

# BIG DATA



## Disciplina: Introdução ao Big Data

**Tema da Aula: Big Data – Oportunidades e Tecnologias Habilitadoras**

### Coordenação:

Prof. Dr. Adolfo Walter  
Pimazzi Canton

Profa. Dra. Alessandra de  
Ávila Montini

**Profa. Rosangela de Fátima Pereira**

**Maio de 2016**

# Um pouco sobre mim

## Formação

- Mestrado em Engenharia de Computação pela Escola Politécnica da Universidade de São Paulo (Poli-USP) ([em andamento](#))
- Especialização em Tecnologia Java pela Universidade Tecnológica Federal do Paraná (UTFPR) (2011)
- Tecnologia em Análise e Desenvolvimento de Sistemas pela UTFPR ([2011](#))
- Bacharelado em Administração de Empresas pela Universidade Estadual do Norte do Paraná (UENP) (2007)

## Experiência

- Professora de Big Data Analytics em empresas e programas de MBA - FIA ([2013 - atual](#))
- Pesquisadora no Laboratório de Arquitetura e Redes de Computadores (LARC) – USP ([2013 - atual](#))
- Professora de cursos de engenharia na UTFPR ([2011 -2012](#))
- Analista de sistemas na BSI Tecnologia ([2009-2010](#))

LinkedIn: <https://br.linkedin.com/pub/rosangela-de-fatima-pereira/68/a10/b56>

Apaixonada por Big Data!

# Conteúdo da Aula

- Os 3 Vs de Big Data
- Tecnologias de Big Data
- Oportunidades a partir de Big Data

# Objetivo da Aula

Apresentar o conceito de Big Data, as tecnologias necessárias para se atuar com Big Data e as oportunidades existentes nesse contexto.

# Antes de definirmos Big Data...

# Resultado



E se não tivéssemos  
mais a Internet?

# Resultado



# Resultado



Vivemos em um  
mundo **conectado**

# Novos desafios

Como extrair valor  
desses dados?



# Novos desafios

Como extrair valor  
desses dados

# BIG DATA

# Definição de Big Data

“Big Data faz referência ao grande volume, variedade e velocidade de dados que demandam formas inovadoras e rentáveis de processamento da informação, para melhor percepção e tomada de decisão.”

Gartner, 2012

# Definição de Big Data

“Big Data faz referência ao grande **volume, variedade e velocidade** de dados que demandam formas inovadoras e rentáveis de processamento da informação, para melhor percepção e tomada de decisão.”

Gartner, 2012

# 6 Vs de Big Data

VOLUME

VARIEDADE

VELOCIDADE

VERACIDADE

VALOR

VULNERABILIDADE

BIG  
DATA

VOLUME

Volume  
Geramos  
2,5 exabytes  
de dados por dia

Fonte: <https://hbr.org/2012/10/big-data-the-management-revolution/ar>

# Volume

VOLUME

De toda a quantia de dados disponível globalmente, aproximadamente **90%** foram criados nos últimos **2 anos**.

Espera-se que esse volume de dados **dobre** a cada ano, pelo menos nos próximos **5 anos**.

Fonte: <http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>

# Variedade

80%

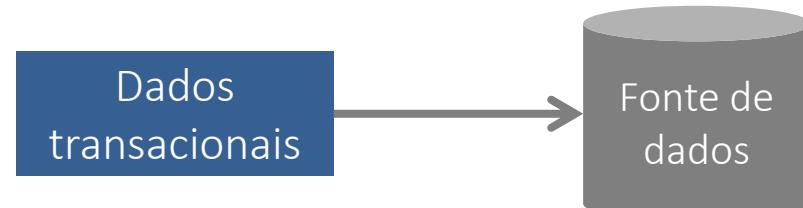
VARIEDADE

dos dados globais atualmente são  
não-estruturados

Fonte: <http://www.computerweekly.com/blog/CW-Developer-Network/IBM-80-percent-of-our-global-data-is-unstructured-so-what-do-we-do>

# Variedade

VARIEDADE



ANTERIORMENTE

# Variedade



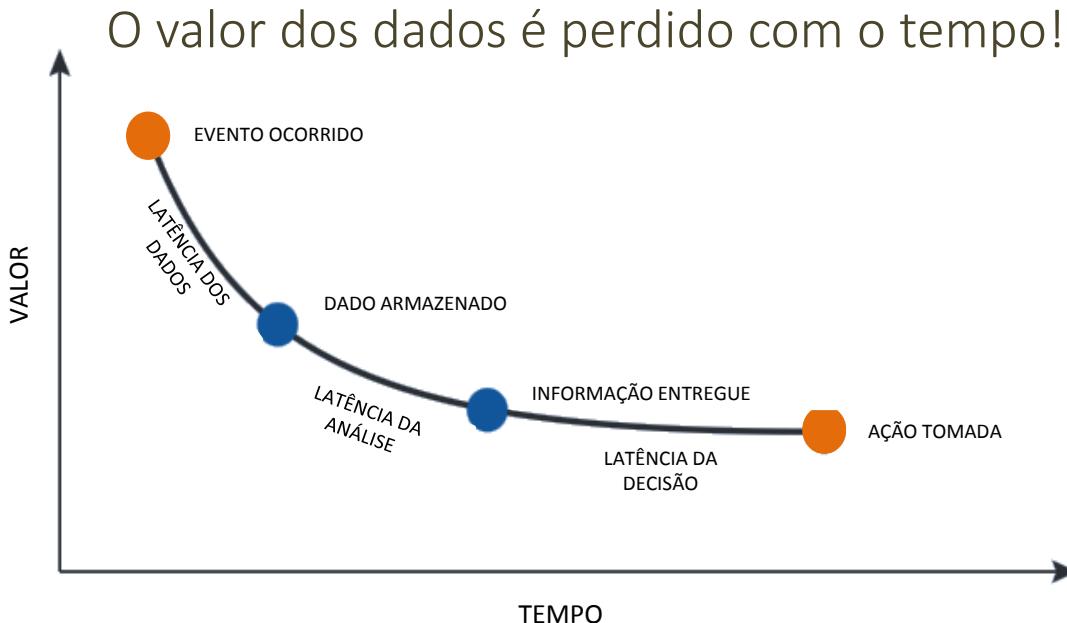
# Velocidade

O que acontece em 60 segundos na Internet?



Fonte: <http://pennystocks.la/internet-in-real-time/>

# Velocidade



Fonte: Hackathorn 2004.

# Tecnologias de Big Data...

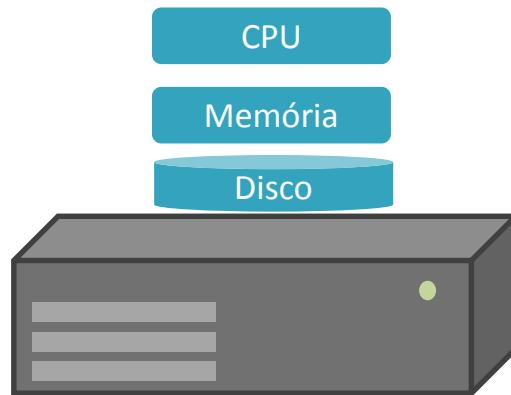
“Big Data faz referência ao grande volume, variedade e velocidade de dados que demandam **formas inovadoras e rentáveis de processamento da informação** para melhor percepção e tomada de decisão.”

Gartner, 2012

# Processamento de dados tradicional

## Escalabilidade vertical

Aumento de recursos computacionais em uma única máquina



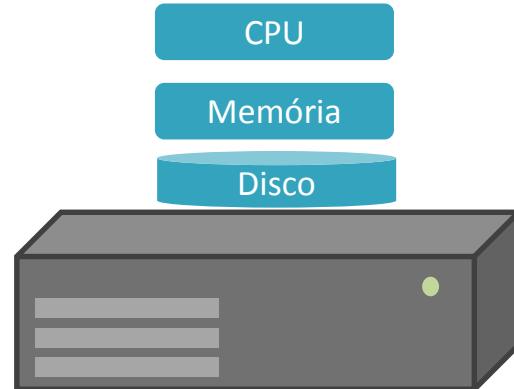
# Processamento de dados tradicional

## Escalabilidade vertical

Aumento de recursos computacionais em uma única máquina

### Problemas:

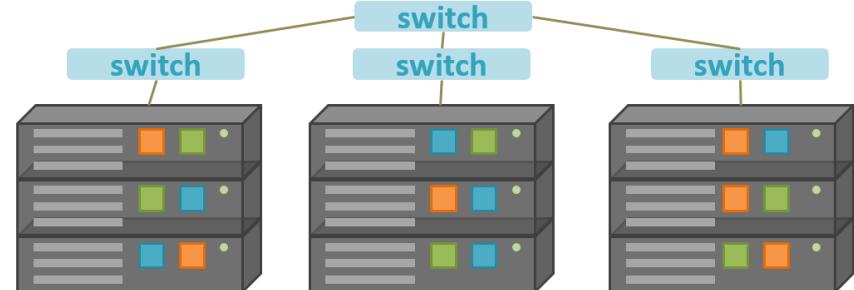
- Custo
- Desempenho
- Escalabilidade limitada



# Processamento de dados em Big Data

## Escalabilidade horizontal

Adição de máquinas ao cluster



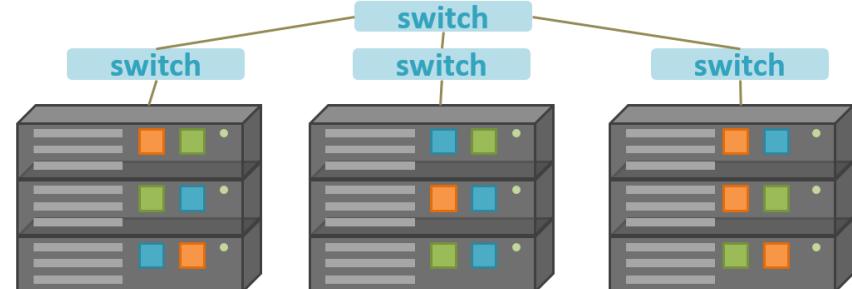
# Processamento de dados em Big Data

## Escalabilidade horizontal

Adição de máquinas ao cluster

### Problemas:

- Adaptação de tecnologias tradicionais a esse ambiente
- Desenvolvimento e manutenção complexa



# Necessidade atual

Novas tecnologias capazes de oferecer  
escalabilidade, disponibilidade, flexibilidade e  
desempenho para a manipulação de grande  
volume de dados

# Inovação no ARMAZENAMENTO

# Tipos de dados

REDES SOCIAIS



VÍDEO



ÁUDIO



EMAIL



MENSAGEM



ONDE  
ARMAZENAR  
ESSES DADOS?



TRANSAÇÕES



MOBILE



IMAGEM



DOCUMENTOS



GEOLOCALIZAÇÃO

# NoSQL - Definição

NoSQL

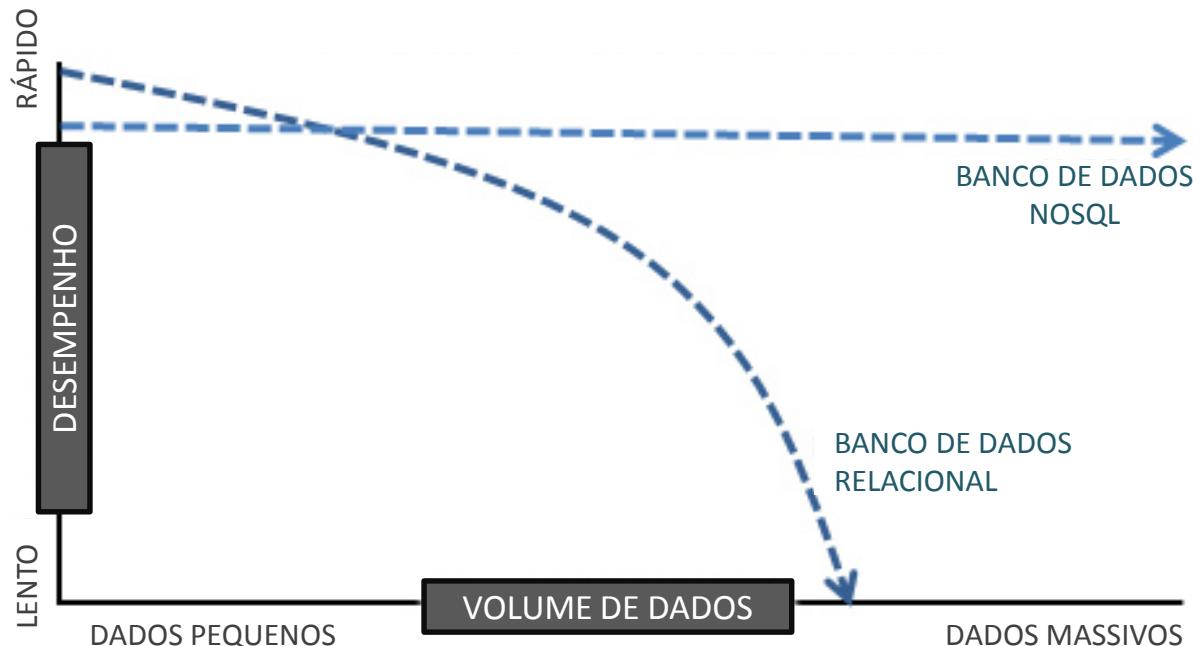
=

Not Only SQL

*“Conjunto de conceitos que permite o processamento rápido e eficiente de conjuntos de dados com foco em desempenho, confiabilidade e agilidade”.*

*Making sensing of NoSQL*

# NoSQL versus Banco de dados relacional



FONTE: DataJobs.com

# NoSQL - Características

## CARACTERÍSTICAS

Não-relacional

*Cluster-friendly*

Sem esquema

Escala horizontalmente

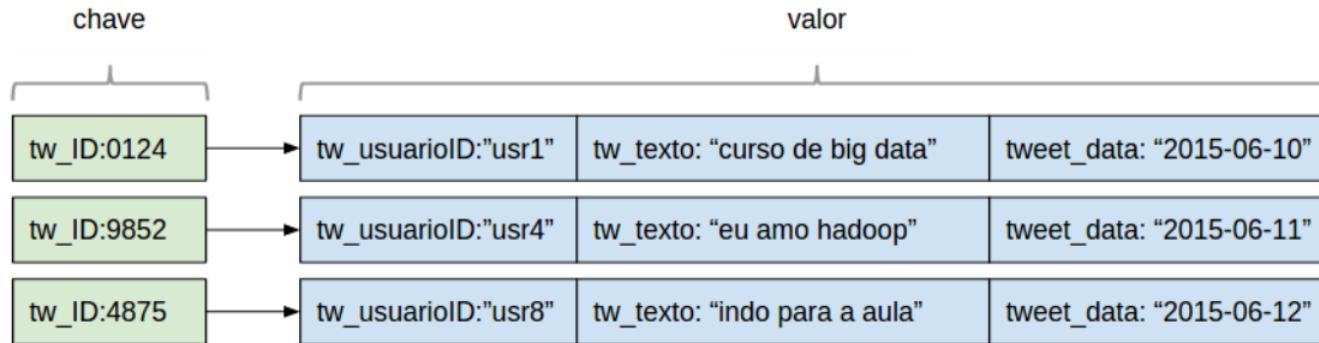
Interface de consulta simples

# NoSQL

Diferentes formas de armazenamento

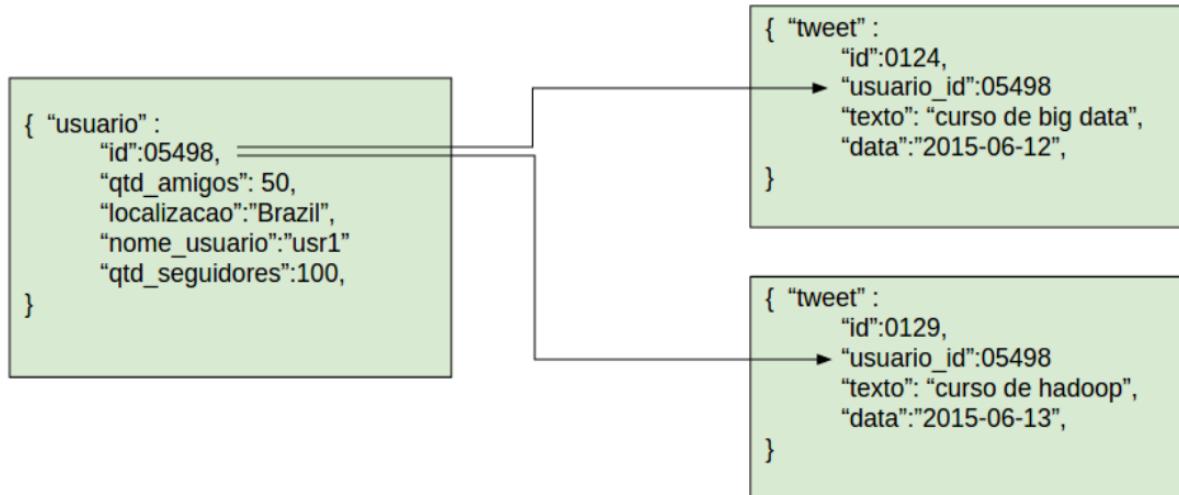


# NoSQL



Orientado a chave-valor  
DynamoDB, Redis, Voldemort

# NoSQL



Orientado a documentos  
MongoDB, ElasticSearch, CouchDB

# NoSQL

ID	Nome	Idade
0011	João	42
0012	José	34
0013	Leila	29

banco de dados relacional

0011	João	42	0012	José	34	0013	Leila	29
------	------	----	------	------	----	------	-------	----

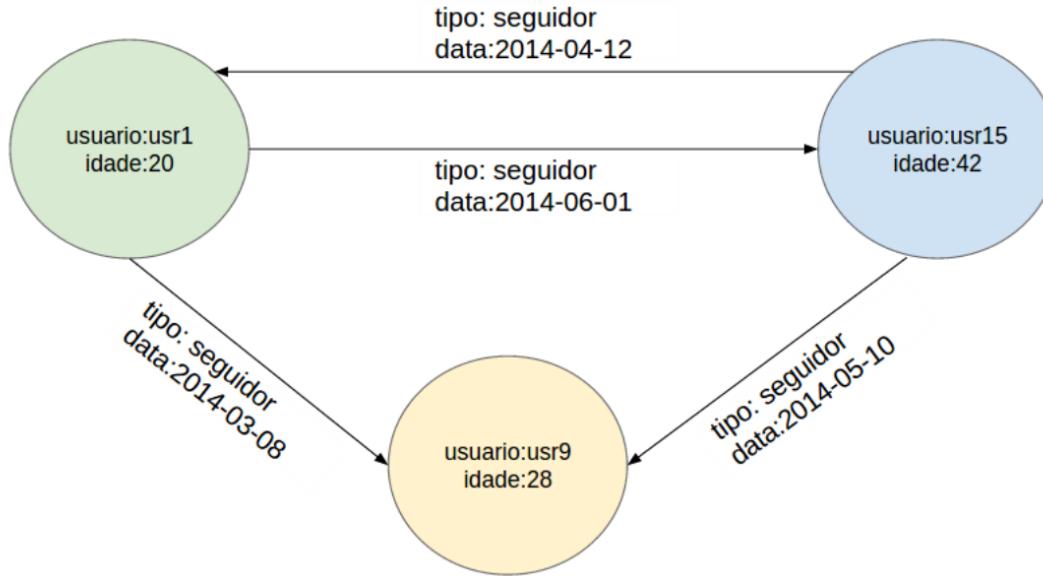
0011	0012	0013	João	José	Leila	42	34	29
------	------	------	------	------	-------	----	----	----

banco de dados orientado a colunas

Orientado a colunas

Hbase, Cassandra, Accumulo

# NoSQL

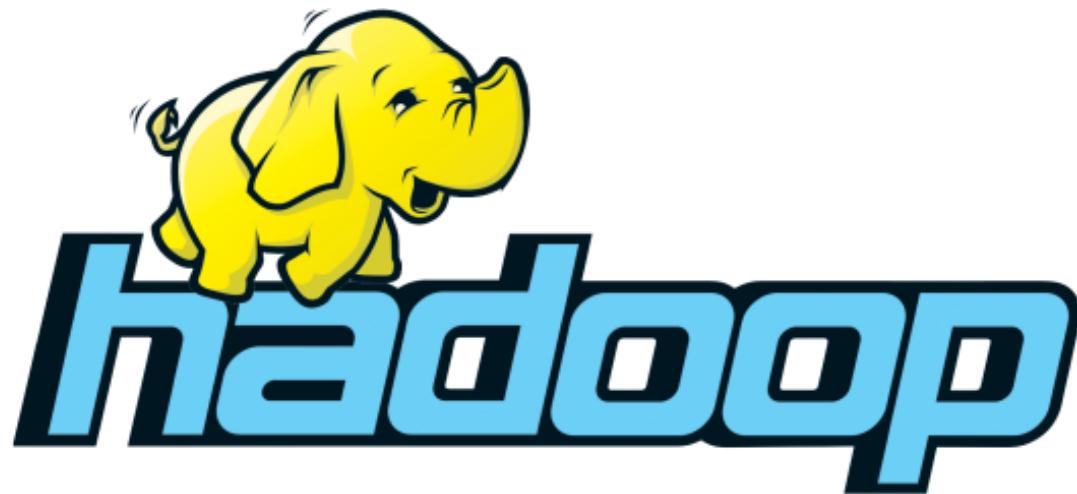


Orientado a grafos

Neo4J, AllegroGraph, InfinitGraph

# Inovação no PROCESSAMENTO

# Inovação no processamento



# Inovação no processamento



# Hadoop - Caso de uso

## Globo.com

**Desafio:** capturar cada interação que os milhões de usuários fazem nos sites e agir instantaneamente gerando **recomendações personalizadas** do conteúdo da Globo.com

**Solução:** adoção de um cluster de distribuição de mensagens (bilhões de mensagens por dia) utilizando Hadoop, HBase, Kafka e Spark

**Resultados:** maior escalabilidade, tolerância a falhas e desempenho do sistema.

**Cluster:** 10 nós totalizando 1 TB de RAM e 500 TB de HD

Fonte: <https://www.infoq.com.br/presentations/construindo-uma-plataforma-com-hadoop-e-elasticsearch>

# Inovação na COLABORAÇÃO

# Organização tradicional dos dados

Data warehouse  
Silos de dados



# Organização dos dados para Big Data

Data Lake (Infraestrutura Hadoop)  
Dados armazenados em sua forma original



Que soluções posso gerar a partir de  
Big Data?

“Big Data faz referência ao grande volume, variedade e velocidade de dados que demandam formas inovadoras e rentáveis de processamento da informação para **melhor percepção e tomada de decisão.**”

Gartner, 2012

# Oportunidades



# Oportunidades



# Big Data Analytics

## Quatro abordagens

DESCRITIVA

PREDITIVA

DIAGNÓSTICA

PRESCRITIVA

# Análise descritiva

Foco em sumarizar **fatos** ocorridos

Permite compreender como a organização está sendo operada

Uso de ferramentas de *Business Intelligence* (BI)

Alertas

Dashboards

Relatórios



O que aconteceu?

# Análise descritiva

Baseado em métricas

- Visualização de tendências
- Descoberta de padrões

**Mais de 80% da análise de negócios são descritivas**

Exemplos:

- Relatório diário/semanal de vendas
- Lista de cancelamento de assinantes

# Análise diagnóstica

Foco em determinar o **motivo** de um evento ter ocorrido

Demonstra variações de desempenho positivas e negativas

Uso de outros métodos estatísticos

- Análise de correlação
- Análise de variância
- Testes de hipóteses



Por que aconteceu?

# Análise preditiva

Foco em **previsões** de eventos futuros

Extrai padrões encontrados em dados históricos

Utilizado para diferentes aplicações

- Detecção de fraude
- Gerenciamento de risco
- Fidelização de clientes



O que acontecerá?

# Análise preditiva

Requer uso de diversos métodos e ferramentas

- Análises estatísticas
- Técnicas de simulação
- Mineração de dados
- Aprendizado de máquina

Análise descritiva

Relatório climático

Análise preditiva

Previsão do tempo

# Análise prescritiva

Foco em prever ações futuras e possíveis consequências

Sugere ações baseadas no conhecimento extraído dos dados

Envolve regras de negócios, conhecimentos matemáticos, mineração de dados



Como fazer acontecer?

# Análise prescritiva

Segundo a consultora Gartner, somente **3%** das empresas utilizam análise prescritiva

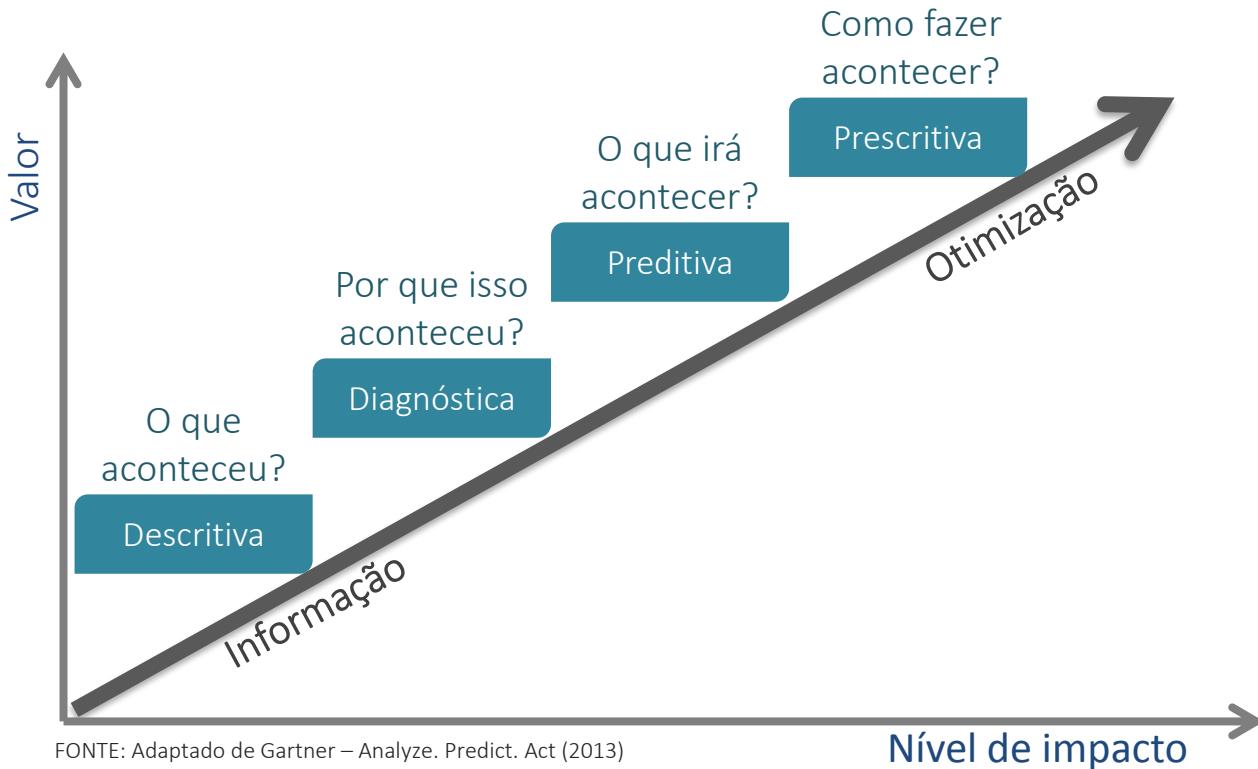
Necessário grandes quantidades de dados

Exemplos:

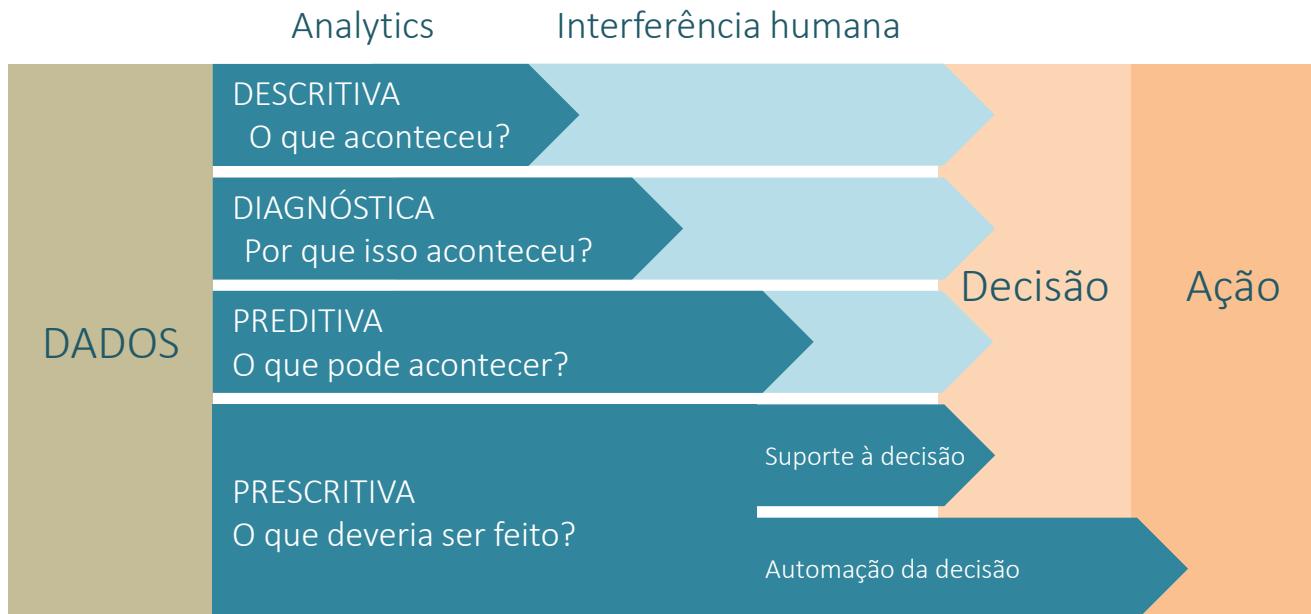
- Manufatura inteligente
- Carro autônomo do Google

Fonte: <http://data-informed.com/future-big-data-prescriptive-analytics-changes-game/>

# Resumo



# Resumo



FONTE: Adaptado de <http://www.datasciencecentral.com/profiles/blogs/prescriptive-versus-predictive-analytics-a-distinction-without-a>

# Big Data Analytics



**TARGET**

Detecção de mudanças ocorridas na vida das pessoas, como a gravidez

Dados históricos

+ Mineração de dados

+ Aprendizado de máquina

= Descoberta de padrões

Fonte: <http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>

# Big Data Analytics

- Máquinas da linha de produção medem centenas de parâmetros no processo de tomada de pneus a cada segundo oferecendo uma manufatura inteligente
- Mais de 4.000 sensores são capturados para o gerenciamento de eficiência de frota



*“Putting data analytics at the core of a business involves a new way of thinking and a new process of decision-making”*

*Carlo Torniai – cientista de dados da Pirelli*

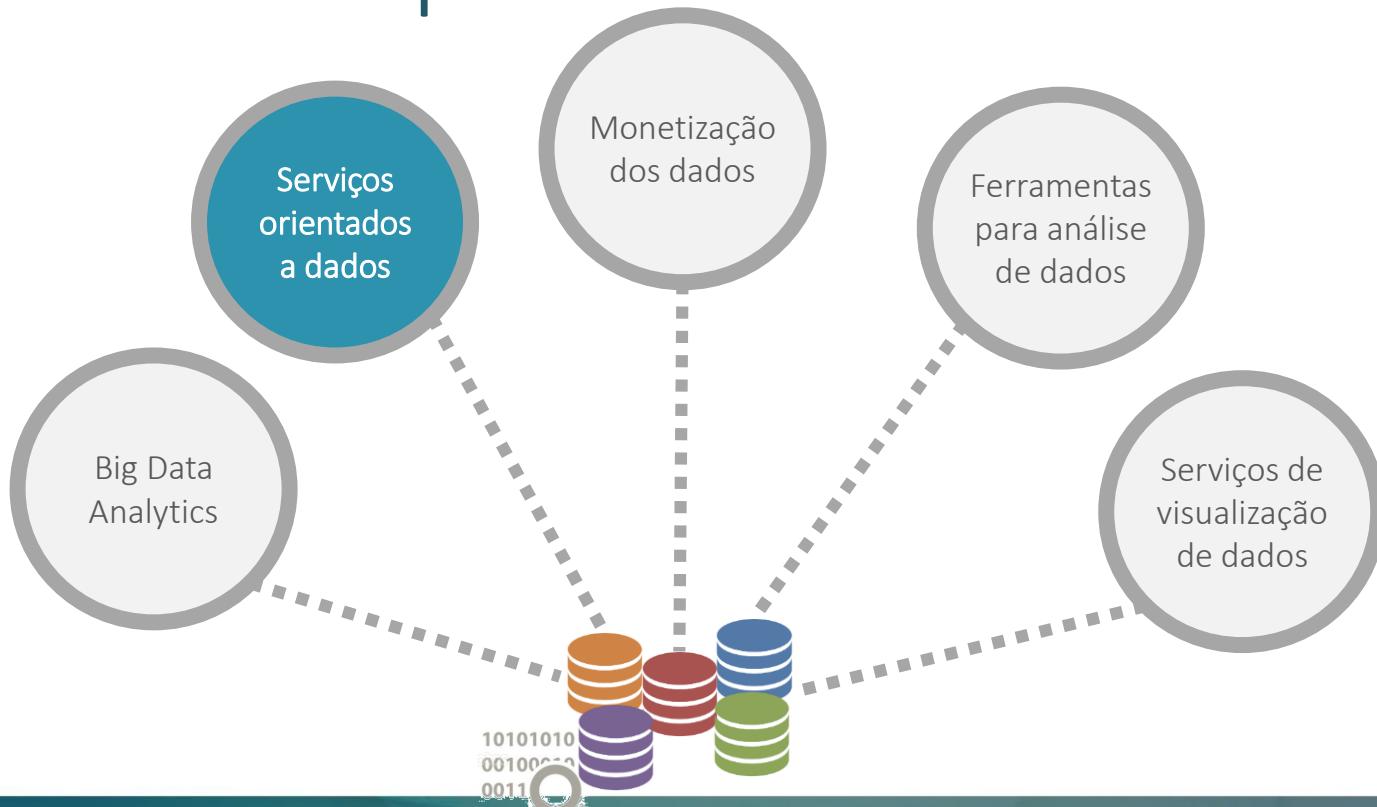
Fonte: <http://business.pirelli.com/global/en-ww/how-pirelli-is-becoming-data-driven>

# Big Data Analytics

Big Data Analytics pode oferecer informações valiosas, porém é necessário uma análise criteriosa das informações...



# Oportunidades



# Serviços orientados a dados

Vantagens competitivas utilizando Big Data

CONSOLIDAÇÃO

MARCA

VELOCIDADE DE  
PROCESSAMENTO

DIFERENCIAMENTO

EXPANSÃO

Fonte: Data and Analytics - Data-Driven Business Models: A Blueprint for Innovation. University of Cambridge

# Serviços orientados a dados



# Serviços orientados a dados

Onde mais esses serviços podem ser aplicados?

TRANSPORTE  
URBANO

CUIDADOS DA  
SAÚDE

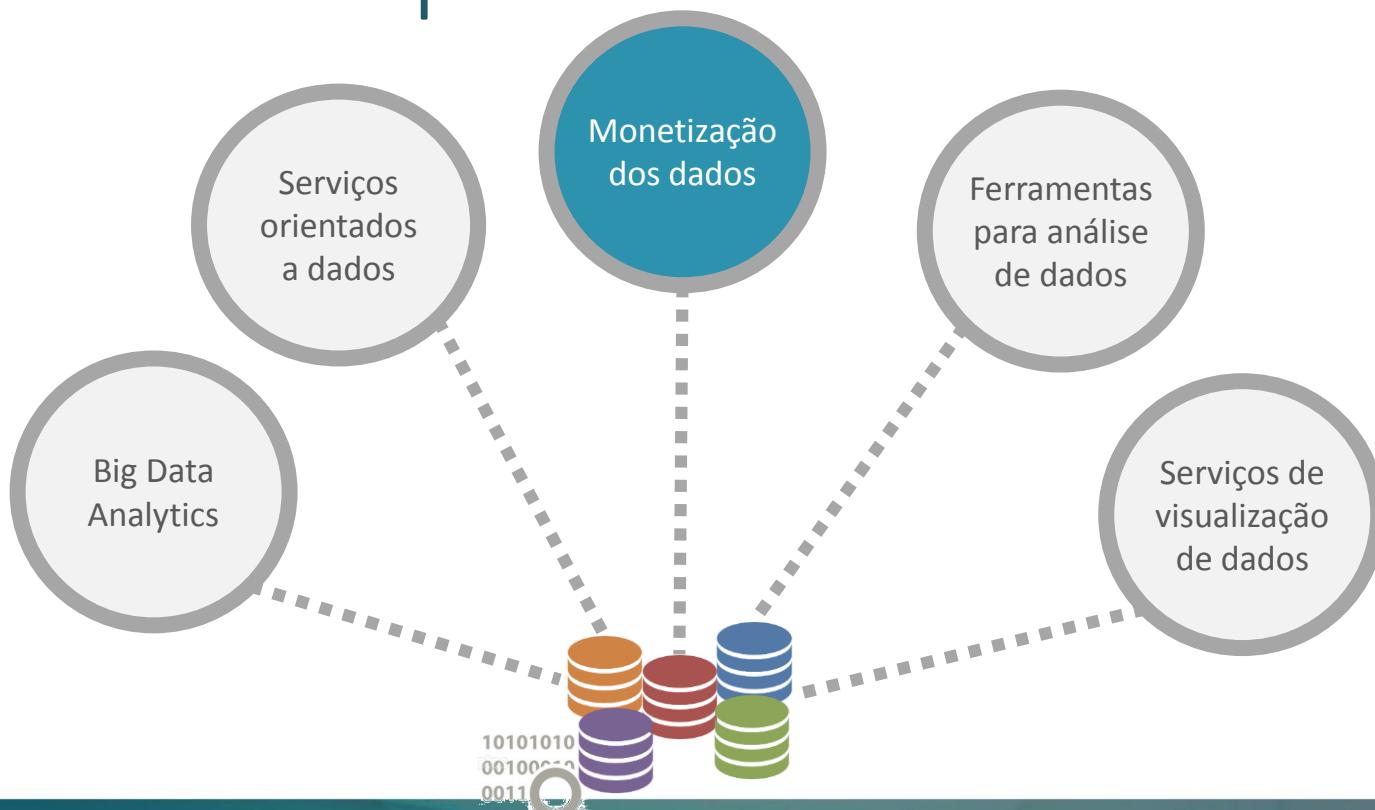
DESENVOLVIMENTO  
ECONÔMICO

GERENCIAMENTO  
DE RESÍDUOS

SEGURANÇA  
PÚBLICA

AGRICULTURA DE  
PRECISÃO

# Oportunidades



# Monetização dos dados

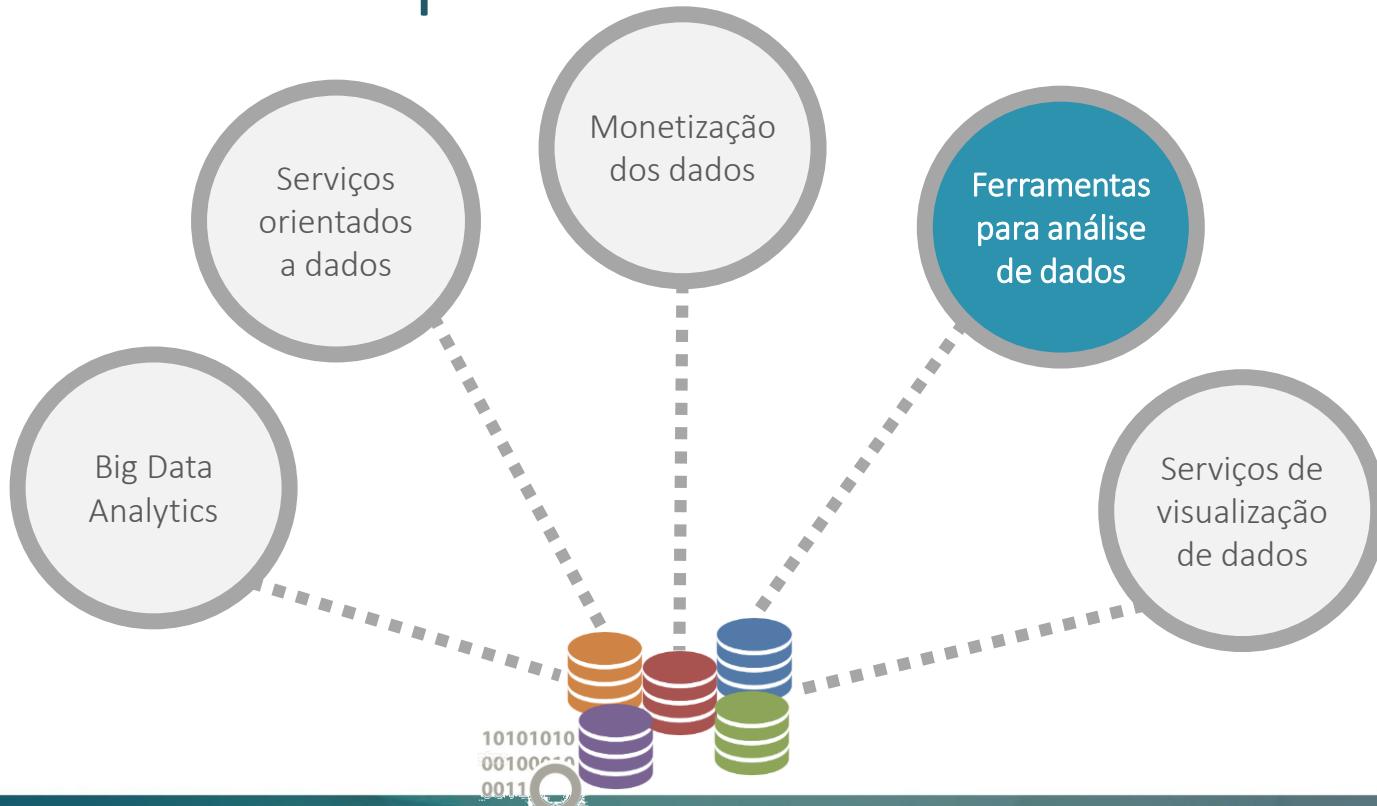
Transformar ativos de informação em dinheiro, direta ou indiretamente, por meio de troca, comercialização ou venda direta

Exemplos:

- Dados de comportamento de usuários para campanhas de marketing
- Métricas de uso de transações de cartão de crédito para comerciantes
- Dados de sensores para seguradoras
- Dados genéticos para pesquisadores
- Dados abertos!



# Oportunidades



# Ferramentas para análise de dados

Processamento  
distribuído dos  
dados

Algoritmos de  
aprendizado de  
máquina

Processamento em  
tempo real

Construção de  
modelos

Tendência: soluções de Big Data como plataformas e serviços de computação em nuvem (PaaS/SaaS)

# Ferramentas para análise de dados

**Caso de uso:** [NextBio](#), empresa líder na análise e agregação de dados de genomas. Oferece uma plataforma como serviço de nuvem para análise de dados para pesquisadores

**Desafio:** acessar e analisar mais de 10.000 estudos sobre os genomas em uma base com mais de 100 TB de dados

**Requisitos:** escalabilidade da plataforma e disponibilidade dos dados

**Solução:** adoção da plataforma Hadoop e o banco de dados NoSQL HBase

# Oportunidades



# Serviços de visualização de dados

NOSSO FORMATO VISUAL É PROCESSADO

# 60.000 vezes

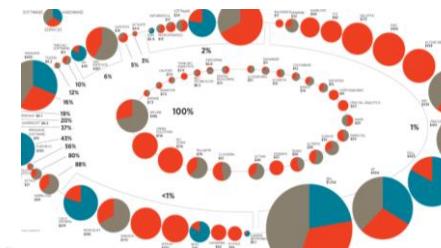
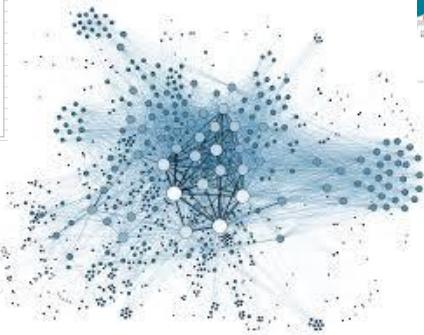
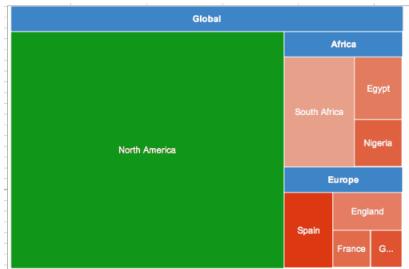
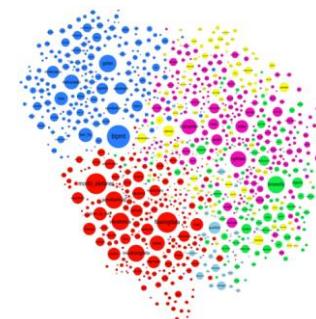
MAIS RÁPIDO QUE O FORMATO DE TEXTO

Fonte: <http://www.t-sciences.com/news/humans-process-visual-data-better>

# Serviços de visualização de dados

## VISUALIZAÇÃO DE DADOS

Exibição gráfica de informações para duas finalidades: construção de sentido (também chamada de análise de dados) e comunicação



design  
information  
help  
user  
interface  
process  
elements  
different  
product  
different  
form  
use  
different  
button  
use  
just  
users  
application

# Serviços de visualização de dados

TOMADA DE  
DECISÃO  
APERFEIÇOADA

MELHORIA NA  
COLABORAÇÃO

REDUÇÃO DE  
TEMPO

MONITORAMENTO  
DE  
PRODUTIVIDADE

MELHORIA NA  
ANÁLISE DE DADOS

MELHOR  
EXPERIÊNCIA AO  
USUÁRIO

Fonte: IDC Research

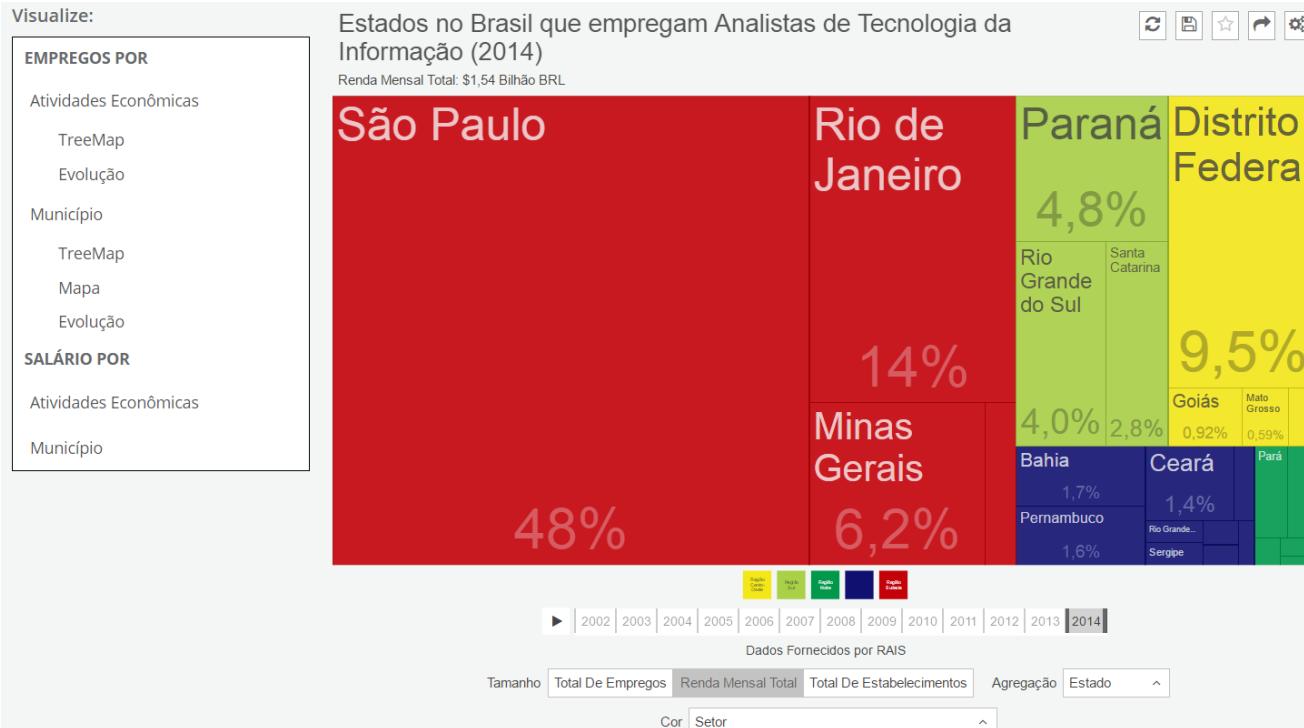
# Serviços de visualização de dados

**Caso de uso:** DataViva – portal de visualização de dados da economia brasileira

**Base de dados:** Relação Anual de Informações Sociais - Ministério do Trabalho e Emprego – TEM

Disponibiliza dados oficiais sobre exportações, atividade econômica, localidade, educação e ocupações de todo o Brasil. São 11 aplicativos, que, juntos, formam mais de 1 bilhão de possibilidades de visualizações

# Serviços de visualização de dados



Fonte: <http://dataviva.info/>

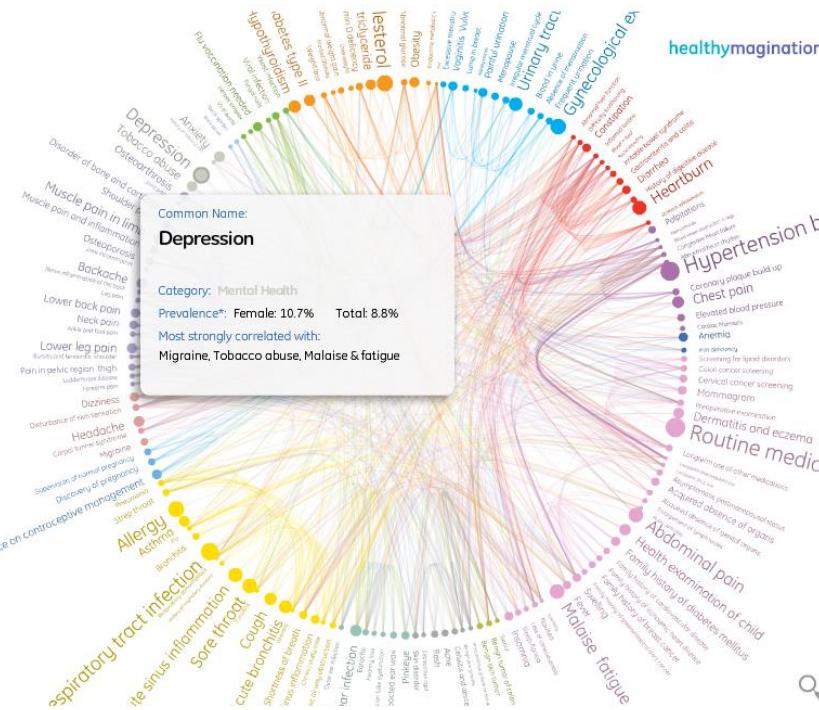
# Serviços de visualização de dados

**Caso de uso:** General Electric – visualizações do relacionamento entre sintomas de doenças

**Base de dados:** registros de 7.2 milhões de pacientes capturados nos equipamentos da GE

Apresenta categoria dos sintomas, como sintomas estão relacionados e prevalência entre homens e mulheres

# Serviços de visualização de dados



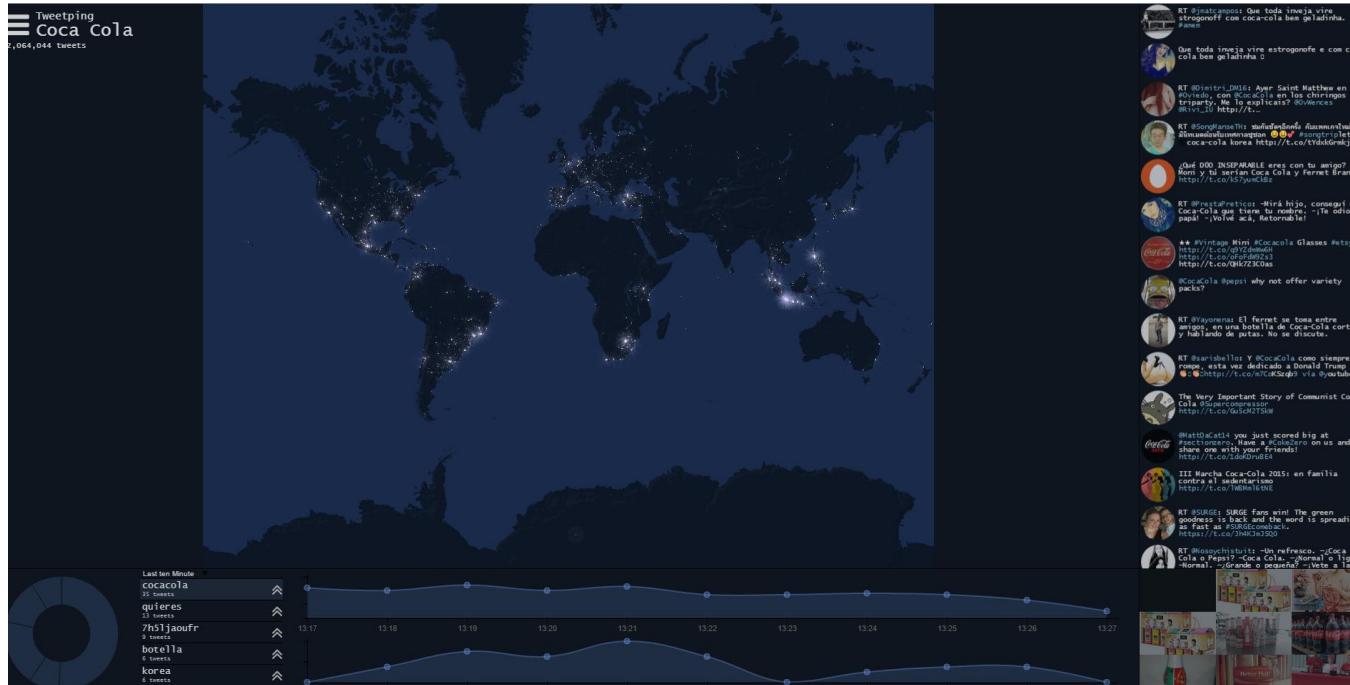
Fonte: <http://senseable.mit.edu/healthinfoscape/interactive/HealthInfoScape.html>

# Serviços de visualização de dados

**Caso de uso:** TweetPing – visualizações em tempo real de mensagens do Twitter

Apresenta em tempo real mensagens de usuários de diversas localidades do mundo, com base em uma palavra chave

# Serviços de visualização de dados



Fonte: <http://pro.tweetping.net/stream/1555>

# POR ONDE COMEÇAR?

# Por onde começar?

COMECE COM UMA  
PERGUNTA!

DESCUBRA O QUE VOCÊ  
DESEJA RESPONDER  
COM OS DADOS

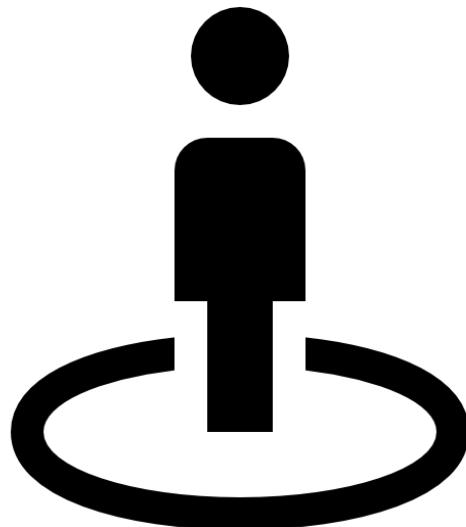


# Por onde começar?

1. Qual é o objetivo do projeto?
2. Quais dados são necessários?
3. Quais são os obstáculos para chegar lá?
4. Quem são os principais interessados e quais são suas funções?

# Por onde começar?

UM PROJETO DE BIG DATA NÃO SE FAZ SOZINHO!



# Por onde começar?

UM PROJETO DE BIG DATA SE FAZ COM UMA EQUIPE HETEROGÊNEA



# Por onde começar?

## PROFISSÃO DO ANO: CIENTISTA DE DADOS

salário médio anual: US\$ 128.240

Eleita a **profissão do ano** pelo site Americano de empregos CareerCast.com, considerando critérios como: ambiente de trabalho, renda, nível de stress e perspectiva de contratação

Profissão nº 2: estatístico  
salário médio anual: US\$ 79.990



Fonte: <http://g1.globo.com/economia/concursos-e-emprego/noticia/2016/05/veja-lista-com-10-melhores-e-piores-profissoes-para-2016.html>

# Por onde começar?

## PROFISSÃO DO ANO: CIENTISTA DE DADOS

*“A elevada procura de cientistas de dados e estatísticos vem de uma crescente ênfase na coleta e avaliação de quantidades massivas de dados”.*

*“As oportunidades para profissionais treinados nestas áreas são enormes, como o setor de TI, saúde, negócios - e qualquer setor que coleta a informação do consumidor pode colocar esses números para uso”.*

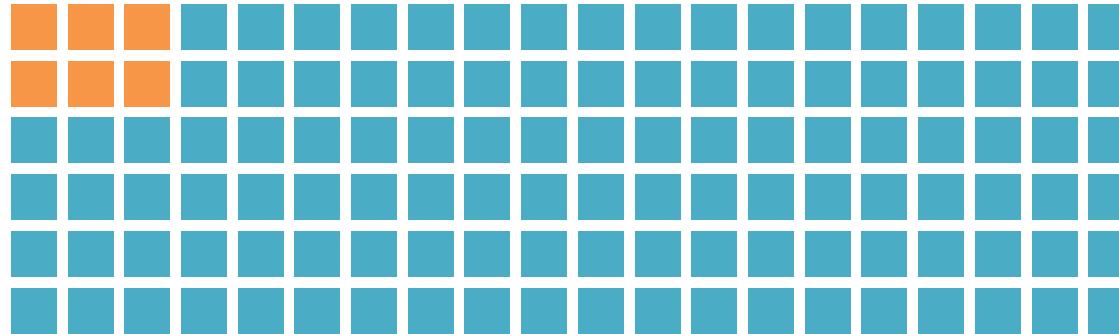


Fonte: <http://www.careercast.com/jobs-rated/best-jobs-2016>

# CONSIDERAÇÕES FINAIS

# Considerações finais

Big Data está apenas em seu início



Em 2012 a IDC estimou que somente **0.5%**  
dos dados globais eram analisados

Fonte: <http://www.theguardian.com/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume>

# Considerações finais

**BIG DATA** =

**BIG** INOVAÇÃO

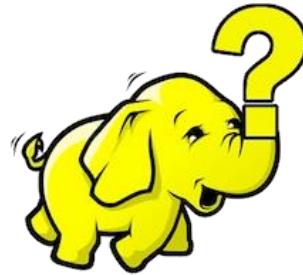
**BIG** AGILIDADE

**BIG** CONHECIMENTO

**BIG** OPORTUNIDADE

# Perguntas

[rpereira@larc.usp.br](mailto:rpereira@larc.usp.br)



# Referências

Gantz, J. and Reinsel, D. ***The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East***, IDC Research and EMC Corporation, Dezembro, 2012

MANYIKA, James et al. **Big data: The next frontier for innovation, competition, and productivity**. 2011.

PROVOST, Foster; FAWCETT, Tom. Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.", 2013.

MCCREARY, Dan; KELLY, Ann. **Making sense of NoSQL**. Greenwich, Conn.: Manning Publications, 2013.

FRY, Ben. **Visualizing data: Exploring and explaining data with the processing environment**. " O'Reilly Media, Inc.", 2007.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. **Big data: A revolution that will transform how we live, work, and think**. Houghton Mifflin Harcourt, 2013.

WHITE, Tom. **Hadoop: The definitive guide**. " O'Reilly Media, Inc.", 2012.

OHLHORST, Frank J. **Big data analytics: turning big data into big money**. John Wiley & Sons, 2012.