

# DATA SCIENCE

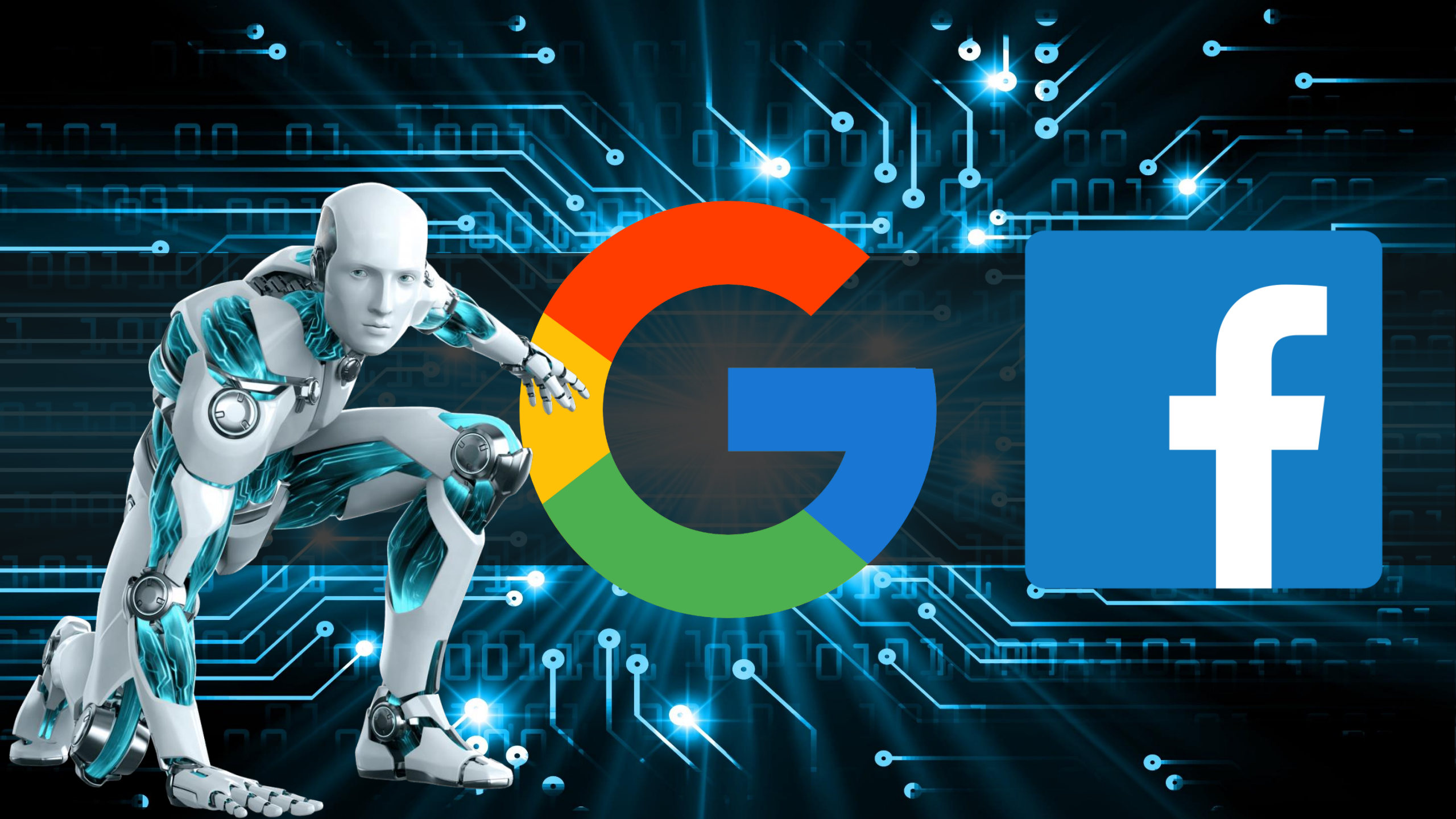






O que é  
Big Data?

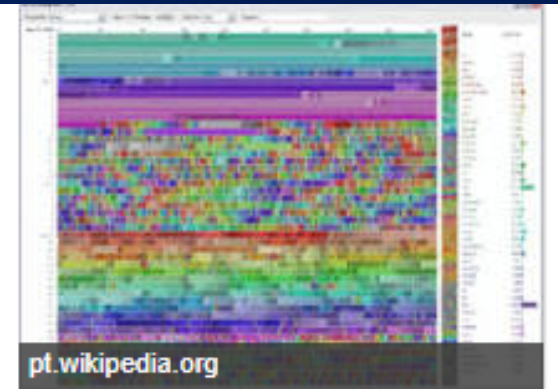








**Big Data** é o termo que descreve o imenso volume de dados – estruturados e não estruturados – que impactam os negócios no dia a dia. Mas o importante não é a quantidade de dados. E sim o que as empresas fazem com os dados que realmente importam.



O que é Big Data? | SAS

[https://www.sas.com/pt\\_br/insights/big-data/what-is-big-data.html](https://www.sas.com/pt_br/insights/big-data/what-is-big-data.html)

Segundo a SAS

## 40 ZETTABYTES

[ 43 TRILLION GIGABYTES ]  
of data will be created by 2020, an increase of 300 times from 2005



## Volume

SCALE OF DATA

### It's estimated that 2.5 QUINTILLION BYTES

[ 2.3 TRILLION GIGABYTES ]  
of data are created each day

Most companies in the U.S. have at least 100 TERABYTES [ 100,000 GIGABYTES ] of data stored

# 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



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



## Variety

DIFFERENT FORMS OF DATA

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



The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

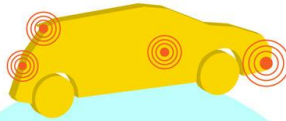
during each trading session



## Velocity

ANALYSIS OF STREAMING DATA

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



1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



Poor data quality costs the US economy around

\$3.1 TRILLION A YEAR



27% OF RESPONDENTS

## Veracity

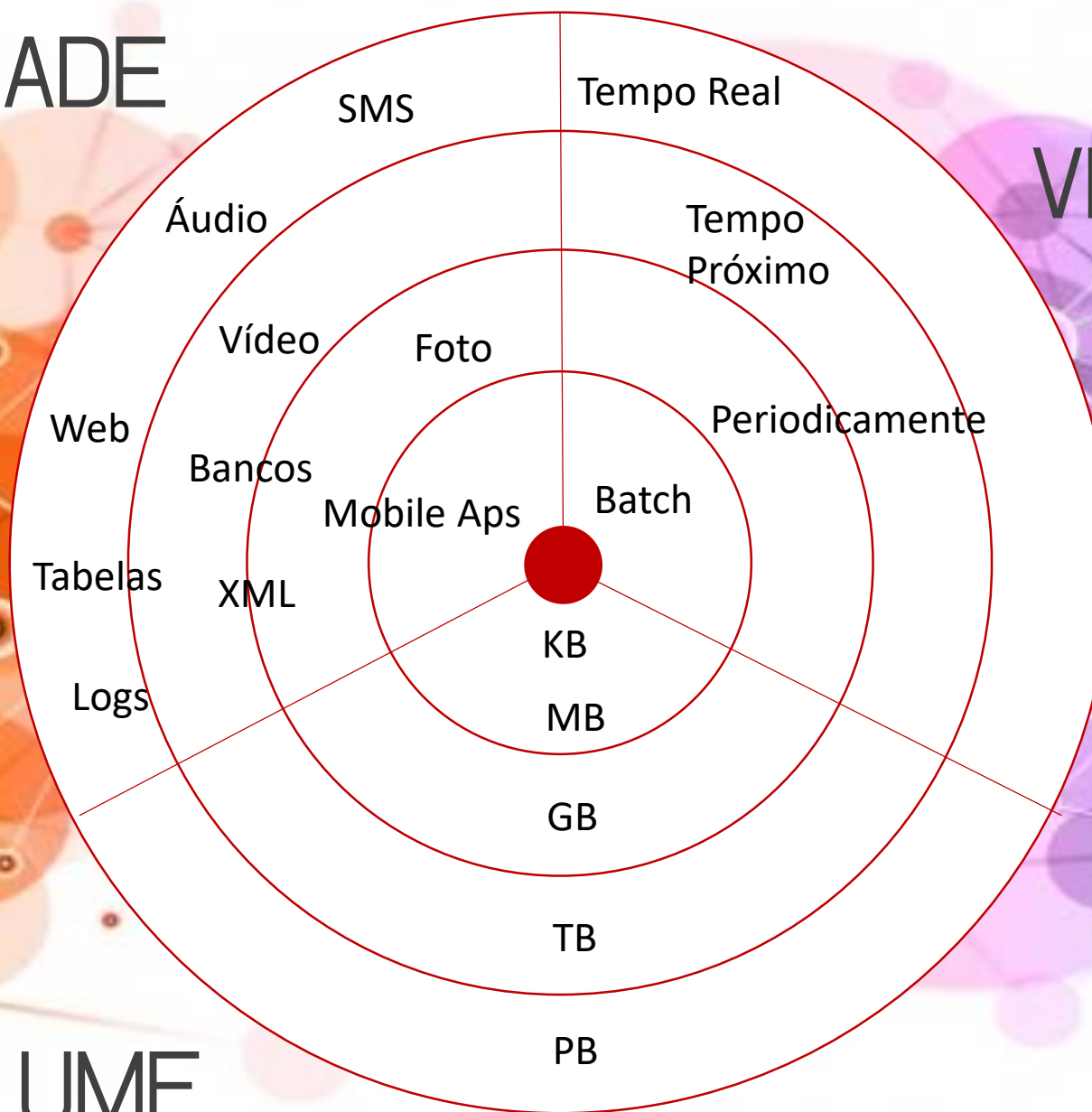
UNCERTAINTY OF DATA

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



VARIEDADE

VELOCIDADE



VOLUME



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



## Big Data Landscape 2016

## Infrastructure

The collage features logos for various data science and cloud computing technologies, organized into four columns:

- Column 1:** Hadoop On-Premise, cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, splice, jethro.
- Column 2:** Hadoop in the Cloud, amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CA PENA, splunk, blue, xplenty.
- Column 3:** Spark, databricks, GridGain, TACHYON.
- Column 4:** Cluster Services, amazon, microsoft, cloudera, doctor, HADOPIWARE, Core OS, BlackIQ.

## Analytics

The image displays a collection of logos for various data science and analytics platforms, organized into four distinct categories, each with a red border and a title. The categories and their respective logos are as follows:

- Analyst Platforms:** Includes Palantir, AYASDI, Quid, and Tableau.
- Analytics Platforms:** Includes Microsoft, gunvus, Domo, and Inverness.
- Data Science Platforms:** Includes Consensus, DataRobot, Alpine, MADR, Dataiku, and others.
- Visualization:** Includes Google, PowerBI, Xignite, Qlik, and CARTO.

## Applications

Sales & Marketing	Customer Service	Human Capital	Legal
 <b>RADIUS Gainsight</b>  <b>bloomreach Zeta</b>  <b>livefyre</b>  <b>bluepond</b>  <b>Lattice</b>  <b>SALSBURY</b>  <b>persado</b>  <b>After Connect</b>  <b>AVITO</b>  <b>ACTIONIQ</b>	 <b>MEDALLIA</b>  <b>ATTERBURY</b>  <b>STELLA Service</b>  <b>NGDATA</b>  <b>Proact</b>  <b>Lixie</b>  <b>Ezupact</b>	 <b>GILD</b>  <b>Connect Kar</b>  <b>Lexipic</b>  <b>Everlane</b>  <b>Breville</b>	 <b>RAVEL</b>  <b>Everlane</b>  <b>Breville</b>

**NoSQL Databases**

- Amazon DynamoDB
- Google Cloud Platform
- Microsoft Azure
- Oracle
- MarkLogic
- MongoDB
- Cassandra
- Hadoop
- Redis
- others

**NewSQL Databases**

- SAP
- Clustrix
- Pivotal
- memsql
- paradigm4
- nuodb
- MariaDB
- VOLTDB
- cloudata
- deep4b
- Trafalgar
- Cockroach Labs

The image displays a collection of logos for various data science and analytics platforms, organized into four distinct categories:

- BI Platforms:** Includes logos for Power BI, Amazon, Tableau, QlikView, Google Data Studio, Alteryx, and others.
- Statistical Computing:** Includes logos for SAS, SPSS, and MATLAB.
- Log Analytics:** Includes logos for Splunk, Sumologic, Hadoop, and Loggly.
- Social Analytics:** Includes logos for NetBase, DataSift, and others.

Ad Optimization

MacMath

Integral

OpenX

theTradeDesk

Security

CYCLANCE

CounterTrack

ThreatMatrix

Recorded Future

Vertical AI Applications

X

Clara

KASIST

The image displays a collection of logos for various AI and data science companies, organized into four columns. The columns are labeled: Real-Time, Machine Learning, Speech & NLP, and Horizontal AI. The logos include: Amazon, MIT, IBM, Google, Microsoft, Facebook, Twitter, LinkedIn, and many others.

<p><b>Publisher Tools</b></p> <p>Outbrain</p> <p>mixpanel</p> <p>Chartbeat</p>	<p><b>Govt / Regulation</b></p> <p>Socrata</p> <p>OPENGOV</p> <p>FiscalNote</p> <p>enigma</p>	<p><b>Finance</b></p> <p>Affirm</p> <p>LendingClub</p> <p>OnDeck</p> <p>Kreditech</p> <p>tidemark</p> <p>INSIKT</p> <p>Dataminr</p>
--	---	---

Management / Monitoring	Security	Storage	App Dev	Crowd-sourcing
 New Relic  APDYNAMICS  Amazon CloudWatch  achtio  Humidity  splunk  Google Cloud  Pivotal	 TANIUM  Cisco Duo  COBALT.io  DataGravity  CyberCortex  VECTRA  Palo Alto Networks  Cisco Duo	 PANASAS  Google Cloud Storage  Microsoft Azure  PANASAS  NetScout Systems  Quanto	 apigee  Cisco Duo  Typepath  Conviva	 Amazon Mechanical Turk  CrowdFlower  iStockphoto  Ponemon  iStockphoto

The image displays a collection of 16 logos for data science and business intelligence companies, organized into four distinct categories:

- Search:** Includes logos for Apache Solr, Elasticsearch, Lucidworks, Elastic, Thompson, Swiftype, Algolia, and Synonym.
- Data Services:** Includes logos for Oracle, OPERA, IBM, and KAGGLE.
- For Business Analysts:** Includes logos for Original, ClearStory, CIRRO, and Import.
- SMB / Commerce:** Includes logos for Google Analytics, Amazon, Bluecore, Sumo, Grify, Retention, and Custom.

**Education/Learning**

Yellinco mark43 KENSHO AIDA ISENTIUM  
Quantopian eSentient

**Life Sciences**

X Courag Recombi FLATIRON cymogen

**Industries**

OPower eHarmony RetailNext STITCH FIX Workfusion TACHYUS durtto

### Cross-Infrastructure/Analytics

Amazon Google Microsoft IBM SAP SAS HP VMware Talend TIBCO TERADATA ORACLE NetApp

Open source

Framework: Hadoop, Spark, YARN, Mesos, Tez, Flink, CDAP, Hive, Pig, Mahout, Hama

Query / Data Flow: Hive, Pig, Mahout, Hama

Data Access: HBase, Cassandra, MongoDB, Redis, Riak, Aerospike, Aerosole

Coordination: Hadoop Distributed Cache, Apache ZooKeeper, Apache HBase

Real-Time: Storm, Spark, Flink, Druid

Stat Tools: Scala, R, Python, SoPy

Machine Learning: mllib, Apache Mahout, Weka, Caffe, FeatureFusion, DIMSUM

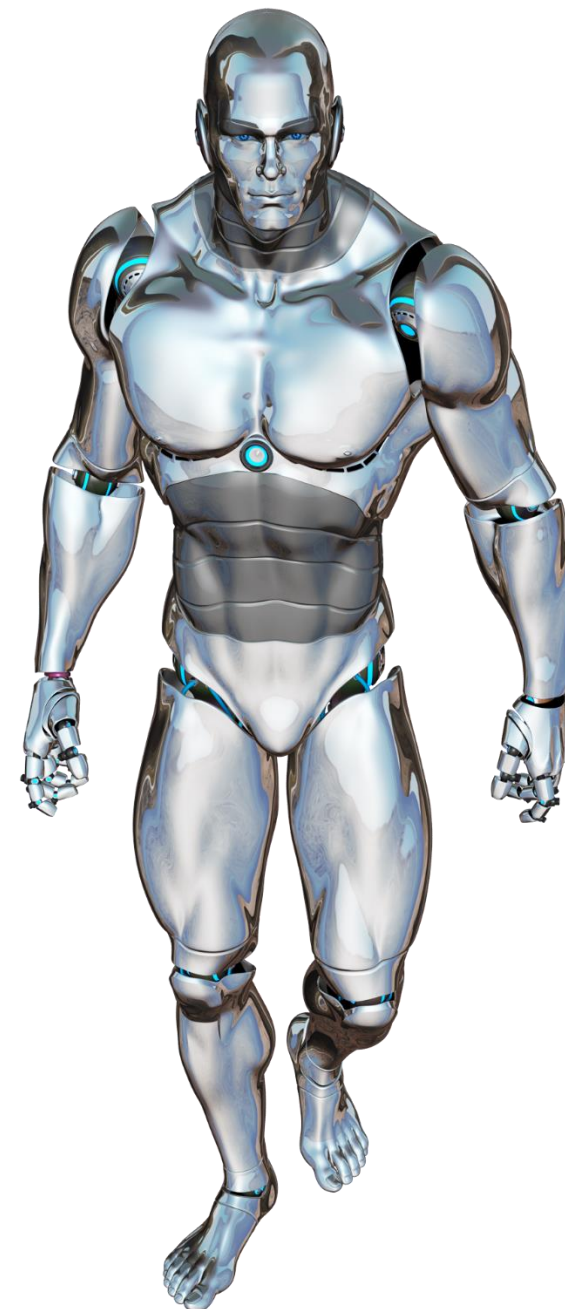
Search: ELK, Solr, Elasticsearch, DL4J

Security: Apache Ranger, Visualization

## Data Sources & APIs

The bar chart displays the following data sources categorized by type:

- Health:** Apple, Jawbone, Garmin, Fitbit, Withings, Validic, Humana API.
- IOT:** Uptake, ThingWorx, Samsara, Inetum.
- Financial & Economic Data:** Bloomberg, Dow Jones, Yahoo!, Premise, S&P Capital IQ, Quandl, Xignite, CB Insights, Morningstar, Gooddata, Plaid.
- Air / Space / Sea:** Planet Labs, Spire, Earthstar, Airbus, Cruise, Bluebird, Sanjit.
- Location / People / Entities:** Garmin, FourSquare, InsideView, Esri, Streetline, Commvault, Factful, Placemeter, Calsonic Navigator, Glomax, Basis, Samsara, DataCamp, Insight, DataElio, The Data Incubator.
- Other:** GA, DataCamp, Insight, DataElio, The Data Incubator.







# What Happens in an Internet Minute?

1,572,877 GB of global IP data transferred<sup>1</sup>



## And Future Growth is Staggering



By 2017, mobile traffic will have grown **13X** in just 5 years<sup>1</sup>



In 2017, there will be **3X** more connected devices than people on Earth<sup>1</sup>

All digital data created reached **4 zettabytes** in 2013<sup>1</sup>





# IoT

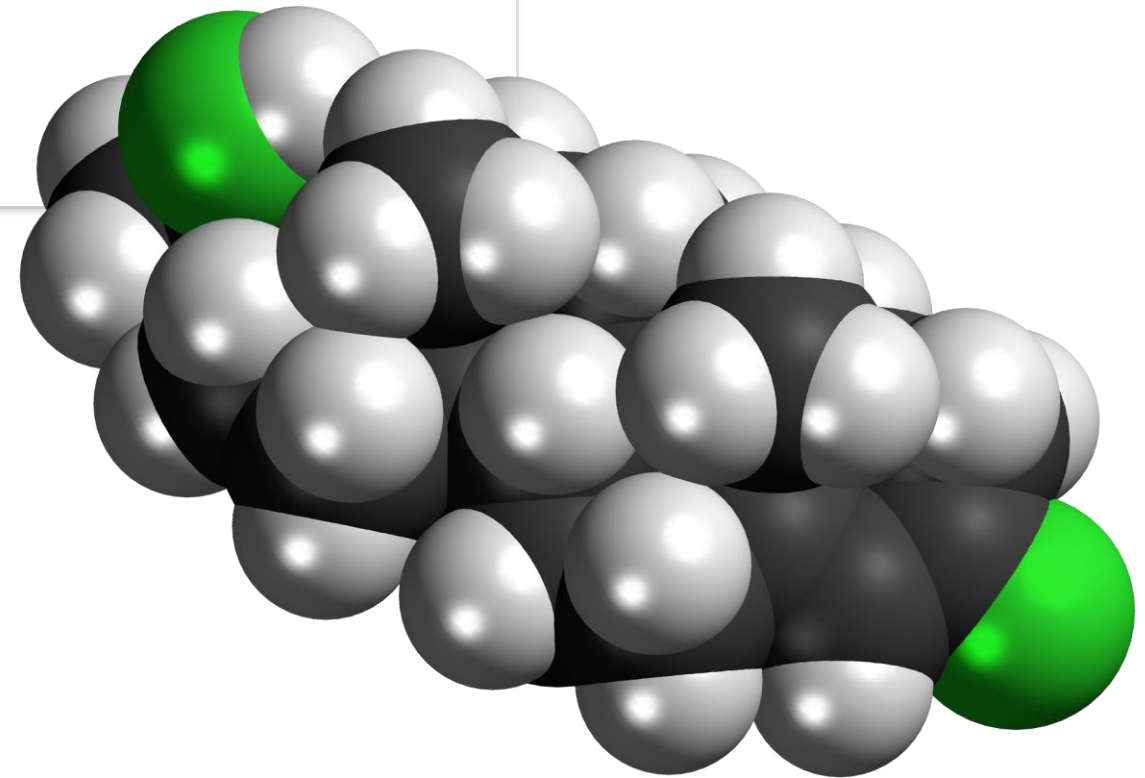
## Internet of Things





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](https://pt.wikipedia.org/wiki/Internet_das_coisas)





## WHAT ARE INTELLIGENT SYSTEMS?

7 Connected Devices per Person

By 2020 each person will own an average of 7 connected devices.

71%

**of Shoppers  
are Multi-Channel...**  
based on respondents  
planning their 2011  
holiday shopping).

23.6M  
Connected  
Cars

8.7

23.6 million cars will have Internet access by 2016, rising from 8.7 million in 2010.

## #2

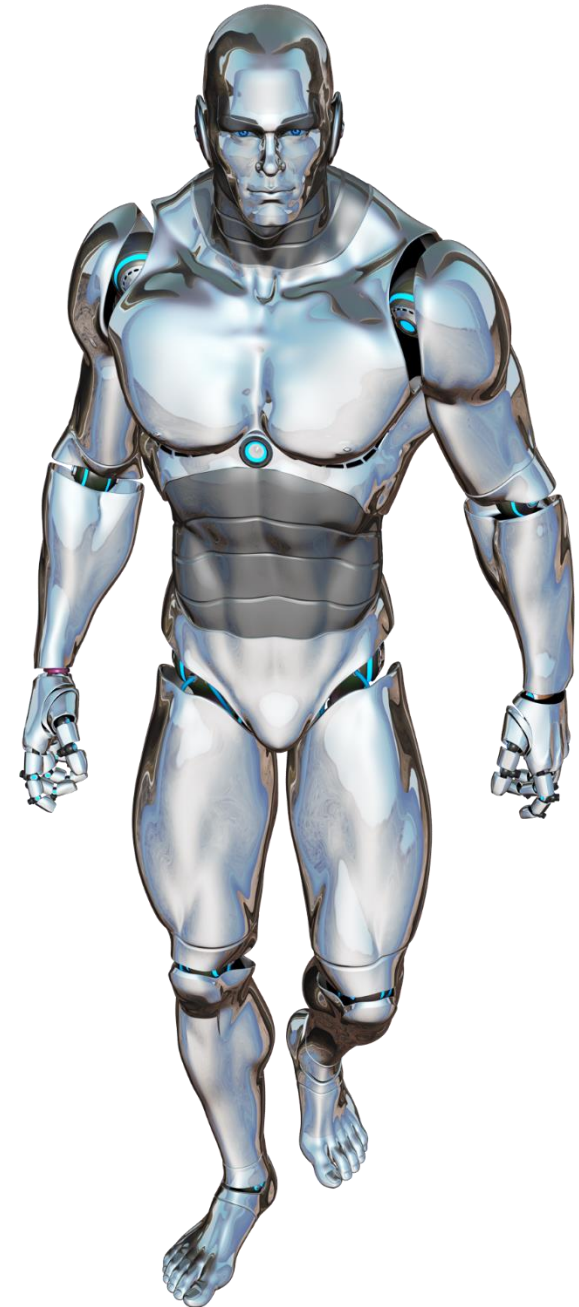
**Data Breach**  
Medical data disclosure is the second most breach

30%

**Annual Growth Rate**  
Projected increase in connected machine-to-machine devices over the next 5 years\*



Intel  
Intelligent  
Systems

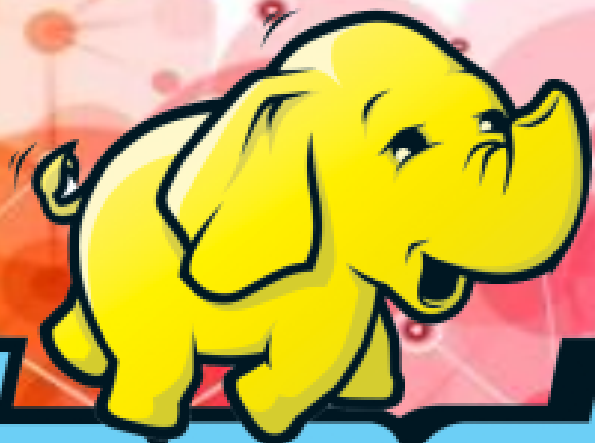






# Aplicacoes de Big Data





**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.**



**1999**

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

**2002**

Nutch created by Doug Cutting and Mike Cafarella

**2008**

Nutch divided and Hadoop is born

**2006**

Cutting joins Yahoo, takes Nutch with him

**2008**

Yahoo releases Hadoop as open-source project to ASF

# HADOOP timeline

**2008**

Hadoop-based start-up Cloudera incorporated

**2009**

Cutting leaves Yahoo for Cloudera

**2011**

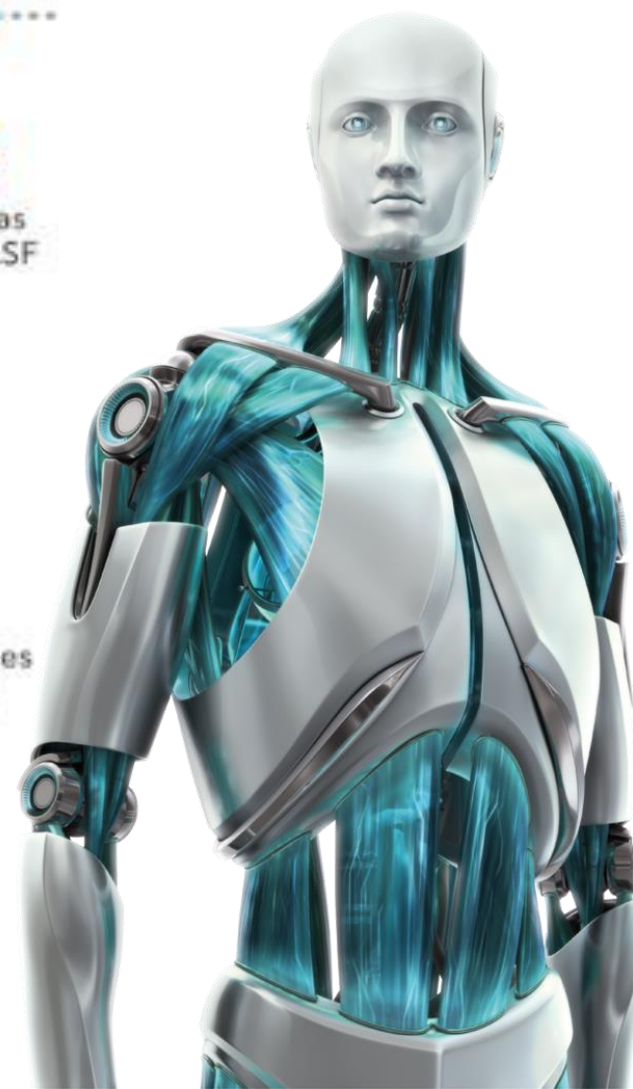
MapR Technologies releases Hadoop distro

**2013**

Greenplum releases Hadoop distro, Pivotal HD

**2011**

Yahoo spins off Hortonworks as commercial Hadoop distro





The background features a complex network of overlapping circles in various colors (green, yellow, orange, red, purple, blue) connected by thin lines, creating a molecular or network-like structure. A large, solid yellow circle is centered on the page, serving as a backdrop for the main title.

# FAMÍLIA DE PROJETOS RELACIONADOS

O Had  
relaci  
distr  
hosp  
*Found*

Projetos





ALTA DISPONIBILIDADE

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





# 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%



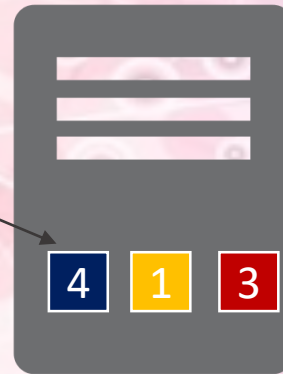
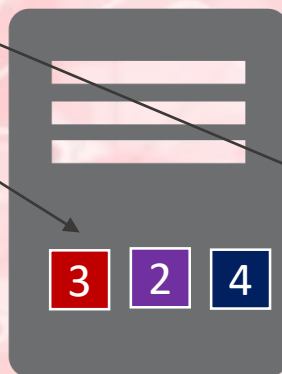
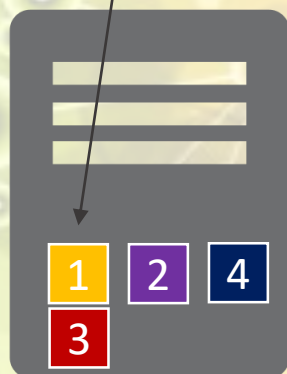






NAME  
NODE

ARQUIVO01



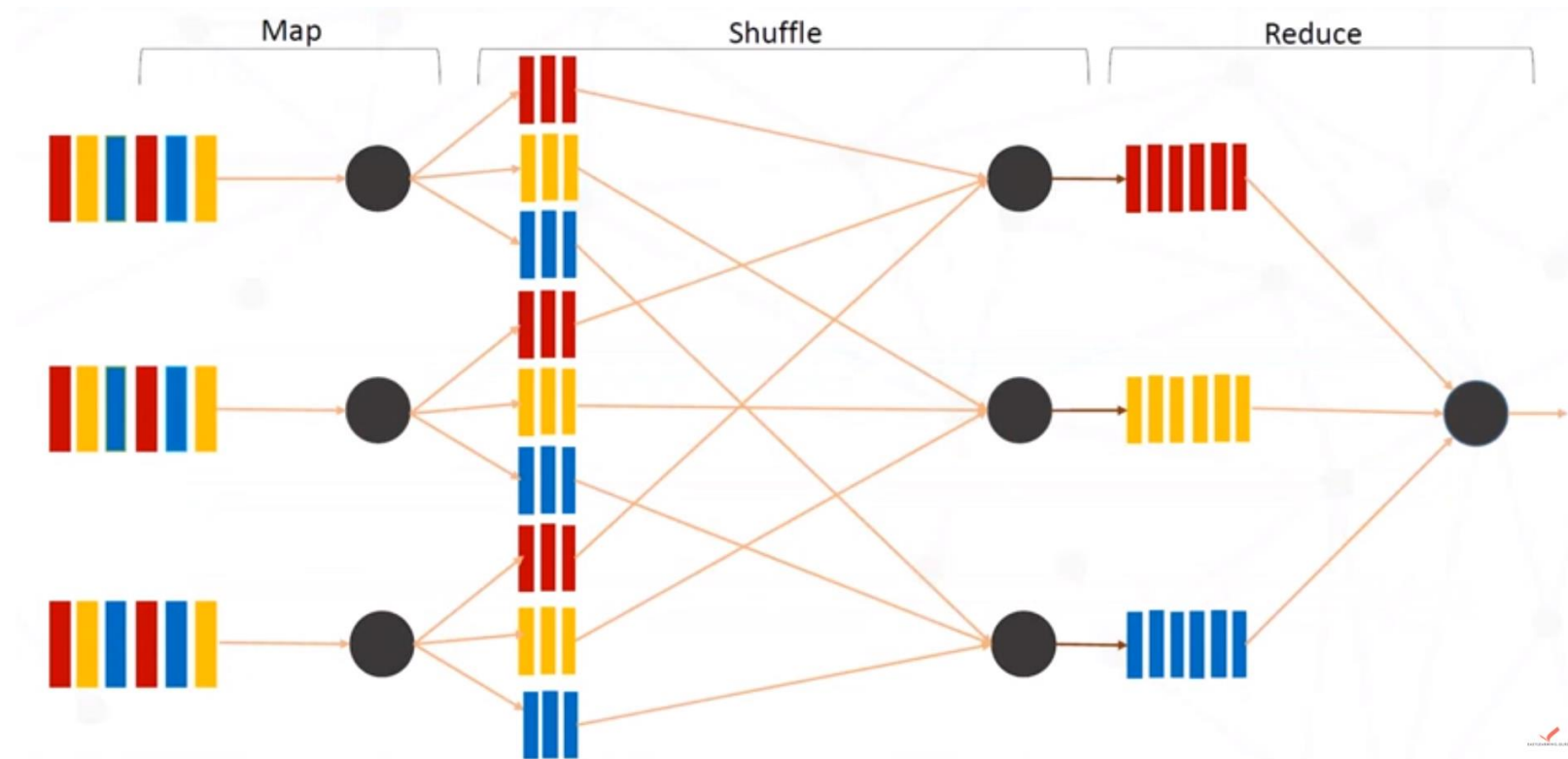
PROCESSAMENTO DISTRIBUIDO E PARALELO  
CLUSTER

Originado pela Google

Modelo de Programação

Utilizado para processar grandes data sets

Altamente escalável





INPUT

MAP

REDUCE



ITALIANO



PERU



PRESUNTO



LISTAS ORDENADAS EM  
NÍVEIS AINDA MENORES

# 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\_italiano",  
Key: "alface",  
Value: 3

Id: "sand\_peru",  
Key: "queijo",  
Value: 1

Id: "sand\_peru",  
Key: "peru",  
Value: 3

Id: "sand\_peru",  
Key: "pão",  
Value: 1

Id: "sand\_presunt",  
Key: "tomate",  
Value: 1

Id: "sand\_italiano",  
Key: "queijo",  
Value: 1

Id: "sand\_italiano",  
Key: "peru",  
Value: 2

Id: "sand\_italiano",  
Key: "pão",  
Value: 1

Id: "sand\_peru",  
Key: "tomate",  
Value: 1

Id: "sand\_presunt",  
Key: "presunto",  
Value: 4

Id: "sand\_presunt",  
Key: "alface",  
Value: 3

Id: "sand\_italiano",  
Key: "tomate",  
Value: 1

Id: "sand\_italiano",  
Key: "presunto",  
Value: 3

Id: "sand\_peru",  
Key: "alface",  
Value: 3

Id: "sand\_presunt",  
Key: "queijo",  
Value: 1

Id: "sand\_italiano",  
Key: "salame",  
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

