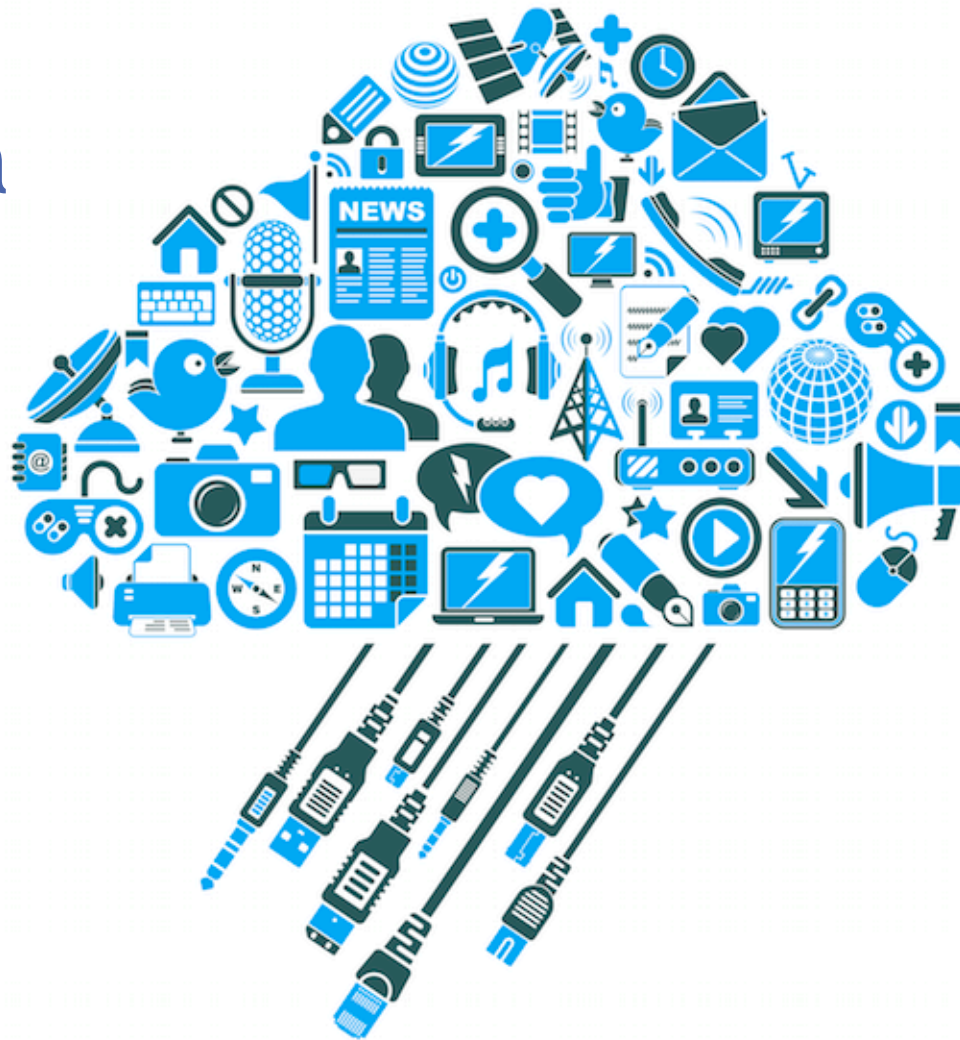


Data Science e Big Data



Prof. Dr. Vladimir Costa Alencar
LANA - UEPB

valencar@servidor.uepb.edu.br

www.valencar.com



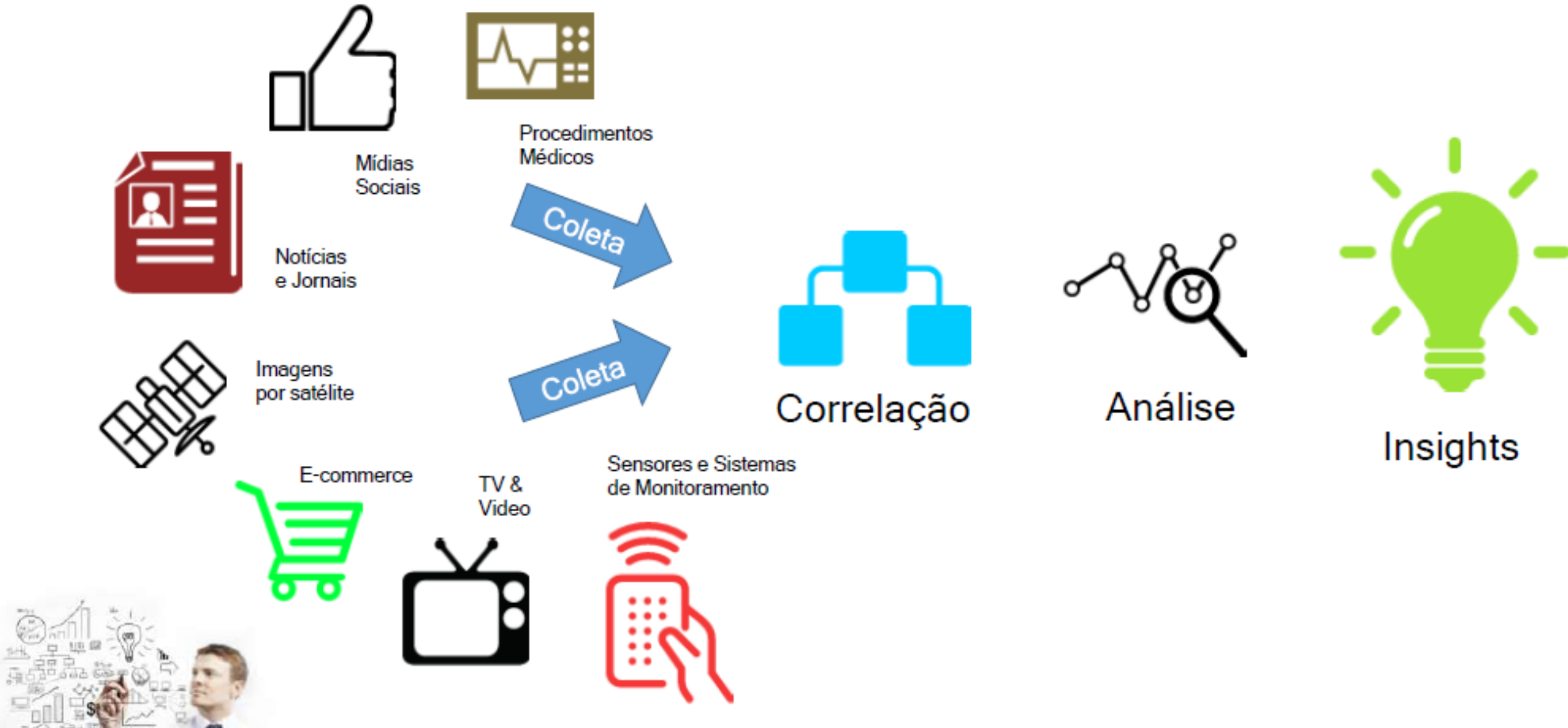
LANA
DATA ANALYSIS LABORATORY

Data Science e Big Data



E O que são dados?

Data Science X Big Data



APR
2020

DIGITAL AROUND THE WORLD IN APRIL 2020

THE ESSENTIAL HEADLINE DATA YOU NEED TO UNDERSTAND MOBILE, INTERNET, AND SOCIAL MEDIA USE

TOTAL
POPULATION



**7.77
BILLION**

URBANISATION:
55%

UNIQUE MOBILE
PHONE USERS



**5.16
BILLION**

PENETRATION:
66%

INTERNET
USERS



**4.57
BILLION**

PENETRATION:
59%

ACTIVE SOCIAL
MEDIA USERS



**3.81
BILLION**

PENETRATION:
49%



we
are
social



2020 *This Is What Happens In An Internet Minute*



2019 *This Is What Happens In An Internet Minute*



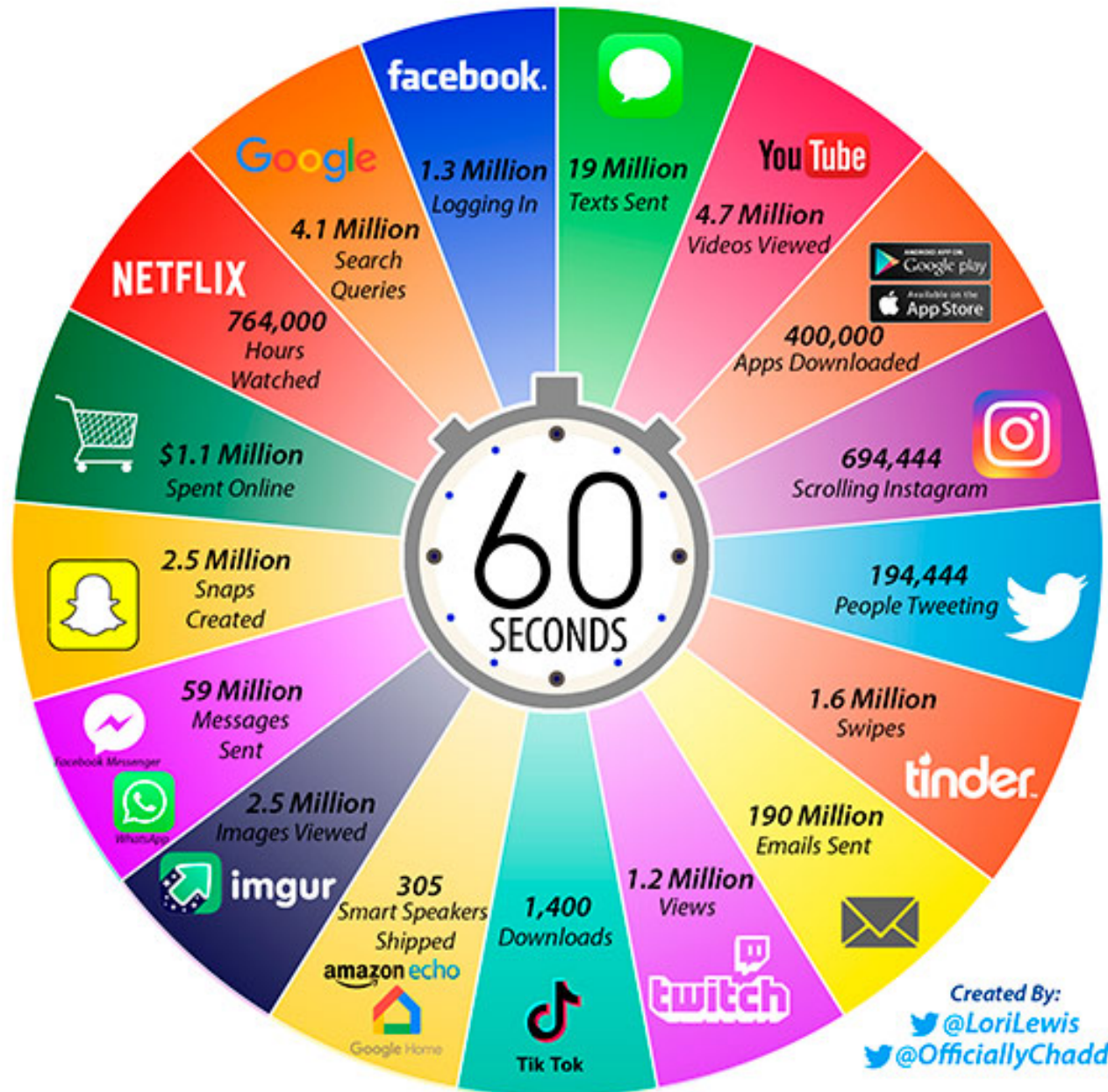
2018 *This Is What Happens In An Internet Minute*



2019 *This Is What Happens In An Internet Minute*



2020 *This Is What Happens In An* Internet Minute



APR
2020

DIGITAL AROUND THE WORLD IN APRIL 2020

THE ESSENTIAL HEADLINE DATA YOU NEED TO UNDERSTAND MOBILE, INTERNET, AND SOCIAL MEDIA USE

TOTAL
POPULATION



7.77
BILLION

URBANISATION:
55%

UNIQUE MOBILE
PHONE USERS



5.16
BILLION

PENETRATION:
66%

INTERNET
USERS



4.57
BILLION

PENETRATION:
59%

ACTIVE SOCIAL
MEDIA USERS



3.81
BILLION

PENETRATION:
49%

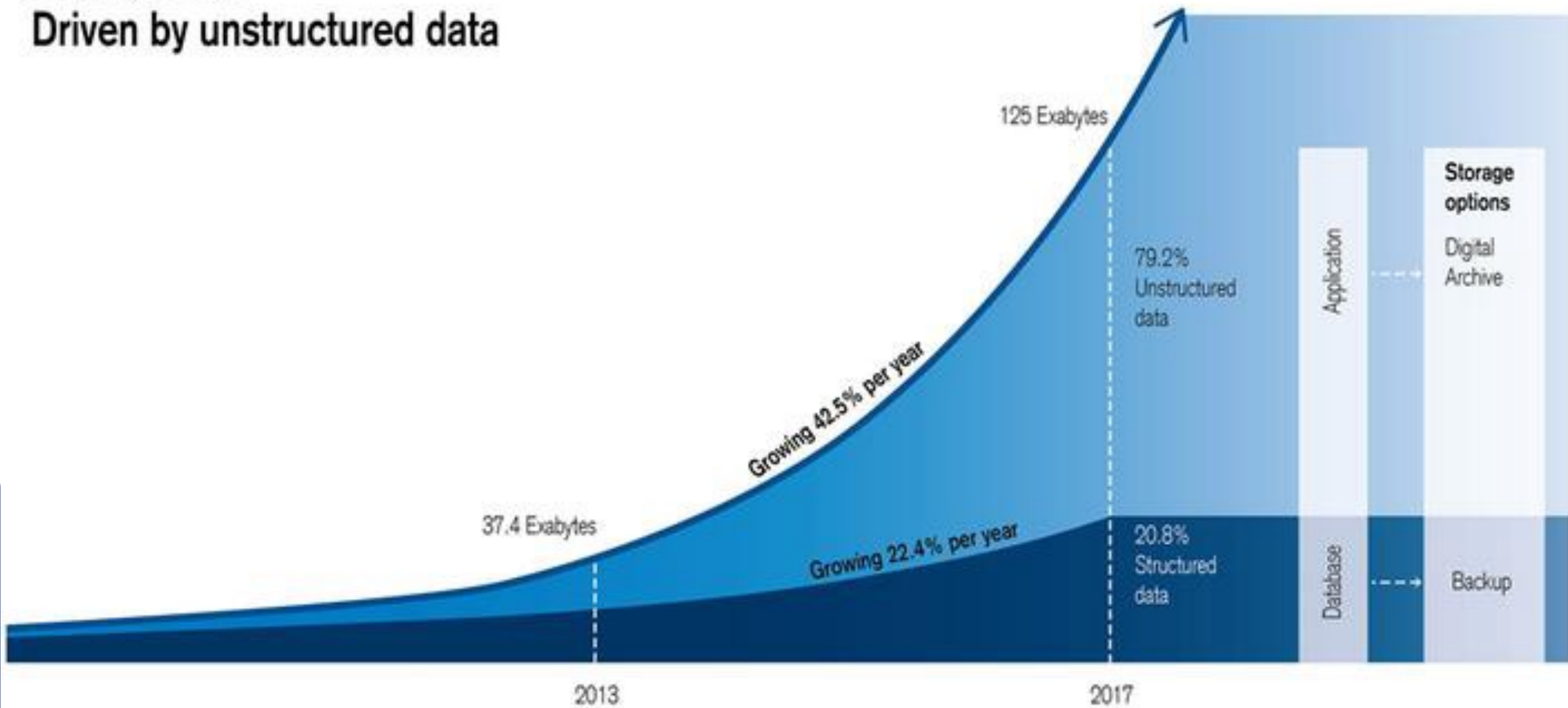


we
are
social



Data growth

Driven by unstructured data



● **big data**
Termo de pesquisa

● **artificial intelligence**
Termo de pesquisa

● **Data Science**
Termo de pesquisa

+ Adicionar comparação

Todo o mundo ▼

2004 - presente ▼

Todas as categorias ▼

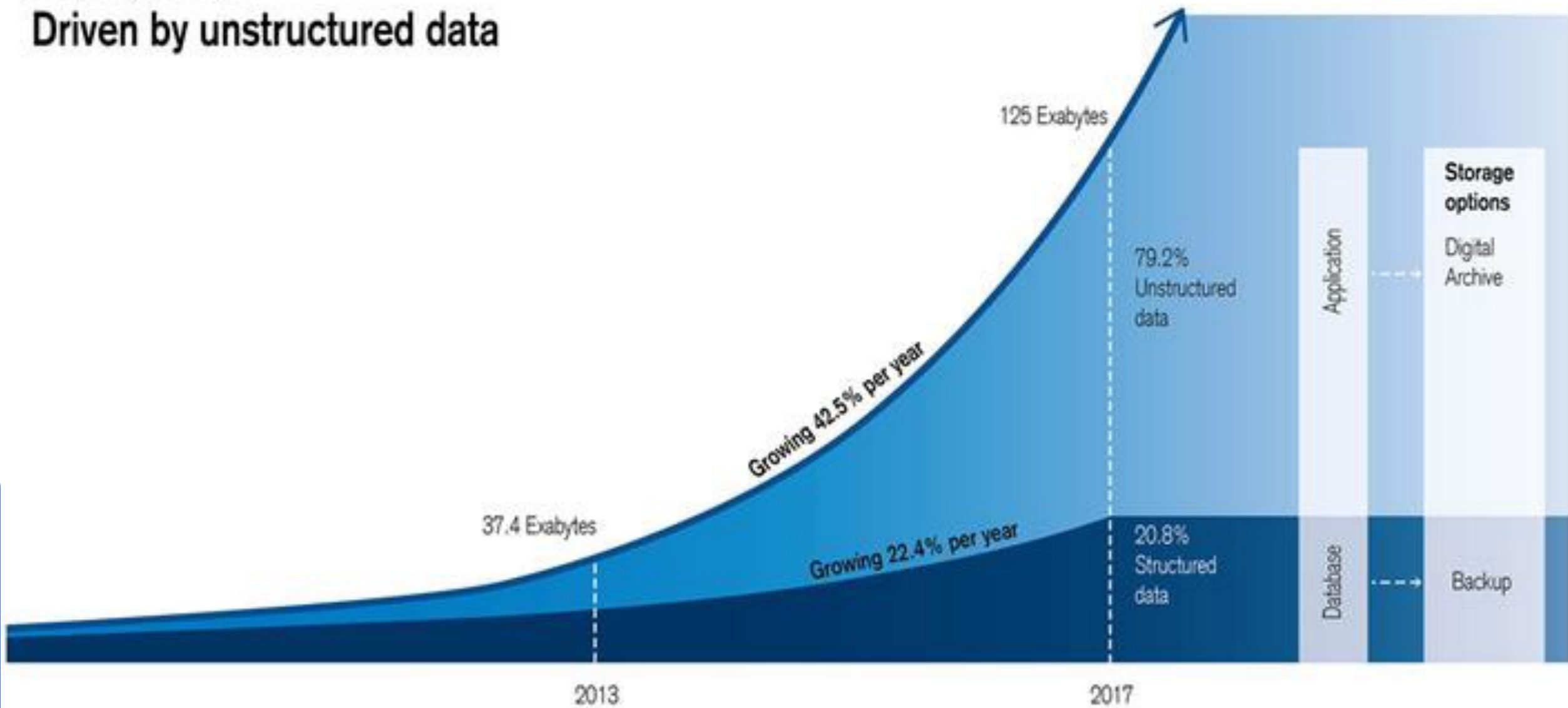
Pesquisa na Web ▼

Interesse ao longo do tempo



Data growth

Driven by unstructured data





110
Accounts Created
57000
Tweets



23140
Video Hours
Watched

20
Video Hours
Uploaded



1820
User Searches



231480
Minutes Used



185190 Likes
10000 Comments
6940 Uploaded



57870
+1s



230
Blog Posts



46080
Searches
\$16020
Ad Revenue



10 Posts
130 Comments
2120 Votes



4630
Posts



2380
Pins



510
Items Purchased
\$23590
Money Spent



350
Check-Ins



5
Reviews



34027780
Emails Sent



115740
Files Saved



57870
Stories Viewed
81020
Messages Sent



6340
App Downloads



12360
App Downloads



521960 Likes
549760 Posts
60 GB of Data



WhatsApp

120
Accounts Created
2199070
Messages Sent



3860
Hours Watched



10190
Hours Streamed



By the way, in the 10 seconds you've been on this page,
approximately 225740 GB of data was transferred over the internet.



\$11740
Profit
\$54190
Revenue



\$480
Profit
\$2490
Revenue



\$90
Profit
\$23610
Revenue



\$4090
Profit
\$17610
Revenue



\$6930
Profit
\$24690
Revenue



\$1120
Revenue



\$1370
Revenue



\$430
Profit
\$1480
Revenue



\$1780
Revenue



\$1740
Revenue



\$630
Revenue



\$1940
Revenue



\$40
Profit
\$1390
Revenue



\$40
Revenue



\$1130
Revenue



\$10
Profit
\$490
Revenue



-\$10
Profit
\$140
Revenue



\$40
Profit
\$70
Revenue



-\$200
Profit
\$210
Revenue



-\$2.5
Profit
\$80
Revenue

By the way, in the 10 seconds you've been on this page, these internet giants have profited \$23630, and around 50% of that went to Apple.

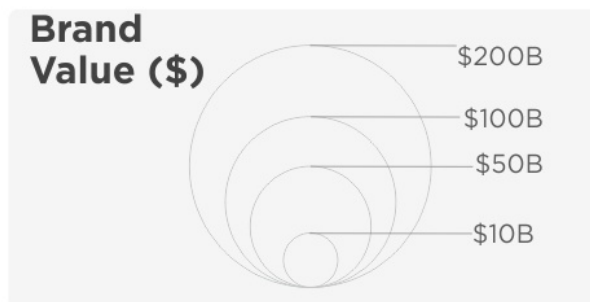
Top 10 Most Valuable Brands

	1	← 1			2020: \$220,791m 2019: \$187,905m	+17.5%
	2	↑ 3			2020: \$159,722m 2019: \$142,755m	+11.9%
	3	↓ 2			2020: \$140,524m 2019: \$153,634m	-8.5%
	4	← 4			2020: \$117,072m 2019: \$119,595m	-2.1%
	5	← 5			2020: \$94,494m 2019: \$91,282m	+3.5%

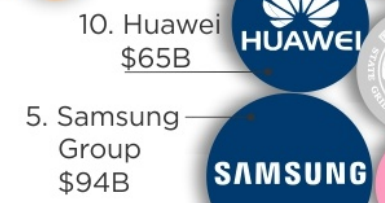
	6	↑ 8			2020: \$80,791m 2019: \$79,823m	+1.2%
	7	← 7			2020: \$79,804m 2019: \$83,202m	-4.1%
	8	↑ 11			2020: \$77,520m 2019: \$67,867m	+14.2%
	9	↑ 14			2020: \$69,041m 2019: \$57,626m	+19.8%
	10	↑ 12			2020: \$65,084m 2019: \$62,278m	+4.5%

The Top 100 Most Valuable Brands in 2020

Companies by Brand Valuation Around the World



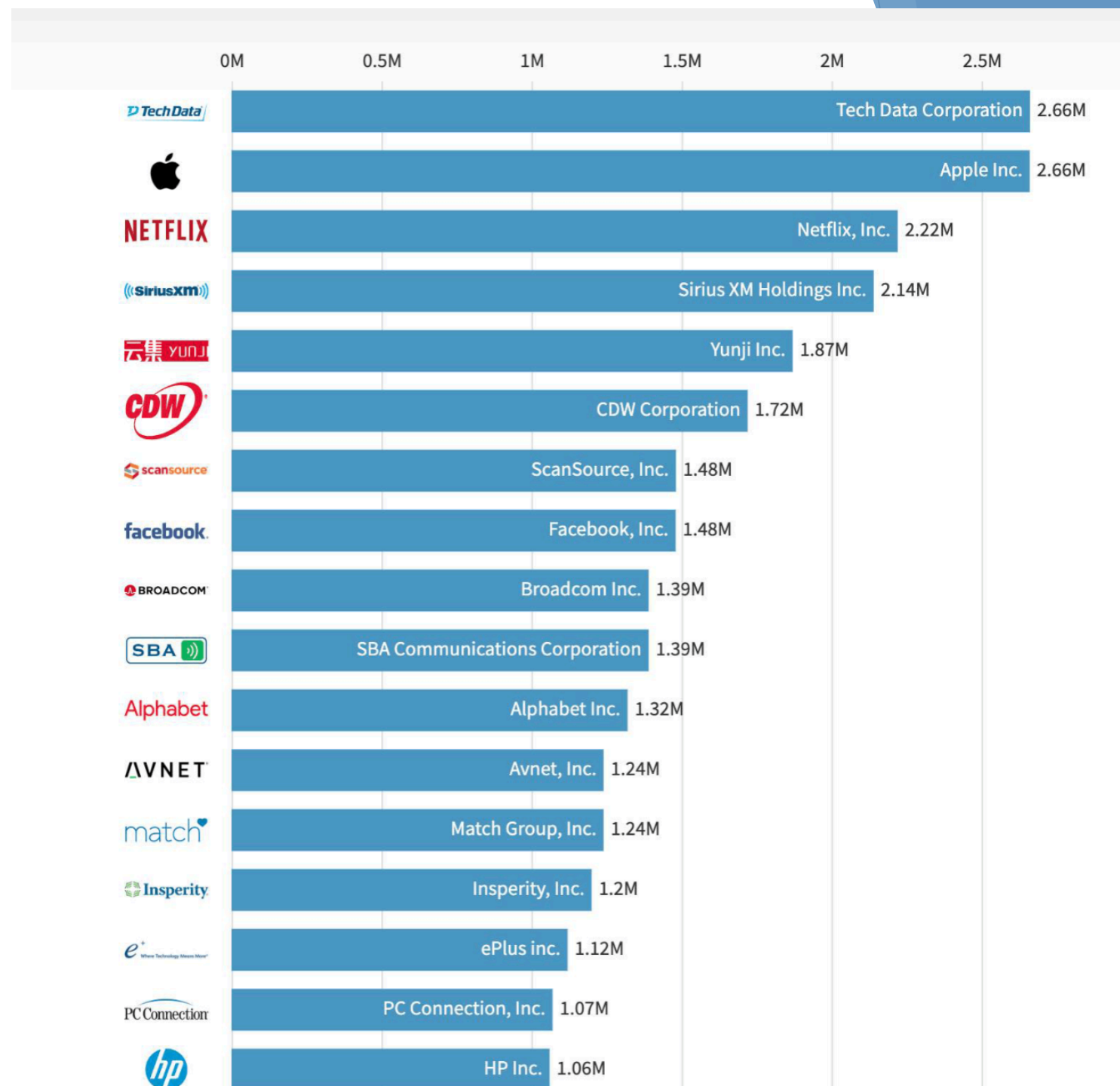
Europe



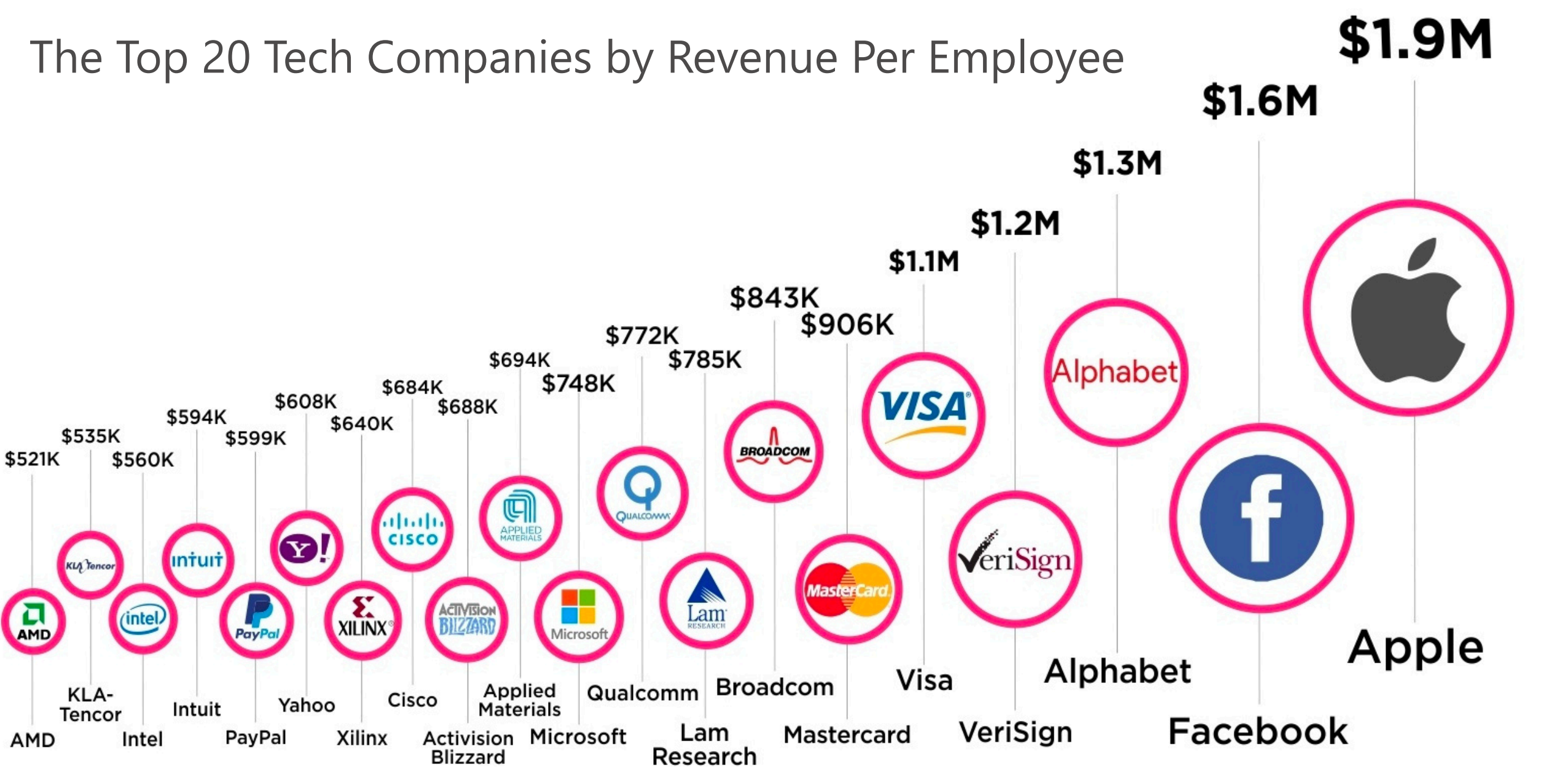
Asia



The Top Tech Companies by Revenue Per Employee (por receita)



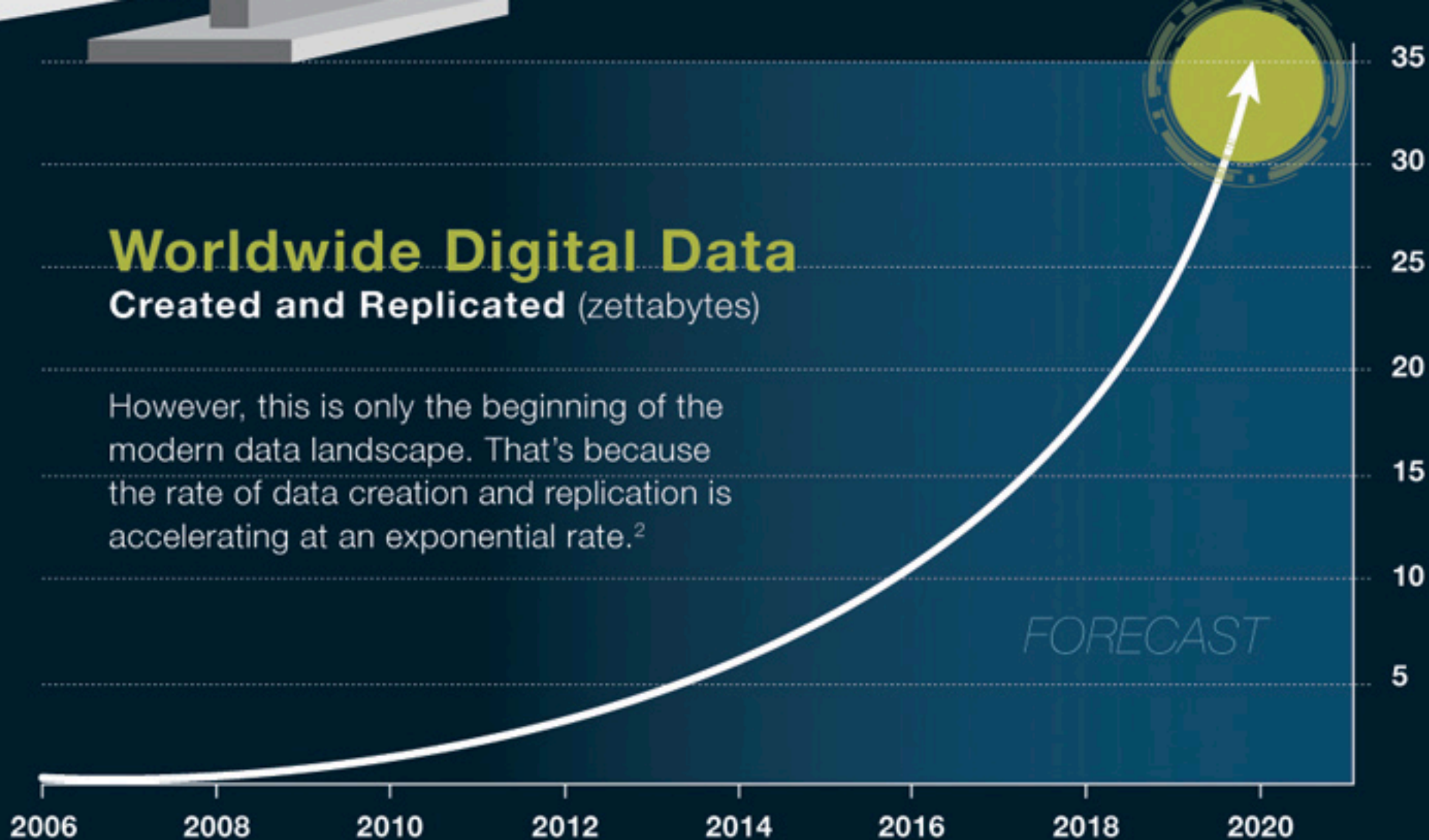
The Top 20 Tech Companies by Revenue Per Employee



Sources:
<https://howmuch.net/articles/tech-companies-revenue-per-employee>
<https://craft.co/reports/s-p-500-revenue-per-employee-perspective>
Data Sourced from Company Filings

Worldwide Digital Data Created and Replicated (zettabytes)

However, this is only the beginning of the modern data landscape. That's because the rate of data creation and replication is accelerating at an exponential rate.²



WHERE DOES ALL OF THIS DATA COME FROM?

There are more sources of data than ever before:

Internal

Archives

Scanned documents and records between organizations and customers

External

Public Web

Government, economic, census, and other data sources published

Both

Social Media

Twitter, LinkedIn, Facebook, YouTube, and Slideshare data

Documents

Email, Word, Excel, PDF, PPT, HTML, and plain text data

Business Apps

Project management, HR, marketing, and CRM data

Machine Log Data

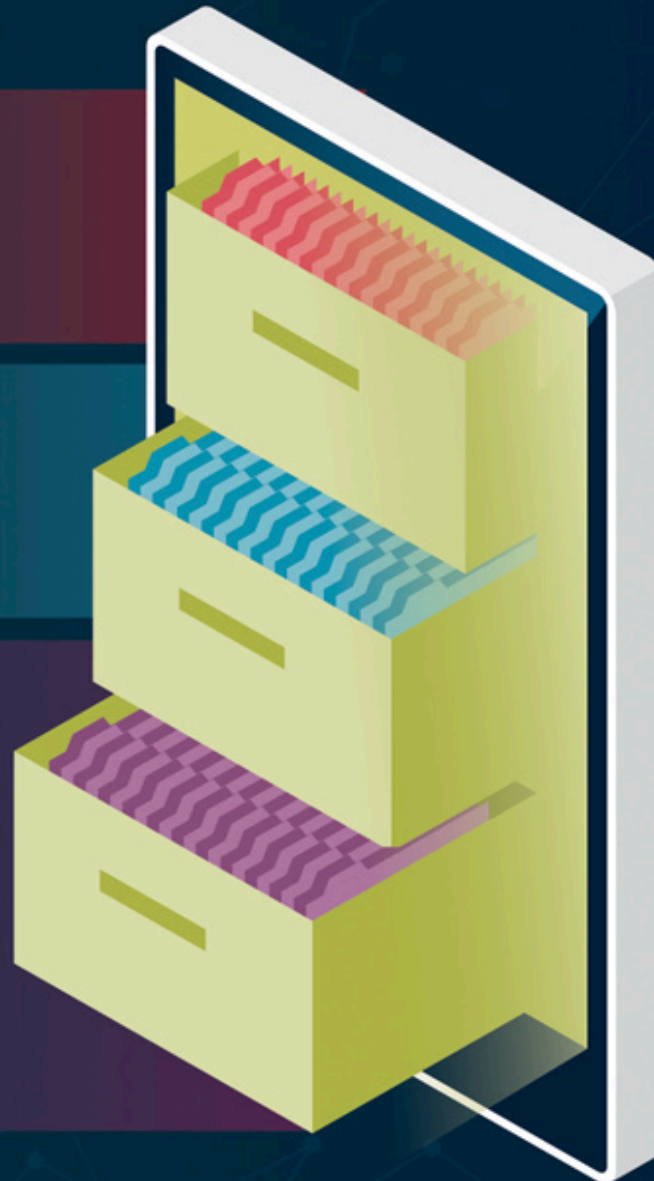
Event logs, server data, and application logs

Media

Images, video, infographics, podcasts, and live stream data

Sensor Data

Medical devices, geotracking, surveillance, smart electric motors, and industrial internet





To get a sense of the challenge of Big Data, here's what's created every minute:³

72 hours
*of Videos Uploaded on
YouTube*

2.5 Million
*Shares on
Facebook*

4 Million
*Search Inquiries on
Google*

300,000
Tweets

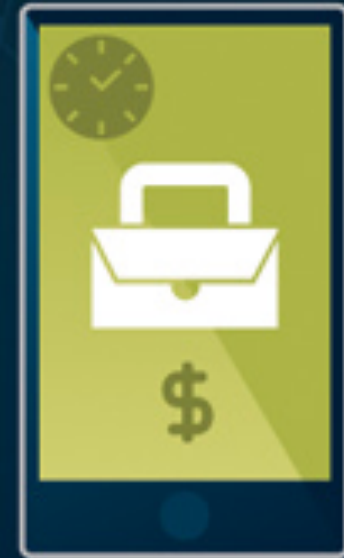
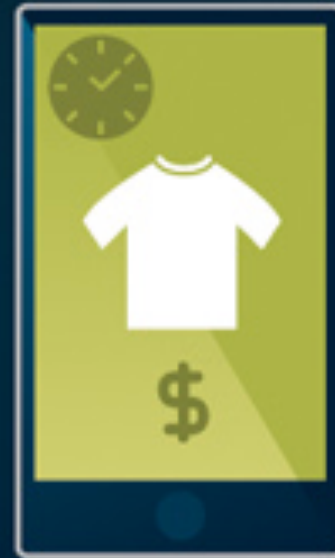
200 Million
Emails Sent



CASE STUDIES



Macy's adjusts pricing in near-real time for **73 million** items based on demand and inventory.⁴





Walmart

Walmart's latest search engine for Walmart.com includes semantic data relying on text analysis, machine learning and even synonym mining to produce relevant search results.

Walmart says it has improved online shoppers completing a purchase by 10% to 15%, increasing revenue by billions of dollars.

LOS ANGELES / SANTA CRUZ POLICE DEPARTMENTS

The Los Angeles and Santa Cruz police departments have taken an algorithm used to predict earthquakes, tweaked it and started feeding it crime data. The software can predict where crimes are likely to occur down to 500 square feet.

In LA, there's been a **33% reduction** in burglaries and **21% reduction** in violent crimes in areas where the software is being used.





American Express started looking for indicators that could really predict loyalty and developed sophisticated predictive models to analyze historical transactions and 115 variables to forecast potential churn.

The company launched a pilot program in Australia, and can now identify **24%** of accounts in the country that will close within the next four months.

MANUFACTURING



FINANCE



MEDICINE



MARKETING



GOVERNMENT



WHO IS USING BIG DATA?

There is no shortage of companies looking to make use of Big Data to make better decisions. These sectors include:

CRIME PREVENTION



AGRICULTURE



ENGINEERING





valencar
Campina Grande >



[Ver informações](#)

Promover novamente



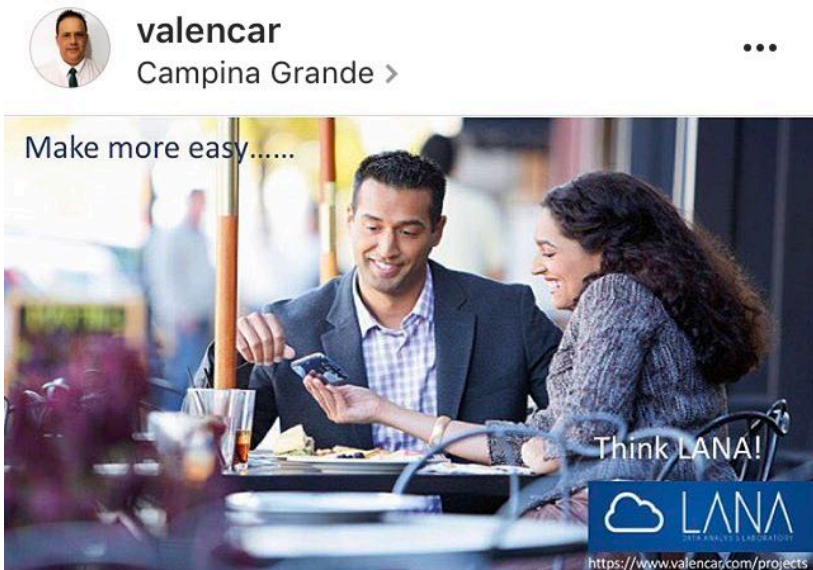
Curtido por **paularabello1**, **georgiakarenia** e outras **153 pessoas**

valencar LANA: <https://www.valencar.com/projects>

10 DE SETEMBRO



Big Data - Marketing Instagram



Ver informações

Promover novamente



Curtido por paularabello1, georgiakarenia e outras 153 pessoas

valencar LANA: <https://www.valencar.com/projects>

10 DE SETEMBRO



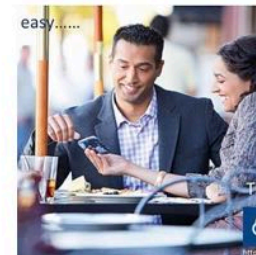
Publicado em 10/9/2017, 12:50

Promoção

Original

Esta publicação foi visualizada **3,135 vezes** e alcançou **3,120** contas.

145 Curtidas 0 Comentári... 0 Salvo



Publicado em 10/9/2017, 12:50

Promoção

Original

145 Curtidas 0 Comentári... 0 Salvo





Publicado em 10/9/2017, 12:50

Promoção

Original

♡ 145

Curtidas

💬 0

Comentári...

🔖 0

Salvo



3,135

Impressão



3,120

Alcance



161

Envolviment
o



Mar



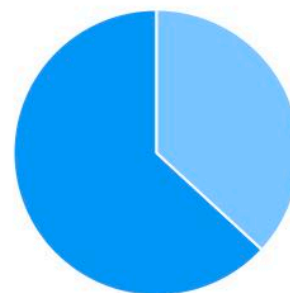
Publicado em 10/9/2017, 12:50

Promoção

Original

PÚBLICO

Gênero



63%

Mulheres

37%

Homens



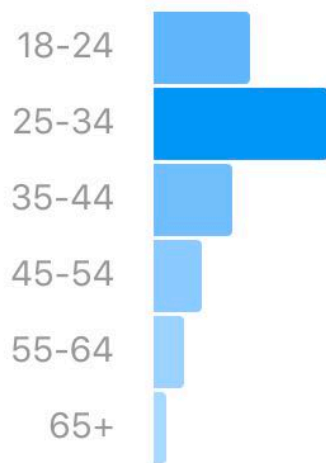


Publicado em 10/9/2017, 12:50

Promoção

Original

Faixa etária

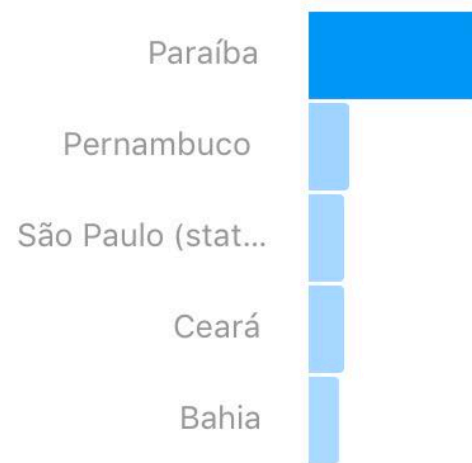


Publicado em 10/9/2017, 12:50

Promoção

Original

Principais localizações



gram



Publicado em 10/9/2017, 12:50

Promoção

Original

63%
Mulheres

37%
Homens



Esta promoção gerou **14 cliques**



Publicado em 10/9/2017, 12:50

Promoção

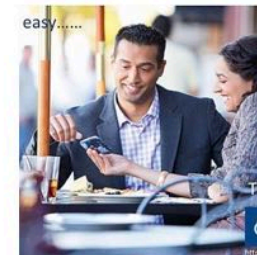
Original

Esta publicação foi visualizada
245 vezes e alcançou **172**
contas.

10
Curtidas

0
Comentári...

0
Salvo



Publicado em 10/9/2017, 12:50

Promoção

Original

Esta publicação foi visualizada
3,135 vezes e alcançou **3,120**
contas.

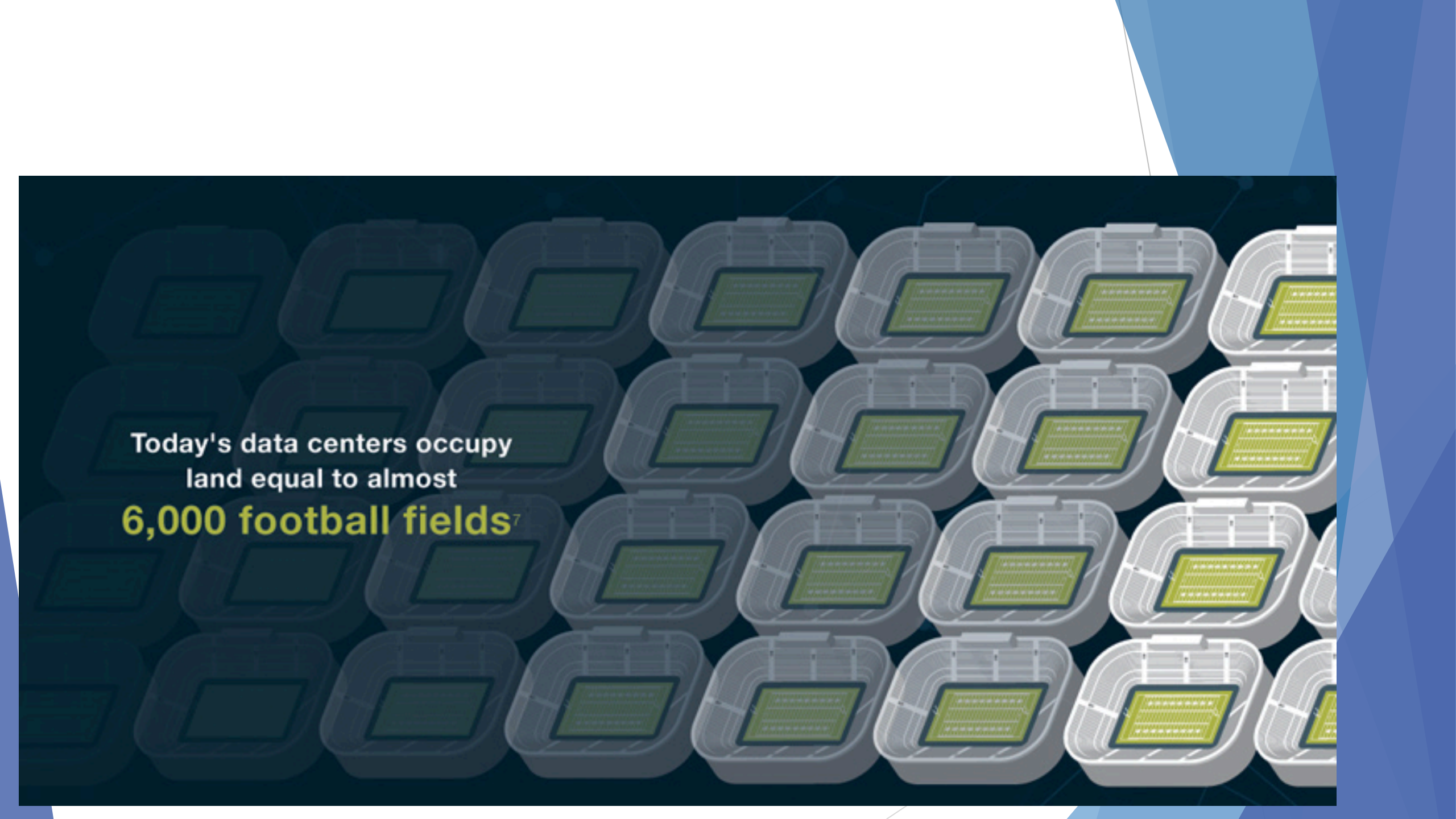
145
Curtidas

0
Comentári...

0
Salvo



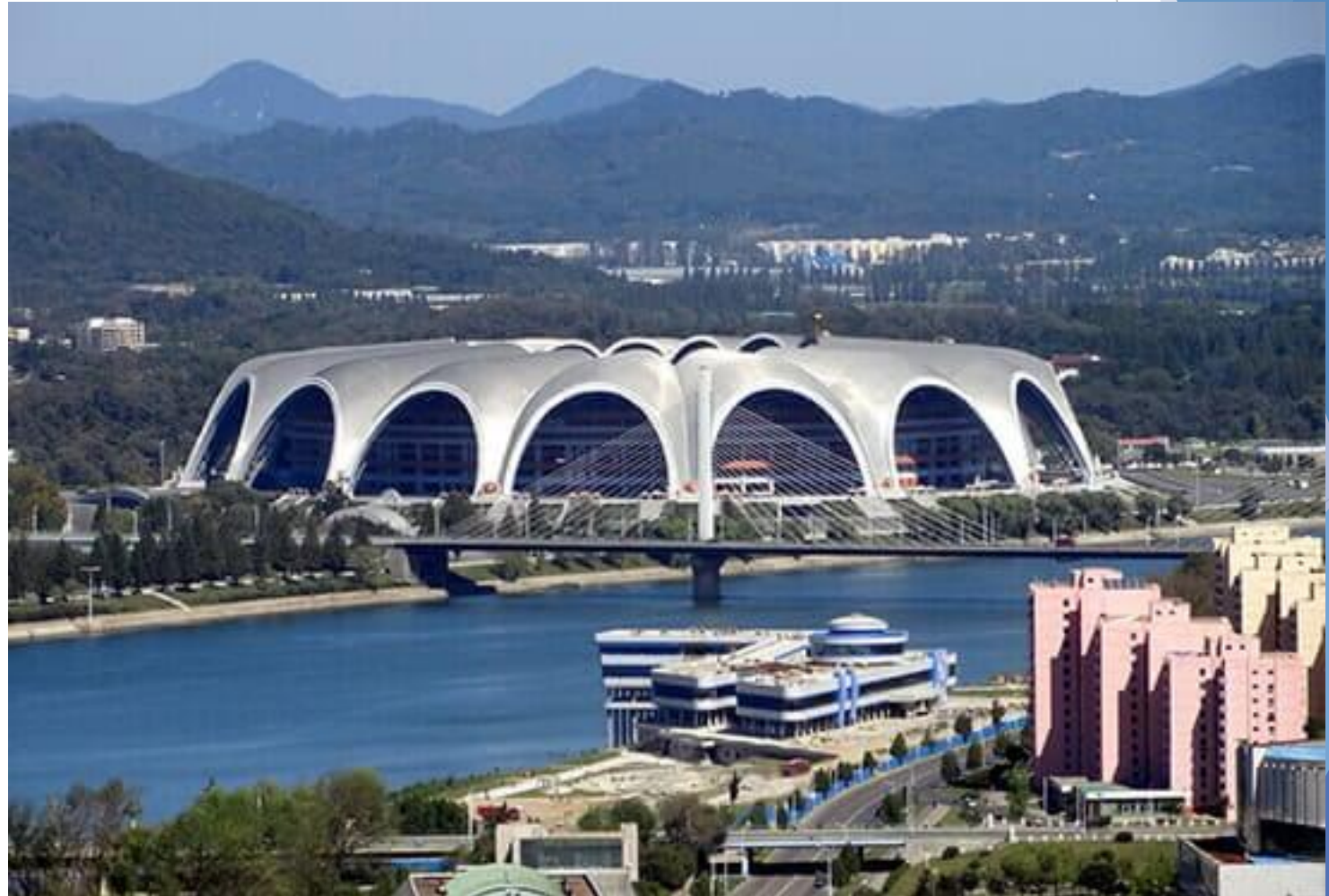
BY 2020,
the amount of digital information
will have grown from 3.2 zettabytes
today to **40 zettabytes⁶**



Today's data centers occupy
land equal to almost
6,000 football fields⁷

Comparção de Medidas

Maior estádio do mundo: May Day, em Pyongyang, capital da Coréia do Norte.
Ele tem impressionantes **207.000 metros quadrados**.
A arena tem capacidade para **150.000** pessoas



Switch SuperNAP (Las Vegas, Nevada) - Data Center

Occupying an area of 3.5 million square feet in Nevada is the Switch SuperNAP. The choice of its location in Las Vegas, Nevada was strategic to protect it from natural disasters.

The Citadel Campus sits on 2,000 acres in the Tahoe Reno Industrial Center, adjacent to the Tesla Gigafactory, and is totally powered by renewable energy. As such, Switch is the only company with a 100% Clean Energy Index.

The Tahoe Reno Citadel Campus is surrounded by a solid concrete wall 20 feet high, guaranteeing security and reliability.

At a cost of \$1 billion, 500 miles of fiber optic network cables will connect San Francisco, Reno, Los Angeles and Las Vegas. The 50 million people living in the four cities will be able to access data in just 14 milliseconds.



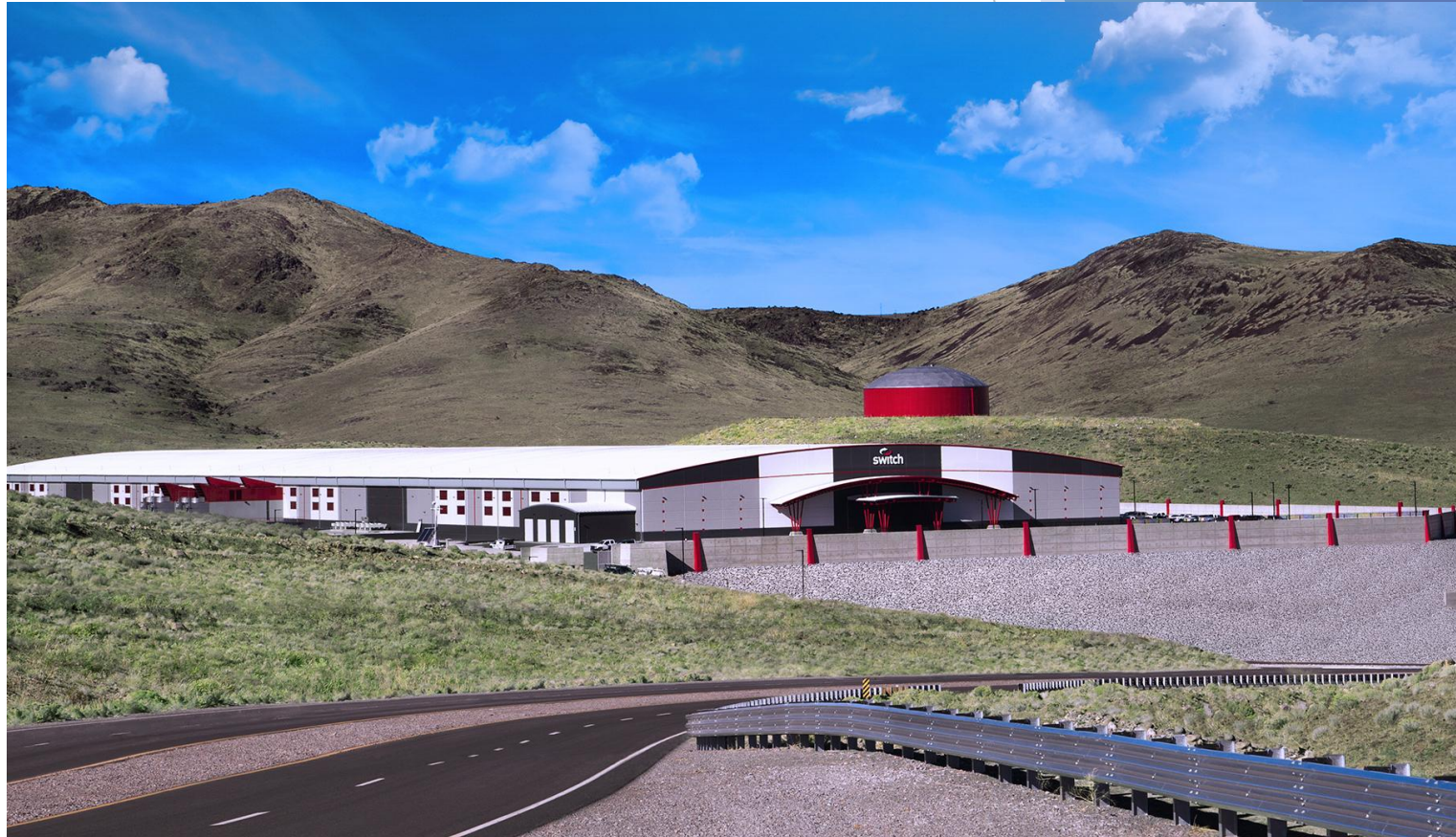
Data Center 2020

Switch TAHOE RENO - The Citadel Campus - Nevada, USA

It is designed for up to 7.2 million square feet of data center space and up to 650 megawatts (MW) of power (3 cities, 200k hab). 1 city = 260 MW

Switch, a globally recognized leader in future-proof data center design, superscale cloud, unparalleled telecom gateways and energy sustainability, today announced the opening of the largest, most advanced data center campus in the world - also known as The Citadel Campus.

The 2,000 acre campus, located in the Tahoe Reno Industrial Center next to the Tesla Gigafactory, is powered by 100-percent renewable energy, commitment to industry-leading green energy.

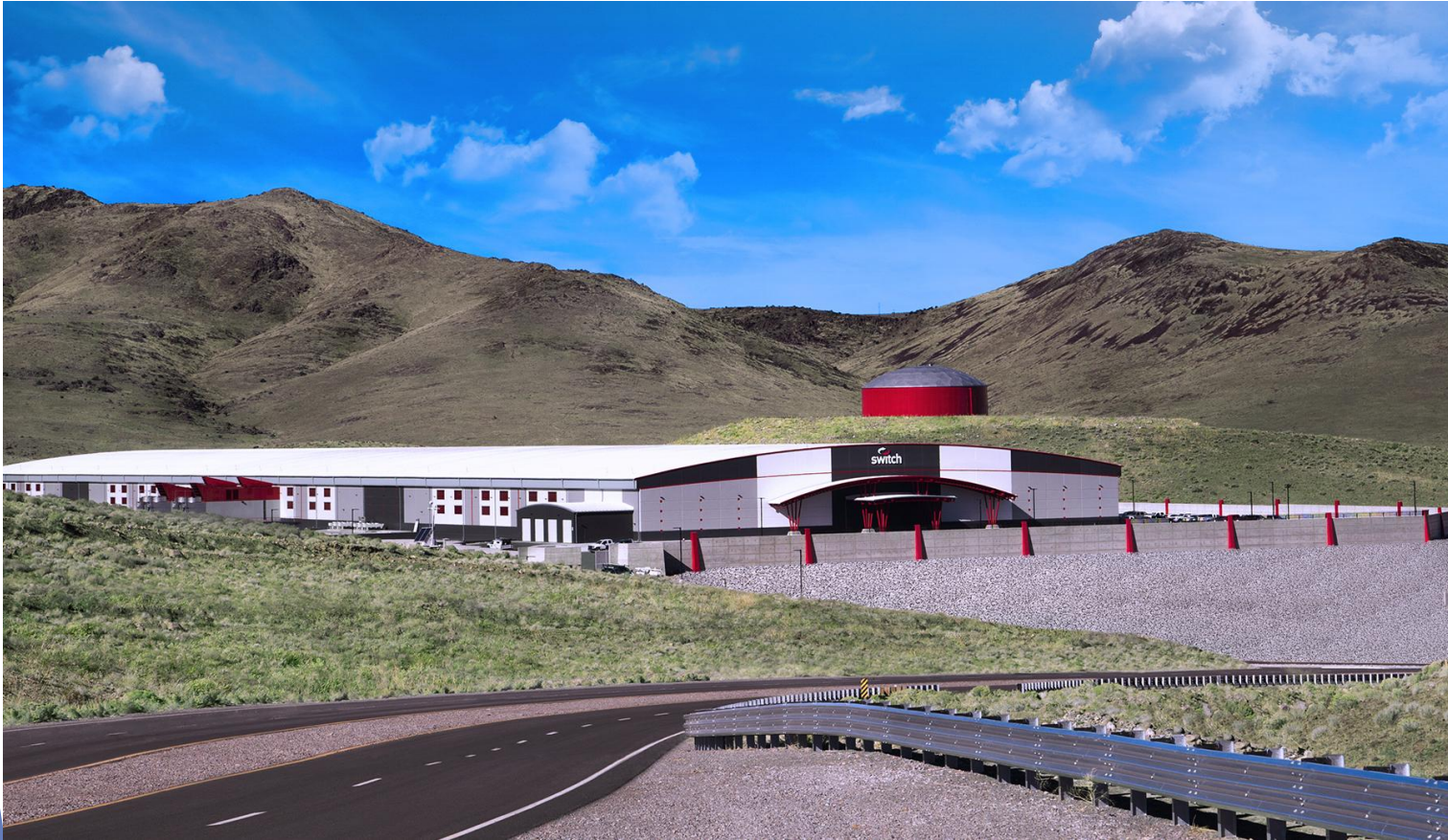


Data Center 2020

Switch TAHOE RENO - The Citadel Campus - Nevada, USA

<https://youtu.be/46vGxdbOIRk>

<https://www.switch.com/tahoe-reno/>



THESE INVESTMENTS COULD HAVE
GAME-CHANGING
IMPLICATIONS

The potential to save
\$300 billion
per year in healthcare
through analytics⁹

Retailers increasing
their profit margin by
up to 60%
through big data
analytics¹⁰

“Information is the oil of the
21st century, and analytics is
the combustion engine.”

- Peter Sondergaard, Gartner Research

Fonte:

Data Science e Big Data



Big Data é uma coleção de conjuntos de dados, grandes e complexos, que não podem ser processados por bancos de dados ou aplicações de processamento tradicionais

Data Science e Big Data



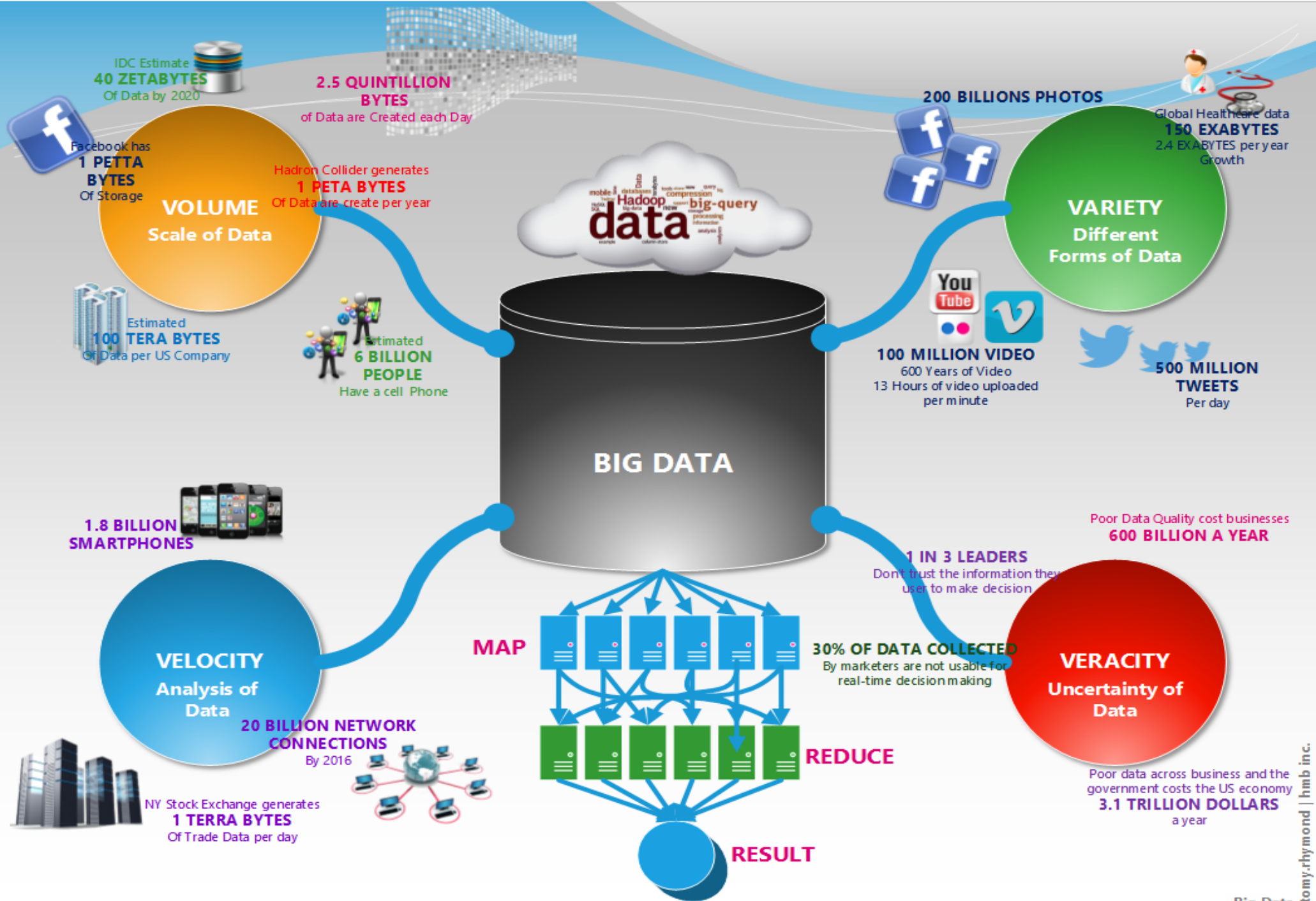
Big
Data



Data Science

Os 4 Vs do Big Data





Os 4 Vs do Big Data

Processar de forma eficiente e com baixo custo grandes volumes de dados

Transformar 12 TB de tweets gerados cada dia em produtos de análise de sentimento



Responder ao aumento da velocidade de geração dos dados

Investigar 5 milhões de eventos de trade nas bolsas de valores a fim de identificar fraudes



Coletar e analisar dados de diferentes formatos e fontes

Monitorar milhares de videos de segurança a fim de identificar pontos perigosos em uma cidade



Garantir que os dados sejam confiáveis

Os 4 Vs do Big Data - Velocidade

1 terabyte de informação é criado durante uma única sessão da bolsa de valores Americana, A New York Stock Exchange (NYSE)

Aproximadamente 100 sensors estão instalados nos carros modernos para monitorar: nível de combustível, pressão dos pneus, estados das peças, motor, etc;

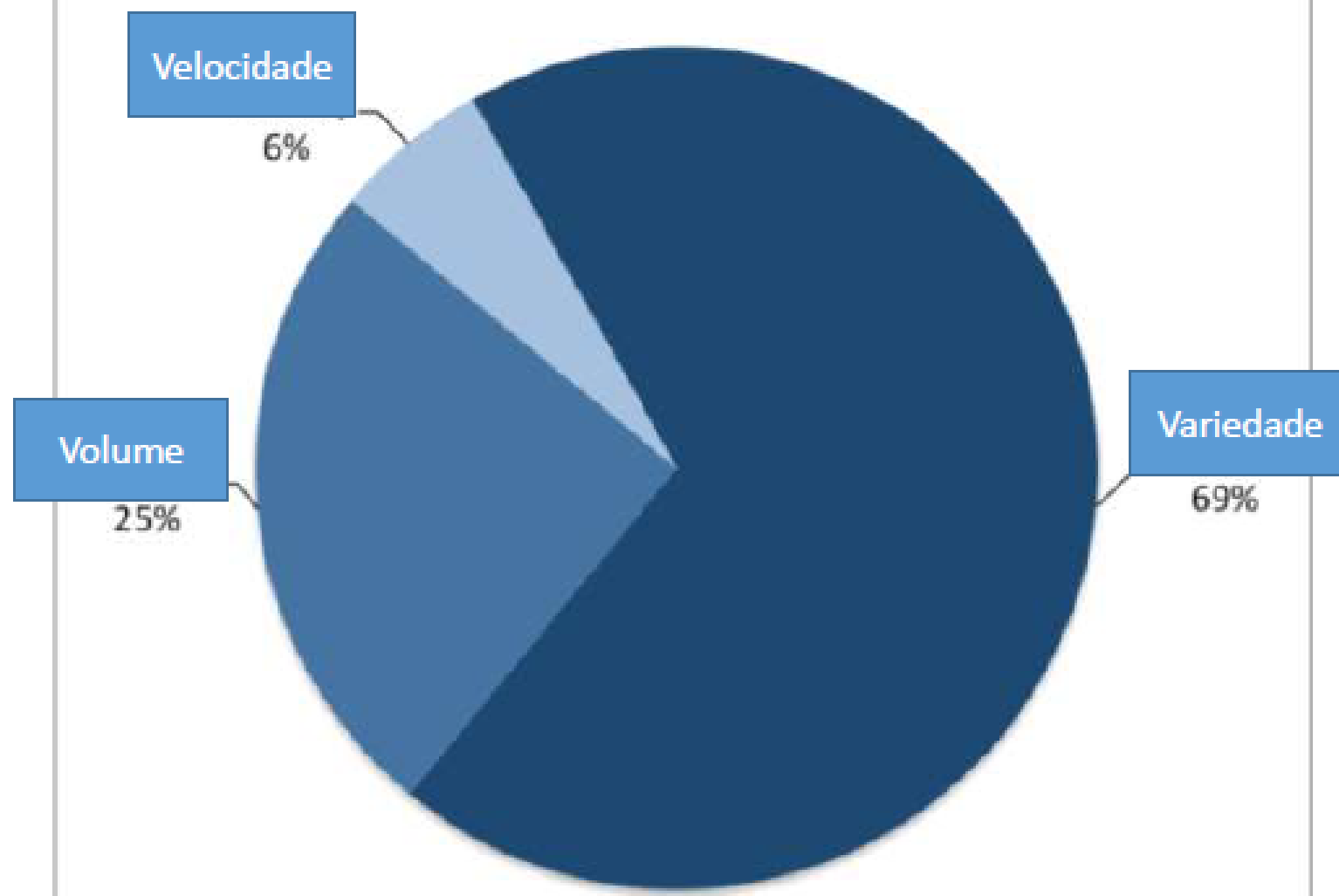
Mais de 20 Bilhões de Conexões de rede

Os 4 Vs do Big Data - Veracidade

Atualmente, 1 em cada 3 gestores tem experimentado problemas relacionados a veracidade dos dados para tomar decisões de negócios.

Além disso, estima-se que 3.1 trilhões de dólares por ano sejam desperdiçados devido a problemas de qualidade dos dados.

Importância: Volume, Velocidade, Variedade



Time spent in a life of a Data Scientist

@datavizzdom

Gulrez

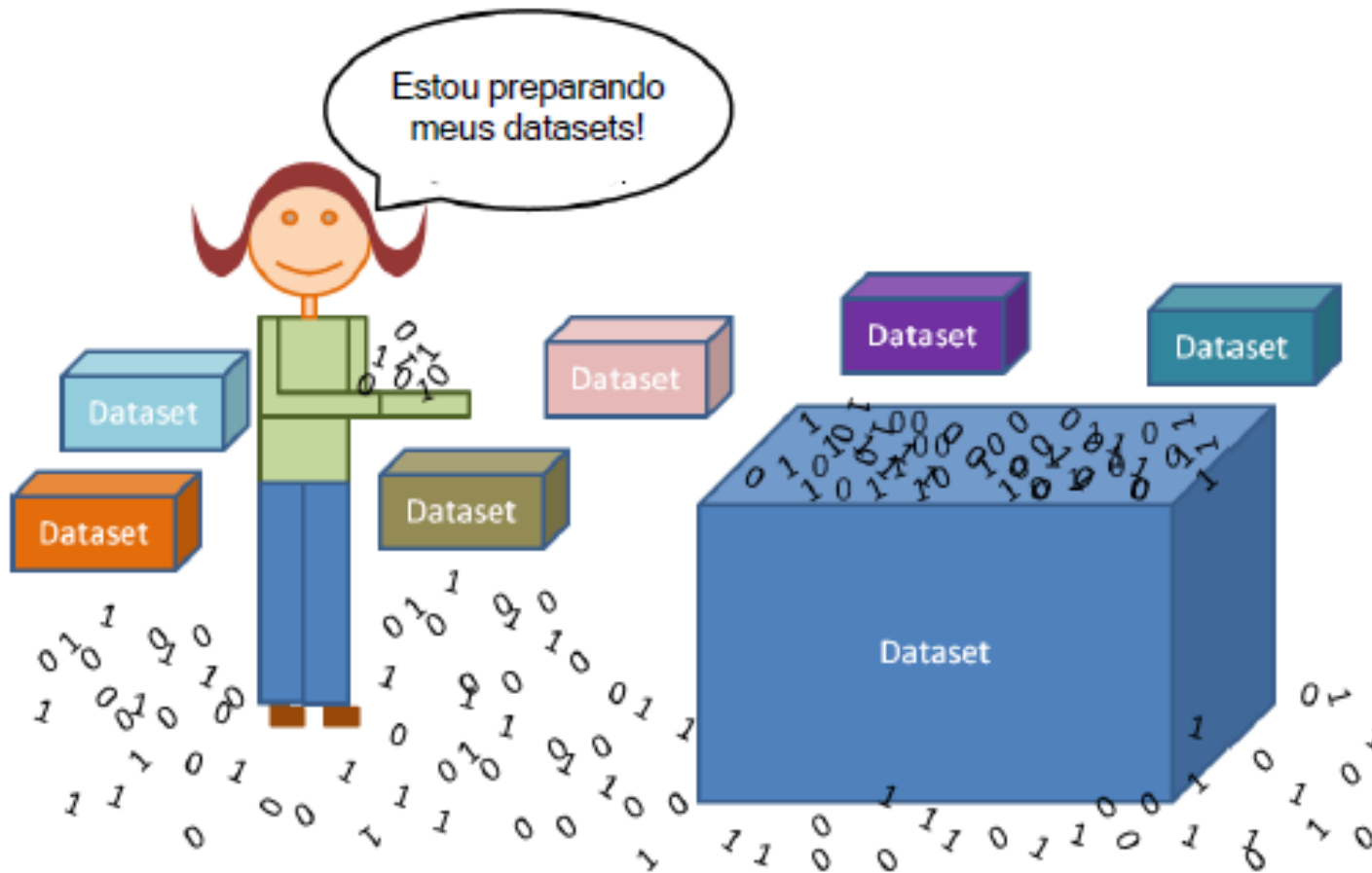
Perception



Reality



Data Science e Big Data



O que é um Dataset?
Conjunto de Dados

Data Science e Big Data



O que é um Dataset?

- Coleção de observações
- Cada observação é tipicamente chamada de registro
- Cada registro tem um conjunto de atributos que apontam características, comportamentos ou resultados

Data Science e Big Data

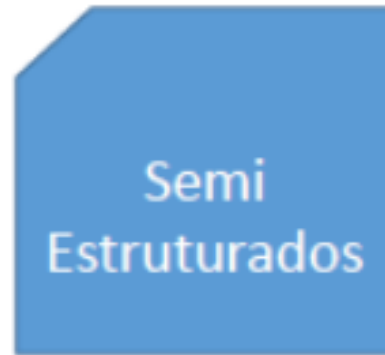


Tipos de Dados



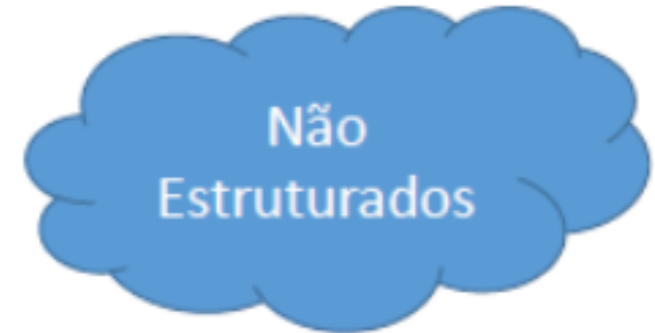
Estruturados

Banco de Dados



Semi
Estruturados

XML, JSON



Não
Estruturados

Twitter, Posts do
Facebook, Fotos, Vídeos,
Música.

Data Science e Big Data



Tipos de Dados: Estruturados

Tabela

Veículo

Atributos

Placa	Fabricante	Marca	Ano	Cor
IOS-0078	Renault	Sandero	2009	Vermelho
ITO-1314	Volkswagen	Fox	2010	Azul
IJM-1453	Hyundai	I30	2014	Pérola
IVA-2018	Chevrolet	Onix	2015	Branco
MAI-1852	Citroen	C3	2013	Preto

Tuplas

Domínio



Banco de Dados Relacionais

- ▶ Armazena dados em **tabelas**
- ▶ Possuem atributo tipo **chave**
- ▶ implementam funcionalidades simples do tipo **CRUD** (do inglês *Create, Read, Update e Delete*)
- ▶ controla o armazenamento, recuperação, exclusão, segurança e integridade dos dados

Tabela Empregado

cod_empr	nome	cod_depto
1	paula	1
2	jaque	1
3	andre	2
4	luiz	2

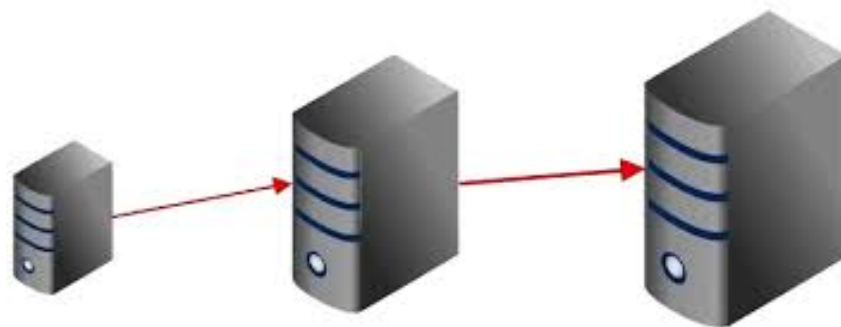


Banco de Dados Não-Relacionais (NoSQL)

Escalabilidade

Volume de dados crescente

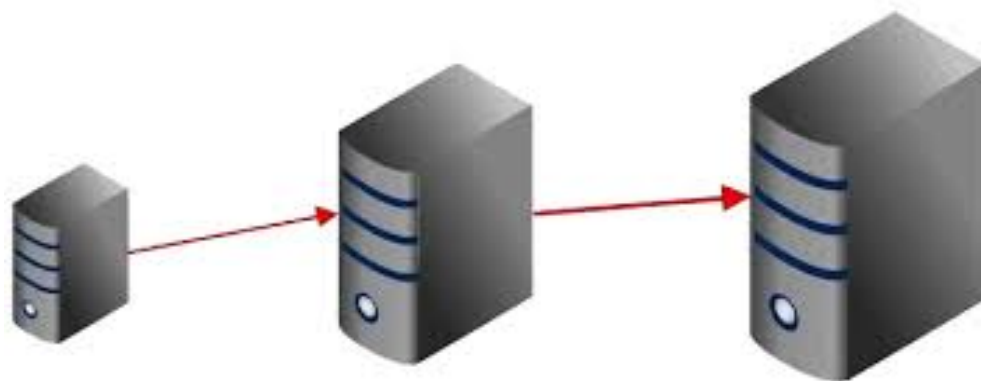
Disponibilidade dos dados (tempo real)



Escalabilidade

Característica de **Aumentar a capacidade** do sistema (seja processamento, armazenamento, E/S)

Há um Aumento da **Performance** depois da adição de **Hardware ou Serviço**.



Escalabilidade Vertical

Adicionar novos componentes na máquina.

É quando você coloca mais memória, mais disco, mais CPU no seu servidor.

Geralmente requer desligar a máquina, adicionar recursos e ligar novamente.



Escalabilidade Horizontal

Adicionar máquinas em Paralelo

Coloca mais servidores para atender a demanda

Carga é balanceada entre os servidores
(ex. Cluster)



NoSQL - Escalabilidade



**Escalabilidade Vertical
(Scale Up)**



**Escalabilidade Horizontal
(Scale Out)**

NoSQL - Não Relacionais

São Facilmente escaláveis Horizontalmente

Trabalham com quantidades maciças de dados (Big Data)

Não utilizam o modelo relacional para suas estruturas de dados

NoSql – Not Only Sql





Banco de Dados Não-Relacionais (NoSQL)

Gerenciar os grandes volumes de dados

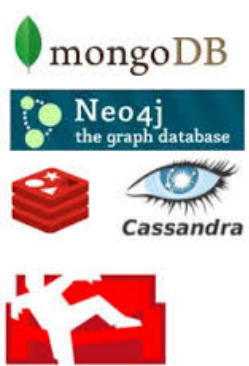
Buscar alto desempenho e disponibilidade

Permitem uma escalabilidade mais barata e menos trabalhosa

Características de poder trabalhar com dados semi-estruturados ou crus vindos de diversas origens

(arquivos de log, web-sites, arquivos multimídia, JSON, etc...)

Modelos Baseados em Documentos, Colunas, Chave-Valor e Grafos.



Tipos (NoSQL)

Chave/Valor (Key/Value)

Orientados a Documentos

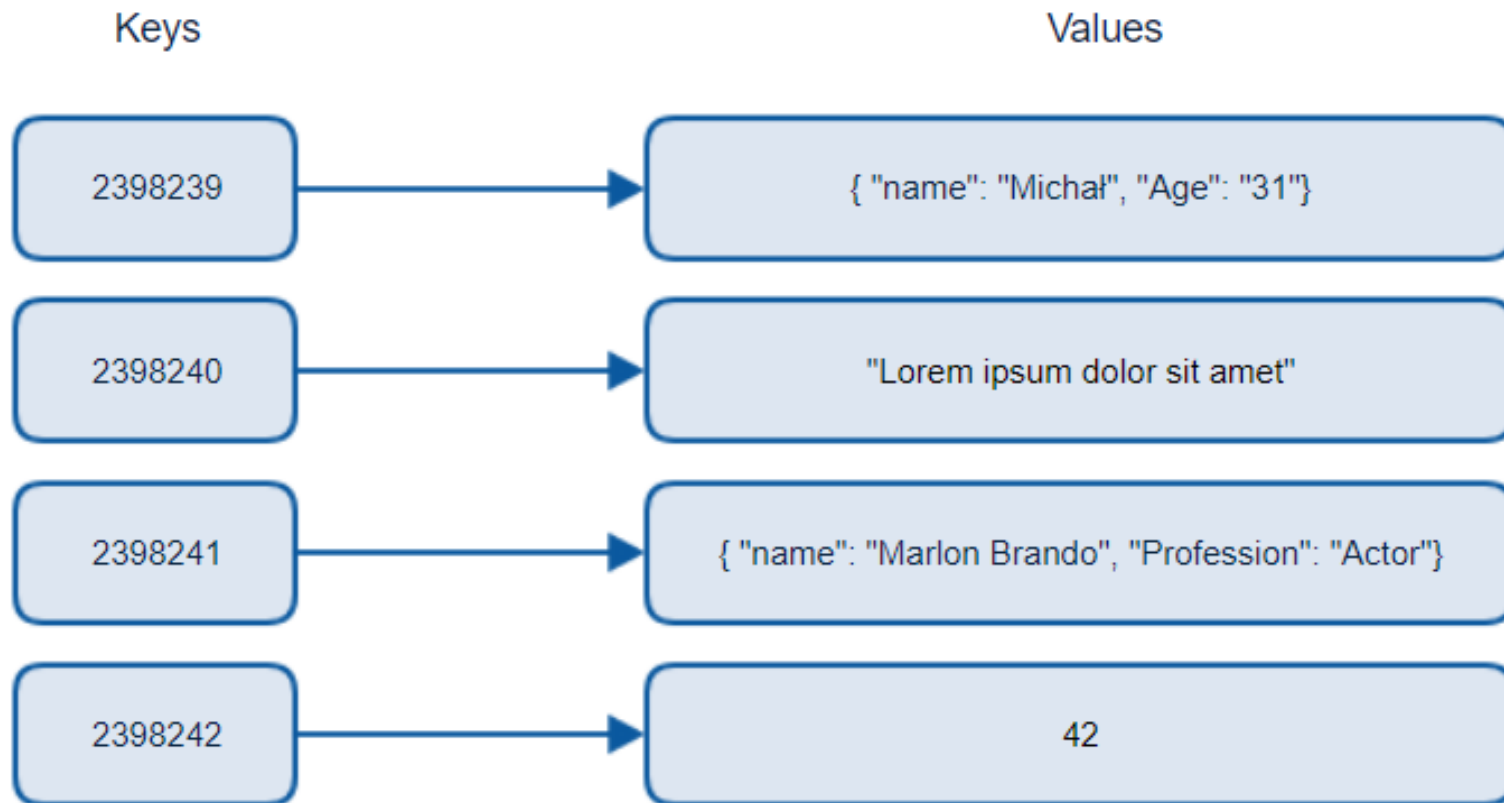
Orientados a Colunas

Grafos

NoSQL - Chave/Valor

O modelo chave-valor é o mais simples e fácil de implementar

Usa uma tabela hash na qual há uma chave única e um indicador de um dado ou de um item em particular



NoSQL - Chave/Valor - Aplicações:

Cache de conteúdo
(grande quantidade de dados, carregamento massivo)

Pesquisas rápidas

Logging
(registro de eventos relevantes)

NoSQL - Chave/Valor - uso:



NoSQL - Orientados a Colunas

Enquanto um banco de dados relacional é otimizado para armazenar **linhas** de dados, geralmente para aplicativos transacionais, um banco de dados colunar é otimizado para **recuperação rápida de colunas** de dados, normalmente em aplicativos analíticos

Armazena e processa grandes quantidades de dados distribuídos em muitas máquinas.

Reduz expressivamente os requisitos gerais de E/S de disco e diminui a quantidade de dados que você precisa carregar do disco.

NoSQL - Orientados a Columnas

Original data

oid	pid	cust	date	price
1	12	Sam	1/1/06	\$100
2	17	Mike	3/4/06	\$87
3	18	Joe	1/2/06	\$12
4	4	Andy	8/4/06	\$125

sales



Physically Stored as Columns

oid	pid	cust	date	price
1	12	Sam	1/1/06	\$100
2	17	Mike	3/4/06	\$87
3	18	Joe	1/2/06	\$12
4	4	Andy	8/4/06	\$125



Split into Several Projections



NoSQL - Orientados a Colunas - Aplicação

Aplicações distribuídas com uso intensivo de dados

Ex. Facebook



NoSQL - Orientados a Columnas - Uso



NoSQL - Orientados a Documentos

Um documento é:

- um objeto
- tem um identificador único
 - tem um conjunto de campos, que podem ser strings, listas ou documentos aninhados

Em cada documento temos um conjunto de campos (chaves) e o valor deste campo

NoSQL - Orientados a Documentos

Documento: archivo .JSON

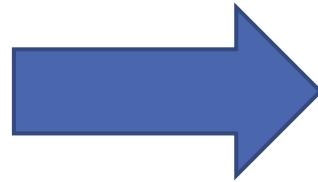
```
{ "screen name": "@infomagazine"  
  "text": "The 8 most stressful jobs in tech. Is yours one of them?"  
  "url": "https://t.co/qOfPJJrfcw",  
  "screen name": @infomagazine  
  "lang": "En",  
  "created_at": "Tue Feb 05 02:31:38 +0000 2019",  
  "place": "United States", ...  
}
```


NoSQL - Orientados a Documentos - Aplicação

Aplicações Web
(Similar ao armazenamento chave-valor)

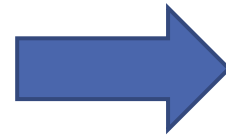
Tolerante a dados incompletos

NoSQL - Orientados a Documentos - Uso



Linked in

Mc
Graw
Hill
Education



Expedia

eBay

NOKIA
Connecting People

foursquare

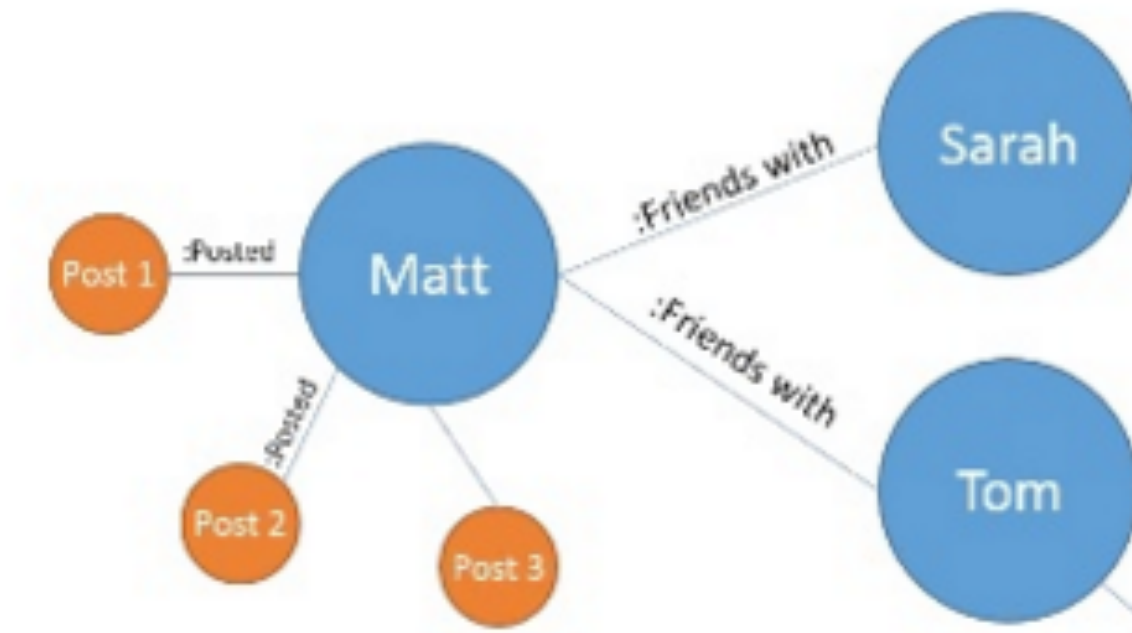
NoSQL - Grafos

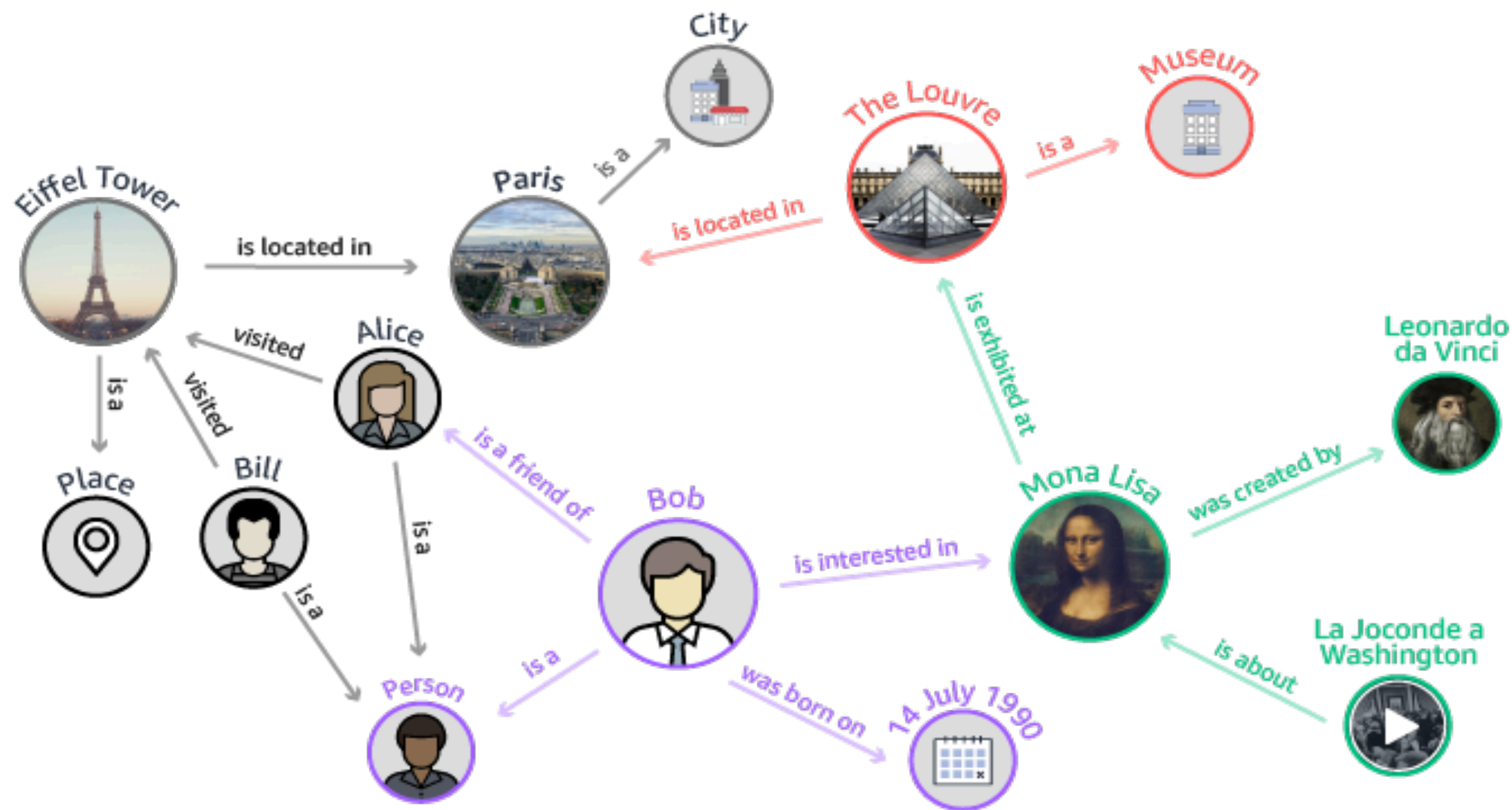
Possuem:

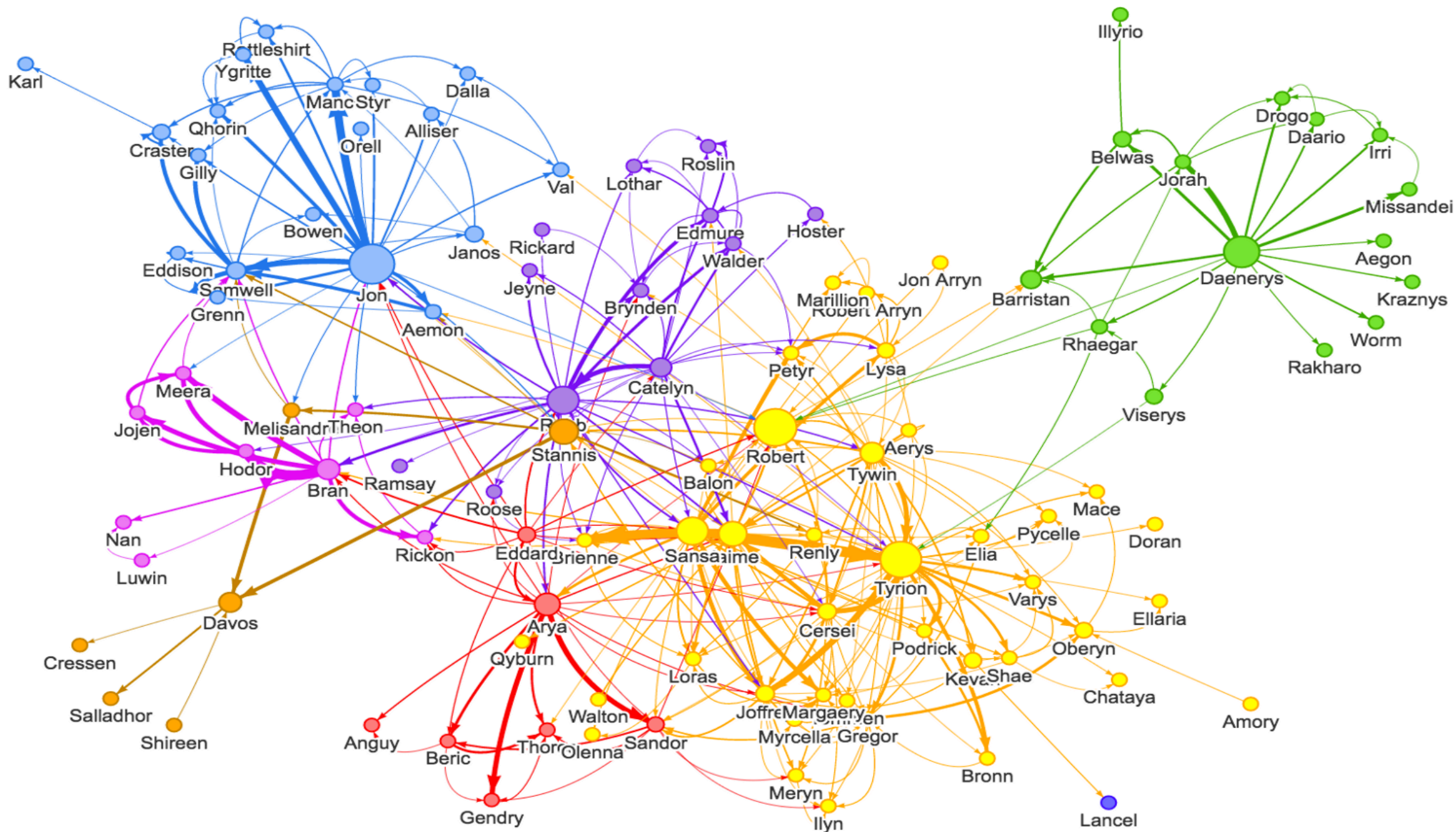
os nós (são os vértices do grafo)

os relacionamentos (são as arestas)

e as propriedades (ou atributos) dos nós e relacionamentos







NoSQL - Grafos



NoSQL - Grafos - Aplicação

Manipulação de dados estatísticos, frequentemente escritos mas raramente lidos

(como um contador de hits na Web)

Aplicações que exigem alto desempenho em consultas com muitas junções (Tabelas)

Redes Sociais, Recomendações (Foco em modelar a estrutura dos dados – interconectividade)

NoSQL - Grafos - Uso



SGBD's NoSQL



50 GB de dados para armazenar

Mysql (Relacional)

- ▶ 300 ms – Write
- ▶ 350 ms – Read

Cassandra (NoSQL)

- ▶ 0.12 ms – Write
- ▶ 15 ms – Read
- ▶ **Diferença:**
 - ▶ Write – **2.500** vezes mais rápido o NoSQL
 - ▶ Read – **23** vezes mais rápido o NoSQL

Quando Usar?

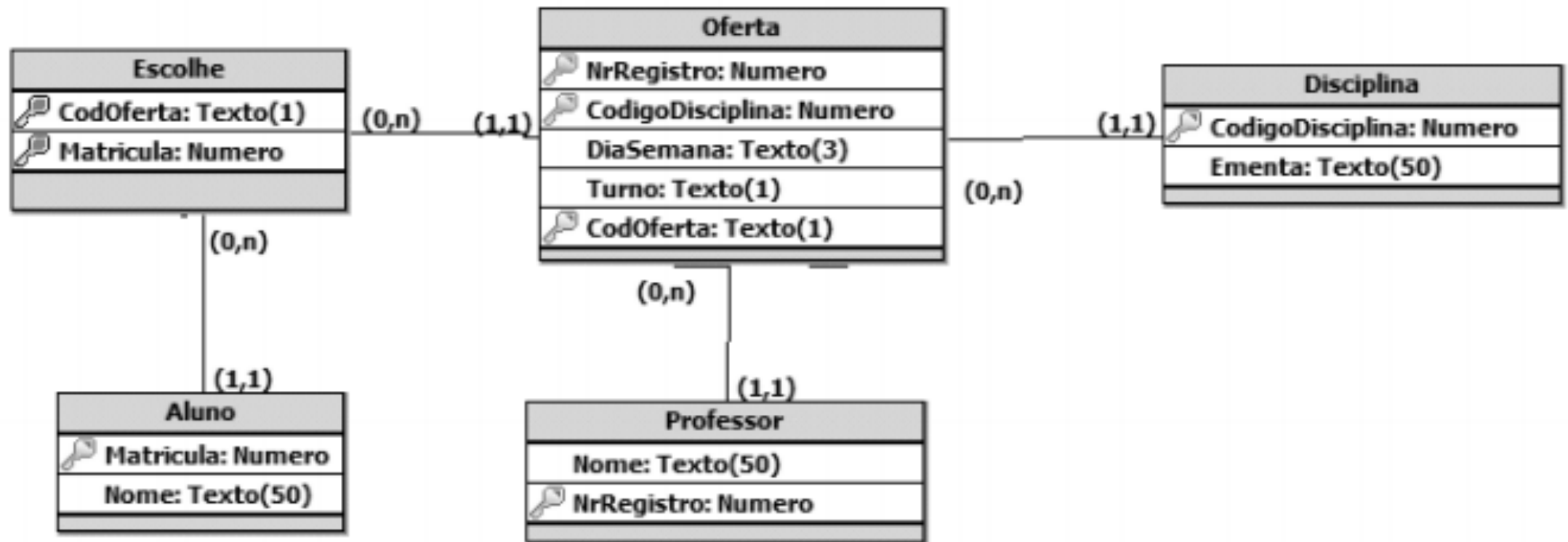
SGDB-Relacional	NoSQL
	O armazenamento deve ser capaz de lidar com carregamentos pesados
Armazenamento é esperado para apresentar carregamento pesado também, mas consiste principalmente na leitura de operações	Você executa muitas operações de escrita no armazenamento
Você prefere performance a uma estrutura de dados sofisticada	Você quer um armazenamento que seja escalável horizontalmente
Você precisa de uma linguagem SQL query poderosa	Simplicidade é bom, como em uma linguagem query bem simples (sem joins)

359 systems in ranking, August 2020

Rank			DBMS	Database Model	Score		
Aug 2020	Jul 2020	Aug 2019			Aug 2020	Jul 2020	Aug 2019
1.	1.	1.	Oracle 	Relational, Multi-model 	1355.16	+14.90	+15.68
2.	2.	2.	MySQL 	Relational, Multi-model 	1261.57	-6.93	+7.89
3.	3.	3.	Microsoft SQL Server 	Relational, Multi-model 	1075.87	+16.15	-17.30
4.	4.	4.	PostgreSQL 	Relational, Multi-model 	536.77	+9.76	+55.43
5.	5.	5.	MongoDB 	Document, Multi-model 	443.56	+0.08	+38.99
6.	6.	6.	IBM Db2 	Relational, Multi-model 	162.45	-0.72	-10.50
7.	 8.	 8.	Redis 	Key-value, Multi-model 	152.87	+2.83	+8.79
8.	 7.	 7.	Elasticsearch 	Search engine, Multi-model 	152.32	+0.73	+3.23
9.	9.	 11.	SQLite 	Relational	126.82	-0.64	+4.10
10.	 11.	 9.	Microsoft Access	Relational	119.86	+3.32	-15.47
11.	 10.	 10.	Cassandra 	Wide column	119.84	-1.25	-5.37
12.	12.	 13.	MariaDB 	Relational, Multi-model 	90.92	-0.21	+5.96
13.	13.	 12.	Splunk	Search engine	89.91	+1.64	+4.03
14.	 15.	 15.	Teradata 	Relational, Multi-model 	76.78	+0.81	+0.14
15.	 14.	 14.	Hive	Relational	75.29	-1.14	-6.51
16.	16.	 18.	Amazon DynamoDB 	Multi-model 	64.75	+0.17	+8.18

[illegible]

Relacionais



[illegible]

```
<recipe id="117" category="sobremesa">
<title> Bolo de banana </title>
<author>
<email> Miguel Furtado furtado@predialnet.com.br </email>
</author>
<date>Sunday,04 Jun 2000</date>
<description> O bolo de banana é feito com banana prata e
possui um sabor maravilhoso.
Pode ser servido quente ou frio. </description>
<ingredients> ... </ingredients>
<preparation>
Combine tudo no liquidificador e bata até misturar bem.
Leve ao forno por 20 minutos. Pronto. É só servir! </preparation>
<related url="#BoloChocolate">Bolo de Chocolate</related></recipe>
```

[illegible]

```
<?xml version="1.0"?>
- <funcionarios>
  - <funcionario ID="2107">
    <nome>Alessandra</nome>
    <salario>1000</salario>
  </funcionario>
  - <funcionario ID="0200">
    <nome>Cid Onir</nome>
    <salario>700</salario>
  </funcionario>
  - <funcionario ID="1020">
    <nome>Bianca</nome>
    <salario>980</salario>
  </funcionario>
</funcionarios>
```

Semi-Estruturado – Arquivo .JSON (Java Script Object Notation)

```
{
  "id_str": "694877215301873666",
  "text": "When Ebola performs duty in Sierra Leone,US UK \"Warships\" were sent 2 treat patients. Now Zica: sending 2 #Brazil???",
  "url": "https://t.co/wTW61kKsl0",
  "id": 694877215301873666,
  "source": "<a href='\"http://twitter.com\"' rel='\"nofollow\"'>Twitter Web Client</a>",
  "coordinates": 10000,20000,
  "retweet_count": 0,
  "followers_count": 722,
  "description": "Portfolio Manager BehindtheScenes Money Maker Management+",
  "friends_count": 165,
  "location": "ATHENS",
  "screen_name": "anthonysamaha",
  "lang": "en",
  "favourites_count": 4963,
  "name": "Anthony Samaha",
  "url": "http://t.co/o6uil766Ds",
  "time_zone": "Athens",
  "lang": "en",
  "created_at": "Wed Feb 03 13:36:52 +0000 2016",
  "place": Athens,
  "iso_language_code": "en" }
```



Data Science e Big Data



Tipos de Dados: Semi-Estruturados

TODOS OS DIAS:



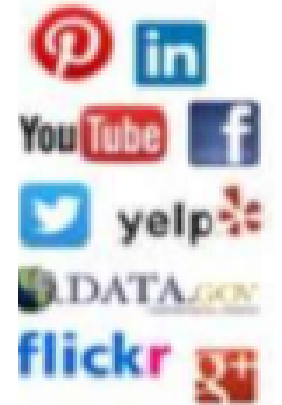
Twitter sozinho gera mais de 12 terabytes de dados



Facebook: 500TB

Fonte dos Dados

Formato dos Dados

		Estruturado	Não-estruturado
Fonte dos Dados	Interno	 • Resultados de pesquisas • Registros de vendas • Medidas de controle de processos • Bancos de dados de sistemas internos (ERP, CRM)	 • E-mails, cartas, mensagens de texto • Legendas de vídeos • Comentários de clientes • Mensagens de voz • Imagens / ilustrações • Avaliação de funcionários
	Externo	 • Likes do Facebook, retweets • Horário de publicação de posts, tweets, updates • Pontuação em sites de classificação	• Conteúdo publicado em redes sociais • Comentários em fóruns online • Imagens • Vídeos de câmeras de segurança

Data Science e Big Data



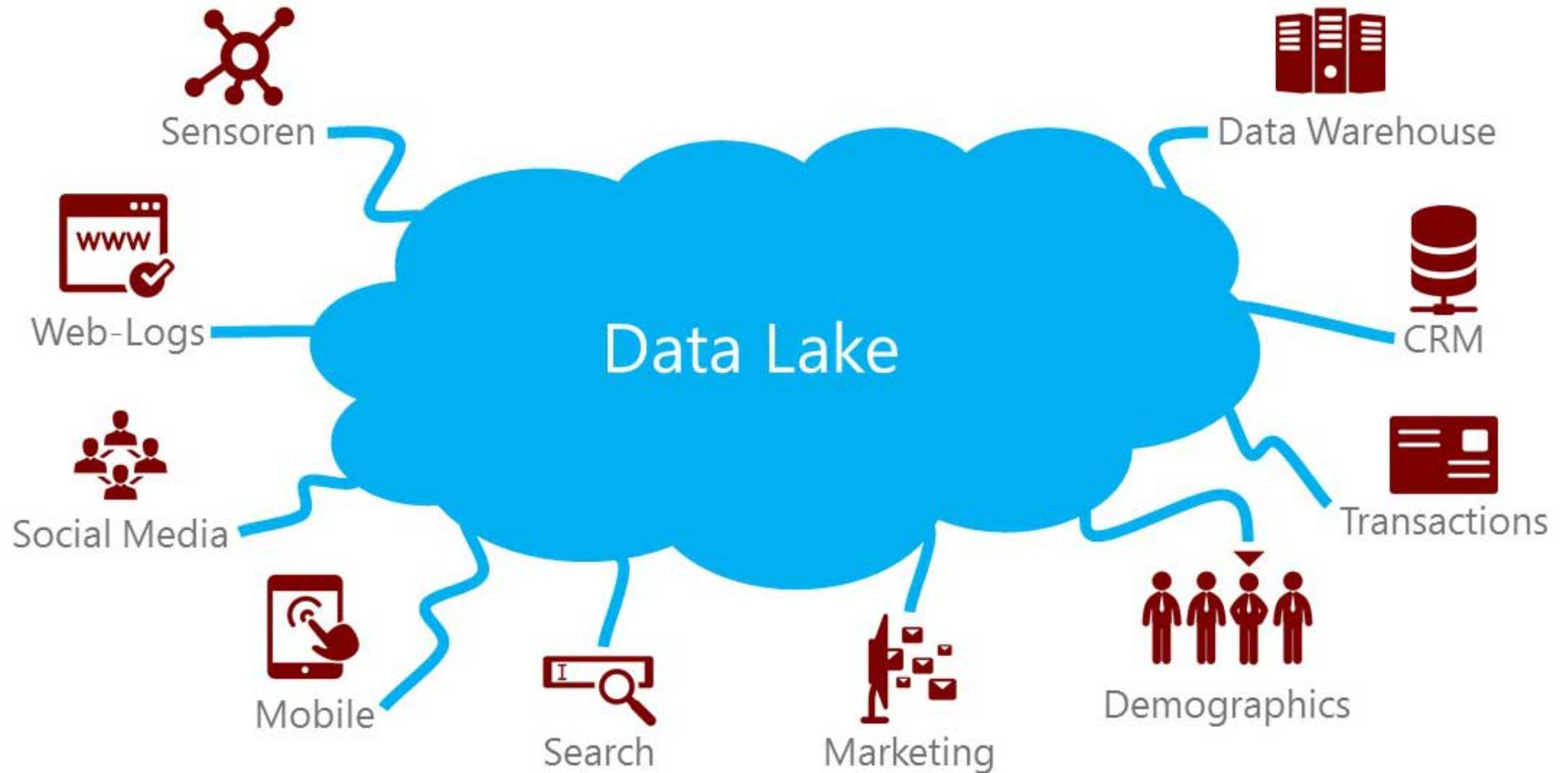
E o Big Data?



Big Data é uma coleção de conjuntos de dados, grandes e complexos, que não podem ser processados por bancos de dados ou aplicações de processamento tradicionais

Big Data

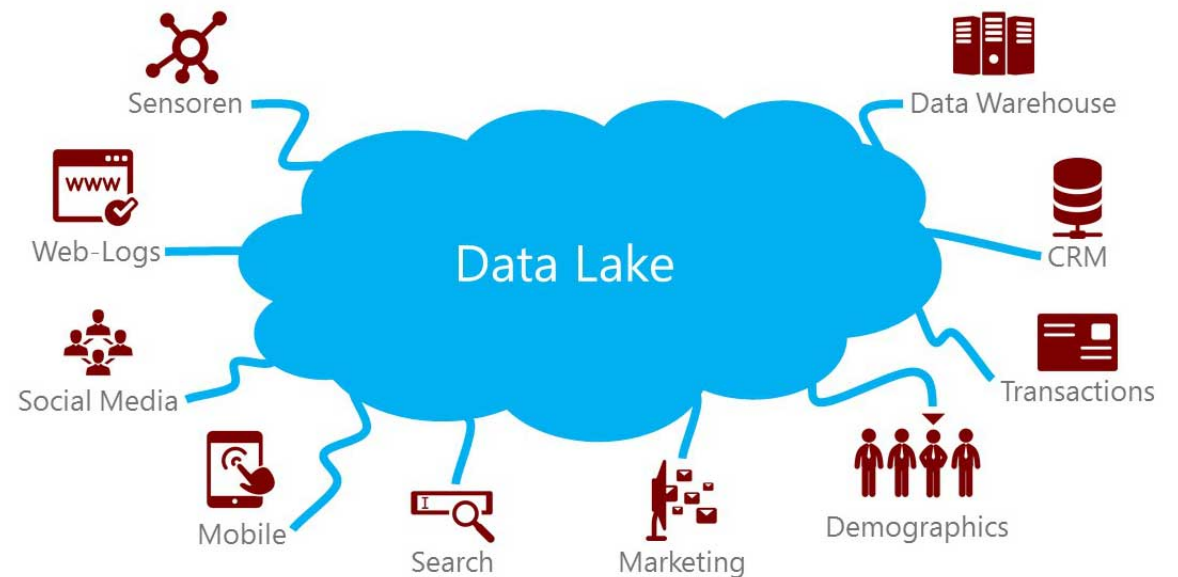
Big Data é mais do que uma palavra da moda no mundo dos negócios. Todo mundo deixa “rastros” quando cria uma conta no Facebook, abre conta em banco, faz compras em supermercados, etc. São dados como frequência de compra, consumo médio, número de curtidas em posts de determinado assunto, tweets. Multiplique isso bilhões de vezes e você começa a perceber o que é o Big Data.



Data Lake é um componente importante no universo da análise de dados e do Big Data.

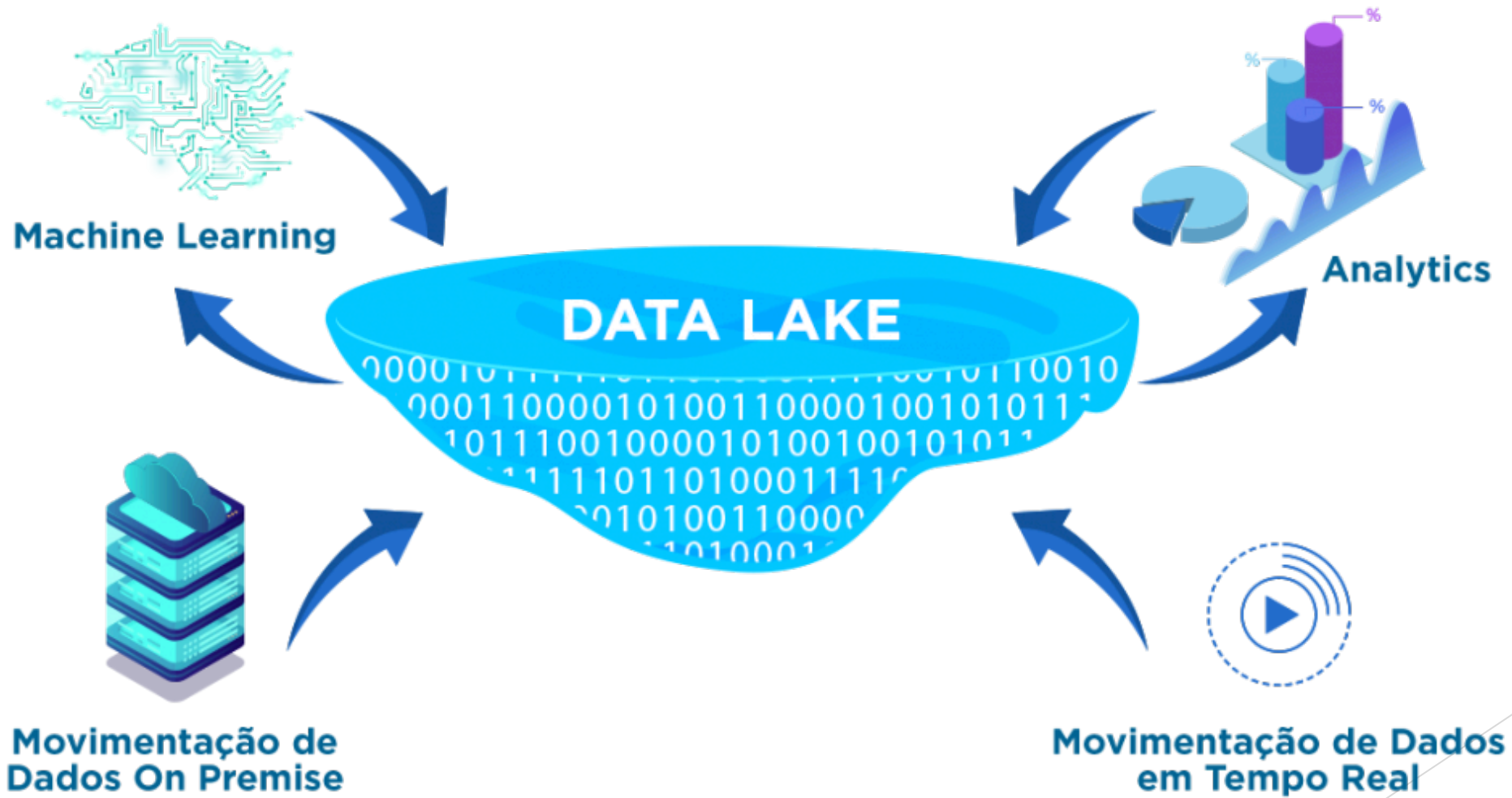
A ideia é ter um único repositório dentro da empresa, para que todos os dados brutos estejam disponíveis a qualquer pessoa que precise fazer análise sobre eles.

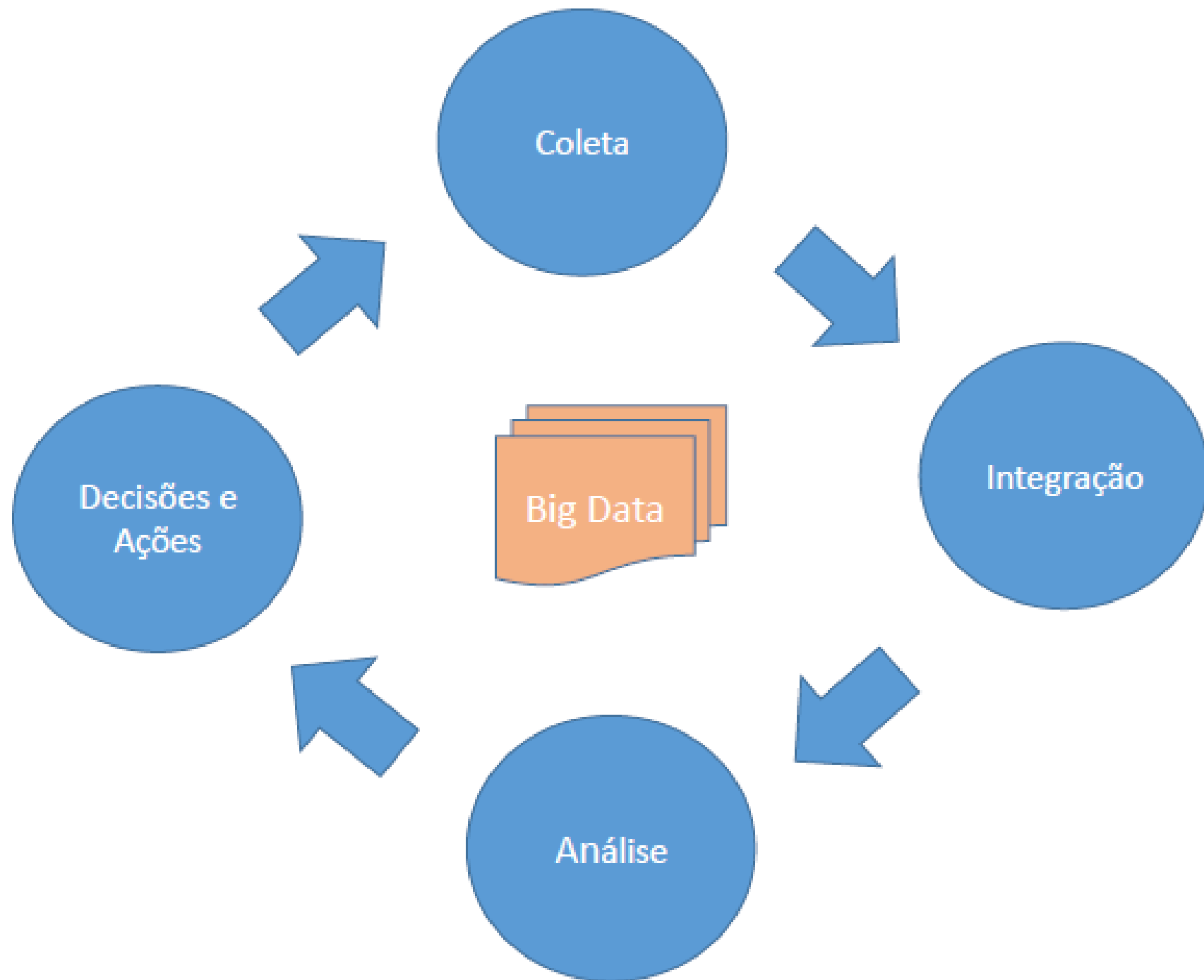
Comumente utiliza-se o Hadoop para trabalhar com os Data Lakes, mas o conceito é bem mais amplo do que apenas Hadoop.





A large, stylized brain shape filled with numerous small, colorful icons. The icons represent various aspects of modern life and technology, including social media (Facebook 'f' logo, Twitter bird), communication (speech bubbles, email envelope), music (musical notes, headphones), travel (globe, airplane), business (bar chart, pie chart), and general digital concepts (Wi-Fi symbol, smartphone, laptop). The overall theme is digital connectivity and information processing.





BI

SCM

RH

Mídia Social

Análise de Textos

API's

Reconhecimento de Voz

Processamento de Linguagem Natural

Internet das Coisas

Automação

SaaS

PaaS



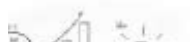
Dados Internos

Estruturados, previsíveis, permanentes,
fáceis de obter

Dados Externos

Não-estruturados, randômicos, voláteis,
difíceis de obter

A Oportunidade do Big Data



[illegible]

Analisar dados semi-estruturados e não-estruturados de uma variedade de fontes

Para análises interativas e exploratórias

Preserva a fidelidade dos dados

2020

4

BILLION

Connected People



\$4

TRILLION

Revenue Opportunity



25+

MILLION

Apps



25+

BILLION

Embedded and
Intelligent Systems



50

TRILLION

GBs of Data



Libelium Smart World

Air Pollution

Control of CO₂ emissions of factories, pollution emitted by cars and toxic gases generated in farms.

Forest Fire Detection

Monitoring of combustion gases and preemptive fire conditions to define alert zones.

Wine Quality Enhancing

Monitoring soil moisture and trunk diameter in vineyards to control the amount of sugar in grapes and grapevine health.

Offspring Care

Control of growing conditions of the offspring in animal farms to ensure its survival and health.

Sportsmen Care

Vital signs monitoring in high performance centers and fields.

Structural Health

Monitoring of vibrations and material conditions in buildings, bridges and historical monuments.

Quality of Shipment Conditions

Monitoring of vibrations, strokes, container openings or cold chain maintenance for insurance purposes.

Smartphones Detection

Detect iPhone and Android devices and in general any device which works with Wifi or Bluetooth interfaces.

Perimeter Access Control

Access control to restricted areas and detection of people in non-authorized areas.

Radiation Levels

Distributed measurement of radiation levels in nuclear power stations surroundings to generate leakage alerts.

Electromagnetic Levels

Measurement of the energy radiated by cell stations and WiFi routers.

Traffic Congestion

Monitoring of vehicles and pedestrian affluence to optimize driving and walking routes.

Smart Roads

Warning messages and diversions according to climate conditions and unexpected events like accidents or traffic jams.

Smart Lighting

Intelligent and weather adaptive lighting in street lights.

Intelligent Shopping

Getting advices in the point of sale according to customer habits, preferences, presence of allergic components for them or expiring dates.

Noise Urban Maps

Sound monitoring in bar areas and centric zones in real time.

Water Leakages

Detection of liquid presence outside tanks and pressure variations along pipes.

Vehicle Auto-diagnosis

Information collection from CanBus to send real time alarms to emergencies or provide advice to drivers.

Item Location

Search of individual items in big surfaces like warehouses or harbours.

Waste Management

Detection of rubbish levels in containers to optimize the trash collection routes.

Smart Parking

Monitoring of parking spaces availability in the city.

Golf Courses

Selective irrigation in dry zones to reduce the water resources required in the green.

Water Quality

Study of water suitability in rivers and the sea for fauna and eligibility for drinkable use.

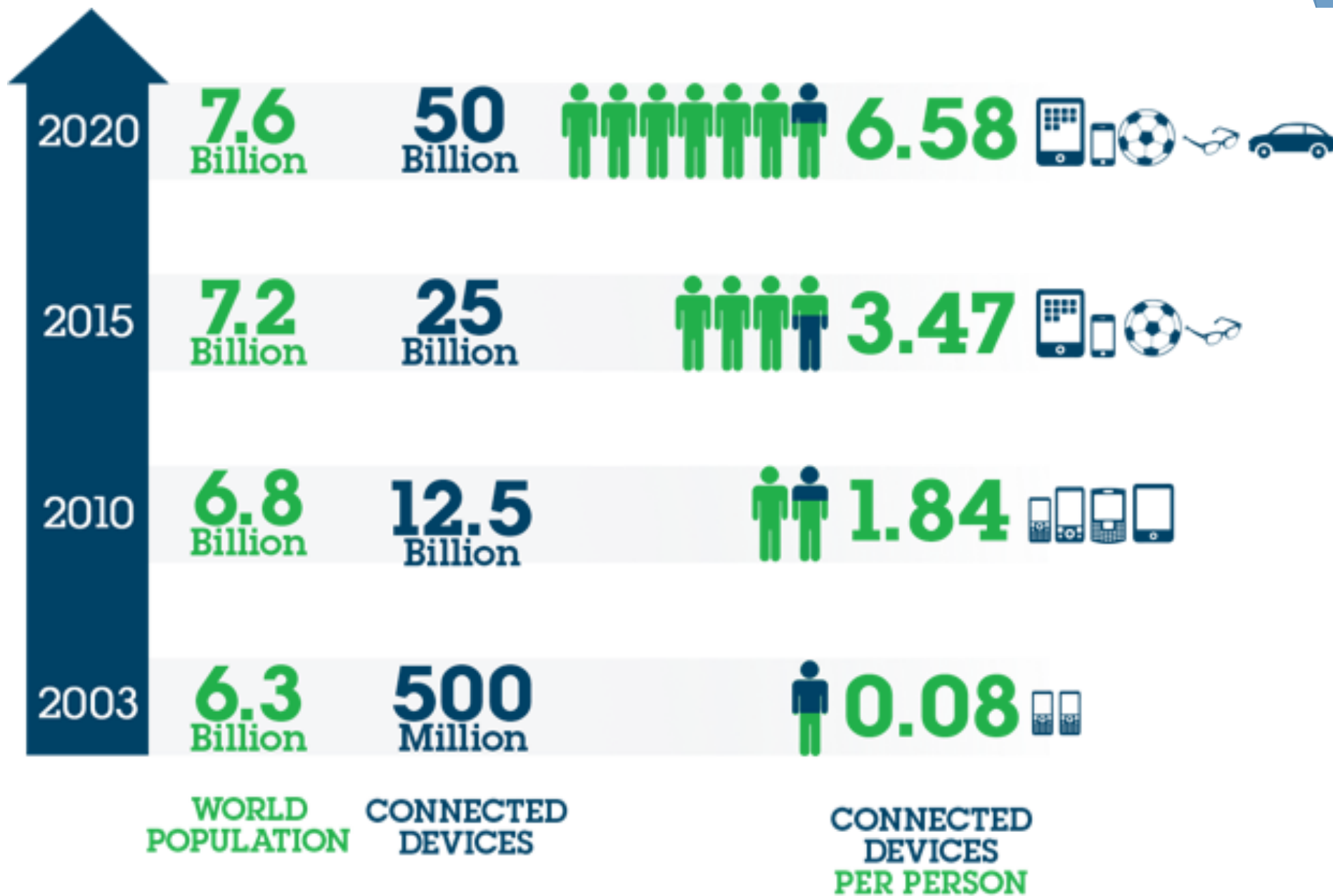


Image source: Cisco

DATA & AI LANDSCAPE 2019

INFRASTRUCTURE

The diagram illustrates the Hadoop ecosystem components, categorized into three main groups:

- HADOOP ON-PREMISE:** Includes Cloudera Hortonworks, MapR, Pivotal, IBM InfoSphere, and jethro.
- HADOOP IN THE CLOUD:** Includes AWS Microsoft Azure, Google Cloud, SAP Cloud Platform, IBM InfoSphere BigInsights, ARM (Hortonworks), Oracle, and CAZENA.
- STREAMING / IN-MEMORY:** Includes Amazon Kinesis, Databricks, SAP Cloud Platform, Oracle, Confluent, Striim, Hazelcast, GridGain, Gigaspaces, Watlaroo, FastData, and Kox.

The collage displays logos for various database and cloud technologies, organized into six categories:

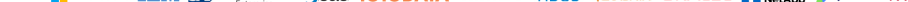
- NoSQL DATABASES:** Includes Google Cloud, AWS, Oracle, Microsoft Azure, MongoDB, MarkLogic, Couchbase, DataStax, Redis Labs, and Aerospike.
- NewSQL DATABASES:** Includes SAP, Clustrix, Fliqlo, MongoDB, Data Sphere, MEMSQL, InfluxData, Cockroach Labs, VoltDB, Splice, and IBM.
- GRAPH DBs:** Includes Neo4j, Amazon Neptune, IBM, Oracle, Orient DB, and InfluxGraph.
- MPP DBs:** Includes Teradata, AWS, Vertica, IBM Data Warehouse Systems, Cxton, Kognito, Exasol, dremio, and Inforworks.
- CLOUD EDW:** Includes Google Cloud, Microsoft Azure, Pivotal, Snowflake, Pulsar, and Nucleo.
- SERVERLESS:** Includes Amazon, Google Cloud, and Microsoft Azure.

The diagram is organized into four columns, each representing a different category of data management tools. Each column has a header and a list of companies with their logos.

- DATA TRANSFORMATION**
 - talend
 - pentaho
 - alteryx
 - TRIFORCE
 - tmr
 - Paxata
 - StreamSets
 - UNIFI
- DATA INTEGRATION**
 - SAP Data Services
 - Informatica
 - MuleSoft
 - TEALUM
 - inlogistics
 - Qoos Data Center
 - Segment
 - ATTUNITY
 - spinity
 - ZALONI
 - import.io
 - InfoWorks
 - FiveTears
 - SNOWFLOW
 - MATILLION
- DATA GOVERNANCE**
 - Informatica
 - SailPoint
 - IBM
 - McAfee
 - Stylight Security Center
 - collibra
 - Alation
 - Watermark
 - dremio
 - WINTERMUTE
 - OKERA
 - data.world
- MGMT / MONITORING**
 - AWS
 - New Relic
 - octrino
 - Rubrik
 - APPLIEDYNAMICS
 - dynatrace
 - WAVEFORM by VERVO
 - outforce
 - Signifix
 - droru
 - splunk
 - Microsoft
 - pagerduty
 - Asana
 - univerval
 - Numerify
 - datacatalyst
 - VERIFAN
 - zenoss
 - OpsRamp
 - MAXIMUS

[illegible]

CROSS-INFRASTRUCTURE/ANALYTICS



ANALYTICS & MACHINE INTELLIGENCE

The diagram is divided into two main sections by a horizontal line. The top section, titled 'DATA ANALYST PLATFORMS', lists several logos: Microsoft, pentaho, alteryx, Digital Reasoning, QVADAT, and AYASDI. The bottom section, titled 'DATA SCIENCE PLATFORMS', lists logos for IBM, databricks, dataiku, DOMINO, rapidminer, TIBCO, ATTIVO, Datameer, incorta, interana, MODE, ENDOR, ANAconda, SSAS, Altair, KNIME, and MathWorks.

BI PLATFORMS

- looker
- amazon analytics
- aws
- domo
- ARCADA DATA
- Thoughtspot
- ATSCALE
- qlik
- business
- Google Data
- Information Builders
- birst
- MicroStrategy
- Keen IO

VISUALIZATION

- tableau
- Power BI
- SAP LUMINA
- Google Cloud
- Periscope Data
- celonis
- zepl
- YODATA
- plioty
- CHARTIO
- FLUXUS TCO

MACHINE LEARNING

- Azure Machine Learning
- Amazon SageMaker
- Google Cloud AI Platform
- H2O
- DatRobot
- gamalon
- visenze
- deeppearl
- ELEMENT

The diagram illustrates the relationship between three main AI domains: Computer Vision, Horizontal AI, and Speech & NLP. Each domain is represented by a large colored box with a list of associated companies and services.

- COMPUTER VISION** (Blue box):
 - Microsoft Azure
 - Amazon Rekognition
 - Cloud Vision
 - darfai
 - 商汤
 - EVER AI
 - deeptomatic
 - neura
 - twentybin
 - UBIQUITY
 - ADCE
 - vision
 - YITU
 - trax
 - 商汤
 - synthesis
 - ser
- HORIZONTAL AI** (Green box):
 - IBM Watson
 - Cortana
 - Face++
 - 旷视
 - sentient
 - VoyagerLife
 - vicarious
 - PROPHESIES
 - CognitiveScale
 - Numenta
 - PETUUM
 - SI
 - SEAN
 - PARALOGIX
 - CIRCUITS AI
 - OSARO
 - blue vision
 - Fortinet
- SPEECH & NLP** (Orange box):
 - Google Cloud
 - amazon alexa
 - Amazon Translate
 - twilio
 - narative
 - science
 - openlemmas
 - MBovel
 - Eigen Technologies
 - SoundHound Inc.
 - PRIMER
 - Mindfield
 - centricall
 - Neural
 - coqito
 - snips
 - SHARPLING
 - AI
 - Polysyl

SEARCH	LOG ANALYTICS	SOCIAL ANALYTICS	WEB / MOBILE / COMMERCE ANALYTICS
 elasticsearch  algolia  ACTIVIO  swiftype  alphanumeric  omni:us	 ORACLE  splunk  sumologic  solarwinds  timber  hibana  logz.io	 Hootsuite  NETBASE  synthesio  simple reach  bitly  Sprinklr  truck	 Google Analytics  mixpanel  AMPITUDE  Airtable  SIGOPT  granity  custora

OPEN SOURCE

The diagram illustrates the Big Data ecosystem landscape, categorized into 12 main areas:

- FRAMEWORKS:** Spark, Flink, YARN, TEZ, MESOS, Kubernetes, CDAP, Hadoop, RedHut, Hightouch, HELIX.
- QUERY / DATA FLOW:** Spark SQL, presto, DORIS, SLAMDATA, GraphX, GraphQL, Flink.
- DATA ACCESS & DATABASES:** cassandra, mongoDB, redis, COUCHDB, Cockroach LABS, druid, ScyllaDB, Ceph, Ceph, Ceph, ScioDB, riak, Faisla, Cloud Spanner, SCOUTLID.
- ORCHESTRATION & MGMT:** talend, Apache Zookeeper, Apache Ambari, Apache Airflow, MESOS, etcd, Kong.
- STREAMING & MESSAGING:** Spark Streaming, nifi, Flink, beam, kafka, STORM, Apache RocketMQ.
- STAT TOOLS & LANGUAGES:** Python, R, Scala, Tableau, Studio, SciPy, julia.
- AI OPS & INFRA:** mflow, KubeFlow, mlearn, DVC, SELDON, Polyaxon.
- AI / MACHINE LEARNING / DEEP LEARNING:** TensorFlow, Keras, PyTorch, MxNet, Caffe, Microsoft Cognitive Toolkit, OpenAI, DM, theano, H2O, Apache SENGA, DIMSUM, FeatureFu, VELES, Chainer, Microsoft, ONNX, WESA, LUDWIG, PyTorch.
- SEARCH:** elasticsearch, Solr.
- LOGGING & MONITORING:** elasticsearch, kibana, SENTRY, logstash, Prometheus, seaborn, fluentbit, fluentd, Grafana, ELK, VICTOR.
- VISUALIZATION:** matplotlib, TensorBoard, seaborn, Bokeh, Jupyter, D3.js, Anaconda.
- COLLABORATION:** BeakerX, Jupyter, D3.js, Anaconda.
- SECURITY:** Apache Ranger, KNOX, Sentry, ACCURUM.

DATA SOURCES & APIs

HEALTH

Apple, VALIDIC, practicefusion, fitbit, GARMIN, HUMAN AP, kinsa, MIMIC

IOT

GE Digital, UPTAKE, thingworx, helium, samsara, estimate

FINANCIAL & ECONOMIC DATA

Bloomberg, THOMSON REUTERS, DOW JONES, S&P CAPITAL IQ, CB INSIGHTS, PLAID, SECOND MEASURE, ENVESTNET YODLEE, THE WORLD BANK, GEstimize, PREMISE, Quantd, Engage Alpha, StockTwits, xignite, Thinknum, earnest, predata

AIR / SPACE / SEA

Orbital Insight, planet, SPOTCAT, AIRBOTICS, spire, FREEDOM OF NAVIGATION, kespri, UNDERSTORY, telluslabs, WINDWARD, DroneDeploy, MarineTraffic, LOFT CAPITAL, AERIS

PEOPLE / ENTITIES

acxiom, experian, EPSILON, InsideView, Crimson Hexagon, BASIS RESEARCH, Quantcast, SAFEGRAPH

LOCATION INTELLIGENCE

FOURSQUARE, Mapbox, mapbox, sense360, plinkeybowl, HEXAGON, PlaceIQ, esri, factual, CAR2, Mapillary, Streetline, cuebic, Radar, OpenStreetMap, A

OTHER

DATA.GOV, IMAGENET, LOBBY, APOLLO, CRUX, laffratillo

APPLICATIONS – ENTERPRISE

SALES

- CH@RUS
- INSIDESALES.com
- conversica
- clari
- aviso
- tact.ai
- TROOPS
- fuse|machines
- Clearbit
- ci

MARKETING - B2B

- RADIUS
- EVERSTING
- MINTIGO
- sense
- tubular
- ENGANGIO
- KNOTCH
- mrp
- App Annie
- Lattice
- OK
- Refurbition

MARKETING - B2C

- zeta
- bloomreach
- SendGrid
- brnze
- ACTIONIQ
- BLUECROW
- CONTENTRESEARCH
- TEALUMIA
- mpartice
- Amplero
- amperity
- QUANTIFIND
- Simon
- Lytica
- PERSADO
- remesh

CUSTOMER EXPERIENCE / SERVICE

- qualtrics
- MEDALLIA
- SurveyMonkey
- Ustream Testing
- CLARABridge
- zendesk
- Customer
- freshdesk
- INTERCOM
- Dynatrace
- LIVEPERSON
- Gainsight
- pendo
- Heap
- Amplitude
- Watson Assistant
- Dialogflow
- DigitalGenius
- ASAP
- ade
- automat
- ahmhi
- CallDesk
- metomi
- Q&A
- framel

ENTERPRISE PRODUCTIVITY

- slack
- ORACLE
- GURU lumio
- DIFFBOT
- ok
- talla
- Kassita

HUMAN CAPITAL 	LEGAL JUDICATA ROSS 	REGTECH & COMPLIANCE PARTNERSHIPS 	FINANCE ZUORO TRADESHIFF SCALEFACTOR pilot	BACK OFFICE AUTOMATION & RPA VIDADO NETWORKS KRYON 	SECURITY
--	--	--	---	---	---

– APPLICATIONS – INDUSTRY

[illegible]

The word cloud displays the following company names and logos, organized by category:

- HEALTHCARE:** flatiron, Clover, Xyruus, HealthTap, METABOTA, Gingerio, Glow, babyon, 3DMed, zebra, PIRNA, ovia, TEMPUS, patientslikeme, iAicure, insitro, LianDuo, acifitune, citizen, notable, Immovest, DISCUSION, prognos, eEntic, imig, bookbom, Qventus, ARTERY, IMA GEN, innovacore, PAIGE, DATAVAN, LianTao.
- LIFE SCIENCES:** Zentiva, Carotus, verily, bencor, Bencor, WuxiNextCODE, Clear Labs, fraxnoma, NAANOPIPH, DNANexus, Phosphor, CITRINE, twoAR, Altemus, genomics, OWINK, Zentiva, Carotus, verily, bencor, Bencor, WuxiNextCODE, Clear Labs, fraxnoma, NAANOPIPH, DNANexus, Phosphor, CITRINE, twoAR, Altemus, genomics, OWINK.
- TRANSPORTATION:** uber, tesla, CLEARPATH, CRUISE, nuRO, drive.ai, CAMBRIDGE, Aurora, nauto, AMOTIVE, G7, PILOT.AI, NIO, OPTIMUS, moovit, Ibc, nexar, Kodiak, comma.ai, netradyne, German, Civil Maps, cognata, thind, INRIX.
- AGRICULTURE:** FARMERS, ZODAX, Granular, John Deere, BLUEBERRY, FarmersEdge, AggroStar, FarmLogs, TARANIS, GAMAYA, terram, prospera.
- COMMERCE:** Instacart, STITCH FIX, Du & Co, HealthHero, Hearst, eharmony, stem, SOIERN, BYOTDance, verdigris, duoeto, Electric, ZINIER, Moveworks.
- INDUSTRIAL:** AVIZA, SIEMENS, PREDIX, UPTAKE, SCORTEX, KOMU, TACHYUS.

DATA RESOURCES

The diagram is organized into three columns, each representing a different sector of the AI ecosystem:

- DATA SERVICES:** Includes logos for OPERA, U.S. MAGNET, DATA SCIENCE, fractol, kaggle, EXL, DataKind, and innopocus.
- INCUBATORS & SCHOOLS:** Includes logos for PLURALSIGHT, DataCamp, DataElite, INSIGHT, The Data Incubator, and METIS.
- RESEARCH:** Includes logos for OpenAI, facebook research, MIRI, ML4, VECTOR INSTITUTE, CSAIL, DEKA, Qi, AIZ, and ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE.

INFRASTRUCTURE

HADOOP ON-PREMISE



HADOOP IN THE CLOUD



STREAMING / IN-MEMORY



NoSQL DATABASES



NewSQL DATABASES



GRAPH DBs



MPP DBs



CLOUD EDW



SERVERLESS

