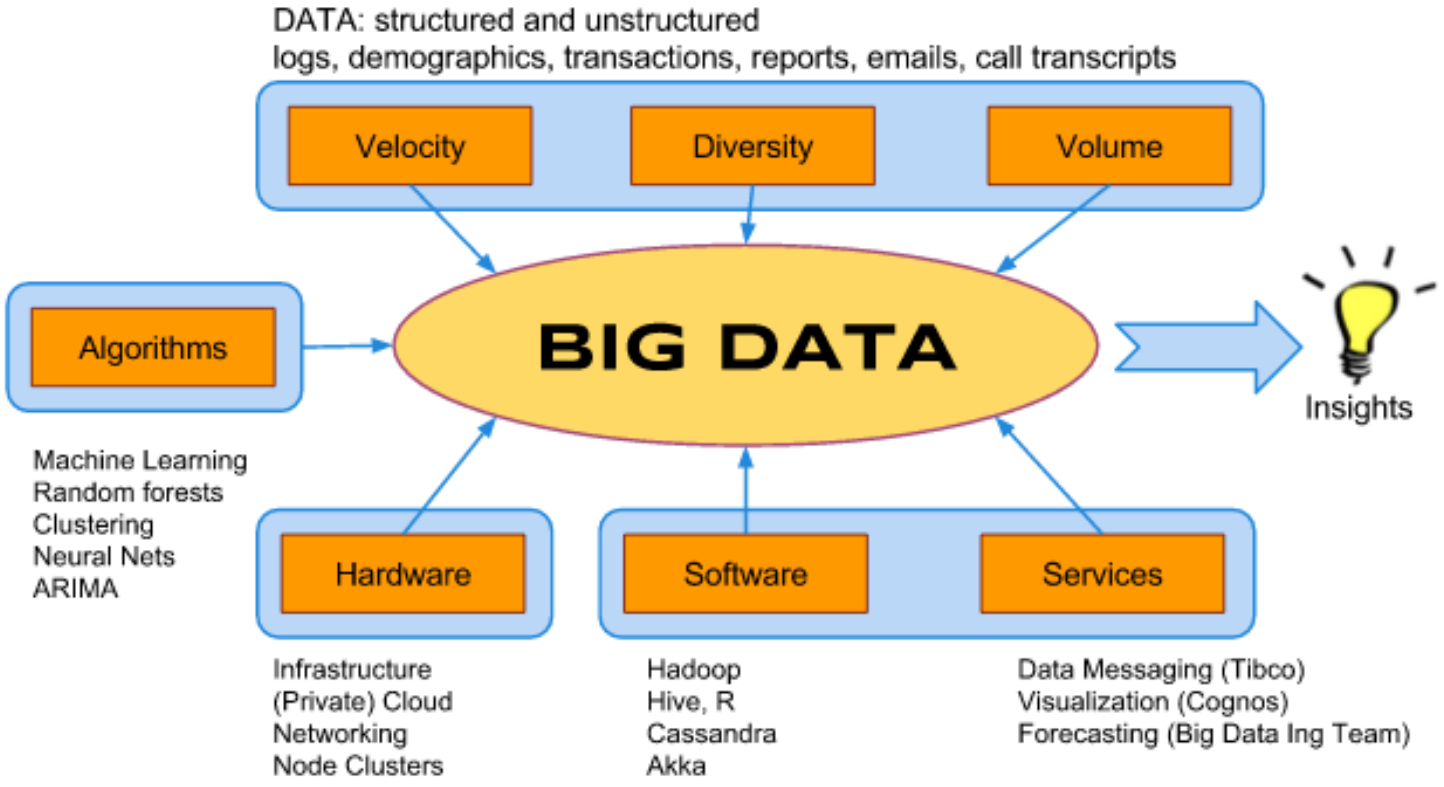
O que é Big Data?



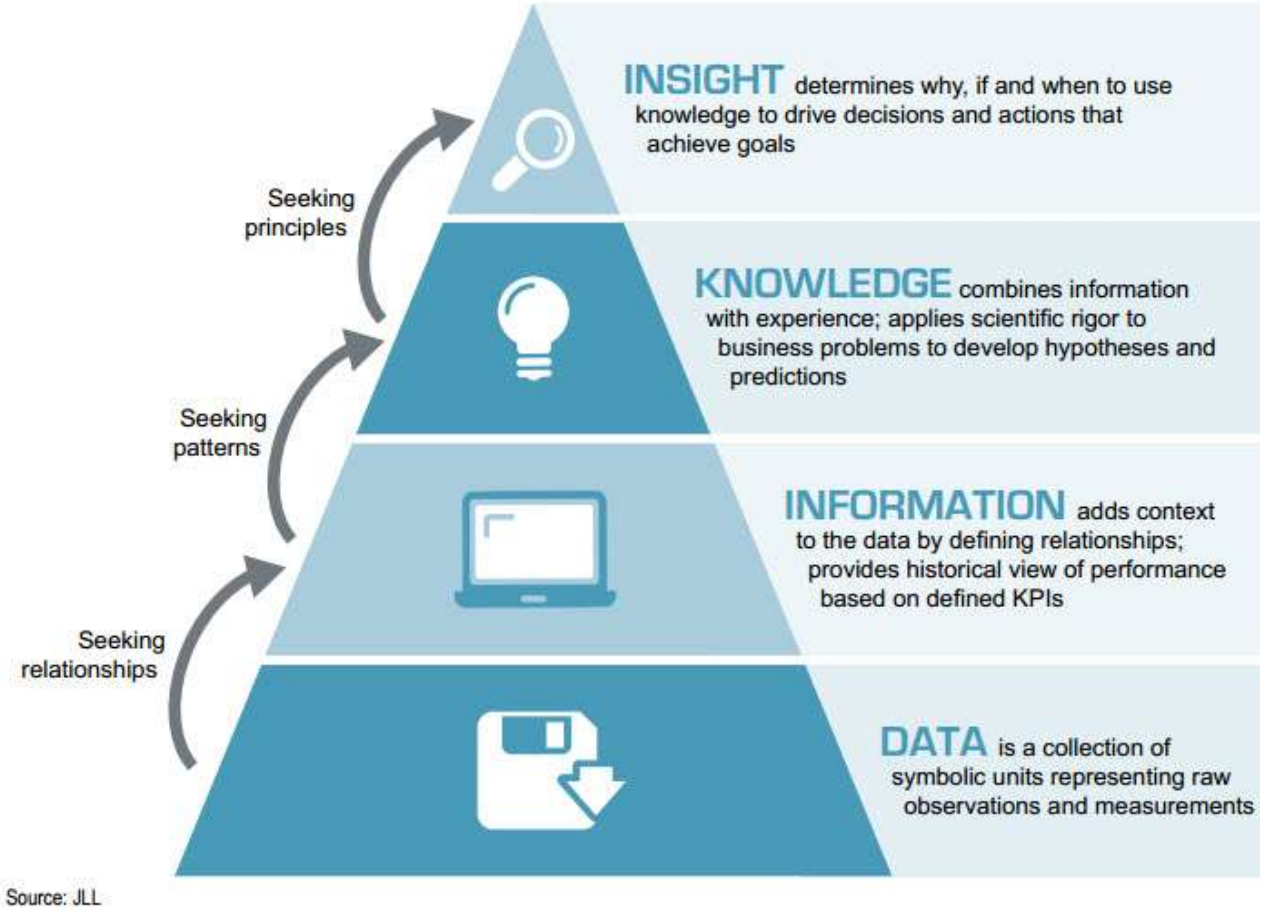
O que é Big Data?

Poucas tendências
tecnológicas ofereceram às empresas
tanto potencial de transformação.

O que é Big Data?



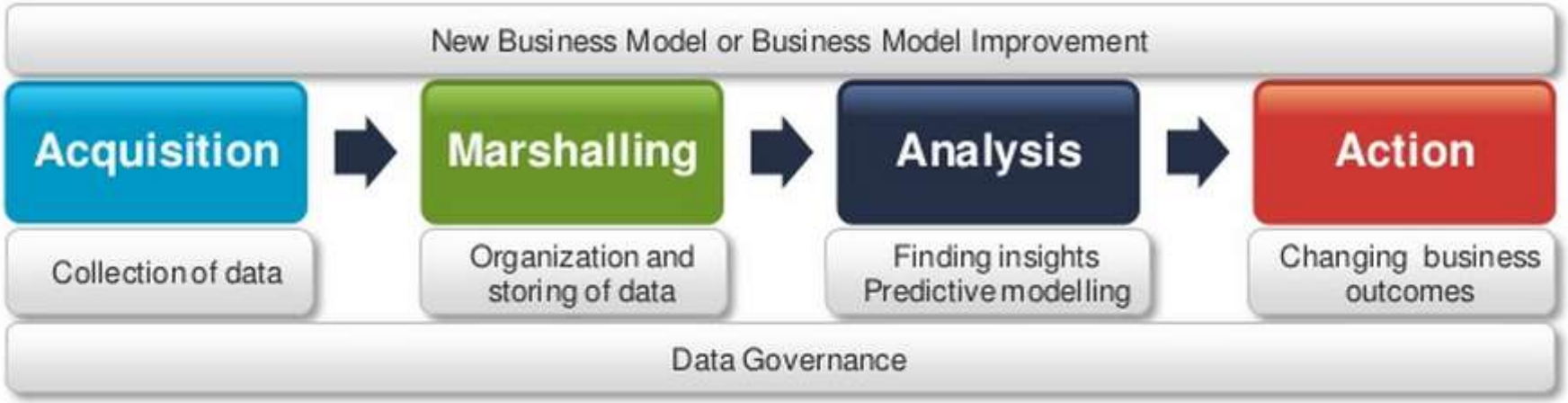
O que é Big Data?



O que é Big Data?

We have developed a Big Data strategy, methodology and delivery capability to help clients take advantage of big data:

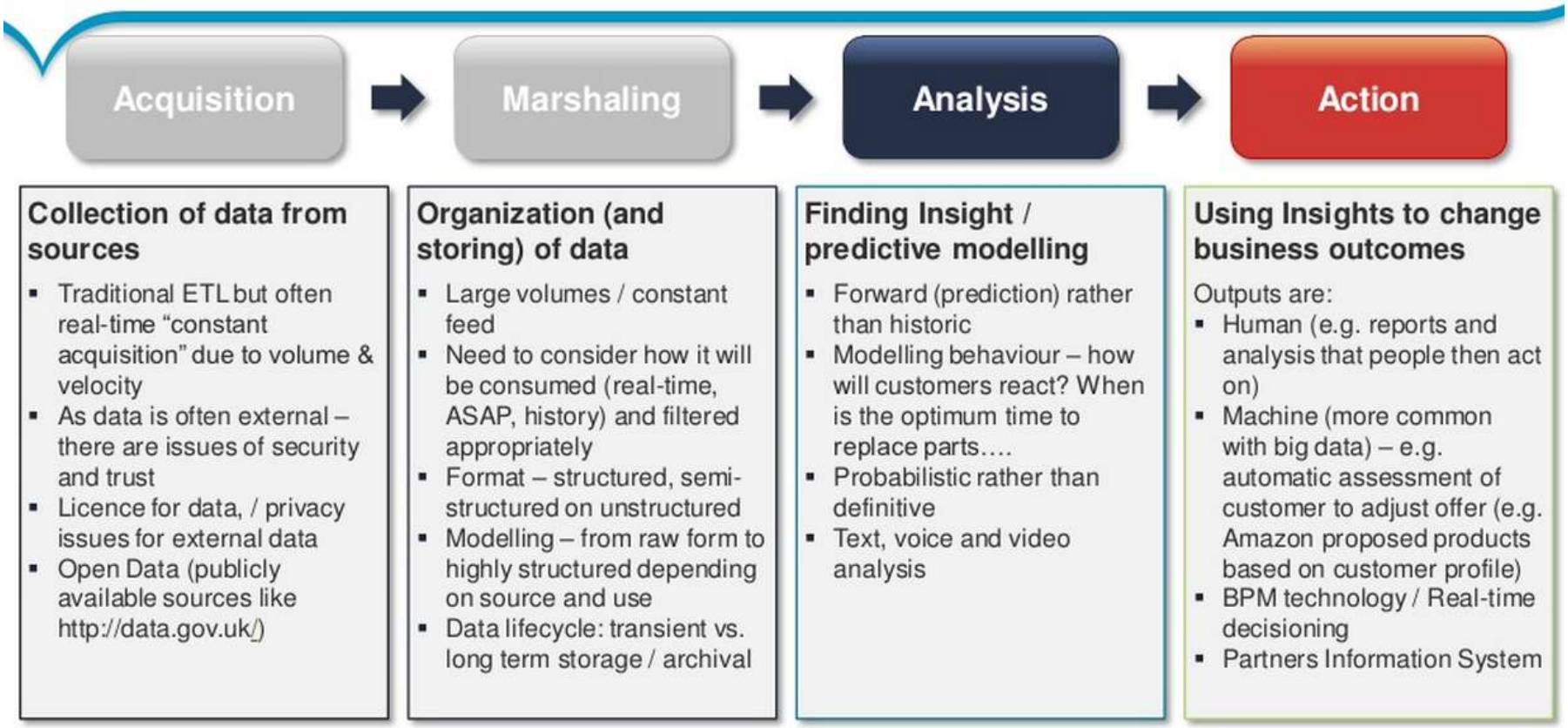
- **Big Data Process Model**



http://pt.slideshare.net/capgemini/understanding-big-data-14598849

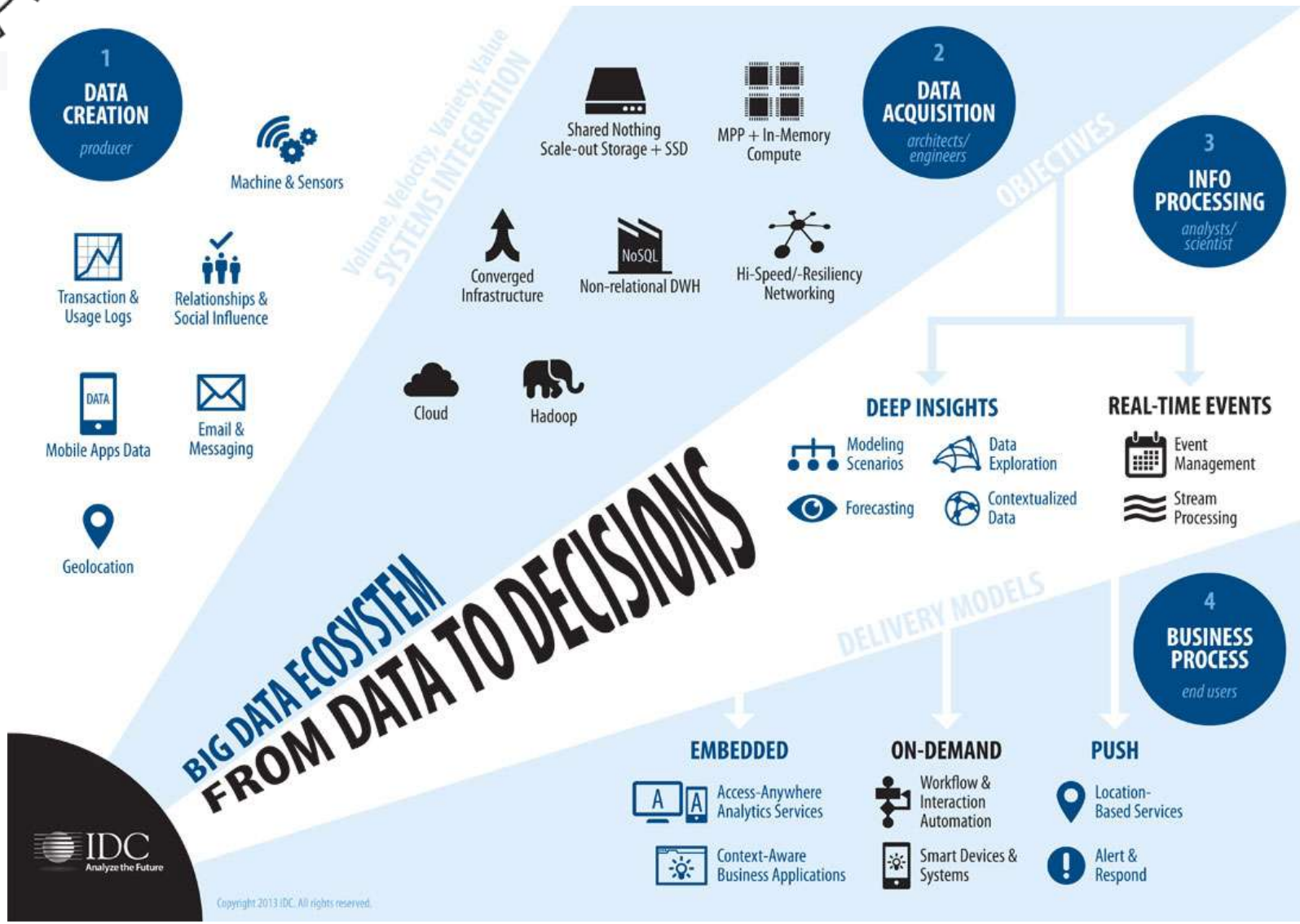
O que é Big Data?

Capgemini Big Data Process Model



http://pt.slideshare.net/capgemini/understanding-big-data-14598849

O que é Big Data?

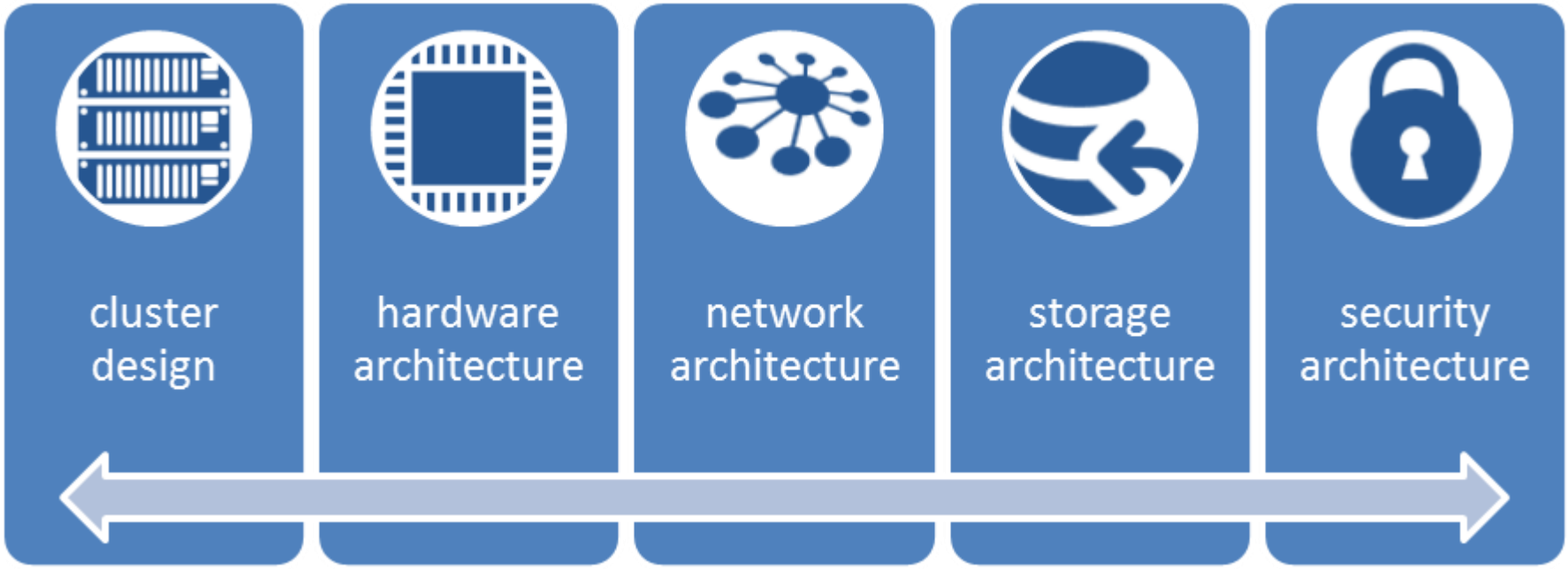


O que é Big Data?

- ✓ Aquisição

O que é Big Data?

- ✓ Armazenamento



O que é Big Data?

- ✓ Análise

- Google x H1N1
- Farecast

O que é Big Data?

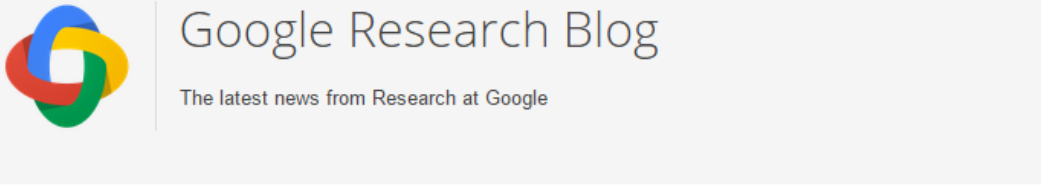
- Google x H1N1
  - Estatísticas de disseminação da doença baseadas nas pesquisas realizadas no Google
  - Dados sanitários oficiais eram atrasados em, no mínimo, 2 semanas
  - Google conseguia rodar estatísticas em tempo real
    - Rodou 450 milhões de modelos matemáticos diferentes a fim de testar os termos de busca, comparando com os casos reais de gripe registrados pelo CDC em 2007 e 2008.
    - O programa descobriu uma combinação de 45 termos de busca que, quando usados juntos num modelo matemático, tinham forte correlação entre a previsão e os números oficiais.

O que é Big Data?

- Google x H1N1
  - *“Como o CDC, eles podiam ver por onde a gripe havia se espalhado, mas, ao contrário do CDC, podiam apontar a disseminação quase em tempo real, e não com uma ou duas semanas de atraso.”*

O que é Big Data?

https://www.google.org/flutrends/



The Next Chapter for Flu Trends

Posted: Thursday, August 20, 2015

Posted by The Flu Trends Team

When a small team of software engineers first started working on Flu Trends in 2008, we wanted to explore how real-world phenomena could be modeled using patterns in search queries. Since its launch, Google Flu Trends has provided useful insights and served as one of the early examples for "nowcasting" based on search trends, which is increasingly used in health, economics, and other fields. Over time, we've used search signals to create prediction models, updating and improving those models over time as we compared our prediction to real-world cases of flu.

Instead of maintaining our own website going forward, we're now going to empower institutions who specialize in infectious disease research to use the data to build their own models. Starting this season, we'll provide Flu and Dengue signal data directly to partners including Columbia University's Mailman School of Public Health (to update their dashboard), Boston Children's Hospital/Harvard, and Centers for Disease Control and Prevention (CDC) Influenza Division. We will also continue to make historical Flu and Dengue estimate data available for anyone to see and analyze.

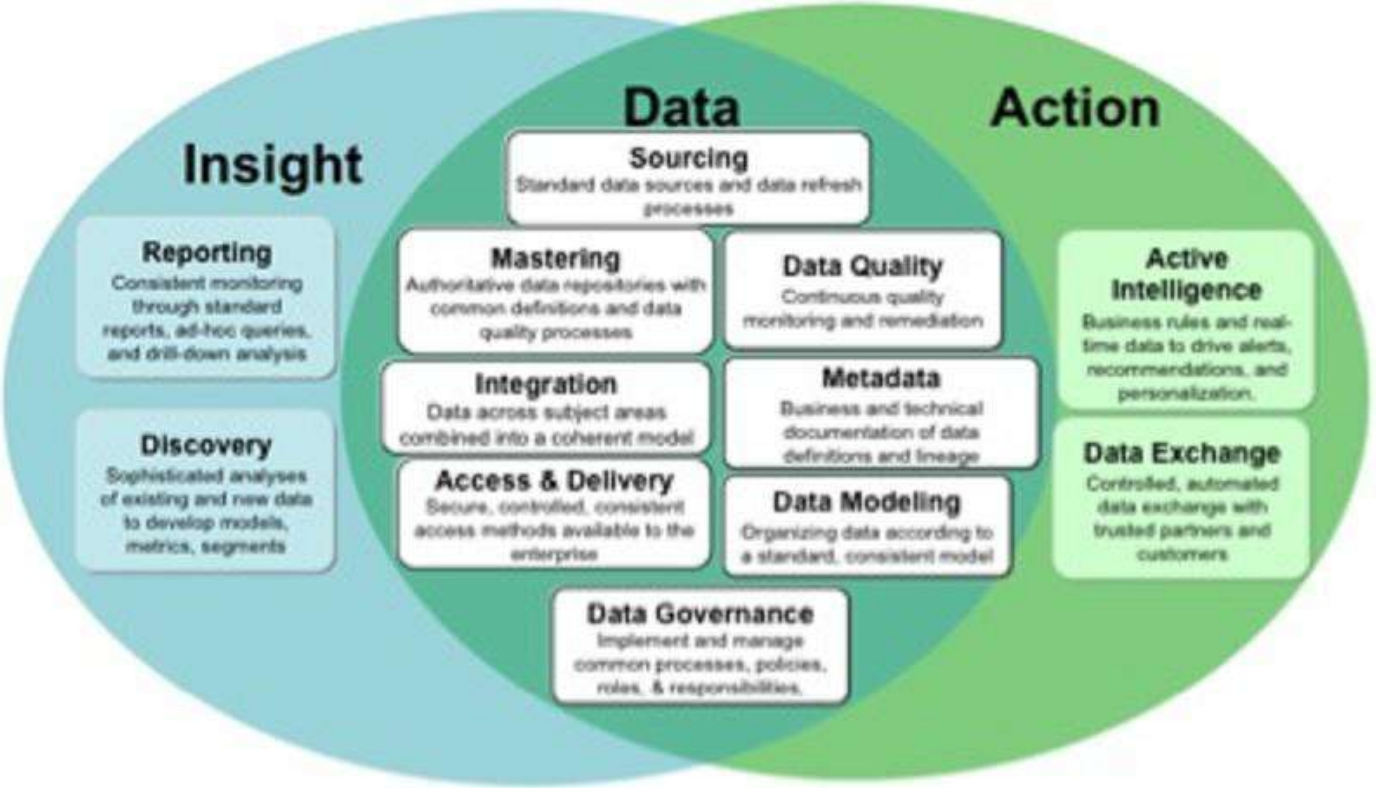
Flu continues to affect millions of people every year, and while it's still early days for nowcasting and similar tools for understanding the spread of diseases like flu and dengue fever—we're excited to see what comes next. To download the historical data or learn more about becoming a research partner, please visit the Flu Trends web page.

O que é Big Data?

- Farecast
  - Em 2003, Oren Etzioni desenvolveu um sistema que previa se e quando o preço de uma passagem aérea aumentaria ou diminuiria.
  - Comprada pela Microsoft e integrada ao sistema de buscas Bing
  - Em 2012, o sistema acertava 75% das previsões e os passageiros economizavam, em média, US$50 por passagem.

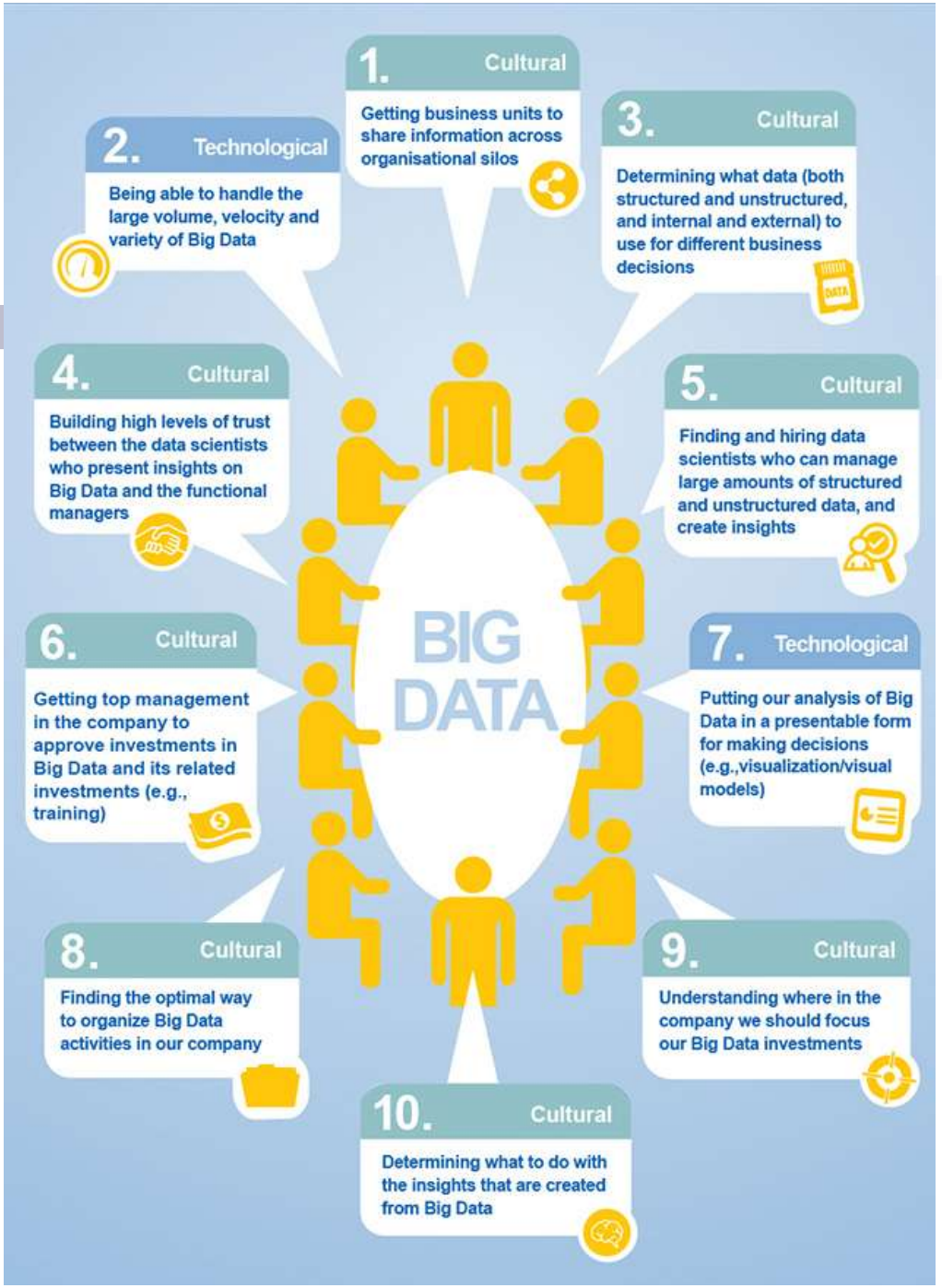
O que é Big Data?

- ✓ Tomada de Decisão



O que é Big Data?

We asked what are the top 10 greatest challenges preventing businesses from capitalizing on Big Data



O que é Big Data?

O Big Data desafia a maneira como vivemos e interagimos com o mundo. Mais importante, a sociedade precisará conter um pouco da obsessão pela causalidade e trocá-la por correlações simples : sem saber o *porquê*, apenas *o quê*.

O que é Big Data?

- O predomínio do big data representa três mudanças na forma como analisamos informações que transformam a maneira como entendemos e organizamos a sociedade:
  - Usar o conjunto total de dados, e não amostragem
  - Maior % de erros
  - Afastamento da busca pela causalidade

O que é Big Data?

No mundo do big data, por sua vez, não temos de nos fixar na causalidade : podemos descobrir padrões e correlações nos dados que podem não nos dizer com exatidão **por que** algo está acontecendo, mas nos alertam que **algo** está acontecendo.

UNIVERSIDADE FEDERAL DE ITAJUBÁ

BIG DATA E TECNOLOGIAS ASSOCIADAS

Tecnologias Associadas ao Big Data

- As aplicações “Big Data” fazem da computação o mecanismo para criar soluções capazes de analisar grandes bases de dados, processar seus pesados cálculos, identificar comportamentos e disponibilizar serviços especializados em seus domínios, porém, quase sempre esbarram no **poder computacional** das máquinas atuais.
  - a **computação paralela e distribuída** acena como alternativa para amenizar alguns dos grandes desafios computacionais.
  - essa computação é normalmente realizada em aglomerados (***clusters***) e grades computacionais, que com um conjunto de computadores comuns, conseguem agregar alto poder de processamento a um custo associado relativamente baixo.

Tecnologias Associadas ao Big Data

- Problemas do sistema distribuído
  - Dividir uma tarefa em sub-tarefas e então executá-las paralelamente em diversas unidades de processamento não é algo trivial.
    - extrair a dependência entre os dados da aplicação;
    - determinar um algoritmo de balanceamento de carga e de escalonamento para as tarefas, para garantir a eficiência do uso dos recursos computacionais;
    - garantir a recuperação ou a não interrupção da execução da aplicação caso uma máquina falhe.

Tecnologias Associadas ao Big Data

- *MapReduce*
  - Paradigma de programação para gerenciar grandes quantidades de dados (mais que 1TB), popularizado pelo Google em 2004.
  - O principal interesse desse paradigma é que as duas primitivas *Map* e *Reduce*, sejam facilmente paralelizadas e capazes de lidar com a grande quantidade de dados.
  - Todos os programas desenvolvidos sobre esse paradigma realizam o processamento paralelo de conjuntos de dados e podem, portanto, ser executados em servidores simples, sem muito esforço.

Tecnologias Associadas ao Big Data

- *MapReduce*
  - A primitiva *Map* consiste no processamento de uma lista de dados, a fim de criar pares chave/valor.
  - A primitiva *Reduce* irá processar cada par, a fim de criar um novo agrupamento chave/valor. Em geral, a primitiva Reduce utiliza alguma função de agregação (SUM, AVG, etc).

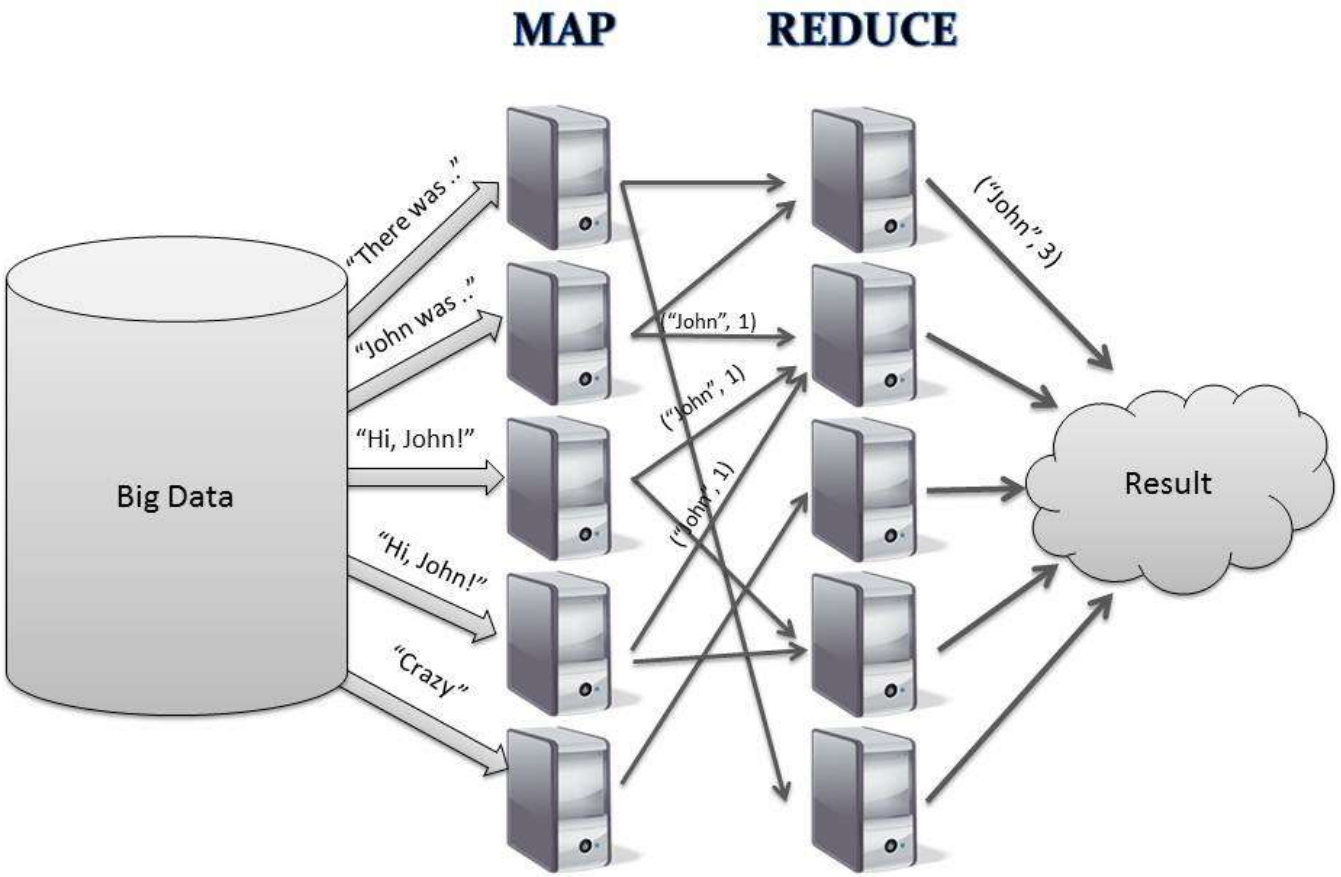
Tecnologias Associadas ao Big Data

- *MapReduce*

$$
\begin{aligned}
List &: (a;2)(a;4)(b;4)(c;5)(b;2)(a;1) \\
After\ mapping &: (a;[2,4,1]),(b;[4,2]),(c[5]) \\
After\ reducing &: (a;7),(b;6),(c;5)
\end{aligned}
$$

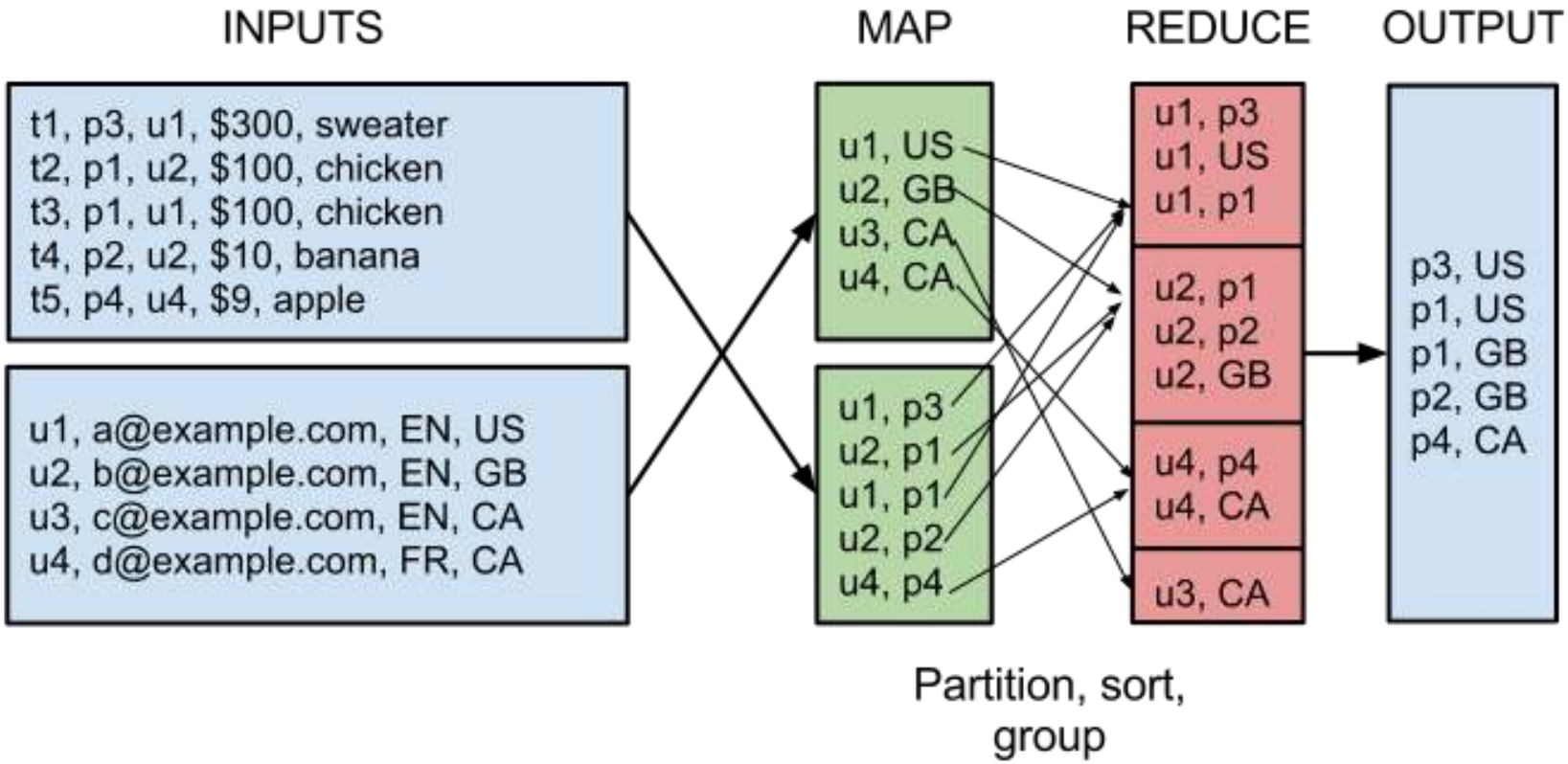
Tecnologias Associadas ao Big Data

- *MapReduce*



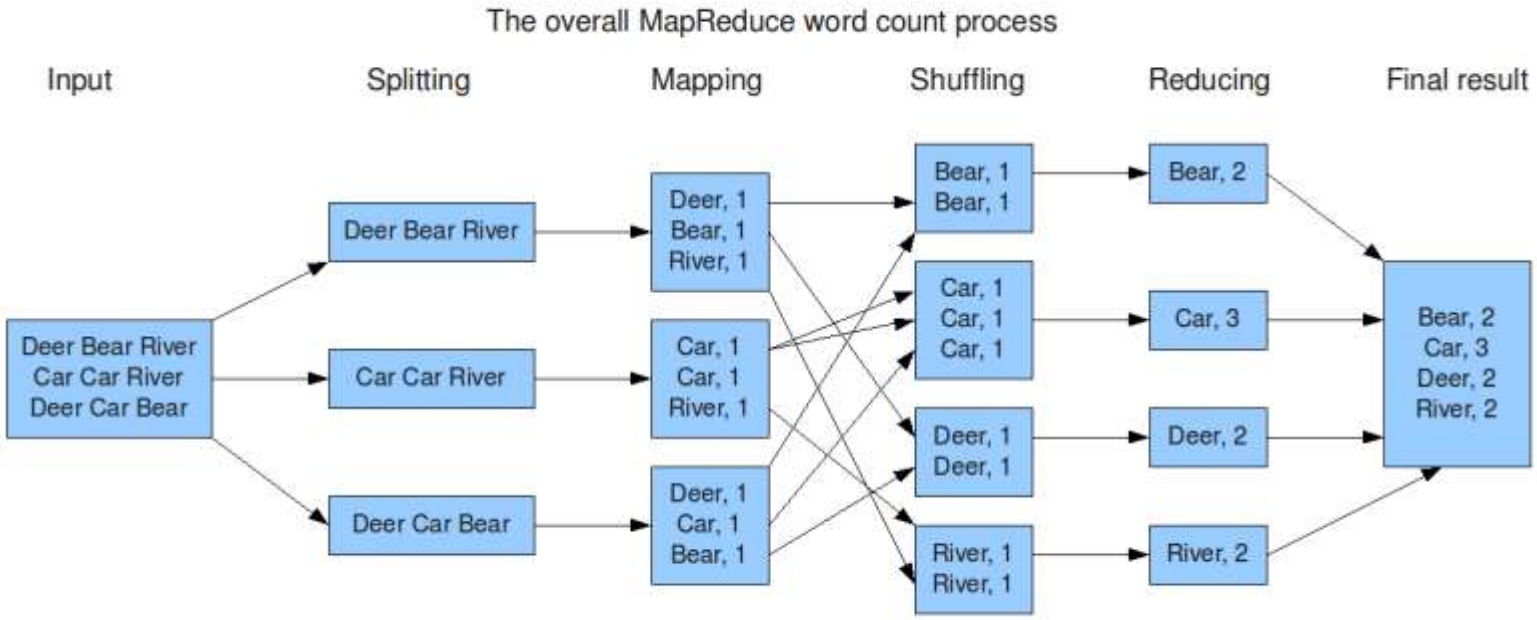
Tecnologias Associadas ao Big Data

- *MapReduce*



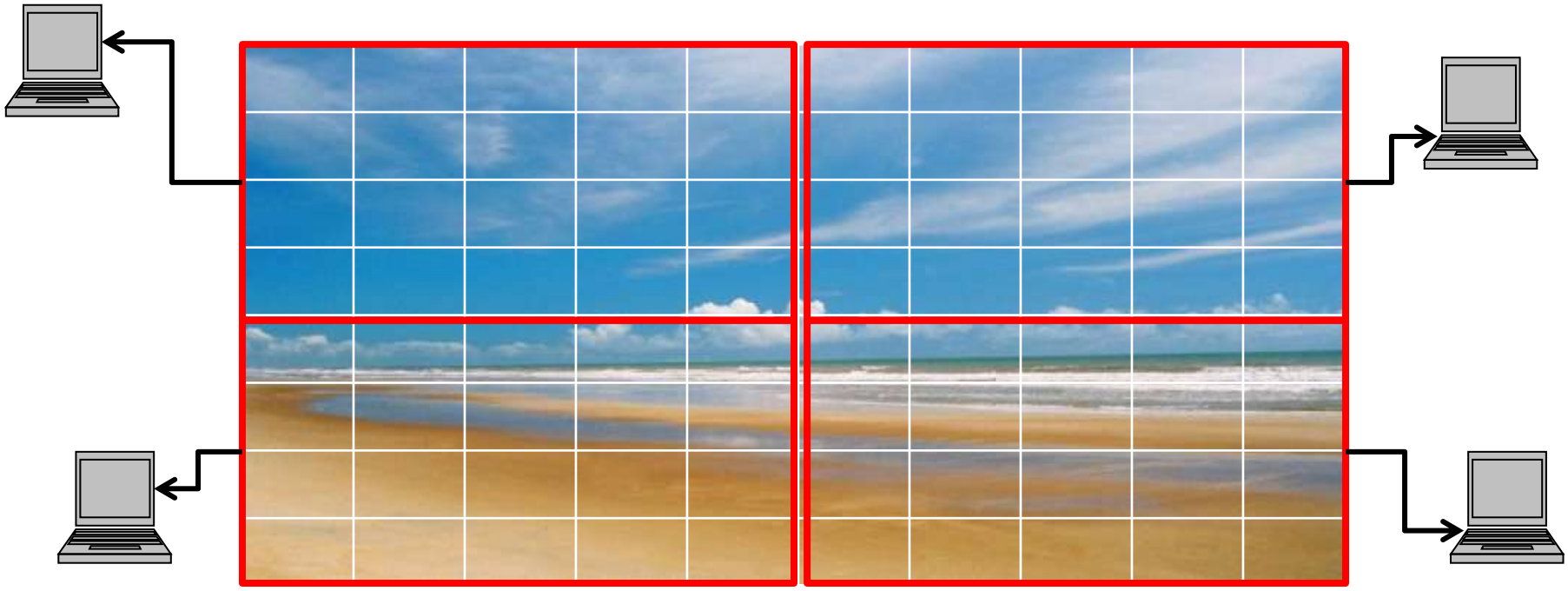
Tecnologias Associadas ao Big Data

- *MapReduce*



Tecnologias Associadas ao Big Data

- *MapReduce*

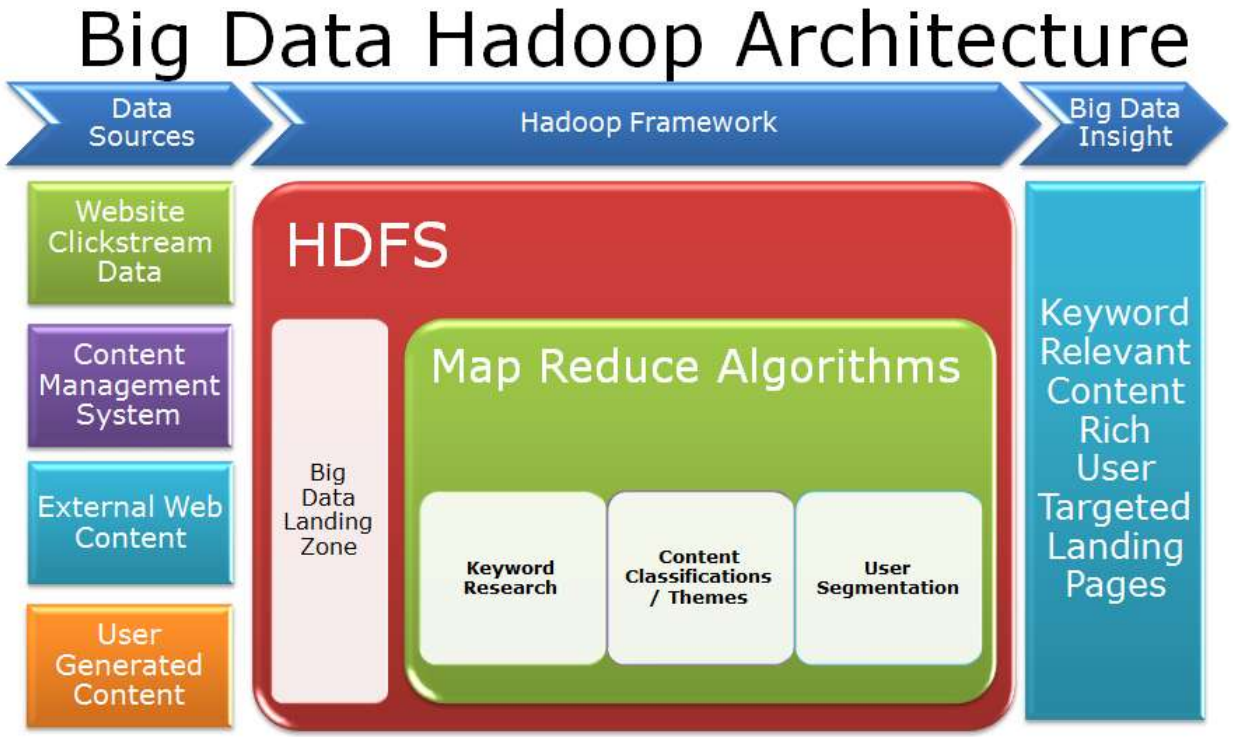


Tecnologias Associadas ao Big Data

- Hadoop
  - Uma das implementações mais famosas do MapReduce
  - *Framework* para o processamento de grandes quantidades de dados em aglomerados e grades computacionais.
  - Promover soluções para os desafios dos sistemas distribuídos é o ponto central do projeto Hadoop.
  - Seu modelo de programação e sistema de armazenamento dos dados promovem um rápido processamento, muito superior às outras tecnologias similares.

Tecnologias Associadas ao Big Data

- O core do Hadoop tem 2 sistemas principais:
  - *Hadoop Distributed File System* (HDFS)
  - *MapReduce*



Tecnologias Associadas ao Big Data

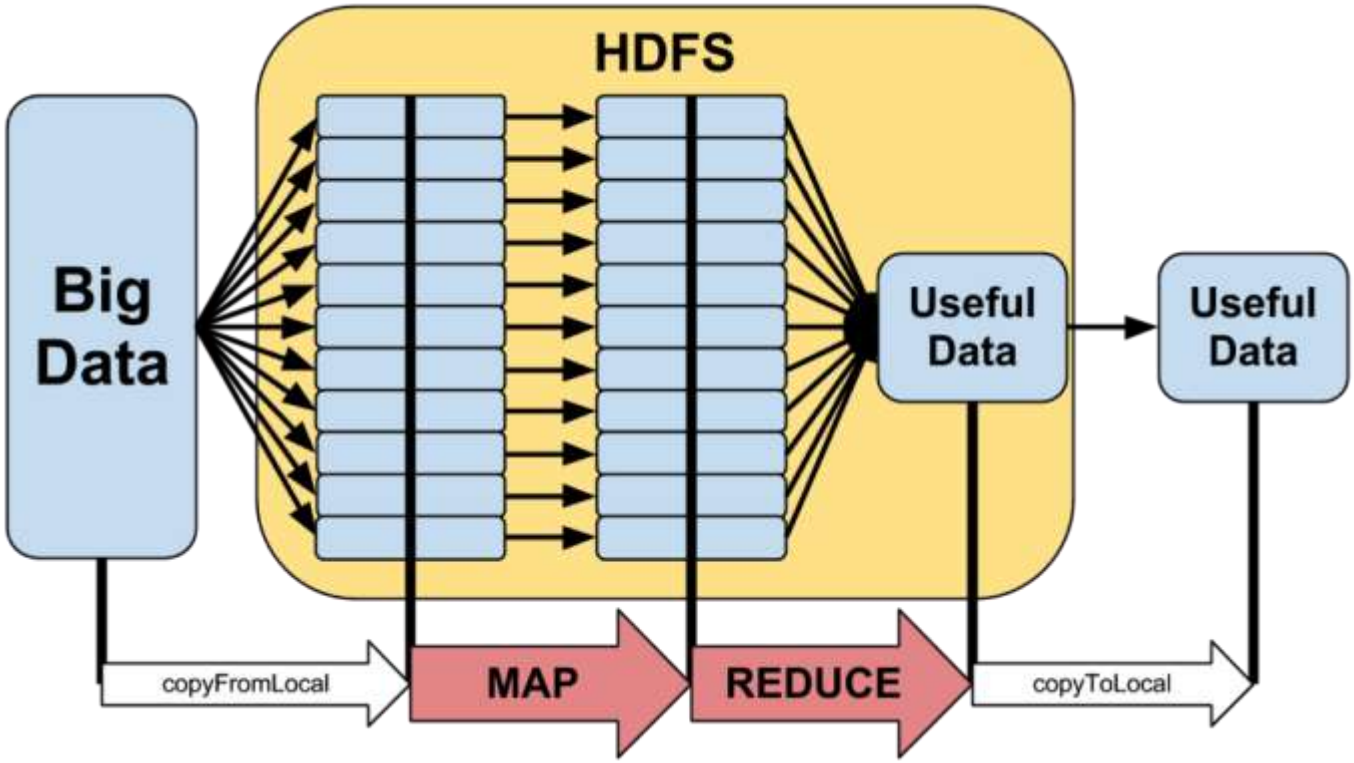
- *Hadoop Distributed File System* (HDFS)
  - Sistema de arquivos distribuído, projetado para armazenar arquivos muito grandes, com padrão de acesso aos dados streaming, utilizando clusters de servidores simples.
  - Não deve ser usado para aplicações que precisem de acesso rápido a um determinado registro, mas sim para aplicações nas quis é necessário ler uma quantidade muito grande de dados.

Tecnologias Associadas ao Big Data

- *Hadoop Distributed File System* (HDFS)
  - Sistema de arquivos distribuído, projetado para armazenar arquivos muito grandes, com padrão de acesso aos dados streaming, utilizando clusters de servidores simples.
  - Não deve ser usado para aplicações que precisem de acesso rápido a um determinado registro, mas sim para aplicações nas quais é necessário ler uma quantidade muito grande de dados.

Tecnologias Associadas ao Big Data

- *Hadoop Distributed File System* (HDFS)



Tecnologias Associadas ao Big Data

- Sistemas Distribuídos
- Computação em Nuvem
- Virtualização
- Armazenamento