Big Data

Diplomatura de Vinculación en Introducción a Ciencia de Datos

Edición 2023

Ing. Román Bond – Mg. Diego Encinas

Agenda-Clase 1

¿Qué es Big Data?

- Marea de información digital
- ¿Big Data o no Big Data?
- Definición y dimensiones en Big Data
- Datos
- Desafíos

Tecnologías de Big Data

- Tecnologías
- Pila

Casos de uso

Herramientas

¿Qué es Big Data?

"El Big data es como el sexo en la adolescencia: todo el mundo habla de ello, nadie sabe realmente cómo hacerlo, todos piensan que los demás lo están haciendo, así que todos dicen que también lo hacen..."

- Dan Ariely - Professor of Psychology and Behavioral Economics at Duke University

¿Qué es Big Data?

Big Data no es fácil de definir, es un término que fue “inventado por el marketing” y que involucra múltiples tecnologías.

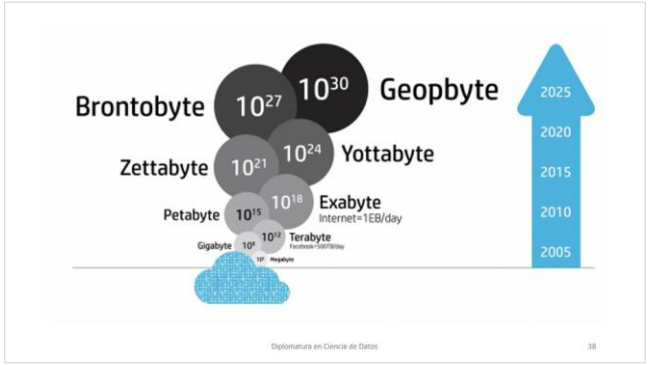
Muy utilizado en las **redes sociales** por los departamentos de marketing.

Con el auge de internet surgió un continuo crecimiento de las redes sociales, los sitios de "archivos multimediales" y los sitios de e-comercio.

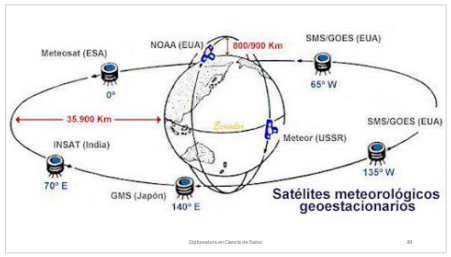
El avance tecnológico permitió generar y capturar datos de sensores de tiempo real, lo que involucró un **crecimiento exponencial** del **volumen de datos**.

Marea de información digital

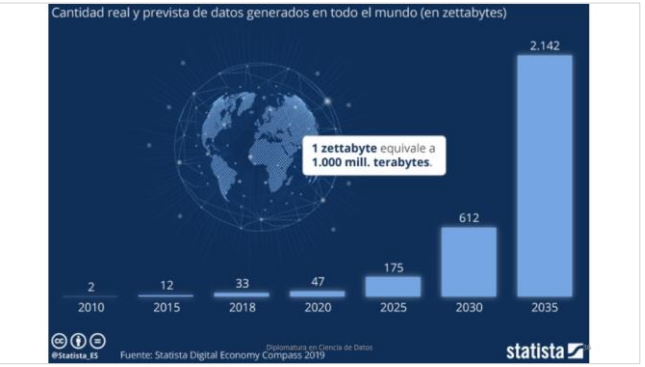
- En 2015 el universo digital estaba compuesto por 6 ZB de datos
  - 1 Zettabyte =1000 Hexabyte
  - 1 Hexabyte = 1000 Petabyte
  - 1 Petabyte = 1000 Terabyte

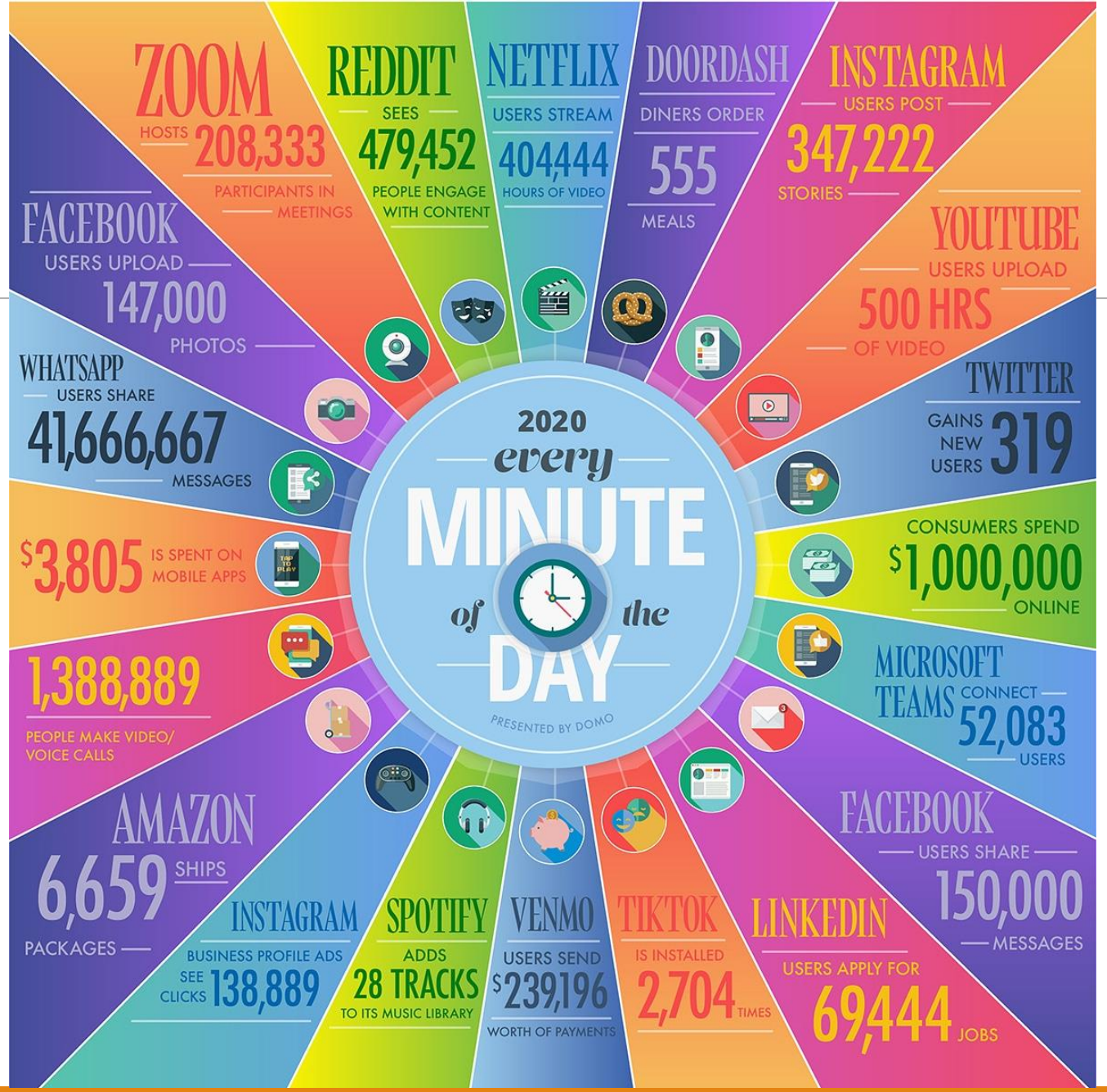


- 6 ZB en discos de 10TB ⟹ 644.245.094 discos
  - Peso: 322.122 toneladas (3 portaaviones)
  - Altura: 16.106 Km



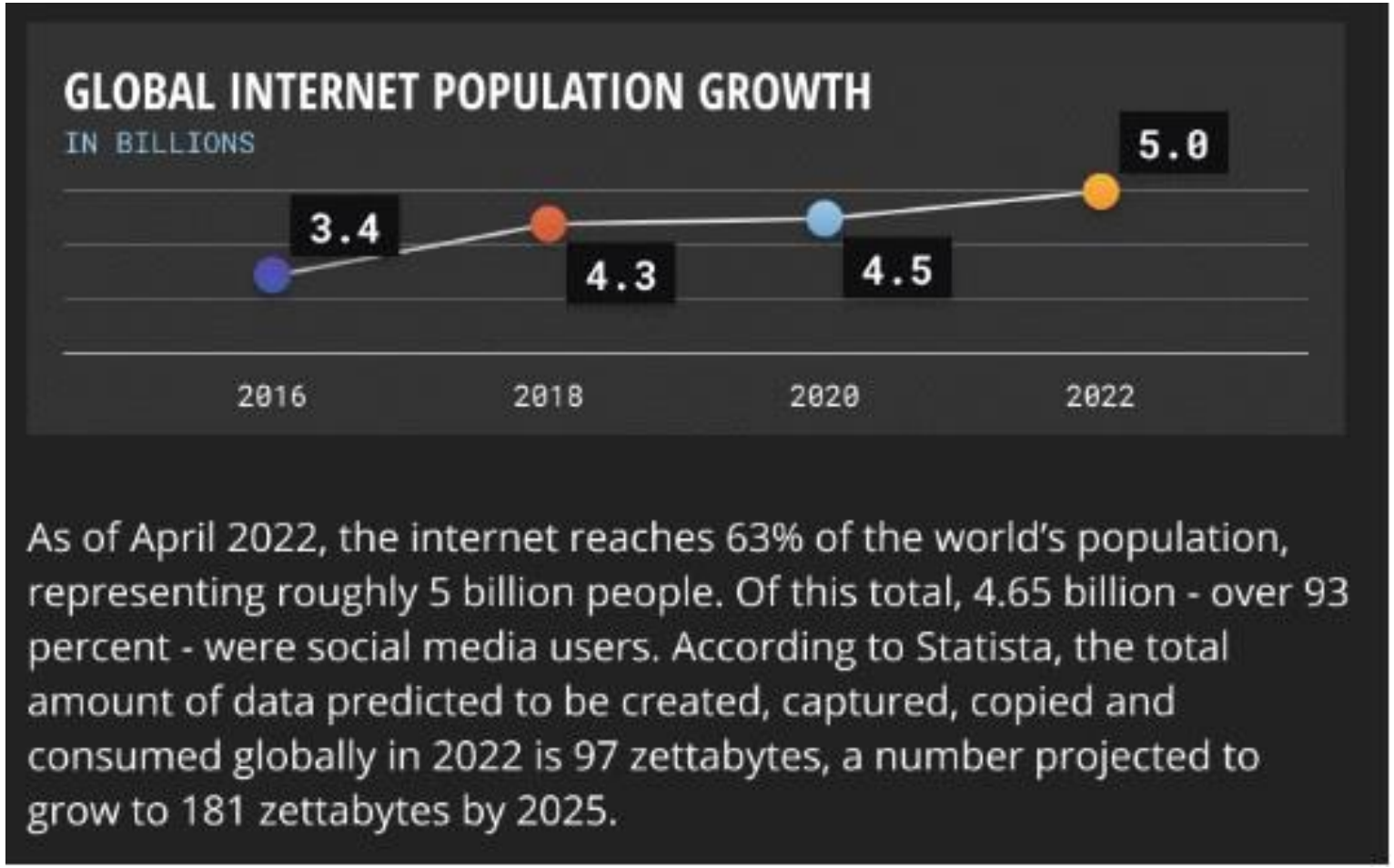
Velocidad



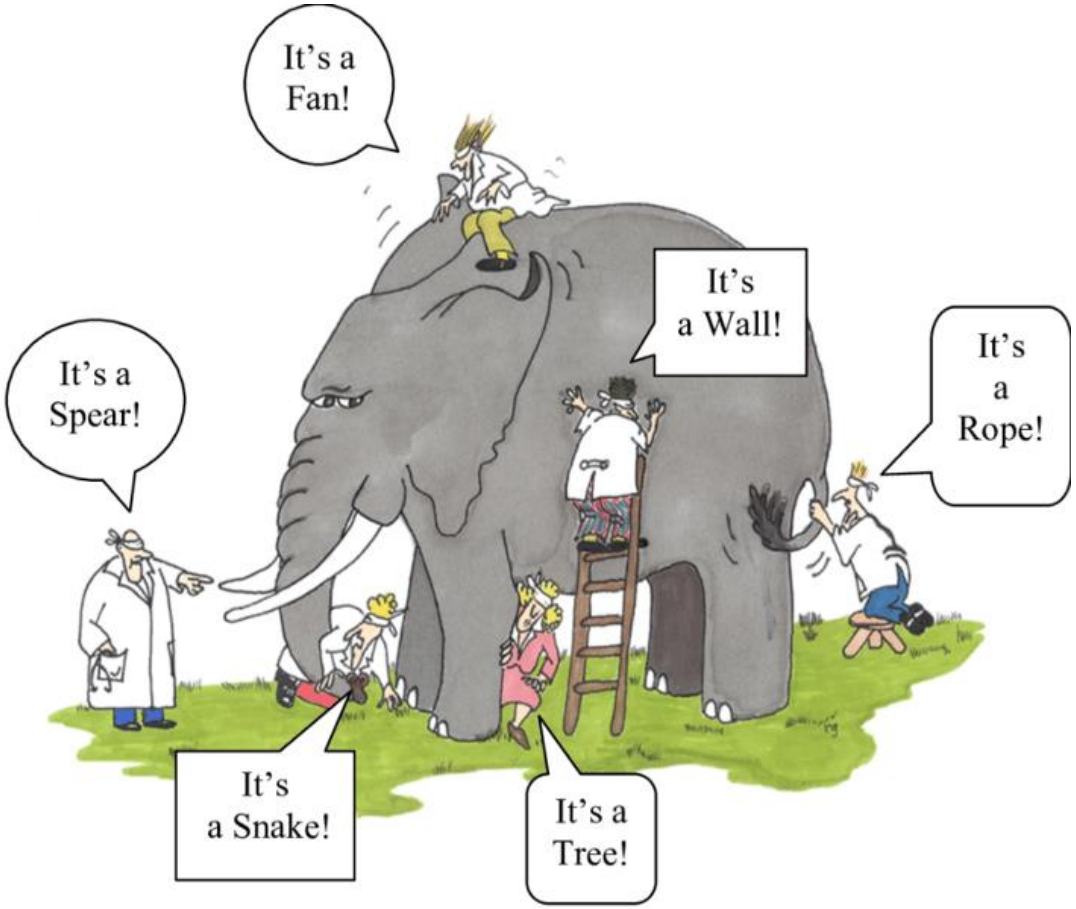

2020
every
MINUTE
of the
DAY
PRESENTED BY DOMO
ZOOM
HOSTS 208,333
PARTICIPANTS IN MEETINGS
REDDIT
SEES 479,452
PEOPLE ENGAGE WITH CONTENT
NETFLIX
USERS STREAM 404,444
HOURS OF VIDEO
DOORDASH
DINERS ORDER 555
MEALS
INSTAGRAM
USERS POST 347,222
STORIES
YOUTUBE
USERS UPLOAD 500 HRS
OF VIDEO
FACEBOOK
USERS UPLOAD 147,000
PHOTOS
WHATSAPP
USERS SHARE 41,666,667
MESSAGES
TWITTER
GAINS NEW USERS 319
$3,805 IS SPENT ON MOBILE APPS
CONSUMERS SPEND $1,000,000 ONLINE
1,388,889
PEOPLE MAKE VIDEO/ VOICE CALLS
MICROSOFT TEAMS CONNECT 52,083
USERS
AMAZON
6,659 SHIPS
PACKAGES
FACEBOOK
USERS SHARE 150,000
MESSAGES
INSTAGRAM
BUSINESS PROFILE ADS SEE CLICKS 138,889
SPOTIFY
ADDS 28 TRACKS
TO ITS MUSIC LIBRARY
VENMO
USERS SEND $239,196
WORTH OF PAYMENTS
TIKTOK
IS INSTALLED 2,704 TIMES
LINKEDIN
USERS APPLY FOR 69,444 JOBS

Variedad

Marea de información digital

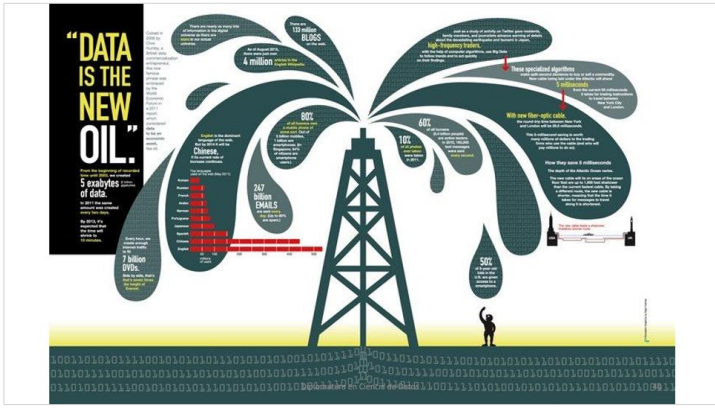


Marea de información digital



¿Big Data o no Big Data?

- No es fácil determinar el límite entre un problema de Big Data del que no lo es.
- Depende de los datos, fuentes, tipo, recolección, etc.
- Depende del procesamiento, almacenamiento, consultas
- Depende del costo



¿Big Data o no Big Data?

¿Cloud or not Cloud?

Amazon AWS

- 1 cluster de 4 instancias
  - 2 vCPU con 4 GB de RAM y 1 TB de almacenamiento 478,51 USD x mes

Amazon.com

- 4 CPU con 4 GB de RAM y 1 TB de almacenamiento [501,11(CPU) + 45,99 (disco)] 547 USD **c/u**

¿Big Data o no Big Data?

¿Cloud or not Cloud? ¿Too big or not too big?

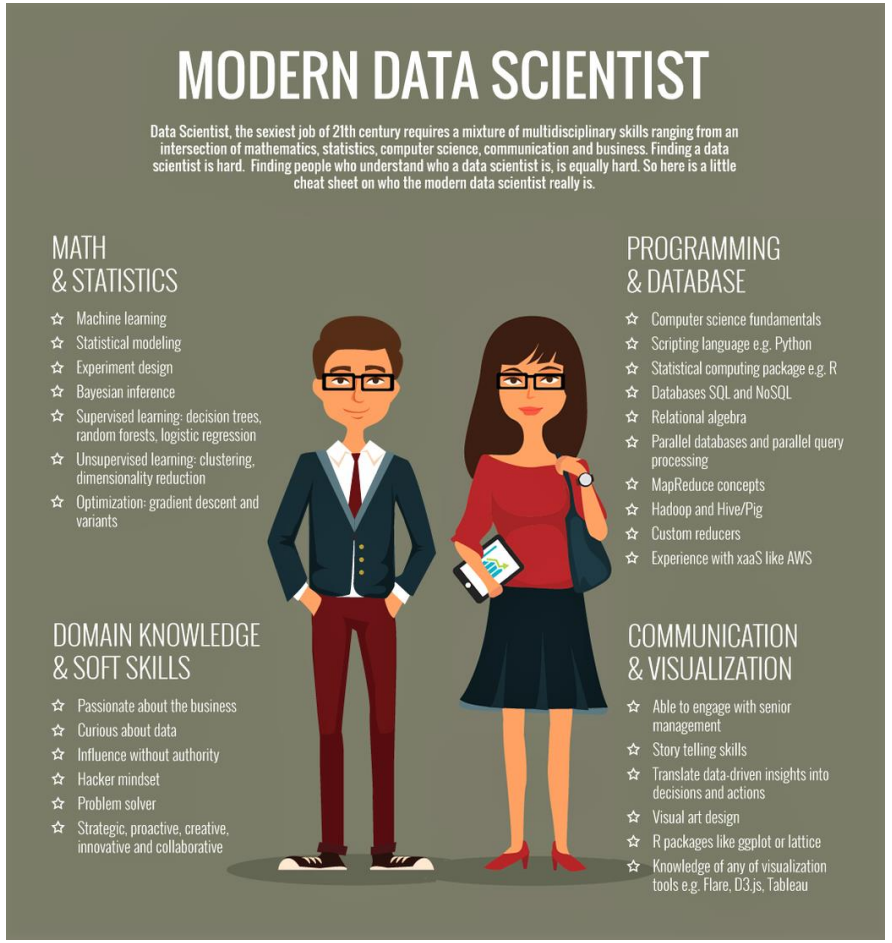
Cloud

- 478.51 x 12 meses = U$D 5742.12

No Cloud

- (501.11+45.99) * 4 = U$D 2188.4

¿Qué es Big Data?



- Arquitecturas en la Nube
- Procesamiento para Grandes Datos

Big Data - Definición

Según IDC (International Data Corporation) www.idc.com

Big data representa una nueva generación de tecnologías y arquitecturas, diseñadas para extraer **valor** económicamente de **volúmenes** muy grandes de una amplia **variedad** de datos, al permitir la captura, el descubrimiento y/o análisis de alta **velocidad**.

Las tres ‘V’ de Big Data

- **V**olumen: el universo digital sigue expandiendo sus fronteras.
- **V**elocidad: la velocidad a la que generamos datos es muy elevada, y la proliferación de sensores es un buen ejemplo de ello. Además, los datos en tráfico –datos de vida efímera, pero con un alto valor para el negocio crecen más deprisa que el resto del universo digital.
- **V**ariedad: los datos no solo crecen sino que también cambian su patrón de crecimiento, a la vez que aumenta el contenido desestructurado

La cuarta ‘V’ de Big Data

- **V**alor: Extraer valor de toda esta información marcará el futuro del manejo de información.

- El valor lo podremos encontrar en diferentes formas:
  - mejoras en el rendimiento del negocio
  - segmentación de clientes
  - tomas de decisiones
  - automatización de decisiones tácticas
  - etc.

Datos

- Datos estructurados
  - Bases de datos relacionales
- Datos no estructurados
  - Texto escrito en lenguaje natural
  - Contenido multimedia, imágenes, fotos, audio y video
- Datos semiestructurados
  - Archivos de texto plano, planillas de cálculo

Datos estructurados

- Generados por humanos
  - Ingreso de datos
  - Actividad web (sites, pages, clicks)
  - Datos generados por juegos
- Generados por computadoras
  - Sensores
  - Logs de aplicaciones o servidores
  - Productos con códigos de barra
  - Operaciones bancarias

Datos no estructurados

- Generados por humanos
  - Informes, reportes
  - Redes sociales
- Generados por computadoras
  - Imágenes satelitales
  - Monitoreo (sísmicos, atmosféricos)
  - Fotografía
  - Video
  - Radares

Datos

Estructurados

No estructurados

DBMS

- Relacionales
  - MySQL
  - PostgreSQL
  - Derby
- No relacionales noSQL (Not only SQL)
  - MongoDB

DBMS no relacional

- Clave/valor
  - No requieren un esquema
  - No tipadas (por lo general todo se almacena como string)
  - Ofrecen el manejo de colecciones de clave/valor
  - Ej: Riak
- Documentos
  - La estructura de los documentos se almacena en formato JSON
  - Útiles cuando se generan muchos reportes
  - Ej: MongoDB, CouchDB

DBMS no relacional

- Orientadas a columnas
  - Permite el agregado simple de columnas, estas se pueden ir llenando fila a fila
  - Es modelado usando BigTable de Google
    - Cada elemento se indexa con una fila, una columna y un timestamp
  - Ej: Hbase
- Orientadas a grafos
  - Su elemento básico es el nodo-relación
  - Se navega de nodo a nodo siguiendo las relaciones
  - Orientado a problemas con naturaleza de grafos
  - Ej: Neo4J

¿Tiempo real o no tiempo real?

- Problemas de tiempo real
  - Detección de fraudes
  - Detección de fallas
  - Determinar eventos en redes sociales para detectar alertas tempranas
  - Publicidad web
- Problemas de no tiempo real (batch)
  - Segmentación de clientes
  - Tomas de decisiones (semanales, mensuales, anuales)

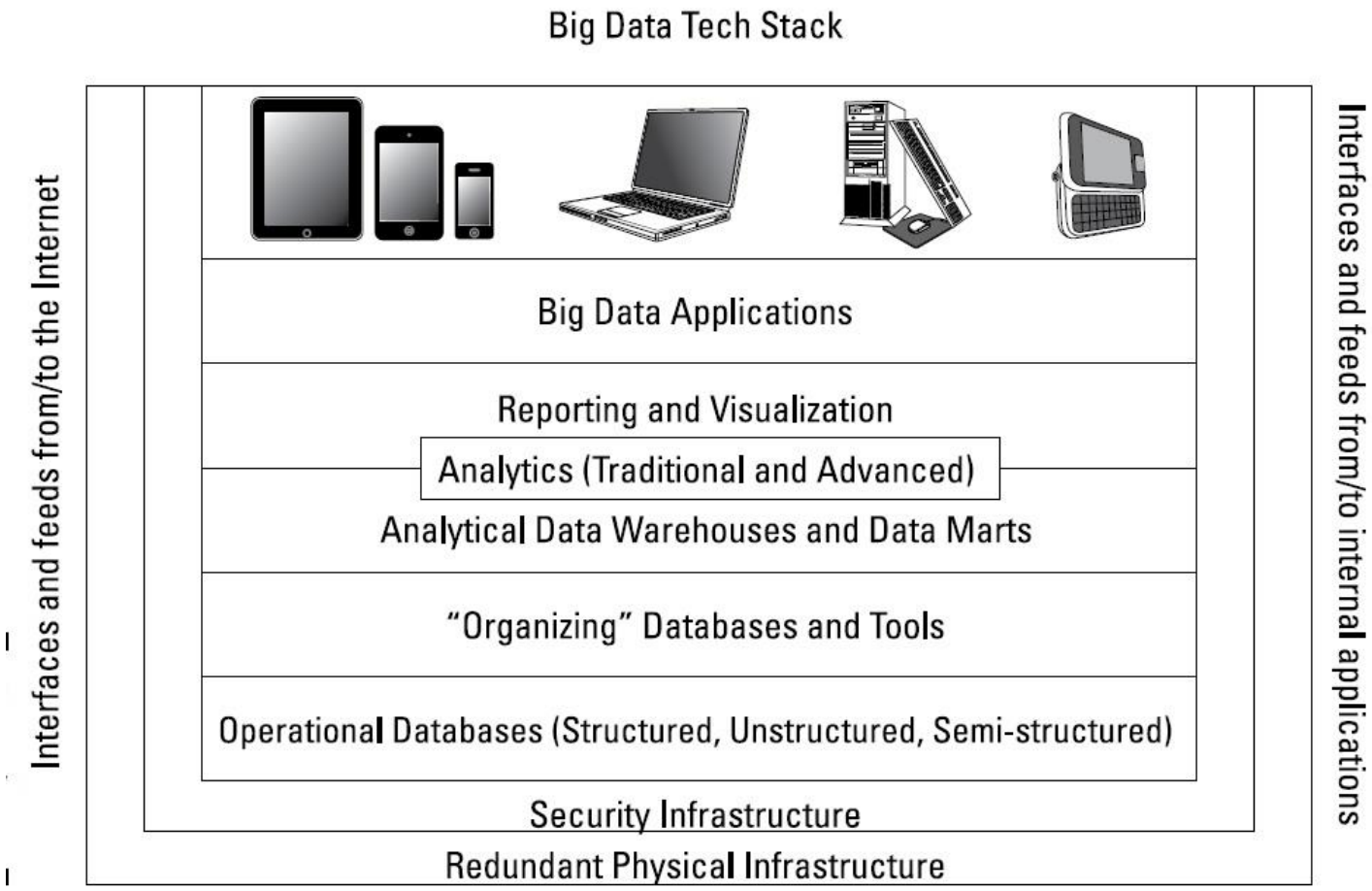
Big Data - Desafíos

- Almacenamiento
- Procesamiento (debe ser rápido y efectivo)
- Diversidad de los datos (estructurados, no estructurados, semiestructurados)

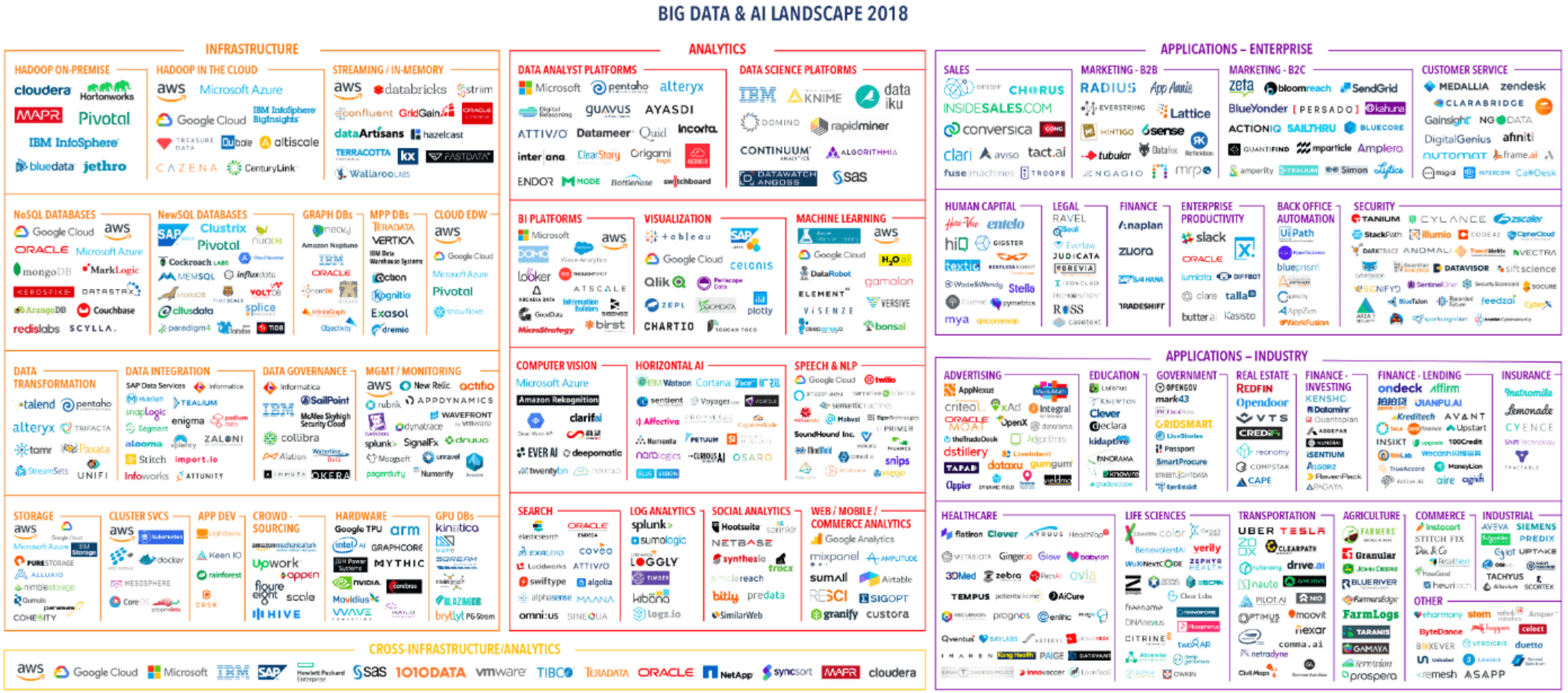
Tecnologías

- Big Data no es una tecnología, es la combinación de varias tecnologías para hacer más fácil el tratamiento de los datos con los que contamos hoy en día.
- Para la ejecución de aplicaciones de Big Data es necesario contar con hardware y software específico
- Clusters, sistemas distribuidos, etc.
- Cloud computing

Tecnologías - Pila



Tecnologías - Pila



Casos de uso

- Segmentación de clientes
  - Marketing
  - Ventas
  - Churn de clientes
- ¿Quién lo hace?
  - Empresas de comunicación
  - Hipermercados
  - Aseguradoras
- Campañas electorales

Casos de uso

- Optimizando procesos de negocio
  - Manejo de stock
  - Manejo de recursos humanos
  - Optimización de rutas de reparto
- ¿Quién lo hace?
  - Cadena de puntos de venta
  - Correo

Casos de uso

- Optimización de rendimiento personal
  - Consumo de calorías
  - Nivel de condición física
  - Patrones de sueño
- ¿Quién lo hace?
  - Google Fit
  - Apple Swatch
  - Jawbone (recolecta 60 años de sueño en una sola noche)

Casos de uso

- Salud
  - Codificación de material genético
  - Dietas y alimentos adecuados
  - Descubrir la activación de genes
- ¿Quién lo hace?
  - Laboratorios
  - Farmacias
  - Hospitales

Casos de uso

- Rendimiento deportivo
  - Patrones de juego
  - Análisis del juego.
  - Imágenes y sensores
- ¿Quién lo hace?
  - SlamTracker (Tenis)
  - NBA
  - Beisbol

Casos de uso

- Seguridad
  - Fraudes
  - Cyber-ataques
  - Perfil criminal.
- Optimización de ciudades
  - Tráfico
  - Optimización de suministro (electricidad)

Herramientas

- Hadoop MapReduce
- Spark
- Gridgane
- HPCC
- Storm
- Hana
- Hive
- Kafka
- Flume

Preguntas? O …