ENERO 31, 2023

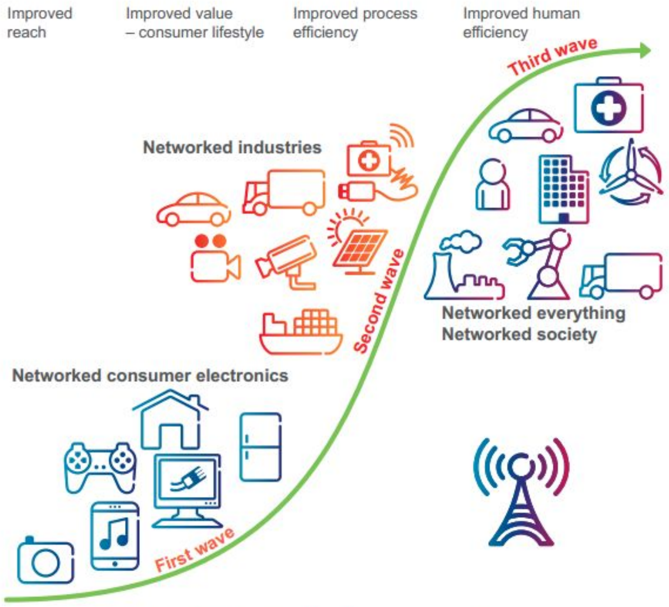
FUNDAMENTOS DE BIG DATA

UNIVERSIDAD Panamericana

Qué es BigData

Nuevo enfoque para entender los datos, dar valor y tomar decisiones, a partir de la descripción de datos (estructurados, no estructurados o semi estructurados) analizados desde un punto de vista no relacional (costo, tiempo, recursos), utilizando herramientas de supercómputo nativo o en la nube.

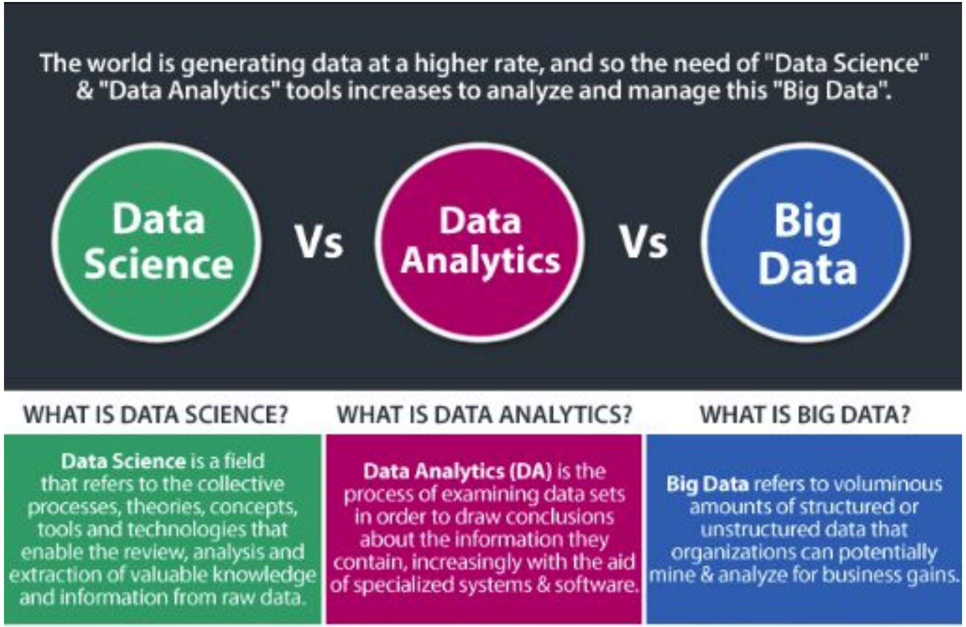
La relación del Big Data con otras tecnologías


Improved reach
Improved value – consumer lifestyle
Improved process efficiency
Improved human efficiency
Third wave
Networked industries
Second wave
Networked everything
Networked society
Networked consumer electronics
First wave

The three waves of connected device development.

- **Gran masa**
- **Diversidad**
- **Almacenamiento**
- **Seguridad**
- **Tomar decisiones y responder**

Qué es BigData



The world is generating data at a higher rate, and so the need of "Data Science" & "Data Analytics" tools increases to analyze and manage this "Big Data".

WHAT IS DATA SCIENCE?

Data Science is a field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis and extraction of valuable knowledge and information from raw data.

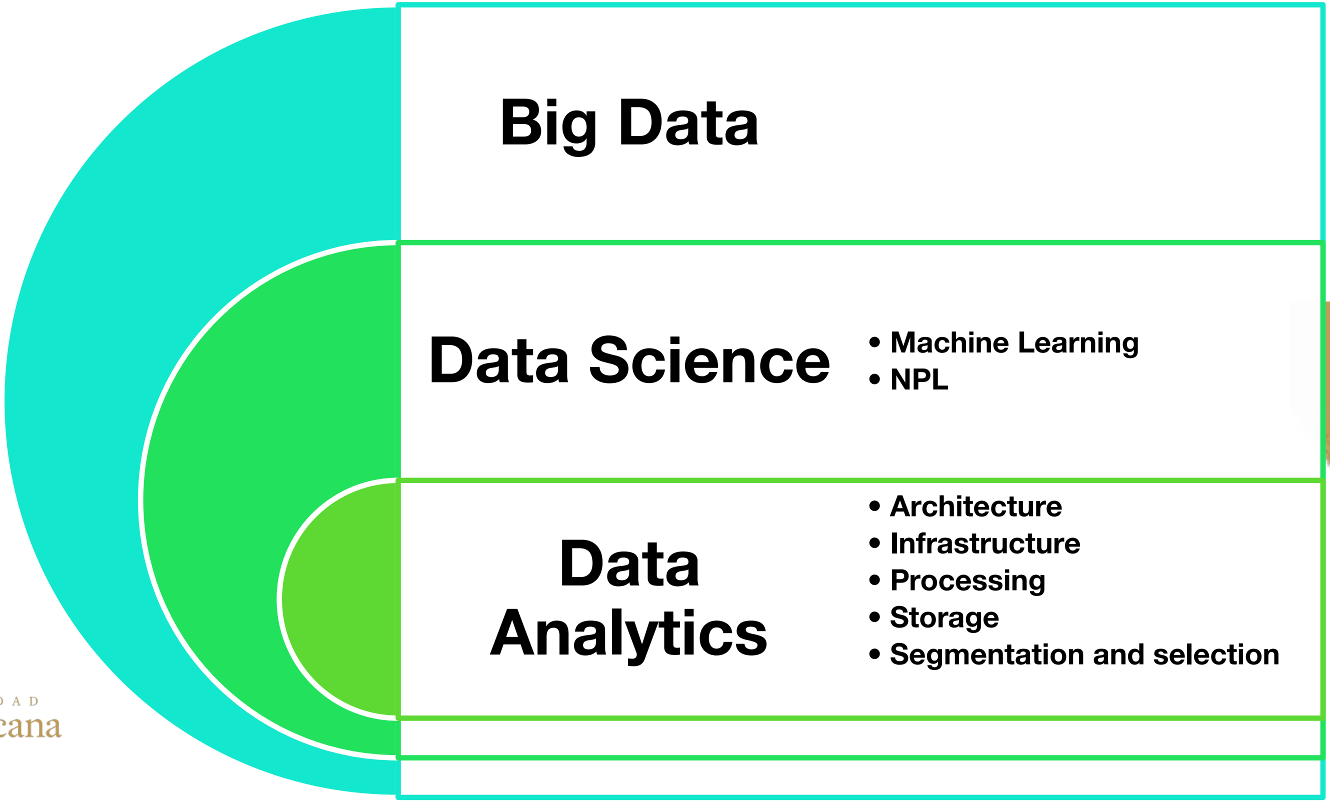
WHAT IS DATA ANALYTICS?

Data Analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems & software.

WHAT IS BIG DATA?

Big Data refers to voluminous amounts of structured or unstructured data that organizations can potentially mine & analyze for business gains.

Qué es BigData



Big Data

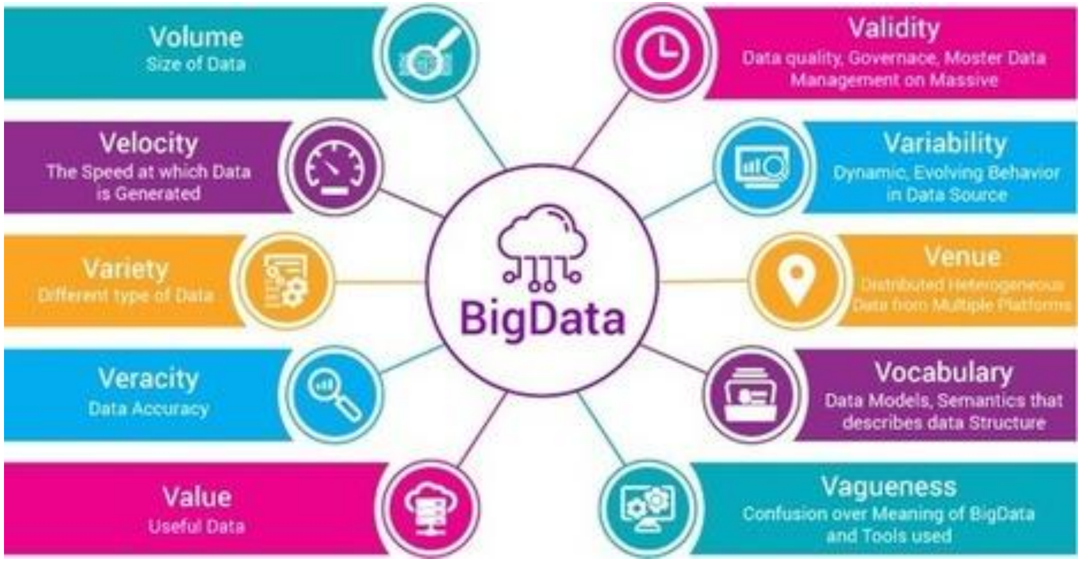
Data Science

- Machine Learning
- NPL

Data Analytics

- Architecture
- Infrastructure
- Processing
- Storage
- Segmentation and selection

Qué es BigData, 10 V’s



BigData

Volume
Size of Data

Velocity
The Speed at which Data is Generated

Variety
Different type of Data

Veracity
Data Accuracy

Value
Useful Data

Validity
Data quality, Governace, Moster Data Management on Massive

Variability
Dynamic, Evolving Behavior in Data Source

Venue
Distributed Heterogeneous Data from Multiple Platforms

Vocabulary
Data Models, Semantics that describes data Structure

Vagueness
Confusion over Meaning of BigData and Tools used

Qué es Big Data

Google BigQuery

mongoDB®

Qué es necesario

Infraestructura

Fuente(s) de datos

Modelo matemático

Big Data, generalidades

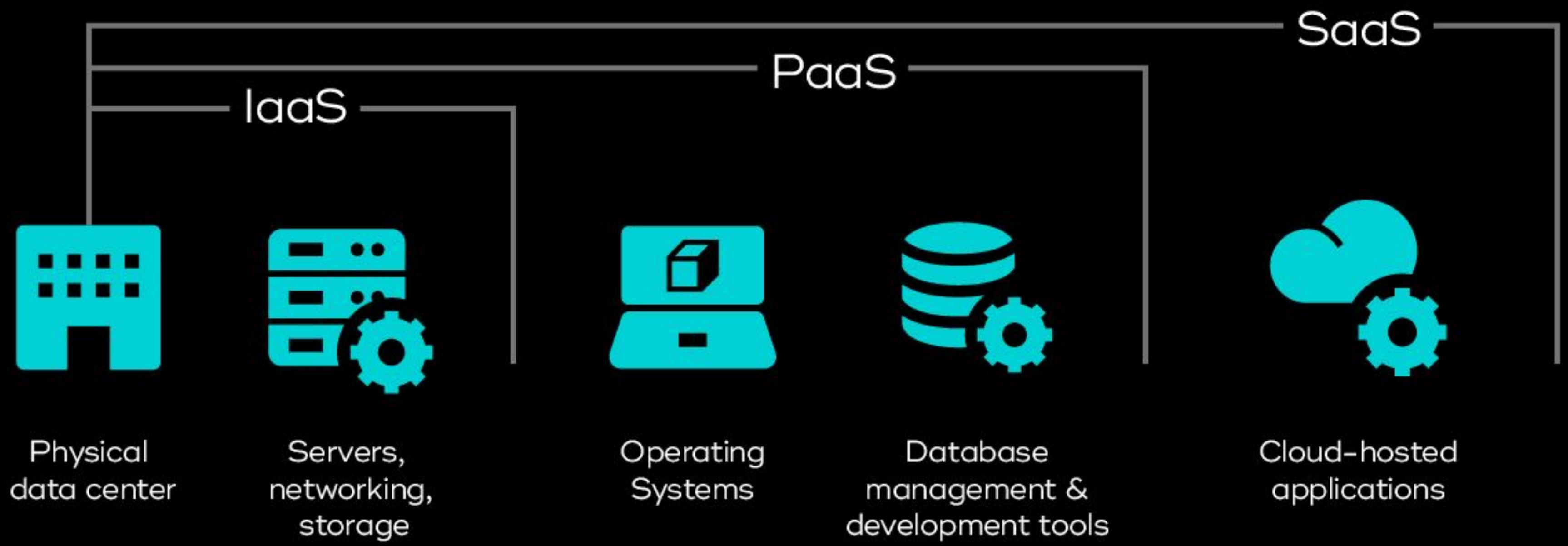
- Los motores de BD tradicionales no pueden tratar altas cantidades de datos, ni reaccionar inmediatamente a ellos. Almacena **datos no relacionales y relacionales**.

- Esquemas de consulta:

1. **Key-value**
2. **Columnal**
3. **Documental**

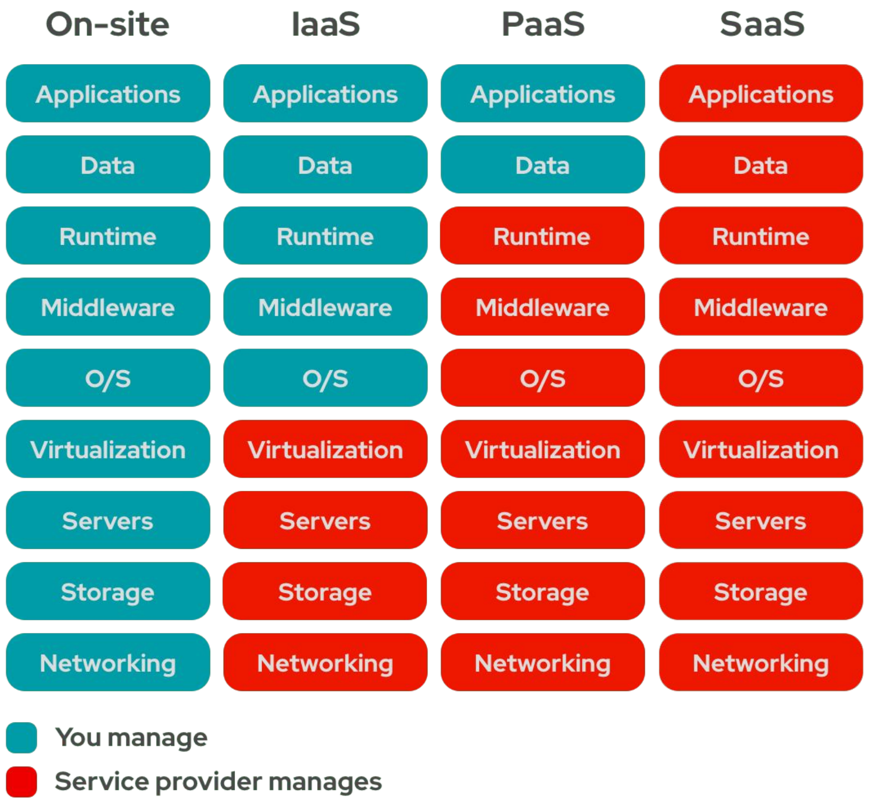
UNIVERSIDAD
Panamericana

Cloud Computing Overview


SaaS
PaaS
IaaS
Physical data center
Servers, networking, storage
Operating Systems
Database management & development tools
Cloud-hosted applications

IaaS, Infrastructure as a Service
PaaS, Platform as a Service
SaaS, Software as a Service

Escenarios de cómputo en la nube



On-site	IaaS	PaaS	SaaS
🟦 Applications	🟦 Applications	🟦 Applications	🟥 Applications
🟦 Data	🟦 Data	🟦 Data	🟥 Data
🟦 Runtime	🟦 Runtime	🟥 Runtime	🟥 Runtime
🟦 Middleware	🟦 Middleware	🟥 Middleware	🟥 Middleware
🟦 O/S	🟦 O/S	🟥 O/S	🟥 O/S
🟦 Virtualization	🟥 Virtualization	🟥 Virtualization	🟥 Virtualization
🟦 Servers	🟥 Servers	🟥 Servers	🟥 Servers
🟦 Storage	🟥 Storage	🟥 Storage	🟥 Storage
🟦 Networking	🟥 Networking	🟥 Networking	🟥 Networking

🟦 You manage

🟥 Service provider manages

UNIVERSIDAD
Panamericana

Tecnologías líderes Big Data engine

ORACLE
NOSQL DATABASE

hadoop

amazon
web services

IBM

mongoDB

Google Cloud Platform

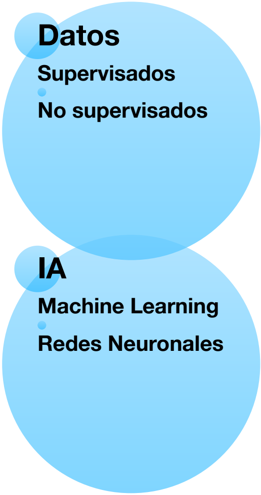
Cassandra

Microsoft
Azure

Qué se puede hacer con BigData

La nube consumía software, la nube ahora consume datos e integrando software desde donde se puede obtener información inteligente:

Análisis retrospectivo
Análisis en tiempo real
Análisis predictivo
App Inteligentes SaaS

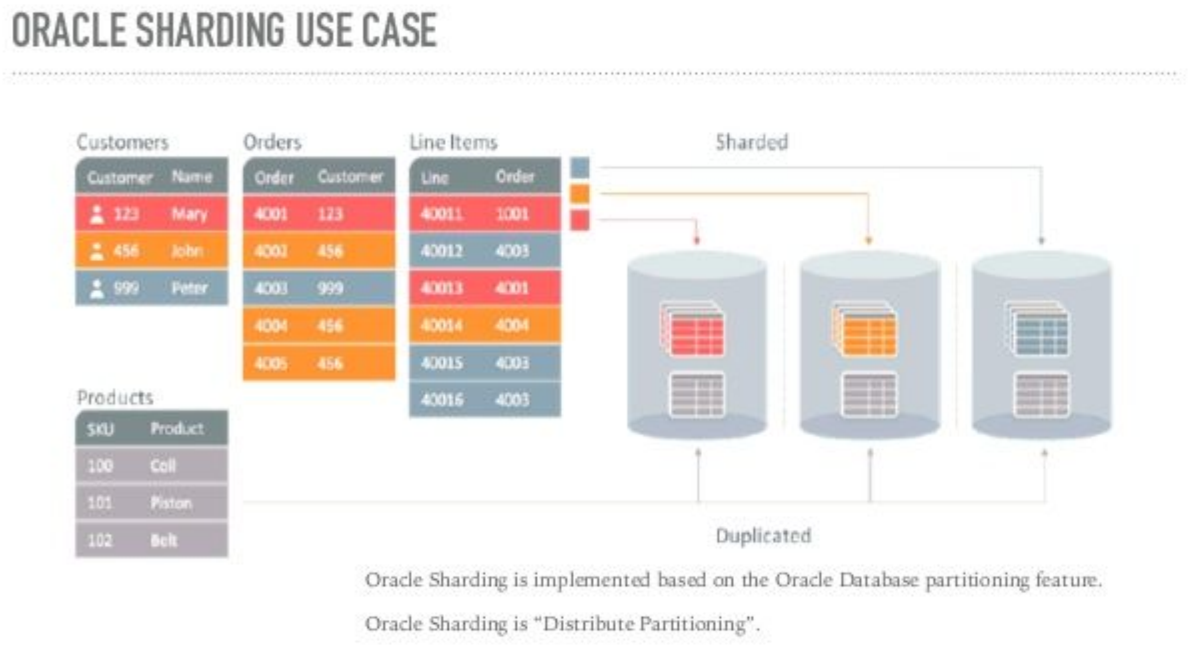


Datos
- Supervisados
- No supervisados

IA
- Machine Learning
- Redes Neuronales

Queue adicional y backup

- Sharding



ORACLE SHARDING USE CASE

Customers

Customer	Name
123	Mary
456	John
999	Peter

Orders

Order	Customer
4001	123
4002	456
4003	999
4004	456
4005	456

Line Items

Line	Order
40011	1001
40012	4003
40013	4001
40014	4004
40015	4003
40016	4003

Products

SKU	Product
100	Coil
101	Piston
102	Belt

Oracle Sharding is implemented based on the Oracle Database partitioning feature.

Oracle Sharding is "Distribute Partitioning".

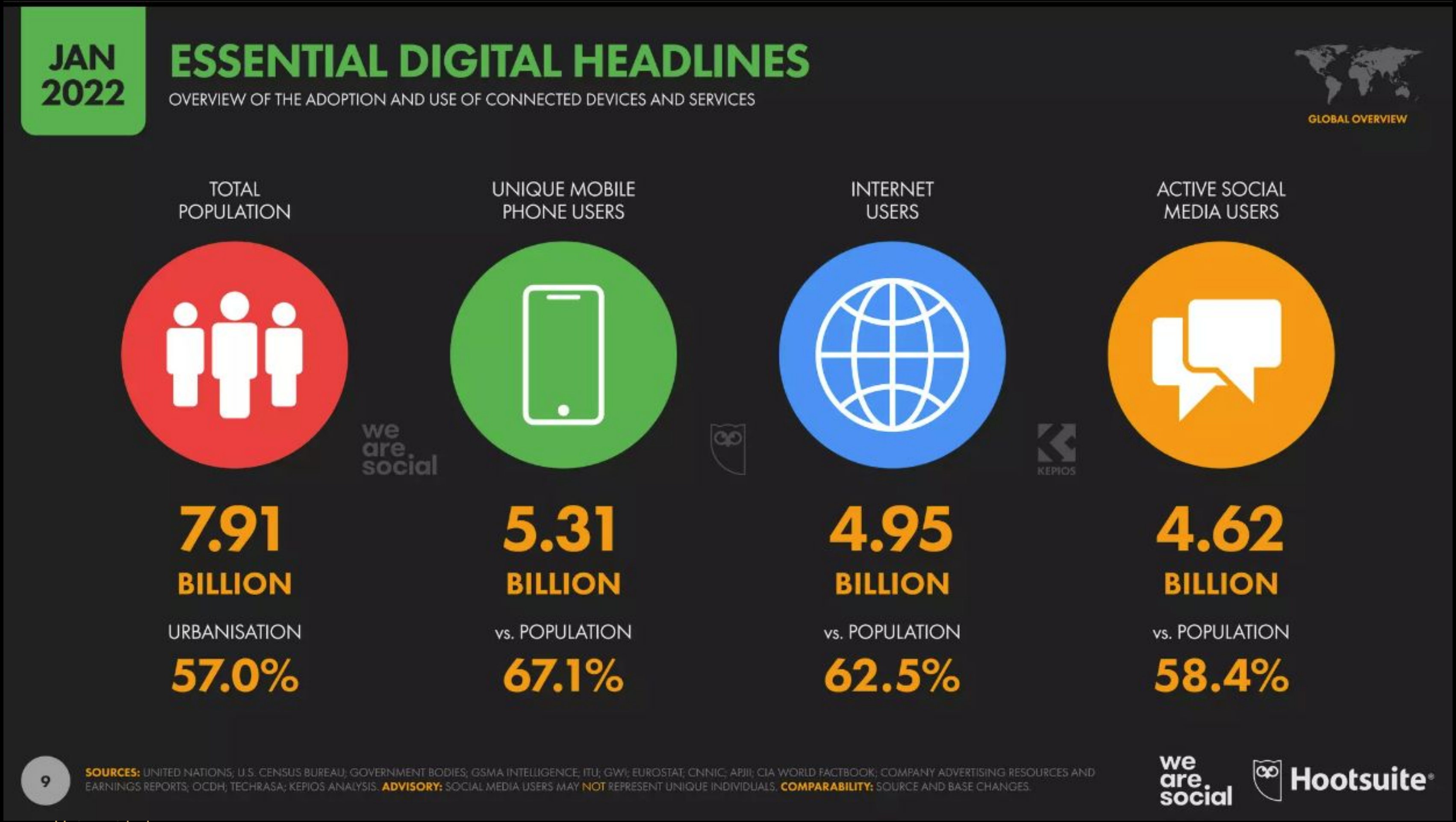
Algo de historia

- Big Data fue precedido por Google (mapreduce), Amazon (Dynamo) y Software Open Source (Hadoop, MongoDB, Cassandra, RabbitMQ, etc)

- RDBMS tienen funciones específicas donde las relaciones entre los datos toman valor único

Propiedades deseadas de un BD

1. Robustes y tolerancia a fallas
2. Baja latencia y actualizaciones
3. Escalabilidad
4. Generalización
5. Extensabilidad
6. Queries ad-hoc
7. Mantenimiento mínimo
8. Debugabilidad
9. Problemas de Arq. Incremental (nuevas funciones)



JAN 2022

ESSENTIAL DIGITAL HEADLINES

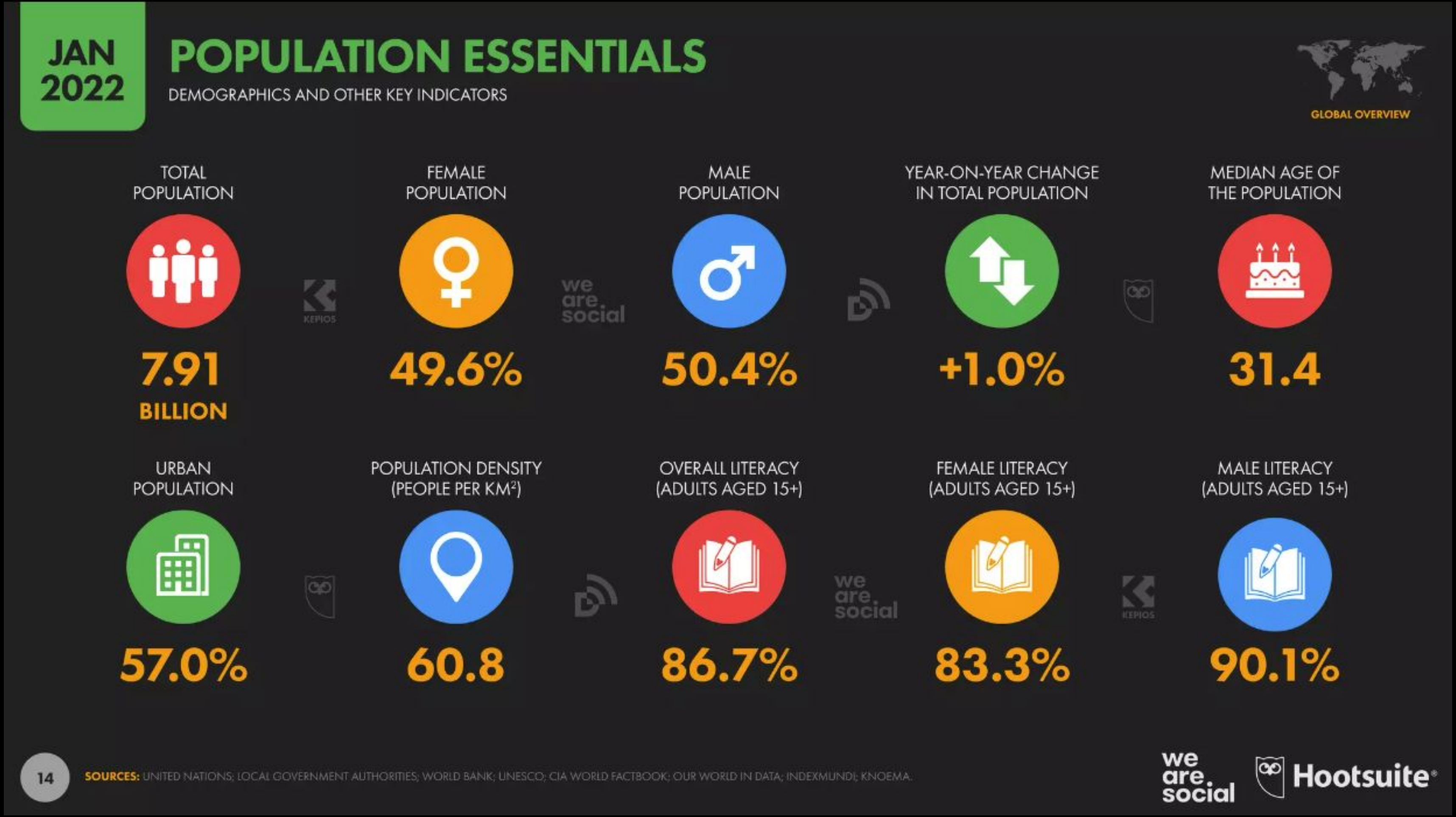
OVERVIEW OF THE ADOPTION AND USE OF CONNECTED DEVICES AND SERVICES

GLOBAL OVERVIEW

TOTAL POPULATION	UNIQUE MOBILE PHONE USERS	INTERNET USERS	ACTIVE SOCIAL MEDIA USERS
7.91 BILLION	5.31 BILLION	4.95 BILLION	4.62 BILLION
URBANISATION	vs. POPULATION	vs. POPULATION	vs. POPULATION
57.0%	67.1%	62.5%	58.4%

SOURCES: UNITED NATIONS; U.S. CENSUS BUREAU; GOVERNMENT BODIES; GSMA INTELLIGENCE; ITU; GWI; EUROSTAT; CNNIC; APJII; CIA WORLD FACTBOOK; COMPANY ADVERTISING RESOURCES AND EARNINGS REPORTS; OCDH; TECHRASA; KEPIOS ANALYSIS. **ADVISORY:** SOCIAL MEDIA USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. **COMPARABILITY:** SOURCE AND BASE CHANGES.

we are social Hootsuite



JAN 2022

POPULATION ESSENTIALS

DEMOGRAPHICS AND OTHER KEY INDICATORS

GLOBAL OVERVIEW

TOTAL POPULATION

7.91 BILLION

FEMALE POPULATION

49.6%

MALE POPULATION

50.4%

YEAR-ON-YEAR CHANGE IN TOTAL POPULATION

+1.0%

MEDIAN AGE OF THE POPULATION

31.4

URBAN POPULATION

57.0%

POPULATION DENSITY (PEOPLE PER KM²)

60.8

OVERALL LITERACY (ADULTS AGED 15+)

86.7%

FEMALE LITERACY (ADULTS AGED 15+)

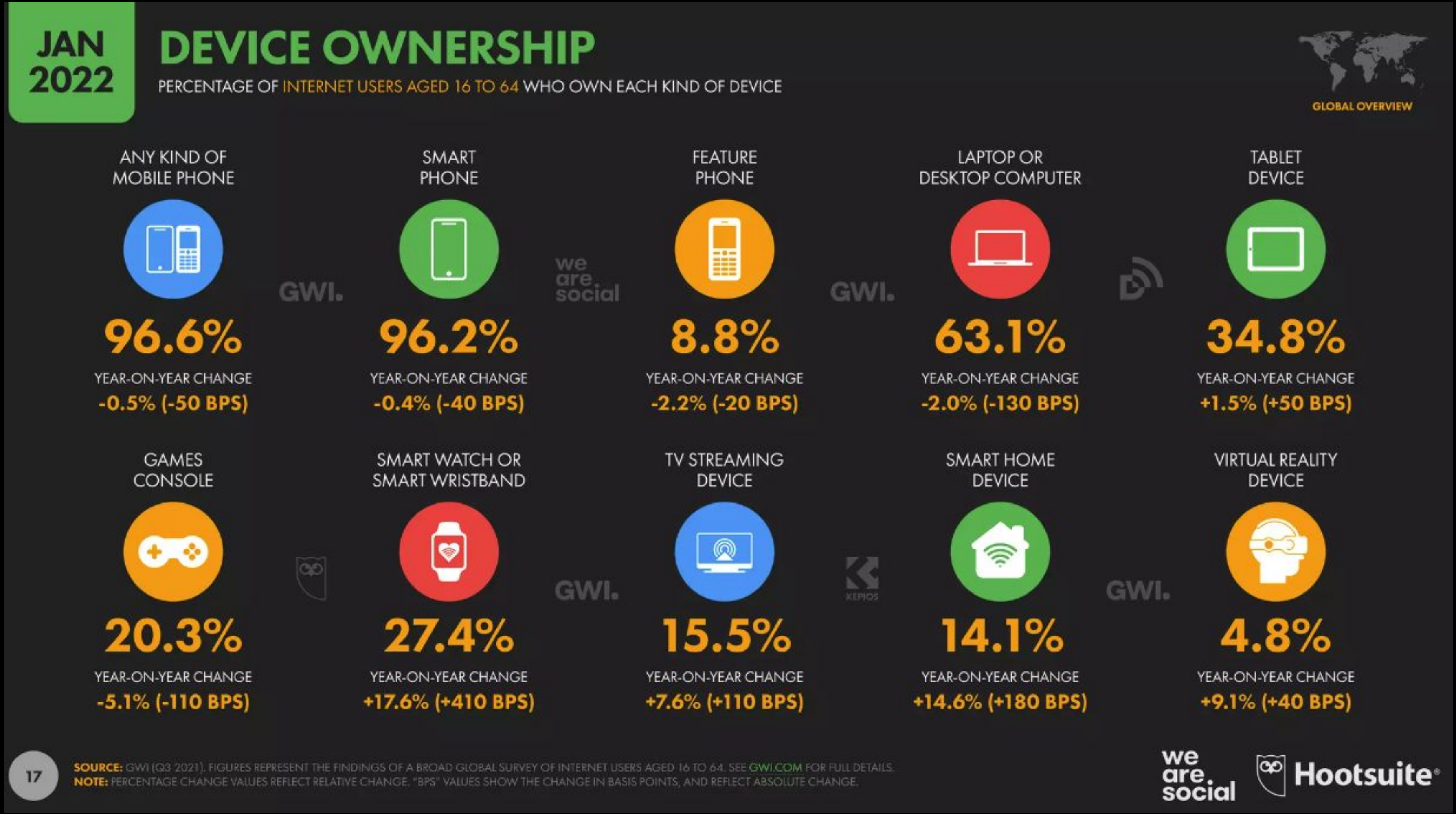
83.3%

MALE LITERACY (ADULTS AGED 15+)

90.1%

SOURCES: UNITED NATIONS; LOCAL GOVERNMENT AUTHORITIES; WORLD BANK; UNESCO; CIA WORLD FACTBOOK; OUR WORLD IN DATA; INDEXMUNDI; KNOEMA.

we are social Hootsuite®



JAN 2022

DEVICE OWNERSHIP

PERCENTAGE OF INTERNET USERS AGED 16 TO 64 WHO OWN EACH KIND OF DEVICE

GLOBAL OVERVIEW

ANY KIND OF MOBILE PHONE
96.6%
YEAR-ON-YEAR CHANGE
-0.5% (-50 BPS)

SMART PHONE
96.2%
YEAR-ON-YEAR CHANGE
-0.4% (-40 BPS)

FEATURE PHONE
8.8%
YEAR-ON-YEAR CHANGE
-2.2% (-20 BPS)

LAPTOP OR DESKTOP COMPUTER
63.1%
YEAR-ON-YEAR CHANGE
-2.0% (-130 BPS)

TABLET DEVICE
34.8%
YEAR-ON-YEAR CHANGE
+1.5% (+50 BPS)

GAMES CONSOLE
20.3%
YEAR-ON-YEAR CHANGE
-5.1% (-110 BPS)

SMART WATCH OR SMART WRISTBAND
27.4%
YEAR-ON-YEAR CHANGE
+17.6% (+410 BPS)

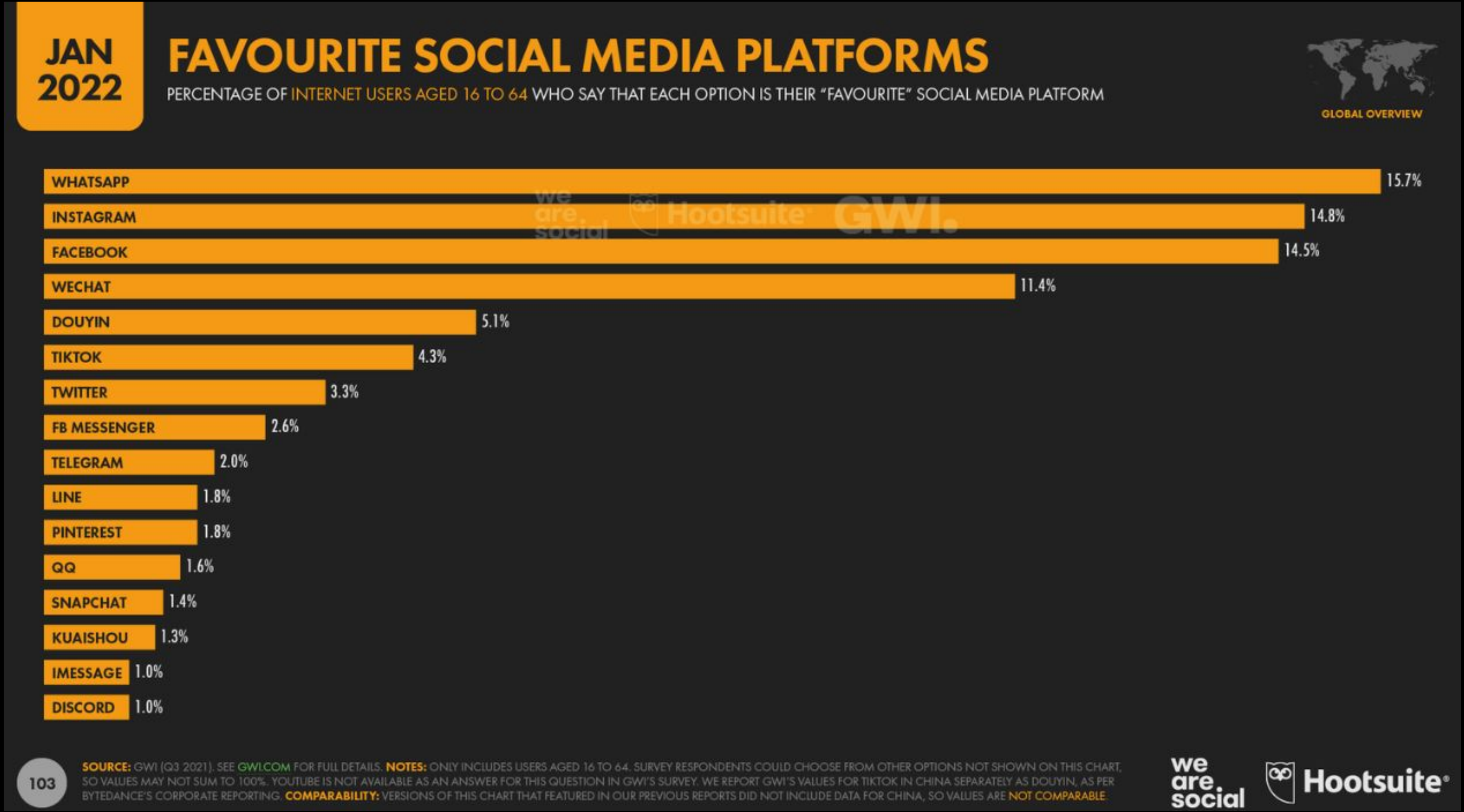
TV STREAMING DEVICE
15.5%
YEAR-ON-YEAR CHANGE
+7.6% (+110 BPS)

SMART HOME DEVICE
14.1%
YEAR-ON-YEAR CHANGE
+14.6% (+180 BPS)

VIRTUAL REALITY DEVICE
4.8%
YEAR-ON-YEAR CHANGE
+9.1% (+40 BPS)

SOURCE: GWI (Q3 2021). FIGURES REPRESENT THE FINDINGS OF A BROAD GLOBAL SURVEY OF INTERNET USERS AGED 16 TO 64. SEE GWI.COM FOR FULL DETAILS.
NOTE: PERCENTAGE CHANGE VALUES REFLECT RELATIVE CHANGE. "BPS" VALUES SHOW THE CHANGE IN BASIS POINTS, AND REFLECT ABSOLUTE CHANGE.

we are social Hootsuite



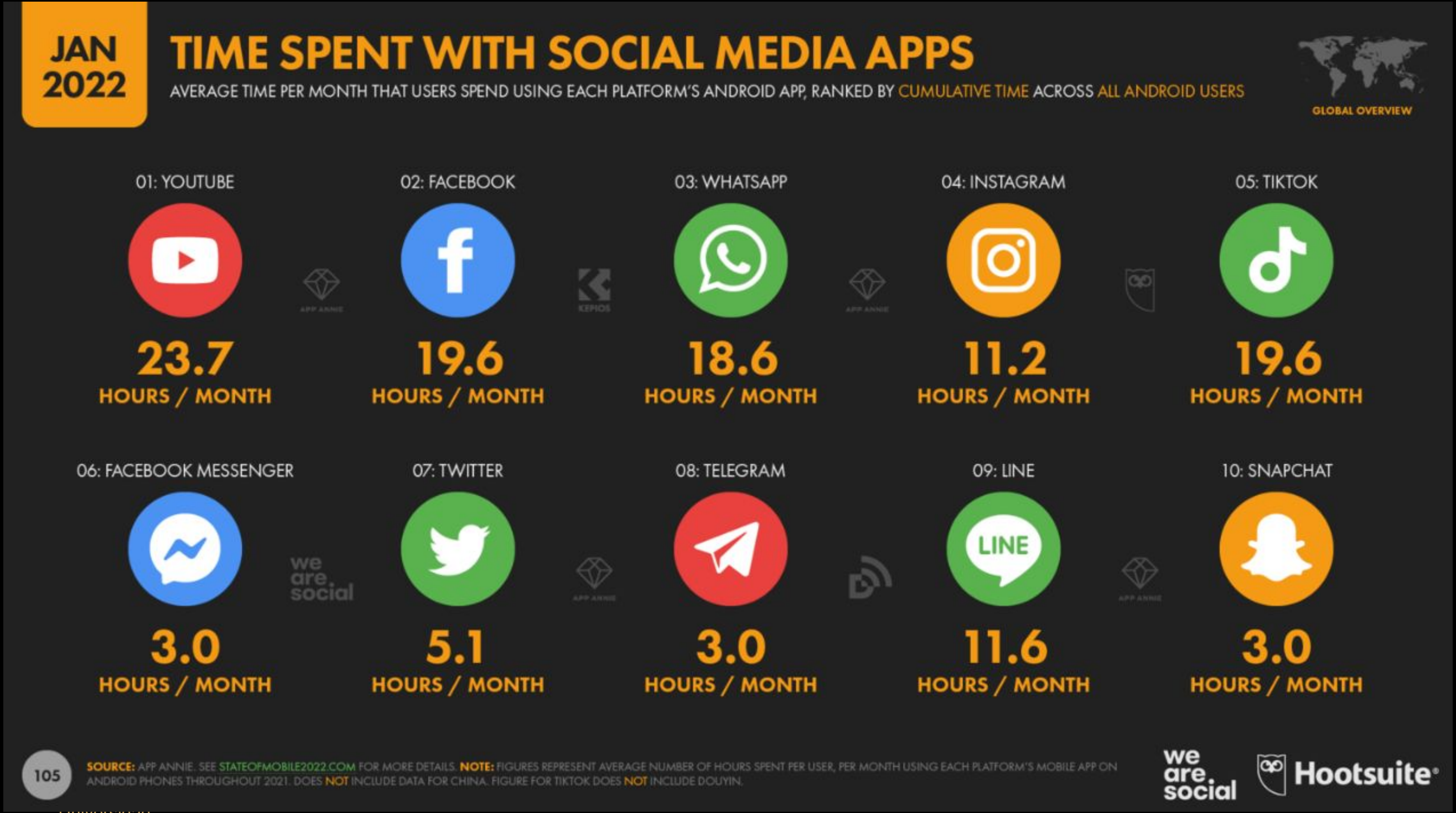
JAN 2022

FAVOURITE SOCIAL MEDIA PLATFORMS

PERCENTAGE OF INTERNET USERS AGED 16 TO 64 WHO SAY THAT EACH OPTION IS THEIR "FAVOURITE" SOCIAL MEDIA PLATFORM

SOURCE: GWI (Q3 2021). SEE GWI.COM FOR FULL DETAILS. **NOTES:** ONLY INCLUDES USERS AGED 16 TO 64. SURVEY RESPONDENTS COULD CHOOSE FROM OTHER OPTIONS NOT SHOWN ON THIS CHART, SO VALUES MAY NOT SUM TO 100%. YOUTUBE IS NOT AVAILABLE AS AN ANSWER FOR THIS QUESTION IN GWI'S SURVEY. WE REPORT GWI'S VALUES FOR TIKTOK IN CHINA SEPARATELY AS DOUYIN, AS PER BYTEDANCE'S CORPORATE REPORTING. **COMPARABILITY:** VERSIONS OF THIS CHART THAT FEATURED IN OUR PREVIOUS REPORTS DID NOT INCLUDE DATA FOR CHINA, SO VALUES ARE NOT COMPARABLE.

Panamericana



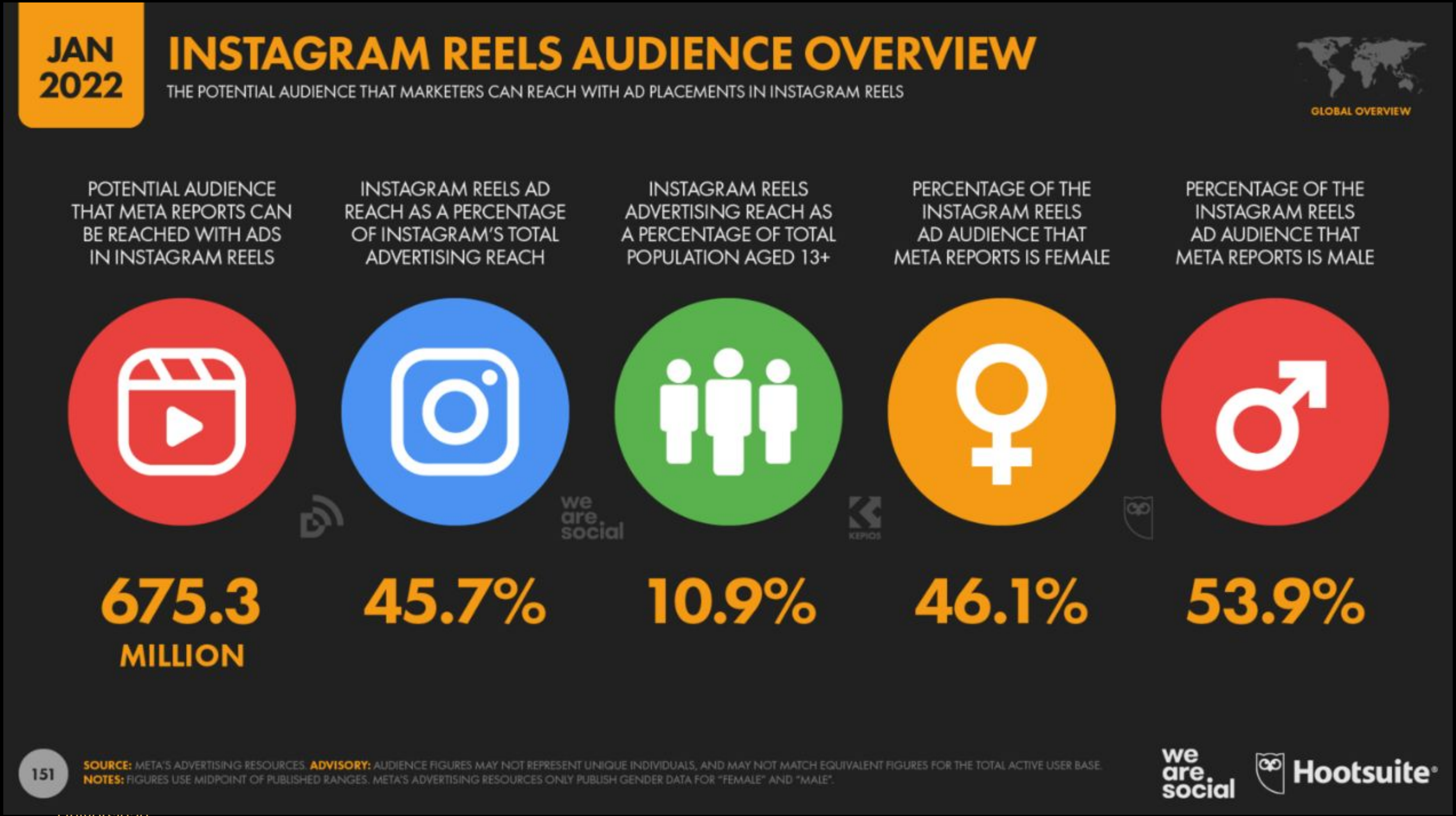
JAN 2022

TIME SPENT WITH SOCIAL MEDIA APPS

AVERAGE TIME PER MONTH THAT USERS SPEND USING EACH PLATFORM'S ANDROID APP, RANKED BY CUMULATIVE TIME ACROSS ALL ANDROID USERS

01: YOUTUBE	23.7 HOURS / MONTH
02: FACEBOOK	19.6 HOURS / MONTH
03: WHATSAPP	18.6 HOURS / MONTH
04: INSTAGRAM	11.2 HOURS / MONTH
05: TIKTOK	19.6 HOURS / MONTH
06: FACEBOOK MESSENGER	3.0 HOURS / MONTH
07: TWITTER	5.1 HOURS / MONTH
08: TELEGRAM	3.0 HOURS / MONTH
09: LINE	11.6 HOURS / MONTH
10: SNAPCHAT	3.0 HOURS / MONTH

SOURCE: APP ANNIE. SEE STATEOFMOBILE2022.COM FOR MORE DETAILS. NOTE: FIGURES REPRESENT AVERAGE NUMBER OF HOURS SPENT PER USER, PER MONTH USING EACH PLATFORM'S MOBILE APP ON ANDROID PHONES THROUGHOUT 2021. DOES NOT INCLUDE DATA FOR CHINA. FIGURE FOR TIKTOK DOES NOT INCLUDE DOUYIN.



JAN 2022

INSTAGRAM REELS AUDIENCE OVERVIEW

THE POTENTIAL AUDIENCE THAT MARKETERS CAN REACH WITH AD PLACEMENTS IN INSTAGRAM REELS

GLOBAL OVERVIEW

POTENTIAL AUDIENCE THAT META REPORTS CAN BE REACHED WITH ADS IN INSTAGRAM REELS	INSTAGRAM REELS AD REACH AS A PERCENTAGE OF INSTAGRAM'S TOTAL ADVERTISING REACH	INSTAGRAM REELS ADVERTISING REACH AS A PERCENTAGE OF TOTAL POPULATION AGED 13+	PERCENTAGE OF THE INSTAGRAM REELS AD AUDIENCE THAT META REPORTS IS FEMALE	PERCENTAGE OF THE INSTAGRAM REELS AD AUDIENCE THAT META REPORTS IS MALE
675.3 MILLION	45.7%	10.9%	46.1%	53.9%

151

SOURCE: META'S ADVERTISING RESOURCES. **ADVISORY:** AUDIENCE FIGURES MAY NOT REPRESENT UNIQUE INDIVIDUALS, AND MAY NOT MATCH EQUIVALENT FIGURES FOR THE TOTAL ACTIVE USER BASE.
NOTES: FIGURES USE MIDPOINT OF PUBLISHED RANGES. META'S ADVERTISING RESOURCES ONLY PUBLISH GENDER DATA FOR "FEMALE" AND "MALE".

Necesidades Globales de datos:

1. Crecimiento de logs exponencial
2. Crecimiento de demanda
3. Crecimiento de popularidad
4. Mayor inteligencia

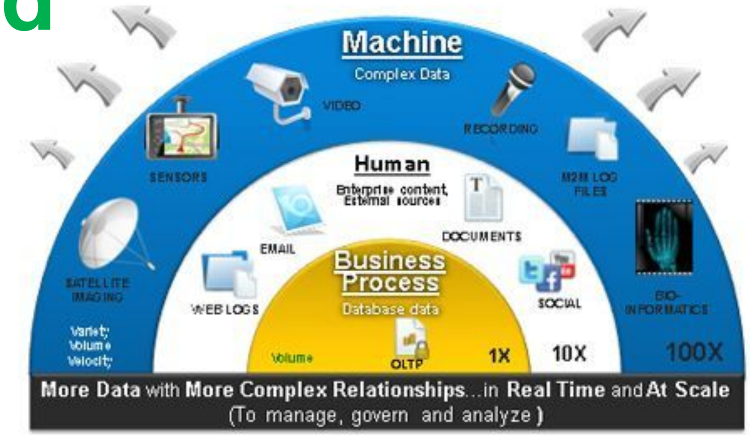
¿Existe una correlación entre el nivel de vida de un país medido por el ingreso econoómico de su población, edad y el acceso a internet?

Universidad
Panamericana

https://global-internet-map-2022.telegeography.com/
https://datareportal.com/reports/digital-2022-global-overview-report

Data sources types in Big Data

1. **Structured**
2. **None structured**
3. **Semi-structured**



More Data with More Complex Relationships...in Real Time and At Scale
(To manage, govern and analyze)

Structured Data

Microsoft SQL Server
ORACLE
PostgreSQL
MySQL
SYBASE
IBM DB2.

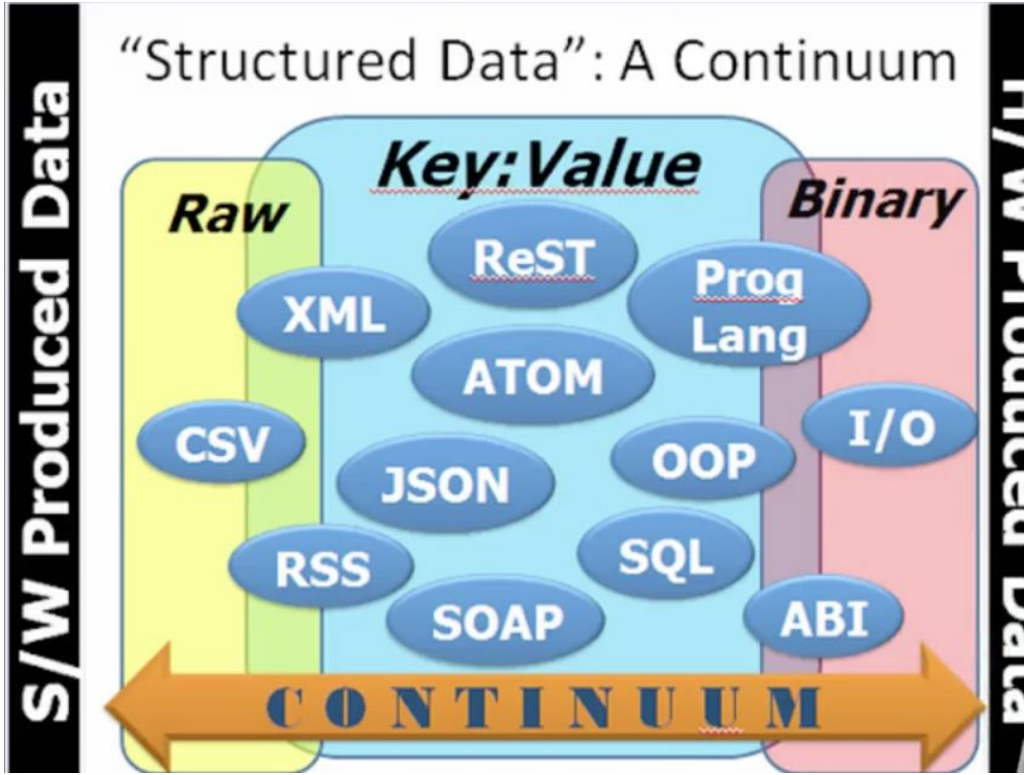
- **Big mass**
- **Diversity**
- **Storage**
- **Security**
- **Decision Support Systems and Answers Recovery**

None Structured Data, Organizations

- **80 -90% world wide data**
- **Images, videos, email, searches, text, pdfs, etc.**
- **V =** $\frac{\Delta X}{\Delta t}$

UNIVERSIDAD
Panamericana

Data schemas, samples



S/W Produced Data

"Structured Data": A Continuum

Raw

Key:Value

Binary

ReST

XML

Prog Lang

ATOM

CSV

I/O

JSON

OOP

RSS

SQL

SOAP

ABI

CONTINUUM

H/W Produced Data

Structured Data, **Organizations**

Past stored in SILOS

- Unconnected stored islands
- Hindered stored and connect silos for pattern recognition
- Outdated and unsynchronized

RDMS (Highly structured data) – SILOS □ Value

UNIVERSIDAD
Panamericana

Structured Data, Organizations

Using pattern recognition a organizations can use it for:

1.- Detect correlated products

2.- Estimated demand

3.- Dapture fraudulent actions

Commerce+Open Data+Analytics□Better predictions (Business Intelligence)

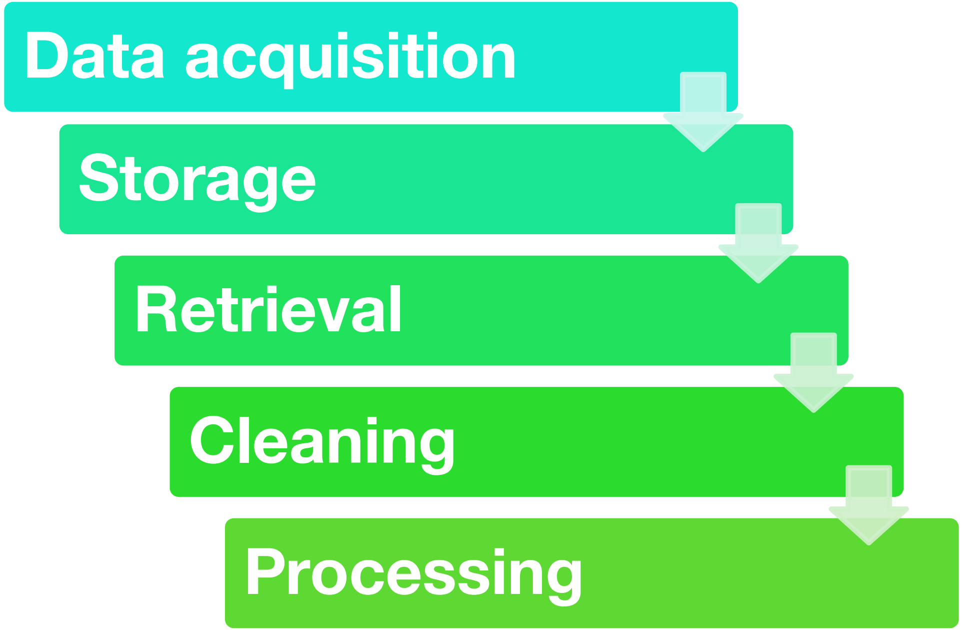
UNIVERSIDAD Panamericana

Semi- Structured Data

16 PB data per year
1 mile per driver route optimized, savings 50 mdd

250 millions clients, 10000 stores
2.5 PB per hour
New products, customize recommendations, predictive support

None-Structured Data, work path



Data acquisition

Storage

Retrieval

Cleaning

Processing

For all data workload

ETL data
(Extract, Transform and Load)

To integrate different sources and loaded it into:
Data Warehouse-Data Lake-Data Mart

4 key properties of a data transaction

Atomicity

All changes to data are performed as if they are a single operation

Consistency

Data is in a consistent state when a transaction starts and when it ends.

Isolation

The intermediate state of a transaction is invisible to other transactions. As a result, transactions that run concurrently appear to be serialized.

Durability

After a transaction successfully completes, changes to data persist and are not undone, even in the event of a system failure.

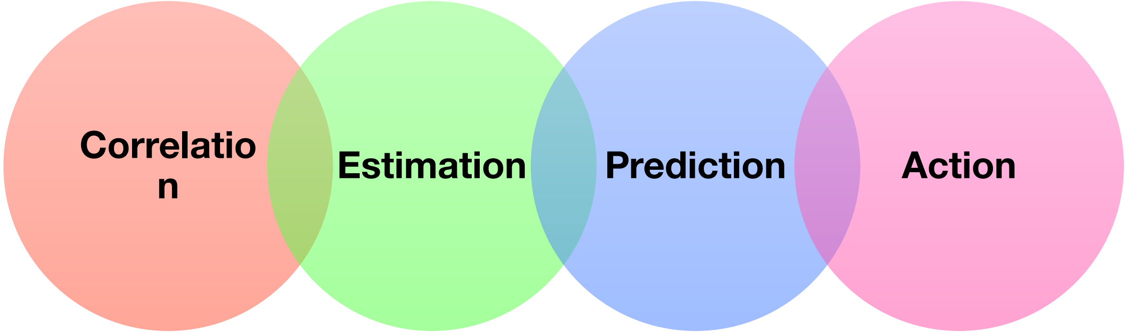
None Structured Data, NoSQL

1. Graph Data Base.– used to find connections between data sets (Neo4j)
2. Key Value Pairs.- access and process data with key value pairs (Cassandra)

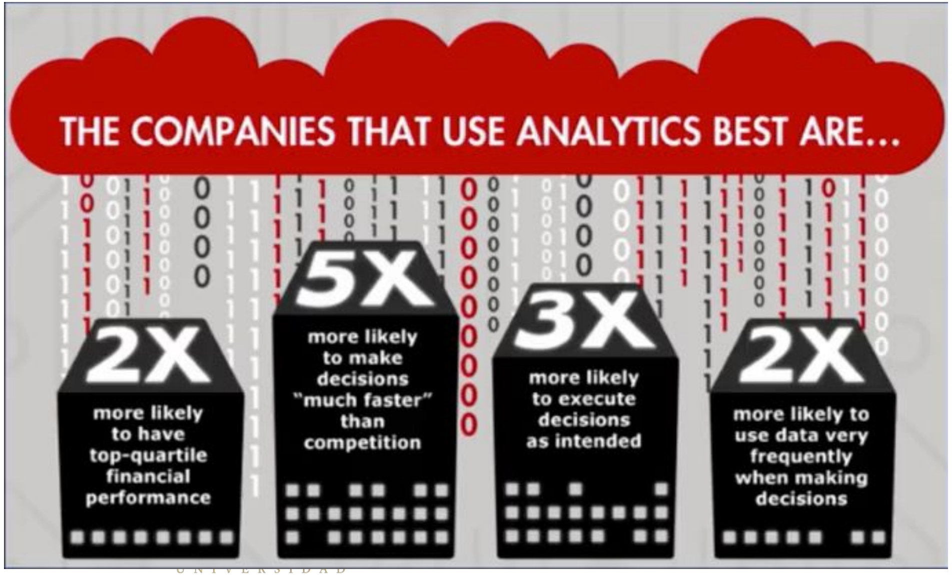
- □Layers for Value Big Data
  - □Retrival and Storage
    - □Pre-processing
      - □Analysis

12 TB/day representative data to sentiment analysis for a product or service (crisis mappers)

None Structured Data, NoSQL
Flow decision making



Strategical Big Data growth, new economical system approach



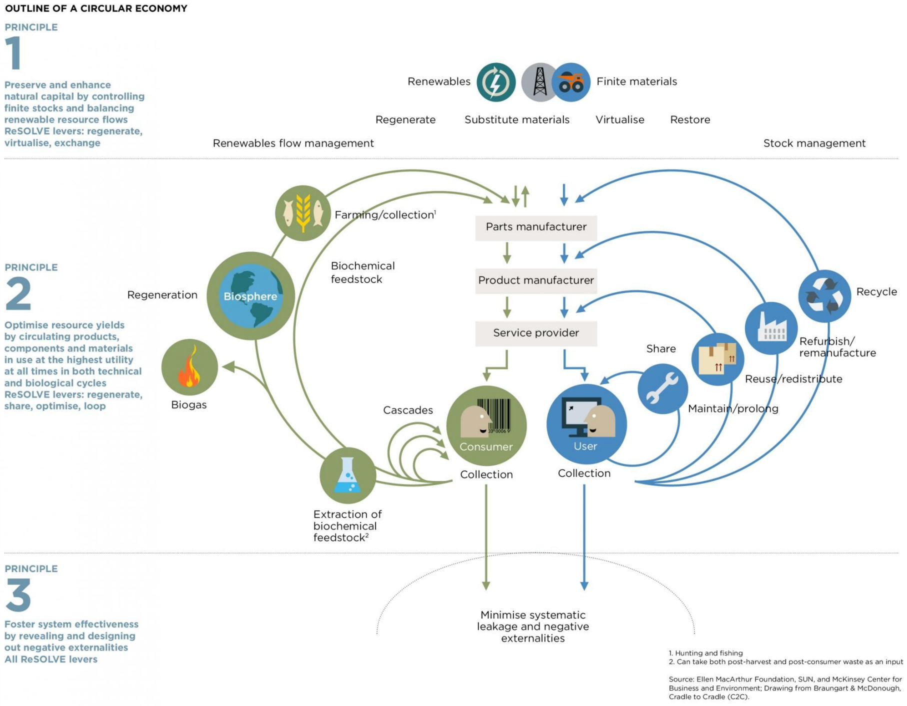
THE COMPANIES THAT USE ANALYTICS BEST ARE...

2X more likely to have top-quartile financial performance

5X more likely to make decisions "much faster" than competition

3X more likely to execute decisions as intended

2X more likely to use data very frequently when making decisions



OUTLINE OF A CIRCULAR ECONOMY

PRINCIPLE 1

Preserve and enhance natural capital by controlling finite stocks and balancing renewable resource flows
ReSOLVE levers: regenerate, virtualise, exchange

PRINCIPLE 2

Optimise resource yields by circulating products, components and materials in use at the highest utility at all times in both technical and biological cycles
ReSOLVE levers: regenerate, share, optimise, loop

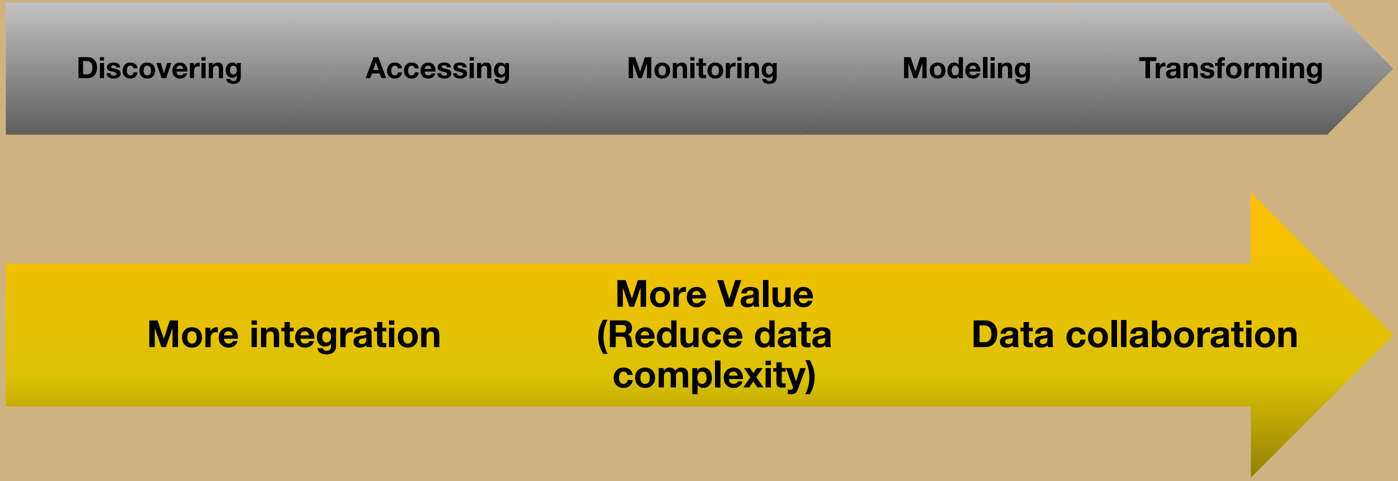
PRINCIPLE 3

Foster system effectiveness by revealing and designing out negative externalities
All ReSOLVE levers

[1] Hunting and fishing
[2] Can take both post-harvest and post-consumer waste as an input

Source: Ellen MacArthur Foundation, SUN, and McKinsey Center for Business and Environment; Drawing from Braungart & McDonough, Cradle to Cradle (C2C).

Data integration Process



Discovering

Accessing

Monitoring

Modeling

Transforming

More integration

More Value
(Reduce data
complexity)

Data collaboration

Activity 2:

1. Read the appropiate article according the assignated team.
2. Answer the questiones asignated to the article.
3. Every team must prepare and present a 5 minute resume

Team 1

- **Link**

Team 2

- **Link**

Team 3

- **Link**

Team 4

- **Link**

Team 5

- Link

Questions Team 1:

1. Does the Reels are changing the consumer experiences and it reflect a new data Vs value?
2. Wich one's and how?
3. What trends (3) do you identify after Covid-19 over data user production?
4. Are those data commonly Structured, NoStructured or SemiStructured? Explain, why?

Questions Team 2:

1. What are the relations between Data Analytics and Business Intelligence (3)?
2. What are the relations between Data Science and Business Intelligence (3)?
3. How do you explain this for a new user or Company who wants to be involved in BI?
4. How could they start with it?

Questions Team 3:

1. For the 12 data keypoint, how many of them are affeted by AI?
2. In conclusion, the D&A should be focused on people or metaverse? Explain, why?
3. How do you resume this article in one paragraph?
4. Does geopolitics are changing data production and consumtion?

Questions Team 4:

1. How do you represent graphically the relation between Data Science-ML-AI?
2. How do you explain the role of Big Data in Data Science, ML and AI?
3. Describe three agrees, and three disagrees about the article's content
4. According the study case: "Self-driving car". Do you agree with the ML, AI and Data Science descriptions? Explain, What is the Big Data role in this case according 10 V's?

Questions Team 5:

1. Does Data Scientist and Data Engineer would do the same job? Explain, why?
2. Define five key skills that make difference between Data Scient and Data Engineer
3. Does DASCA is leading changing the employment bases for Data Scientist and Data Engineers?
4. How has been involved the Big Data in this new industrial revolution called 4.0?