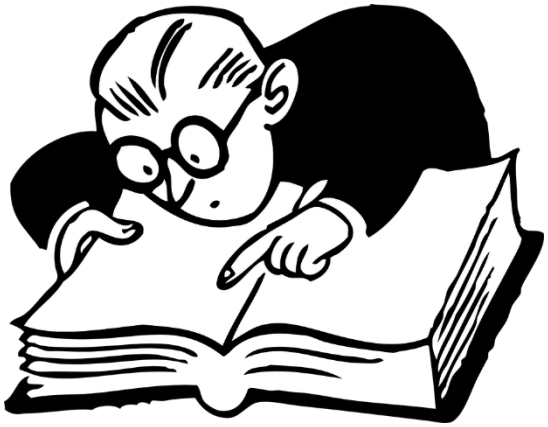
El Diccionario del Big Data



Sergio Ilarri Artigas

Jornada ProCom 2016, Zaragoza, 26/10/2016

Escuela de
Ingeniería y Arquitectura
Universidad Zaragoza

This presentation has been created for education purposes and no commercial intention. His author has made his best to properly reference and link the relevant sources (e.g., for images and videos publicly available) and avoid the use of private materials from third parties (e.g., by using images with public domain licenses available at http://pixabay.com). The copyrights of the different works and images, as well as the templates used for the presentation, belong to their respective authors.

You are free to use this material for your private use under the following terms:

Attribution — You must give appropriate credit to its creator and this license, not suggesting that the licensor endorses you or your use.

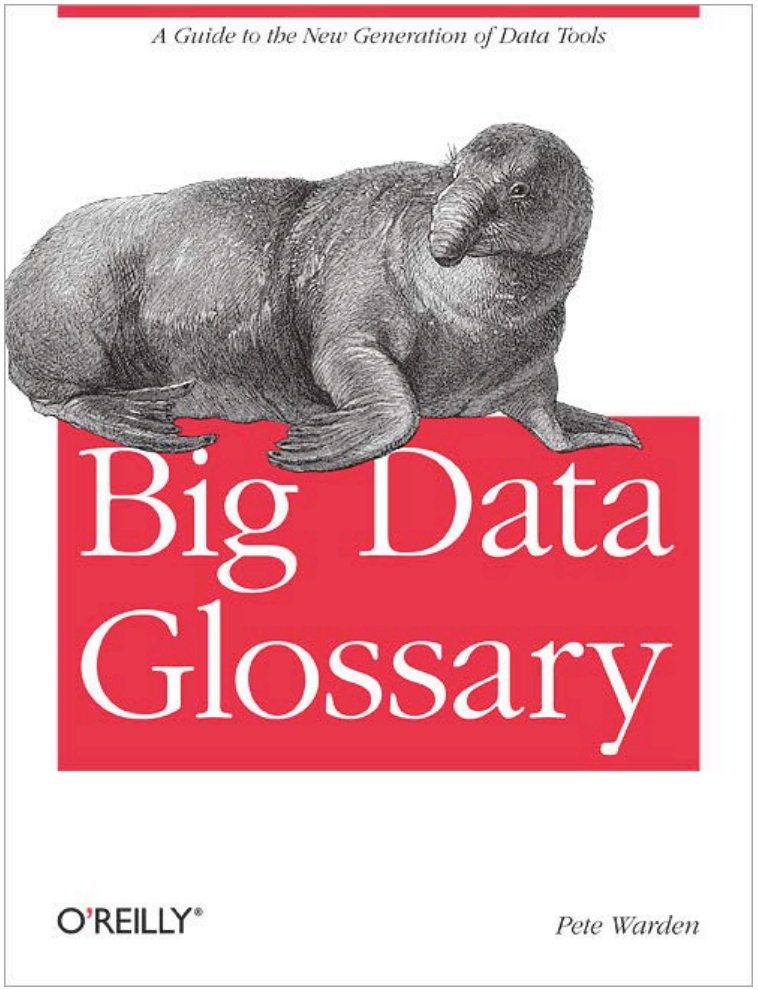
Noncommercial — You may not use the material for commercial purposes.

No Derivatives — If you remix, transform, or build upon the material, you may not distribute the modified material.

No Warranties — You assume full responsibility in the use of this material.

Share — Sharing (parts of) this presentation is not explicitly prohibited by the author, as long as you assume full responsibility and respect intellectual property rights. However, it must be observed that this presentation has been conceived exclusively for private use and education purposes. Any unauthorized use is strictly prohibited.

Esta presentación tiene únicamente fines educativos. Está concebida para un uso exclusivamente privado y no para su difusión fuera del ámbito en el que tiene lugar la presentación. Las imágenes utilizadas pueden pertenecer a terceros y, por tanto, son propiedad de sus autores. El autor de la presentación ha hecho todo lo posible para respetar los derechos de propiedad intelectual y citar apropiadamente las fuentes y autores.

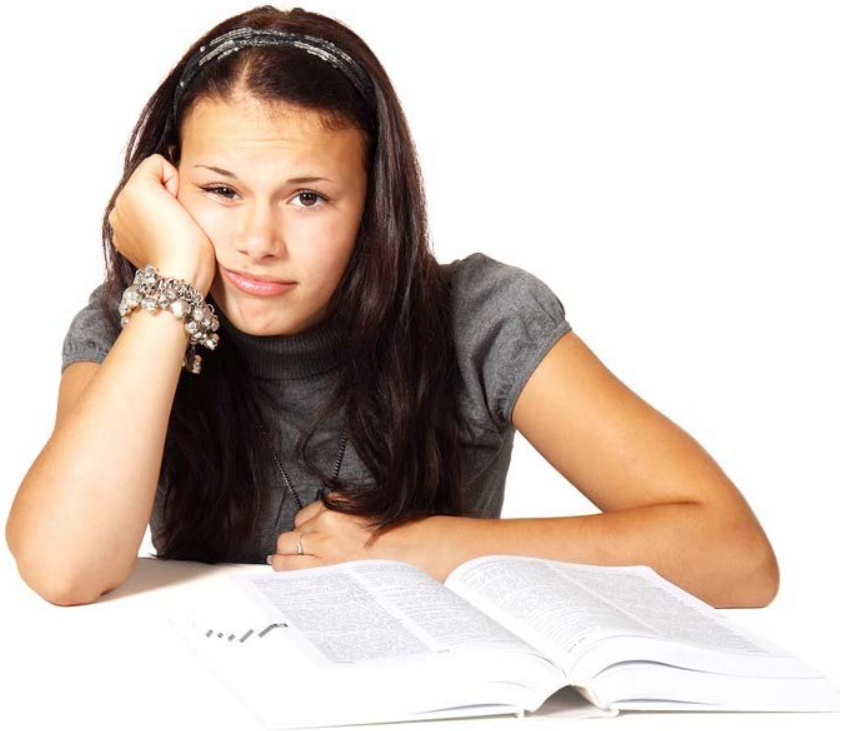


http://shop.oreilly.com/product/0636920022466.do
September 2011, 62 pages

¿Un Diccionario de Big Data?

“Dictionaries are like watches, the worst is better than none and the best cannot be expected to go quite true”

Samuel Johnson (English writer, 1709-1784)



Big Data

Dan Ariely lo compara con el sexo entre adolescentes:

-Todo el mundo habla de ello

-Nadie sabe realmente cómo hacerlo

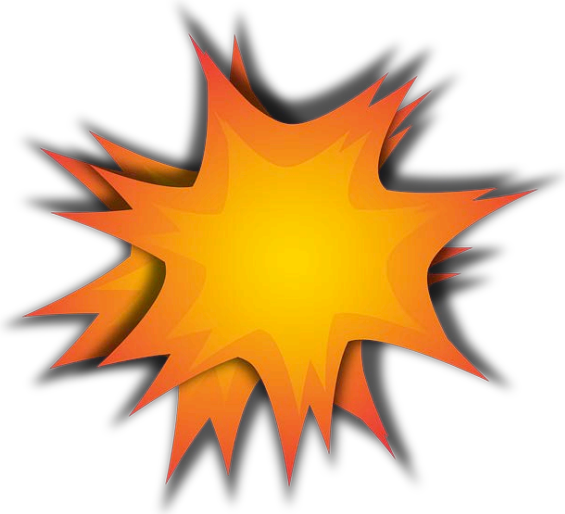
-Todos piensan que los demás lo están haciendo

-Y por ello todos dicen que lo hacen

https://www.facebook.com/dan.ariely/posts/904383595868

Duke University, Professor of Psychology and Behavioral Economics

¿Una Cuestión de Tamaño?



Unidad	Abreviatura	Equivalencia
Byte/Octeto	B	8 bits
Kilobyte	KB	1024 bytes
Megabyte	MB	1024 KB
Gigabyte	GB	1024 MB
Terabyte	TB	1024 GB
Petabyte	PB	1024 TB
Exabyte	EB	1024 PB
Zettabyte	ZB	1024 EB
Yottabyte	YB	1024 ZB
Brontobyte	BB	1024 YB
Geopbyte	GeB	1024 BB
…	…	…

¿Una Cuestión de Tamaño?

Según IDC (2014):

En 2013 teníamos 4.4 ZB

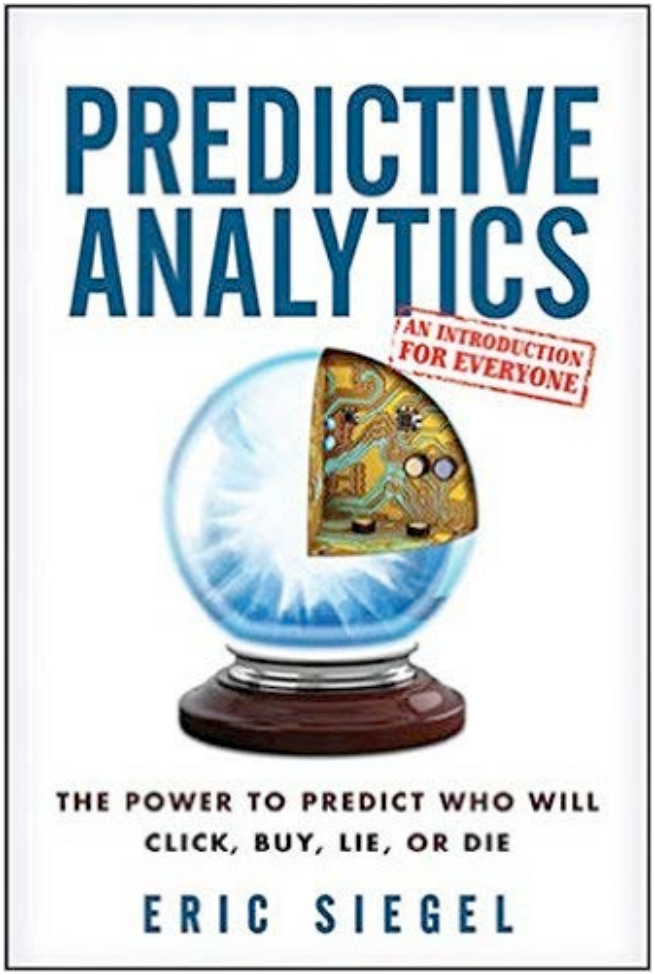
En 2020 se espera que tengamos 44 ZB

Los datos del universo digital se duplican cada 2 años

The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, April 2014, IDC
http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm
http://www.emc.com/leadership/digital-universe/2014iview/digital-universe-of-opportunities-vernon-turner.htm

¿Una Cuestión de Tamaño?

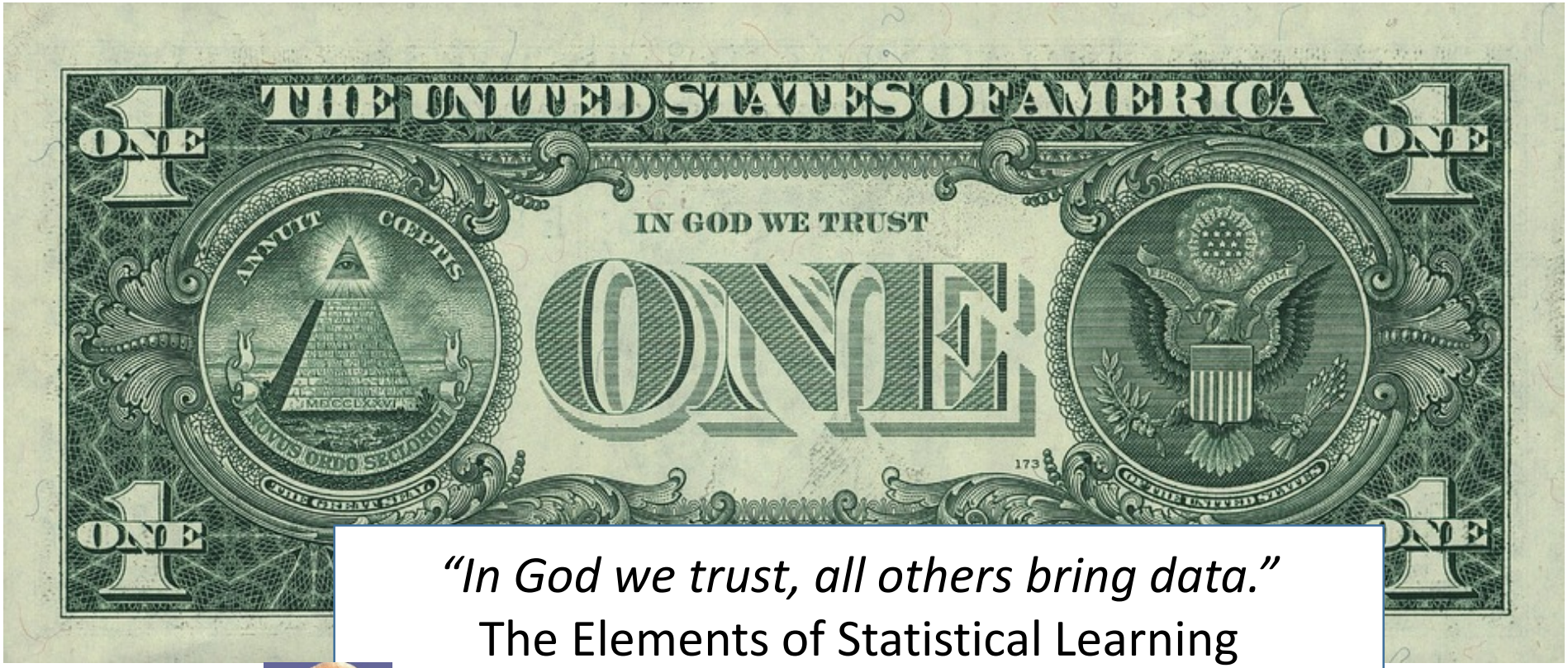
“Big Data does not exist. […] What’s exciting about data isn’t how much of it there is but how quickly it is growing.”



https://www.amazon.es/Predictive-Analytics-Power-Predict-Click/dp/1118356853

¿Es esto Big Data?: https://www.andertoons.com/math/cartoon/6517/does-this-count-as-big-data

Lo Realmente Importante: Los Datos



"In God we trust, all others bring data."
The Elements of Statistical Learning
William Edwards Deming (1900-1993)

Áreas del Big Data

1. HPC e infraestructura
2. Gestión de datos (BD, DW)
3. Minería de datos
4. Captura e integración de datos
5. Dominio del problema

Big Data, un diamante con muchas aristas:
http://www.philipchircop.com/post/25783275888/seeing-the-full-elephant-its-a-tree-its-a

Las 3 V's

- Volumen
- Velocidad
- Variedad

IBM

"Understanding Big Data", Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, IBM, **2012**.

https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=500016891&S_CPM=is_bdebook1_biginsightsfp

Las 3 V's

- Volumen
- Velocidad
- Variedad

"3D Data Management: Controlling Data Volume, Velocity, and Variety", Doug Laney, Application Delivery Strategies, META Group Inc., **2001**.
http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Las 4 V’s

- Volumen
- Velocidad
- Variedad
- Veracidad

Las tareas de preparación de datos son las que más tiempo consumen en una estrategia de *customer analytics*

(The Next Generation of Customer Analytics, Ventana Research, 2014)

Las 5 V's

- Volumen
- Velocidad
- Variedad
- Veracidad
- Valor → *smart data*

Demchenko, Y.; Grosso, P.; De Laat, C.; Membrey, P., "Addressing big data issues in Scientific Data Infrastructure," 2013 International Conference on Collaboration Technologies and Systems (CTS), pp.48,55, 20-24 May 2013

Las 6 V’s

- Volumen
- Velocidad
- Variedad
- Veracidad
- Validez
- Volatilidad

Inderpal Bhandar (Chief Data Officer, Express Scripts), Big Data Innovation Summit 2013 (Boston)

Y Así Sucesivamente…

- Volumen
- Velocidad
- Variedad
- Veracidad
- Validez
- Volatilidad
- Valor
- Visualización

Y Así Sucesivamente…

- Volumen
- Velocidad
- Variedad
- Veracidad
- Validez
- Volatilidad
- Valor
- Visualización
- …

Resumiendo…

Según Samuel Madden, las carácterísticas más importantes son:

-Volumen

-Velocidad

-Variedad

-Vexing

Samuel Madden (MIT), “Going Big On Big Data”, 2015

Resumiendo…

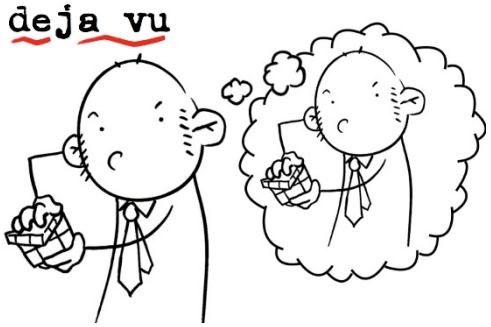
Según Michael Stonebraker, Big Data puede significar al menos una de estas 4 cosas:

- Grandes volúmenes de datos pero “análisis pequeños”
  - -SQL analítico sobre muchos datos
- Grandes volúmenes de datos y “análisis grandes”
  - -Minería de datos con muchos datos
- Gran velocidad
  - -Datos en *streaming*, procesamiento en tiempo real
- Gran variedad
  - -ETL con muchas fuentes de datos

What Does 'Big Data' Mean?, Michael Stonebraker, BLOG@CACM, September 21, 2012, http://cacm.acm.org/blogs/blog-cacm/155468-what-does-big-data-mean/fulltext

¿Es Esto Realmente Nuevo?

- *International Conference on Very Large Data Bases (VLDB)* → primera edición en 1975 (edición 42 en 2016)
- *Data streams* → término de finales de los 90, estudio intensivo en DSMSs a partir del 2000
- Integración de datos → a comienzos de los 80 ya se comienza a investigar en integración de BD heterogéneas
- …



Entonces, ¿Qué Ha Cambiado?

- Capacidad de almacenamiento
- Capacidad de procesamiento
- Computación en la nube (*cloud computing*)
- Fuentes de datos

Fuentes de Datos





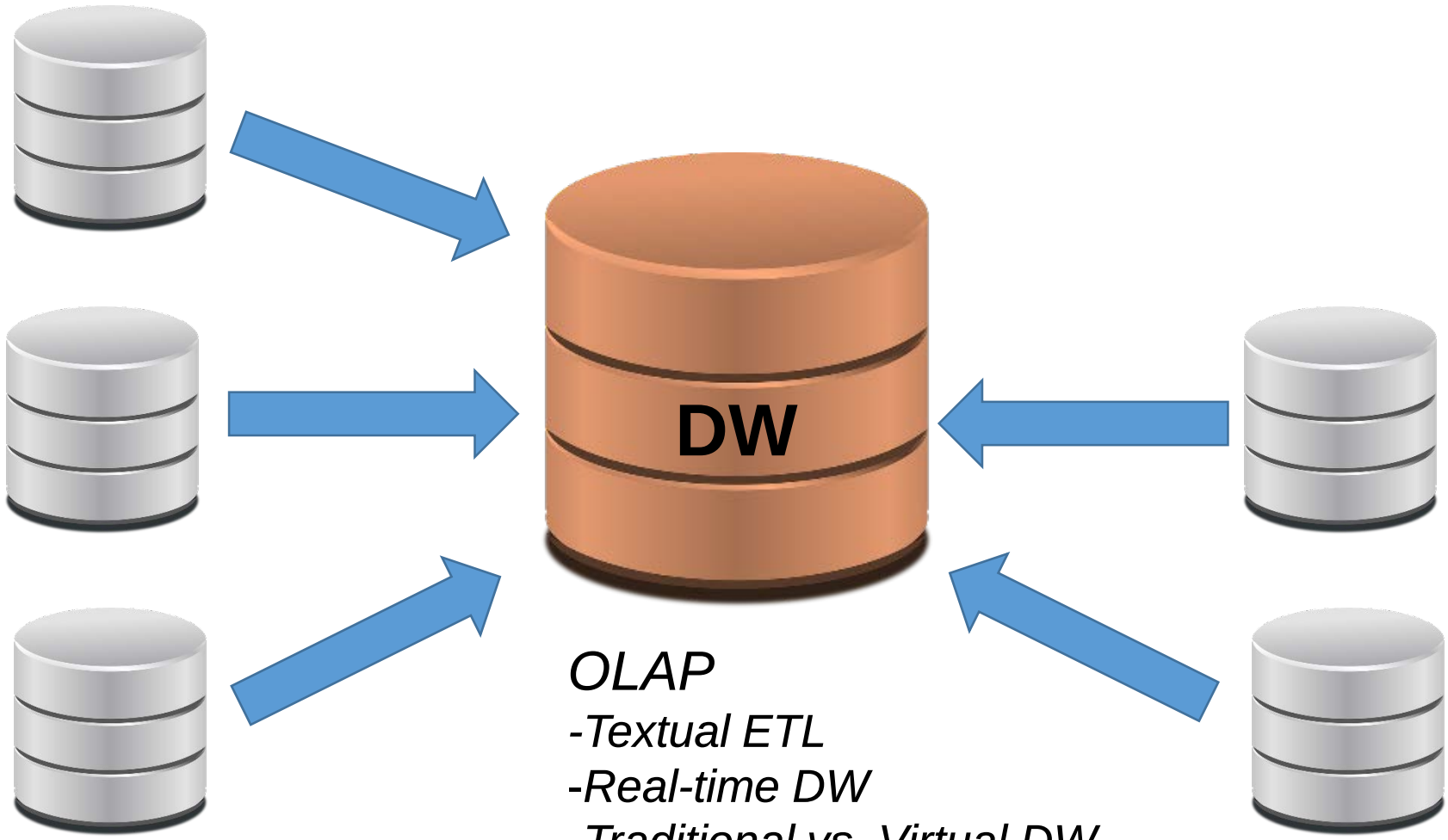
Fuentes de Datos

-BD relacionales, NoSQL, NewSQL

-BD distribuidas
-BD federadas

-Distintos propósitos (OLTP)

Fuentes de Datos



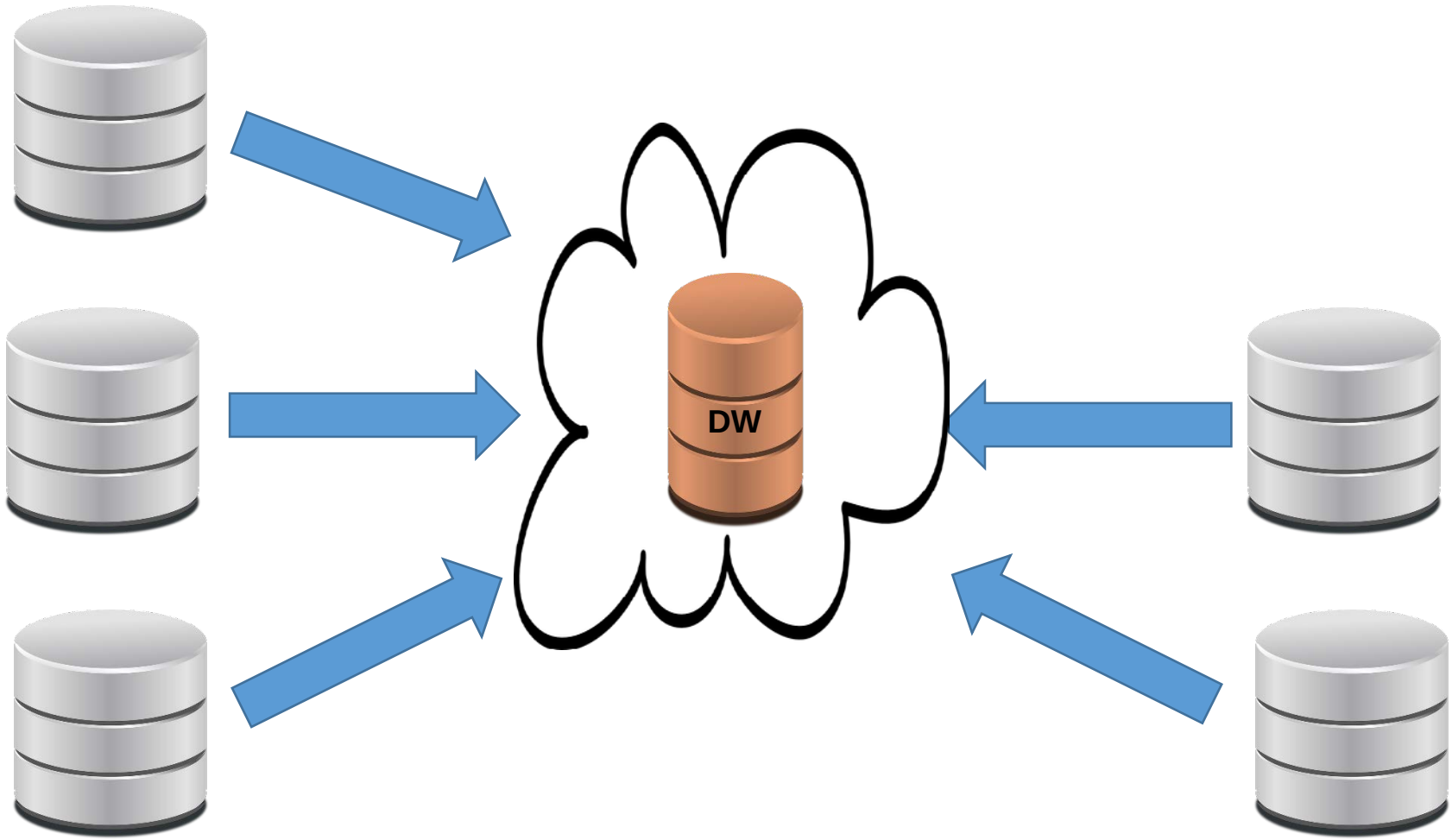
OLAP

-Textual ETL

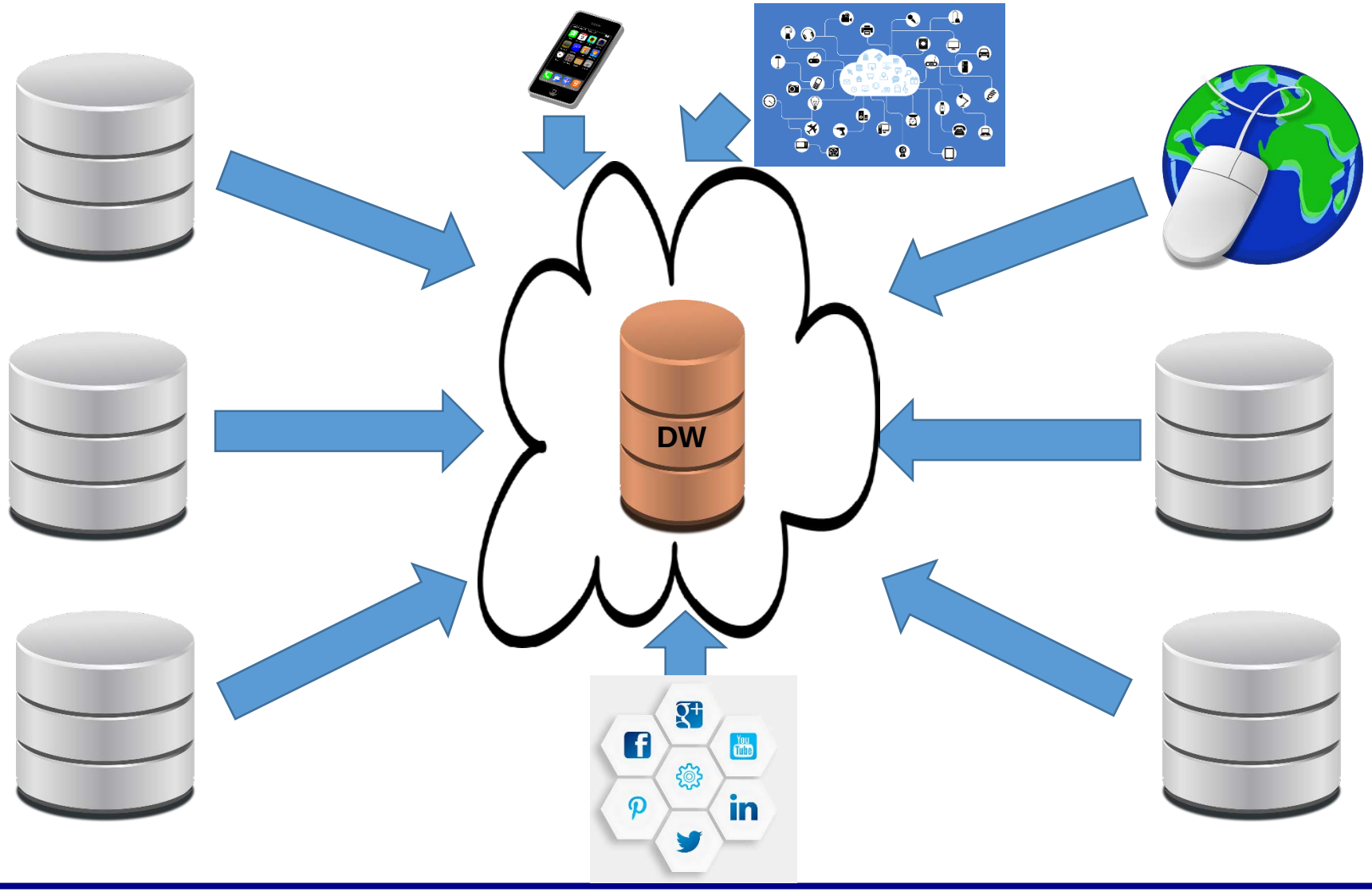
-Real-time DW

-Traditional vs. Virtual DW

Fuentes de Datos



Fuentes de Datos



Ejemplo: Computación Móvil

- Multitud de dispositivos y sensores
- *Fog Computing / Edge Computing / Mobile Cloud Computing*

  -Explotar dispositivos cercanos, cyber-foraging (¿agentes móviles?), cloudlets
- *Vehicular clouds*
- *Mobile crowdsensing*

Big Data y *Business Intelligence*

Business Intelligence (BI):

- Combinación de:
  - -Datos
  - -Tecnología: DW, herramientas de BI
  - -Analítica
  - -Conocimiento humano
- Para optimizar decisiones de negocio

Big Data y *Business Intelligence*

Big Data en relación al BI:

- Nuevas fuentes de datos: sensores, redes sociales, …
- Mayor volumen de datos y mayores requerimientos
- Nuevas tecnologías: procesamiento paralelo, analítica predictiva, …
- Nueva combinación de habilidades → científico de datos

Aproximaciones

- *Data Mining / Knowledge Discovery from Databases*

  -Análisis descriptivo

  -Aplicaciones: segmentación clientes, análisis del carrito de la compra, …

- *Predictive Analytics / Machine Learning*

  -Análisis predictivo

  -Aplicaciones: predicción micro (probabilidad de que algo suceda), …

- *Data Science*

  -Análisis prescriptivo

  -Aplicaciones: sistemas de recomendación, …

- *Big Data Landscape*

  http://mattturck.com/big-data-landscape-2016-v18-final/

El Científico de Datos

Ingeniero de datos con mente analítica

Steven Hillion (vicepresidente de Analytics en EMC Greenplum)

Ingeniero que emplea el método científico para extraer información de los datos

Gil Press (Managing Partner en gPress)

-Capturar y gestionar grandes volúmenes datos → ingeniería

-Extraer valor → estadística

-Presentar resultados → comunicación

John Rauser (ingeniero de Amazon)