

Nuevos paradigmas de datos - Big Data

¿Qué? ¿Cuándo? ¿Cómo?

Seminario ML 2018

Juan Galán Páez
Dpto. Ciencias de la Computación e Inteligencia Artificial
Datrik Intelligence

Junio - 2018

- Big Data y Ciencia de Datos
 - Introducción
 - Conceptos básicos
 - Soluciones Big Data
- Ecosistema Big Data
 - Introducción
 - Ecosistema Hadoop
 - MapReduce
- Spark y casos prácticos en python
 - Introducción
 - RDDs (Resilient Distributed Datasets)
 - Spark SQL y DataFrames
 - MLlib
 - Otras APIs como Spark Streaming o GraphFrames

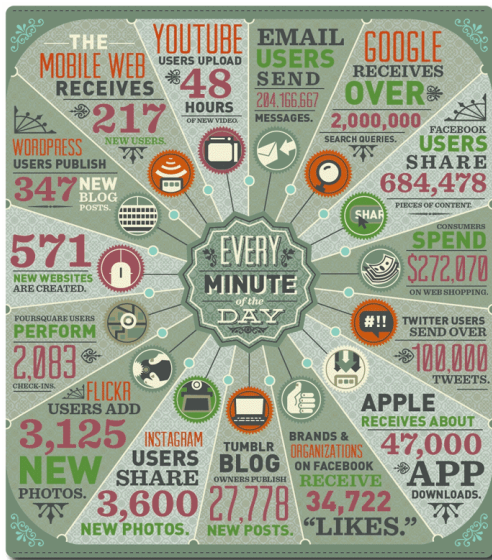
- 1 ¿Qué es Big Data? - ¿QUÉ?
- 2 Big Data: arquitecturas y escalabilidad
 - Evolución de las arquitecturas
 - Escalabilidad
- 3 Limitaciones e indicaciones de uso - ¿CUÁNDO?
 - Limitaciones Big Data
 - ¿Cuándo usar Big Data?
- 4 Soluciones Big Data - ¿CÓMO?
 - Implantación y despliegue
 - Clústers - Ejemplos



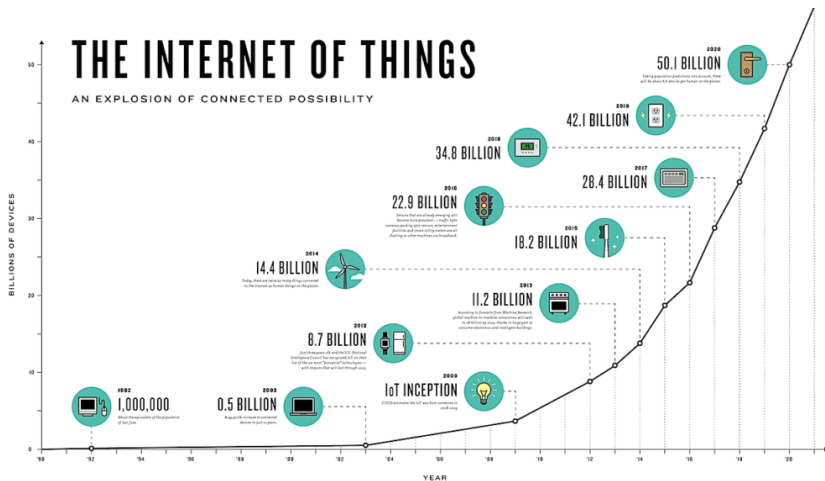
Fuente: <https://www.youtube.com/watch?v=TzxmjbL-i4Y>

- Término maltratado
 - Se ha convertido en humo y marketing
 - Si no haces Big Data no existes
- ¿Qué es?
 - ¿Análisis de datos y machine learning?
 - ¿Análisis de grandes volúmenes de datos?
 - ¿**Magia negra**?
- ¿Se trata de muchos datos o es algo mas?
- **Nueva era del dato. ¿Qué ha cambiado?**

¿Qué ha cambiado? - Internet social



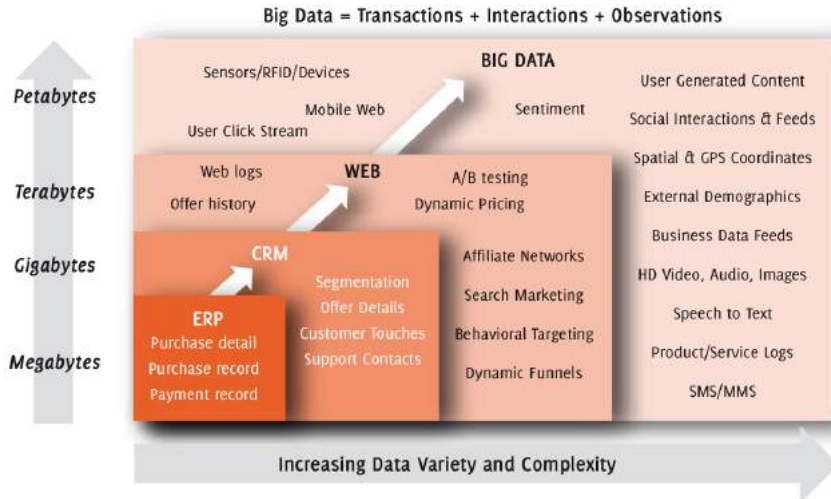
Fuente: <http://www.business2community.com/marketing/>



Fuente:

<https://disruptionhub.com/disrupted-electronics-internet-things-may-create-moores-law-steroids/>

¿Qué ha cambiado? - Fuentes



Source: Contents of above graphic created in partnership with Teradata, Inc.

Fuente: http://vint.sogeti.com/wp-content/uploads/2013/11/Sogeti_NoMoreSecrets.pdf

Antes

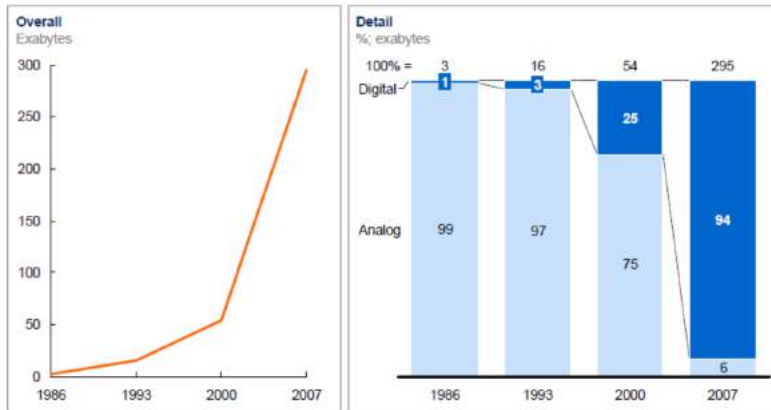
- Hardware caro al alcance de unos pocos
- Prestaciones limitadas
- Software solo para usuarios expertos
- Almacenar solo lo necesario

Ahora

- Hardware barato. Accesible
- Altas prestaciones
- Software amigable. **Democratización**
- HDD grande y barato. Almacenamos todos los datos que generamos
- Se ha descubierto el valor potencial de los datos en cualquier ámbito

Data storage has grown significantly, shifting markedly from analog to digital after 2000

Global installed, optimally compressed, storage



NOTE: Numbers may not sum due to rounding.

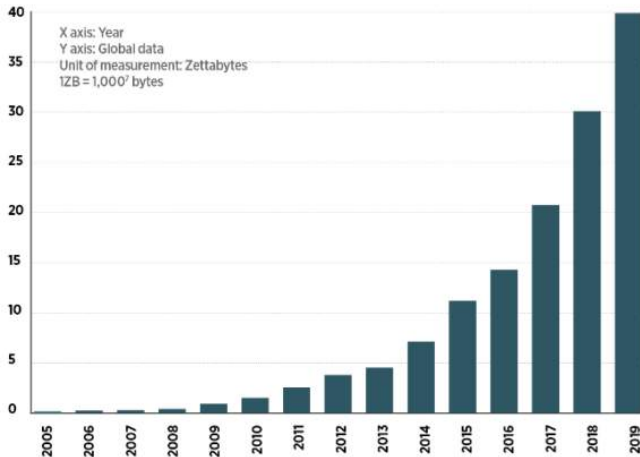
SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Fuente: http://www.planet-data.eu/sites/default/files/presentations/Big_Data_Tutorial_part4.pdf

Generamos y consumimos más datos

- **Auge digital:** Democratización del acceso a la tecnología (social y económicamente). Ahora cualquiera genera y consume datos
- Internet
- Dispositivos móviles
- El hecho de poder almacenarlos y explotarlos hace que generemos más datos
 - Tecnología para producir más datos. Sensores (IoT, SmartCities)
- El aumento del consumo de tecnología es en parte la causa de la reducción de costes
- Más usuarios = más consumidores (retroalimentación)

DATA GROWTH



Note: Post-2013 figures are predicted. Source: UNECE

Fuente: <https://www.nojitter.com/post/240170228/the-network-impact-of-big-data>

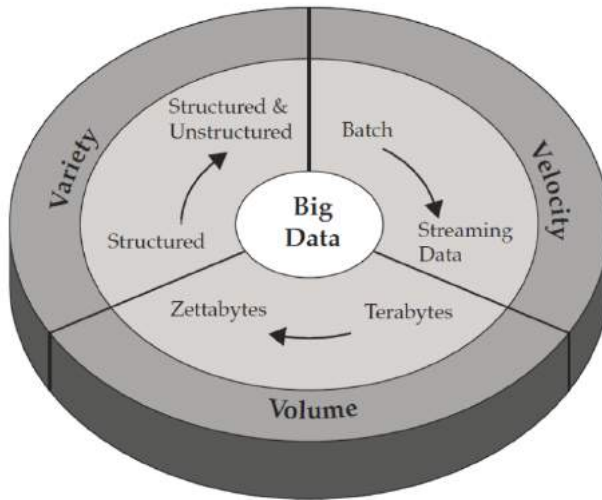
Velocidad

- Otra de las claves es la **velocidad** en los datos son generados y deben ser procesados
- IoT, transacciones, imágenes etc.
- Recoger y almacenar grandes cantidades en tiempo real puede suponer un gran reto

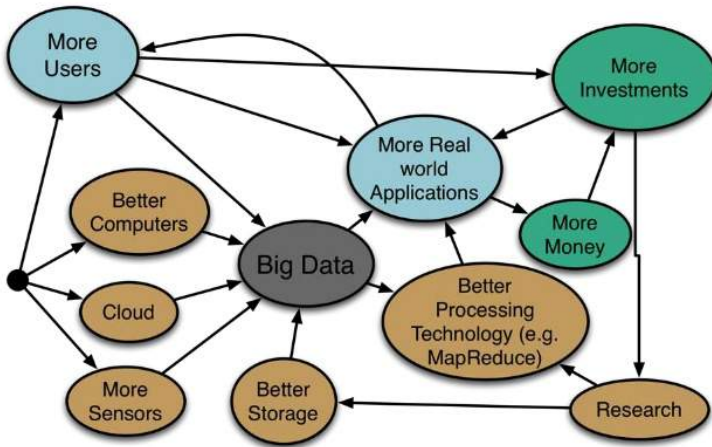
Variedad

- La disponibilidad de más y mejor tecnología permite analizar nuevos tipos de datos **no estructurados**
 - Textos libre (libros, artículos, blogs, redes sociales, etc.)
 - Todo tipo de registros (logs) de actividad (desplazamientos, llamadas, navegación web, etc.)
 - Multimedia (fotos, vídeos, audio)

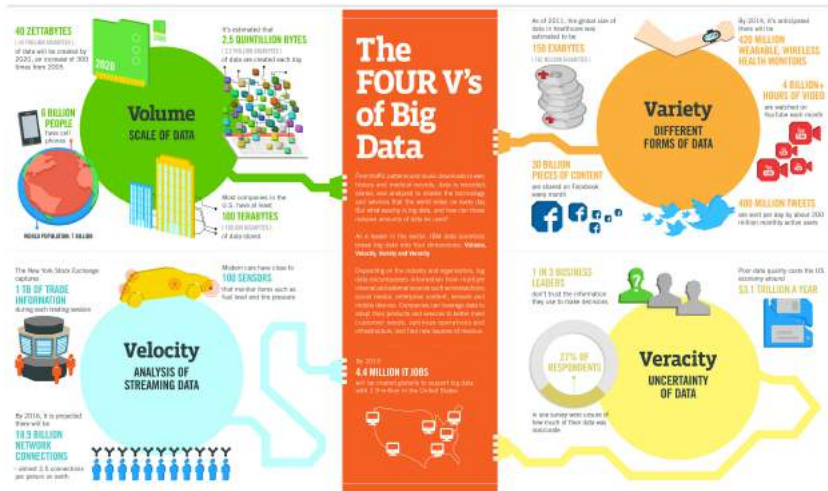
¿Qué ha cambiado? - Resumen: las 3 Vs de Big Data



Drivers of Big Data



Características - Punto de vista del dato: Las 3 Vs (o más)



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MPTIC, SAP

IBM

Fuente: <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>

El termino Big Data

- El **nuevo panorama del dato** requiere **nuevas tecnologías**
- **Una definición:** *"Big data es aplicar herramientas de ciencia de datos diseñadas para procesar grandes volúmenes de datos"*

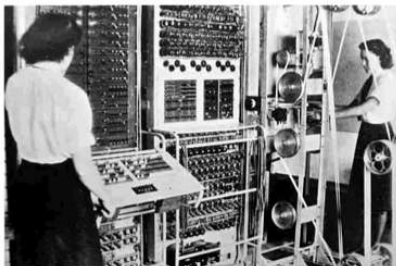
Las nuevas tecnologías

- No hardware nuevo o especializado
- No hardware de altas prestaciones (en la mayoría de los casos)
- ¿Revolución en las técnicas y algoritmos? **No**
 - Usamos las mismas técnicas (en realidad un **subconjunto**)
- ¿Cuál es la revolución del big data?
 - Arquitecturas y software para **gran volumen de datos**
 - Algoritmos (de toda la vida) adaptados
- **Arquitecturas distribuidas.** ¿Nuevas?
 - Sistemas de propósito general y transparentes

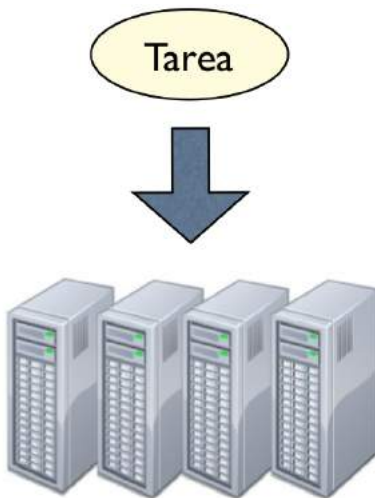
- 1 ¿Qué es Big Data? - ¿QUÉ?
- 2 **Big Data: arquitecturas y escalabilidad**
 - Evolución de las arquitecturas
 - Escalabilidad
- 3 Limitaciones e indicaciones de uso - ¿CUÁNDO?
 - Limitaciones Big Data
 - ¿Cuándo usar Big Data?
- 4 Soluciones Big Data - ¿CÓMO?
 - Implantación y despliegue
 - Clústers - Ejemplos

- 1 ¿Qué es Big Data? - ¿QUÉ?
- 2 **Big Data: arquitecturas y escalabilidad**
 - Evolución de las arquitecturas
 - Escalabilidad
- 3 Limitaciones e indicaciones de uso - ¿CUÁNDO?
 - Limitaciones Big Data
 - ¿Cuándo usar Big Data?
- 4 Soluciones Big Data - ¿CÓMO?
 - Implantación y despliegue
 - Clústers - Ejemplos

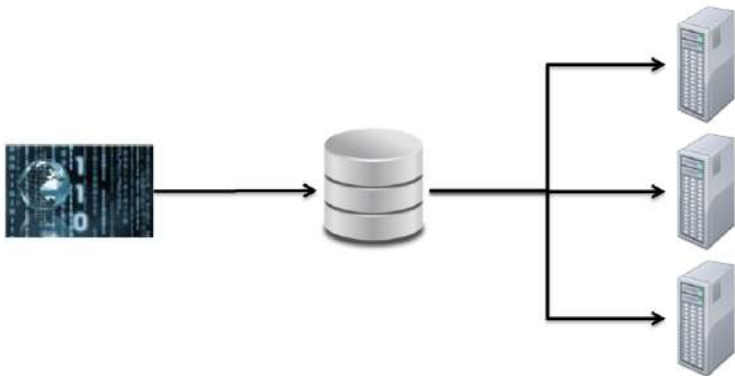
- Inicios de la computación, hasta el día de hoy
- Un único supercomputador. Proceamiento y almacenamiento centralizado
- Tipo de problemas: Pocos datos, mucho procesamiento
- Crecimiento vertical:
 - Si necesitamos más almacenamiento/procesamiento, añadimos más memoria/procesadores



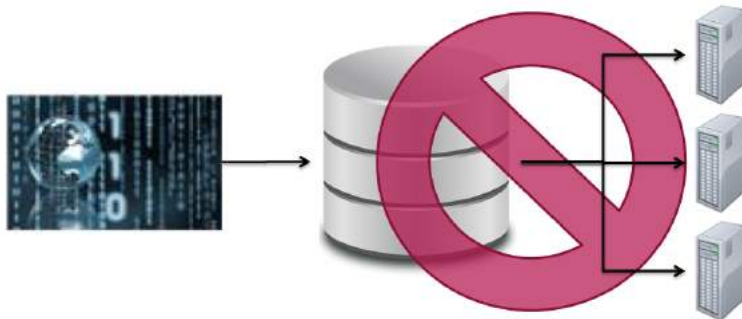
- Usar muchas máquinas para completar una única tarea



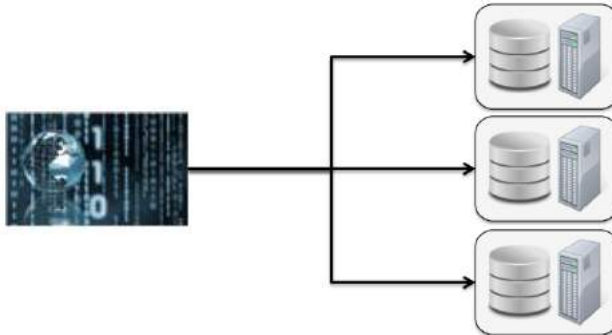
- Arquitectura frecuente hoy en día
- Almacenamiento centralizado (**altas prestaciones**)
- Unidades de procesamiento (pueden ser **convencionales**)
- Volumen de datos moderado



- Actualidad: datos del orden de **TeraBytes** o **PetaBytes**
- Saturación del almacenamiento centralizado
- El punto crítico está en las **comunicaciones**
 - Los datos y la capacidad de procesarlos han crecido más que la velocidad de las comunicaciones

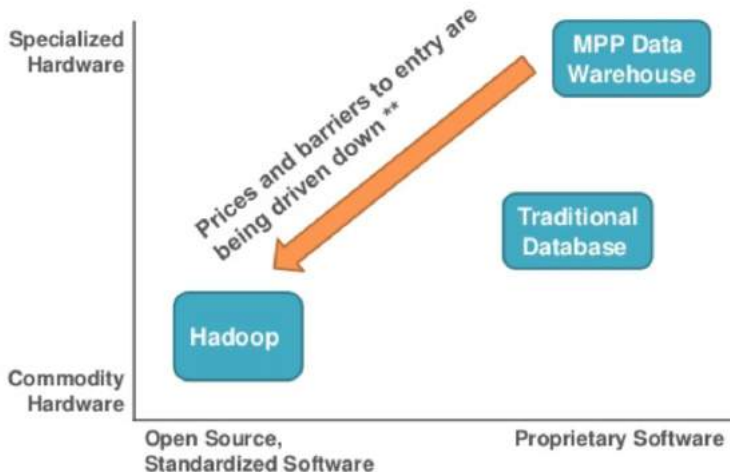


- También llamadas **arquitecturas Big Data**
- **Almacenamos los datos donde van a ser procesados**
 - Comunicación: se reduce todo lo posible
- Arquitectura **maestro/esclavos**
 - Maestro solo dirige, no procesa
- **Escalado horizontal**



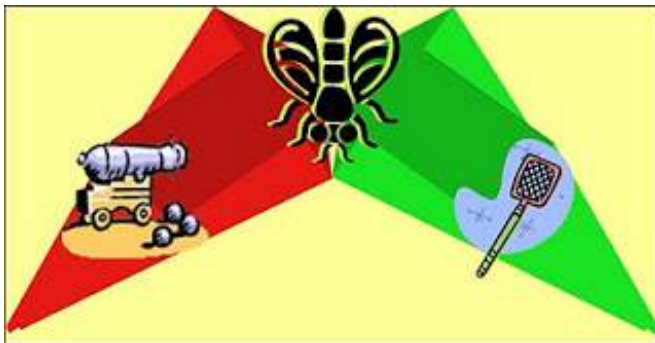
¿Seguro que esto es nuevo y único?

So the trend looks something like this



- Robusto ante **fallos** en los nodos
 - **Redundancia** de datos
 - Control de trabajos perdidos
- Procesamiento en paralelo y distribuido **transparente**. El analista no sabe si por detrás hay una gran máquina o 100 pequeñas
- Gran ecosistema de herramientas
 - **Automatización** de flujos de trabajo
- Escalado lineal de la capacidad de procesamiento y almacenamiento
 - **Crecimiento sostenible**. Según necesidad
 - Crecimiento casi **ilimitado**
- Y no menos más importante:
 - El **volumen** de datos almacenar y procesar
 - **Velocidad** de ingestión

-



- 1 ¿Qué es Big Data? - ¿QUÉ?
- 2 **Big Data: arquitecturas y escalabilidad**
 - Evolución de las arquitecturas
 - Escalabilidad
- 3 Limitaciones e indicaciones de uso - ¿CUÁNDO?
 - Limitaciones Big Data
 - ¿Cuándo usar Big Data?
- 4 Soluciones Big Data - ¿CÓMO?
 - Implantación y despliegue
 - Clústers - Ejemplos

Tecnología

- En tecnología se refiere a *la propiedad de aumentar la capacidad de trabajo o de tamaño de un sistema sin comprometer el funcionamiento y calidad normales del mismo*

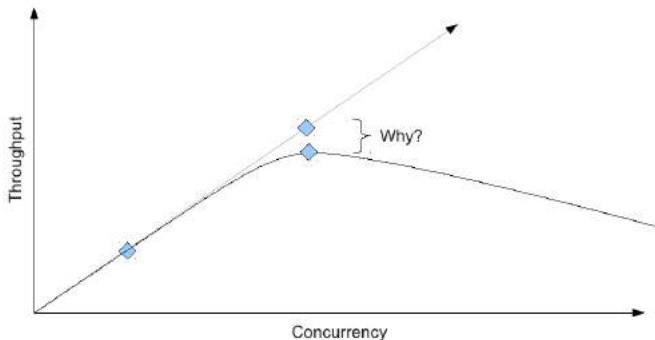
Computación

- En computación se refiere a *la propiedad de aumentar el número de usuarios, de **datos que procesa** o de solicitudes que recibe, sin que se afecte significativamente su **velocidad de respuesta***

Escalabilidad Lineal

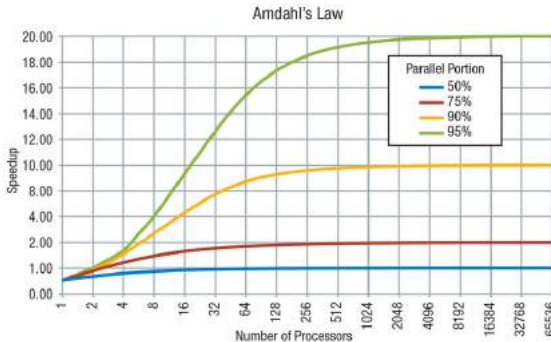
- *Un sistema cuyo rendimiento es mejorado después de haberle añadido más capacidad hardware, proporcionalmente a la capacidad añadida, se dice que **escala linealmente***
- **La escalabilidad ideal no es realista**

System Scalability



www.percona.com

- Estima la mejora que se producirá en un programa si lo paralelizamos
- Muestra que la **aceleración** de un programa estará **limitada por la porción serial** del mismo
- Cuanto más procesadores añadimos más notable es la penalización de la parte serial
- Ley teórica **¿Y las comunicaciones?**



Fuente: <http://rtcmagazine.com/articles/view/103209>

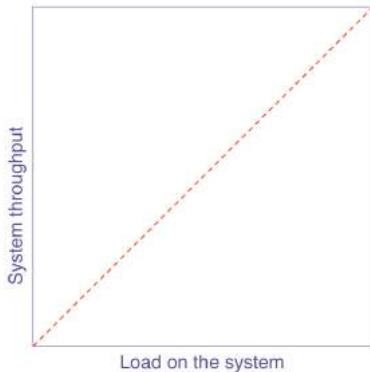
- Neil J. Gunther, 1993
- Gunther dice que *La escalabilidad debe ser cuantificada. Si no puedes cuantificarla no puedes garantizarla*
- El objetivo de la USL (Universal Scalability Law) es proporcionar una cuantificación teórica de la escalabilidad
- Esta ley se aplica tanto a software como a hardware
- **Escalabilidad software:** Mide el rendimiento del sistema a medida que aumenta la carga de trabajo (e.g. usuarios del software). El hardware se mantiene fijo
- **Escalabilidad hardware:** Mide el rendimiento del sistema a medida que aumenta el número de procesadores del manteniendo fija la carga de trabajo por procesador

Formulación

$$C(N) = \frac{N}{1 + \alpha(N - 1) + \beta N(N - 1)}$$

Parámetros

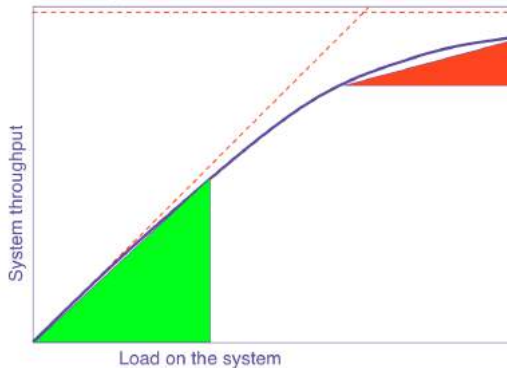
- N es el número de procesadores del sistema
- $C(N)$ es el rendimiento teórico del sistema
- **Escalado lineal (1):** escalado ideal
- **Contención ($\alpha(N - 1)$):** latencia producida por componentes hardware compartidos o elementos software no paralelizables
- **Coherencia ($\beta N(N - 1)$):** latencia hasta alcanzar la consistencia de datos



Fuente: <http://www.perfdynamics.com/Manifesto/USLscalability.html>

Situación ideal

$$\alpha = 0, \beta = 0$$



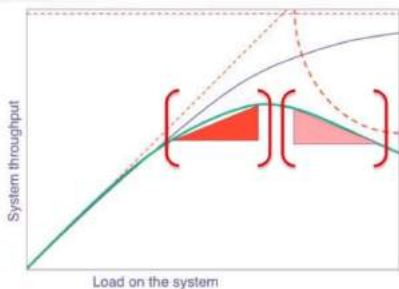
Fuente: <http://www.perfdynamics.com/Manifesto/USLscalability.html>

Contención

$$\alpha > 0, \beta = 0$$

System throughput is limited by

- Contention
 - Queueing on a shared resource, $O(N)$
- Coherence
 - Coordination and communication between all nodes, $O(N^2)$



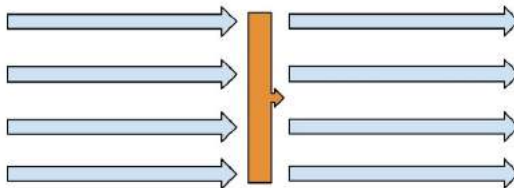
$$C(N) = \frac{1}{1 + \left(\frac{\sigma(N-1)}{\kappa N(N-1)} \right)}$$

<http://www.perfdynamics.com/Manifesto/USLscalability.html>

Coherencia y contención

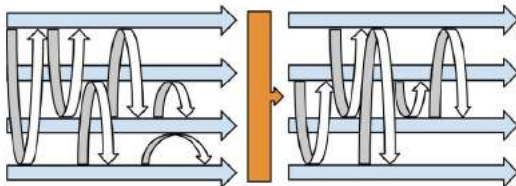
$$\alpha > 0, \beta > 0$$

Amdahl Scalability



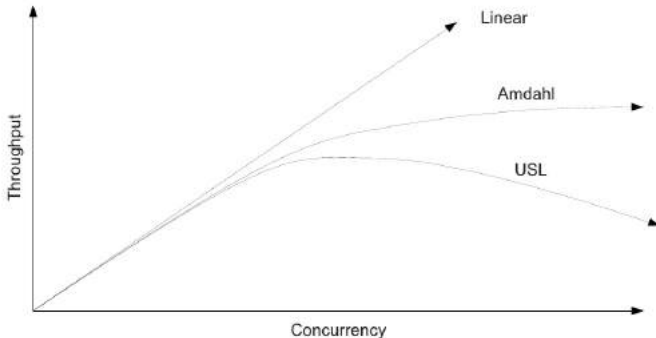
$$C(N) = \frac{N}{1 + \sigma(N - 1)}$$

USL Scalability



$$C(N) = \frac{N}{1 + \sigma(N-1) + \boxed{\kappa N(N-1)}}$$

Universal Scalability Law

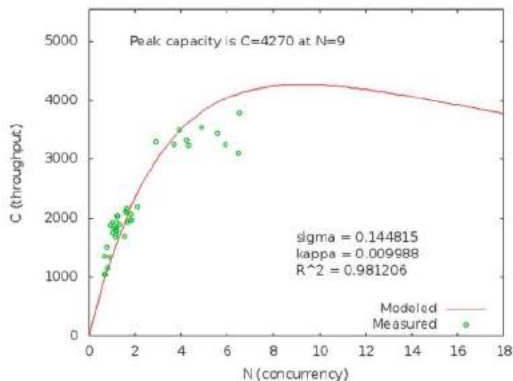


www.percona.com

Aplicación

- Tomar **muestras** del rendimiento de nuestro sistema en diferentes situaciones:
 - Carga de trabajo
 - Capacidad
- Ajustamos los datos a la ecuación para obtener los parámetros para nuestro sistema
- Ya tenemos un **modelo teórico** de nuestro sistema. Ahora podemos:
 - **Simular** diferentes situaciones de carga
 - Estimar el beneficio/coste de ampliar nuestro sistema
 - Obtener el número de procesadores óptimo en diferentes tareas

USL Scalability Modeling



- 1 ¿Qué es Big Data? - ¿QUÉ?
- 2 Big Data: arquitecturas y escalabilidad
 - Evolución de las arquitecturas
 - Escalabilidad
- 3 Limitaciones e indicaciones de uso - ¿CUÁNDO?
 - Limitaciones Big Data
 - ¿Cuándo usar Big Data?
- 4 Soluciones Big Data - ¿CÓMO?
 - Implantación y despliegue
 - Clústers - Ejemplos

- 1 ¿Qué es Big Data? - ¿QUÉ?
- 2 Big Data: arquitecturas y escalabilidad
 - Evolución de las arquitecturas
 - Escalabilidad
- 3 Limitaciones e indicaciones de uso - ¿CUÁNDO?
 - Limitaciones Big Data
 - ¿Cuándo usar Big Data?
- 4 Soluciones Big Data - ¿CÓMO?
 - Implantación y despliegue
 - Clústers - Ejemplos



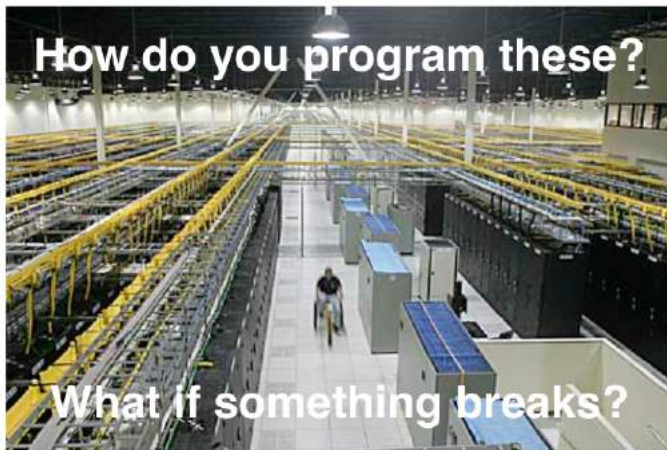
Fuente: <https://es.linkedin.com/pulse/la-burbuja-del-bigdata-alberto-conde-mellado>

*“Las nuevas aptitudes que tanto atraen la atención de los medios no sirven para resolver más eficazmente el problema de la inferencia; **son puras técnicas de supervivencia para gestionar los artefactos inducidos por la computación distribuida a gran escala.** Lidian con las enormes restricciones que impone el mundo de los sistemas multiproceso y distribuidos sobre los algoritmos. En este mundo tan constreñido, el elenco de algoritmos utilizables es tan limitado si se lo compara con el disponible en el de un único procesador, que **es inevitable adoptar técnicas estadísticas que hubieran sido tachadas de rudimentarias, si no de inadecuadas, en otros tiempos.** Estos problemas consumen nuestro tiempo y energía, deforman nuestro criterio sobre lo que resulta adecuado y nos desvían de las estrategias de análisis de datos que habríamos aplicado de oficio en otras circunstancias.”*

<https://www.datanalytics.com/2017/03/09/un-parrafo-afortunadisimo-sobre-las-nuevas-aptitudes/>

Artículo completo: <http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf>

- ¿Y el software? **Mayor complejidad, menor rendimiento**
 - Procesamiento distribuido, particionado, redundancia, comunicaciones, etc.



Fuente: <http://semantica.cs.lth.se/pyspark/slides/Lecture%201.pdf>

- **Artículo:** <https://www.techrepublic.com/article/how-complexity-is-killing-big-data-deployments/>

Rule of Thumb: Scalability reduces as implementation complexity grows

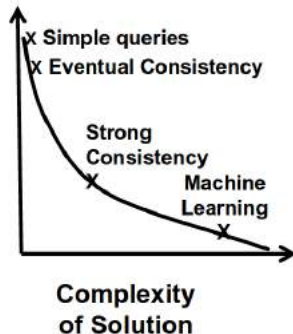
Workload

- # of concurrent sessions and operations
- Operation mix (create, read, update, delete)
- Generally, each system use case represents a distinct and varying workload

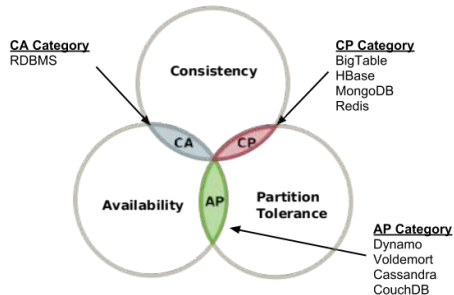
Data Sets

- Number of records
- Record size
- Record structure (e.g., sparse records)
- Homogeneity/heterogeneity of structure/schema
- Consistency

Scalability

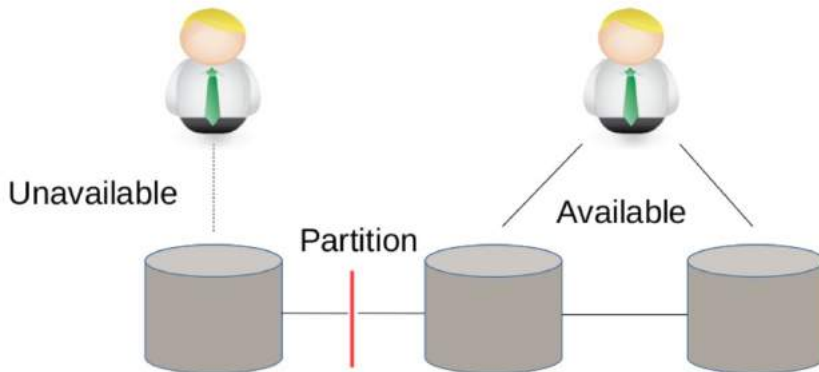


- **Teorema CAP (o de Brewer):** Un sistema puede tener **no más de dos** de estas características simultáneamente
 - **Consistencia:** Todos los nodos ven la misma información al mismo tiempo
 - **Disponibilidad:** Garantiza de que cada petición a un nodo es atendida correctamente
 - **Tolerancia al particionado:** El sistema sigue funcionando incluso si algún nodo falla
- **Problema: las comunicaciones**



- ¿Qué necesitas? Elige: **Consistencia** o Disponibilidad

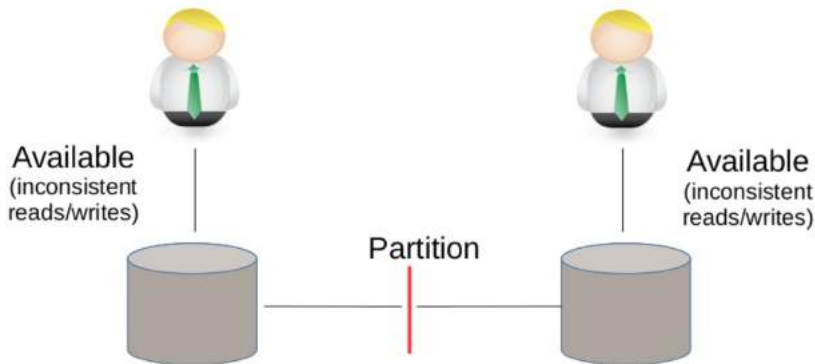
CP systems (Partition)



Fuente: https://www.slideshare.net/Klika_Tech/cap-theorem-and-distributed-systems

- ¿Qué necesitas? Elige: Consistencia o **Disponibilidad**

AP systems (Partition)

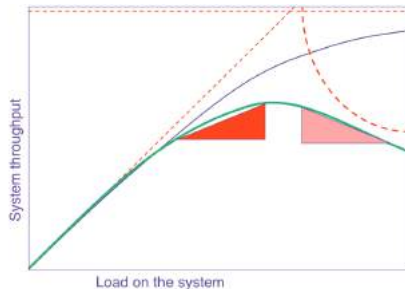
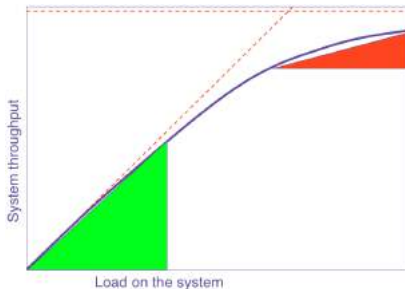


Fuente: https://www.slideshare.net/Klika_Tech/cap-theorem-and-distributed-systems

- MENTIRA**



- Particionado: Según el teorema CAP no podemos tener **disponibilidad** y **consistencia** al mismo tiempo
- Paralelismo: A mayor complejidad menor grado de paralelización
 - **USL - Contención:** Tiempo perdido por código(serial)/recursos compartidos
- Consistencia: Las comunicaciones son **lentas**, y la penalización por intercambio de datos entre nodos **alta**
 - **USL - Coherencia:** Tiempo perdido hasta alcanzar la coherencia de datos entre nodos
- En la mayoría la mejora obtenida estará **muy lejos de la ideal**



- ¿Escalado lineal? **¡¡Si, en costes!!**
- Si tengo un cluster con N nodos:
- **Consumo eléctrico:** multiplica $\times N$
- **Equipamiento:** multiplica $\times N$
- **Licencias Software:** multiplica $\times N$
- **Instalación y mantenimiento (software y hardware):**
 - Multiplica $\times N$
 - Este coste puede ser mayor (complejidad de la instalación)
- **Infraestructura de red**
- Otras limitaciones:
 - Complejidad y limitaciones del paralelismo $1 + 1 < 2$
 - Almacenamiento (redundancia): $Total/3 - OS * N$
 - **El nodo maestro**

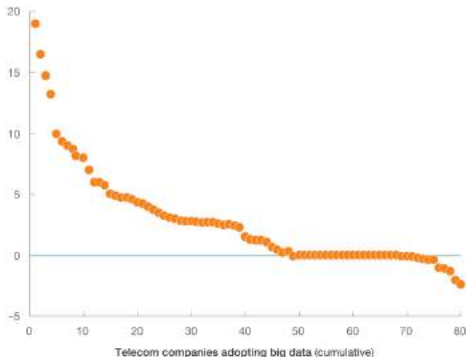
- 1 ¿Qué es Big Data? - ¿QUÉ?
- 2 Big Data: arquitecturas y escalabilidad
 - Evolución de las arquitecturas
 - Escalabilidad
- 3 Limitaciones e indicaciones de uso - ¿CUÁNDO?
 - Limitaciones Big Data
 - ¿Cuándo usar Big Data?
- 4 Soluciones Big Data - ¿CÓMO?
 - Implantación y despliegue
 - Clústers - Ejemplos

¿Todo el mundo usa Big Data?

- En 2015 **80 de 273 (29 %)** telecoms había invertido en BD
- 5 tienen un 10 % o mas de beneficios. La mitad sin beneficios

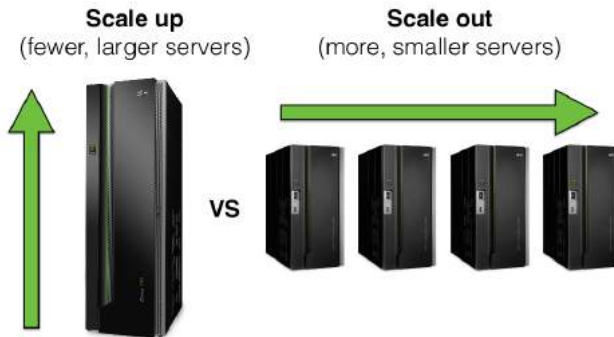
Among telecom companies that invested in big data, only a few enjoyed an incremental profit impact exceeding 10 percent.

Impact of big data on telecom companies' profits,
% of total profit



Source: 2015 McKinsey survey of 273 global telecom companies, 80 of which have made big data analytics investments.

- Horizontal: Si la máquina se queda pequeña, la ampliamos. Si no podemos, la cambiamos por una mayor
- Vertical: Añadimos máquinas. Aumenta la complejidad del software (sistemas distribuidos)



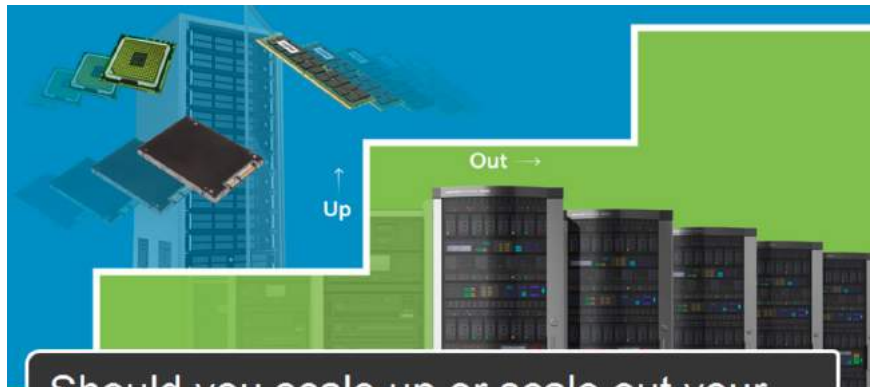
Fuente: <http://semantica.cs.lth.se/pyspark/slides/Lecture%201.pdf>

Módulo memoria RAM (DDR4 - ECC)		
Capacidad	Precio	Precio/Gb
8Gb	80\$	10\$
16Gb	170\$	10,625\$
32Gb	340\$	10,625\$
64Gb	700\$	10,937\$
128Gb	2900\$	22,656\$

Disco Duro SATA 3.5"		
Capacidad	Precio	Precio/Tb
1Tb	100€	100€
2Tb	160€	80€
4Tb	230€	57,5€
6Tb	330€	55€
8Tb	420€	52,5€
10Tb	520€	52€
12Tb	690€	57,5€

Disco Duro SSD		
Capacidad	Precio	Precio/Tb
250Gb	94€	376€
500Gb	167€	334€
1Tb	310€	310€
2Tb	703€	351,5€
4Tb	1400€	350€

Procesadores Intel Xeon E5-26xx-v3			
Modelo	Cores (hilos)	Precio	Precio/Core
Xeon E5-2620-v3	6 (12)	417\$	69,5\$
Xeon E5-2630-v3	8 (16)	667\$	83,375\$
Xeon E5-2650-v3	10 (20)	1166\$	116,6\$
Xeon E5-2670-v3	12 (24)	1589\$	132,417\$
Xeon E5-2695-v3	14 (28)	2424\$	173,143\$
Xeon E5-2698-v3	16 (32)	3226\$	201,625\$
Xeon E5-2699-v3	18 (36)	4115\$	228,61\$



Should you scale up or scale out your data center?

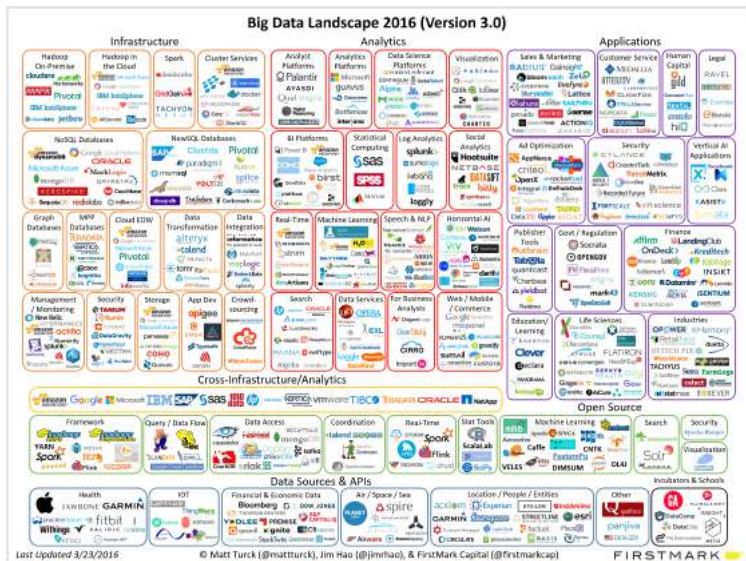
Fuente: <http://www.microlandusa.com/microland/marketing/Crucial-Scale-Up-or-Scale-Out.html>

Algunas claves

- Volumen de datos
- **Crecimiento** del volumen de datos **en el tiempo**
- Ingestion y almacenamiento en **tiempo real**
- Complejidad de las tareas. Grado de paralelismo
 - Transformación y almacenamiento
 - Algoritmos de aprendizaje automático
- **Ejemplo (2009):** <https://blog.codinghorror.com/scaling-up-vs-scaling-out-hidden-costs/>
- **Asistente:** <http://www.microlandusa.com/microland/marketing/Crucial-Scale-Up-or-Scale-Out.html>

Conclusión

- Evítalo si puedes
- ¿El escalado vertical cubre mis necesidades a largo plazo?



Fuente: <http://mattturck.com/2016/02/01/big-data-landscape/>

- 1 ¿Qué es Big Data? - ¿QUÉ?
- 2 Big Data: arquitecturas y escalabilidad
 - Evolución de las arquitecturas
 - Escalabilidad
- 3 Limitaciones e indicaciones de uso - ¿CUÁNDO?
 - Limitaciones Big Data
 - ¿Cuándo usar Big Data?
- 4 Soluciones Big Data - ¿CÓMO?
 - Implantación y despliegue
 - Clústers - Ejemplos

- 1 ¿Qué es Big Data? - ¿QUÉ?
- 2 Big Data: arquitecturas y escalabilidad
 - Evolución de las arquitecturas
 - Escalabilidad
- 3 Limitaciones e indicaciones de uso - ¿CUÁNDO?
 - Limitaciones Big Data
 - ¿Cuándo usar Big Data?
- 4 Soluciones Big Data - ¿CÓMO?
 - Implantación y despliegue
 - Clústers - Ejemplos

- Las soluciones Big Data, suelen diseñarse a medida
- Se requiere análisis de la necesidad y planificación del despliegue

Principales cuestiones

- Acotar caso(s) de uso y definir sus **requisitos**
 - Tiempo real
 - Volumen de datos y su crecimiento esperado
 - Seguridad/Privacidad
- Recursos disponibles: económicos, infraestructura (**Red**), etc.
- Personal: ¿equipo IT cualificado? ¿solo analistas?
- Previsión de las necesidades en tiempo y recursos computacionales de cada caso de uso: **¿puntual? ¿estable? ¿fluctua mucho?**

Forma de gasto

- **Gasto de capital** (CAPEX – Capital Expenditure): Tenemos dinero disponible y lo *invertimos* de golpe en adquirir un activo (máquinas o licencias) que pasa a formar parte de la empresa y que será amortizado con el paso del tiempo
- **Gasto operativo** (OPEX – Operating Expenditure): Costes de funcionamiento recurrentes, se pagan según se consumen y se realizan según la necesidad

On Premises - CAPEX

- **Hardware On Premises:** El hardware pertenece a la empresa y es gestionada y mantenida por la misma
- **Software On Premises:** Aplicaciones alojadas, gestionadas y mantenidas por la propia empresa

Servicios Cloud - OPEX

- **Servicios en la nube:**
 - Son alojados, gestionados y mantenidos por un tercero
 - Son contratados y facturados **bajo demanda**
- **IaaS** (Infrastructure as a Service): Capacidad de cálculo y almacenamiento (normalmente son servicios **virtualizados**)
 - Amazon EC2
- **PaaS** (Platform as a Service): Ecosistema de herramientas (para desarrolladores) que cubre el ciclo de vida completo de desarrollo y puesta en marcha de aplicaciones y servicios (**MLaaS, BDaaS, DBaaS**)
 - Amazon EMR
- **SaaS** (Software as a Service): Aplicaciones de dominio específico listas para ser usadas por un usuario final (no técnico)
 - <https://www.salesforce.com>

CLOUD MACHINE LEARNING SERVICES COMPARISON

	Amazon ML	Amazon SageMaker*	Azure ML Studio	Google Prediction API	Google ML Engine**
Classification	✓	✓	✓	✓	✓
Regression	✓	✓	✓	✓	✓
Clustering	✗	✓	✓	✗	✓
Anomaly detection	✗	✓	✓	✗	✓
Recommendation	✗	✓	✓	✗	✓
Ranking	✗	✓	✓	✗	✓
Algorithms	unknown	10 built-in + custom available	100+ algorithms and modules	unknown	TensorFlow-based
Frameworks	✗	TensorFlow, MXNet	✗	✗	TensorFlow
Graphical interface	✗	✗	✓	✗	✗
Automation level	high	medium	low	high	low

*Both out-of-the-box features and possible custom-built features are marked as available in Amazon SageMaker.

**The features available in TensorFlow are respectively marked as available in Google ML Engine.

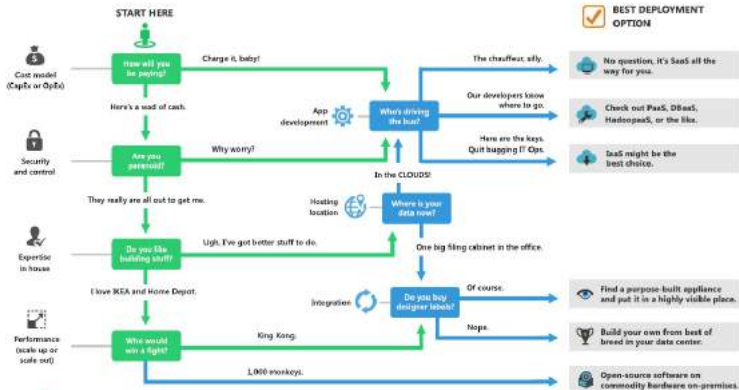


Fuente: <https://www.altexsoft.com/blog/datascience/>



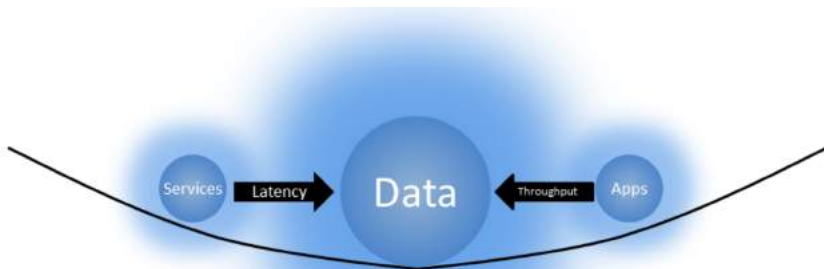
Choosing the Right Big Data Deployment Model

Many enterprises are looking to roll their first large-scale big data initiative out to full production. Yet what may have looked easy on the whiteboard doesn't always meet the operational requirements of the business. Follow our handy (and fun) flowchart to help find the approach that suits you.



© 2014 by the Enterprise Strategy Group, Inc.

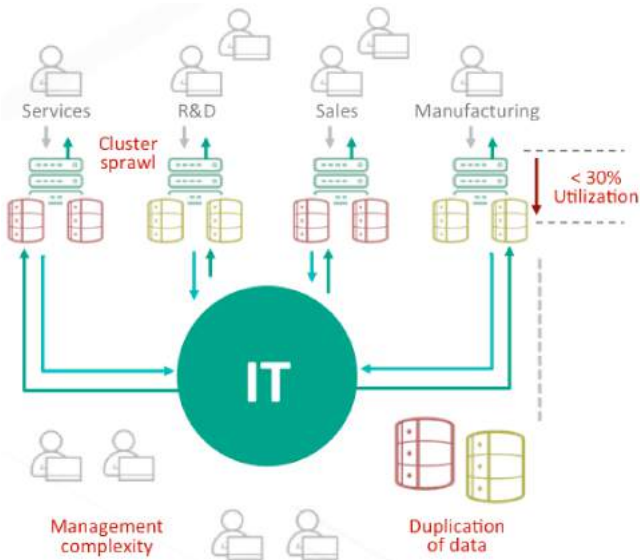
- ¿Donde están tus datos? No son fáciles de mover
- Cuanto mas grandes son, capacidad de atracción sobre **IT** (fuertemente vinculado al almacenamiento)



Fuente: <https://blog.mccrory.me/2010/12/07/data-gravity-in-the-clouds/>

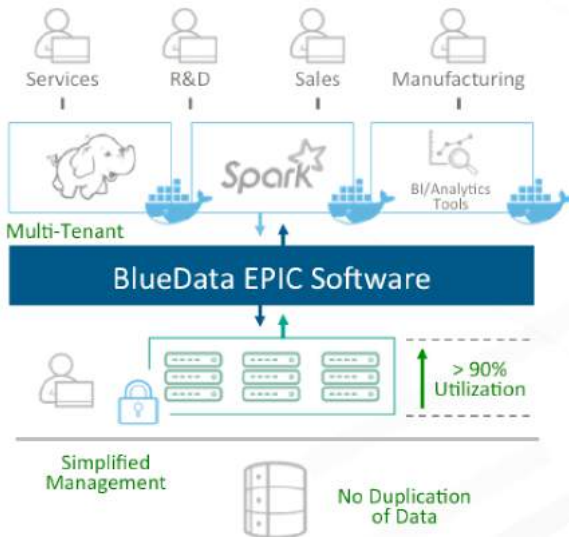
Definición Wikipedia:

- **Server sprawl:** *is a term used in the information technology industry. It describes a set of situations that can occur in data centers that result in poor hardware resource utilization, poor system and software level security, and wasted energy. Various techniques exist to mitigate server sprawl, such as computer virtualization*



Fuente: <https://www.bluedata.com/blog/2016/03/announcing-the-bluedata-epic-spring-release/>

- ¿Como evitar *Cluster Sprawl* si no puedo mudarme a la nube?



Fuente: <https://www.bluedata.com/blog/2016/03/announcing-the-bluedata-epic-spring-release/>

On Premises

- Gran inversión inicial. Barato a largo plazo
- Requiere planificación de las necesidades a largo plazo
- Requiere conocimiento experto en arquitecturas y sistemas
- Adecuado para empresas y proyectos maduros

Cloud

- Sin inversión inicial. Caro a largo plazo
- Usa y paga lo que necesitas
- Improvisación. Amplia o reduce los recursos según necesidad
- Diferentes niveles de conocimiento experto en sistemas:
 - **IaaS** (Alto), **PaaS** (Medio), **SaaS** (Bajo)
- Adecuado para necesidades puntuales o variables
 - I+D, Startups, proyectos concretos, MVP, etc.
 - ML: Entrena un modelo y apaga al terminar

- 1 ¿Qué es Big Data? - ¿QUÉ?
- 2 Big Data: arquitecturas y escalabilidad
 - Evolución de las arquitecturas
 - Escalabilidad
- 3 Limitaciones e indicaciones de uso - ¿CUÁNDO?
 - Limitaciones Big Data
 - ¿Cuándo usar Big Data?
- 4 Soluciones Big Data - ¿CÓMO?
 - Implantación y despliegue
 - Clústers - Ejemplos

Data Storage - NetApp

- Almacenamiento y análisis de *logs* sobre sistemas de almacenamiento
- Clúster: **más de 30 nodos**
- Volumen: **7TB al mes**

Finanzas - Banca

- Validación de calidad y precisión de datos y cumplimiento de regulaciones gubernamentales
- Clúster: **más de 20 nodos**
- Volumen: **1TB al mes**

Fuente (2012): https://hadoopilluminated.com/hadoop_illuminated/Hadoop_Use_Cases.html

Sanidad

- Almacenamiento y análisis de datos sobre reclamación y pago de gastos médicos (entre aseguradoras y hospitales)
- Clúster: **más de 10 nodos**. Era un proyecto piloto
- Volumen: **1TB al día**

Telecomunicaciones - China Mobil Guangdong

- Almacenamiento y consulta de registros de llamadas e información de facturación
- Clúster: **mas de 100 nodos**
- Volumen: **30TB al mes**

Fuente (2012): https://hadoopilluminated.com/hadoop_illuminated/Hadoop_Use_Cases.html

Statistics

- 10's PB of data all across Nokia
- Multi-tenant, multi-petabyte analytics cluster
- 10-20K+ jobs per day
- 600+ internal users
- 250M+ KV queries
- Over a terabyte flowing every day
- Multiple data centers around the world

7

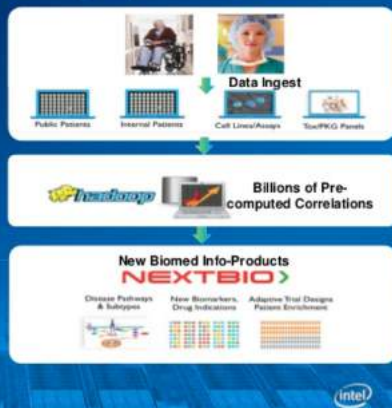
NOKIA

Fuente: https://hadoopilluminated.com/hadoop_illuminated/cached_reports/Nokia_Bigdata.pdf

BioScience: Genomics for Translational Medicine

Hadoop for Data Correlation & Discovery Insights

- **Challenge:** Derive new value added patient discovery services while bringing down genome processing costs
- **Solution:** Dynamically partition & scale correlation of patient data to all public data using Hadoop and Hbase
- **Benefits:** Contributes to **800x** reduction in cost to process 4 Million genome variants
- **Infrastructure and Data Characteristics:**
 - 10 Node Hbase Cluster
 - Billions of pre-computed correlations
 - 1 Genome → 10 Million rows
 - 100 Genomes → 1 Billion rows
 - 1M Genomes → 10 Trillion rows
 - 100M Genomes → 1 Quadrillion 1,000,000,000,000,000 rows



Public Sector- Smart Traffic Intelligent Transport System

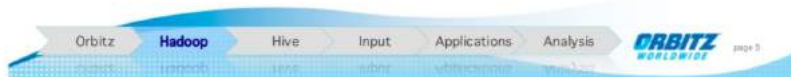
Hadoop for Predictive Analytics

- **Challenge:** Analyze city traffic to derive statistics for crime prevention, info sharing, and predictive traffic analysis
- **Solution:** Embed HBase client in camera for real-time inserts of structured/unstructured data
- **Benefits:**
 - Automated queries for traffic violation
 - Data mining of fake licenses <1 minute for all data captured for a week
 - Predictive traffic forecasting
- **Data Characteristics:**
 - 30000 + camera data collection points
 - Petabytes of traffic data & terabytes of images
 - 2 billion HBase records



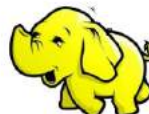
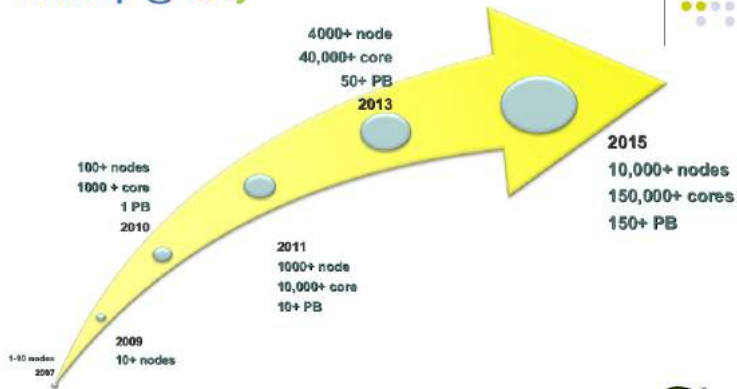
Data Challenges

- Orbitz.com generates ~1.5 million air searches and ~1 million hotel searches every day.
- All of this activity generates massive amounts of data – over 500 GB/day of log data, and even this doesn't capture all of the data we want.
- Expensive and difficult to use existing data infrastructure for storing and processing this data.
- Need an infrastructure that provides:
 - Long term storage of very large data sets.
 - Open access to developers and analysts.
 - Allows for ad-hoc querying of data and rapid deployment of reporting applications.



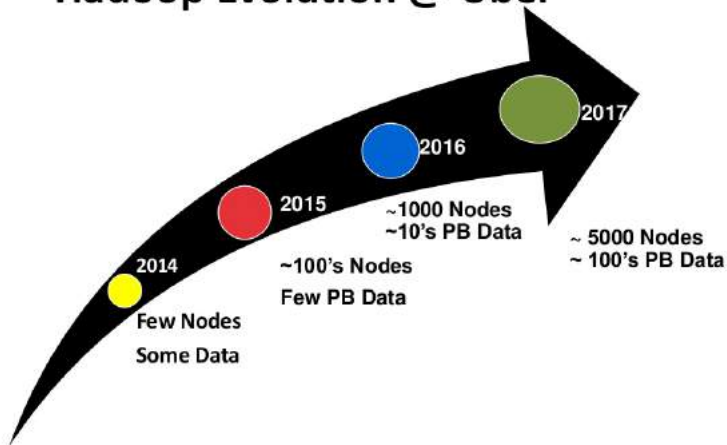


Hadoop @ ebay





Hadoop Evolution @ Uber



- ¿Y en España?

Telecomunicaciones - Telefónica Research

- Data mining and user modeling, multimedia, and internet research groups
- Clúster: **6 nodos. 96 cores**
- Nodo: **8GB** RAM. **2TB** almacenamiento
- Listado completo en:
<https://wiki.apache.org/hadoop/PoweredBy>
- No está completamente actualizado

Big Thanks!!!

