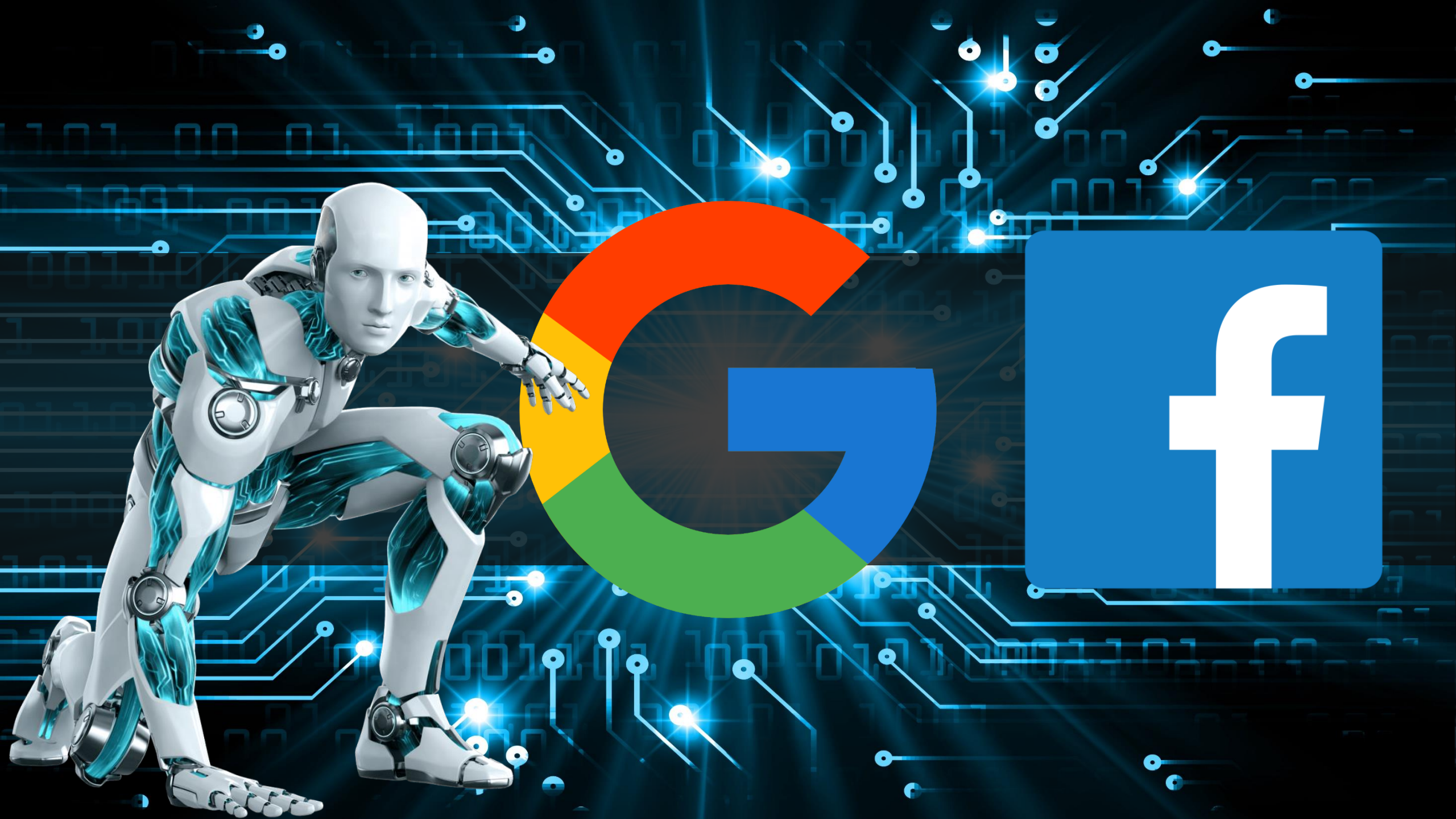
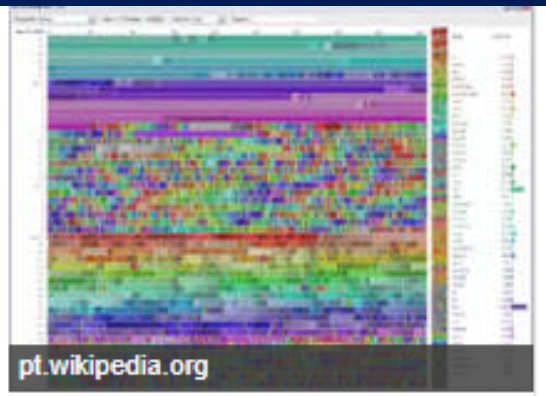
DATA SCIENCE

O que é Big Data?



Big Data é o termo que descreve o imenso volume de dados – estruturados e não estruturados – que impactam os negócios no dia a dia. Mas o importante não é a quantidade de dados. E sim o que as empresas fazem com os dados que realmente importam.



O que é Big Data? | SAS
https://www.sas.com/pt_br/insights/big-data/what-is-big-data.html

Segundo a SAS

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

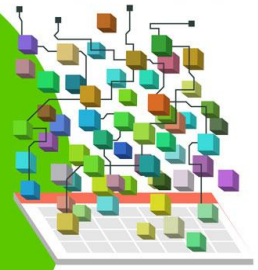
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day



6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT

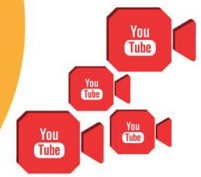
are shared on Facebook every month

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month



400 MILLION TWEETS

are sent per day by about 200 million monthly active users

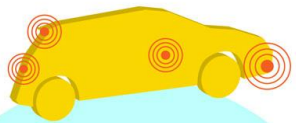
Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session



Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

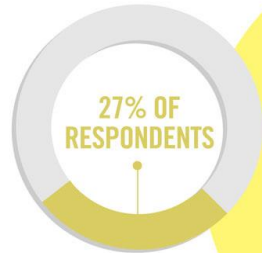
– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



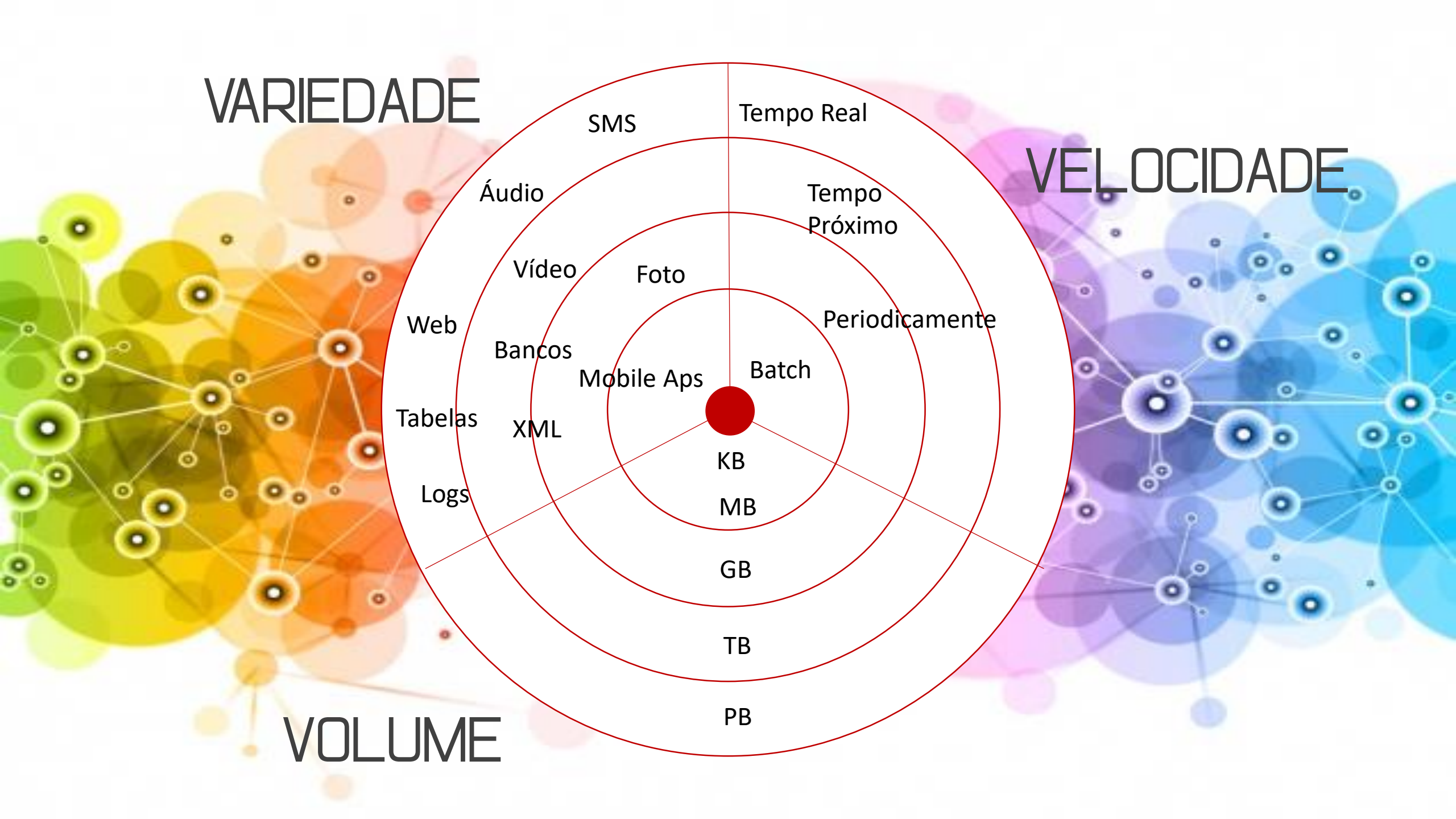
in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM


VARIEDADE
VELOCIDADE
VOLUME
SMS
Áudio
Vídeo
Web
Bancos
Tabelas
XML
Logs
Tempo Real
Tempo Próximo
Periodicamente
Batch
Foto
Mobile Aps
KB
MB
GB
TB
PB

Big Data representa um conjunto de dados que nao Pode mais ser facilmente gerenciado ou analisado com as ferramentas atuais de dados, métodos ou arquiteturas disponíveis até entao.

Big Data Landscape 2016

Infrastructure

Hadoop On-Premise
Hadoop in the Cloud
Spark
Cluster Services
NoSQL Databases
NewSQL Databases
Graph Databases
MPP Databases
Cloud EDW
Data Transformation
Data Integration
Management / Monitoring
Security
Storage
App Dev
Crowd-sourcing

Analytics

Analyst Platforms
Analytics Platforms
Data Science Platforms
Visualization
BI Platforms
Statistical Computing
Log Analytics
Social Analytics
Real-Time
Machine Learning
Speech & NLP
Horizontal AI
Search
Data Services
For Business Analysts
SMB / Commerce

Applications

Sales & Marketing
Customer Service
Human Capital
Legal
Ad Optimization
Security
Vertical AI Applications
Publisher Tools
Govt / Regulation
Finance
Education/ Learning
Life Sciences
Industries

Cross-Infrastructure/Analytics

Open Source

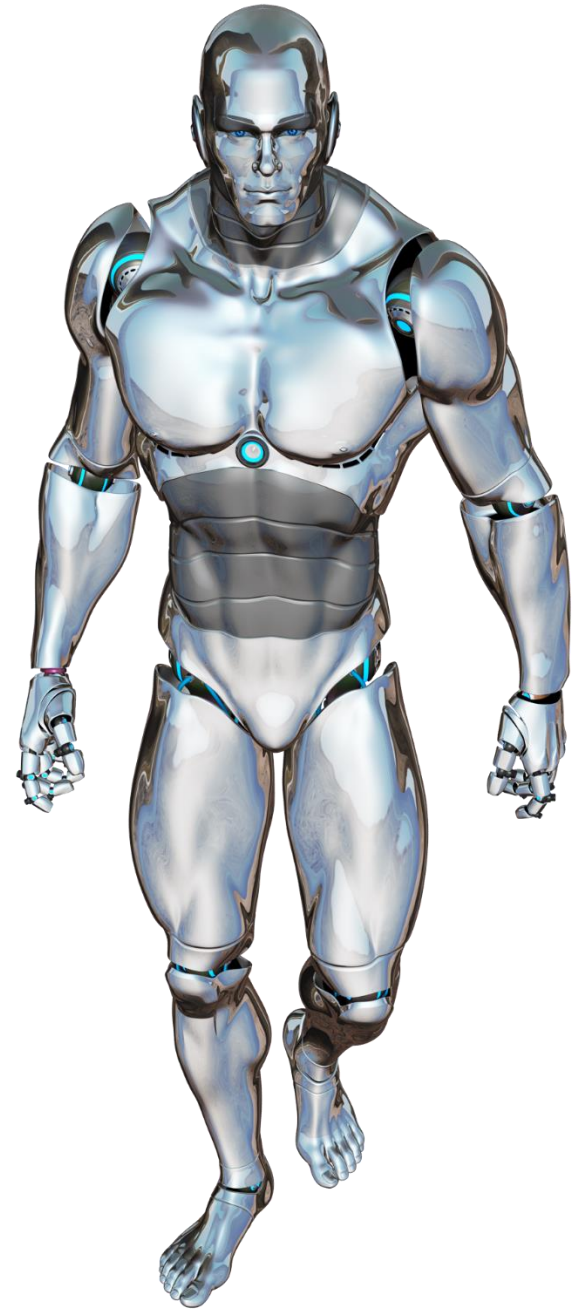
Framework
Query / Data Flow
Data Access
Coordination
Real-Time
Stat Tools
Machine Learning
Search
Security
Visualization

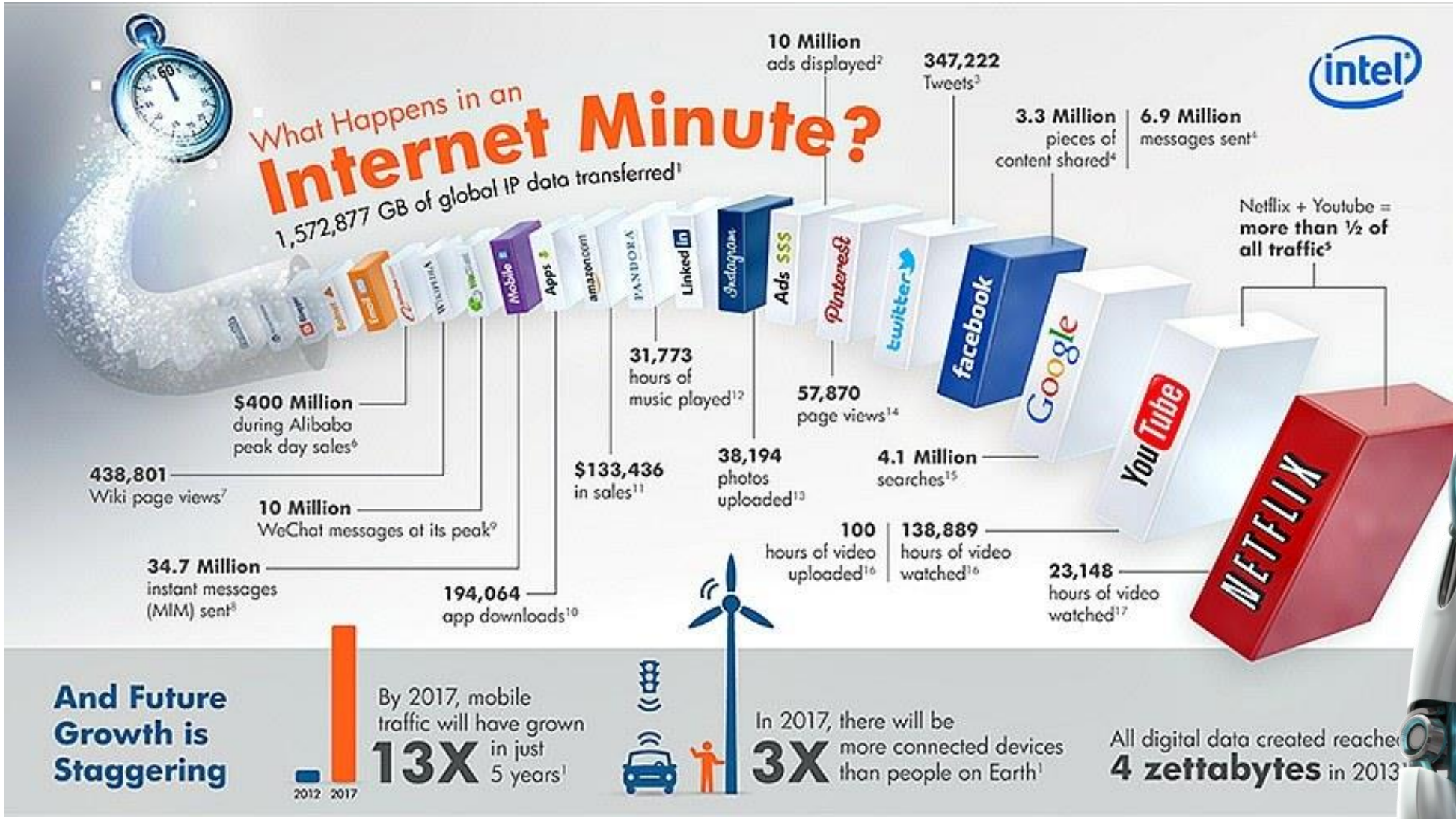
Data Sources & APIs

Health
IOT
Financial & Economic Data
Air / Space / Sea
Location / People / Entities
Other

Incubators & Schools

FIRSTMARK

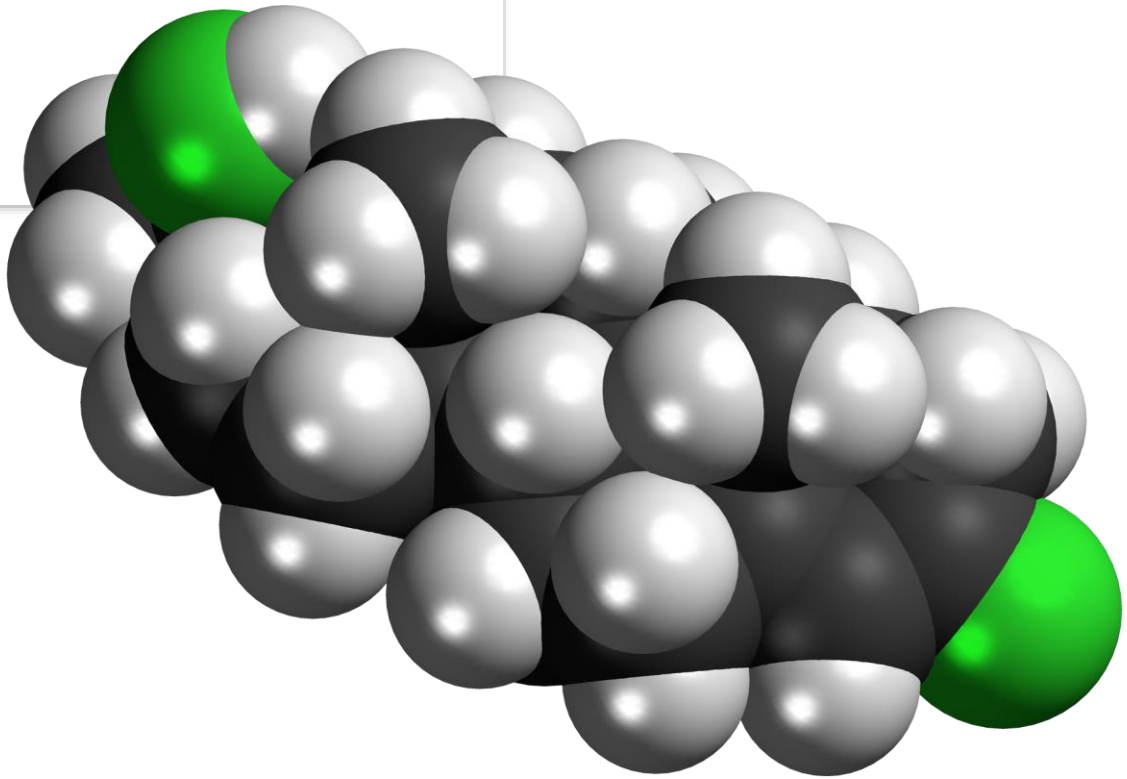



What Happens in an
Internet Minute?
1,572,877 GB of global IP data transferred[1]
intel
10 Million
ads displayed[2]
347,222
Tweets[3]
3.3 Million
pieces of
content shared[4]
6.9 Million
messages sent[4]
Netflix + Youtube =
more than ½ of
all traffic[5]
$400 Million
during Alibaba
peak day sales[6]
438,801
Wiki page views[7]
10 Million
WeChat messages at its peak[9]
34.7 Million
instant messages
(MIM) sent[8]
194,064
app downloads[10]
$133,436
in sales[11]
31,773
hours of
music played[12]
38,194
photos
uploaded[13]
57,870
page views[14]
4.1 Million
searches[15]
100
hours of video
uploaded[16]
138,889
hours of video
watched[16]
23,148
hours of video
watched[17]
Mobile
Apps
amazon.com
PANDORA
LinkedIn
Instagram
Ads $$$
Pinterest
twitter
facebook
Google
YouTube
NETFLIX
And Future
Growth is
Staggering
2012
2017
By 2017, mobile
traffic will have grown
13X in just
5 years[1]
In 2017, there will be
3X more connected devices
than people on Earth[1]
All digital data created reache
4 zettabytes in 2013

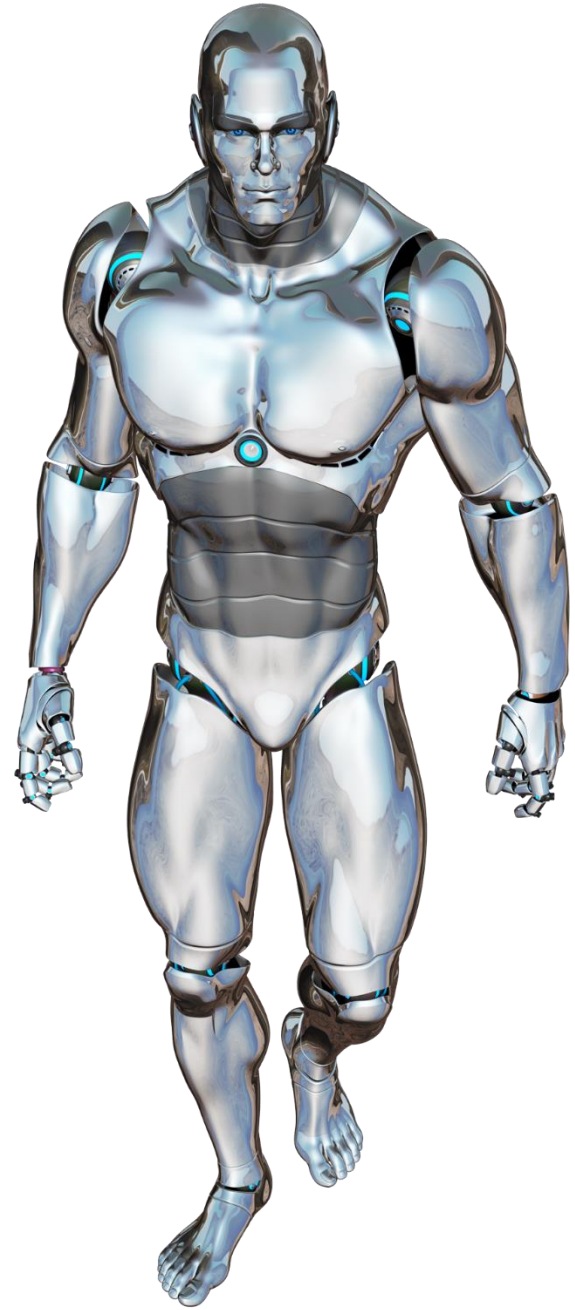
IoT
Internet of Things

A **Internet das Coisas** (do inglês, **Internet** of Things) é uma revolução tecnológica a fim de conectar dispositivos eletrônicos utilizados no dia-a-dia (como aparelhos eletrodomésticos, eletroportáteis, máquinas industriais, meios de transporte etc.) à **Internet**, cujo desenvolvimento depende da inovação técnica dinâmica em ...

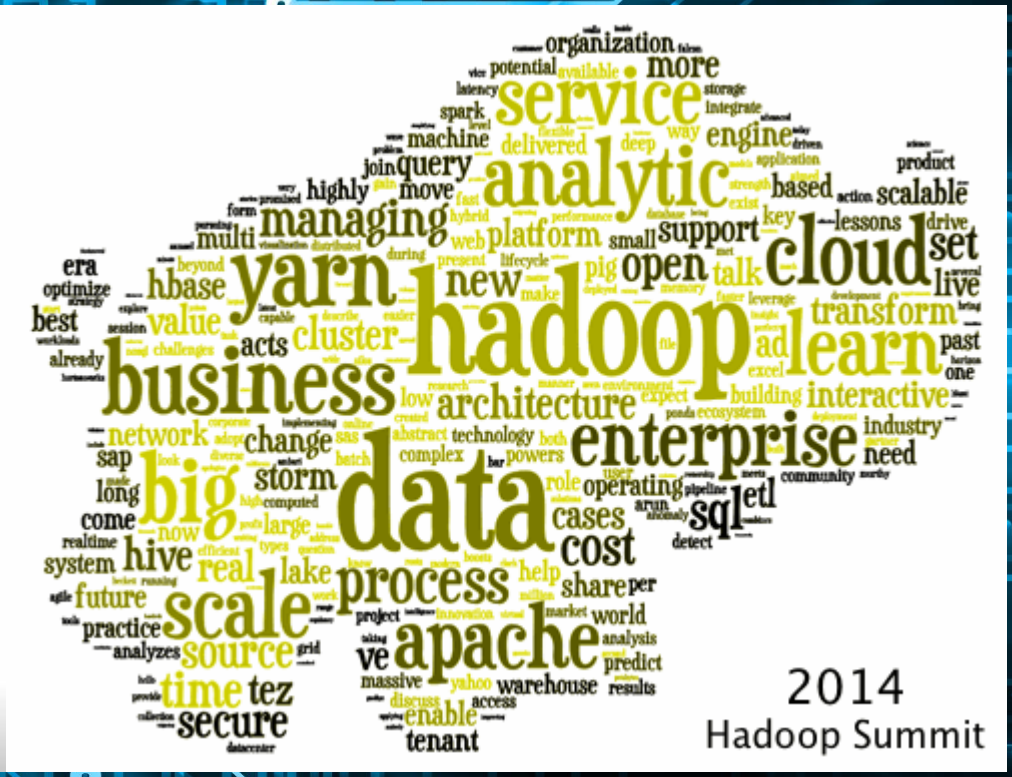
Internet das coisas – Wikipédia, a enciclopédia livre
https://pt.wikipedia.org/wiki/Internet_das_coisas




Intelligent Systems for a More Connected World
WHAT ARE INTELLIGENT SYSTEMS?
Intelligent Systems are devices that transform how we travel, shop, make things and more.
7 Connected Devices per Person
By 2020 each person will own an average of 7 connected devices[1].
COMMUNICATIONS
Connected
Shares data through internet and the cloud
Managed
Can be remotely monitored, updated and power controlled
Secured
Protects data against malware, theft and tampering
#2
Data Breach
Medical data disclosure is the second most breached source of data[2].
MEDICAL
71%
of Shoppers are Multi-Channel...
based on respondents planning their 2011 holiday shopping[3].
RETAIL
23.6M
Connected Cars
8.7
2010
2016
23.6 million cars will have internet access by 2016, rising from 8.7 million in 2010[4].
VEHICLES
30%
Annual Growth Rate
Projected increase in connected machine-to-machine devices over the next 5 years[5].
INDUSTRIAL
intel
Intelligent Systems



- Machine Learning
- Sentiments
- Text Processing
- Image Processing
- Video Analytics
- Log Parsing
- Collaborative Filtering
- Context Search
- Email & Content



Apliçacoes de Big Data

É um Sistema de Armazenamento e Processamento de Dados em alto volume.

Estão presentes em clusters baseados em Linux e compostos por outros softwares Open Sources.

Armazenam Dados, sejam eles Estruturados, semi estruturados ou não estruturados.

HADOOP timeline

1999
Apache Software Foundation (ASF) formed as a non-profit

2002
Nutch created by Doug Cutting and Mike Cafarella

2006
Cutting joins Yahoo, takes Nutch with him

2008
Nutch divided and Hadoop is born

2008
Yahoo releases Hadoop as open-source project to ASF

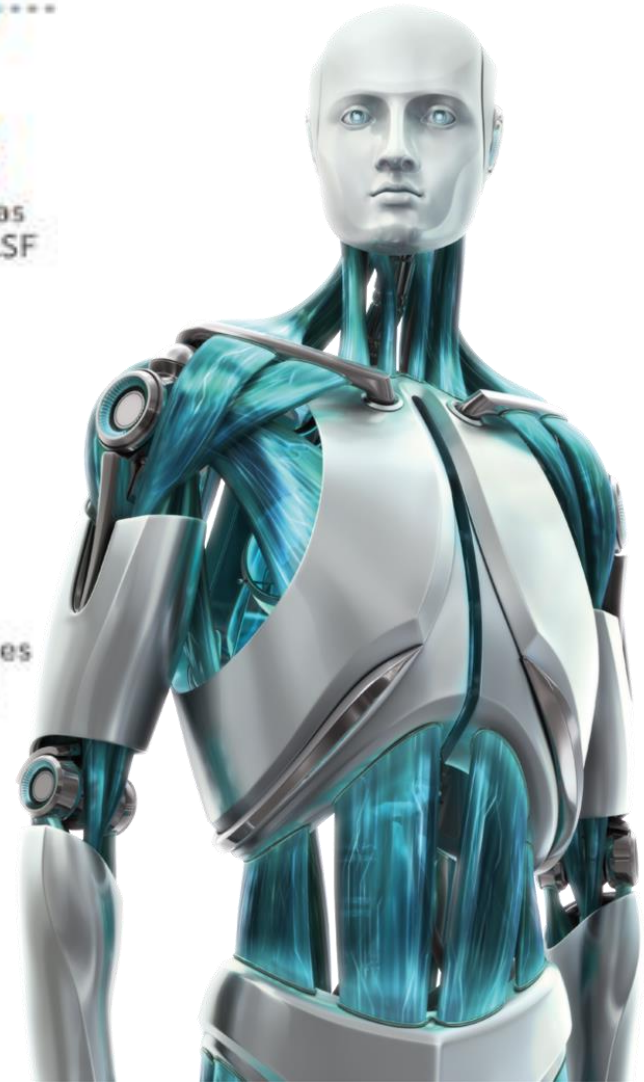
2008
Hadoop-based start-up Cloudera incorporated

2009
Cutting leaves Yahoo for Cloudera

2011
Yahoo spins off Hortonworks as commercial Hadoop distro

2011
MapR Technologies releases Hadoop distro

2013
Greenplum releases Hadoop distro, Pivotal HD



FAMÍLIA DE PROJETOS RELACIONADOS

O Had
relac
distr
hosp
Foun

ojetos

ALTA DISPONIBLIDADE

BALANCEAMENTO DE CARGA

PROCESAMENTO PARALELO

1000 Kilobytes = 1 Megabyte

1000 Megabytes = 1 Gigabyte

1000 Gigabytes = 1 Terabyte

1000 Terabytes = 1 Petabyte

1000 Petabytes = 1 Exabyte

1000 Exabytes = 1 Zettabyte

1000 Zettabytes = 1 Yottabyte

1000 Yottabytes = 1 Brontobyte

1000 Brontobytes = 1 Geobyte

Crescimento Vertical

Memória 64 GB

4 X Pentium I7

HD 10 TB

Utilização de recursos
Em 100%

Troca ou aumento de Processador

Aumento de Memória

Upgrade de HD

hadoop

Crescimento Horizontal

Memória 64 GB

Pentium I7

HD 1 TB

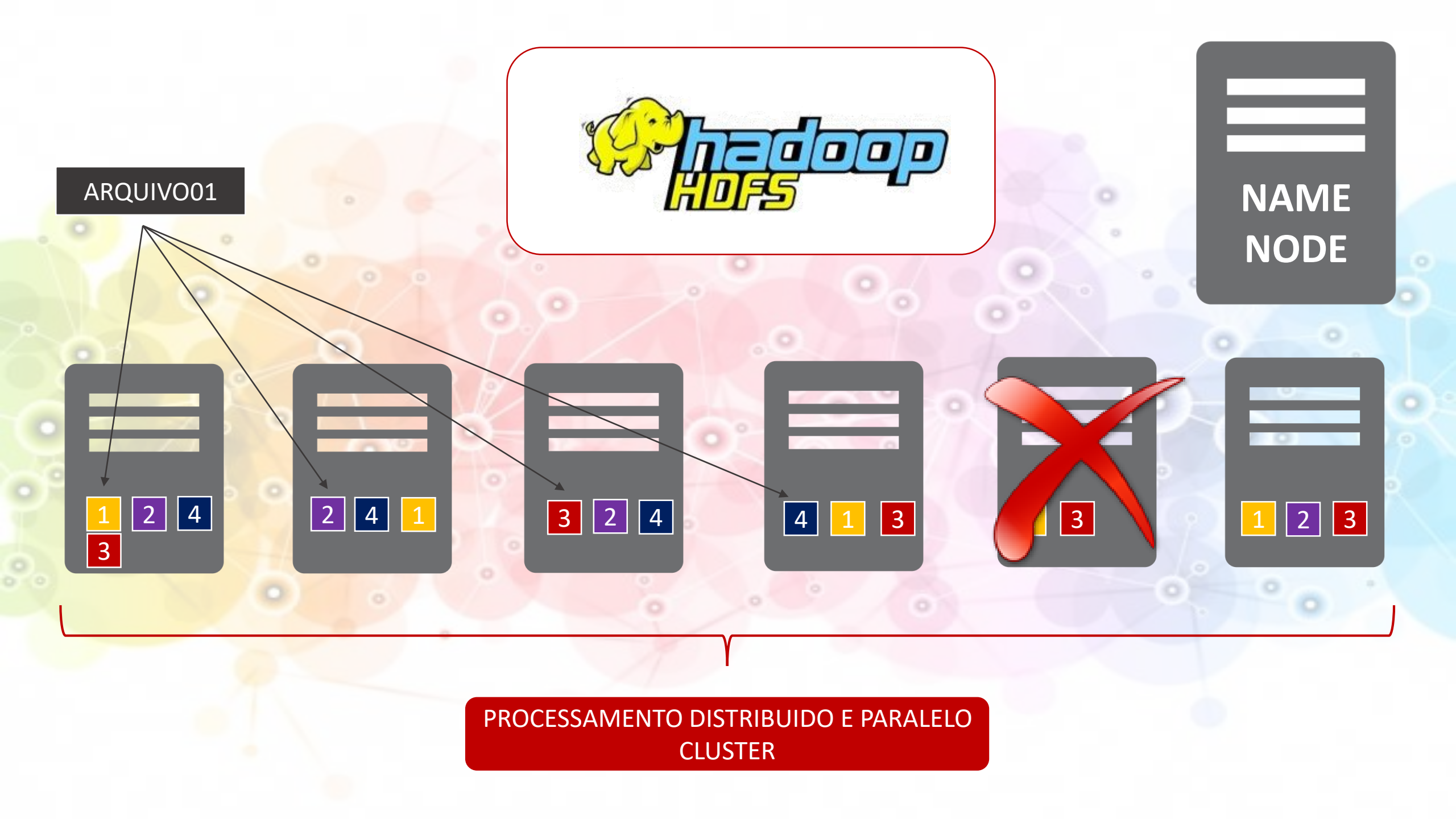
CADA MÁQUINA

Crescimento vertical aumenta a escalabilidade e diminui o custo, uma vez que o hardware utilizado é mais barato.

Utilização de recursos Em 100%

hadoop


hadoop
hadoop
HDFS
hadoop
MapReduce

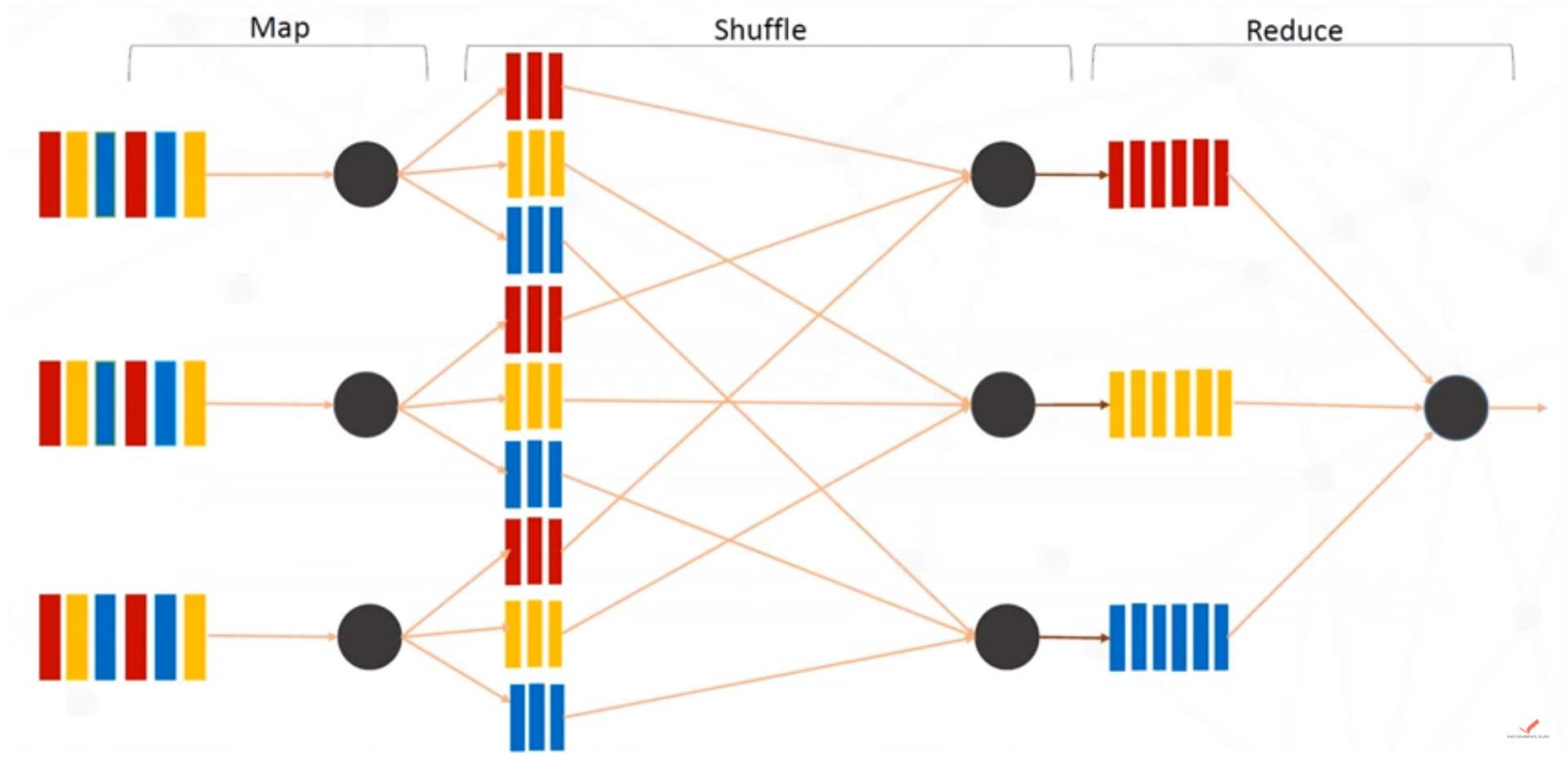

hadoop
HDFS
NAME
NODE
ARQUIVO01
1 2 4 3
2 4 1
3 2 4
4 1 3
3
1 2 3
PROCESSAMENTO DISTRIBUIDO E PARALELO
CLUSTER

Originado pela Google

Modelo de Programação

Utilizado para processar grandes data sets

Altamente escalável



INPUT

MAP

REDUCE

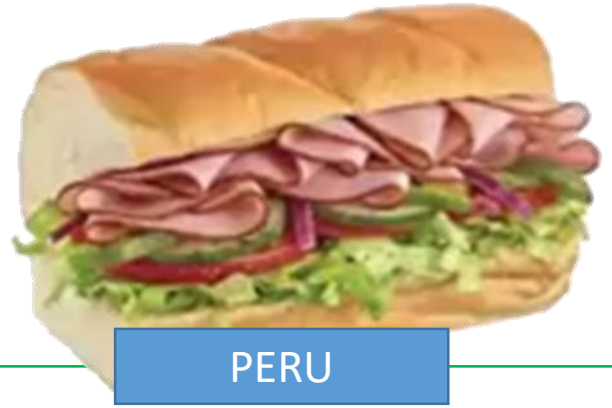
LISTAS ORDENADAS EM NÍVEIS AINDA MENORES

hadoop MapReduce

Fase de Map

ITALIANO


```
{
        “_id”: “sand_italiano”,
        “_ref”: “83722-APO”,
        “pão”: 1,
        “presunto”: 3,
        “salame”: 2,
        “peru”: 2,
        “queijo”: 1,
        “alface”: 3,
        “tomate”:1
}
```

PERU


```
{
        “_id”: “sand_peru”,
        “_ref”: “83722-APA”,
        “pão”: 1,
        “presunto”: 0,
        “salame”: 0,
        “peru”: 3,
        “queijo”: 1,
        “alface”: 3,
        “tomate”:1
}
```


PRESUNTO


```
{
        “_id”: “sand_presunt”,
        “_ref”: “83722-API”,
        “pão”: 1,
        “presunto”: 4,
        “salame”: 0,
        “peru”: 0,
        “queijo”: 1,
        “alface”: 3,
        “tomate”:1
}
```

Fase de Map

Id: “sand_presunt”,
Key: “pão”,
Value: 1

Id: “sand_peru”,
Key: “pão”,
Value: 1

Id: “sand_italiano”,
Key: “pão”,
Value: 1

Id: “sand_presunt”,
Key: “alface”,
Value: 3

Id: “sand_peru”,
Key: “alface”,
Value: 3

Id: “sand_italiano”,
Key: “alface”,
Value: 3

Id: “sand_presunt”,
Key: “tomate”,
Value: 1

Id: “sand_peru”,
Key: “tomate”,
Value: 1

Id: “sand_italiano”,
Key: “tomate”,
Value: 1

Id: “sand_presunt”,
Key: “queijo”,
Value: 1

Id: “sand_peru”,
Key: “queijo”,
Value: 1

Id: “sand_italiano”,
Key: “queijo”,
Value: 1

Id: “sand_presunt”,
Key: “presunto”,
Value: 4

Id: “sand_italiano”,
Key: “presunto”,
Value: 3

Id: “sand_italiano”,
Key: “salame”,
Value: 2

Id: “sand_peru”,
Key: “peru”,
Value: 3

Id: “sand_italiano”,
Key: “peru”,
Value: 2

Fase de Reduce

Key: “pão”,
Value: 3

Key: “alface”,
Value: 3

Key: “tomate”,
Value: 3

Key: “queijo”,
Value: 3

Key: “presunto”,
Value: 7

Key: “salame”,
Value: 2

Key: “peru”,
Value: 5