

Conceptos y Aplicaciones de Big Data

CONCEPTOS DE BIG DATA

PROF. WALDO HASPERUÉ
WHASPERUE@LIDI.INFO.UNLP.EDU.AR

Contenidos de la materia

Fundamentos y conceptos de Big Data

Frameworks para soluciones en Big Data

- MapReduce
- Spark

Temario de la clase

¿Qué es Big Data?

- Definición y dimensiones en Big Data.

Herramientas y tecnologías de Big Data

Casos de uso

¿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.

¿Qué es Big Data?

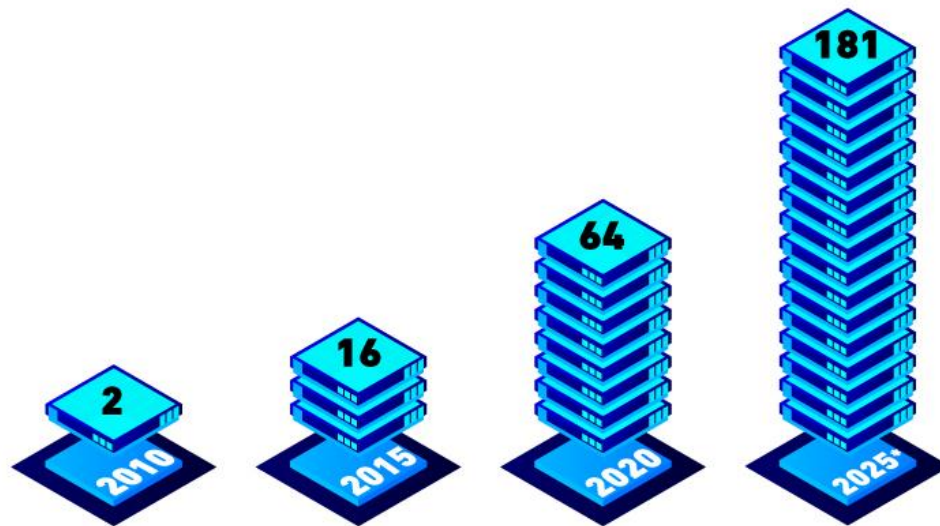
Con el auge de internet surgió un continuo crecimiento de las redes sociales, los sitios de "archivos multimediales" y los sitios de e-comercio

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 16 ZB de datos.

En 2025 se prevee un volumen de 181 ZB.



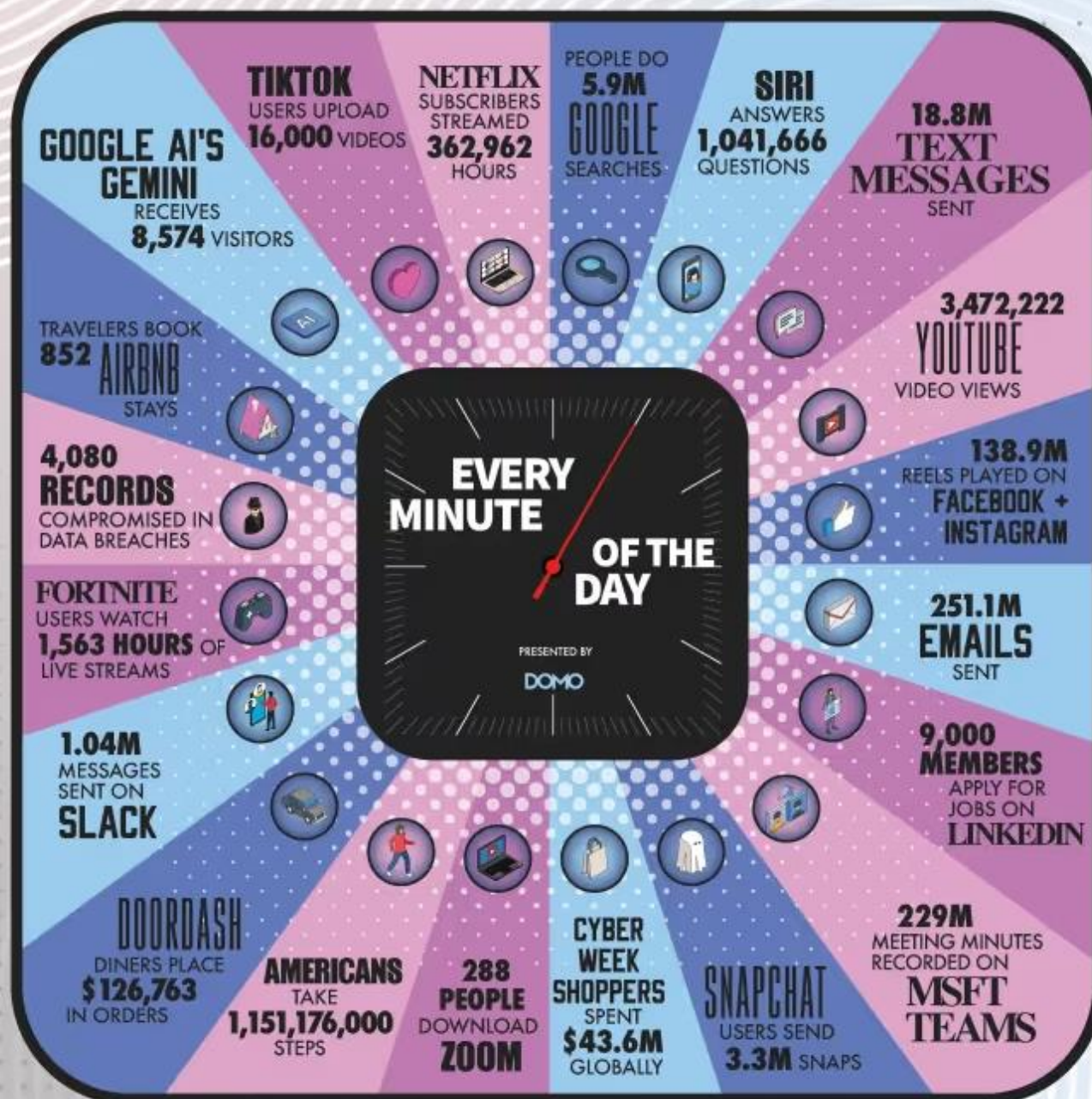
Marea de información digital

- 1 Zettabyte = 1000 Hexabyte
- 1 Hexabyte = 1000 Petabyte
- 1 Petabyte = 1000 Terabyte

181 ZB en discos de 10TB ➔ +18.000.000.000 discos

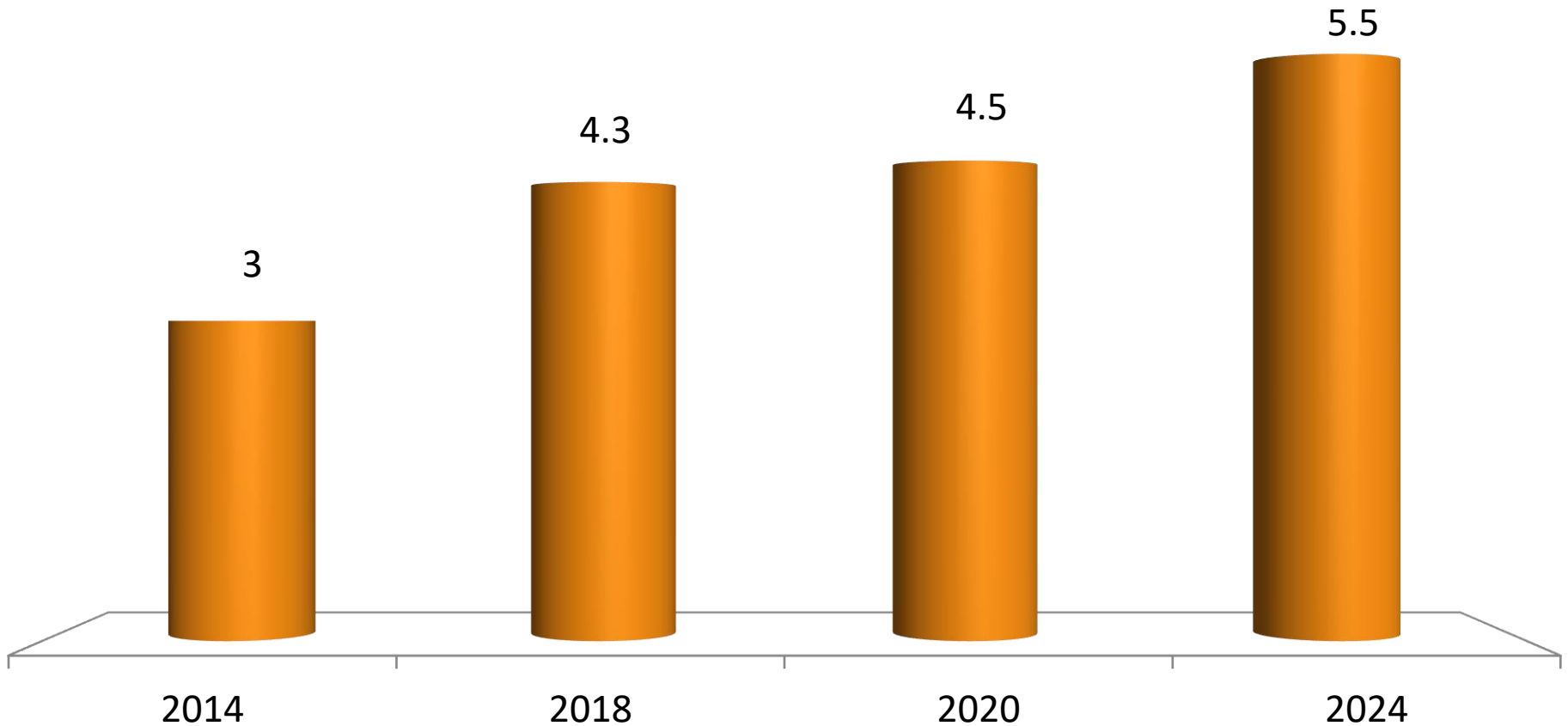
- Peso: +13.500.000 toneladas (\approx 135 portaaviones)
- Altura: +450.000 Km (\approx 35.5 planetas Tierra)
(\approx 3.2 planetas Júpiter)

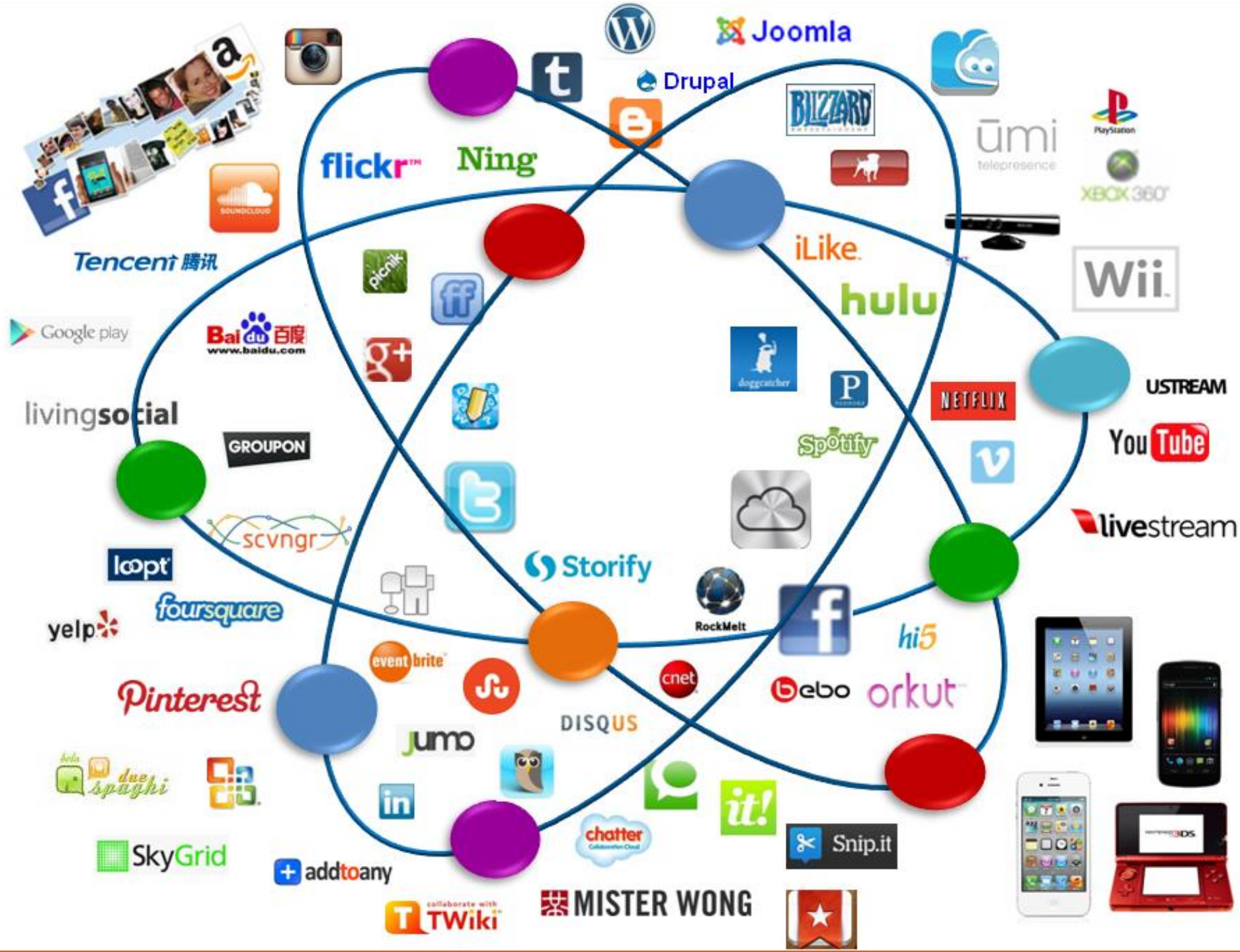


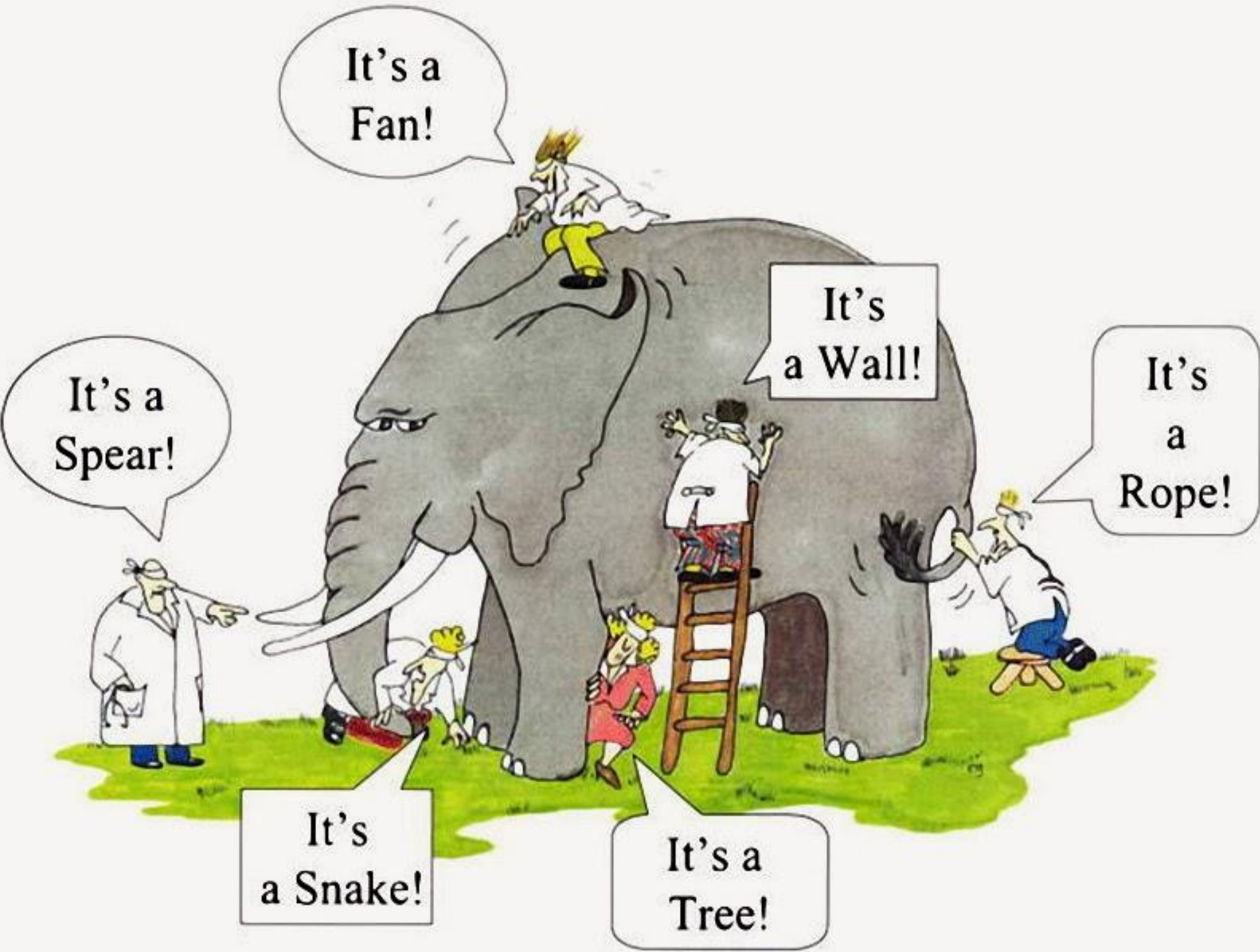


Marea de información digital

Miles de millones de usuarios de internet







¿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?

Una fuente de datos de 4TB almacenada en un disco con velocidad de transferencia de 500 MB/s

Cualquier proceso tardaría más de dos horas en procesar toda esa información

4 TB → 4194304 MB

500 MB en 1 segundo → 4194304 MB en 8388 segundos

8388 segundos \cong 2.3 horas

¿Escala? Proyección a futuro

Costos - Cloud

Amazon AWS

- 1 cluster de 4 instancias
 - 2 vCPU con 4 GB de RAM y 1 TB de almacenamiento

Amazon EC2 estimate

Amazon EC2 Instance Savings Plans instances (monthly)	68.91 USD
Amazon Elastic Block Storage (EBS) pricing (monthly)	409.60 USD
Total monthly cost:	478.51 USD

Costos – No cloud

Amazon.com

- 4 CPU con 4 GB de RAM y 1 TB de almacenamiento



Acer - Aspire - Ordenador de escritorio, procesador Intel Core i5-9400 de novena generación, USB 3.1 tipo C, sistema operativo Windows 10 Home, Negro

★★★★☆ ~ 798

Computadoras personales

US\$ **501¹¹** ~~US\$549.99~~

Con envíos a Argentina



Patrocinado ⓘ

Disco duro externo Seagate Backup Plus, Negro 1TB

★★★★☆ ~ 39,864

Computadoras personales

US\$ **45⁹⁹**

Con envíos a Argentina

Cloud or not cloud?

Too big or not too big?

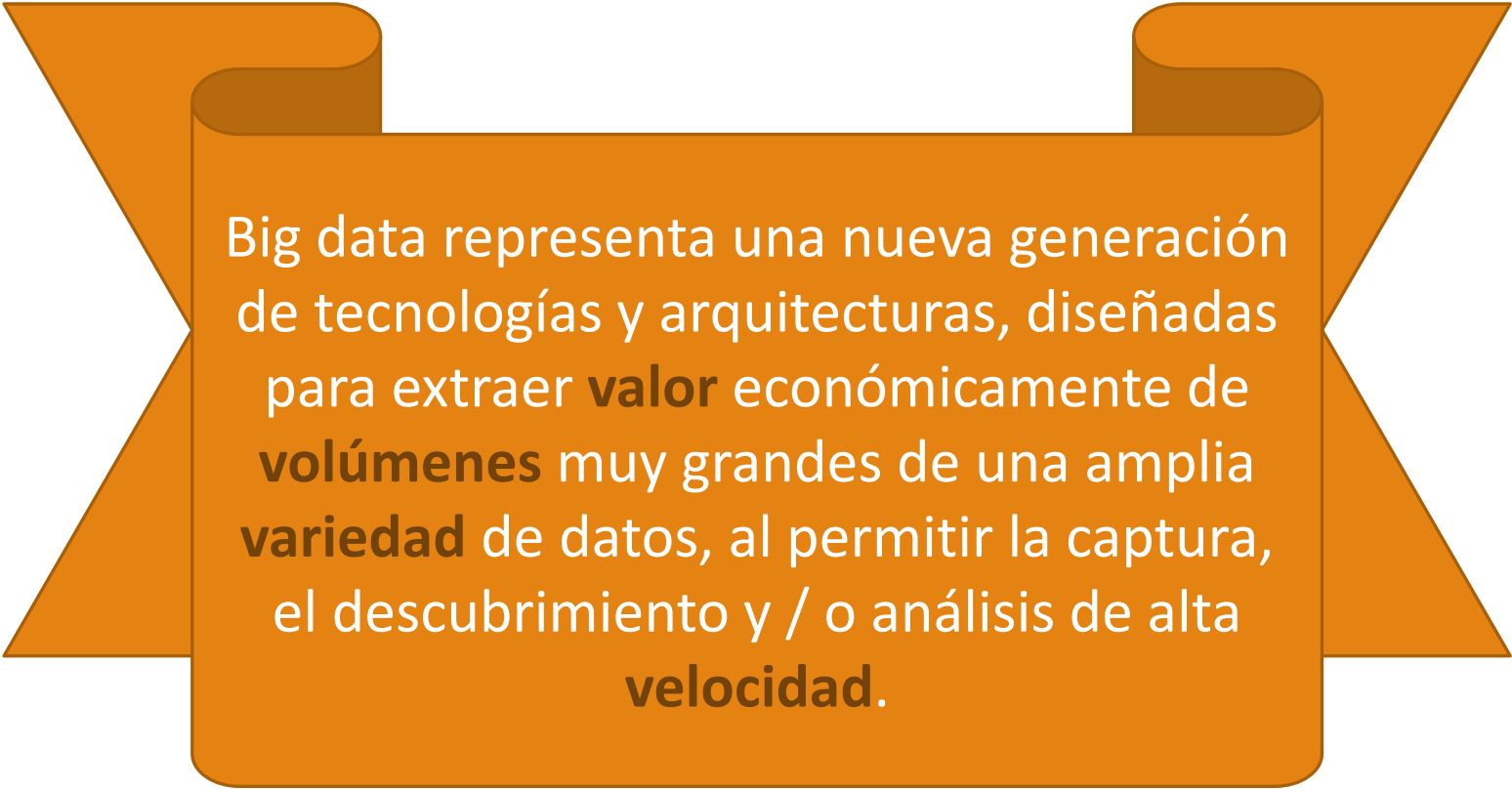
Cloud

- $478.51 \times 12 \text{ meses} = \text{U\$D } 5742.12$

No cloud

- $(501.11 + 45.99) * 4 = \text{U\$D } 2188.4$

Big Data – Definición de IDC



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

Volumen: el universo digital sigue expandiendo sus fronteras.

Velocidad: 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.

Variedad: 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

Valor: 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 semiestructurados

- Archivos de texto plano, planillas de cálculo

Datos no estructurados

- Texto escrito en lenguaje natural
- Contenido multimedia, imágenes, fotos, audio y video

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



DBMS

Relacionales

- MySQL
- PostgreSQL
- Derby

No relacionales noSQL (Not only SQL)

- MongoDB

DBMS no relacional

Clave/valor

- No requieren un esquema
- No son tipadas (por lo general todo se almacena como string)
- Ofrecen el manejo de colecciones de clave/valor
- Ej: Riak

DBMS no relacional

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

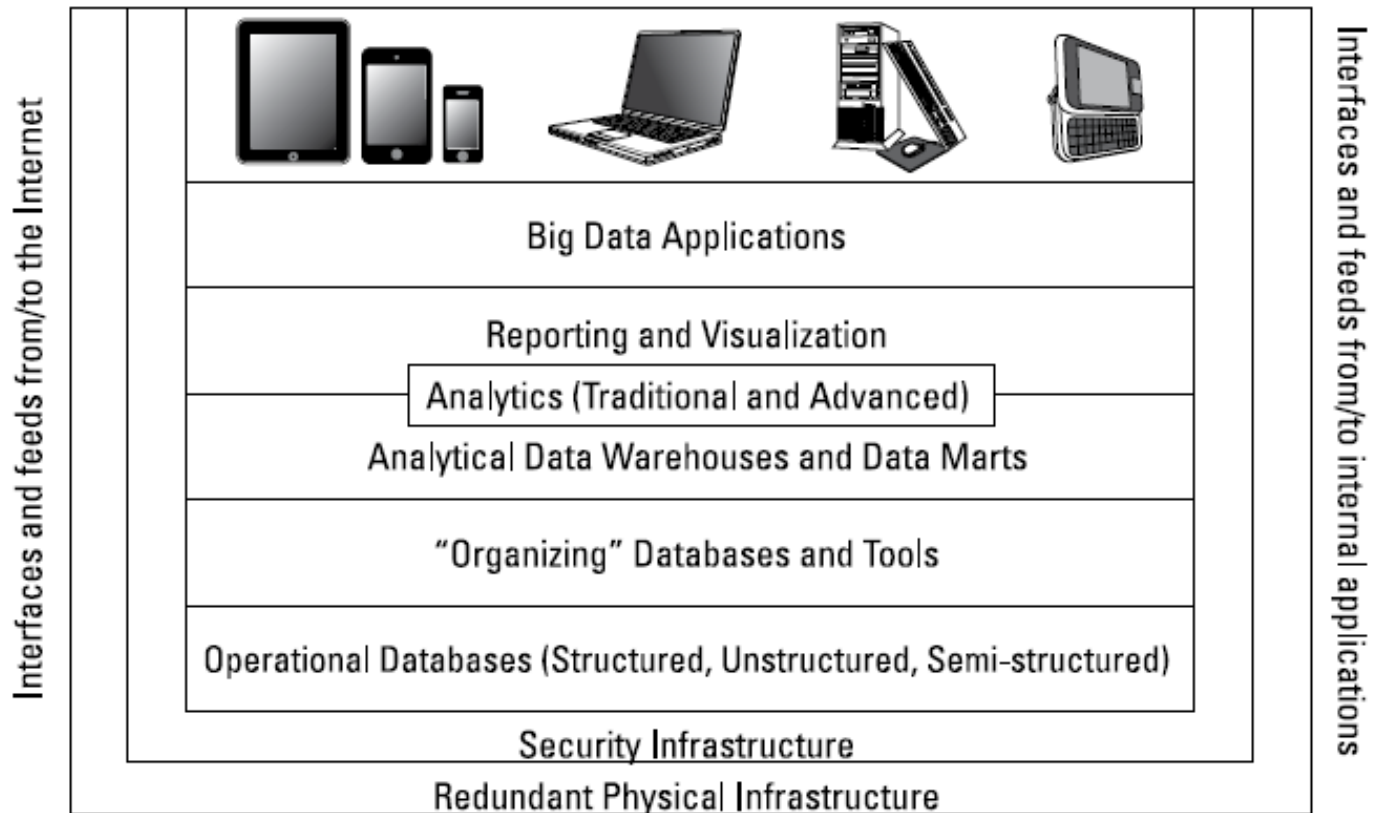
Big Data no es una tecnología, es la combinación de varias tecnologías para hacer más fácil el tratamiento de los datos con los que contamos hoy en día.

Para la ejecución de aplicaciones de Big Data es necesario contar con hardware y software específico.

- Clusters, sistemas distribuidos, etc.
- Cloud computing

Tecnologías

Big Data Tech Stack



Casos reales

Segmentación de clientes

- Marketing
- Ventas
- *Churn* de clientes

¿Quién lo hace?

- Empresas de comunicación
- Hipermercados
- Aseguradoras

Campañas electorales

Casos reales

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 reales

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 reales

Salud

- Codificación de material genético
- Dietas y alimentos adecuados
- Descubrir la activación de genes

¿Quién lo hace?

- Laboratorios
- Farmacias
- Hospitales

Casos reales

Rendimiento deportivo

- Patrones de juego
- Análisis del juego.
- Imágenes y sensores

¿Quién lo hace?

- SlamTracker (Tenis)
- NBA
- Beisbol

Casos reales

Seguridad

- Fraudes
- Cyber-ataques
- Perfil criminal.

Optimización de ciudades (Smart cities)

- Tráfico
- Optimización de suministro (electricidad)

Casos reales

Ciencia



Trading financiero



Auto autónomo



Casos reales

[NASA](#)

[NSA's Data Center](#)

[CERN's Hadron Collider](#)

[Facebook's Big Data](#)

[Big Data Analytics in Obama's Election Campaign](#)

[Internet live stats](#)

Herramientas

Hadoop MapReduce

Spark

Gridgane

HPCC

Storm

Hana

Hive

Kafka

Flume

Elasticsearch

MongoDB

Cassandra

Drill

Oozie

Databricks

Cloud Computing

Servers



Virtual
Desktop



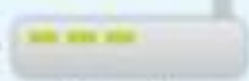
Software
Platform



Applications



Storage/
Data



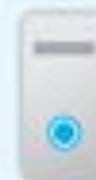
Router



Switch



End
User



Cloud computing



Google Cloud Platform



SOFTLAYER®



¿Quién usa Big Data?



¿Qué veremos?

Procesamiento batch

- Hadoop MapReduce
- Apache Spark

Procesamiento stream

- Spark streaming