Big Data

Desafios e Oportunidades na Análise Massiva de Dados

Eduardo Tenório - embat@cin.ufpe.br

Sumário

- O que é Big Data?
- E o que não é Big Data
- Os 5 V's
- Arquitetura
- Desafios e Oportunidades
- Existe um "Big Market"?
- Críticas

O que é Big Data?

"Big Data é um termo popular usado para descrever o crescimento, a disponibilidade e o uso exponencial de informações estruturadas e não estruturadas."

SAS
(Statistical Analysis System)

O que é Big Data?

"This is a next big thing in computing and generate value from very large datasets that cannot be analysed with traditional computing techniques."

Christopher Barnatt
(Nottingham University Business School)

O que é Big Data?

"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it..."

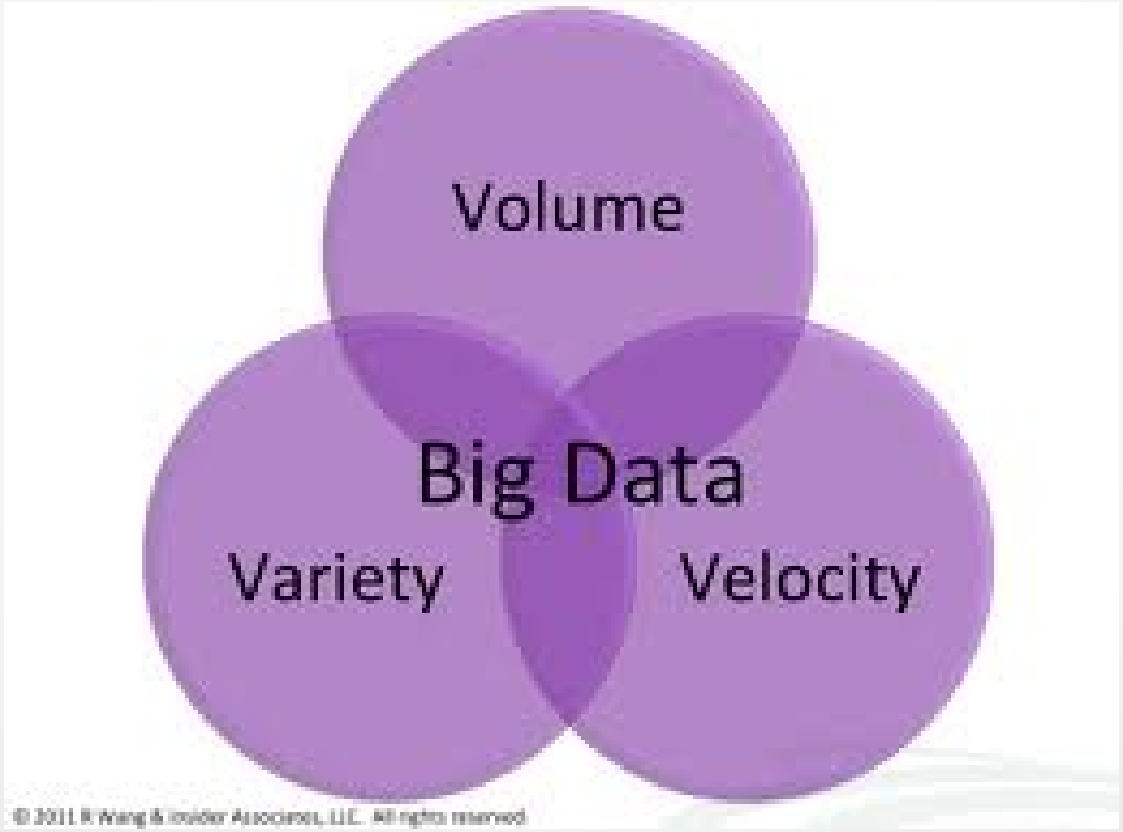
Dan Ariely
(Duke University)

O que é Big Data?

Basicamente, Big Data consiste na **análise estatística** de grandes coleções de dados (**volume**), disponíveis em diversos formatos (**variedade**) e que respondam em tempo quase real (**velocidade**).

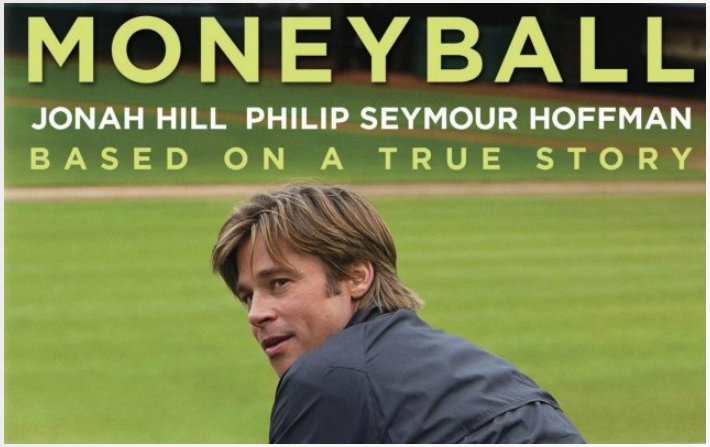
São os "tradicionais" 3 V's do Big Data.

O que é Big Data?

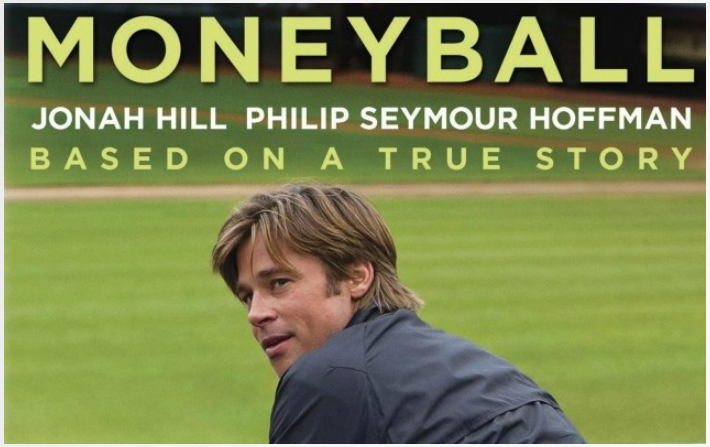


Eduardo Tenório - embat@cin.ufpe.br

O que é Big Data?

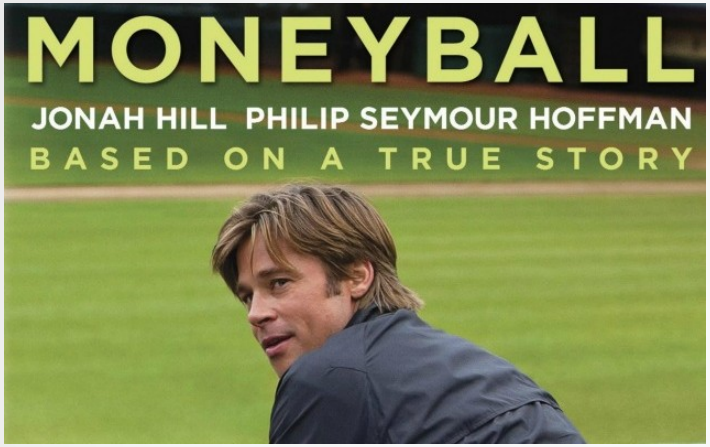


O que é Big Data?



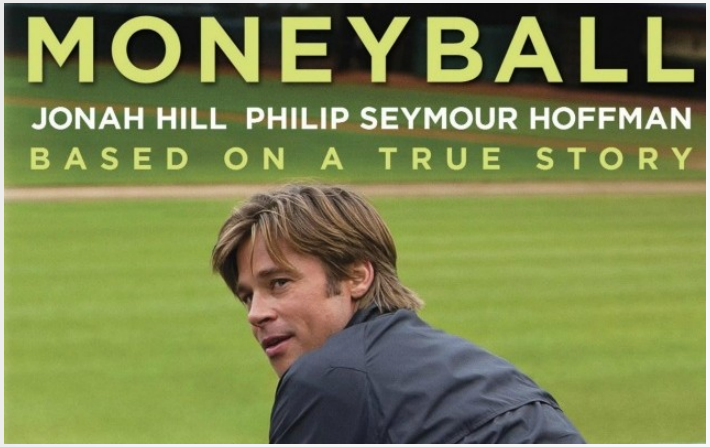
- Como montar um time competitivo a partir de estatísticas?

O que é Big Data?



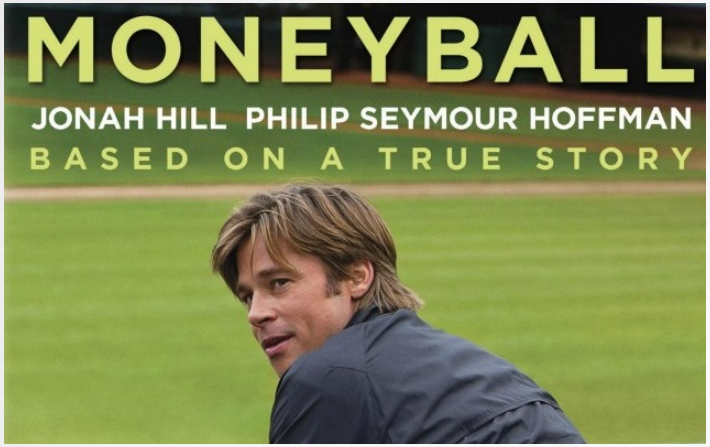
- Como montar um time competitivo a partir de estatísticas?
- Baseado numa história real

O que é Big Data?



- Como montar um time competitivo a partir de estatísticas?
- Baseado numa história real
- Não é bem Big Data...

O que é Big Data?



- Como montar um time competitivo a partir de estatísticas?
- Baseado numa história real
- Não é bem Big Data... mas é um começo!

O que é Big Data?

Eduardo Tenório - embat@cin.ufpe.br

O que é Big Data?

- Já usava Big Data antes de se chamar "Big Data"

O que é Big Data?

- Já usava Big Data antes de se chamar "Big Data"
- Expandiu o cluster Hadoop de 10 para 250 nós (2012)

O que é Big Data?

- Já usava Big Data antes de se chamar "Big Data"
- Expandiu o cluster Hadoop de 10 para 250 nós (2012)
- Database com 2,5 PB

O que é Big Data?

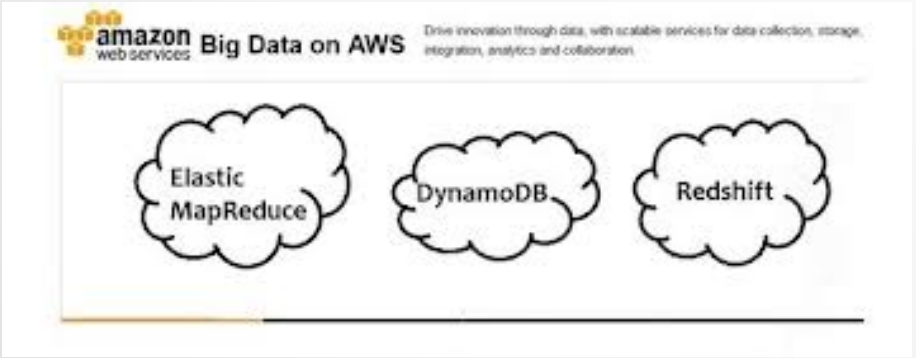
- Já usava Big Data antes de se chamar "Big Data"
- Expandiu o cluster Hadoop de 10 para 250 nós (2012)
- Database com 2,5 PB
- 1 milhão de transações de clientes a cada hora

O que é Big Data?

- Já usava Big Data antes de se chamar "Big Data"
- Expandiu o cluster Hadoop de 10 para 250 nós (2012)
- Database com 2,5 PB

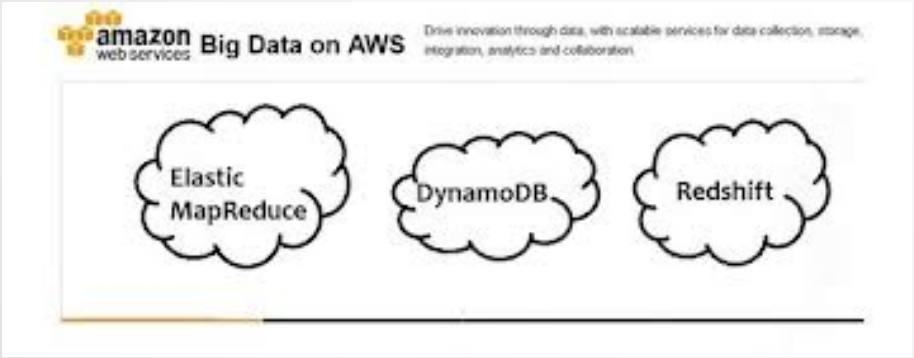
- 1 milhão de transações de clientes a cada hora
- Análise de 100 milhões de palavras-chave para otimizar ofertas, utilizando uma base diária

O que é Big Data?



O que é Big Data?

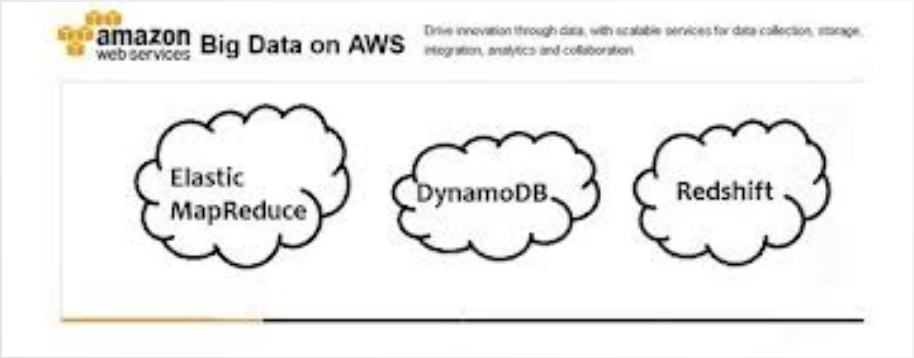
- Milhões de operações/dia



O que é Big Data?

amazon.com

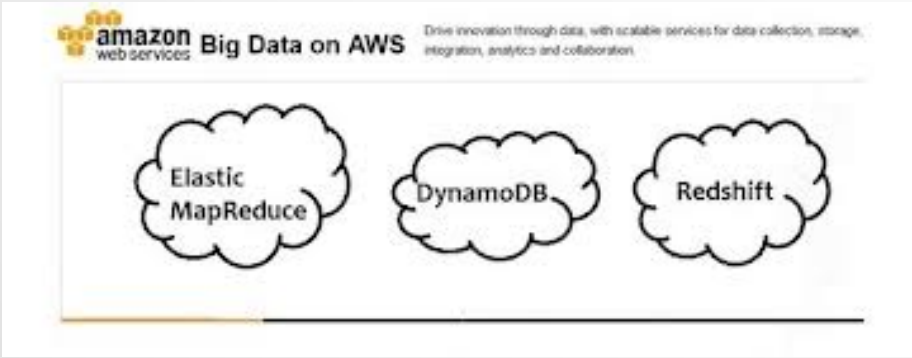
- Milhões de operações/dia
- ½ milhão de consultas para produtos de terceiros



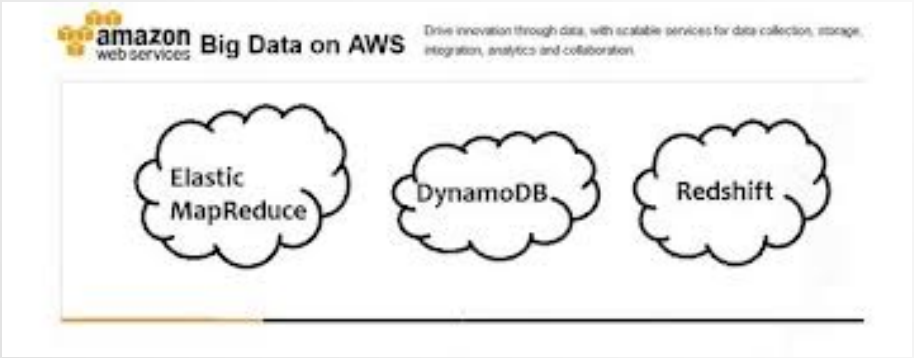
O que é Big Data?

amazon.com

- Milhões de operações/dia
- ½ milhão de consultas para produtos de terceiros
- Os 3 maiores databases Linux do mundo

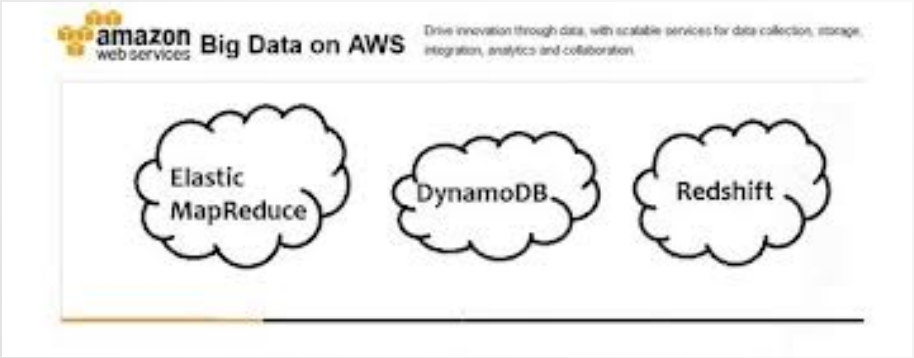


O que é Big Data?



- Milhões de operações/dia
- ½ milhão de consultas para produtos de terceiros
- Os 3 maiores databases Linux do mundo
- Utiliza Hadoop como base do serviço

O que é Big Data?



- Milhões de operações/dia
- ½ milhão de consultas para produtos de terceiros
- Os 3 maiores databases Linux do mundo
- Utiliza Hadoop como base do serviço
- Big Data como serviço

E o que não é Big Data

- Big Data é uma questão de volume

E o que não é Big Data

- Big Data é uma questão de volume
  - "Big" é relativo

E o que não é Big Data

- Big Data é uma questão de volume
  - "Big" é relativo
  - Um computador pessoal pode armazenar TB atualmente

E o que não é Big Data

- Big Data é uma questão de volume
  - "Big" é relativo
  - Um computador pessoal pode armazenar TB atualmente
  - Jogadores de baseball possuem estatísticas padronizadas

E o que não é Big Data

- Big Data é uma questão de volume
  - "Big" é relativo
  - Um computador pessoal pode armazenar TB atualmente
  - Jogadores de baseball possuem estatísticas padronizadas
  - E a resposta pode demorar alguns dias...

E o que não é Big Data

- O enfoque é tecnológico

E o que não é Big Data

- O enfoque é tecnológico
  - Não, é "de negócios"!

E o que não é Big Data

- O enfoque é tecnológico
  - Não, é "de negócios"!
    - É sobre agir de forma inteligente, ter uma vantagem competitiva ao entender melhor o cliente

E o que não é Big Data

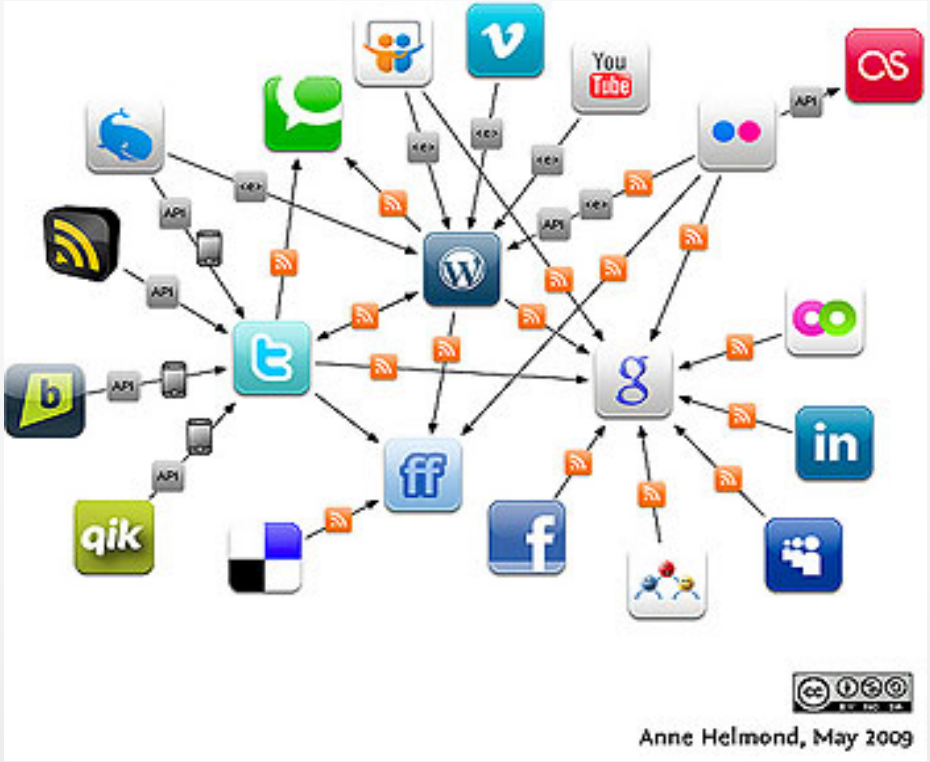
- O enfoque é tecnológico
  - Não, é "de negócios"!
    - É sobre agir de forma inteligente, ter uma vantagem competitiva ao entender melhor o cliente
    - A tecnologia não é algo novo...

E o que não é Big Data

- O enfoque é tecnológico
  - Não, é "de negócios"!
  - É sobre agir de forma inteligente, ter uma vantagem competitiva ao entender melhor o cliente
  - A tecnologia não é algo novo...
  - ... apenas está sendo entendida agora!

E o que não é Big Data

- Big Data é "social data"



E o que não é Big Data

- Big Data é "social data"
  - Se você trabalha na Google ou Facebook, social data é algo realmente grande!

E o que não é Big Data

- Big Data é "social data"
  - Se você trabalha na Google ou Facebook, social data é algo realmente grande!
  - Mas "likes" e "shares" não são ferramentas de análise

E o que não é Big Data

- Big Data é "social data"
  - Se você trabalha na Google ou Facebook, social data é algo realmente grande!
  - Mas "likes" e "shares" não são ferramentas de análise
  - Saber a tendência é o que importa, independente de ser o que é falado nas redes sociais, qual o comportamento esperado do mercado de ações ou como partículas colidem no LHC

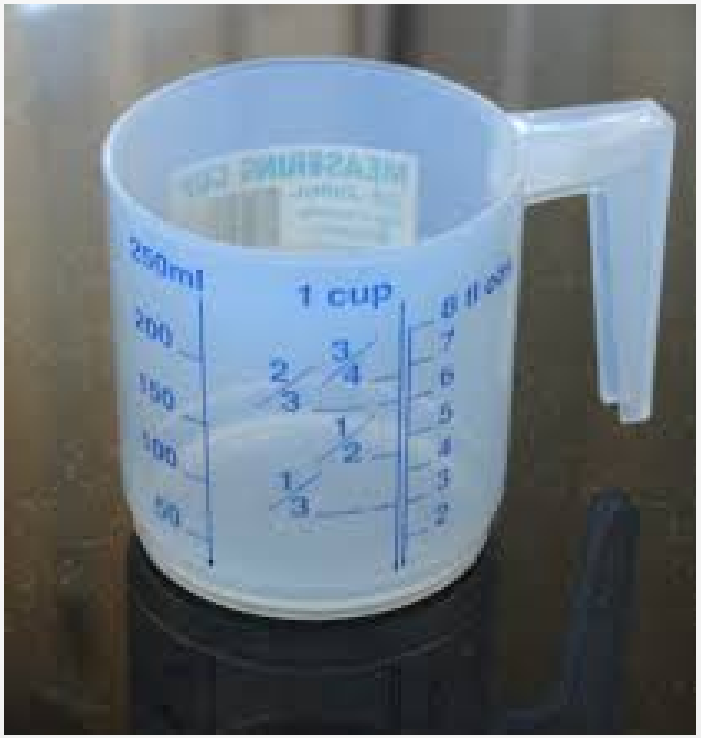
Os 5 V's

Além dos "tradicionais" 3 V's do Big Data: volume, variedade e velocidade;

Existem outras dimensões consideradas por alguns players: veracidade e valor

Os 5 V's

- Volume



Os 5 V's

- Volume
  - O custo de armazenamento está cada vez menor

Os 5 V's

- Volume
  - O custo de armazenamento está cada vez menor
  - O volume de dados gerado aumentou consideravelmente

Os 5 V's

- Volume
  - O custo de armazenamento está cada vez menor
  - O volume de dados gerado aumentou consideravelmente
  - Em 2008 foram produzidos mais dados do que a soma de todos os anos anteriores

Os 5 V's

- Volume
  - O custo de armazenamento está cada vez menor
  - O volume de dados gerado aumentou consideravelmente
  - Em 2008 foram produzidos mais dados do que a soma de todos os anos anteriores
  - E a taxa está aumentando

Os 5 V's

- Variedade

Os 5 V's

- Variedade
  - Databases relacionais, NoSQL, Data Warehouses

Os 5 V's

- Variedade
  - Databases relacionais, NoSQL, Data Warehouses
  - Arquivos de texto, som e vídeo

Os 5 V's

- Variedade
  - Databases relacionais, NoSQL, Data Warehouses
  - Arquivos de texto, som e vídeo
  - Medidores e sensores diversos

Os 5 V's

- Variedade
  - Databases relacionais, NoSQL, Data Warehouses
  - Arquivos de texto, som e vídeo
  - Medidores e sensores diversos
  - Transações financeiras

Os 5 V's

- Variedade
  - Databases relacionais, NoSQL, Data Warehouses
  - Arquivos de texto, som e vídeo
  - Medidores e sensores diversos
  - Transações financeiras
  - Segundo estimativas, 80% dos dados de uma organização não são numéricos

Os 5 V's

- Velocidade

Os 5 V's

- Velocidade
  - Quão rápido os dados são produzidos

Os 5 V's

- Velocidade
  - Quão rápido os dados são produzidos
  - E quão rápido são analisados

Os 5 V's

- Velocidade
  - Quão rápido os dados são produzidos
  - E quão rápido são analisados
  - Etiquetas RFID demandam respostas em tempo quase real

Os 5 V's

- Veracidade

Os 5 V's

- Veracidade
  - Separar o joio do trigo

Os 5 V's

- Veracidade
  - Separar o joio do trigo
  - O mais difícil dos 5 V's

Os 5 V's

- Veracidade
  - Separar o joio do trigo
  - O mais difícil dos 5 V's
  - Garbage in, garbage out

Os 5 V's

- Valor

Os 5 V's

- Valor
  - O dado está correto?

Os 5 V's

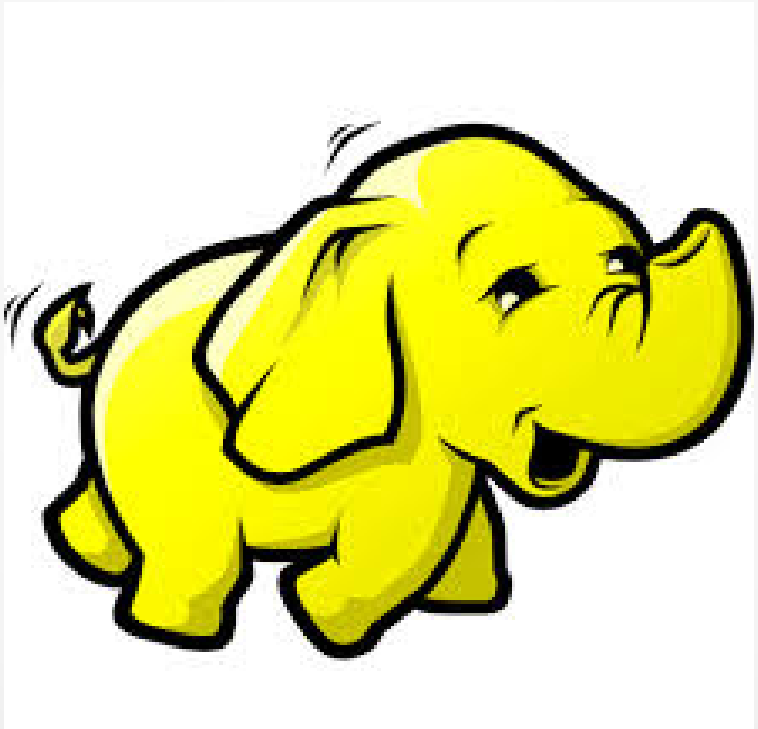
- Valor
  - O dado está correto?
  - Qual a sua acurácia?

Os 5 V's

- Valor
  - O dado está correto?
  - Qual a sua acurácia?
  - Possui um peso estatístico

Arquitetura

- MapReduce



Arquitetura

- MapReduce
  - Um novo paradigma de programação

Arquitetura

- MapReduce
  - Um novo paradigma de programação
  - Processa grandes conjuntos de dados

Arquitetura

- MapReduce
  - Um novo paradigma de programação
  - Processa grandes conjuntos de dados
  - Algoritmo paralelo e distribuído em clusters

Arquitetura

- MapReduce
  - Um novo paradigma de programação
  - Processa grandes conjuntos de dados
  - Algoritmo paralelo e distribuído em clusters
  - A função **Map()** filtra e distribui os dados no cluster

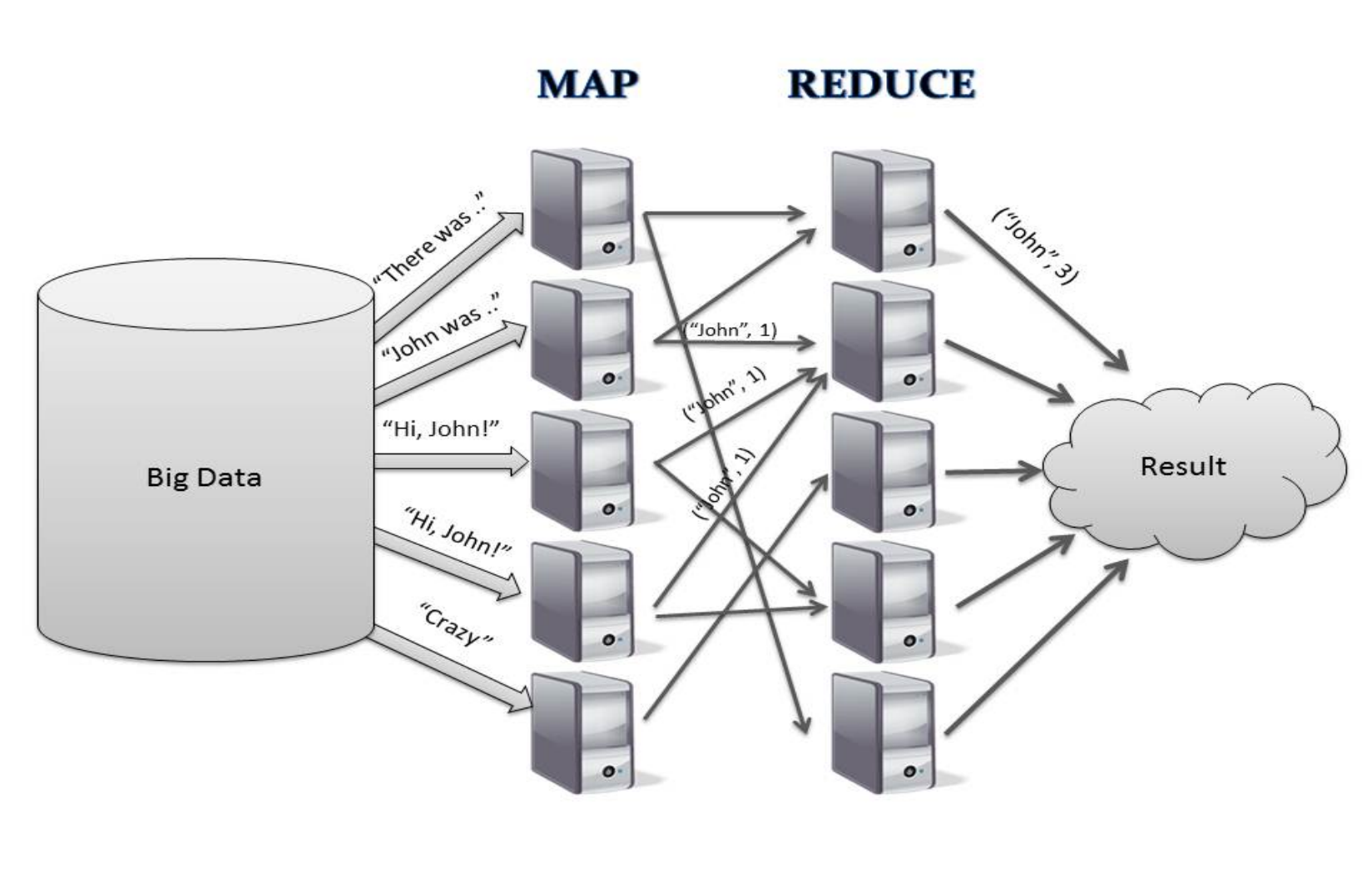
Arquitetura

- MapReduce
  - Um novo paradigma de programação
  - Processa grandes conjuntos de dados
  - Algoritmo paralelo e distribuído em clusters
  - A função **Map()** filtra e distribui os dados no cluster
  - E a **Reduce()** reagrupa o resultado do **Map()**

Arquitetura

- MapReduce
  - Um novo paradigma de programação
  - Processa grandes conjuntos de dados
  - Algoritmo paralelo e distribuído em clusters
  - A função **Map()** filtra e distribui os dados no cluster
  - E a **Reduce()** reagrupa o resultado do **Map()**
  - Inspirado na programação funcional

Arquitetura

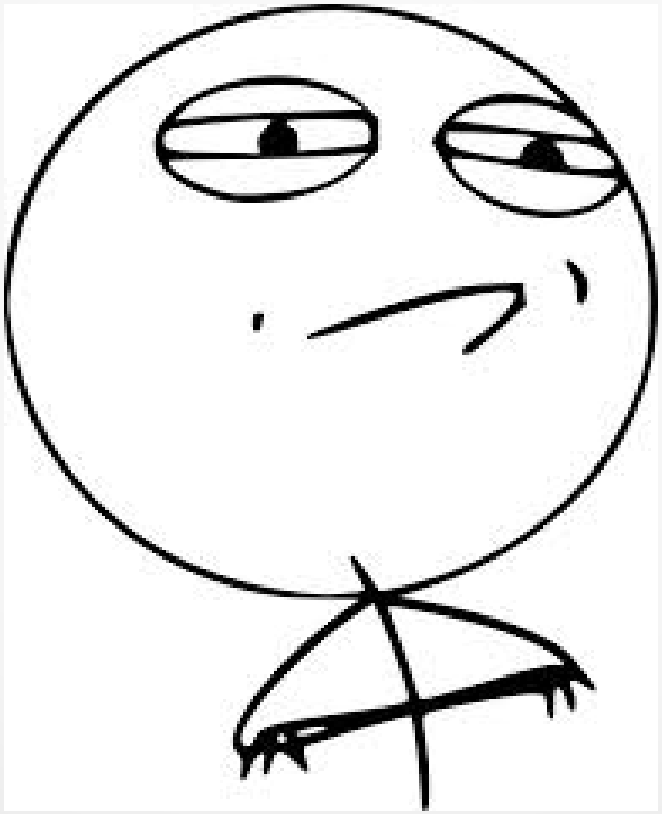


Arquitetura

- MapReduce
  - Um novo paradigma de programação
  - Processa grandes conjuntos de dados
  - Algoritmo paralelo e distribuído em clusters
  - A função **Map()** filtra e distribui os dados no cluster
  - E a **Reduce()** reagrupa o resultado do **Map()**
  - Inspirado na programação funcional
  - A implementação mais utilizada é o Hadoop

Desafios e Oportunidades

- Desafios



Desafios e Oportunidades

- Desafios
  - Entender e utilizar dados desestruturados

Desafios e Oportunidades

- Desafios
  - Entender e utilizar dados desestruturados
  - Capturar os dados realmente importantes

Desafios e Oportunidades

- Desafios
  - Entender e utilizar dados desestruturados
  - Capturar os dados realmente importantes
  - Armazenar uma quantidade de dados crescente

Desafios e Oportunidades

- Desafios
  - Entender e utilizar dados desestruturados
  - Capturar os dados realmente importantes
  - Armazenar uma quantidade de dados crescente
  - Privacidade

Desafios e Oportunidades

- Desafios
  - Entender e utilizar dados desestruturados
  - Capturar os dados realmente importantes
  - Armazenar uma quantidade de dados crescente
  - Privacidade
  - Acesso seguro

Desafios e Oportunidades

- Oportunidades

Desafios e Oportunidades

- Oportunidades
  - "The next frontier for innovation, competition and productivity"

Desafios e Oportunidades

- Oportunidades
  - "The next frontier for innovation, competition and productivity"
  - Extrair compreensão e conhecimento

Desafios e Oportunidades

- Oportunidades
  - "The next frontier for innovation, competition and productivity"
  - Extrair compreensão e conhecimento
  - Identificar tendências

Desafios e Oportunidades

- Oportunidades
  - "The next frontier for innovation, competition and productivity"
  - Extrair compreensão e conhecimento
  - Identificar tendências
  - Agregar valor à economia, saúde e educação

Existe um Big Market?

Será que toda essa "modinha" de Big Data dá mesmo dinheiro?

Existe um Big Market?

- Softwares de auxílio a tomada de decisões mais precisos

Existe um Big Market?

- Softwares de auxílio a tomada de decisões mais precisos
- Aumento da procura por profissionais com conhecimento de estatística

Existe um Big Market?

- Softwares de auxílio a tomada de decisões mais precisos
- Aumento da procura por profissionais com conhecimento de estatística
- Precisa saber sobre Gerenciamento de Dados e Aprendizagem de Máquina

Existe um Big Market?

- Softwares de auxílio a tomada de decisões mais precisos
- Aumento da procura por profissionais com conhecimento de estatística
- Precisa saber sobre Gerenciamento de Dados e Aprendizagem de Máquina
- IBM criou a Big Data University

Críticas

- Privacidade

Críticas

- Privacidade
  - A maioria das pessoas já acha a recomendação de links do Google invasiva demais

Críticas

- Privacidade
  - A maioria das pessoas já acha a recomendação de links do Google invasiva demais
  - Pessoas mais previsíveis (e mais manipuláveis)?

Críticas

- Privacidade
  - A maioria das pessoas já acha a recomendação de links do Google invasiva demais
  - Pessoas mais previsíveis (e mais manipuláveis)?
  - Quem tem acesso a esses dados?

Críticas

- Na Ciência

Críticas

- Na Ciência
  - Negligenciar princípios como a escolha de uma amostra significativa ao se preocupar em demasia com o tratamento de grandes quantidades de dados

Críticas

- Na Ciência
  - Negligenciar princípios como a escolha de uma amostra significativa ao se preocupar em demasia com o tratamento de grandes quantidades de dados
  - Datasets heterogêneos podem distorcer resultados

Referências

• sas.com/offices/latinamerica/brazil/solucoes/bigdata/
• youtube.com/watch?v=7D1CQ_LOizA
• facebook.com/dan.ariely/posts/904383595868
• en.wikipedia.org/wiki/Big_data
• aws.amazon.com/pt/big-data/
• bigdata-startups.com/BigData-startup/walmart-making-big-data-part-dna/
• thedatacreatives.com/2013/12/5-big-data-myths-debunked.html
• readwrite.com/2013/12/26/big-data-myths-reality
• inside-bigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/
• spotfire.tibco.com/blog/?p=6793
• mckinseyquarterly.com/The_challenge_and_opportunity_of_big_data_2806
• cio.com.br/opiniao/2012/05/11/o-caos-conceitual-e-os-5-vs-do-big-data/

Perguntas

Eduardo Tenório - embat@cin.ufpe.br