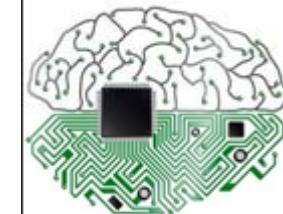




# Escuela de Verano de Inteligencia Artificial



## Big Data



Francisco Herrera

Dpto. Ciencias de la Computación e I.A.  
Universidad de Granada

[herrera@decsai.ugr.es](mailto:herrera@decsai.ugr.es)

Grupo de investigación SCI<sup>2</sup>S  
<http://sci2s.ugr.es>



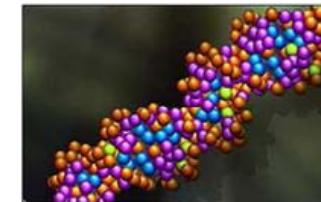
**DECSAI**  
Universidad de Granada



# 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



INTERNET | Campus Party Europa 2013

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



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

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



## Índice

---

- ¿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
- Un caso de estudio:  
ECBDL'14 Competición Big Data
- Comentarios Finales



## Índice

- **¿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
- Un caso de estudio:  
ECBDL'14 Competición Big Data
- Comentarios Finales



# ¿Qué es Big Data?

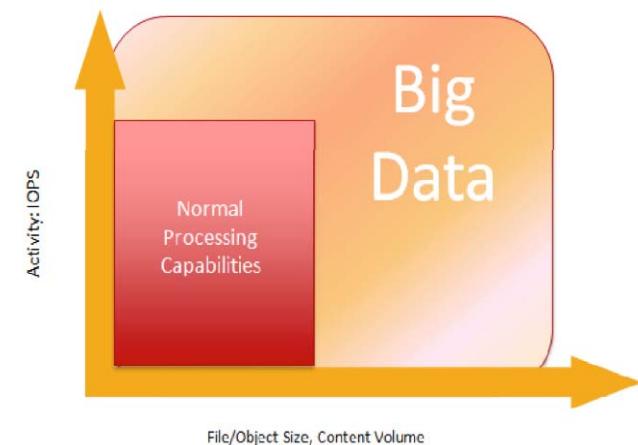
No hay una definición estándar



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

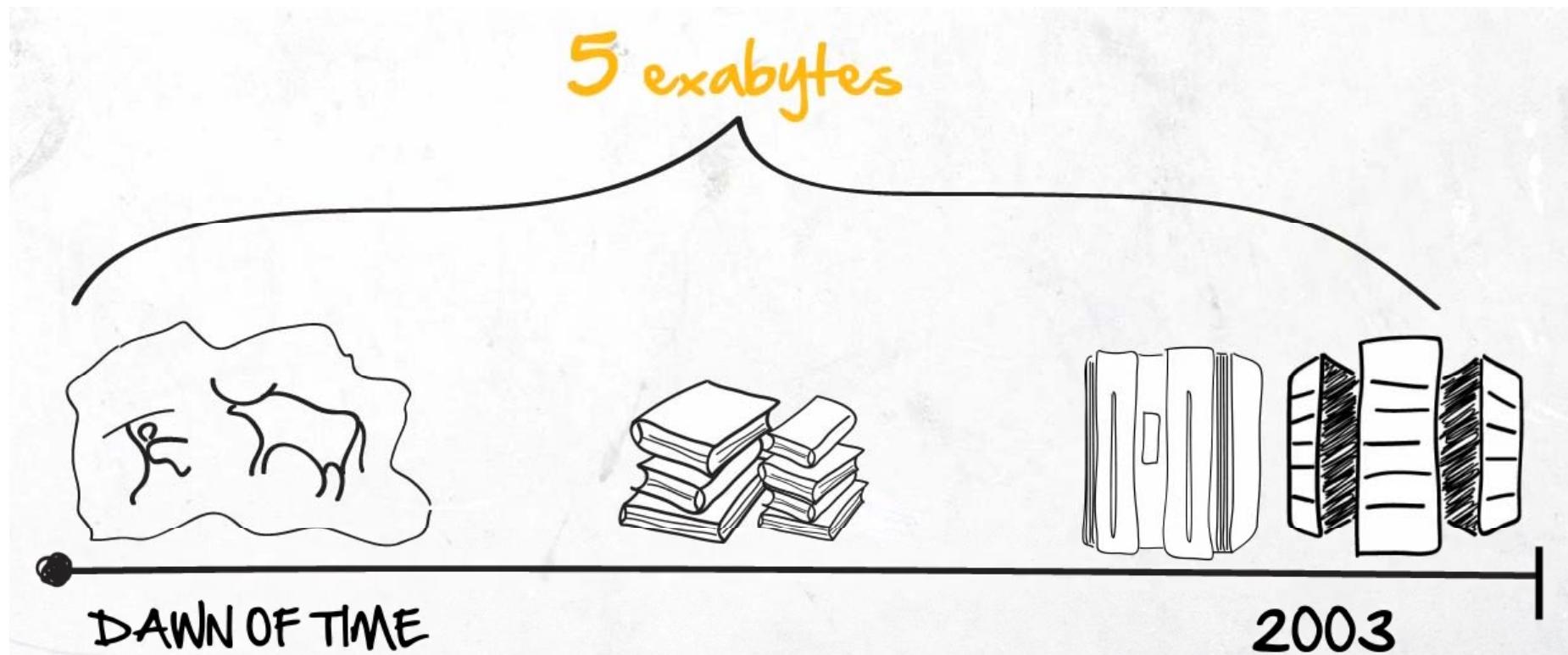


"**Big Data**" son datos cuyo volumen, diversidad y complejidad **requieren nueva arquitectura, técnicas, algoritmos y análisis** para gestionar y extraer valor y conocimiento oculto en ellos ...

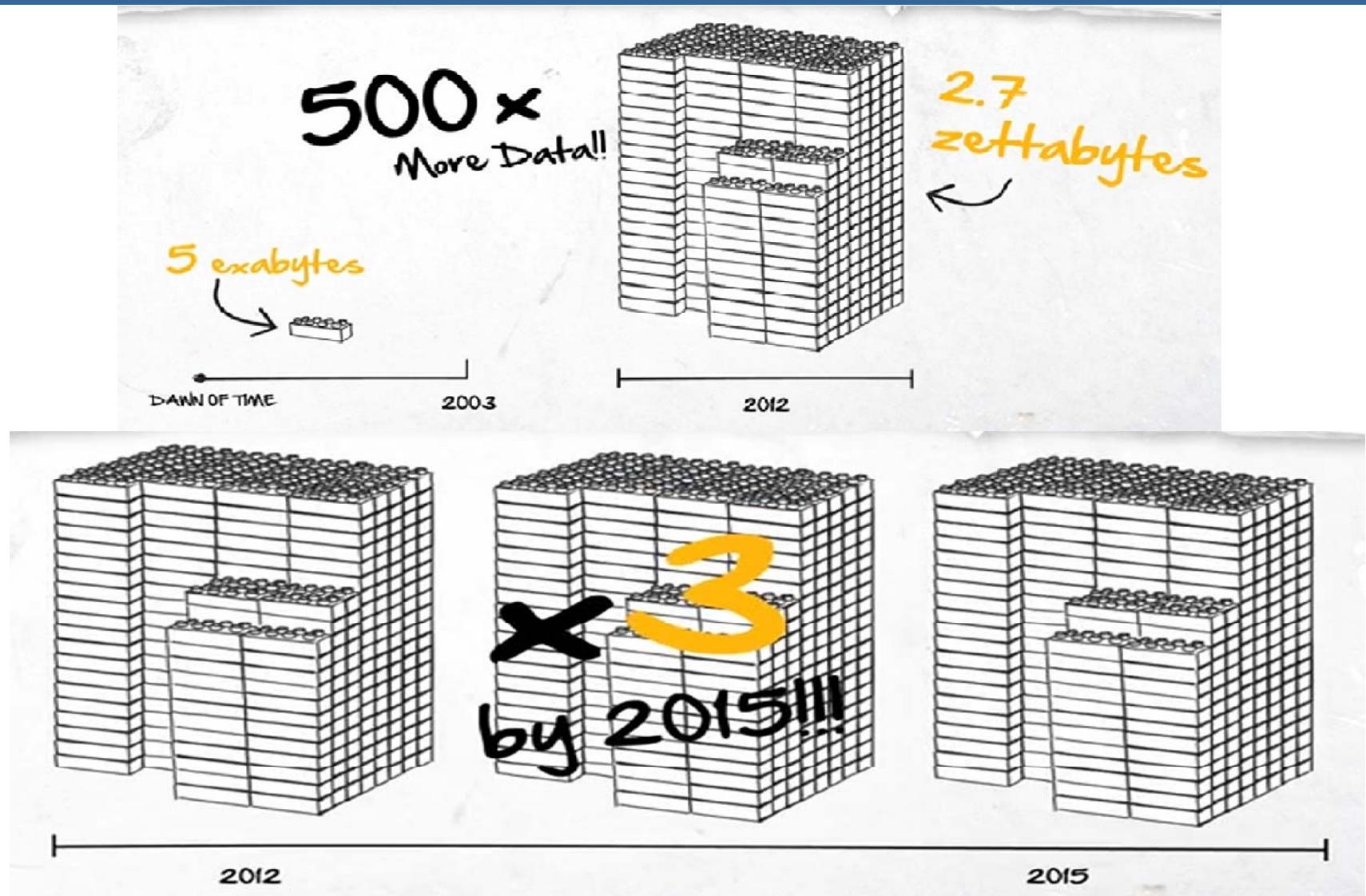


# ¿Qué es Big Data?

---



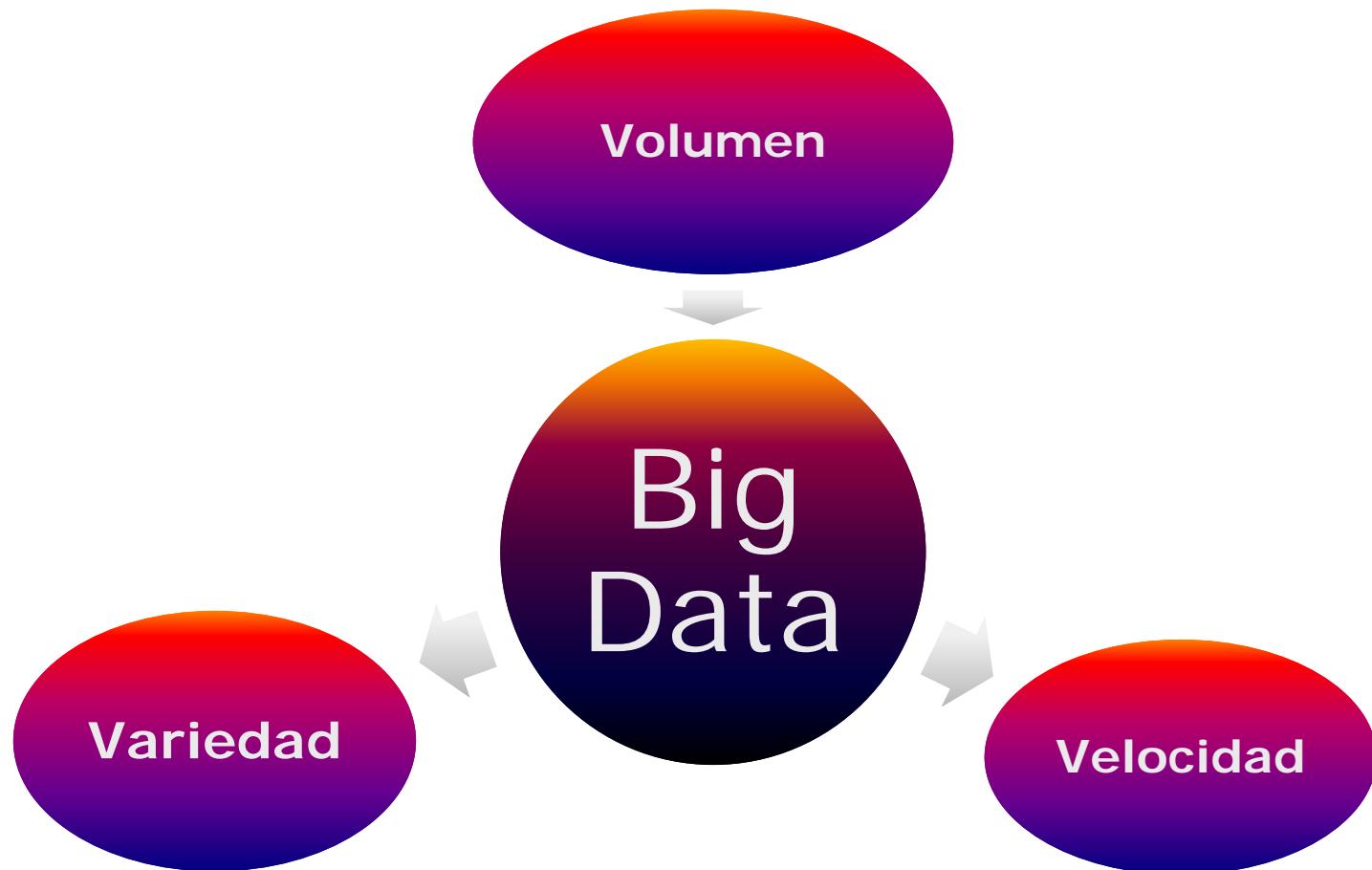
# ¿Qué es Big Data?



# ¿Qué es Big Data?

---

## Las 3 V's de Big Data



# ¿Qué es Big Data? 3 V's de Big Data

---

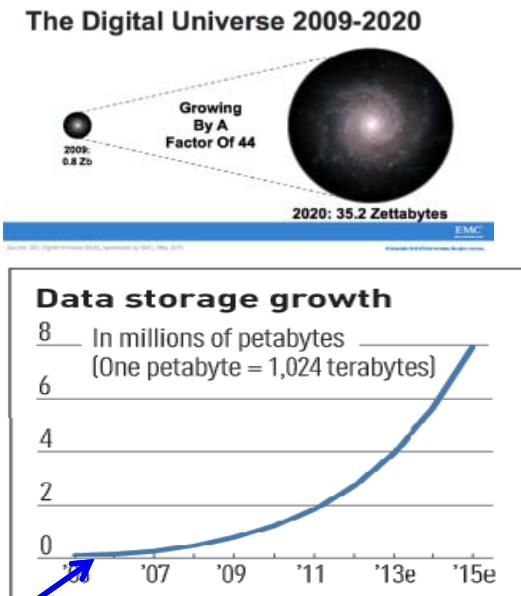
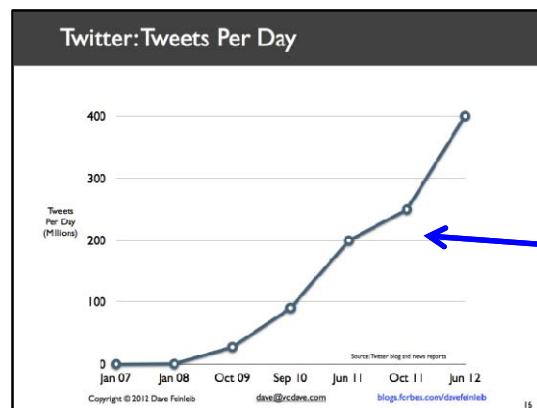
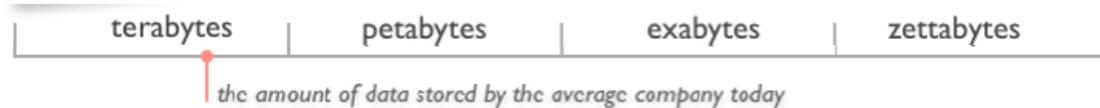


# ¿Qué es Big Data? 3 V's de Big Data

## 1<sup>a</sup>:Volumen

**El volumen de datos crece exponencialmente**

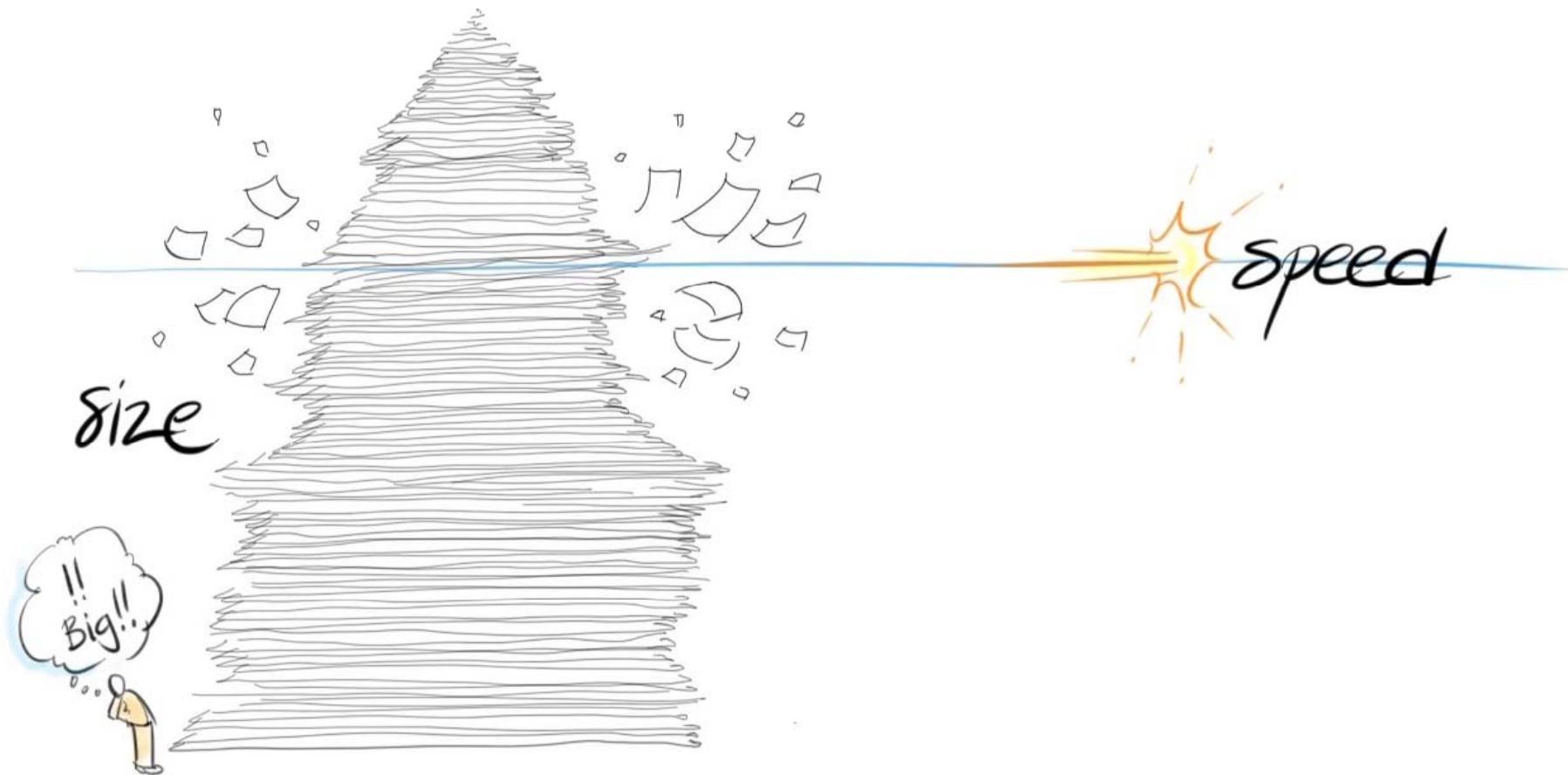
- Crecimiento x 44 de 2009 a 2020
- De 0.8 zettabytes a 35ZB



*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

---

## 2<sup>a</sup>:Velocidad

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

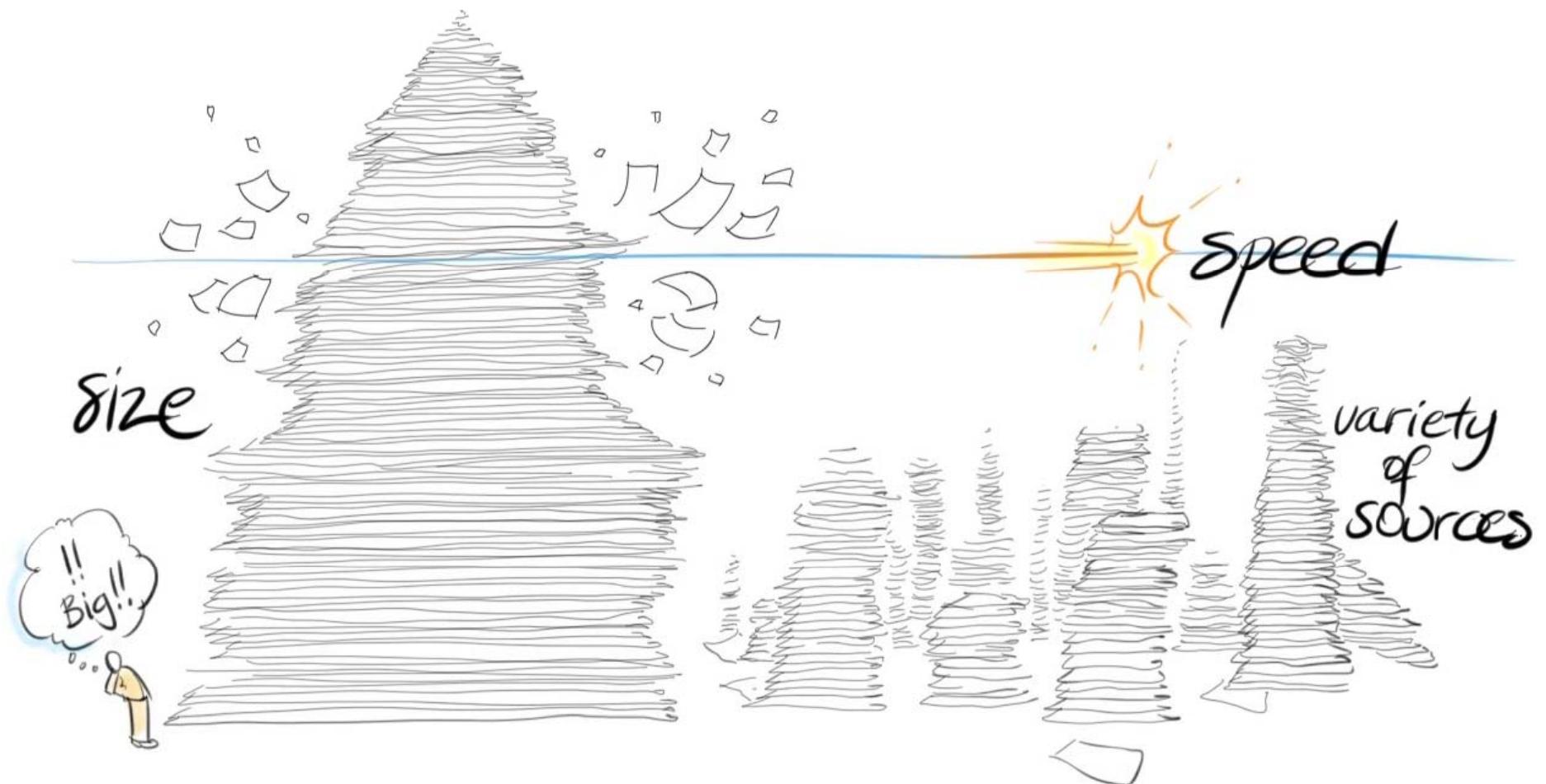


### Ejemplos:

- **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

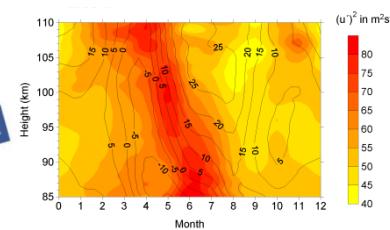
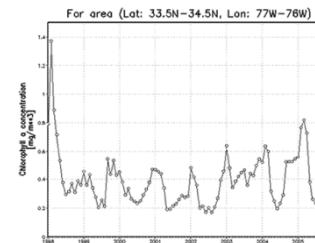
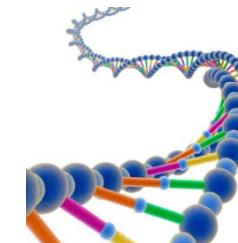
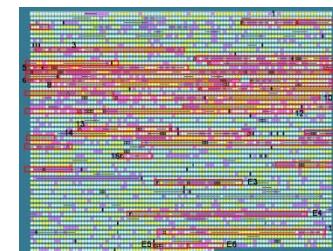
---



# ¿Qué es Big Data? 3 V's de Big Data

## 3<sup>a</sup>:Variedad

- Varios formatos y estructuras:  
Texto, numéricos, imágenes, audio,  
video, secuencias, series temporales  
...
- Una sola aplicación puede generar  
muchos tipos de datos

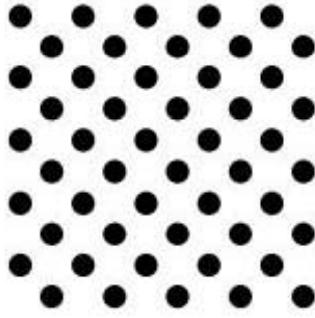
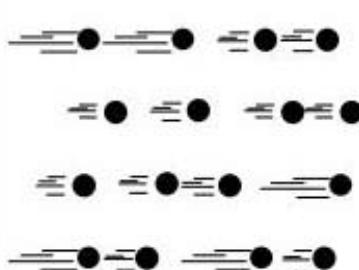
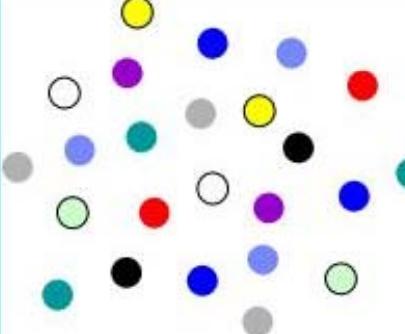
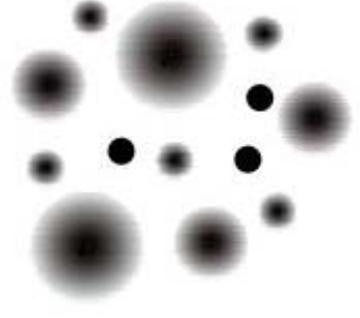


**Extracción de conocimiento →**  
Todos estos tipos de datos  
necesitan ser analizados  
conjuntamente

# ¿Qué es Big Data?

4<sup>a</sup> V → Veracidad

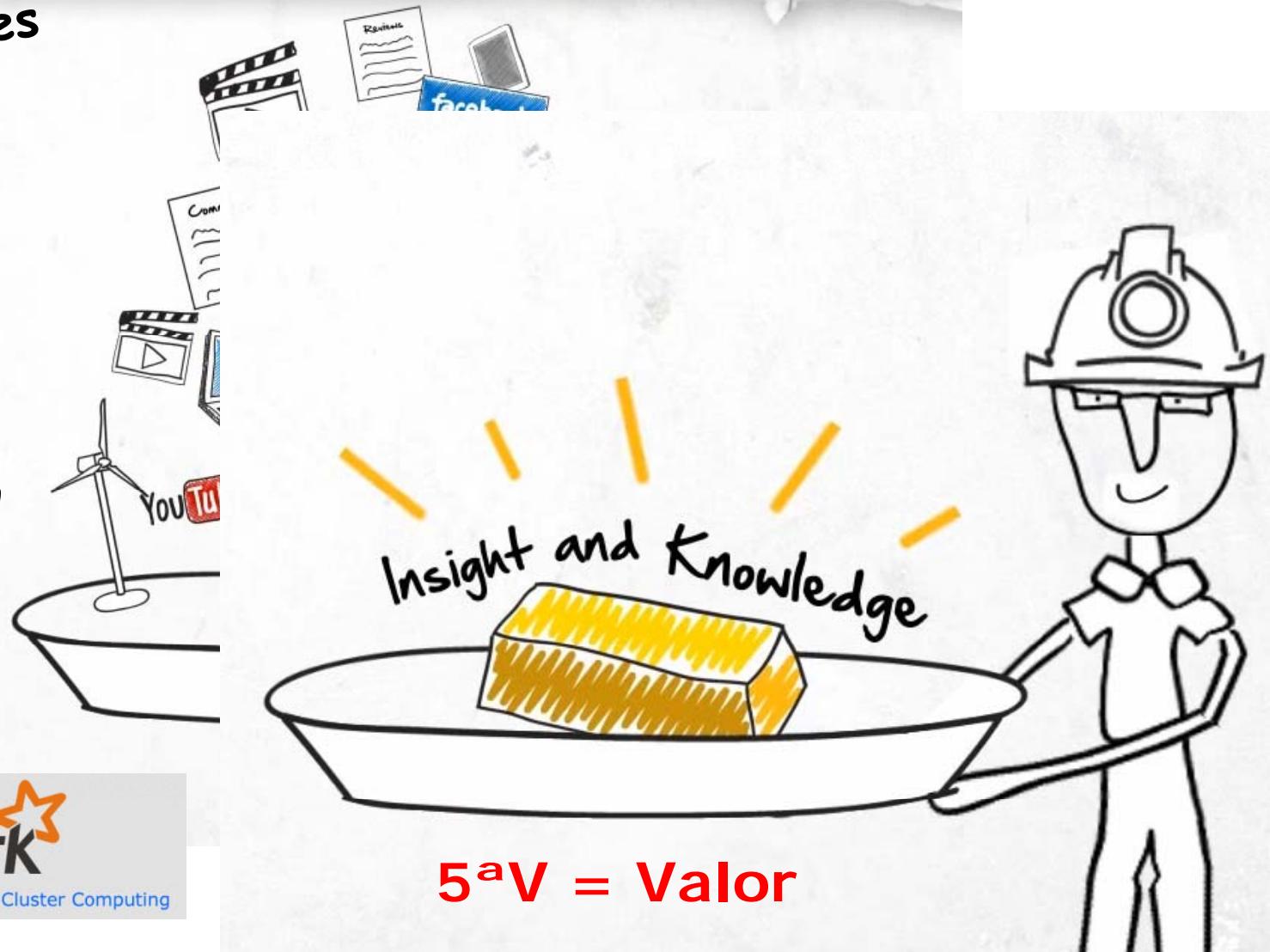
4<sup>a</sup>V  
↓  
Veracidad

Volume	Velocity	Variety	Veracity*
			
Data at Rest	Data in Motion	Data in Many Forms	Data in Doubt
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

# ¿Qué es Big Data?

5 V's --> Valor

Aproximaciones  
y tecnologías  
innovativas



# ¿Qué es Big Data?

Las 8 V's de Big Data

---



# ¿Qué es Big Data?

---



**Big data es cualquier característica sobre los datos que represente un reto para las funcionalidades de un sistema.**

# ¿Qué es Big Data?

---



**Big data incluye datos estructurados con datos no estructurados, imágenes, videos ...**

# ¿Qué es Big Data?

## ¿Quién genera Big Data?



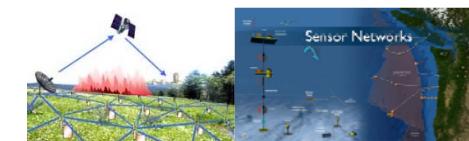
**Redes sociales y multimedia**  
(todos generamos datos)



**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**

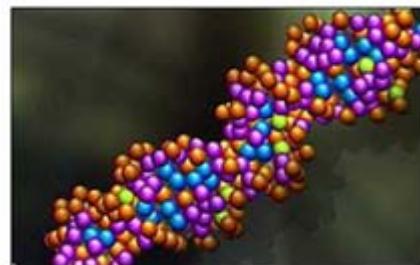
# Big Data. Aplicaciones

## 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



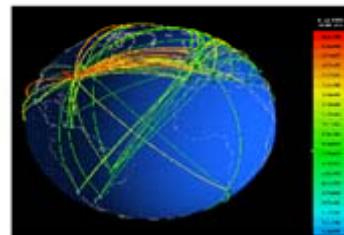
- 250M calls/day
- 60G calls/year
- 40 bytes/call
- 2.5 Terabytes/year

## 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



- 251
- 10k
- 250
- \*De
- timi



# Big Data. Ejemplo

## Evolutionary Computation for Big Data and Big Learning Workshop

**Big Data Competition 2014: Self-deployment track**

**Objective: Contact map prediction**

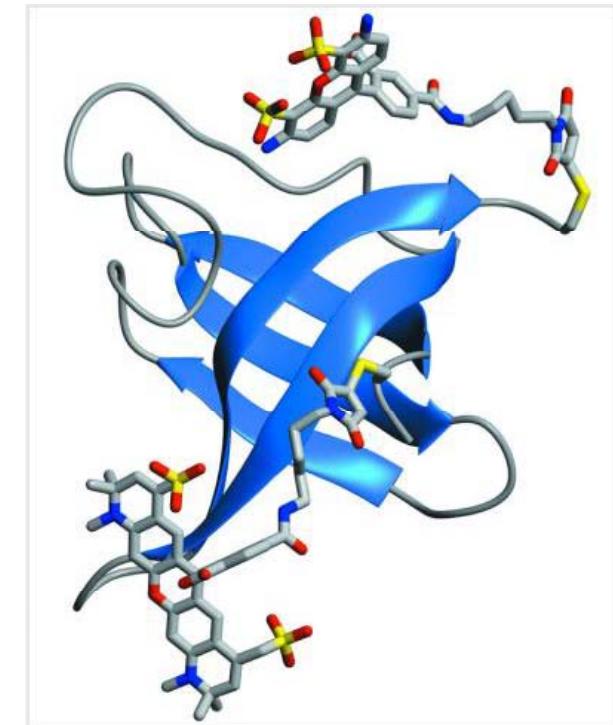
**Details:**

- 32 million instances
- 631 attributes (**539 real & 92 nominal values**)
- 2 classes
- 98% of negative examples
- About 56.7GB of disk space

**Evaluation:**

True positive rate · True negative rate  
TPR · TNR

<http://cruncher.ncl.ac.uk/bdcomp/index.pl?action=data>



## Índice

---

- ¿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
- Un caso de estudio:  
ECBDL'14 Competición Big Data
- 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**

**¿Qué ocurre cuando el tamaño de los datos aumenta y los requerimientos de tiempo se mantiene?**

**Hace unos años:** Había que aumentar los recursos de hardware (número de nodos). Esto tiene limitaciones de espacio, costes, ...

**Google 2004:** Paradigma **MapReduce**



# 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**

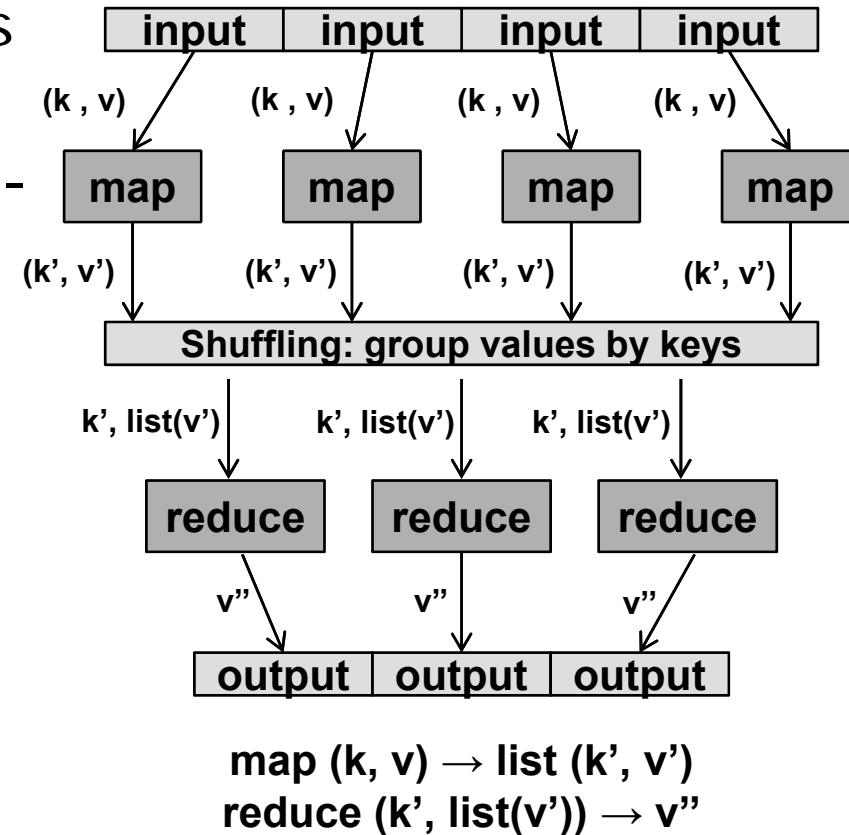
## MapReduce

- Modelo de programación de datos paralela
- Concepto simple, elegante, extensible para múltiples aplicaciones
- **Creado por Google (2004)**
  - Procesa 20 PB de datos por día (2004)
- **Popularizado por el proyecto de código abierto Hadoop**
  - Usado por [Yahoo!](#), [Facebook](#), [Amazon](#), ...



# MapReduce

- MapReduce es el entorno más popular para Big Data
- Basado en la estructura Valor-llave.
- Dos operaciones:
  1. **Función Map :** Procesa bloques de información
  2. **Función Reduce function:** Fusiona los resultados previous de acuerdo a su llave.
- + Una etapa intermedia de agrupamiento por llave (**Shuffling**)



J. Dean, S. Ghemawat, MapReduce: Simplified data processing on large clusters, Communications of the ACM 51 (1) (2008) 107-113.

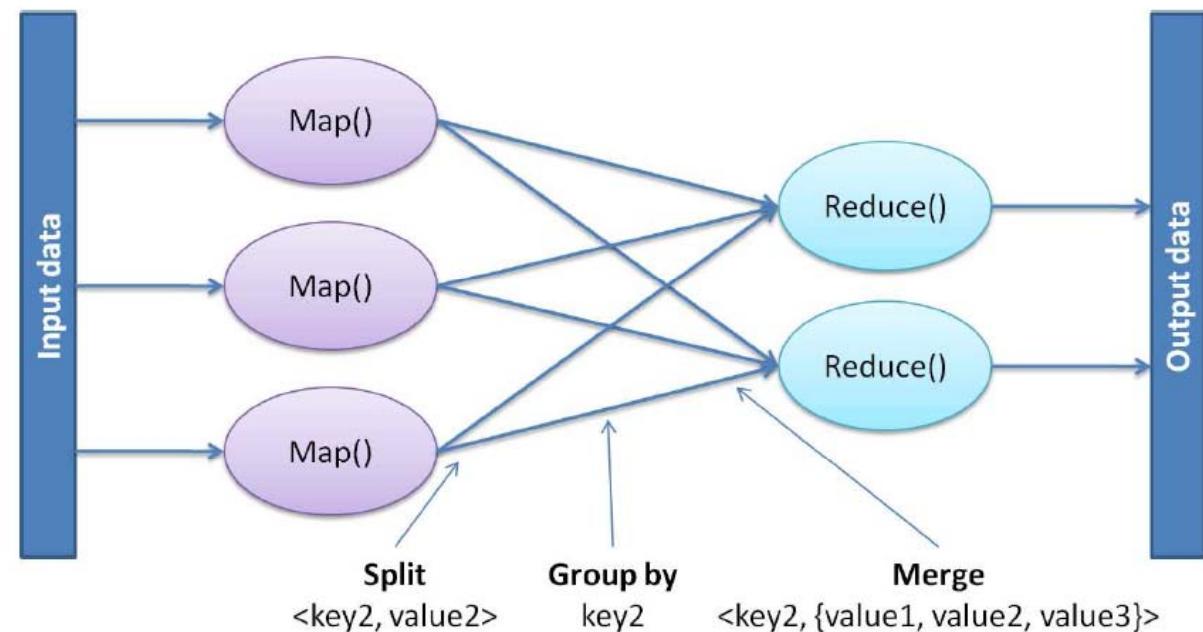


# MapReduce

## Flujo de datos MapReduce

**map**  $(k, v) \rightarrow$   
**list** $(k', v')$

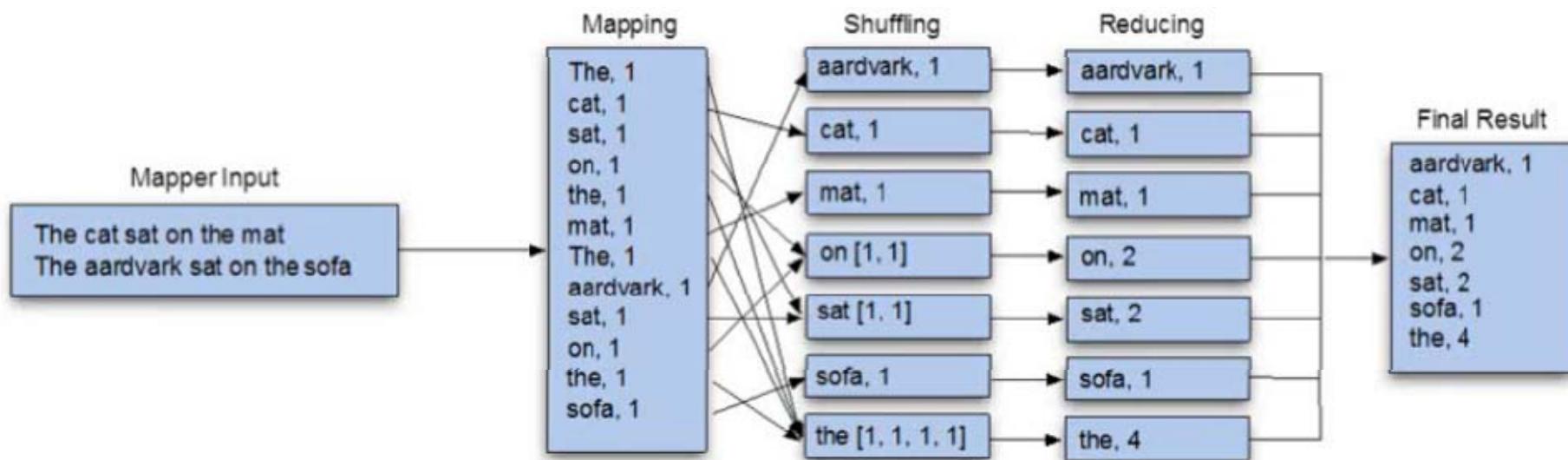
**reduce**  
**(** $k'$ **,** **list** $(v')$ **)**  $\rightarrow$  **v''**





# MapReduce

## Una imagen completa del proceso MapReduce



# 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



# MapReduce

---

## Características

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

# MapReduce

## ¿Número de BLOQUES?

El número de bloques es independiente del número de nodos disponibles.

Va a estar asociado al tamaño y características del problema.

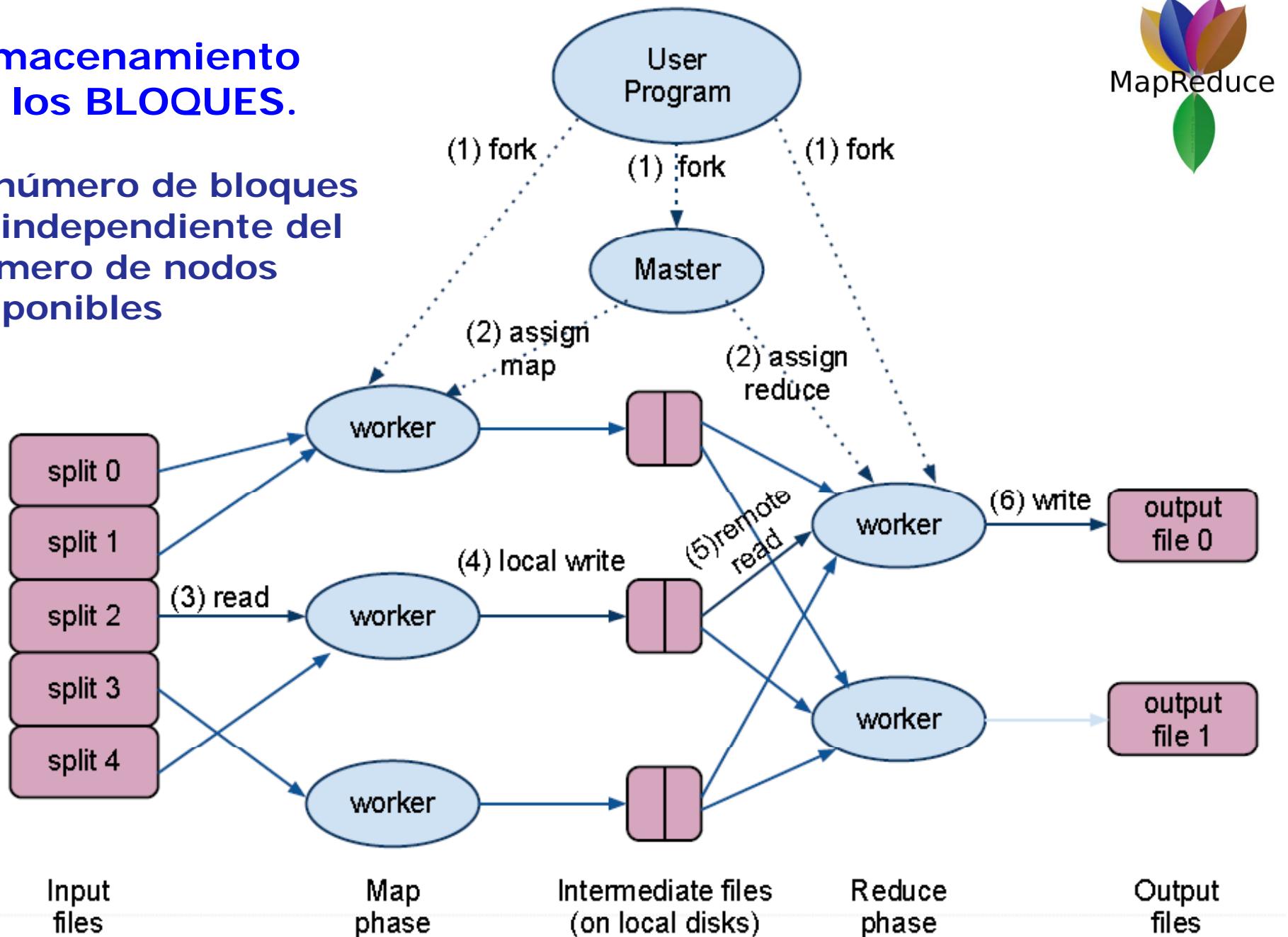


VS



## Almacenamiento de los BLOQUES.

El número de bloques es independiente del número de nodos disponibles

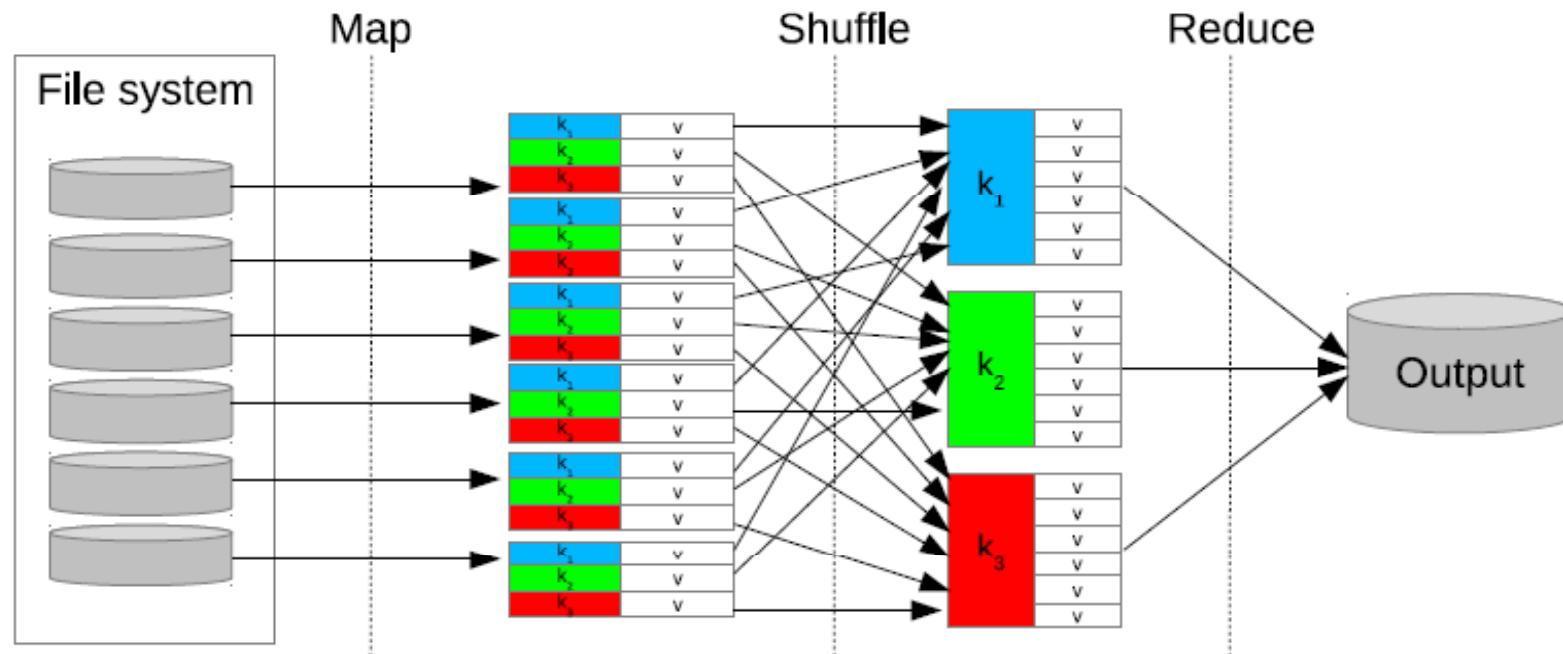


Almacenamiento con copias, normalmente r=3



# MapReduce

Flujo de datos en MapReduce, transparente para el programador



Ficheros de  
entrada  
Particionamiento  
en bloques

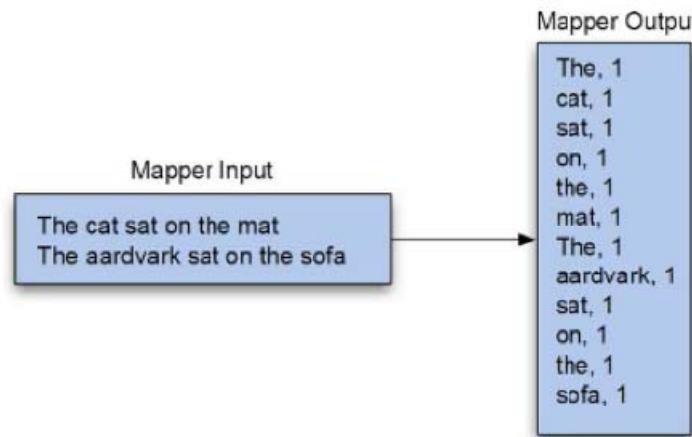
Ficheros Intermedios

Ficheros salida

# MapReduce

## Aspectos a analizar:

**Proceso map:** Puede crear conjuntos de datos muy pequeños, lo cual puede ser un inconveniente para algunos problemas.

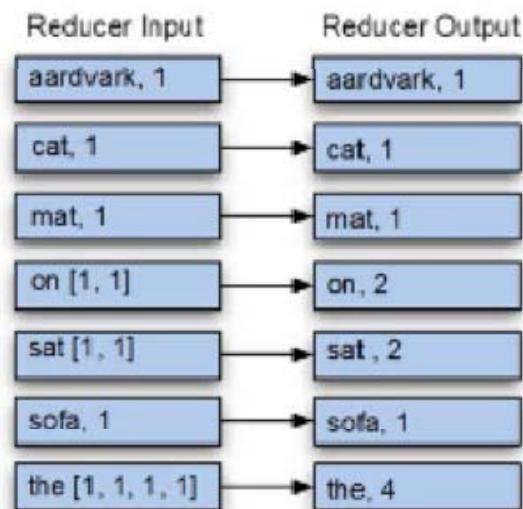


**Ejemplo: Problemas de Clasificación.** Nos podríamos encontrar con el problema de falta de densidad de datos y de clases con muy pocos datos (clasificación no balanceada).

# MapReduce

## Aspectos a analizar:

**Proceso reduce:** Debe combinar las soluciones de todos los modelos/procesos intermedios



Esta es la fase más creativa porque hay que conseguir crear un modelo global de calidad asociado al problema que se desea resolver a partir de los modelos intermedios.



# 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**

# MapReduce

## Limitaciones



**"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

ORIGINAL ARTICLE

Volume 1 - Number 1 - March 2013

ISSN 2167-6461

BIG DATA

Mary Ann Liebert, Inc. & publishers  
www.liebertpub.com/big



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

Iterative Graph Algorithms  
Gradient Descent  
Expectation Maximization



## Índice

---

- ¿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
- Un caso de estudio:  
ECBDL'14 Competición Big Data
- Comentarios Finales



# Hadoop



**Hadoop es una implementación de código abierto del paradigma computacional MapReduce**

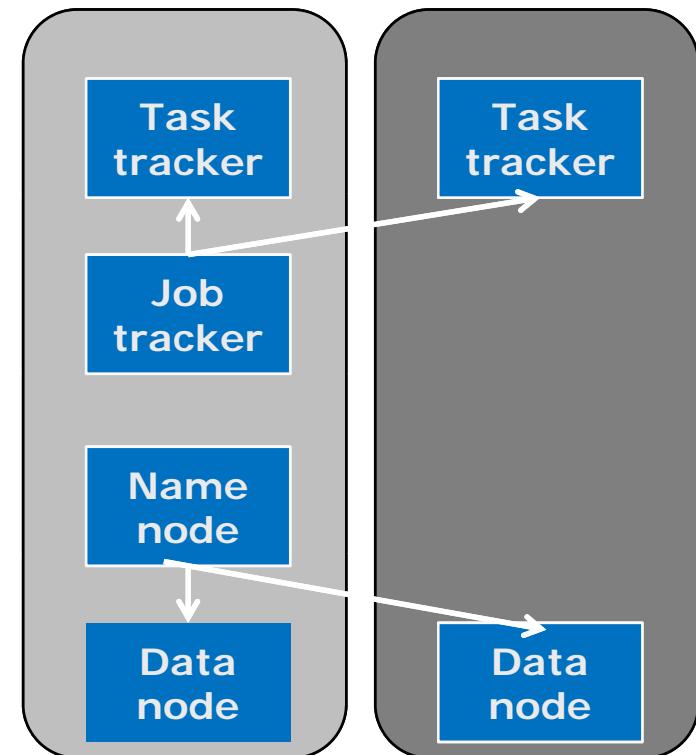
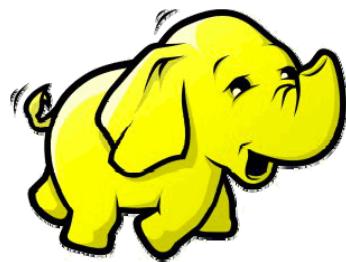


<http://hadoop.apache.org/>

# Hadoop



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



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 O'Malley, 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

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



Bytes	Nodes	Maps	Reduces	Replication	Time
500,000,000,000	1406	8000	2600	1	59 seconds
1,000,000,000,000	1460	8000	2700	1	62 seconds
100,000,000,000,000	3452	190,000	10,000	2	173 minutes
1,000,000,000,000,000	3658	80,000	20,000	2	975 minutes

## Yahoos' Hammer Cluster

- approximately 3800 nodes (in such a large cluster, nodes are always down)
- 2 quad core Xeons @ 2.5ghz per node
- 4 SATA disks per node
- 8G RAM per node (upgraded to 16GB before the petabyte sort)

<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



---

## El proyecto Apache Hadoop incluye los módulos:

**Hadoop Common:** Las utilidades comunes que apoyan los otros módulos de Hadoop.

**Hadoop Distributes File System (HDFS):** El sistema de ficheros que proporciona el acceso

**Hadoop YARN:** Marco para el manejo de recursos de programación y grupo de trabajo.

**Hadoop MapReduce:** Un sistema de basado en YARN o para el procesamiento en paralelo de grandes conjuntos de datos.

<http://hadoop.apache.org/>

## Ecosistema Apache Hadoop incluye más de 150 proyectos:

**Avro:** Un sistema de serialización de datos.

**Cassandra:** Una base de datos escalable multi-master sin puntos individuales y fallo

**Chukwa:** Un sistema de recogida de datos para la gestión de grandes sistemas distribuidos.

**Hbase:** Una base de datos distribuida, escalable que soporta estructurado de almacenamiento de datos para tablas de gran tamaño.

**Hive:** Un almacén de datos que proporciona el Resumen de datos para tablas de gran tamaño.

**Pig:** Lenguaje para la ejecución de alto nivel de flujo de datos para computación paralela.

**Tez:** Sustituye al modelo “MapShuffleReduce” por un flujo de ejecución con grafos acíclico dirigido (DAG)

**Giraph:** Procesamiento iterativo de grafos

**Mahout:** Aprendizaje automático escalable (biblioteca de minería de datos)

Recientemente: Apache Spark



# 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™

Windows Azure

<http://www.windowsazure.com/>

Instalación en un cluster  
Ejemplo ATLAS, infraestructura  
del grupo SCI<sup>2</sup>S



Cluster ATLAS: 4 super servers from Super Micro Computer Inc. (4 nodes per server)

The features of each node are:

- Microprocessors: 2 x Intel Xeon E5-2620 (6 cores/12 threads, 2 GHz, 15 MB Cache)
- RAM 64 GB DDR3 ECC 1600MHz, Registered
- 1 HDD SATA 1TB, 3Gb/s; (system)
- 1 HDD SATA 2TB, 3Gb/s; (distributed file system)

# Hadoop



¿Cómo puedo instalar Hadoop?

cloudera Ask Bigger Questions

Distribución que ofrece Cloudera para Hadoop.

WHY CLOUDERA | PRODUCTS | SOLUTIONS | PARTNERS | RESOURCES | SUPPORT | ABOUT

<http://www.cloudera.com/content/cloudera/en/why-cloudera/hadoop-and-big-data.html>

¿Qué es Cloudera?

Cloudera es la primera distribución Apache Hadoop comercial y no-comercial.

## Índice

- ¿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
- Un caso de estudio:  
ECBDL'14 Competición en Big Data
- Comentarios Finales



# Mahout



## Software de Ciencia de Datos

Generation	1 <sup>a</sup> Generación	2 <sup>a</sup> Generación
Ejemplos	KNIME, SAS, R, Weka, SPSS, KEEL	Mahout, Pentaho, Cascading
Scalabilidad	Vertical	Horizontal (over Hadoop)
Algoritmos disponibles	Huge collection of algorithms	Small subset: sequential logistic regression, linear SVMs, Stochastic Gradient Descent, k-means clustering, Random forest, etc.
Algoritmos No disponibles	Practically nothing	Vast no.: Kernel SVMs, Multivariate Logistic Regression, Conjugate Gradient Descent, ALS, etc.
Tolerancia a Fallos	Single point of failure	Most tools are FT, as they are built on top of Hadoop

# Mahout



## Scalable machine learning and data mining

Apache Mahout has implementations of a wide range of machine learning and data mining algorithms:  
clustering, classification, collaborative filtering and frequent pattern mining



### 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/>

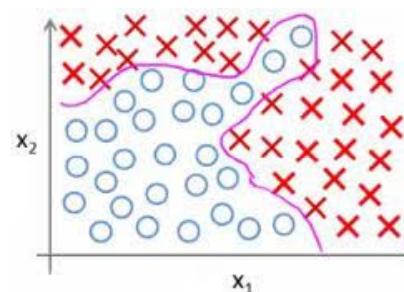
# Mahout



Scalable machine learning and data mining

Apache Mahout has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining

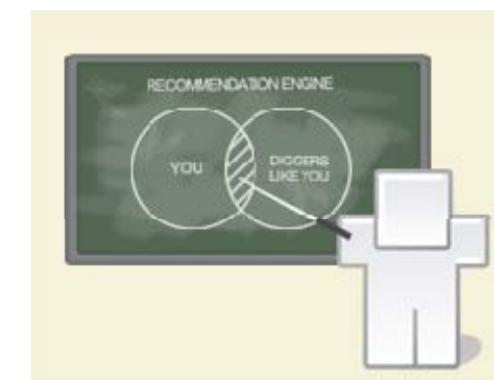
Cuatro grandes áreas de aplicación



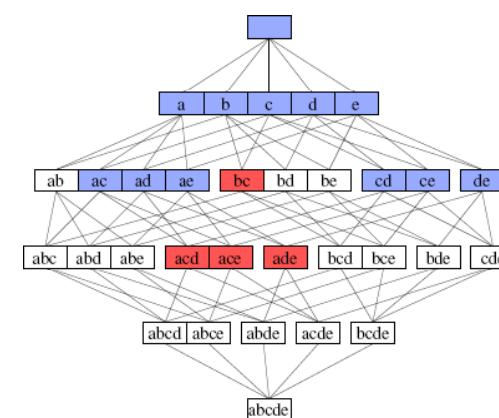
Clasificación

Agrupamiento

Asociación



Sistemas de Recomendaciones



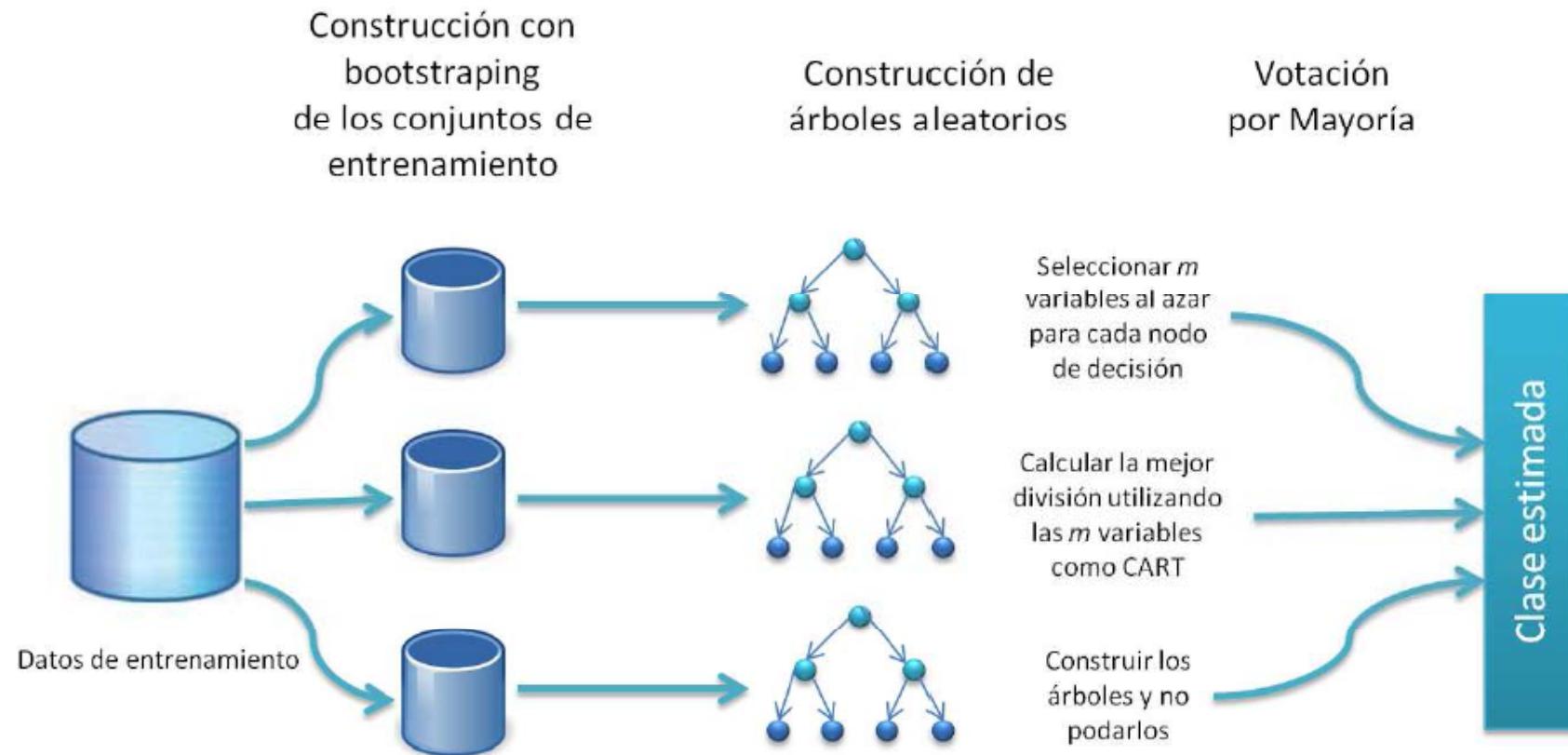
# Mahout



Scalable machine learning and data mining

Apache Mahout has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining

## Caso de estudio: Random Forest para KddCup99



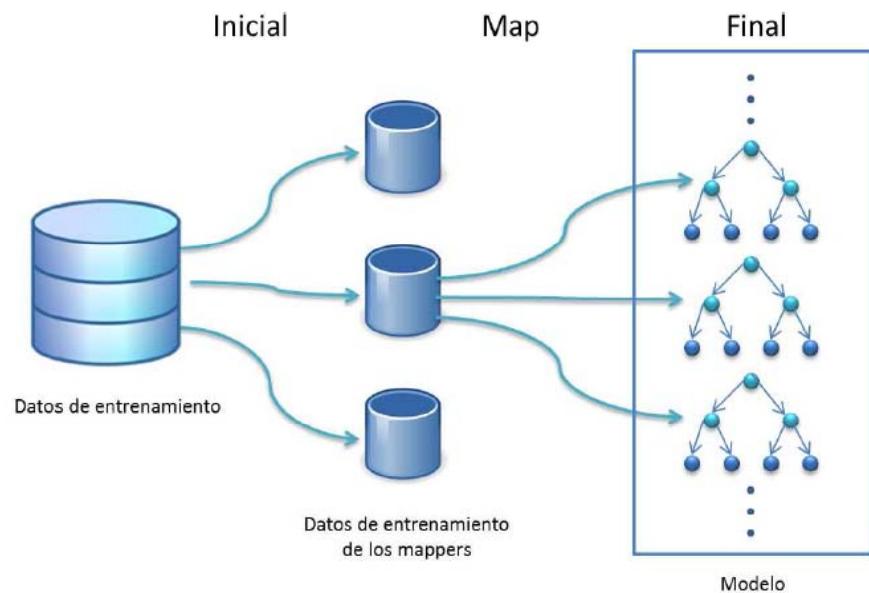
# Mahout



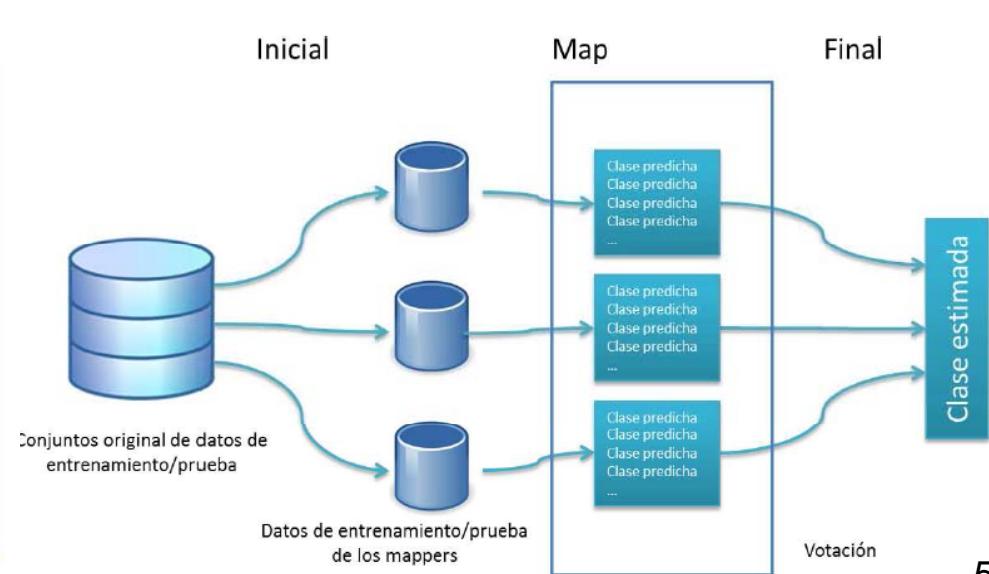
## Caso de estudio: Random Forest para KddCup99

**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



Scalable machine learning and data mining

Apache Mahout has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining

## Caso de estudio: Random Forest para KddCup99

Class	Instance Number
normal	972.781
DOS	3.883.370
PRB	41.102
R2L	1.126
U2R	52

### 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: 16 nodos

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

# Mahout



Scalable machine learning and data mining

Apache Mahout has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining

## Caso de estudio: Random Forest para KddCup99

Class	Instance Number
normal	972.781
DOS	3.883.370
PRB	41.102
R2L	1.126
U2R	52

	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: 16 nodos

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

# 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.rhipe.org/)

<http://www.datadr.org/>

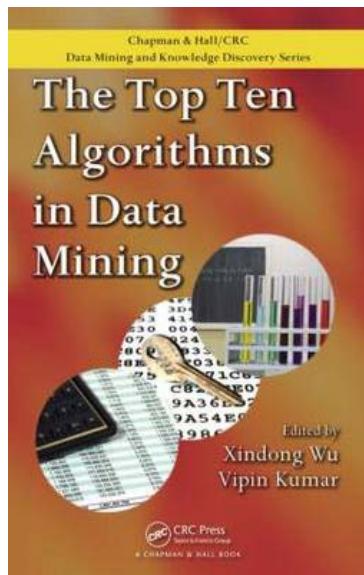
## Índice

- ¿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**
- Un caso de estudio:  
ECBDL'14 Competición Big Data
- Comentarios Finales



¿Qué algoritmos puedo encontrar para Hadoop?

Analizamos 10 algoritmos muy conocidos



Decision trees (C4.5, Cart) (MReC4.5)

K-Means

SVM

Apriori

kNN

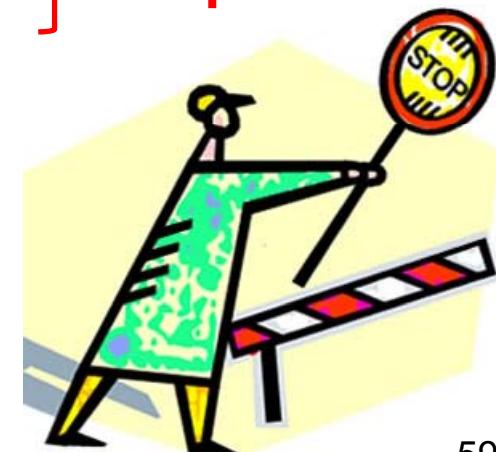
Naïve Bayes

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 iterativos de MapReduce](#))

# Limitaciones de MapReduce

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



**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)

Pregel: A System for Large-Scale Graph Processing

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

Malewicz, G., Austern, M., Bik, A., Dehnert, J., Horn, I., Leiser, N., and Czajkowski, G. Pregel: A system for large escale graph processing. ACM SIGMOD 2010.

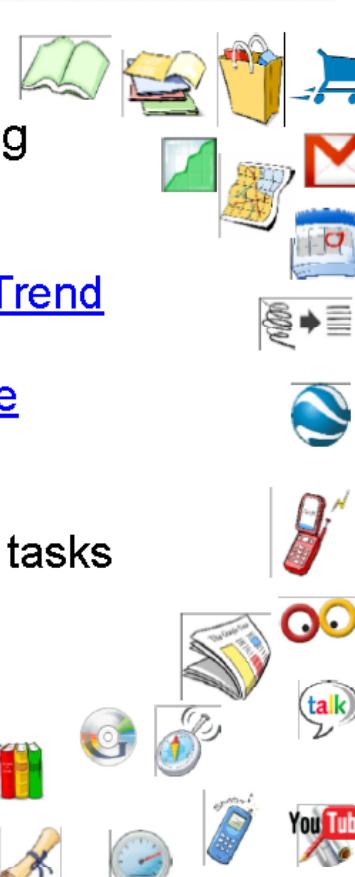
# Limitaciones de MapReduce

# MapReduce inside Google

## Googlers' hammer for 80% of our data crunching

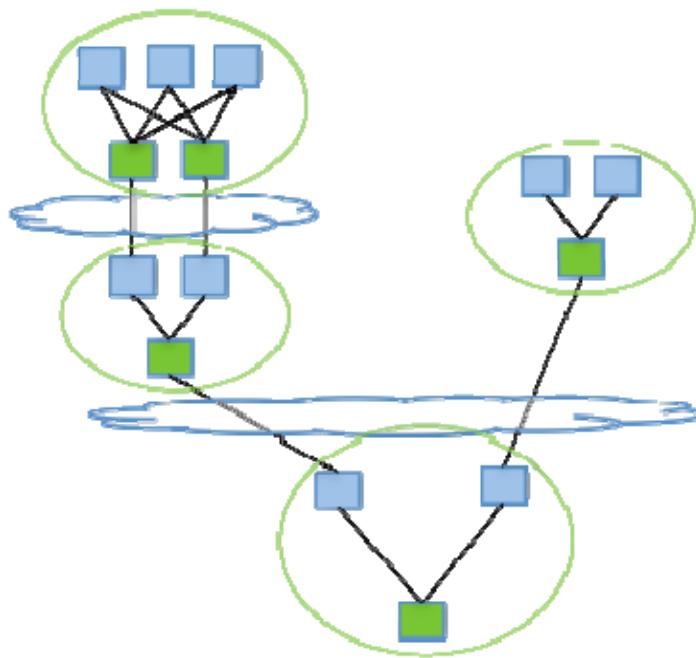
- Large-scale web search indexing
  - Clustering problems for Google News
  - Produce reports for popular queries, e.g. Google Trend
  - Processing of satellite imagery data
  - Language model processing for statistical machine translation
  - Large-scale machine learning problems
  - Just a plain tool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

The other 20%? e.g. [Pregel](#)

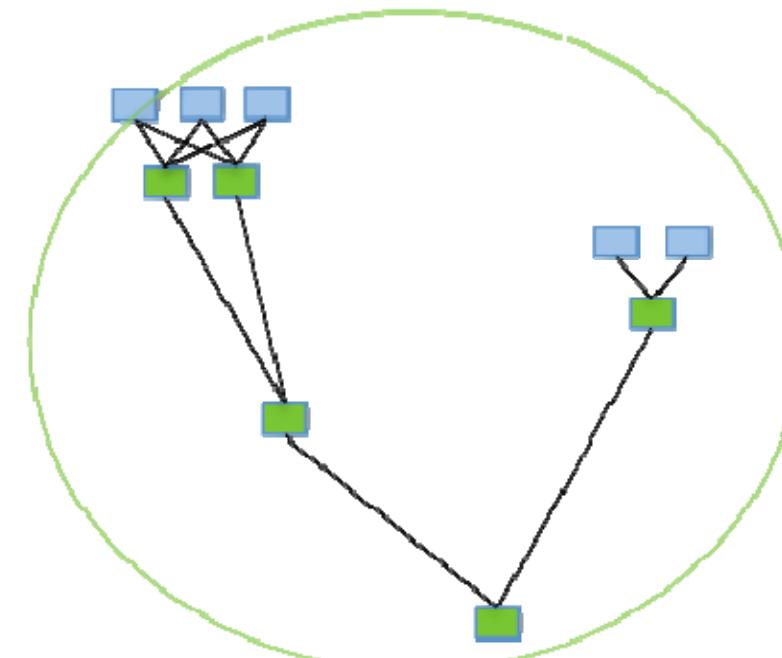


# Limitaciones de MapReduce

## Procesos con flujos acíclicos de procesamiento de datos



Pig/Hive - MR

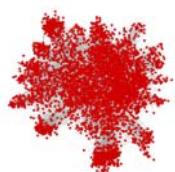


Pig/Hive - Tez

# Limitaciones de MapReduce



**GIRAPH (APACHE Project)**  
[\(http://giraph.apache.org/\)](http://giraph.apache.org/)  
*Procesamiento iterativo 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)**  
(Apache Foundation)  
<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  
GPMR



# Limitaciones de MapReduce



## Spark (UC Berkeley)

<http://spark.incubator.apache.org/research.html>. It

started as a research project at UC Berkeley in the [AMPLab](#), which focuses on big data analytics. It introduces the resilient distributed datasets (RDD) abstraction, allowing iterative algorithms. It's about 100 times faster than Hadoop.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M., Shenker, S., and Stoica, I. Resilient Distributed Datasets: A fault-tolerant abstraction for in-memory cluster computing. NSDI 2012.

## ML Lib Spark

### 2014 - Hadoop Spark

- Classification and regression
  - linear support vector machine (SVM)
  - logistic regression
  - linear least squares, Lasso, and ridge regression
  - decision tree
  - naive Bayes

### Big Data “in-memory”

<https://spark.apache.org/docs/latest/mllib-guide.html>

## Índice

- ¿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
- **Un caso de estudio:  
ECBDL'14 Competición Big Data**
- Comentarios Finales



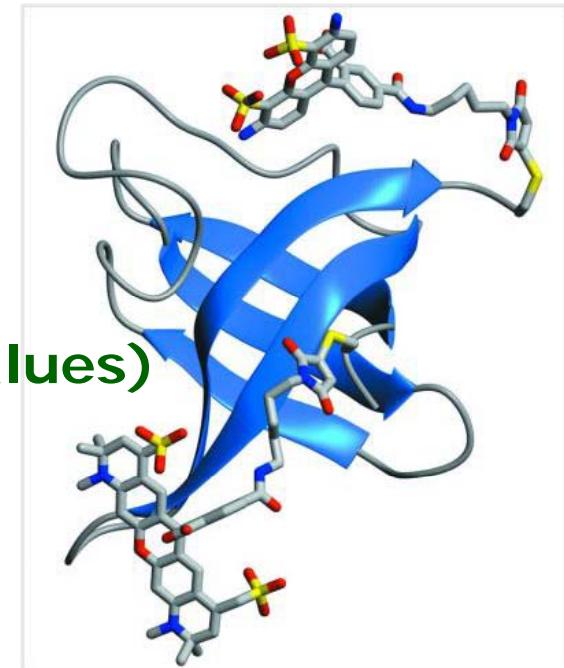
# Evolutionary Computation for Big Data and Big Learning Workshop

ECBDL'14 Big Data Competition 2014: Self-deployment track

**Objective:** Contact map prediction

**Details:**

- 32 million instances
- 631 attributes (**539 real & 92 nominal values**)
- 2 classes
- 98% of negative examples
- About 56.7GB of disk space



**Evaluation:**

**True positive rate · True negative rate**  
**TPR · TNR**

<http://cruncher.ncl.ac.uk/bdcomp/index.pl?action=data>

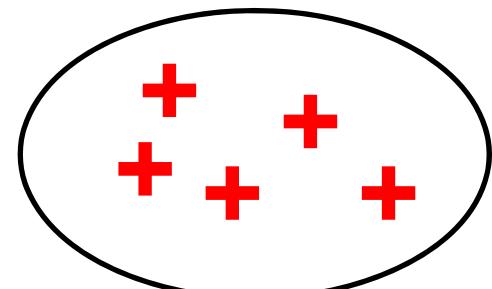
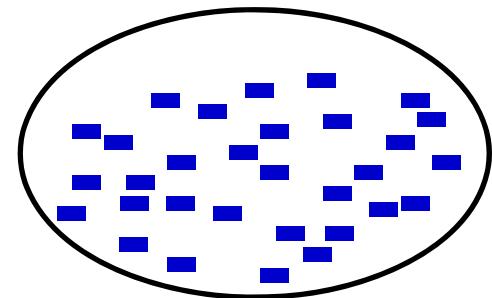
J. Bacardit et al, Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features, Bioinformatics 28 (19) (2012) 2441-2448

# Evolutionary Computation for Big Data and Big Learning Workshop

ECBDL'14 Big Data Competition 2014: Self-deployment track

## The challenge:

- ❑ Very large size of the training set
    - ❑ Does not fit all together in memory.
  - ❑ Even large for the test set (5.1GB, 2.9 million instances)
  - ❑ Relatively high dimensional data.
  - ❑ Low ratio (<2%) of true contacts.  
Imbalance rate: > 49
    - ❑ Unbalanced problem!



# ECBDL'14 Competición Big Data

---

## ECBDL'14 Big Data Competition 2014: Self-deployment track

### Nuestra propuesta:

1. Balancear los datos de entrenamiento  
(primera idea, que fué extendida)
  - Random Oversampling
2. Detección de características relevantes
  1. Evolutionary Feature Weighting
3. Modelo de aprendizaje
  - RandomForest

Clasificar el conjunto de test

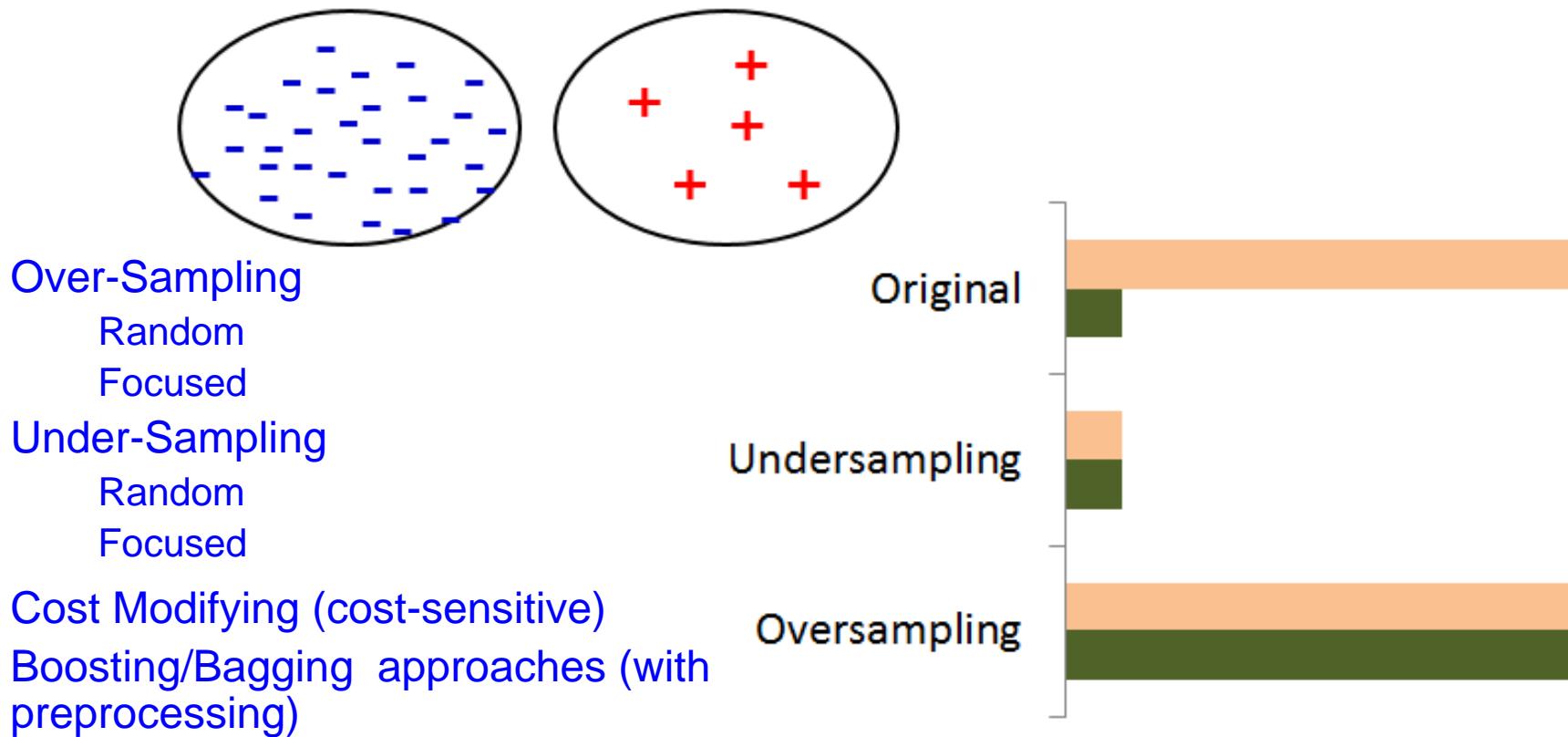


# ECBDL'14 Competición Big Data

## A MapReduce Approach for Random Oversampling

Low ratio of true contacts (<2%).

Imbalance rate: > 49. **Unbalanced problem!**

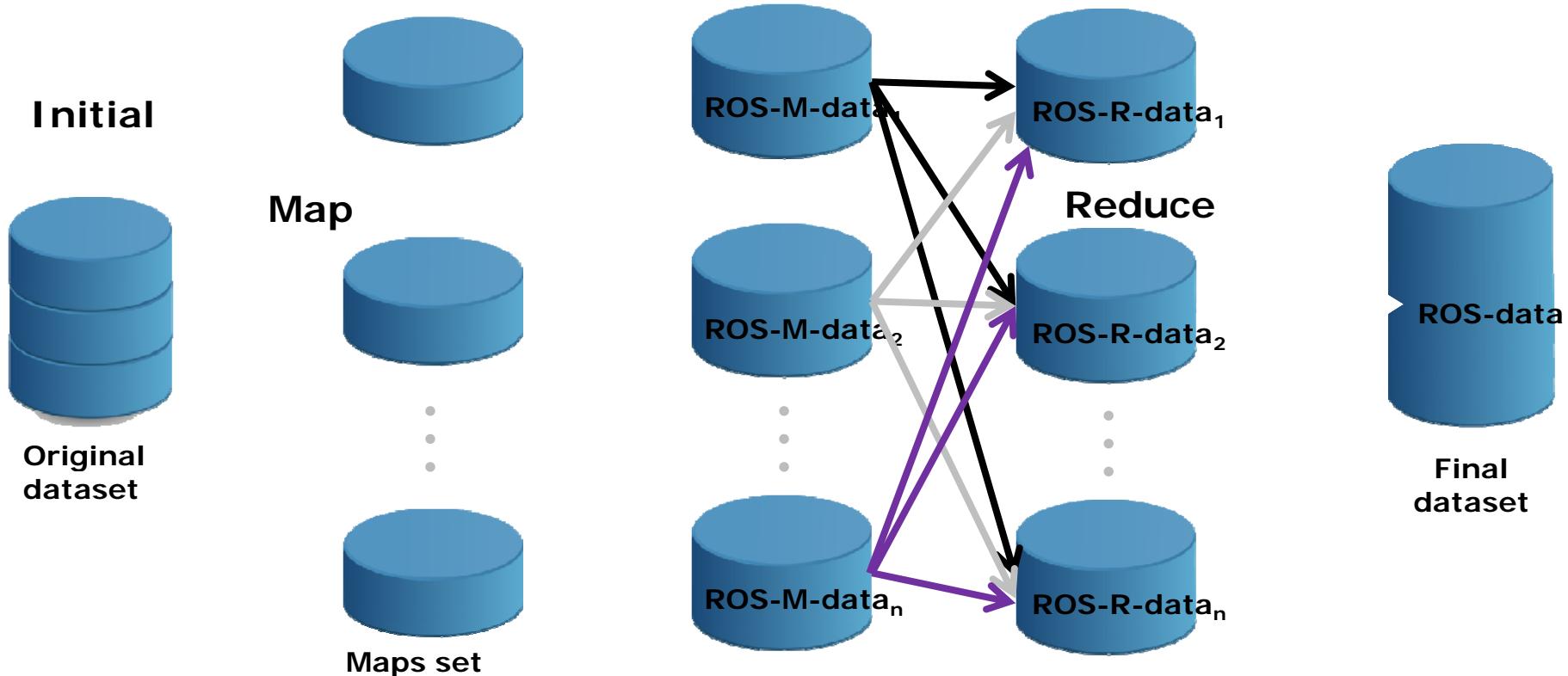


# ECBDL'14 Competición Big Data

## A MapReduce Approach for Random Oversampling

Low ratio of true contacts (<2%).

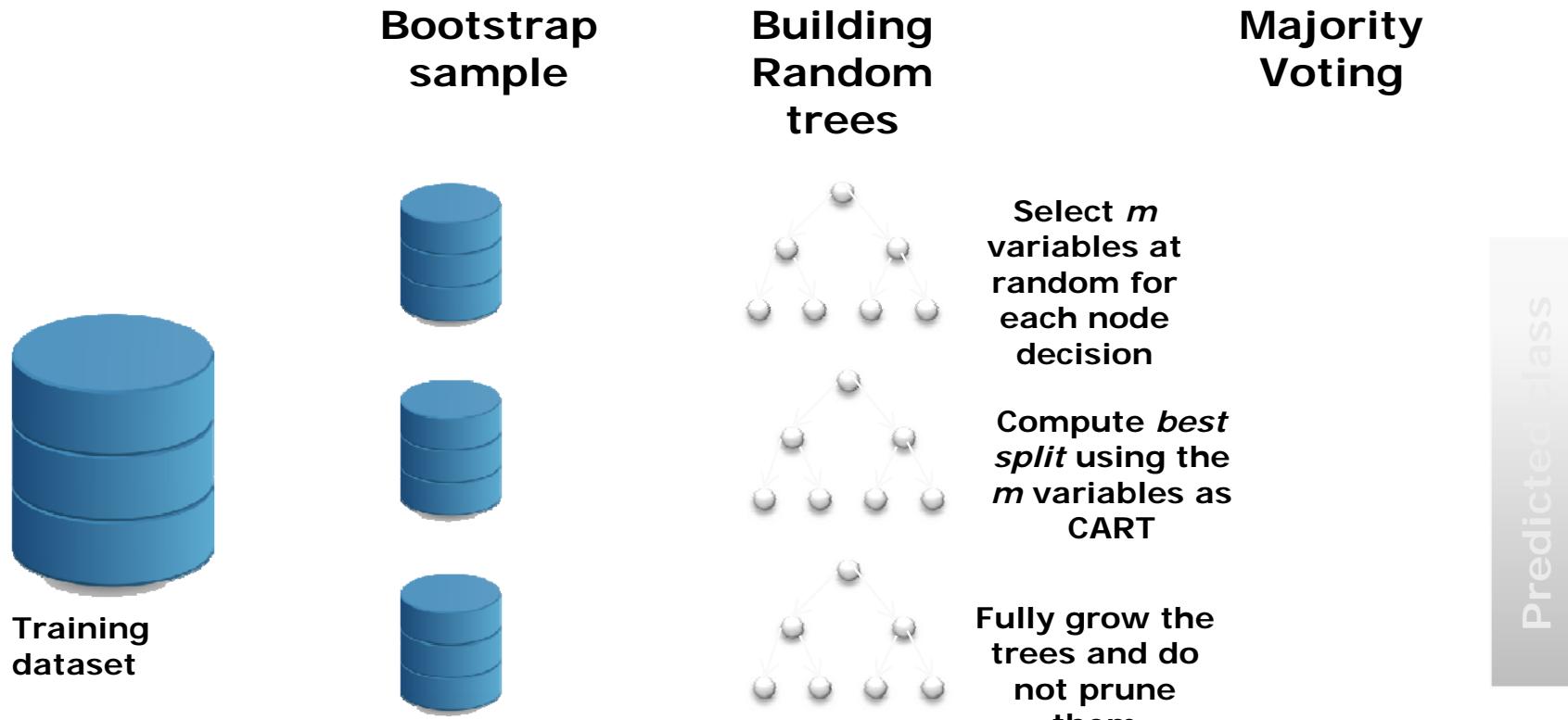
Imbalance rate: > 49. **Unbalanced problem!**



S. Río, V. López, J.M. Benítez, F. Herrera, On the use of MapReduce for Imbalanced Big Data using Random Forest. *Information Sciences*, in press 2014.

# ECBDL'14 Competición Big Data

## Building a model with Random Forest



# ECBDL'14 Competición Big Data

---

We initially focused on

- ❑ Oversampling rate: 100%

RandomForest:

- ❑ Number of used features:  $10 (\log n + 1)$ ; Number of trees: 100
- ❑ Number of maps: {64, 190, 1024, 2048}

Nº mappers	TPR_tst	TNR_tst	TNR*TPR Test
64	0,601723	0,806269	0,485151

**Very low TPR (relevant!)**

**How to increase the TPR rate?**

**Idea: To increase the ROS percentaje**

# ECBDL'14 Competición Big Data

## How to increase the TPR rate?

Idea: To increase the ROS percentaje

- ❑ Oversampling rate: {100, 105, 110m 115,130}

RandomForest:

- ❑ Number of used features: 10; Number of trees: 100

Algorithms	TPR	TNR	TNR*TPR
Test			
<b>ROS+RF (RS: 100%)</b>	<b>0.6351</b>	<b>0.7733</b>	<b>0.491186</b>
ROS+RF (RS: 105%)	0.6568	0.7555	0.496286
ROS+RF (RS: 110%)	0.6759	0.7337	0.495941
<b>ROS+RF (RS: 115%)</b>	<b>0.7041</b>	<b>0.7103</b>	<b>0.500175</b>
ROS+RF (RS: 130%)	0.7472	0.6609	0.493913

The higher ROS percentage, the higher TPR  
and the lower TNR

# ECBDL'14 Competición Big Data

---

## Third component: MapReduce Approach for Feature Weighting for getting a major equilibrium between classes

### Map Side

- ❑ Each map read one block from dataset.
- ❑ Perform an **Evolutionary Feature Weighting** step.
- ❑ **Output:** a real vector that represents the degree of importance of each feature.
- ❑ Number of maps: 32768 (**less than 1000 original data per map**)

### Reduce Side

- ❑ Aggregate the feature's weights
- ❑ A feature is finally selected if it overcomes a given threshold.
- ❑ **Output:** a binary vector that represents the final selection

I. Triguero, J. Derrac, S. García, F. Herrera, Integrating a Differential Evolution Feature Weighting scheme into Prototype Generation. Neurocomputing 97 (2012) 332-343

# ECBDL'14 Competición Big Data

---

## Experimental study

Random Oversampling:

- ❑ Oversampling ratio. Analyzed values: { 100 to 130}

**Feature Weighting:**

- ❑ Threshold --> number of selected features.
- ❑ Set of features: { 19, 63, 90, 146}
- ❑ Number of maps: 32768

RandomForest:

- ❑ Number of used features: { log NumFeatures, 2 \* Log +1 }
- ❑ Number of trees: { 100 }
- ❑ Number of maps: { 32, 64, 128, 190, 256, 512 }

# ECBDL'14 Competición Big Data

---

**Evolutionary Feature Weighting.**  
It allows us to construct several subset of features  
(changing the threshold).

Algorithms	64 mappers			
	TNR*TPR Training	TPR	TNR	TNR*TPR Test
ROS+RF (130% - Feature Weighting 63)	0.726350	0.66949	0.775652	0.519292
ROS+RF (115% - Feature Weighting 63)	0.736596	0.652692	0.790822	0.516163
ROS+RF (100% - Feature Weighting 63)	0.752824	0.626190	0.811176	0.507950

# ECBDL'14 Competición Big Data

---

We decided to investigate:

- a) On the Random Forest: the influence of the Random Forest's parameters (internal features and number of trees)
- b) Higher number of features (90) and ROS with 140%

Algorithms	190 mappers			
	TNR*TPR Training	TPR	TNR	TNR*TPR Test
ROS+ RF (130%+ FW 63+6f+100t)	0.604687	0.698152	0.742462	0.518351
ROS+ RF (130%+ FW 63+6f+200t)	0.632078	0.700064	0.745225	0.521705
ROS+ RF (140%+ FW 63+15f+200t)	0.627409	0.719678	0.728912	0.524582
ROS+ RF (140%+ FW 90+15f+200t)	0.635855	0.722639	0.726397	0.524923
ROS+ RF (140%+ FW 90+25f+200t)	0.629273	0.721652	0.729740	0.526618

Correct decisions with FW 90 and RF with 25f and 200 trees.  
Good trade off between TPR and TNR

# ECBDL'14 Competición Big Data

**Current state:**

Algorithms	190 mappers			
	TNR*TPR Training	TPR	TNR	TNR*TPR Test
ROS+ RF (140%+ FW 90+25f+200t)	0.629273	0.721652	0.729740	0.526618

**Our knowledge:**

Good configuration: “FW 90” and “RF with 25 f + 200t”

The higher ROS percentage, the higher TPR and the lower TNR

The less number of maps, the less TPR and the high TNR and high accuracy.

**4 days to finish the competition:**

Can we take decisions for improving the model?

# ECBDL'14 Competición Big Data

The less number of maps, the less TPR and the high TNR

Algorithms	190 mappers			
	TNR*TPR Training	TPR	TNR	TNR*TPR Test
ROS+ RF (140%+ FW 90+25f+200t)	0.629273	0.721652	0.729740	0.526618

Algorithms	64 mappers			
	TNR*TPR Training	TPR	TNR	TNR*TPR Test
ROS+ RF (130%+ FW 90+25f+200t)	0.736987	0.671279	0.783911	0.526223
ROS+ RF (140%+ FW 90+25f+200t)	0.717048	0.695109	0.763951	0.531029

64 mappers and we got 0.53

ROS 130 – 65 (140 – 68) replications of the minority instances

# ECBDL'14 Competición Big Data

Last decision: We investigated to increase ROS until 180% with the number of mappers (64)

Algorithms	64 mappers			
	TNR*TPR Training	TPR	TNR	TNR*TPR Test
ROS+ RF (130%+ FW 90+25f+200t)	0.736987	0.671279	0.783911	0.526223
ROS+ RF (140%+ FW 90+25f+200t)	0.717048	0.695109	0.763951	0.531029
ROS+ RF (150%+ FW 90+25f+200t)	0.706934	0.705882	0.753625	0.531971
ROS+ RF (160%+ FW 90+25f+200t)	0.698769	0.718692	0.741976	0.533252
ROS+ RF (170%+ FW 90+25f+200t)	0.682910	0.730432	0.730183	0.533349
ROS+ RF (180%+ FW 90+25f+200t)	0.678986	0.737381	0.722583	0.532819

To increase ROS and reduce the mappers number lead us to get a trade-off with good results

ROS 170 – 85 replications of the minority instances

# ECBDL'14 Competición Big Data

---

## Evolutionary Computation for Big Data and Big Learning Workshop

**Results of the competition: Contact map prediction**

Team Name	TPR	TNR	Acc	TPR · TNR
Efdamis	<b>0.730432</b>	<b>0.730183</b>	<b>0.730188</b>	<b>0.533349</b>
ICOS	0.703210	0.730155	0.729703	0.513452
UNSW	<b>0.699159</b>	<b>0.727631</b>	<b>0.727153</b>	<b>0.508730</b>
HyperEns	0.640027	0.763378	0.761308	0.488583
PUC-Rio_ICA	0.657092	0.714599	0.713634	0.469558

**EF DAMIS team ranked first in the ECBDL'14 big data competition**

---

<http://cruncher.ncl.ac.uk/bdcomp/index.pl?action=ranking>

# ECBDL'14 Competición Big Data

## Evolutionary Computation for Big Data and Big Learning Workshop

### Results of the competition: Contact map prediction

Team Name	TPR	TNR	Acc	TPR · TNR
Efdamis	0.730432	0.730183	0.730188	0.533349
ICOS	0.703210	0.730155	0.729703	0.513452
UNSW	0.699159	0.727631	0.727153	0.508730

Algorithms	64 mappers			
	TNR*TPR Training	TPR	TNR	TNR*TPR Test
ROS+RF (130% - Feature Weighting 63)	0.726350	0.66949	0.775652	0.519292
ROS+RF (115% - Feature Weighting 63)	0.736596	0.652692	0.790822	0.516163
ROS+RF (100% - Feature Weighting 63)	0.752824	0.626190	0.811176	0.507950

To increase ROS and to use Evolutionary feature weighting were two good decisions for getting the first position

# ECBDL'14 Competición Big Data

AI comienzo

**ROS+RF (RS: 100%)**

Nº mappers	TPR_tst	TNR_tst	TNR*TPR Test
64	0,601723	0,806269	0,485151

AI final (30 de Junio)

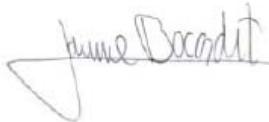
Algorithms	64 mappers			
	TNR*TPR Training	TPR	TNR	TNR*TPR Test
<b>ROS+ RF (160%+ FW 90+25f+200t)</b>	<b>0,698769</b>	<b>0.718692</b>	<b>0.741976</b>	<b>0.533252</b>
<b>ROS+ RF (170%+ FW 90+25f+200t)</b>	<b>0.682910</b>	<b>0.730432</b>	<b>0.730183</b>	<b>0.533349</b>
<b>ROS+ RF (180%+ FW 90+25f+200t)</b>	<b>0,678986</b>	<b>0.737381</b>	<b>0.722583</b>	<b>0.532819</b>

# ECBTL'14 Competición Big Data

---

**ECBTL'14: Evolutionary Computation for  
Big Data and Big Learning Workshop**  
**July 13<sup>th</sup>, 2014**  
**GECCO-2014, Vancouver, Canada**

This is to certify that team EFDAMIS, formed by Isaac Triguero, Sara del Río, Victoria López, José Manuel Benítez and Francisco Herrera, ranked **first** in the ECBTL'14 big data competition



---

Jaume Bacardit, organizer  
ECBTL'14 big data competition



## Índice

---

- ¿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
- Un caso de estudio:  
ECBDL'14 Competición Big Data
- **Comentarios Finales**



# Comentarios Finales

---

- La parallelización de los algoritmos de aprendizaje automático mediante el particionamiento de los datos puede abordarse exitósamente con MapReduce.
- El particionamiento permite aplicar el algoritmo de aprendizaje a cada bloque.
- La fase “**Reduce**” centrada en la combinación de los modelos obtenidos a partir de la fase MAP es el gran reto en el diseño de algoritmos.
- Spark combinado con HDFS Hadoop puede convertirse en la tecnología estrella a medio plazo. Mahout desarrollará su siguiente versión sobre Spark.

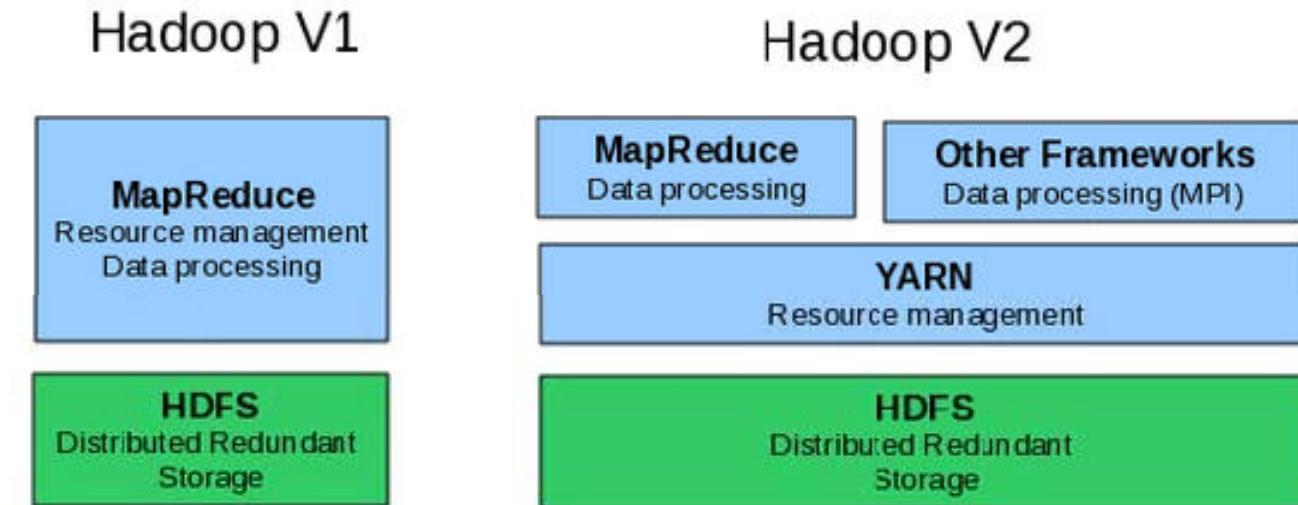
**Bibliografía:** A. Fernandez, S. Río, V. López, A. Bawakid, M.J. del Jesus, J.M. Benítez, F. Herrera, **Big Data with Cloud Computing: An Insight on the Computing Environment, MapReduce and Programming Frameworks.** *WIREs Data Mining and Knowledge Discovery, in press (2014).*

<b>Generation</b>	<b>1st Generation</b>	<b>2nd Generation</b>	<b>3rd Generation</b>
Examples	SAS, R, Weka, SPSS, KEEL	Mahout, Pentaho, Cascading	Spark, Haloop, GraphLab, Pregel, Giraph, ML over Storm
Scalability	Vertical	Horizontal (over Hadoop)	Horizontal (Beyond Hadoop)
Algorithms Available	Huge collection of algorithms	Small subset: sequential logistic regression, linear SVMs, Stochastic Gradient Descent, k-means clustering, Random forest, etc.	Much wider: CGD, ALS, collaborative filtering, kernel SVM, matrix factorization, Gibbs sampling, etc.
Algorithms Not Available	Practically nothing	Vast no.: Kernel SVMs, Multivariate Logistic Regression, Conjugate Gradient Descent, ALS, etc.	Multivariate logistic regression in general form, k-means clustering, etc. – Work in progress to expand the set of available algorithms
Fault-Tolerance	Single point of failure	Most tools are FT, as they are built on top of Hadoop	FT: HaLoop, Spark Not FT: Pregel, GraphLab, Giraph

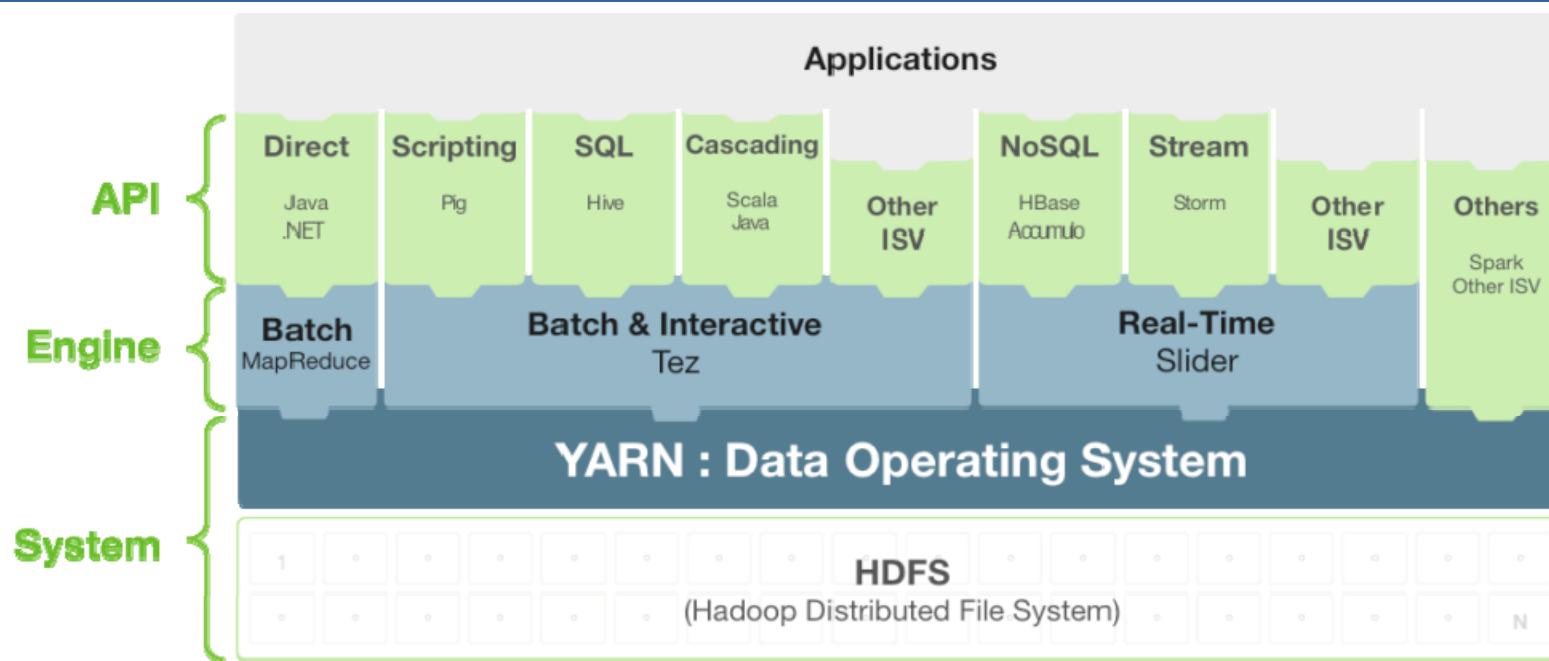
# Comentarios Finales

**BIG  
DATA**

## Evolución de Hadoop



# Comentarios Finales

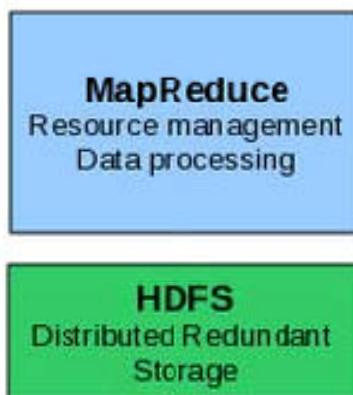


**Apache Hadoop YARN** es el sistema operativo de datos de Hadoop 2, responsable de la gestión del acceso a los recursos críticos de Hadoop. YARN permite al usuario interactuar con todos los datos de múltiples maneras al mismo tiempo, haciendo de Hadoop una verdadera plataforma de datos multi-uso y lo que le permite tomar su lugar en una arquitectura de datos moderna.

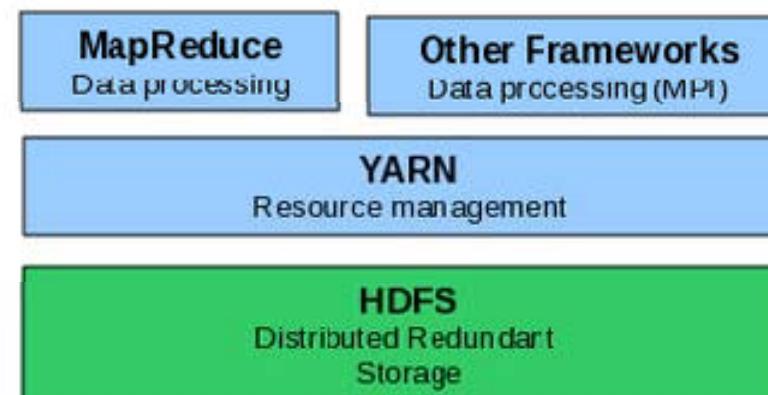
# Comentarios Finales

BIG  
DATA

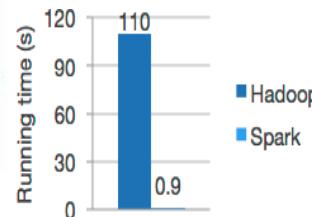
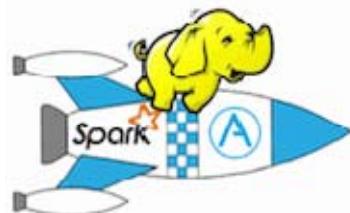
Hadoop V1



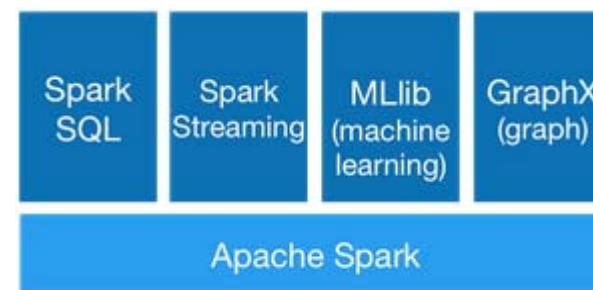
Hadoop V2



Enfoque InMemory  
HDFS Hadoop + SPARK



Ecosistema  
Apache Spark



Futura versión de  
Mahout con Spark

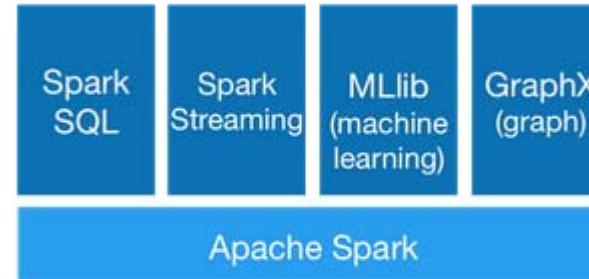


# Comentarios Finales

BIG  
DATA

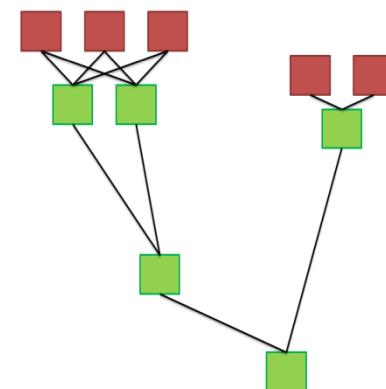


- Ecosistema Apache Spark



• **Big Data “in-memory”.** Spark permite realizar trabajos paralelizados totalmente en memoria, lo cual reduce mucho los tiempos de procesamiento. Sobre todo si se trata de unos procesos iterativos. En el caso de que algunos datos no quepan en la memoria, Spark seguirá trabajando y usará el disco duro para volcar aquellos datos que no se necesitan en este momento (Hadoop “commodity hardware”).

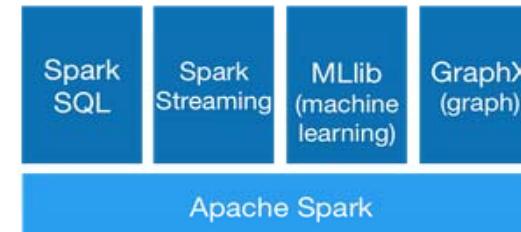
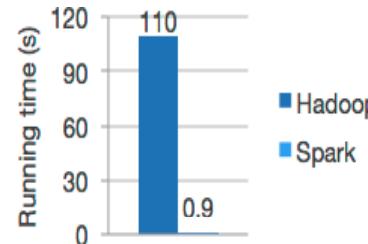
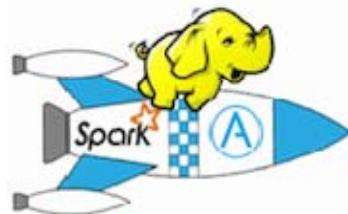
- Spark ofrece una API para Java, Python y Scala
- Esquema de computación más flexible que MapReduce.  
Permite la flujo acíclicos de procesamiento de datos



# Comentarios Finales

**BIG  
DATA**

## Enfoque InMemory



## Futura versión de Mahout con Spark



Spark no es estable y todavía no está listo para la producción

Spark todavía es un software joven y que no goza de toda la estabilidad de Hadoop-MapReduce.

# 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

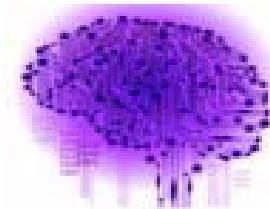
# Comentarios Finales

**BIG  
DATA**

## Desafíos en Big Data

<http://www.kdnuggets.com/2013/12/3-stages-big-data.html>

By Gregory Piatetsky, Dec 8, 2013.



Big Data 3.0:  
Intelligent



Big Data 2.0:  
Networked



Big Data 1.0:  
Transactional

**In many new applications - face recognition, speech understanding, recommendations, or fraud detection - bigger data does produces better results**

To help clarify the different meanings of "Big Data", Dr. Piatetsky proposes to consider 3 stages of Big Data.

# Comentarios Finales

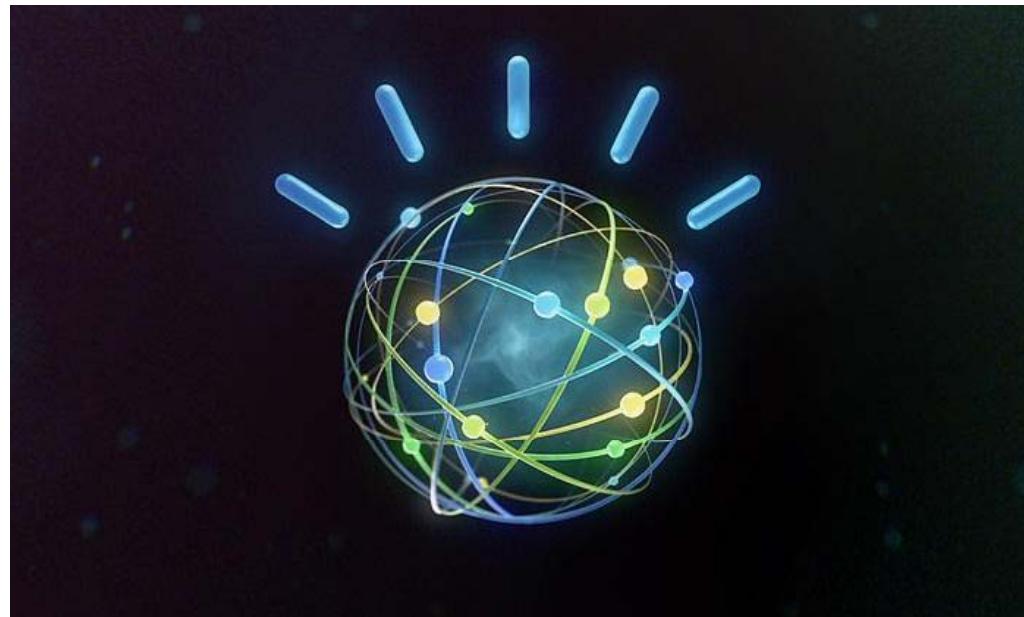


## Desafíos en Big Data



<http://www.ibtimes.com/ibm-watson-api-coming-3-potential-business-applications-ibms-watson-cloud-ecosystem-1470694>

By [Dave Smith](#) on November 14 2013 12:22 PM



IBM is preparing its Watson supercomputer technology to be utilized by third-party developers for the first time via a Watson cloud service called the “Watson Ecosystem.”

1. Watson the Shopping Companion
2. Watson the Journalist.
3. Watson the Nurse.

# Comentarios Finales

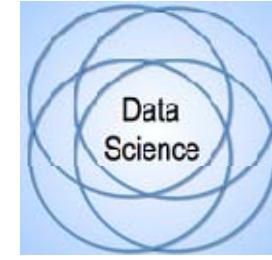
# BIG DATA

# Oportunidades en Big Data

**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.**



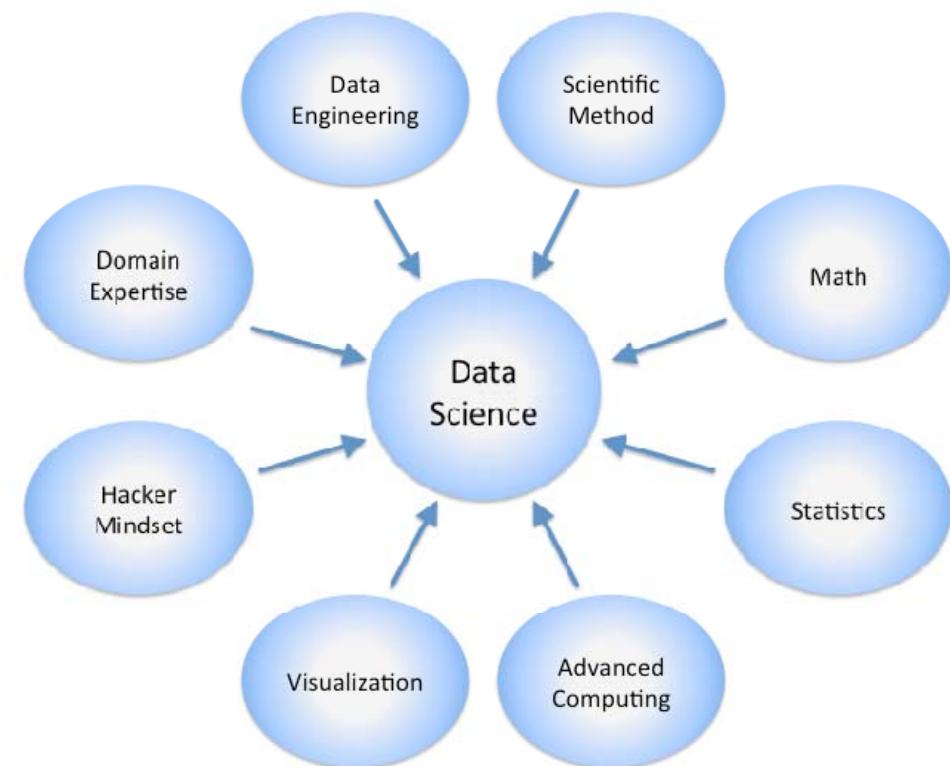
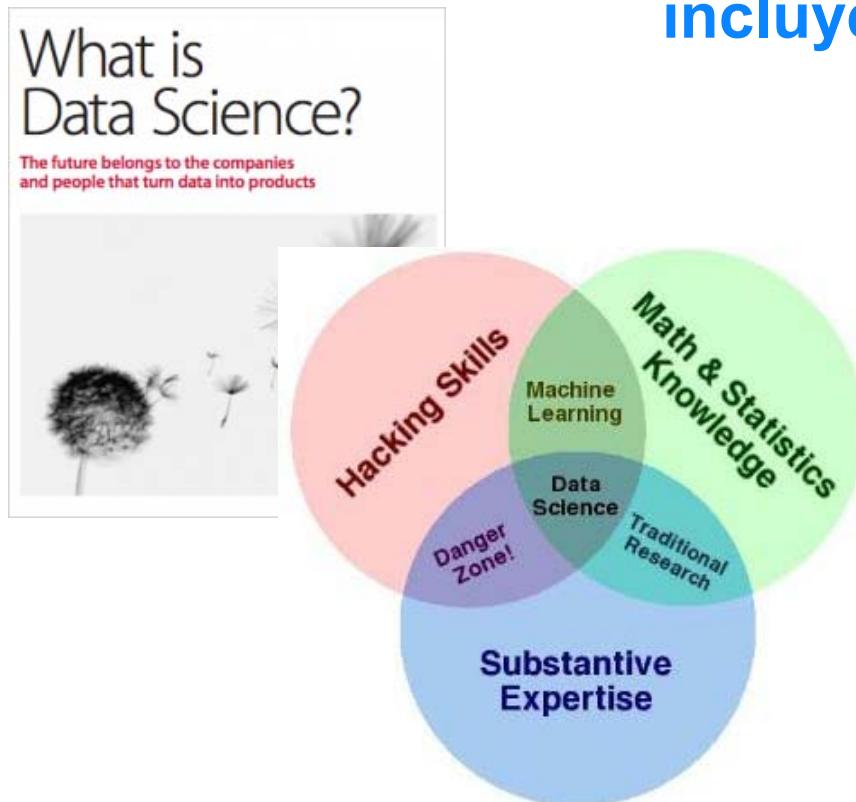
# Comentarios Finales



**BIG  
DATA**

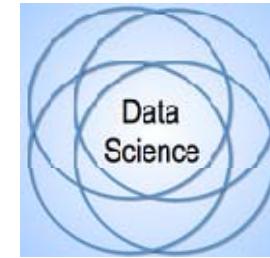
## Data Science

**Ciencia de Datos es el ámbito de conocimiento que engloba las habilidades asociadas al procesamiento de datos, incluyendo Big Data**



# Comentarios Finales

Surge como profesión el “Científico de Datos”



**BIG  
DATA**

## Científico de Datos

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

Gartner

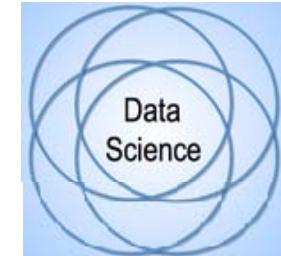
Fuente: <http://www.gartner.com/technology/topics/big-data.jsp>

**BIGdata**  
**dataSCIENTIST** automation predictive torrents  
**DataScience** streams Analytics INSIGHTS logicalDATAwarehouse



# Comentarios Finales

Una demanda creciente de profesionales en “Big Data” y “Ciencia de Datos”



## Oportunidades en Big Data

La demanda de profesionales formados en Ciencia de Datos y *Big Data* es enorme.

Se estima que la conversión de datos en información útil generará un mercado de 132.000 millones de dólares en 2015 y que se crearán más de 4.4 millones de empleos.

España necesitará para 2015 más de 60.000 profesionales con formación en Ciencia de Datos y *Big Data*.

### El maná de los datos

■ La conversión de datos en información útil para las empresas generará un mercado de 132.000 millones de dólares en 2015. La herramienta ‘big data’ sacará del mercado a quien no la use

SUSANA BLÁZQUEZ | Madrid | 29 SEP 2013 - 01:00 CET

10

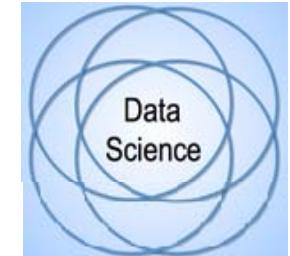
Archivado en: Citigroup Cap Gemini Sogeti SAP Oracle ING Bank BBVA Mapfre Bases datos IBM Telefónica Aplicaciones informáticas Tecnología Empresas Programas informáticos Economía



[http://economia.elpais.com/economia/2013/09/27/actualidad/1380283725\\_938376.html](http://economia.elpais.com/economia/2013/09/27/actualidad/1380283725_938376.html)

# Comentarios Finales

Una demanda creciente de profesionales en “Big Data” y “Ciencia de Datos”



## Oportunidades en Big Data (en España)

[http://www.revistacloudcomputing.com/2013/10/espana-necesitara-60-000-profesionales-de-big-data-hasta-2015/?goback=.gde\\_4377072\\_member\\_5811011886832984067#!](http://www.revistacloudcomputing.com/2013/10/espana-necesitara-60-000-profesionales-de-big-data-hasta-2015/?goback=.gde_4377072_member_5811011886832984067#!)

### España necesitará 60.000 profesionales de Big Data hasta 2015

22 octubre, 2013 Eventos 18



España necesitará 60.000 profesionales de Big Data hasta 2015

“España va a necesitar alrededor de sesenta mil profesionales del Big Data de aquí a 2015”, así lo ha asegurado Francisco Javier Antón, Subdirector General de Tecnologías del Ministerio de Educación, Cultura y Deportes en una mesa redonda sobre beneficio y aplicación de Big Data en pymes, moderada por Daniel Tapias de Sigma Technologies, celebrada durante el 4º Congreso Nacional de CENTAC de

“Existe una demanda mundial para formar a 4,4 millones de profesionales de la gestión Big Data desde ingenieros, gestores y científicos de datos”, comenta Antón. Sin embargo, “las empresas todavía no ven en el Big Data un modelo de negocio”, lamenta. “Solo se extrae un 1% de los datos disponibles en la red”, añade. “Hace falta formación y concienciación.

# Comentarios Finales

**BIG  
DATA**

[http://elpais.com/elpais/2013/12/02/vinetas/  
1386011115\\_645213.html](http://elpais.com/elpais/2013/12/02/vinetas/1386011115_645213.html)

## El Roto

Viñeta de El Roto  
3 de diciembre de 2013



**BIG  
DATA**



**BIG  
DATA**

