Herramientas para Big Data

Juan Pablo Soto & Julio Waissman

Contenido

- Introducción a Big Data
- Contenedores Docker
- Manejo de versiones con Git
- Libretas Jupyter
- Spark
  - Arquitectura
  - Instalación
  - Uso básico
- PySpark
- TensorFlow
  - Arquitectura
  - Uso básico

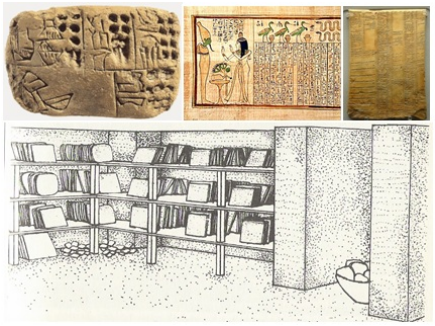
APACHE Spark™

jupyter

TensorFlow

docker

Datos / Información / Conocimiento



Big Data

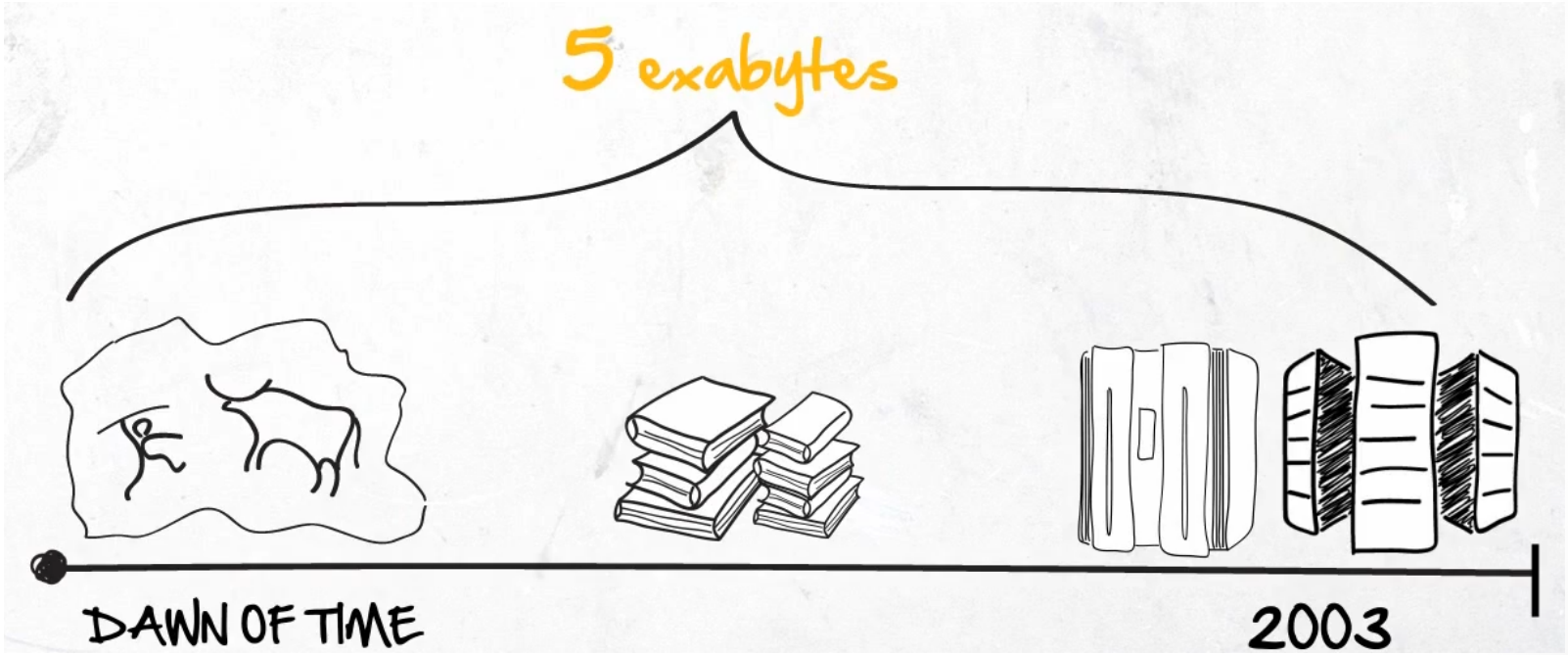
Big Data and the Next Wave of InfraStress Problems, Solutions, Opportunities

Abstract:

Data storage is growing at a higher rate than ever before, and coupled with rapidly increasing demand for instant access, will cause great stress on both the physical and the human infrastructure of computing. System planners and administrators will soon face the interesting challenge of dealing with network and backup issues when office systems hold 100s of GB of disks, and larger servers reach 10s and 100s of TB and even PB. There will also be great opportunities in both research and commercial applications, but the problems must be understood, and solutions anticipated. This talk will give some examples, including some large customer problems that Silicon Graphics has been working on; and examine technology trends in storage capacities, access times, computer architectures, and bandwidths, to see what these portend over the next few years.

John R. Mashey, Silicon Graphics/Cray Research

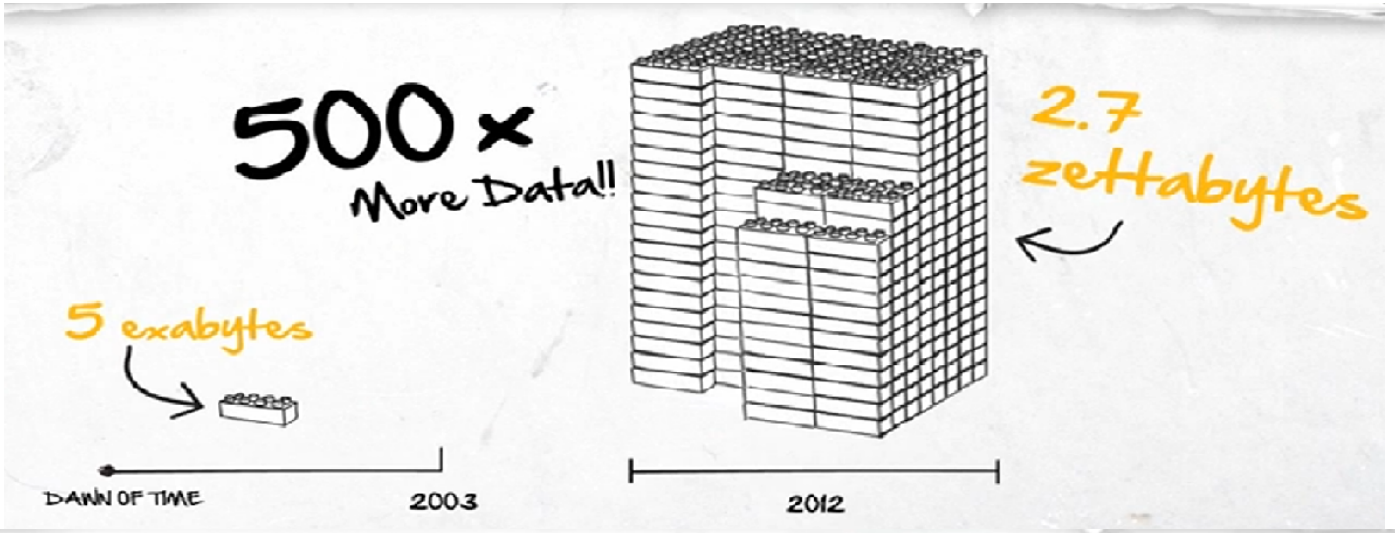
Explosión de los datos



1 EB = 10^3 PB = 10^6 TB = 10^9 GB = 10^{12} MB = 10^{15} KB = 10^{18} bytes.

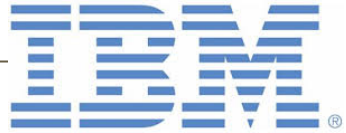
Explosión de los datos

Zettabyes: 10^{21} bytes.



Explosión de los datos

En 2014 el 90% de la información mundial había sido creada en 2012/13 #BigData

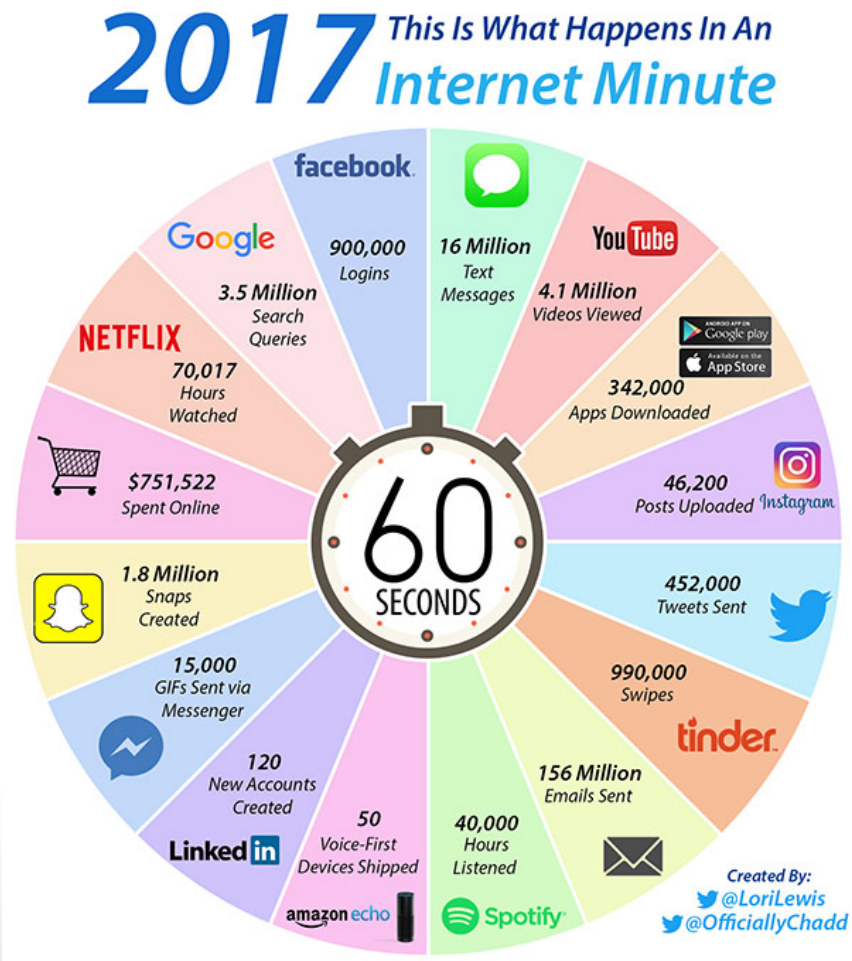


"Generamos **más información en dos días que en toda nuestra historia** hasta antes del 2003"

Erik Schmidt, Director ejecutivo Google

El **consumidor** de hoy navega a diario en Internet **dejando un claro rastro** sobre:

- **Quién** es
- **Qué** le interesa
- Con **quién** se relaciona
- **Dónde** compra
- Y **Cuando**

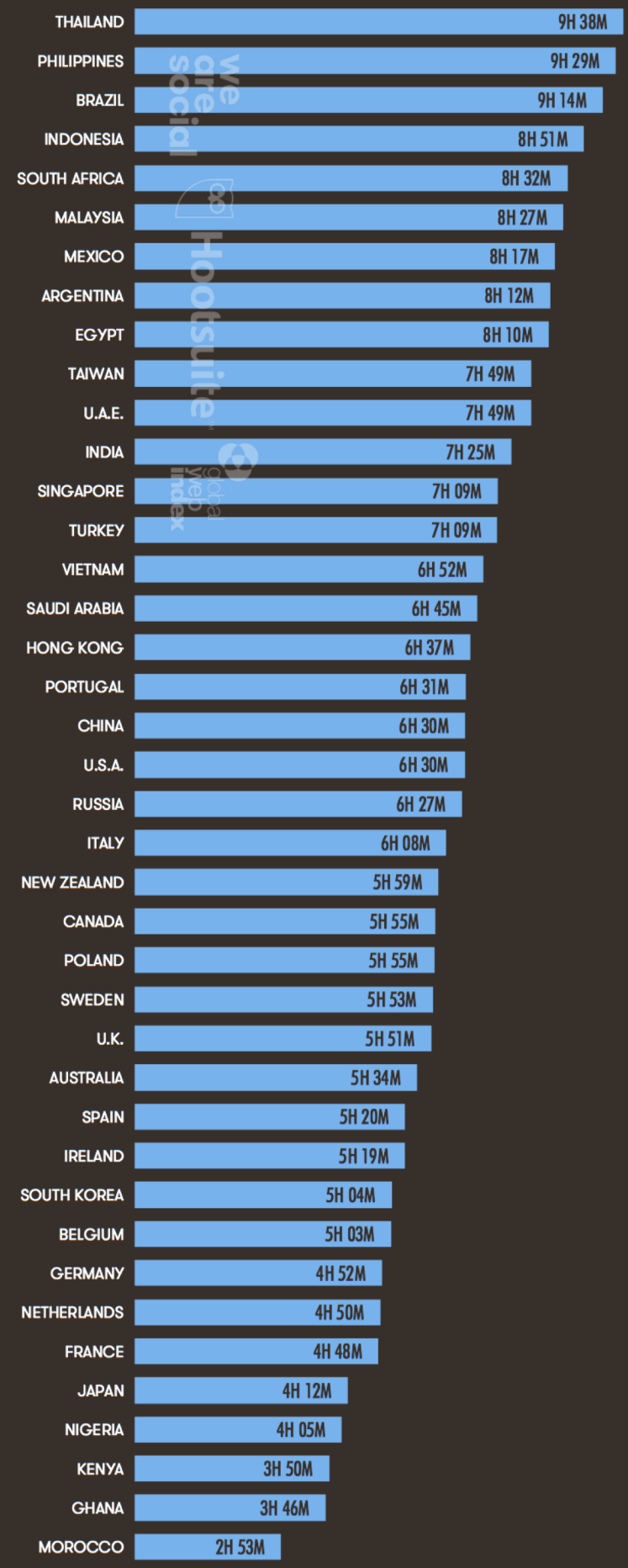

2017 This Is What Happens In An Internet Minute
facebook
900,000 Logins
16 Million Text Messages
YouTube
4.1 Million Videos Viewed
342,000 Apps Downloaded
46,200 Posts Uploaded
Instagram
452,000 Tweets Sent
990,000 Swipes
tinder
156 Million Emails Sent
40,000 Hours Listened
Spotify
50 Voice-First Devices Shipped
amazon echo
120 New Accounts Created
LinkedIn
15,000 GIFs Sent via Messenger
1.8 Million Snaps Created
$751,522 Spent Online
NETFLIX
70,017 Hours Watched
3.5 Million Search Queries
Google
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd


2018 This Is What Happens In An Internet Minute
facebook
973,000 Logins
18 Million Text Messages
YouTube
4.3 Million Videos Viewed
375,000 Apps Downloaded
174,000 Scrolling Instagram
481,000 Tweets Sent
1.1 Million Swipes
tinder
187 Million Emails Sent
936,073 Views
twitch
67 Voice-First Devices Shipped
amazon echo
38 Million Messages
25,000 GIFs Sent via Messenger
2.4 Million Snaps Created
$862,823 Spent Online
NETFLIX
266,000 Hours Watched
3.7 Million Search Queries
Google
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd

TIME SPENT PER DAY ON THE INTERNET

AVERAGE NUMBER OF HOURS SPENT USING THE INTERNET PER DAY VIA ANY DEVICE [SURVEY BASED]

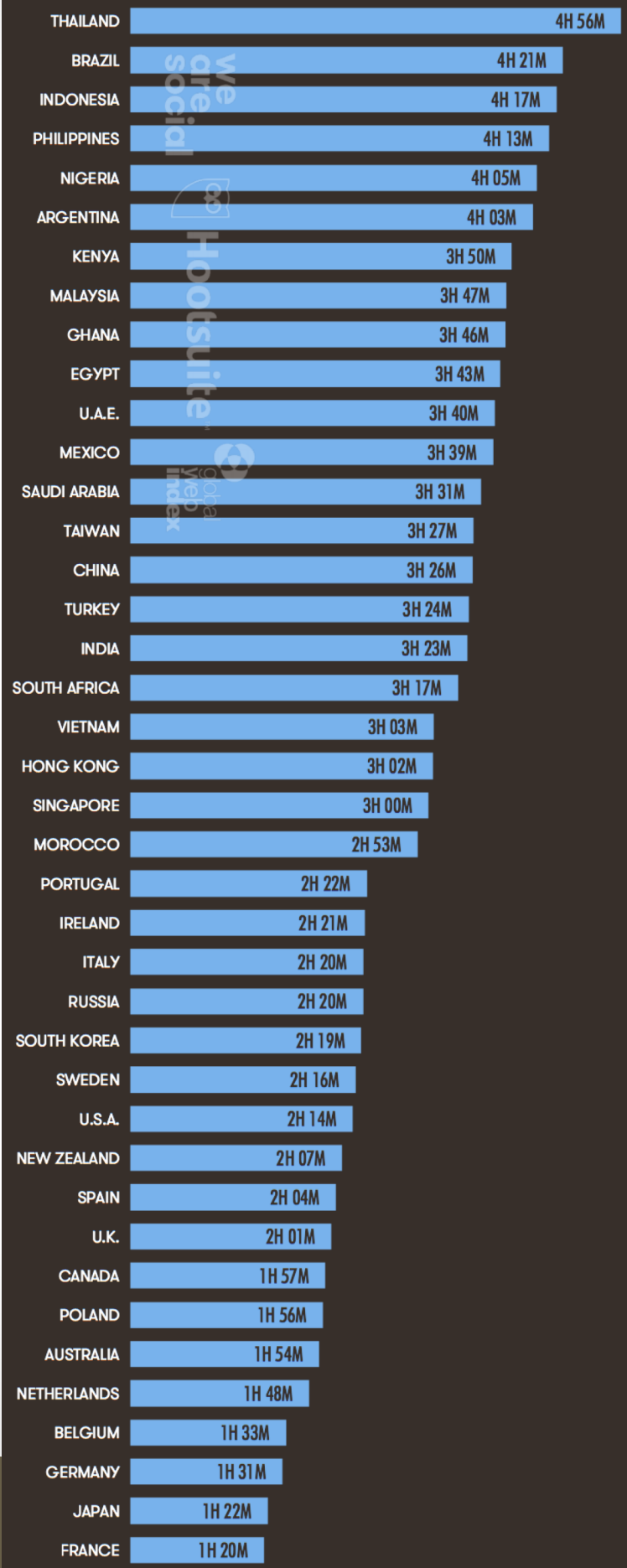
JAN 2018



JAN 2018

TIME SPENT PER DAY USING MOBILE INTERNET

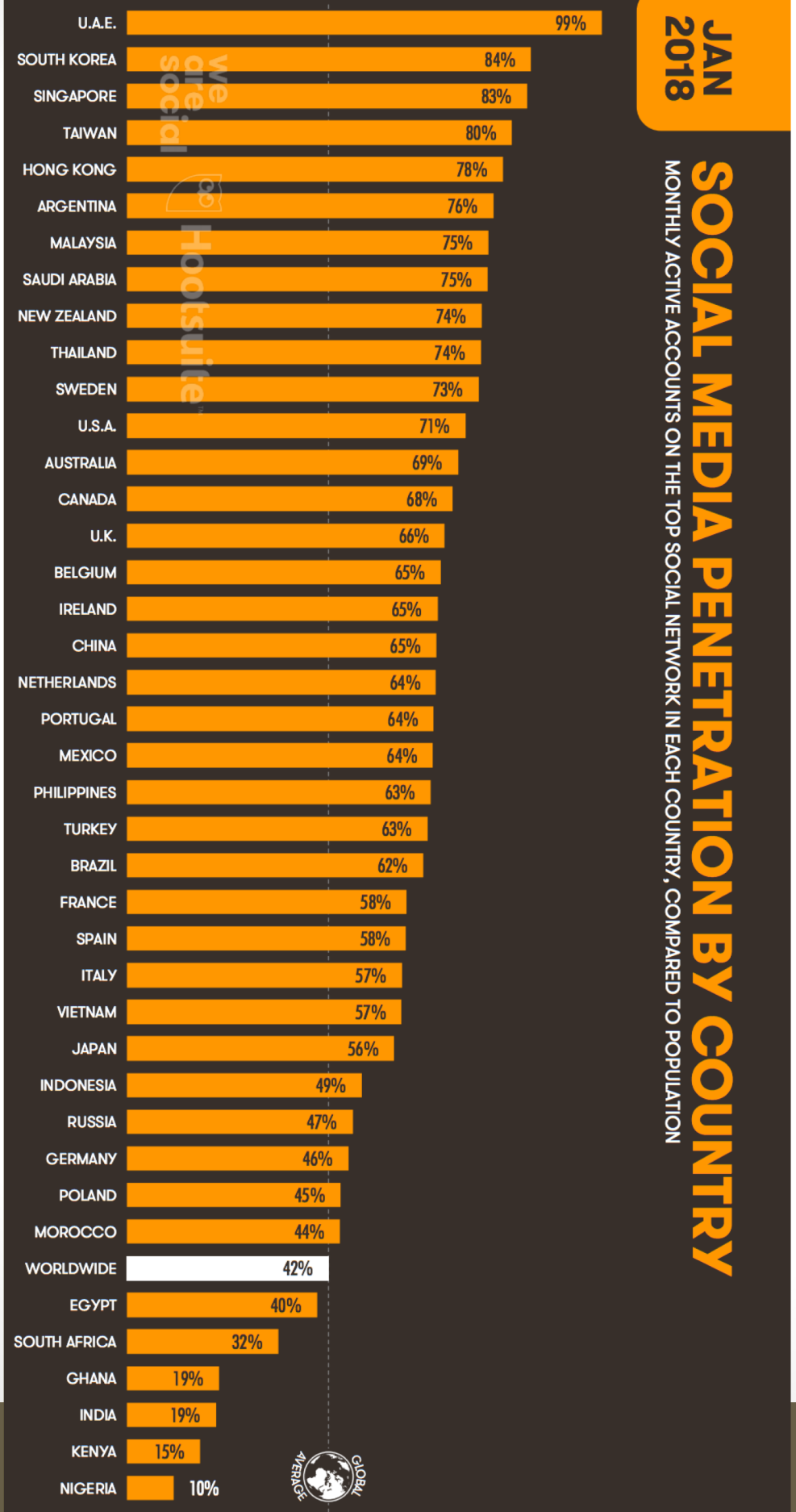
AVERAGE NUMBER OF HOURS PER DAY SPENT ACCESSING THE INTERNET VIA A MOBILE PHONE [SURVEY BASED]



JAN 2018

SOCIAL MEDIA PENETRATION BY COUNTRY

MONTHLY ACTIVE ACCOUNTS ON THE TOP SOCIAL NETWORK IN EACH COUNTRY, COMPARED TO POPULATION



¿Qué es big data?

"La capacidad de la sociedad para asimilar la información mediante vías novedosas con el objetivo de producir conocimientos, bienes y servicios de *valor significativo*".

Mayer-Schonberger y Cukier

"Almacenamiento y gestión de una cantidad elevada de datos".

RAE

Big data

Big data es un término que describe el gran volumen de datos – estructurados y no estructurados – que inundan una empresa todos los días. Pero no es la cantidad de datos lo importante. Lo que importa es lo que las organizaciones hacen con los datos.

El big data puede ser analizado para obtener *insights* que conlleven a mejores decisiones y acciones de negocios estratégicas.

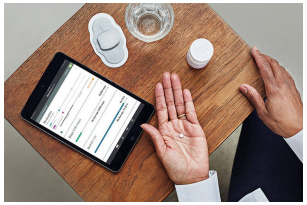
Big data: Características

Ventajas

- Permite obtener respuestas más complejas, ya que dispone de mayor cantidad de información.
- Determinar las causas de origen de fallos, problemas y defectos casi en tiempo real.
- Detectar conductas fraudulentas antes de que afecte a su organización.
- Proporciona ventajas competitivas sobre la competencia.

Aplicaciones

proteus®
DIGITAL HEALTH



Walmart

TARGET

"El big data permitirá, entre otras cosas, hacer recomendaciones cada vez más precisas, basadas en el análisis de nuestro rastro y en la comparación con el de otros usuarios" (Albert Bifet).

Recomendaciones

"Si yo se sobre algo, tengo poderes sobre ti.
El poder está en el código, si no lo entiendes se aprovecharán de ti"
(Marc Goodman)

"Todavía no hemos llegado tan lejos. Pero Si que existe el riesgo de que pensemos que los datos nunca mienten y que sustituyamos las decisiones humanas por predicciones deterministas.