17-20 sept 2013

CAEPIA'13

Madrid

Big Data

Francisco Herrera

Research Group on Soft Computing and Information Intelligent Systems (SCI²S)

Dept. of Computer Science and A.I.

University of Granada, Spain

Email: herrera@decsai.ugr.es

http://sci2s.ugr.es

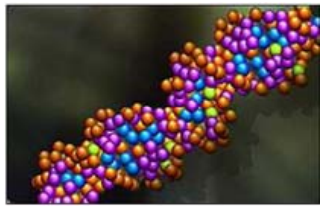
DECSAI

Universidad de Granada

Big Data

Nuestro mundo gira en torno a los datos

- Ciencia
  - Bases de datos de astronomía, genómica, datos medio-ambientales, datos de transporte, …



- Ciencias Sociales y Humanidades
  - Libros escaneados, documentos históricos, datos sociales, …

- Negocio y Comercio
  - Ventas de corporaciones, transacciones de mercados, censos, tráfico de aerolíneas, …
- Entretenimiento y Ocio
  - Imágenes en internet, películas, ficheros MP3, …

- Medicina
  - Datos de pacientes, datos de escaner, radiografías …
- Industria, Energía, …
  - Sensores, …

Big Data

ELMUNDO.es
Líder mundial en español | Miércoles 04/09/2013. Actualizado 16:27h.

Alex ' Sandy' Pentland, director del programa de emprendedores del 'Media Lab' del Massachusetts Institute of Technology (MIT)

INTERNET | Campus Party Europa 2013

'Es la década de los datos y de ahí vendrá la revolución'

Alex 'Sandy' Pentland, durante su ponencia. | M. Sáinz

Considerado por 'Forbes' como uno de los siete científicos de datos más poderosos del mundo

http://www.elmundo.es/elmundo/2013/09/03/navegante/1378243782.html

Outline

- ¿Qué es Big Data?
- MapReduce: Paradigma de Programación para Big Data (Google)
- Plataforma Hadoop (Open access)
- Librería Mahout para Big Data. Otras librerías
- Limitaciones de MapReduce
- Comentarios Finales

Outline

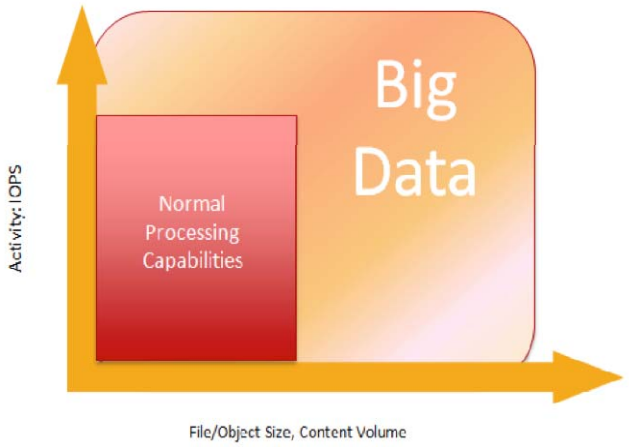
- **¿Qué es Big Data?**
- MapReduce: Paradigma de Programación para Big Data (Google)
- Plataforma Hadoop (Open access)
- Librería Mahout para Big Data. Otras librerías
- Limitaciones de MapReduce
- Comentarios Finales

¿Qué es Big Data?

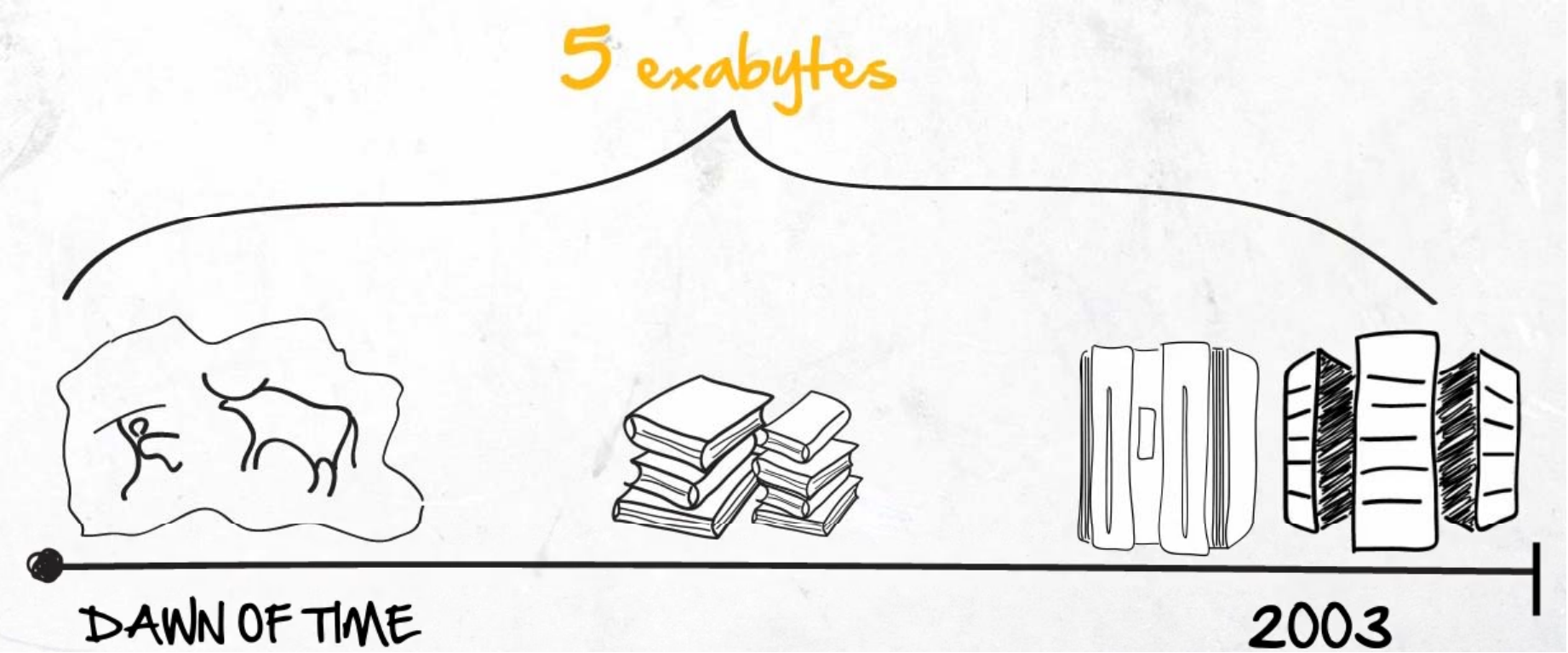
No hay una definición estándar

Big data es una colección de datos grande, complejos, **muy difícil de procesar a través de herramientas de gestión y procesamiento de datos tradicionales**

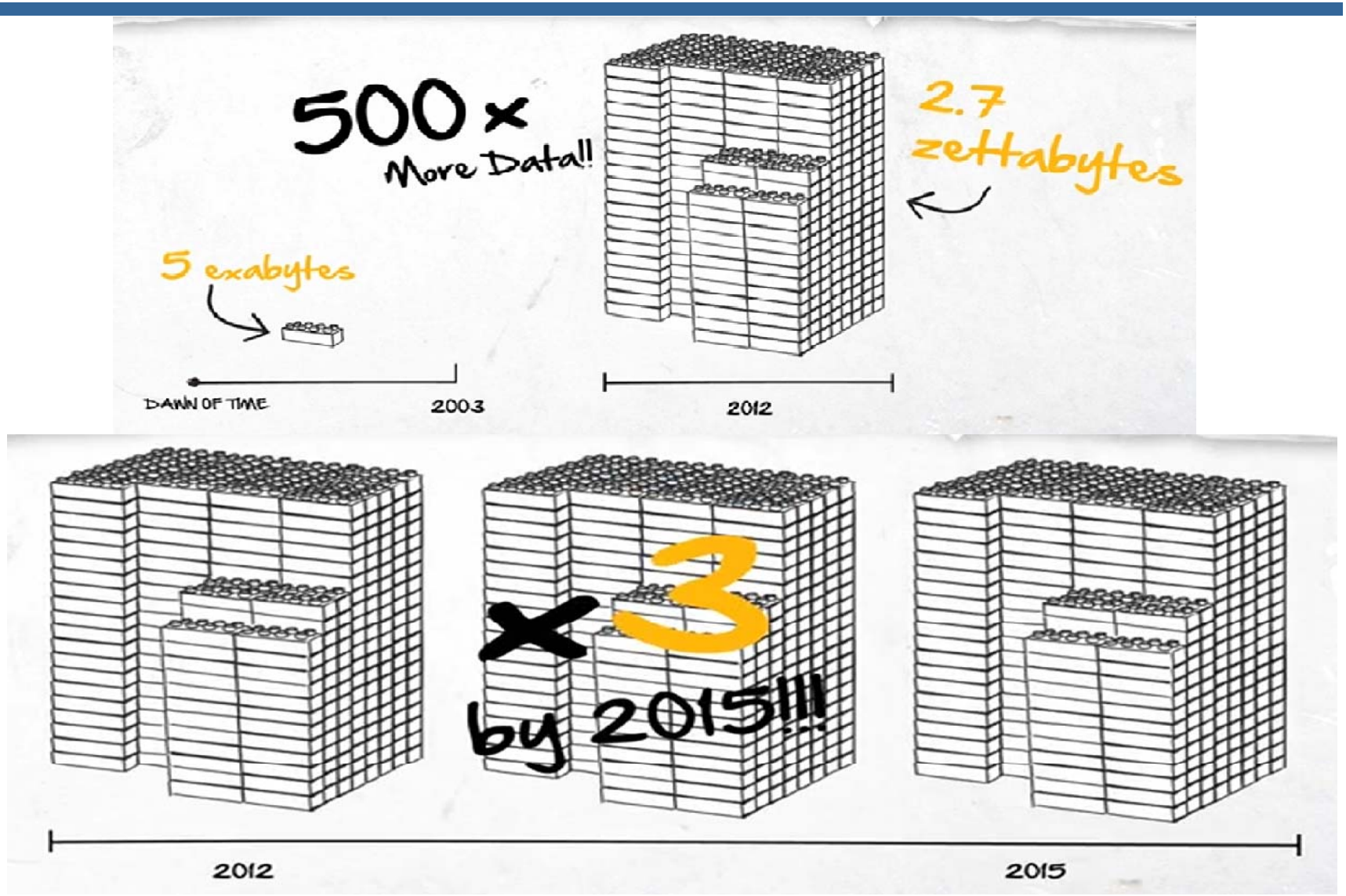
"***Big Data***" son datos cuyo volumen, diversidad y complejidad **requieren nueva arquitectura, técnicas, algoritmos y análisis** para gestionar y extraer valor y conocimiento oculto en ellos …



¿Qué es Big Data?



¿Qué es Big Data?



¿Qué es Big Data? **3 V's de Big Data**

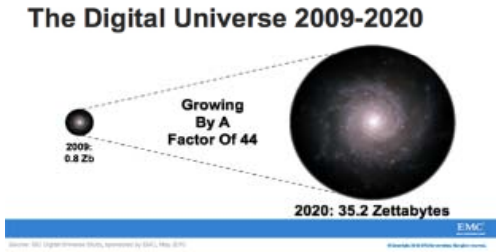
¿Qué es Big Data? 3 V's de Big Data

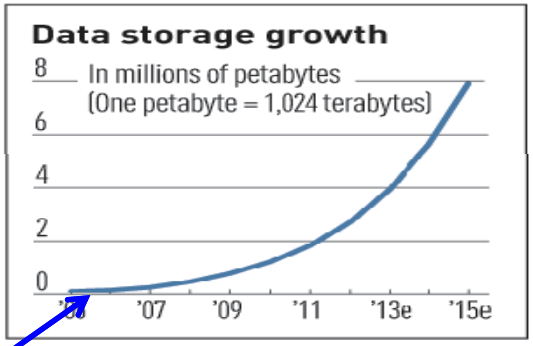
1ª:Volumen

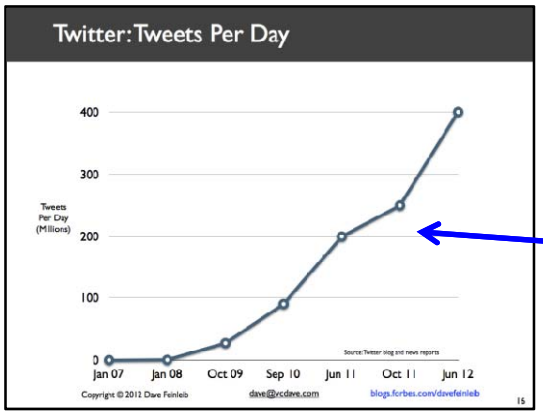
El volumen de datos

crece exponencialmente

- Crecimiento x 44 de 2009 a 2020
- De 0.8 zettabytes a 35ZB

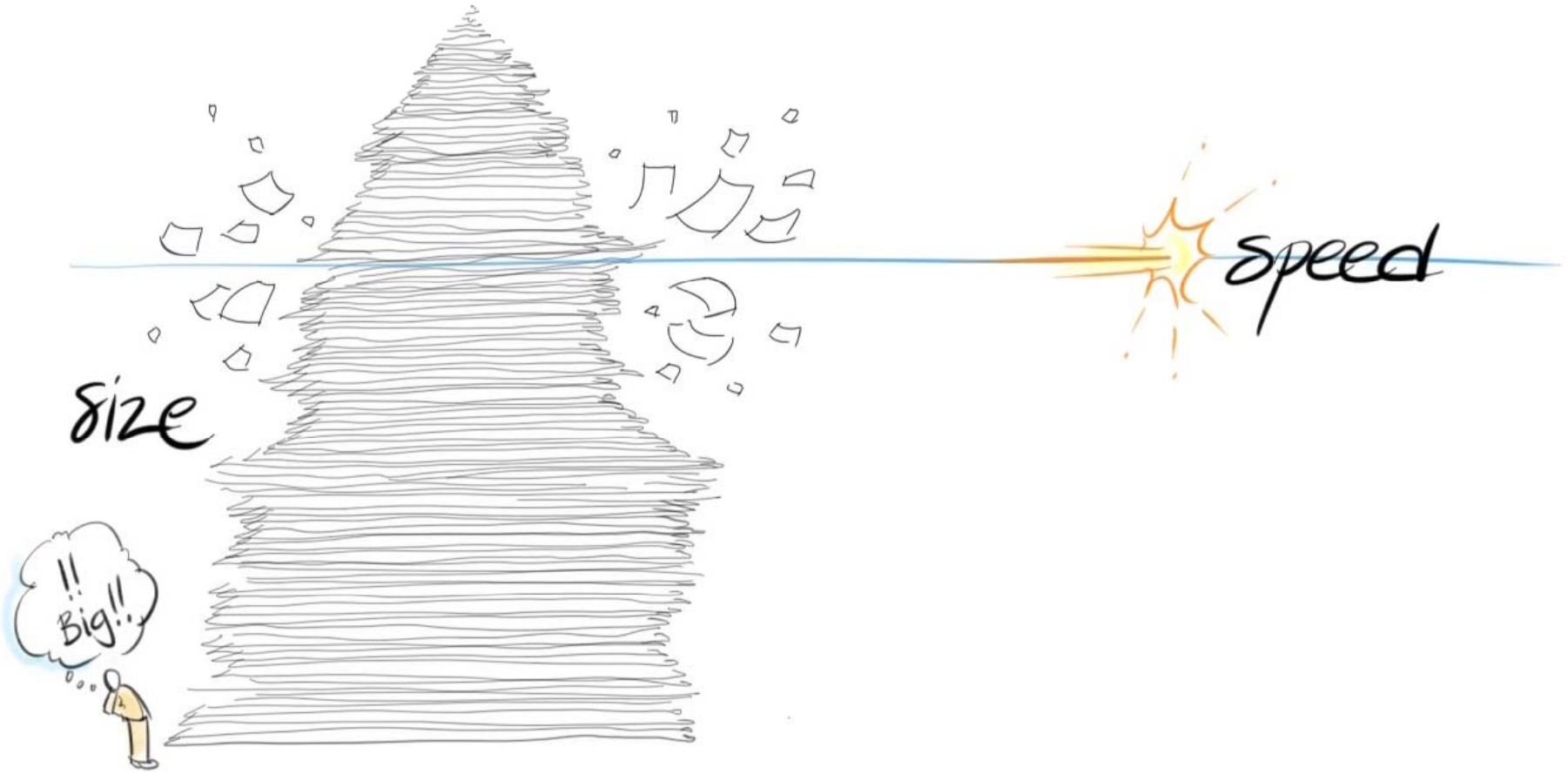






Crecimiento exponencial en los datos generados/almacenados

¿Qué es Big Data? **3 V's de Big Data**



¿Qué es Big Data? 3 V's de Big Data

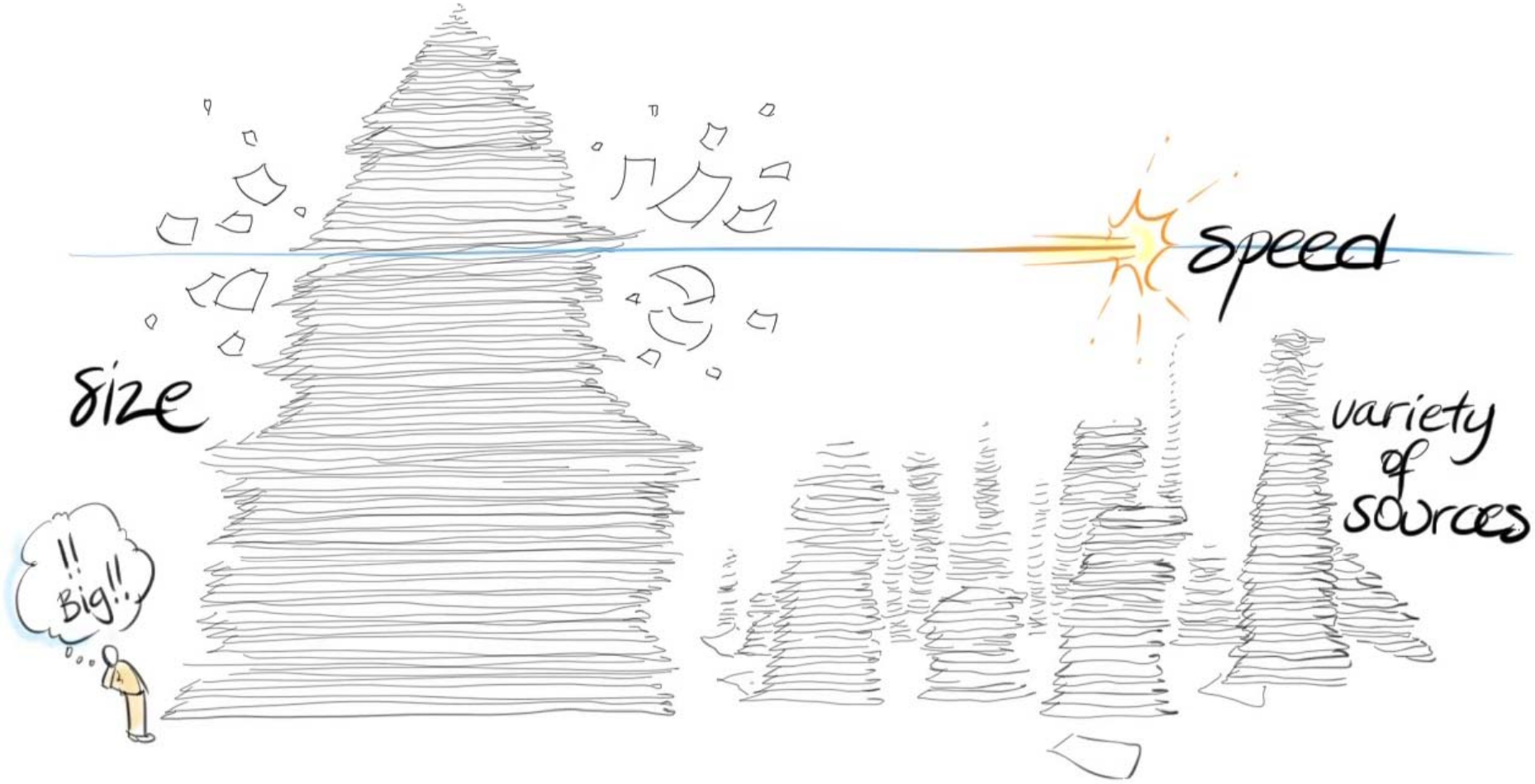
2ª:Velocidad

- **Los DATOS se generan muy rápido y necesitan ser procesados rápidamente**
- **Online Data Analytics**
- **Decisiones tardías ➔ oportunidades perdidas**

Ejemplos:

- **E-Promociones:** Basadas en la posición actual e historial de compra ➔ envío de promociones en el momento de comercios cercanos a la posición
- **Monitorización / vigilancia sanitaria:** Monitorización sensorial de las actividades del cuerpo ➔ cualquier medida anormal requiere una reacción inmediata

¿Qué es Big Data? **3 V's de Big Data**

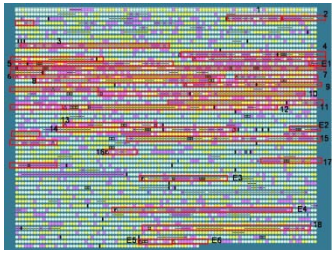


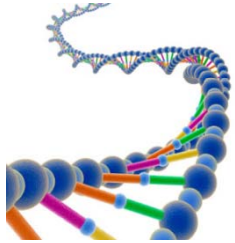
¿Qué es Big Data? 3 V's de Big Data

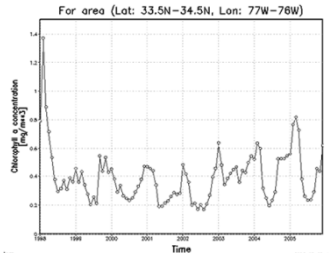
3ª:Variedad

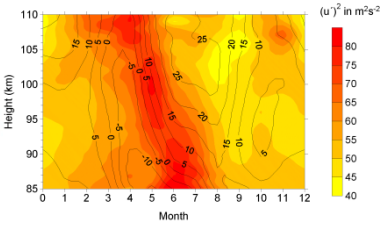
- Varios formatos y estructuras:

  Texto, numéricos, imágenes, audio, video, sequencias, series temporales …
- Una sola aplicación puede generar muchos tipos de datos



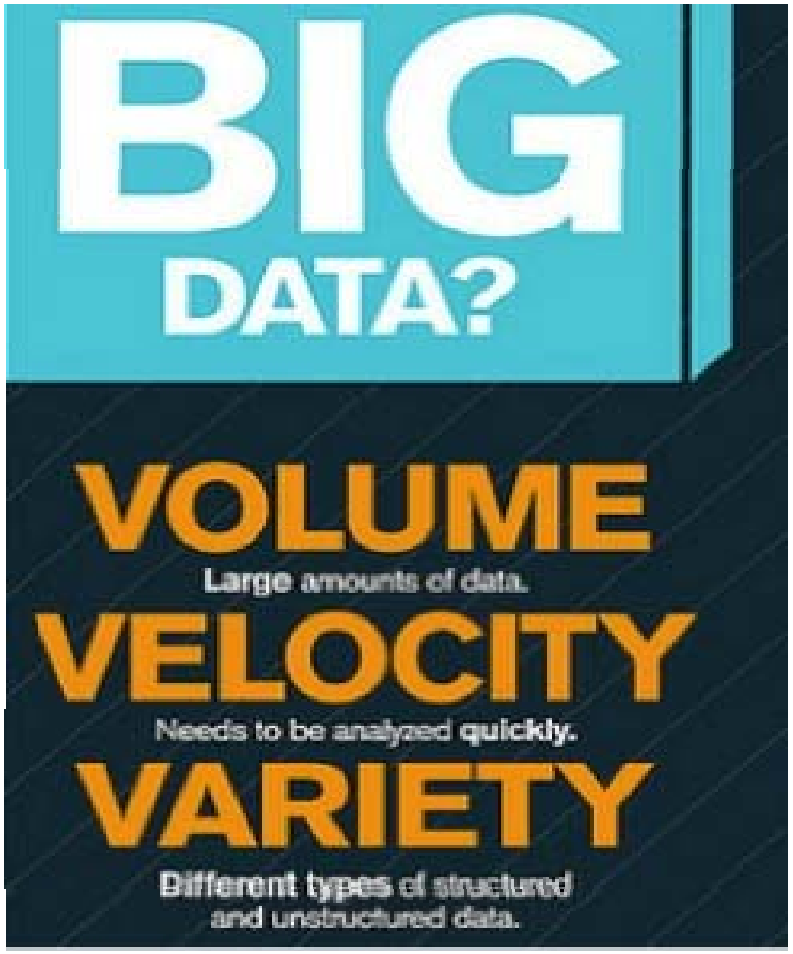






Extracción de conocimiento ➔
Todos estos tipos de datos necesitan ser analizados conjuntamente

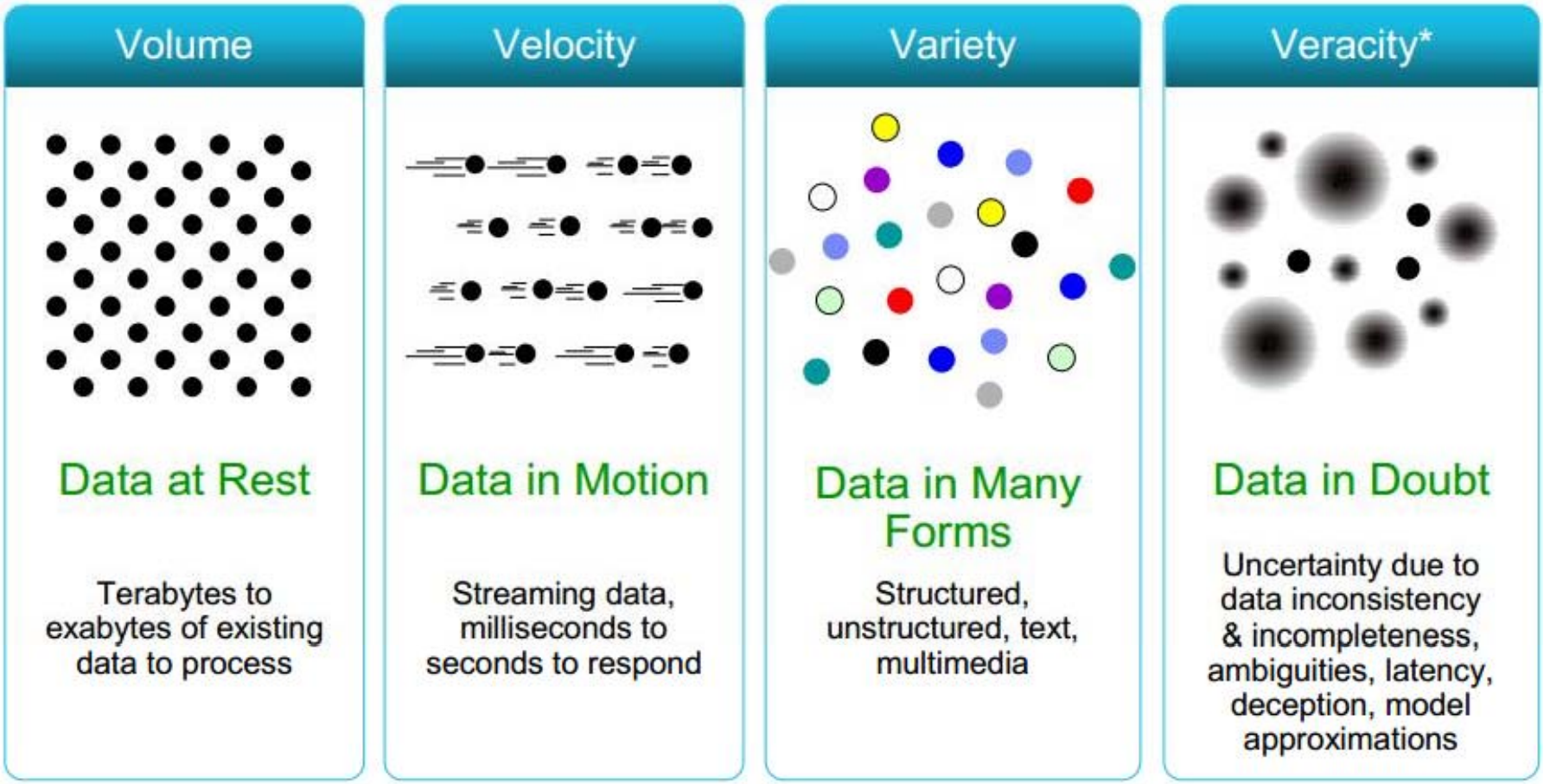
¿Qué es Big Data? **3 V's de Big Data**



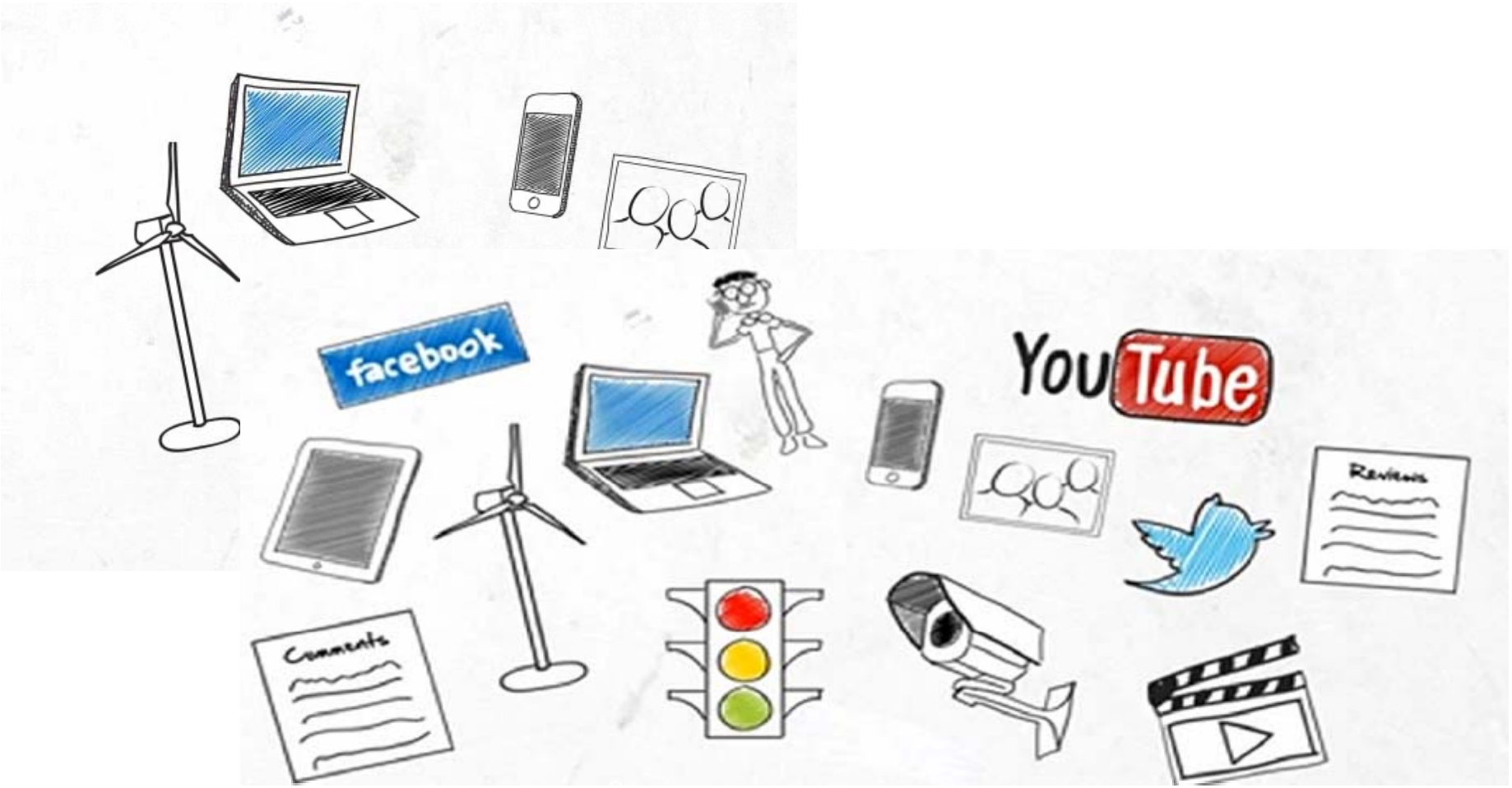
¿5V's?

¿Qué es Big Data? 3 V's de Big Data

4ª V
Veracidad



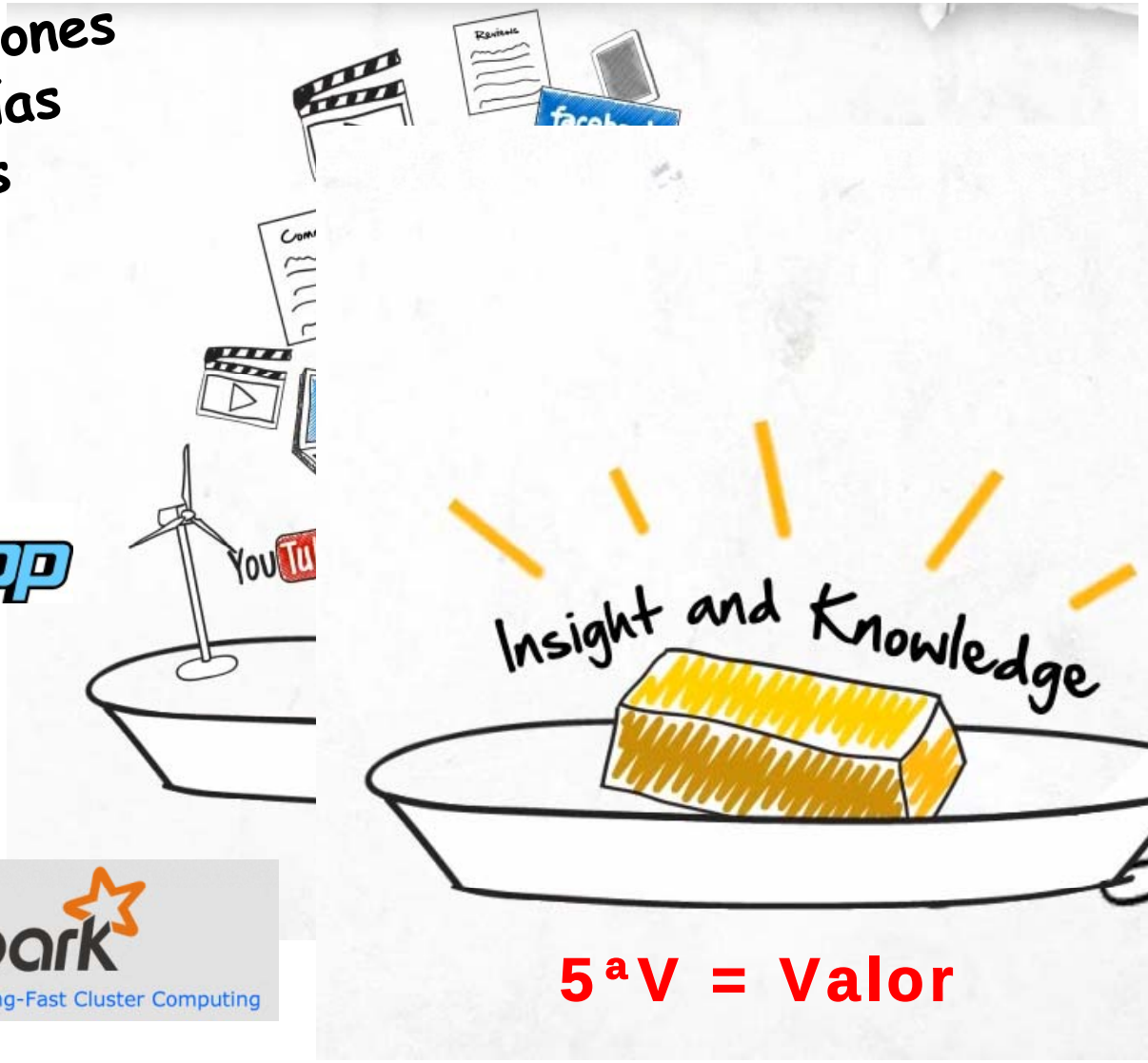
¿Qué es Big Data?



Datos no estructurados

¿Qué es Big Data?

Aproximaciones y tecnologías innovativas



MapReduce

hadoop

mahout

APACHE GIRAPH

Spark
Lightning-Fast Cluster Computing

¿Qué es Big Data?

¿Quién genera Big Data?

Redes sociales y multimedia
(todos generamos datos)

Instrumentos científicos
(colección de toda clase de datos)

Dispositivos móviles
(seguimiento de objetos)

Redes de sensores
(se miden toda clase de datos)

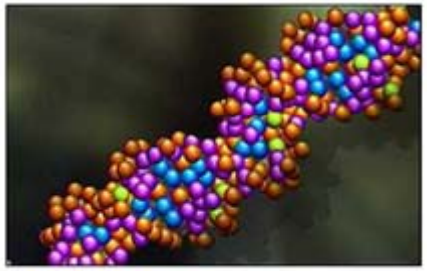
El progreso y la innovación ya no se ven obstaculizados por la capacidad de recopilar datos, sino por la capacidad de gestionar, analizar, sintetizar, visualizar, y descubrir el conocimiento de los datos recopilados de manera oportuna y en una forma escalable

Big Data. **Aplicaciones**

Astronomía

- Astronomical sky surveys
- 120 Gigabytes/week
- 6.5 Terabytes/year

Genómica



- 25,000 genes in human genome
- 3 billion bases
- 3 Gigabytes of genetic data

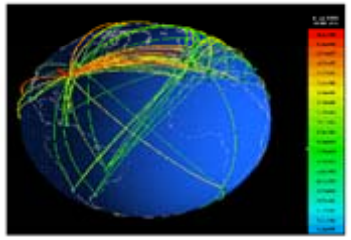
Telefonía

- 250M calls/day
- 60G calls/year
- 40 bytes/call
- 2.5 Terabytes/year

Transacciones de tarjetas de crédito

- 47.5 billion transactions in 2005 worldwide
- 115 Terabytes of data transmitted to VisaNet data processing center in 2004

Tráfico en Internet



Traffic in a typical router:

- 42 kB/second
- 3.5 Gigabytes/day
- 1.3 Terabytes/year

Procesamiento de información WEB

- 25 I
- 10k
- 250
- "De
  tim

Outline

- ¿Qué es Big Data?
- **MapReduce: Paradigma de Programación para Big Data (Google)**
- Plataforma Hadoop (Open access)
- Librería Mahout para Big Data. Otras librerías
- Limitaciones de MapReduce
- Comentarios Finales

MapReduce

- **Problema:** Escalabilidad de grandes cantidades de datos
- **Ejemplo:**
  - Exploración 100 TB en 1 nodo @ 50 MB/sec = 23 días
  - Exploración en un clúster de 1000 nodos = 33 minutos
- **Solución ➔ Divide-Y-Vencerás**

Una sola máquina no puede gestionar grandes volúmenes de datos de manera eficiente

MapReduce

- Escalabilidad de grandes cantidades de datos
  - Exploración 100 TB en 1 nodo @ 50 MB/sec = 23 días
  - Exploración en un clúster de 1000 nodos = 33 minutos

Solución ➔ Divide-Y-Vencerás

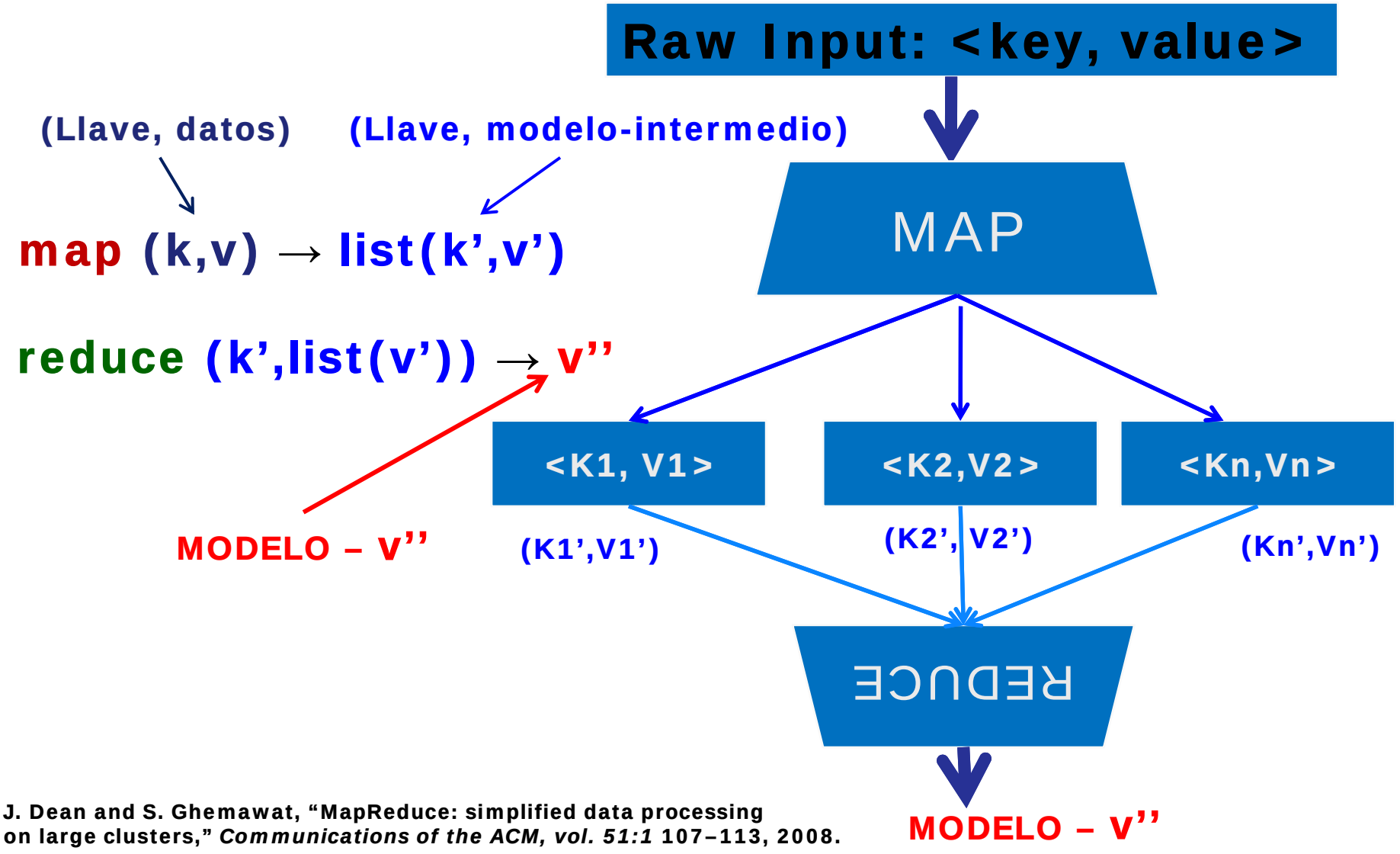
MapReduce

- Modelo de programación de datos paralela
- Concepto simple, elegante, extensible para múltiples aplicaciones

- **Creado por Google (2004)**
  - Procesa 20 PB de datos por día
- **Popularizado por el proyecto de codigo abierto Hadoop**
  - Usado por Yahoo!, Facebook, Amazon, …

MapReduce

Programming Framework



J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM, vol. 51:1* 107–113, 2008.

MapReduce

Características

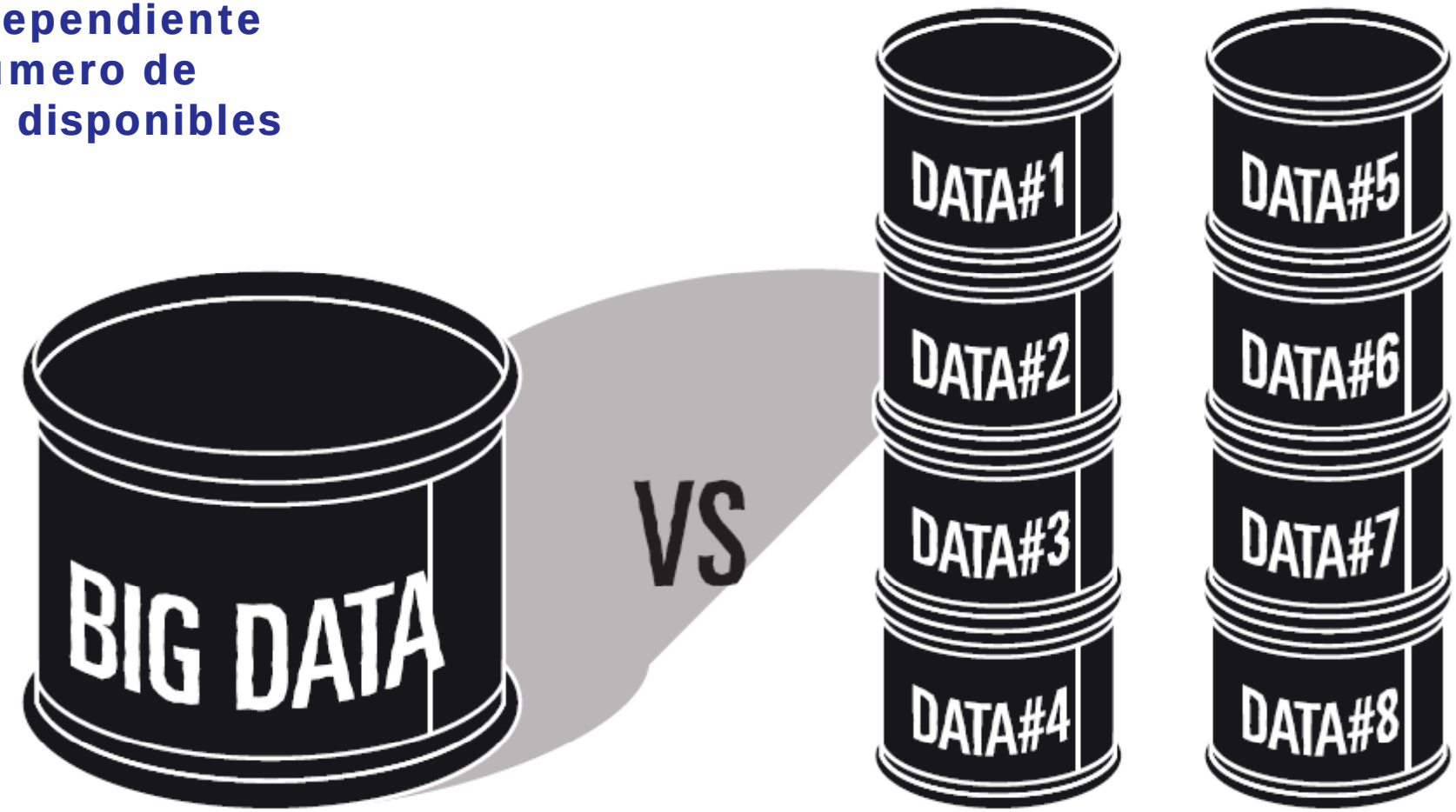
- **Paralelización automática:**
  - Dependiendo del tamaño de ENTRADA DE DATOS➔ se crean mutiples tareas MAP
  - Dependiendo del número de intermedio <clave, valor> particiones ➔ se crean tareas REDUCE
- **Escalabilidad:**
  - Funciona sobre cualquier cluster de nodos/procesadores
  - Puede trabajar desde 2 a 10,000 máquinas
- **Transparencia programación**
  - Manejo de los fallos de la máquina
  - Gestión de comunicación entre máquina

Características

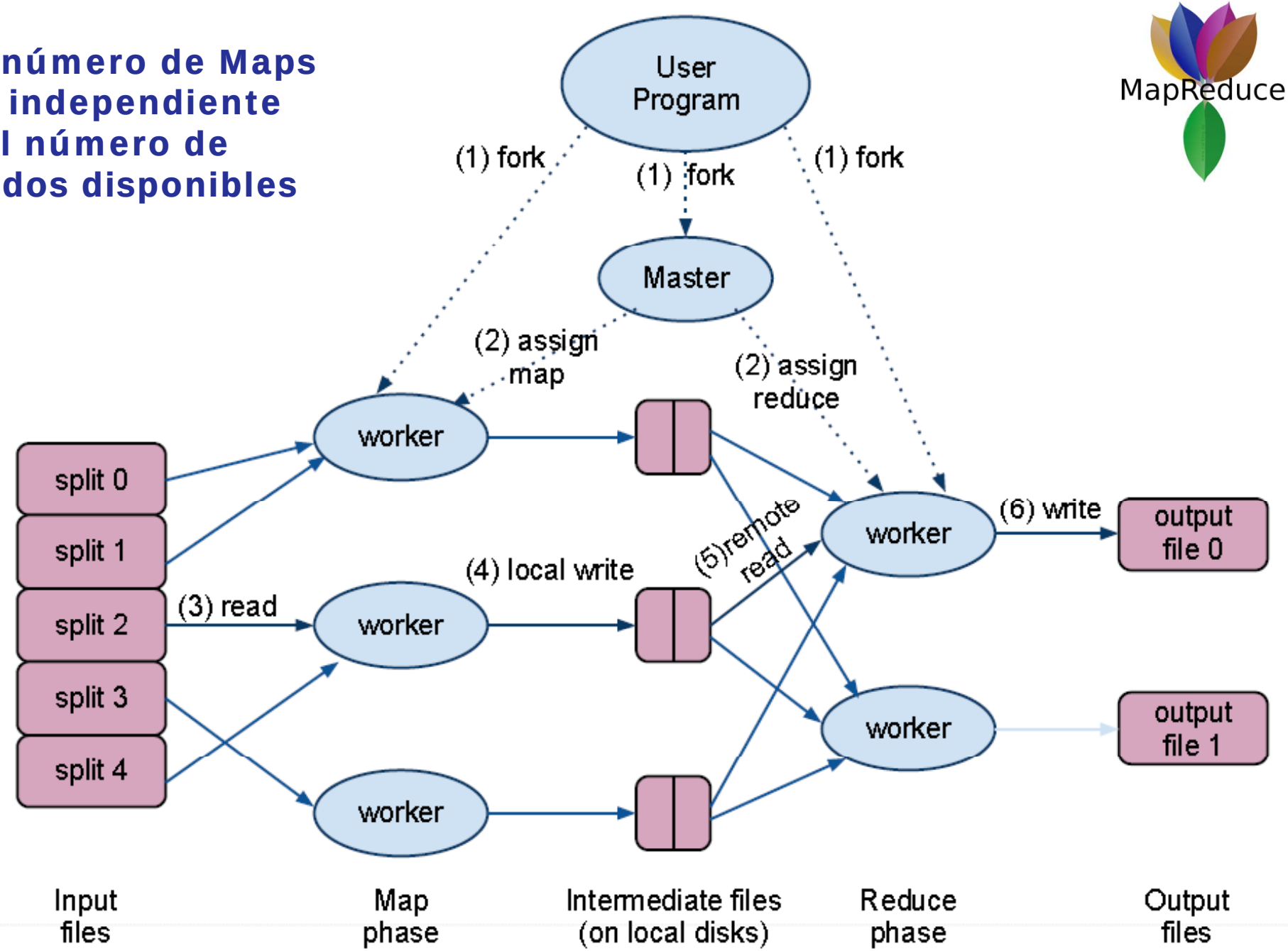
- **Tiempo de ejecución:**
  - Partición de datos
  - Programación de la tarea
  - Manejo de los fallos de la máquina
  - Gestión de comunicación entre máquina

MapReduce

El número de Maps es independiente del número de nodos disponibles



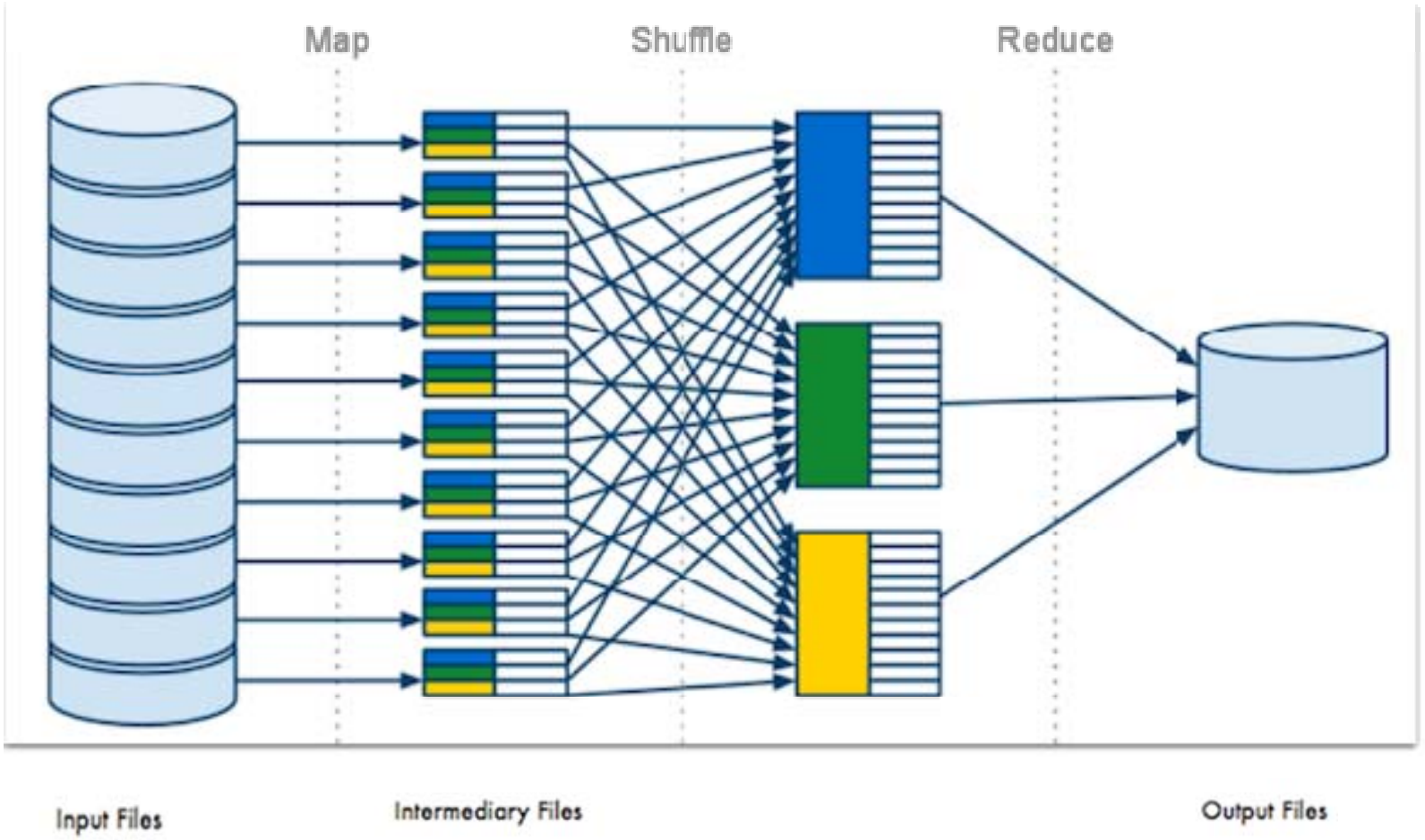
El número de Maps es independiente del número de nodos disponibles



Almacenamiento con copias, normalmente r=3

MapReduce

Completamente transparente para el programador



MapReduce

Resumiendo:

- **Ventaja frente a los modelos distribuidos clásicos:** El modelo de programación paralela de datos de MapReduce oculta la complejidad de la distribución y tolerancia a fallos
- **Claves de su filosofía**: Es
  - ***escalable:*** *se olvidan los problemas de hardware*
  - ***más barato:*** *se ahorran costes en hardware, programación y administración*
- **MapReduce no es adecuado para todos los problemas, pero cuando funciona, puede ahorrar mucho tiempo**

Outline

- ¿Qué es Big Data?
- MapReduce: Paradigma de Programación para Big Data (Google)
- **Plataforma Hadoop (Open access)**
- Librería Mahout para Big Data. Otras librerías
- Limitaciones de MapReduce
- Comentarios Finales

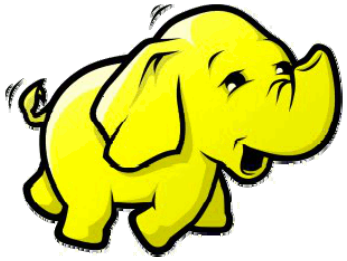
Hadoop

Hadoop es una implementación de código abierto del paradigma computacional MapReduce

http://hadoop.apache.org/

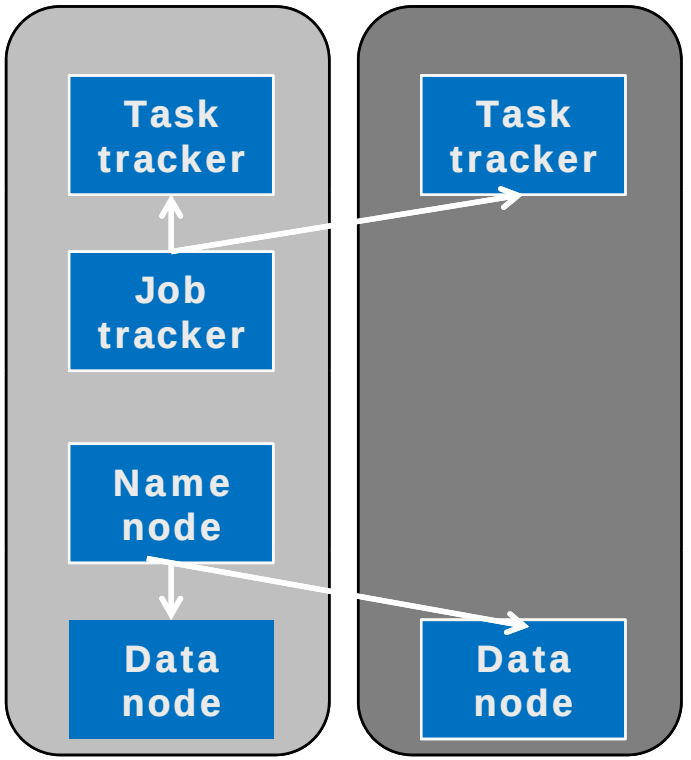
Hadoop

Hadoop Distributed File System (HDFS) es un sistema de archivos distribuido, escalable y portátil escrito en Java para el framework Hadoop



Map Reduce Layer

HDFS Layer



Creado por Doug Cutting (chairman of board of directors of the Apache Software Foundation, 2010)

http://hadoop.apache.org/

Hadoop

http://sortbenchmark.org/

Primer hito de Hadoop: July 2008 - Hadoop Wins Terabyte Sort Benchmark

Uno de los grupos de Yahoo Hadoop ordenó 1 terabyte de datos en 209 segundos, superando el récord anterior de 297 segundos en la competición anual de ordenación de un terabyte (Daytona). Esta es la primera vez que un programa en Java de código abierto ganó la competición.

2008, 3.48 minutes

Hadoop

910 nodes x (4 dual-core processors, 4 disks, 8 GB memory)

Owen OMalley, Yahoo

2007, 4.95 min

TokuSampleSort

tx2500 disk cluster

400 nodes x (2 processors, 6-disk RAID, 8 GB memory)

Bradley C. Kuszmaul , MIT

	Daytona
Gray	2013, 1.42 TB/min **Hadoop** 102.5 TB in 4,328 seconds 2100 nodes x (2 2.3Ghz hexcore Xeon E5-2630, 64 GB memory, 12x3TB disks) Thomas Graves Yahoo! Inc.

http://developer.yahoo.com/blogs/hadoop/hadoop-sorts-petabyte-16-25-hours-terabyte-62-422.html

Hadoop

http://hadoop.apache.org/

¿Qué es Apache Hadoop?

Apache™ Hadoop® es un proyecto que desarrolla software de código abierto fiable, escalable, para computación distribuida

Hadoop se puede ejecutar de tres formas distintas (configuraciones):

1. **Modo Local / *Standalone.*** **Se ejecuta en una única JVM (*Java Virtual Machine*). *Esto es útil para depuración***

2. **Modo Pseudo-distribuido** **(simulando así un clúster o sistema distribuido de pequeña escala)**

3. **Distribuido (Clúster)**

Hadoop

¿Cómo accedo a una plataforma Hadoop?

Plataformas Cloud con instalación de Hadoop

Amazon Elastic Compute Cloud (Amazon EC2)

http://aws.amazon.com/es/ec2/

Windows Azure

http://www.windowsazure.com/

Instalación en un cluster privado

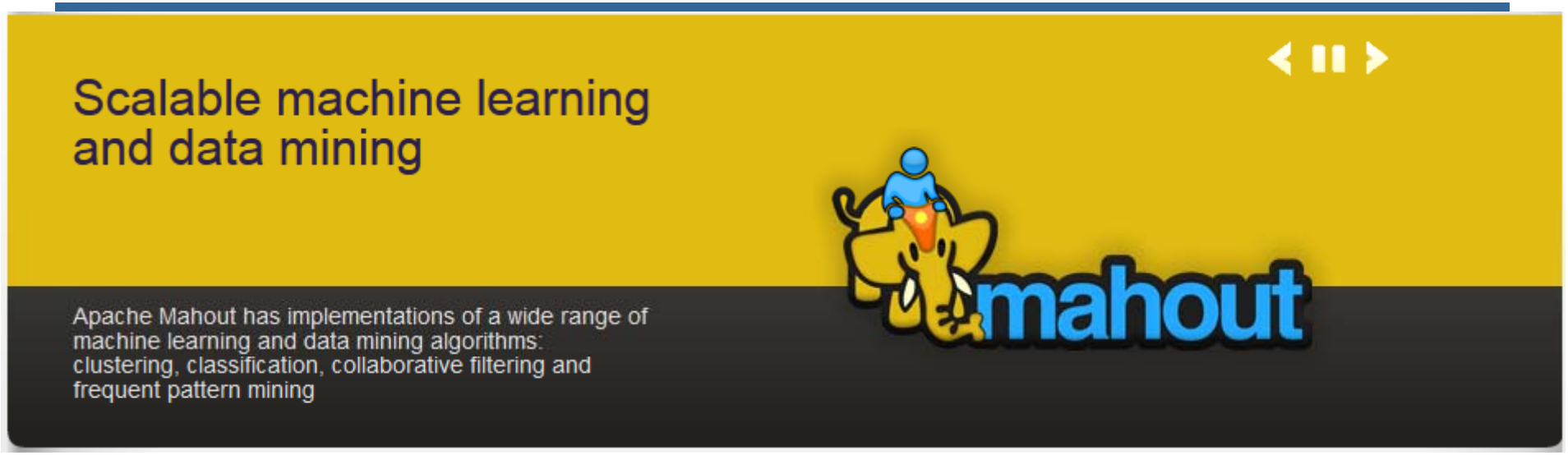
Ejemplo: Cluster ATLAS, 12 nodos (UGR)
-Microprocessors: 2 x Intel E5-2620 (6 cores/12 threads, 2 GHz)
- RAM 64 GB DDR3 ECC 1600MHz

¿Cómo puedo instalar Hadoop?

Outline

- ¿Qué es Big Data?
- MapReduce: Paradigma de Programación para Big Data (Google)
- Plataforma Hadoop (Open access)
- **Librería Mahout para Big Data. Otras librerías**
- Limitaciones de MapReduce
- Comentarios Finales

Mahout



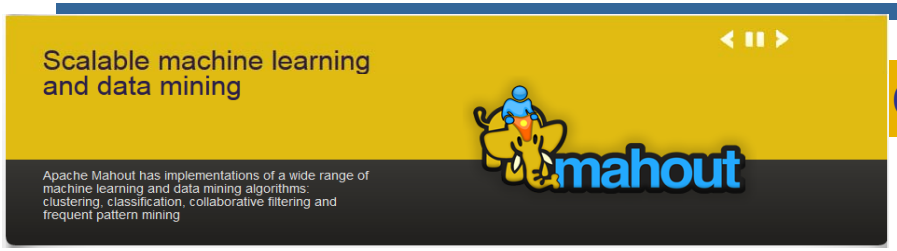
Mahout currently has

- Collaborative Filtering
- User and Item based recommenders
- K-Means, Fuzzy K-Means clustering
- Mean Shift clustering
- Dirichlet process clustering
- Latent Dirichlet Allocation
- Singular value decomposition
- Parallel Frequent Pattern mining
- Complementary Naive Bayes classifier
- Random forest decision tree based classifier
- High performance java collections (previously colt collections)
- A vibrant community
- and many more cool stuff to come by this summer thanks to Google summer of code

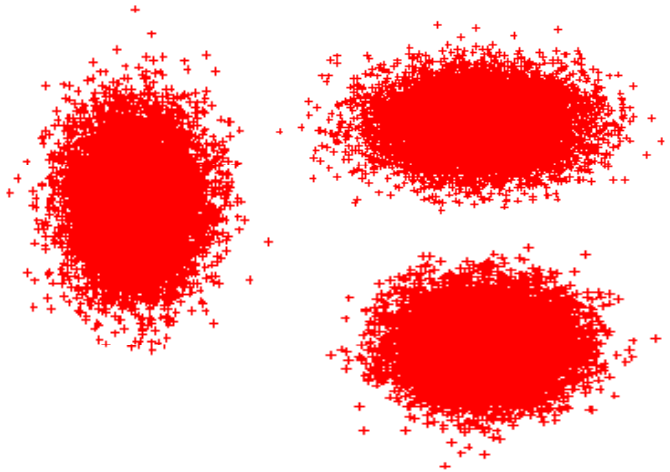
Biblioteca de código abierto en APACHE

http://mahout.apache.org/

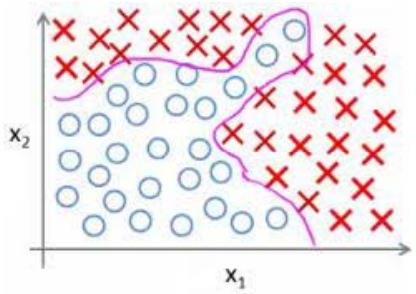
Mahout



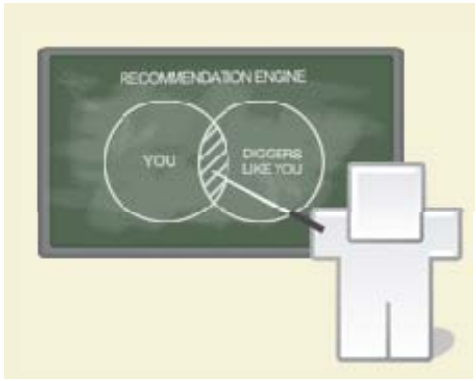
Cuatro grandes áreas de aplicación



Agrupamiento

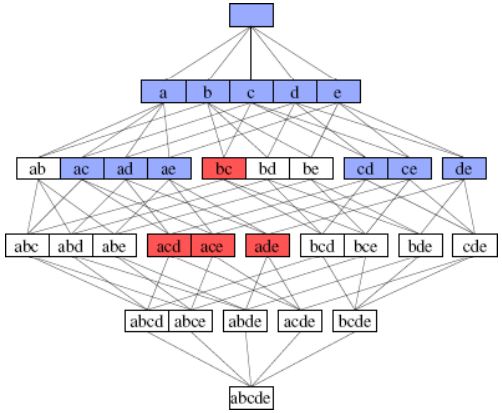


Clasificación

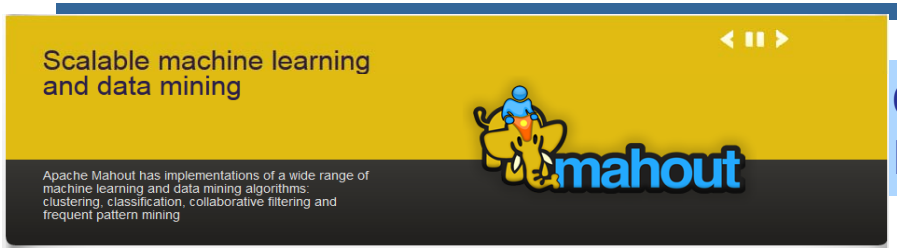


Sistemas de Recomendaciones

Asociación



Mahout



Caso de estudio:
Random Forest para KddCup99

Clase	Nº de instancias
normal	972.781
DOS	3.883.370
PRB	41.102
R2L	1.126
U2R	52

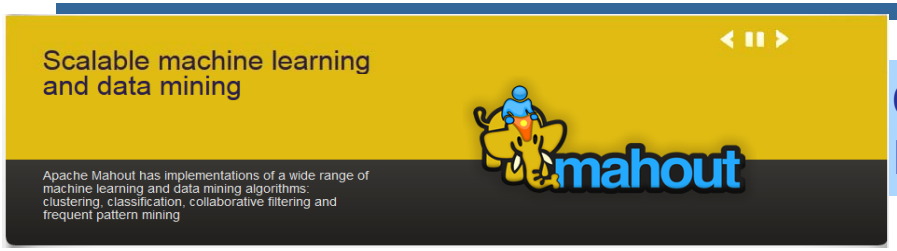
Tabla 2: Número de instancias por clase

Tiempo en segundos para ejecución secuencial

Datasets	RF		
	10%	50%	full
DOS_versus_normal	6344.42	49134.78	NC
DOS_versus_PRB	4825.48	28819.03	NC
DOS_versus_R2L	4454.58	28073.79	NC
DOS_versus_U2R	3848.97	24774.03	NC
normal_versus_PRB	468.75	6011.70	NC
normal_versus_R2L	364.66	4773.09	14703.55
normal_versus_U2R	295.64	4785.66	14635.36

Cluster ATLAS: 12 nodos
-Microprocessors: 2 x Intel E5-2620 (6 cores/12 threads, 2 GHz)
- RAM 64 GB DDR3 ECC 1600MHz
- Mahout version 0.8

Mahout



Caso de estudio:
Random Forest para KddCup99

Clase	Nº de instancias
normal	972.781
DOS	3.883.370
PRB	41.102
R2L	1.126
U2R	52

Tabla 2: Número de instancias por clase

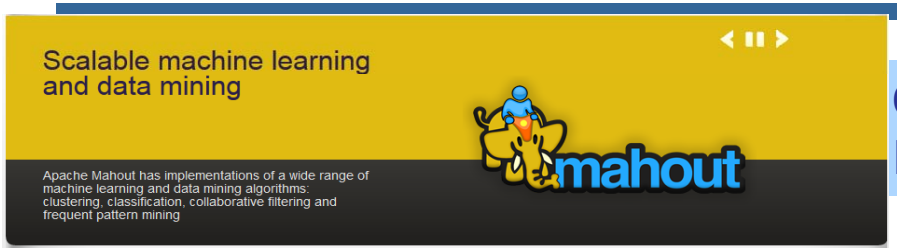
	10%	50%	full
DOS_versus_normal	6344.42	49134.78	NC
DOS_versus_PRB	4825.48	28819.03	NC

Tiempo en segundos para Big Data con 20 particiones

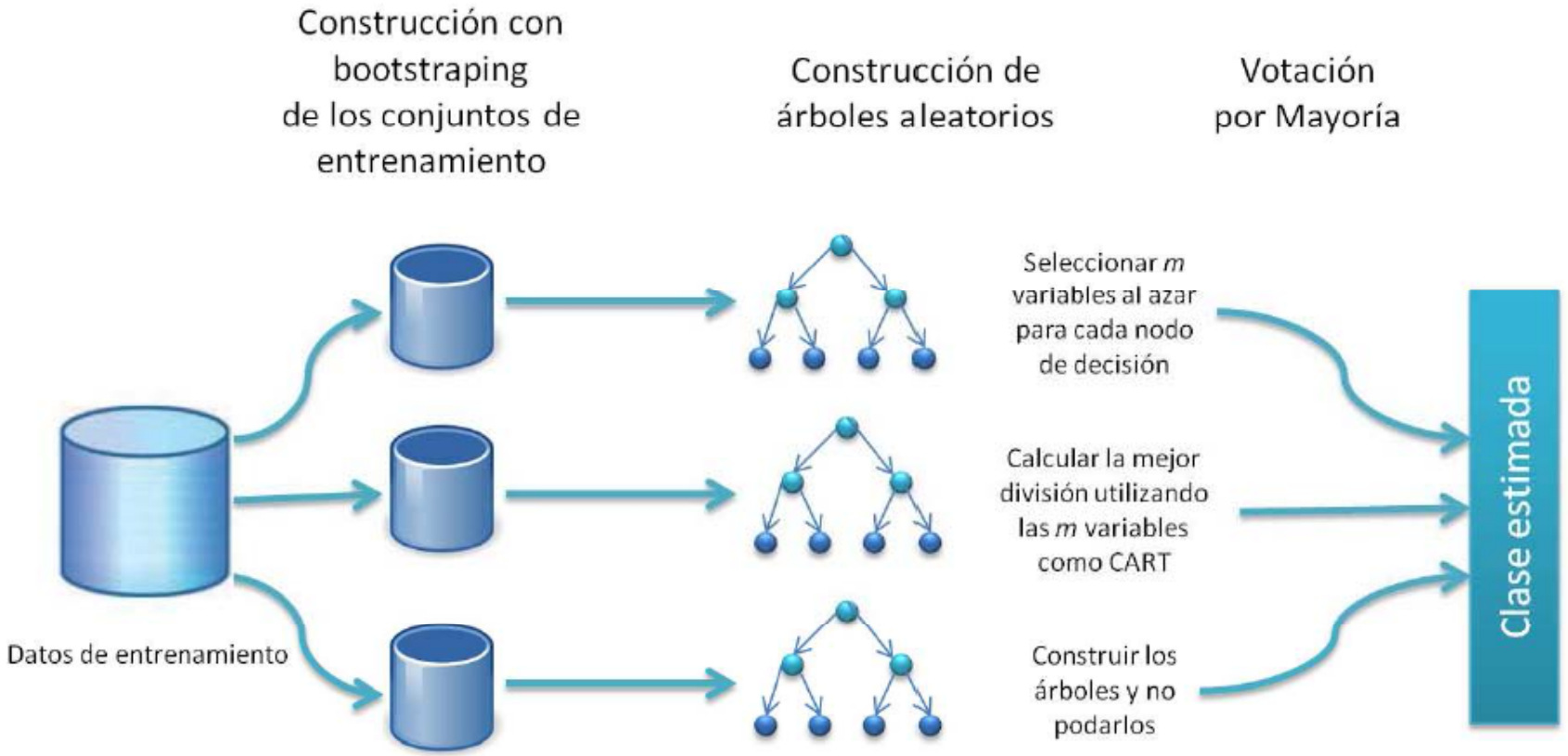
Datasets	RF-BigData		
	10%	50%	full
DOS_versus_normal	98	221	236
DOS_versus_PRB	100	186	190
DOS_versus_R2L	97	157	136
DOS_versus_U2R	93	134	122
normal_versus_PRB	94	58	72
normal_versus_R2L	92	39	69
normal_versus_U2R	93	52	64

Cluster ATLAS: 12 nodos
-Microprocessors: 2 x Intel E5-2620 (6 cores/12 threads, 2 GHz)
- RAM 64 GB DDR3 ECC 1600MHz
- Mahout version 0.8

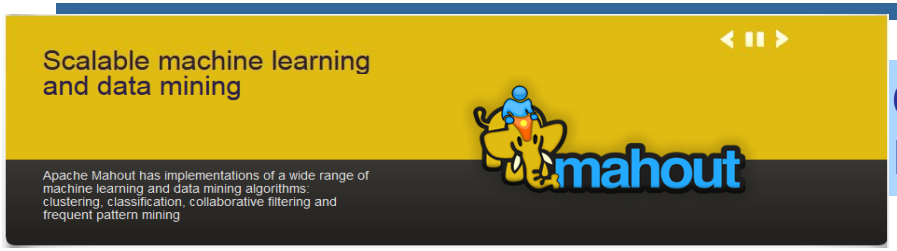
Mahout



Caso de estudio: Random Forest para KddCup99



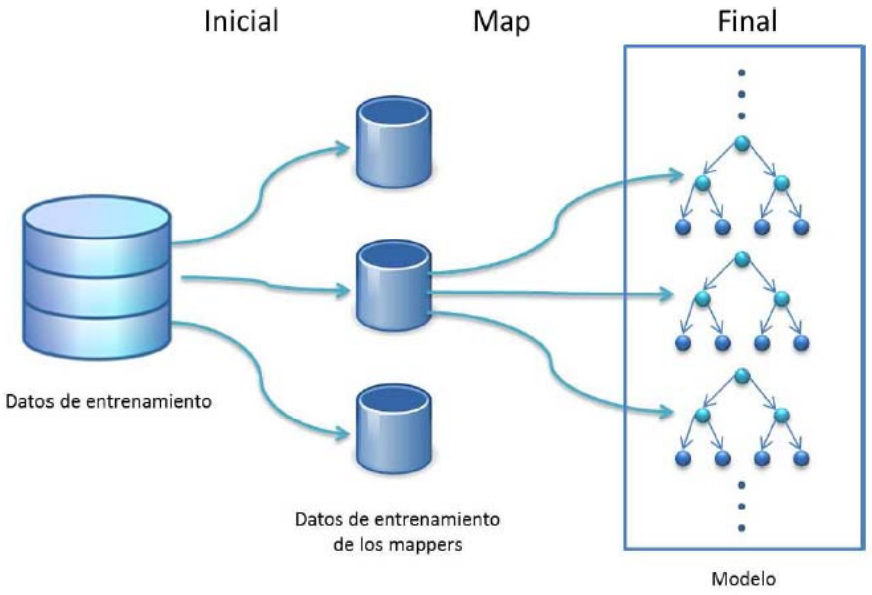
Mahout



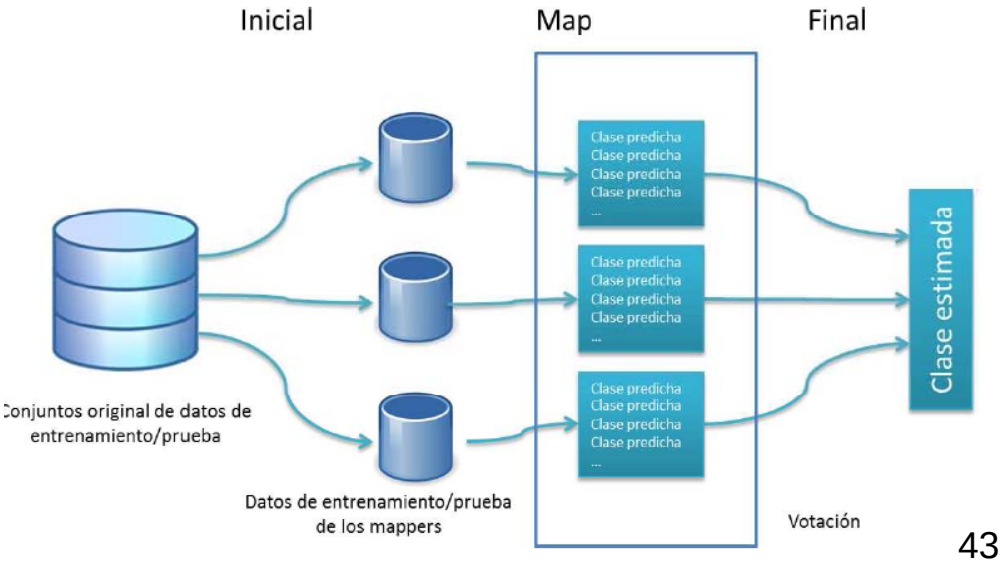
Caso de estudio:
Random Forest para KddCup99

Implementación RF Mahout Partial: Es un algoritmo que genera varios árboles de diferentes partes de los datos (maps). Dos fases:

Fase de Construcción



Fase de Clasificación

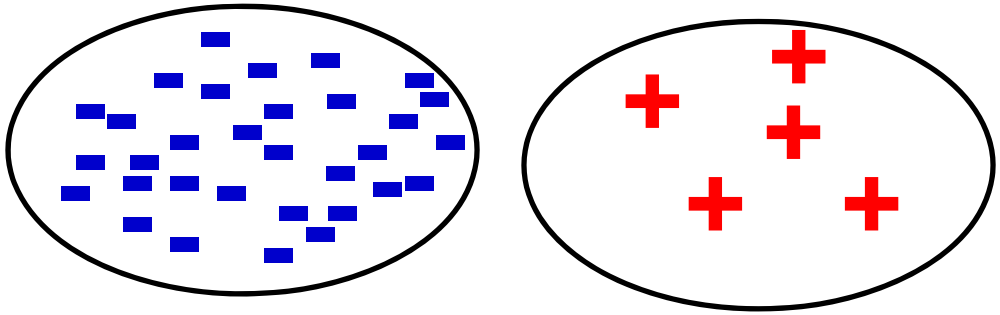


Mahout

Nuestro grupo de investigación trabaja en:

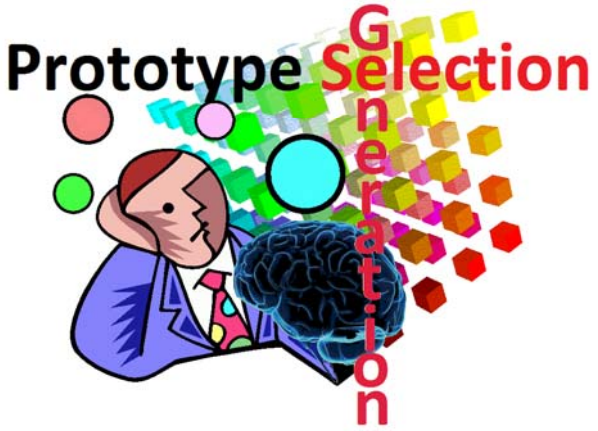
Conjuntos no-balanceados Big Data

Multiclasificadores
Pesos en instancias



Generación de prototipos Para Big Data

Algoritmos evolutivos
Modelos de reducción de prototipos



Mahout vs **Nuevas herramientas**

Herramienta comercial

http://www.pentahobigdata.com/

NIMBLE
(IBM researchers)
ACM SIGKDD, 2011

SystemML
(IBM researchers, DML language, 100-core Amazon EC2)
ICDE 2011

Ricardo
(IBM researchers, Amazon EC2)
R and hadoop
ACM SIGMOD , 2010.

Rhipe
(Purdue University, 2012)
R and hadoop
www.rhipe.org/
http://www.datadr.org/

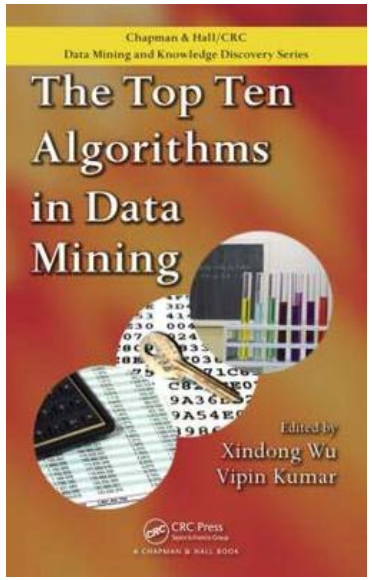
Outline

- ¿Qué es Big Data?
- MapReduce: Paradigma de Programación para Big Data (Google)
- Plataforma Hadoop (Open access)
- Librería Mahout para Big Data. Otras librerías
- **Limitaciones de MapReduce**
- Comentarios Finales

Hadoop Mahout

¿Qué algoritmos puedo encontrar para Hadoop?

Analizamos 10 algoritmos muy conocidos



- **Decision trees (C4.5, Cart)** (MReC4.5)
- **K-Means**
- **SVM**
- **Apriori**
- **kNN**
- **Naïve Bayes**
- **EM (Expectation Maximization)**
- **PageRank**
- **Adaboost**

No disponibles

Limitaciones de MapReduce

Palit, I., Reddy, C.K., 2012. *Scalable and parallel* ***boosting*** *with mapReduce. IEEE TKDE 24 (10), pp. 1904-1916.*

(Amazon EC2 cloud, **CGL-MapReduce:** (modelos de un paso e **iterativos de** MapReduce)

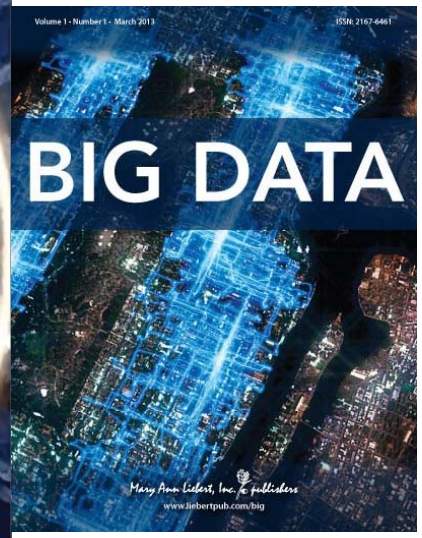
Limitaciones de MapReduce

"If all you have is a hammer, then everything looks like a nail."

MAPREDUCE IS GOOD ENOUGH?

If All You Have is a Hammer, Throw Away Everything That's Not a Nail!

Jimmy Lin
The iSchool, University of Maryland
College Park, Maryland



Los siguientes tipos de algoritmos son ejemplos en los que MapReduce no funciona bien:

Iterative Graph Algorithms
Gradient Descent
Expectation Maximization

Limitaciones de MapReduce

Algoritmos de grafos iterativos. Existen muchas limitaciones para estos algoritmos.

Ejemplo: Cada iteración de PageRank se corresponde a un trabajo de MapReduce.

Se han propuesto una serie de extensiones de MapReduce o modelos de programación alternativa para acelerar el cálculo iterativo:

Pregel (Google)

Implementación: http://www.michaelnielsen.org/ddi/pregel/

Limitaciones de MapReduce

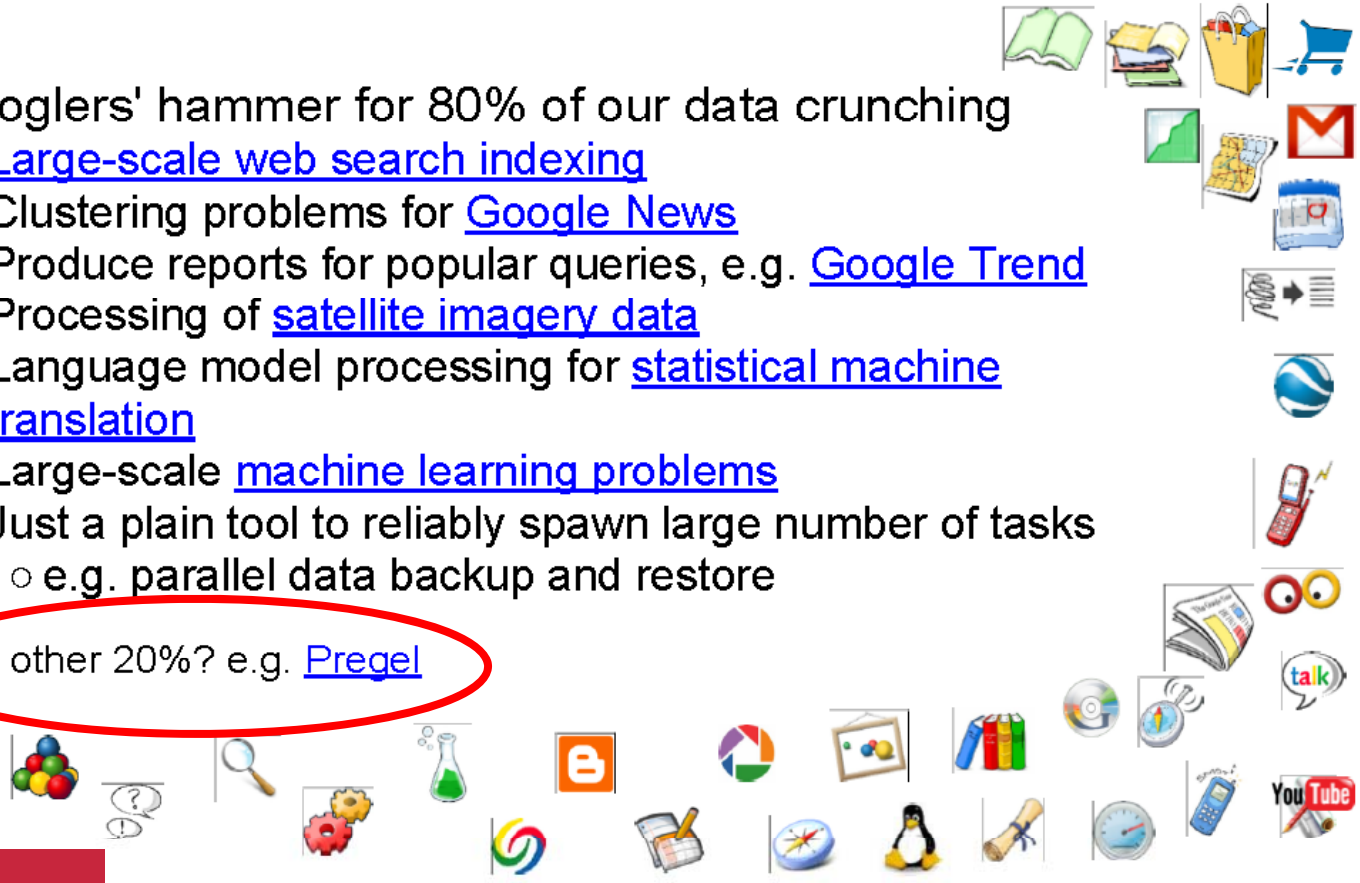
MapReduce inside Google

Google

Googlers' hammer for 80% of our data crunching

- Large-scale web search indexing
- Clustering problems for Google News
- Produce reports for popular queries, e.g. Google Trend
- Processing of satellite imagery data
- Language model processing for statistical machine translation
- Large-scale machine learning problems
- Just a plain tool to reliably spawn large number of tasks
  - e.g. parallel data backup and restore

The other 20%? e.g. Pregel



Limitaciones de MapReduce

GIRAPH (APACHE Project)
(http://giraph.apache.org/)
Procesamiento iterativeo de grafos

GPS - A Graph Processing System, (Stanford)
http://infolab.stanford.edu/gps/
para Amazon's EC2

Distributed GraphLab (Carnegie Mellon Univ.)
https://github.com/graphlab-code/graphlab
Amazon's EC2

Spark (UC Berkeley)
(Se publicita 100 veces más rápido que Hadoop e incluye algoritmos iterativos)
http://spark.incubator.apache.org/research.html

Twister (Indiana University)
http://www.iterativemapreduce.org/
Clusters propios

PrIter (University of Massachusetts Amherst, Northeastern University-China)
http://code.google.com/p/priter/
Cluster propios y Amazon EC2 cloud

HaLoop (University of Washington)
http://clue.cs.washington.edu/node/14
http://code.google.com/p/haloop/
Amazon's EC2

GPU based platforms
Mars
Grex

Outline

- ¿Qué es Big Data?
- MapReduce: Paradigma de Programación para Big Data (Google)
- Plataforma Hadoop (Open access)
- Librería Mahout para Big Data. Otras librerías
- Limitaciones de MapReduce
- **Comentarios Finales**

Comentarios Finales

Desafíos en Big Data

- **Requisitos de rendimiento para el algoritmo**
  - Tradicionalmente, los algoritmos "eficientes"
    - Se ejecutan en tiempo polinomial (pequeño): $O(n \log n)$
    - Utilizar el espacio lineal: $O(n)$
  - Para grandes conjuntos de datos, los algoritmos eficientes
    - Deben ejecutarse en el tiempo lineal o incluso sublineal: $o(n)$
    - Deben utilizar hasta espacio polilogarítmico: $(\log n)^2$
- **Limpieza Big Data**
  - Ruido y datos distorsionados
    - Resultados de cómputo
    - Resultados de búsqueda
  - Necesidad de métodos automáticos para la "limpieza" de los datos
    - Eliminación de duplicados
    - Evaluación de la calidad
- **Modelo de computación**
  - Precisión y aproximación
  - Eficiencia

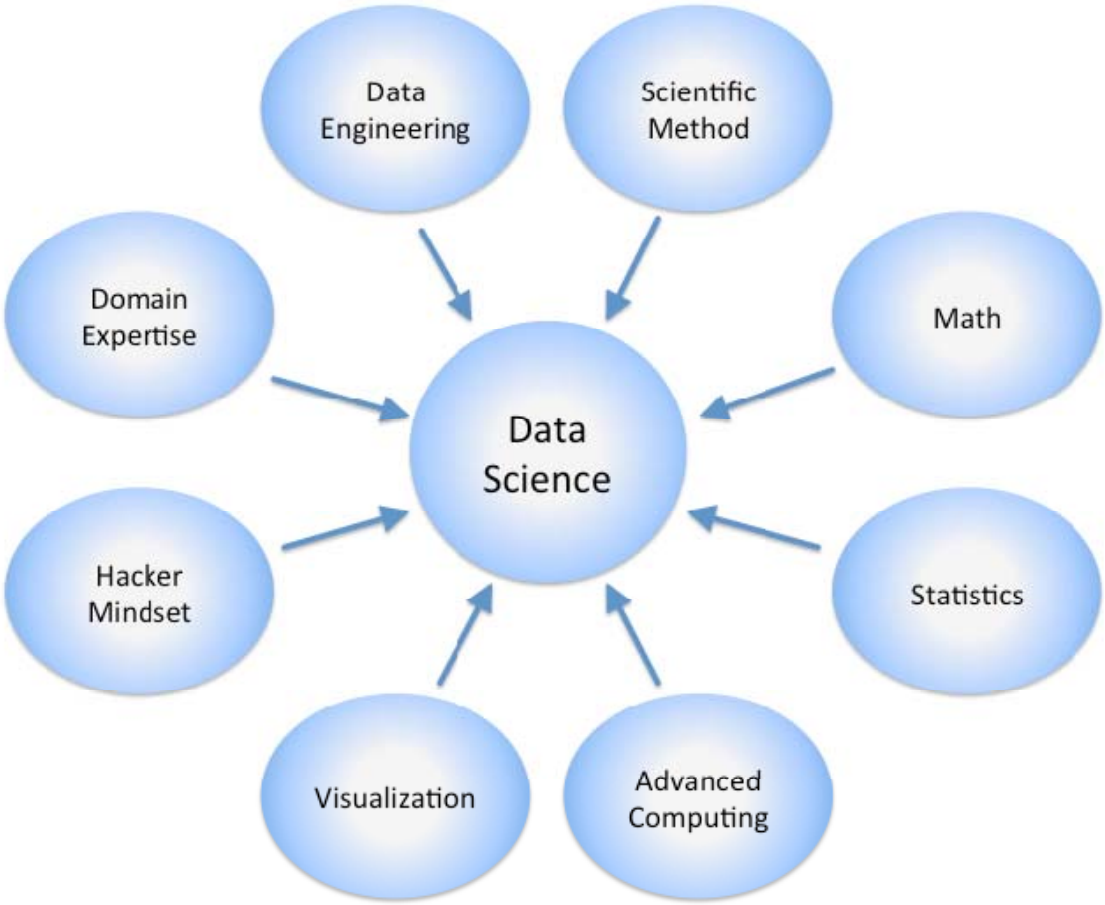
Comentarios Finales

Oportunidades en Big Data

Oportunidad científica: Big Data es un área emergente y en expansión.
Las posibilidades de desarrollo de algoritmos para nuevos datos, aplicaciones reales … es un nicho de investigación y desarrollo en los próximos años.

Comentarios Finales

Oportunidades en Big Data



Comentarios Finales

Oportunidades en Big Data

Oportunidad profesional: En 2015, Gartner predice que 4,4 millones de empleos serán creados en torno a big data. (Gartner, 2013)

Fuente: http://www.gartner.com/technology/topics/big-data.jsp


Gracias!!!