CURSOS DE VERANO 2014

APROXIMACIÓN PRÁCTICA A LA CIENCIA DE DATOS Y BIG DATA: HERRAMIENTAS KNIME, R, HADOOP Y MAHOUT

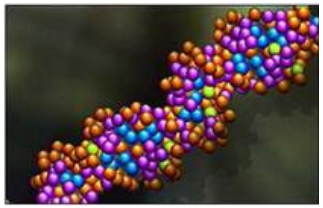
Introducción a Big Data

Francisco Herrera

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

Ben Chams - Fotolia

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


Objetivos de esta sesión:

- Introducir el concepto de Big Data.
- Introducir el paradigma MapReduce
- Introducir la plataforma de cómputo Hadoop y las nuevas tendencias (SPARK, …)

Big Data

- ¿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

Big Data

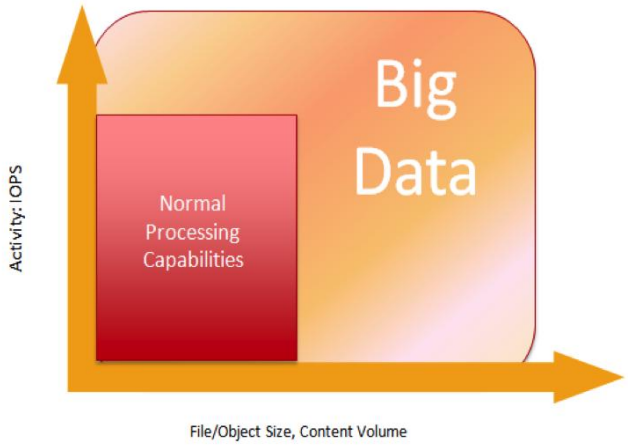
- **¿Qué es Big Data?**
- MapReduce: Paradigma de Programación para Big Data (Google)
- Plataforma Hadoop (Open access)
- Librería Mahout para Big Data. Otras librerías
- Limitaciones de MapReduce
- Un caso de estudio: ECBDL'14 Competición Big Data
- Comentarios Finales

¿Qué es Big Data?

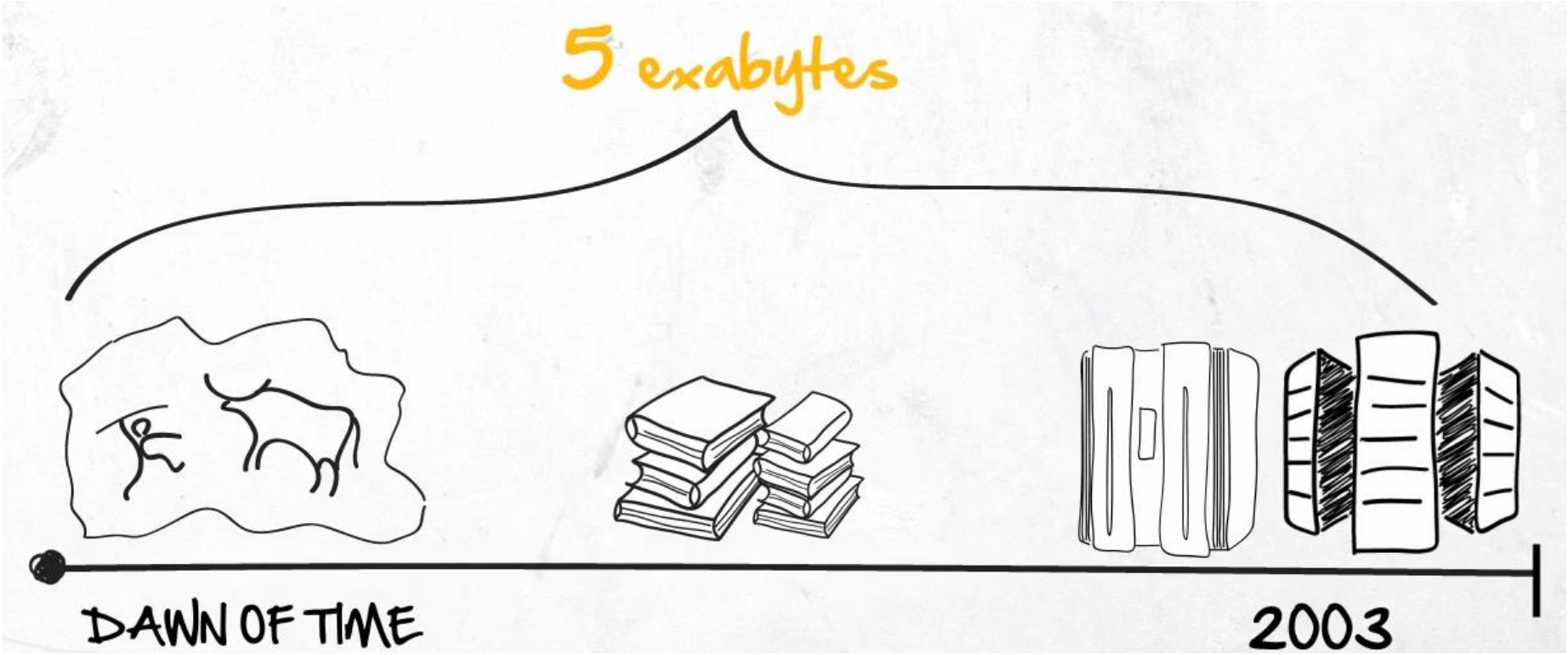
No hay una definición estándar

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

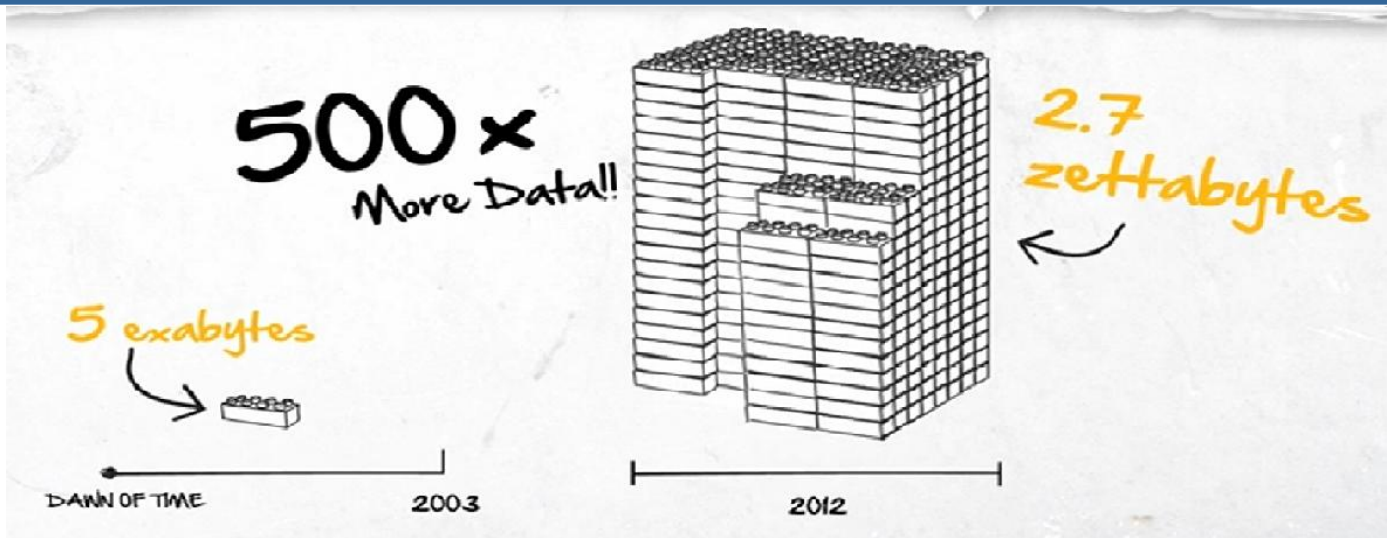
"*Big Data*" son datos cuyo volumen, diversidad y complejidad requieren nueva arquitectura, técnicas, algoritmos y análisis para gestionar y extraer valor y conocimiento oculto en ellos ...

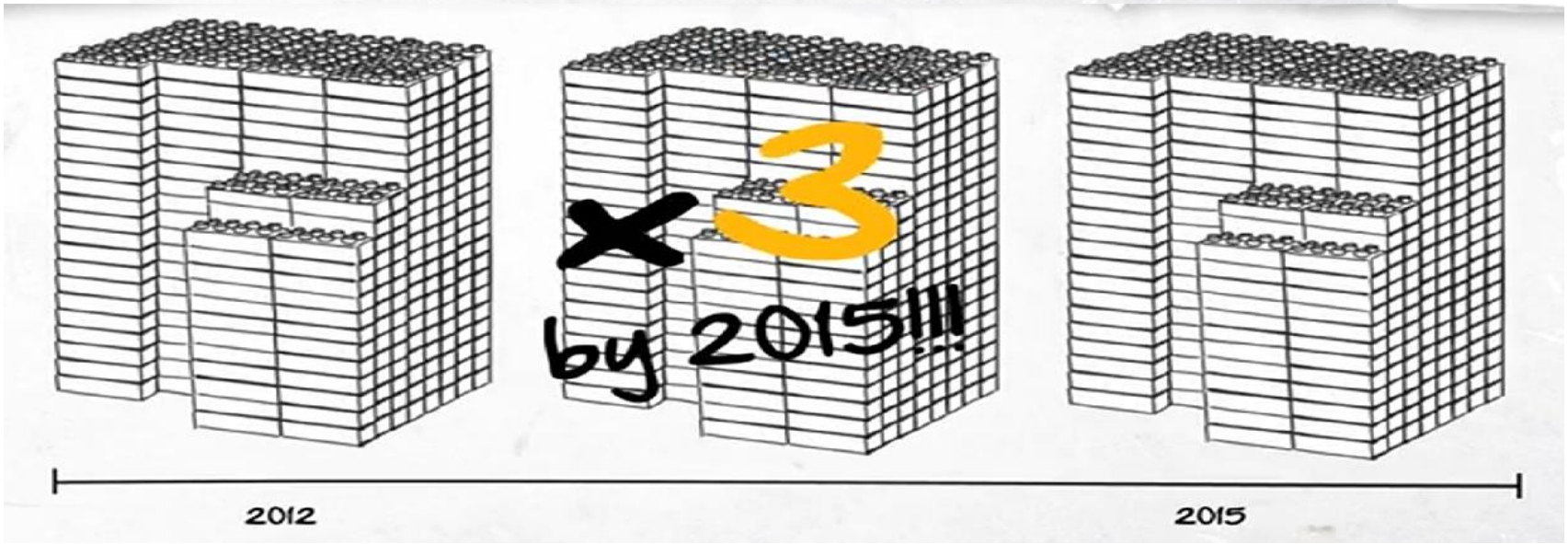


¿Qué es Big Data?



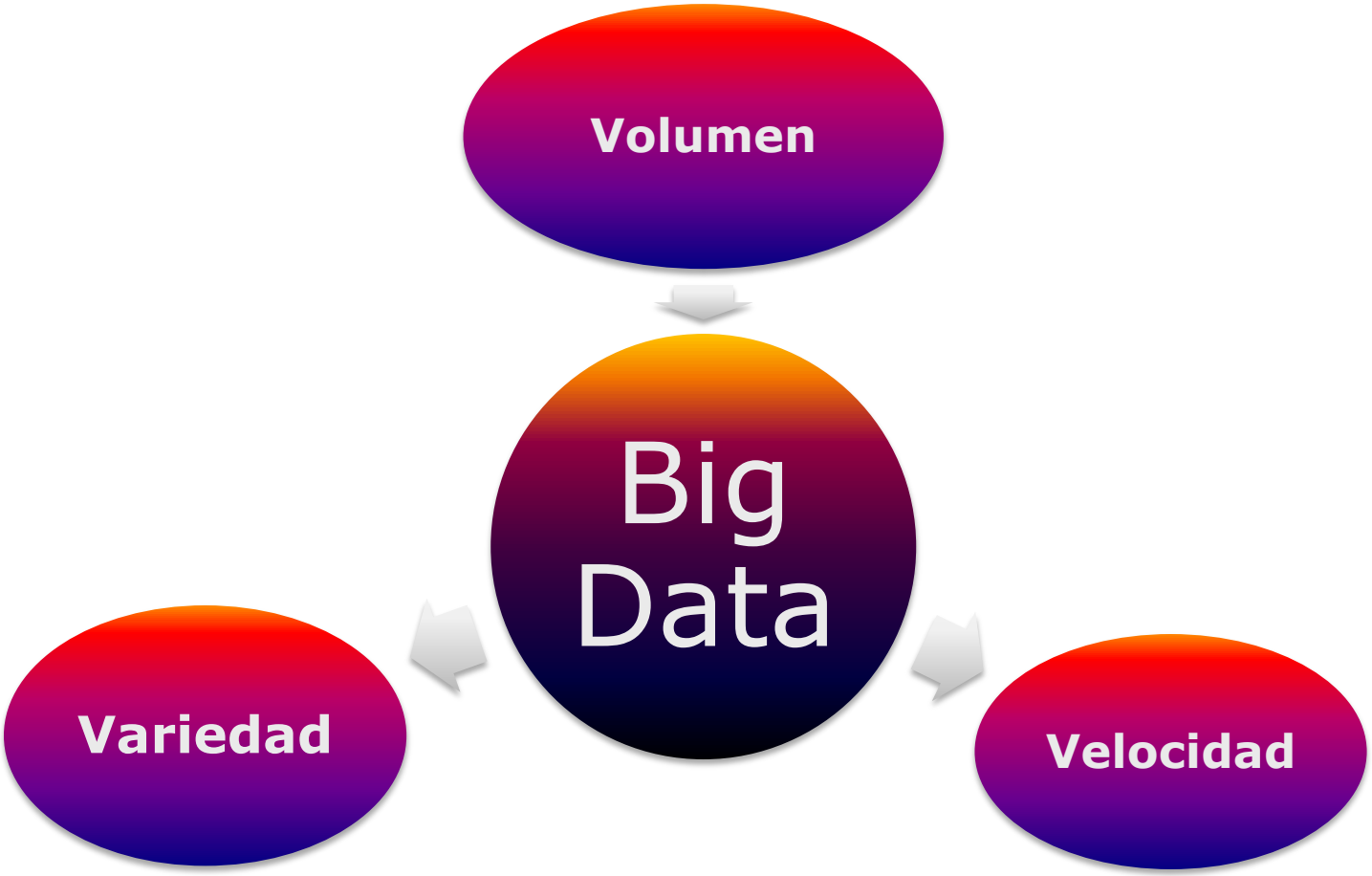
¿Qué es Big Data?





¿Qué es Big Data?

Las 3 V's de Big Data



¿Qué es Big Data? 3 V's de Big Data

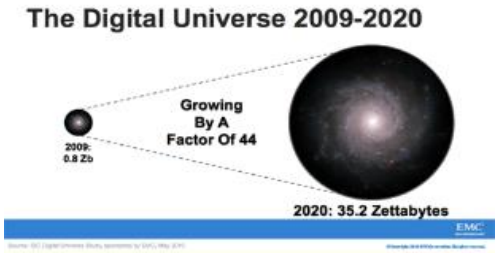
¿Qué es Big Data? 3 V's de Big Data

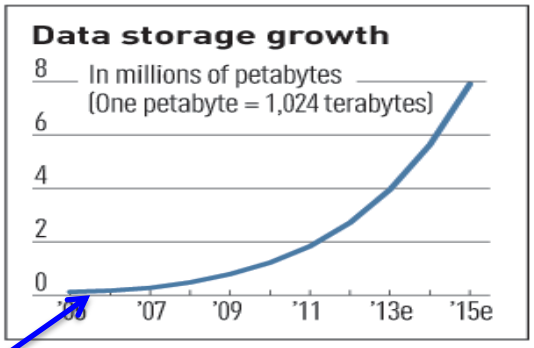
1ª:Volumen

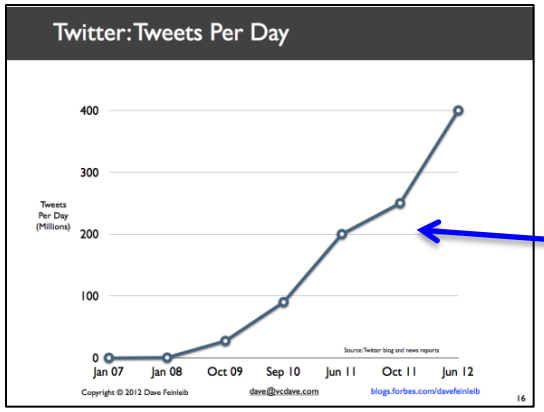
El volumen de datos

crece exponencialmente

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

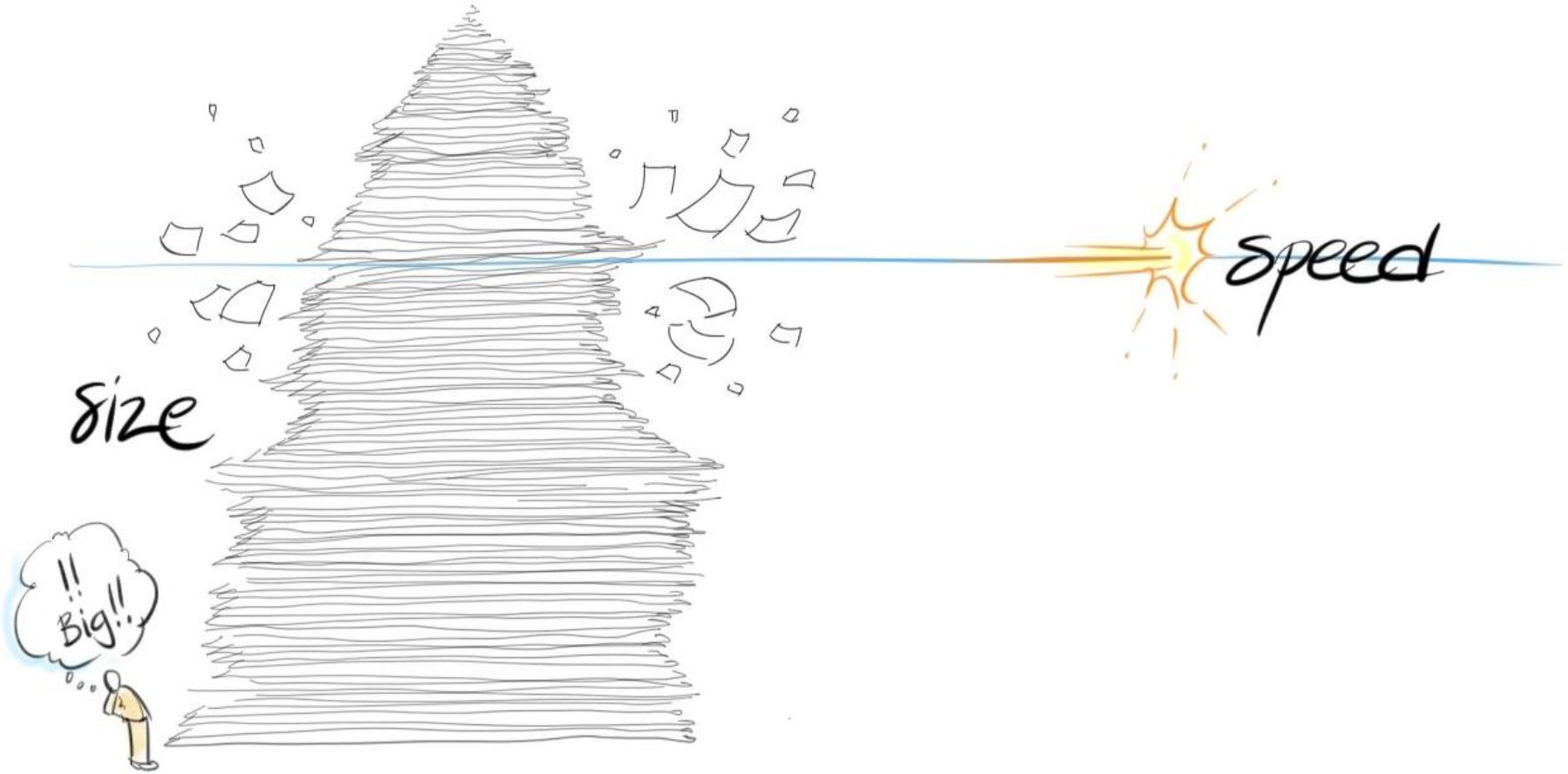






Crecimiento exponencial en los datos generados/almacenados

¿Qué es Big Data? 3 V's de Big Data



¿Qué es Big Data? 3 V's de Big Data

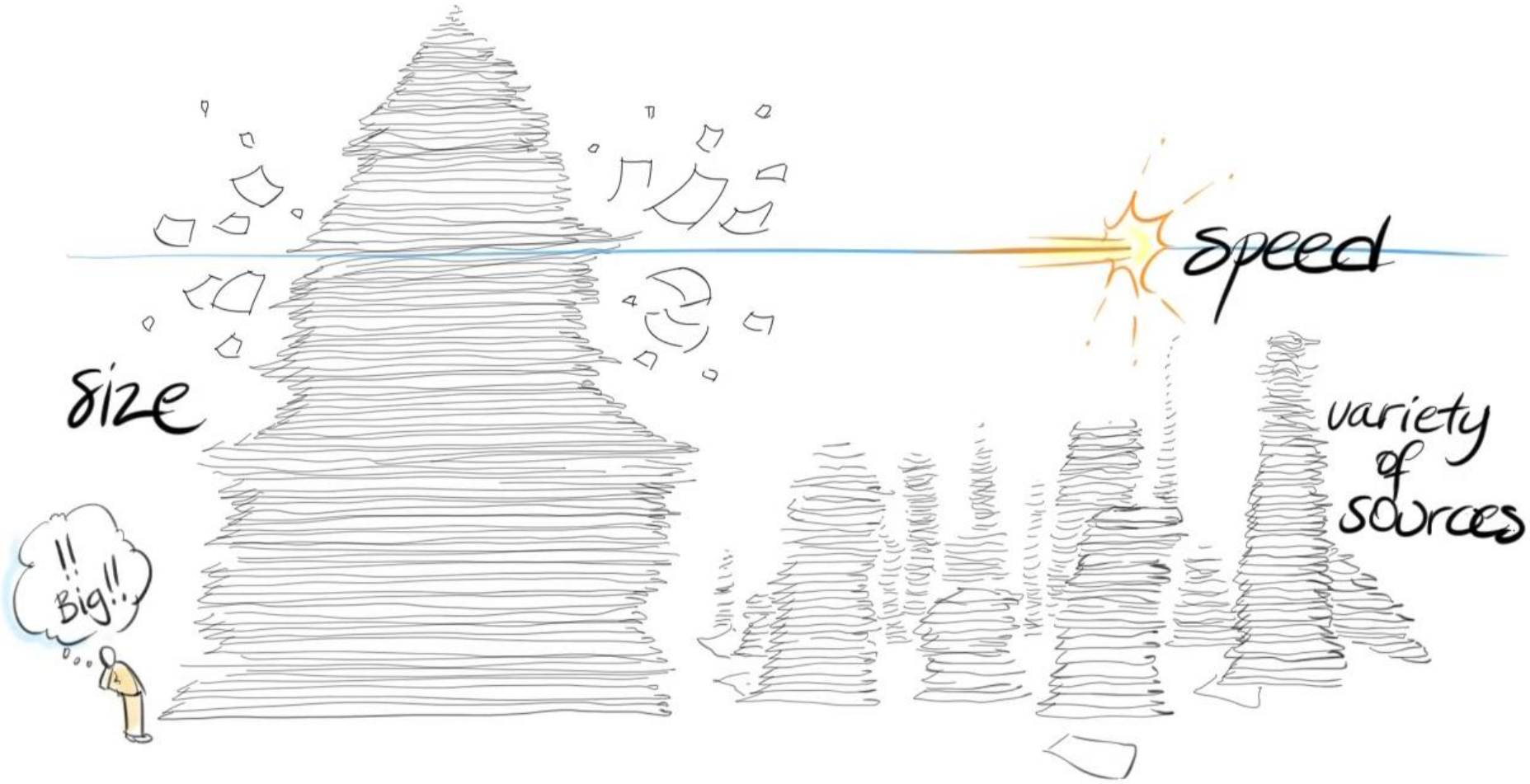
2ª:Velocidad

- **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



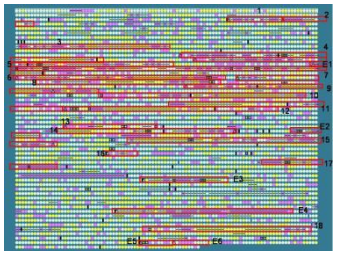
¿Qué es Big Data? 3 V's de Big Data

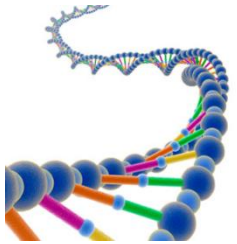
3ª:Variedad

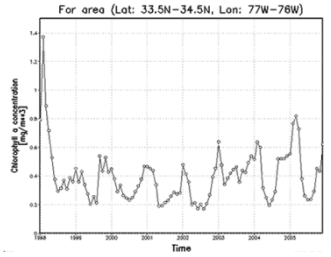
- Varios formatos y estructuras:

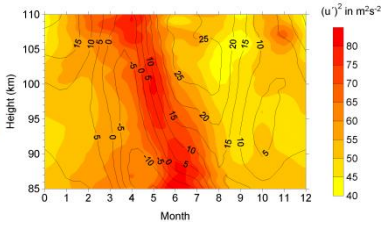
  Texto, numéricos, imágenes, audio, video, sequencias, series temporales

  …
- Una sola aplicación puede generar muchos tipos de datos







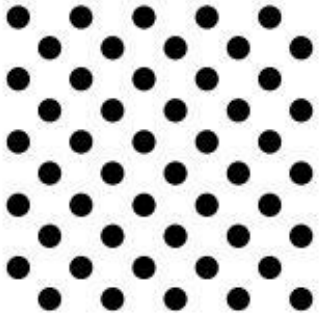
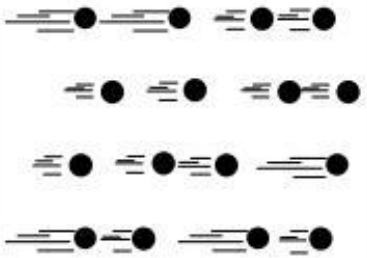
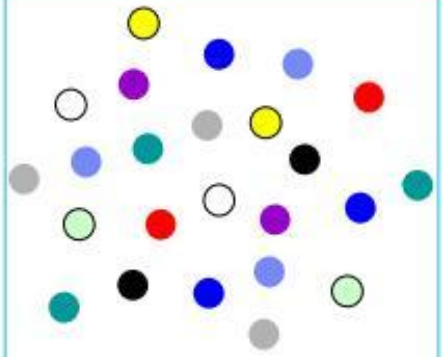
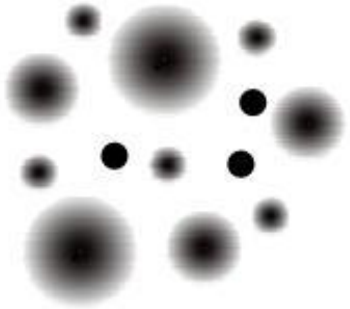


Extracción de conocimiento ➔ Todos estos tipos de datos necesitan ser analizados conjuntamente

¿Qué es Big Data?

4ª V → Veracidad

4ªV
↓ 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?

5ª V's → Valor

Aproximaciones y tecnologías innovativas

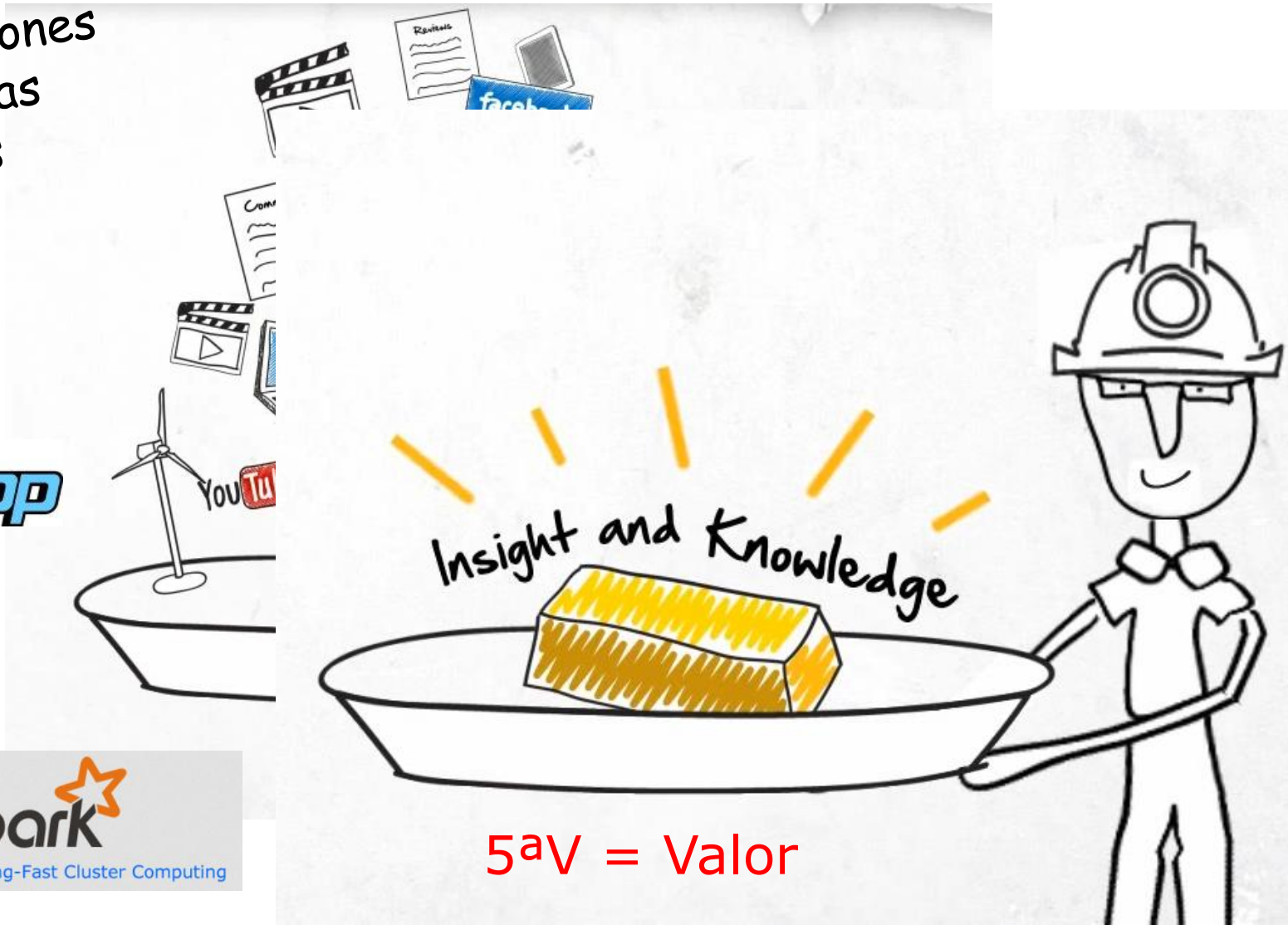
MapReduce

hadoop

mahout

APACHE GIRAPH

Spark
Lightning-Fast Cluster Computing

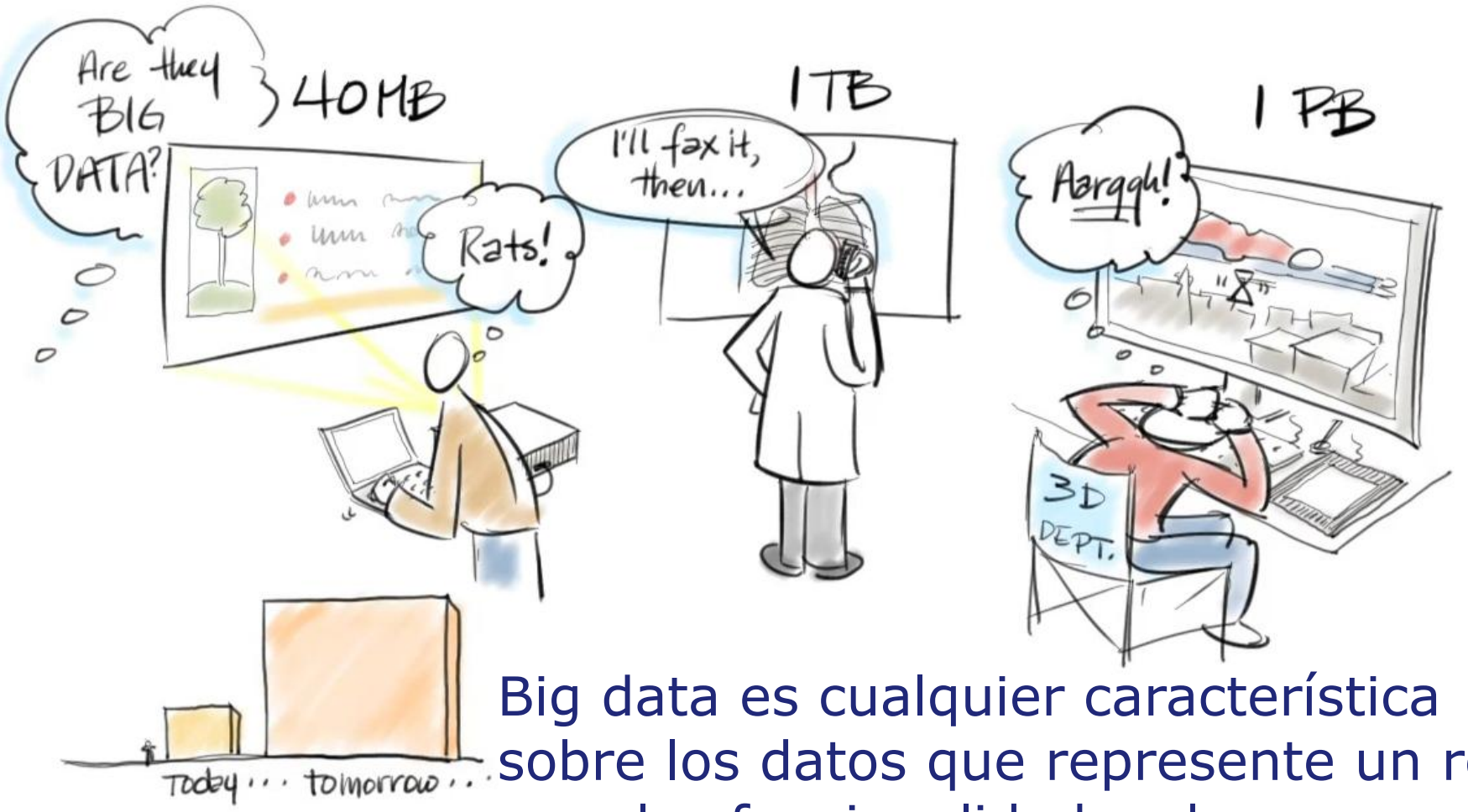


5ªV = Valor

¿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, vídeos …

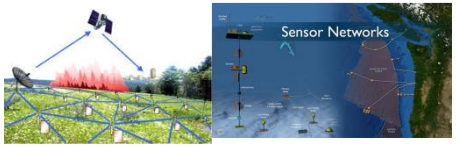
¿Qué es Big Data?

¿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)

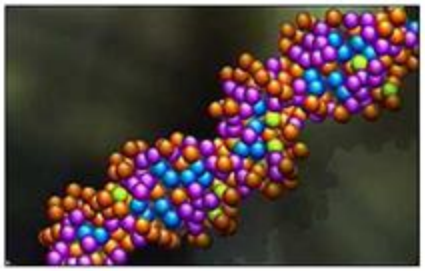
El progreso y la innovación ya no se ven obstaculizados por la capacidad de recopilar datos, sino por la capacidad de gestionar, analizar, sintetizar, visualizar, y descubrir el conocimiento de los datos recopilados de manera oportuna y en una forma escalable

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

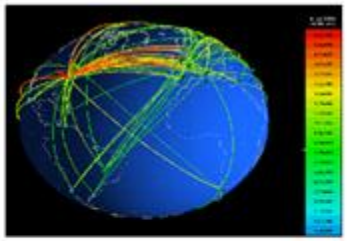
Telefonía

- 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

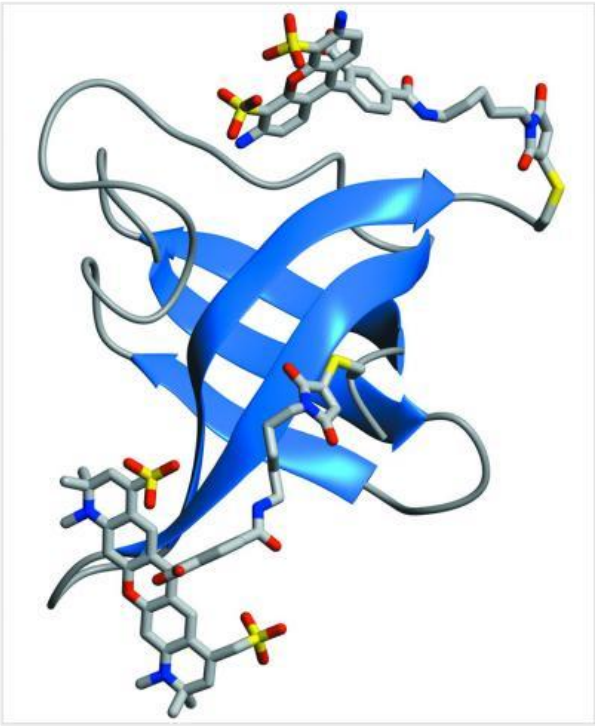
Procesamiento de información WEB

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

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

Big Data

- ¿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

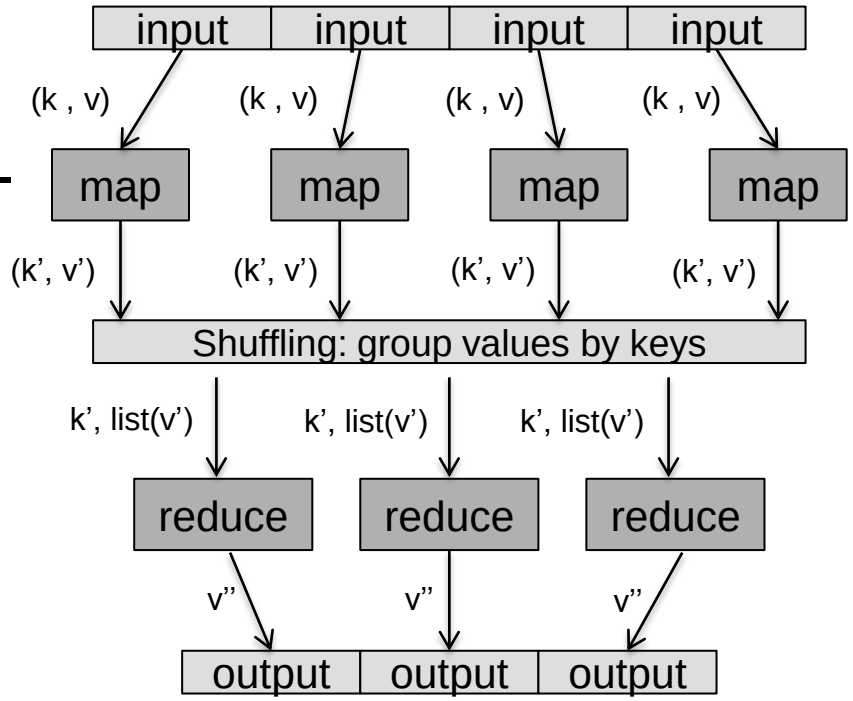
MapReduce

- 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 codigo 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)**



map (k, v) → list (k', v')
reduce (k', list(v')) → v''

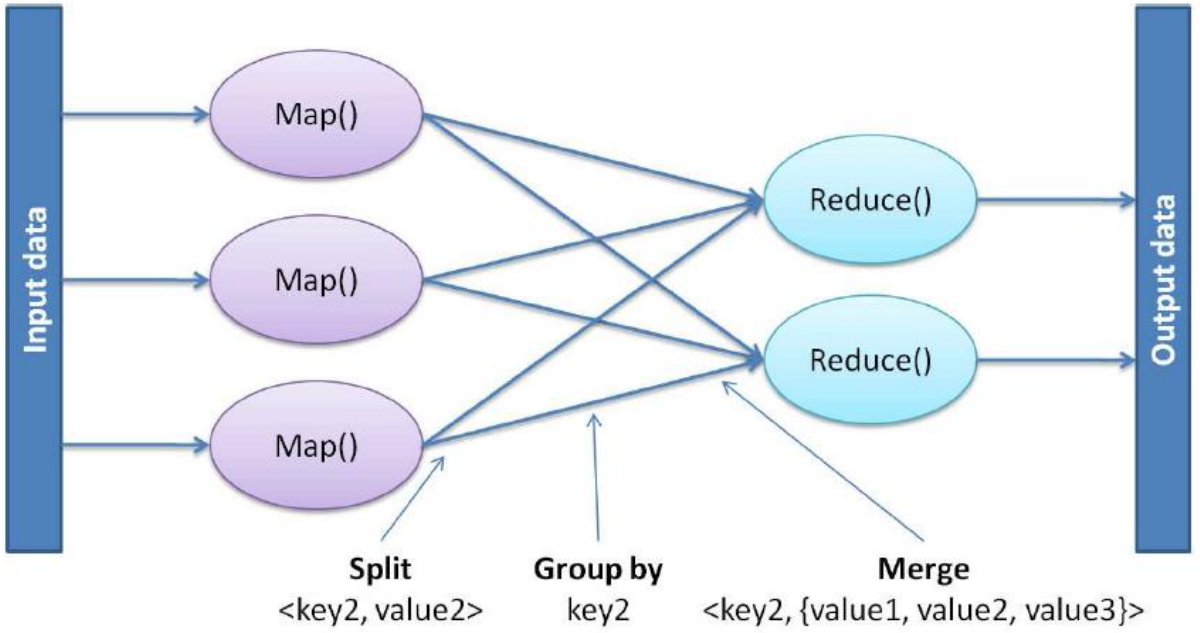
J. Dean, S. Ghemawat, **MapReduce: Simplified data processing on large clusters**, Communications of the ACM 51 (1) (2008) 107-113.

MapReduce

Flujo de datos MapReduce

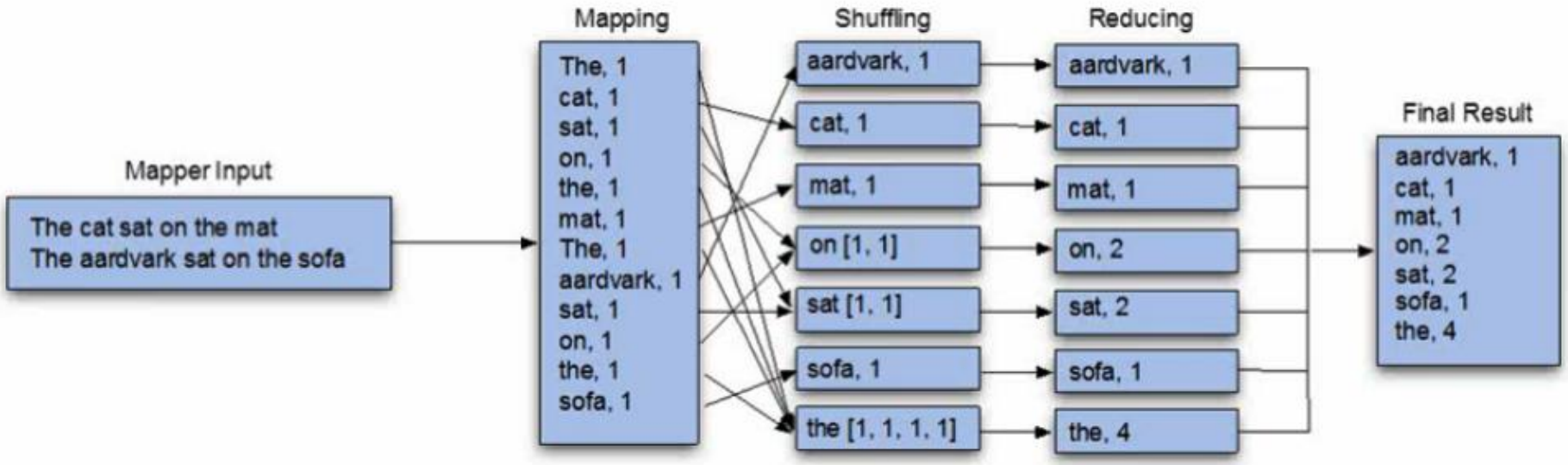
map (k,v) → list(k',v')

reduce (k',list(v')) → v''



MapReduce

Una imagen completa del proceso 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

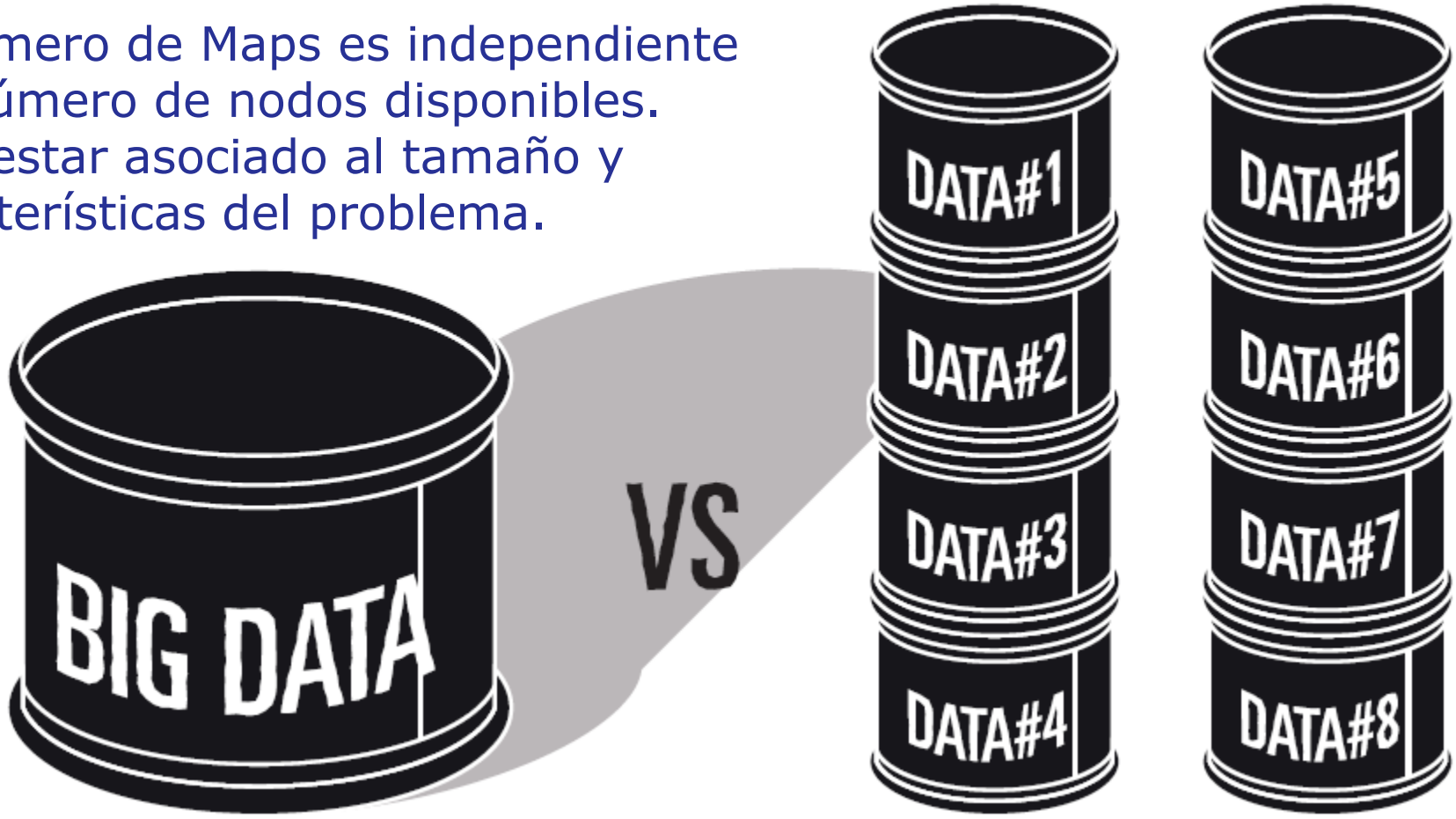
Características

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

MapReduce

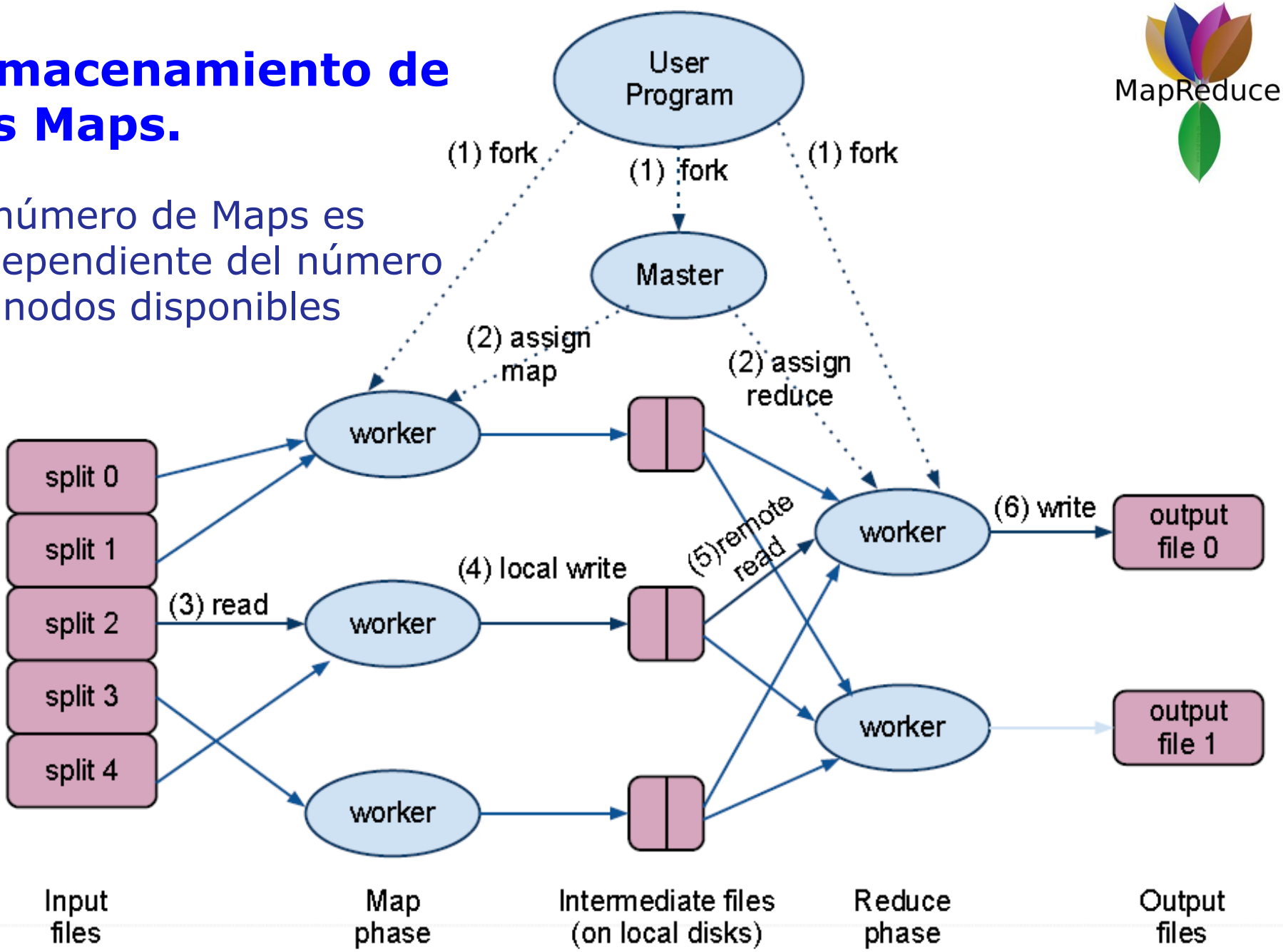
¿Número de Maps?

El número de Maps es independiente del número de nodos disponibles.
Va a estar asociado al tamaño y características del problema.



Almacenamiento de los Maps.

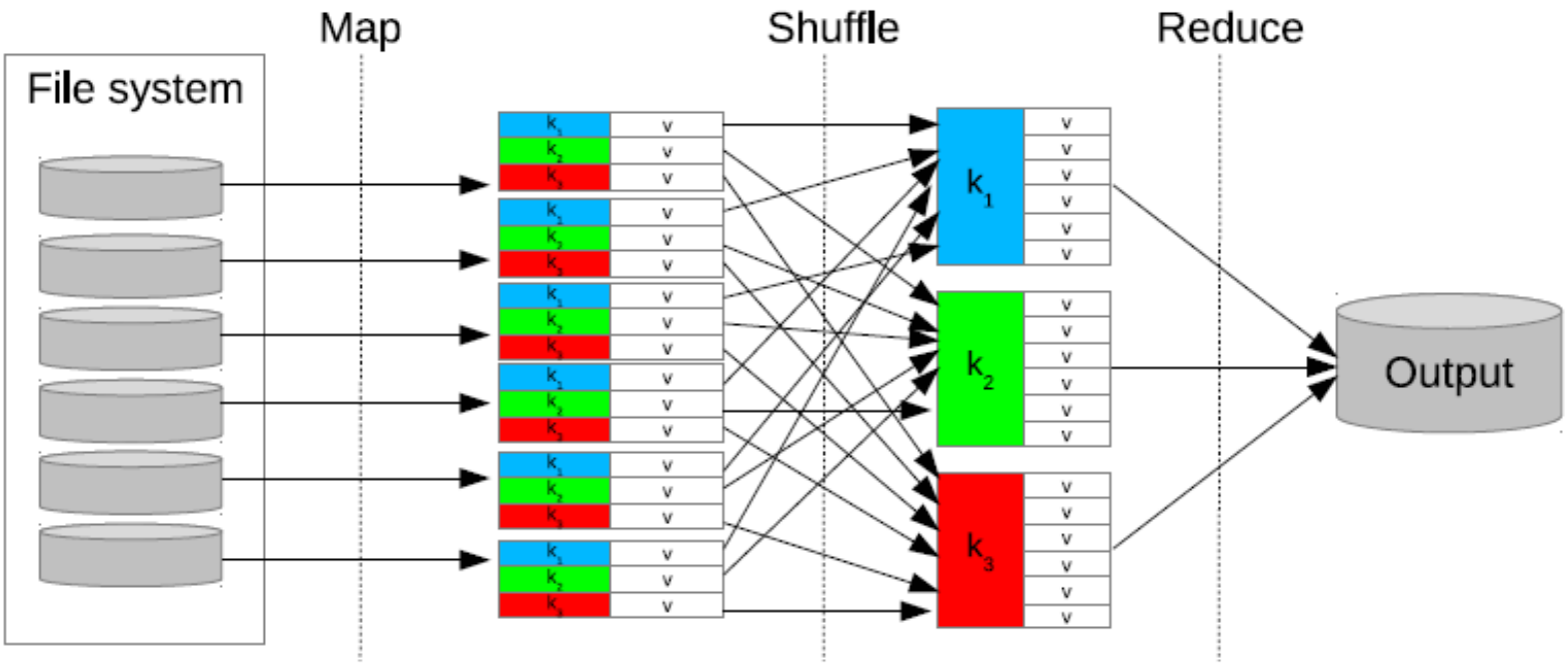
El número de Maps es independiente del número de nodos disponibles



Almacenamiento con copias, normalmente r=3

MapReduce

Flujo de datos en MapReduce, transparente para el programador



Ficheros de entrada

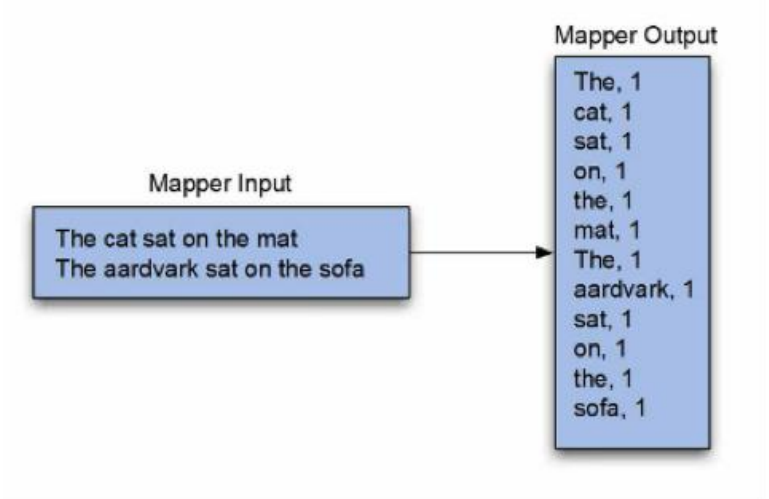
Ficheros Intermedios

Ficheros salida

MapReduce

Aspectos a analizar:

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

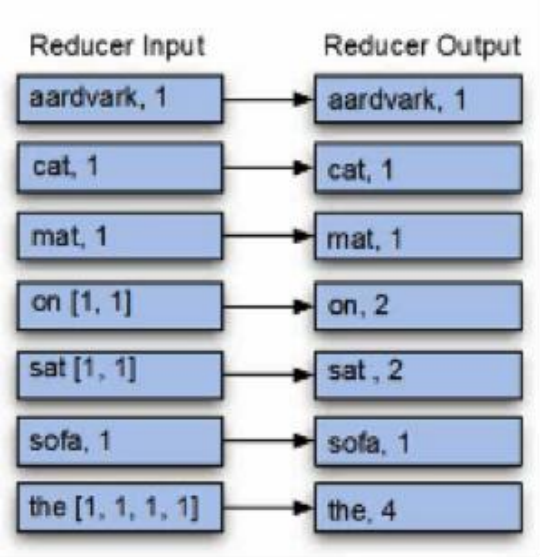


Ejemplo: Problemas de Clasificación. Nos podríamos encontrar con el problema de falta de densidad de datos y de clases con muy pocos datos (clasificación no balanceada).

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:

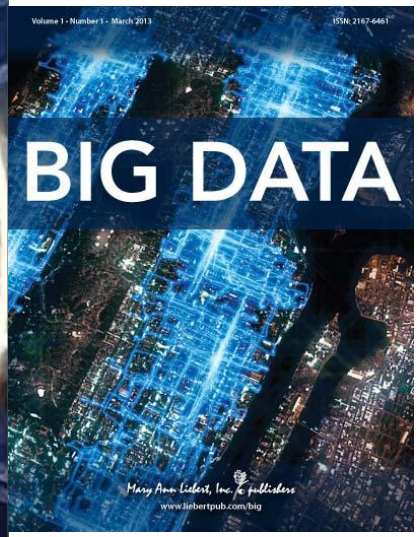
- **Ventaja frente a los modelos distribuidos clásicos:** El modelo de programación paralela de datos de MapReduce oculta la complejidad de la distribución y tolerancia a fallos
- **Claves de su filosofía**: Es
  - ***escalable:*** *se olvidan los problemas de hardware*
  - ***más barato:*** *se ahorran 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?



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

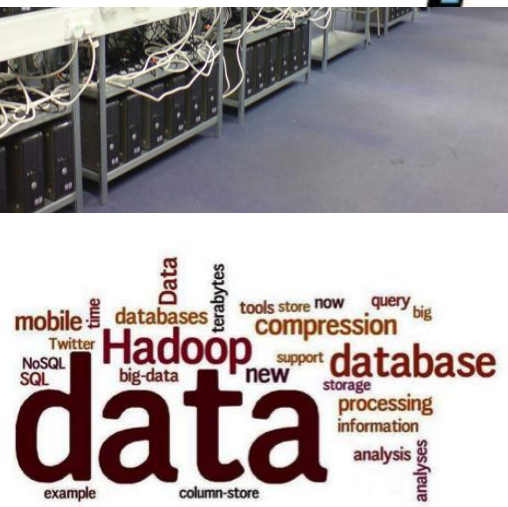
- **Iterative Graph Algorithms: PageRank, ...**
- **Gradient Descent**
- **Expectation Maximization**

Big Data

- ¿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

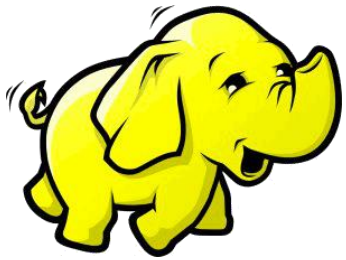
Hadoop es una implementación de código abierto del paradigma computacional MapReduce



http://hadoop.apache.org/

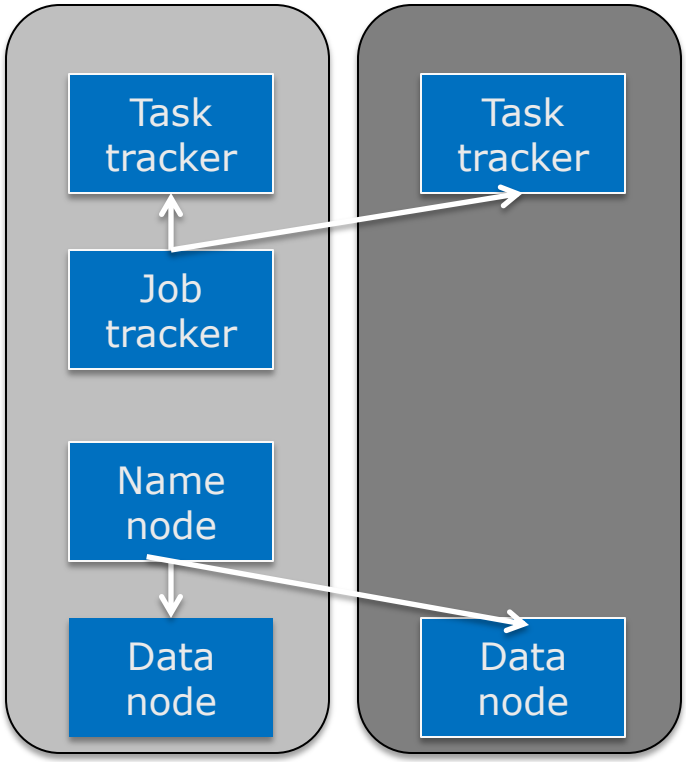
Hadoop

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



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

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*). *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 SCI²S

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?

¿Qué es Cloudera?

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

Big Data

- ¿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

Mahout

Software de Ciencia de Datos

Generation	1ª Generación	2ª 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 Descendent, k-means clustering, Random forest, etc.
Algoritmos No disponibles	Practically nothing	Vast no.: Kernel SVMs, Multivariate Logistic Regression, Conjugate Gradient Descendent, 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

bigml

Biblioteca de código abierto en APACHE

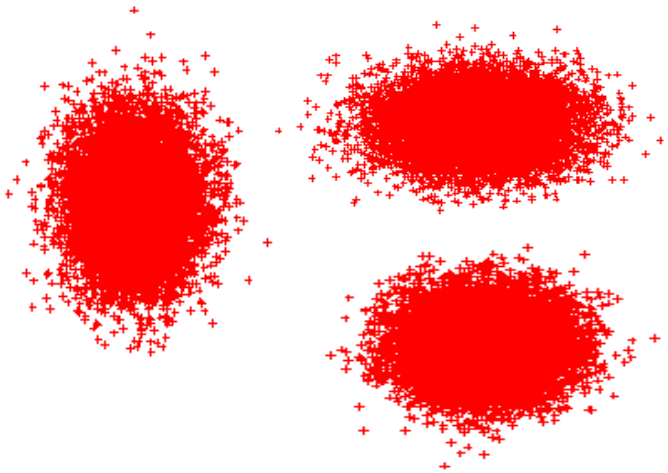
http://mahout.apache.org/

Mahout

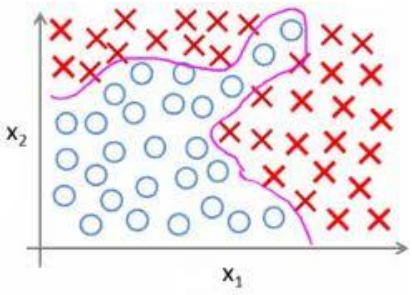
Scalable machine learning and data mining

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

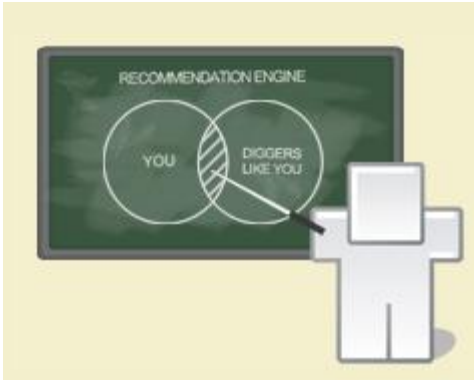
Cuatro grandes áreas de aplicación



Agrupamiento

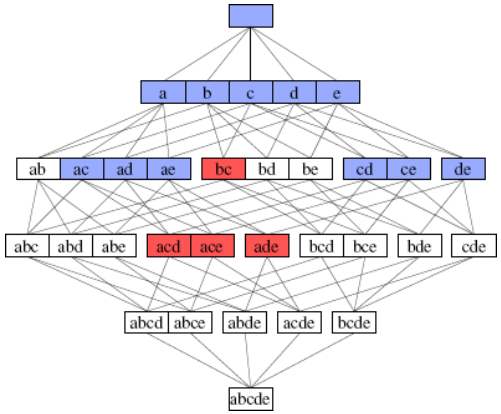


Clasificación



Sistemas de Recomendaciones

Asociación

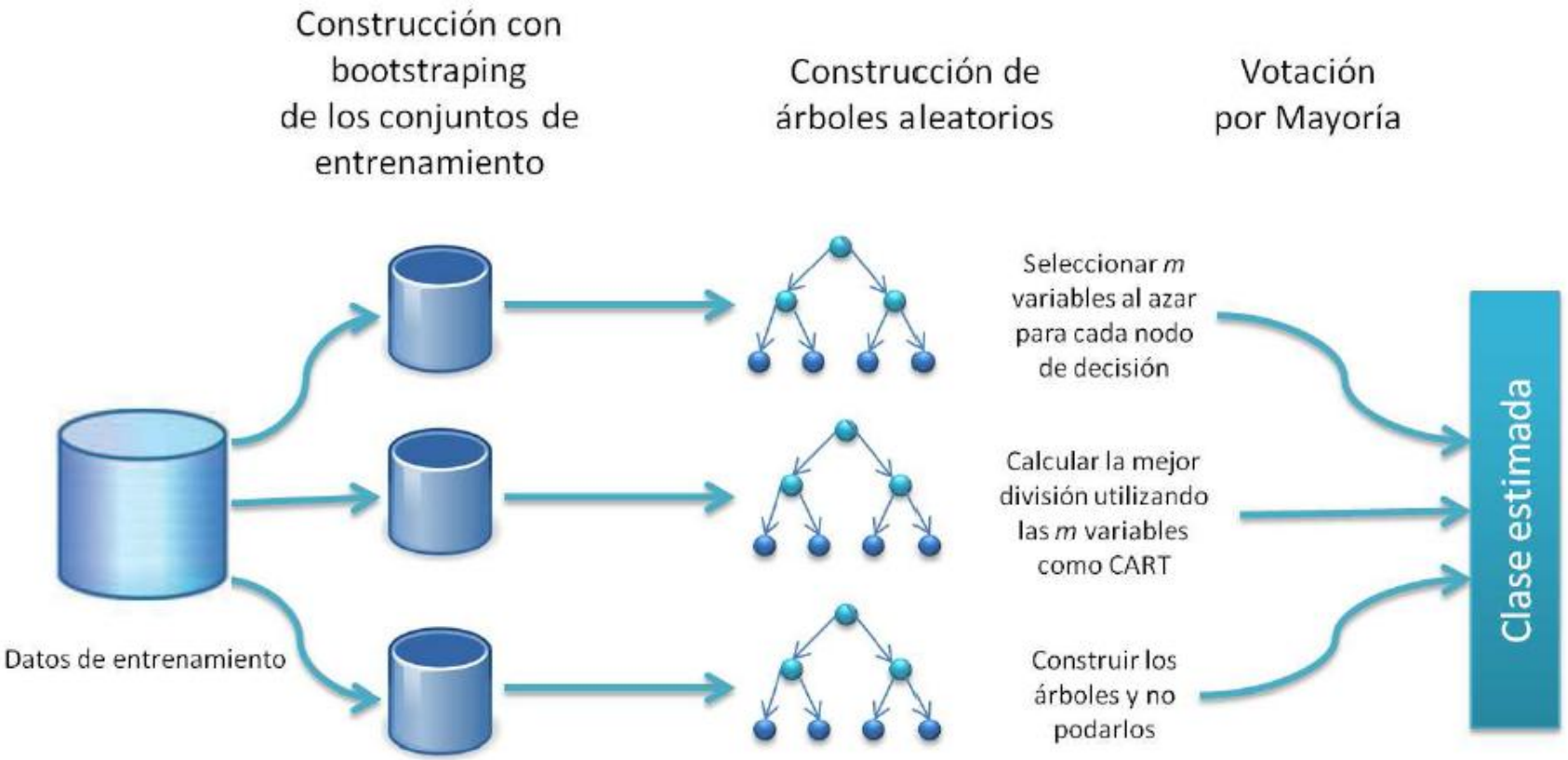


Mahout

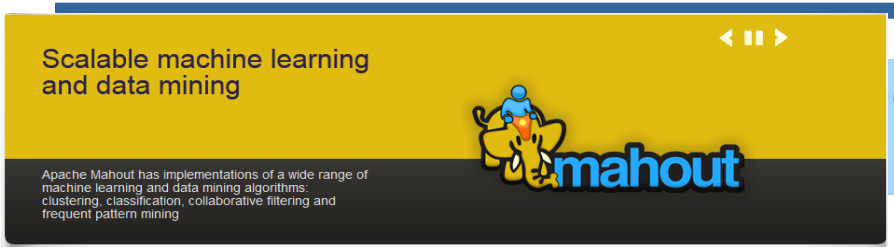
Scalable machine learning and data mining

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



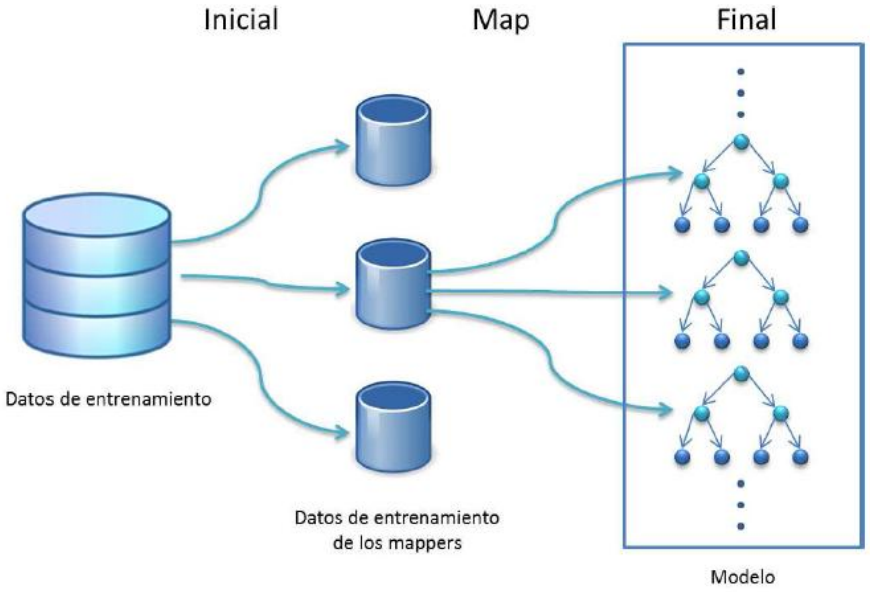
Mahout



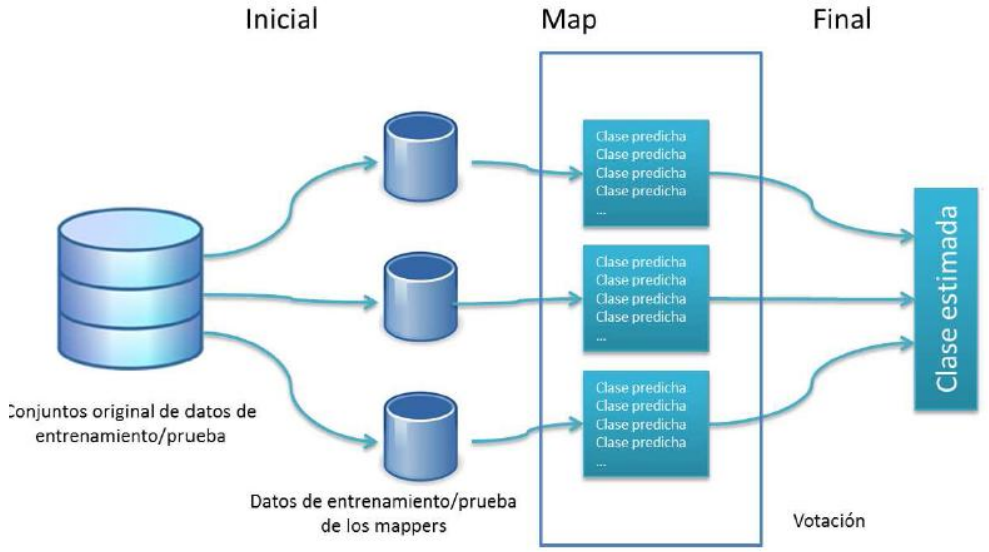
Caso de estudio:
Random Forest para KddCup99

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

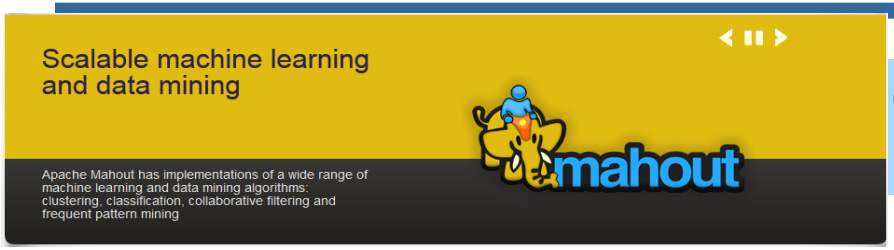
Fase de Construcción



Fase de Clasificación



Mahout



Caso de estudio: Random Forest para KddCup99

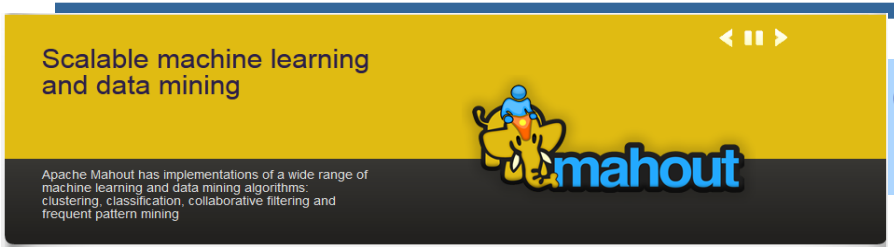
Class	Instance Number
normal	972.781
DOS	3.883.370
PRB	41.102
R2L	1.126
U2R	52

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



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.datadr.org/

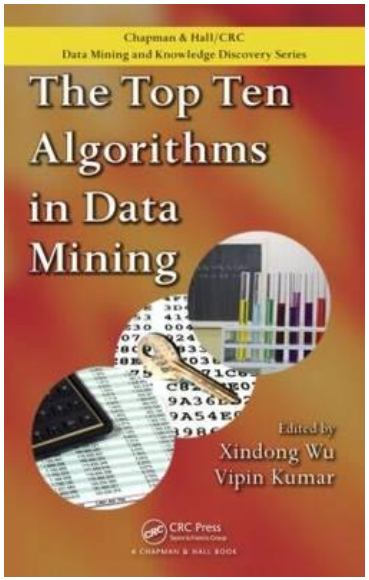
Big Data

- ¿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 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), 1904-1916.***

(Amazon EC2 cloud, CGL-MapReduce: (modelos iterativos de MapReduce)

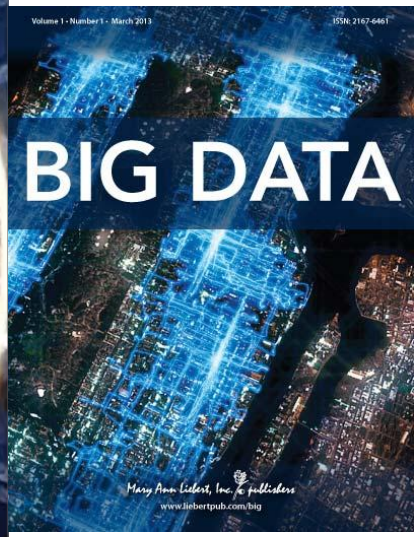
Limitaciones de MapReduce

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

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/

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

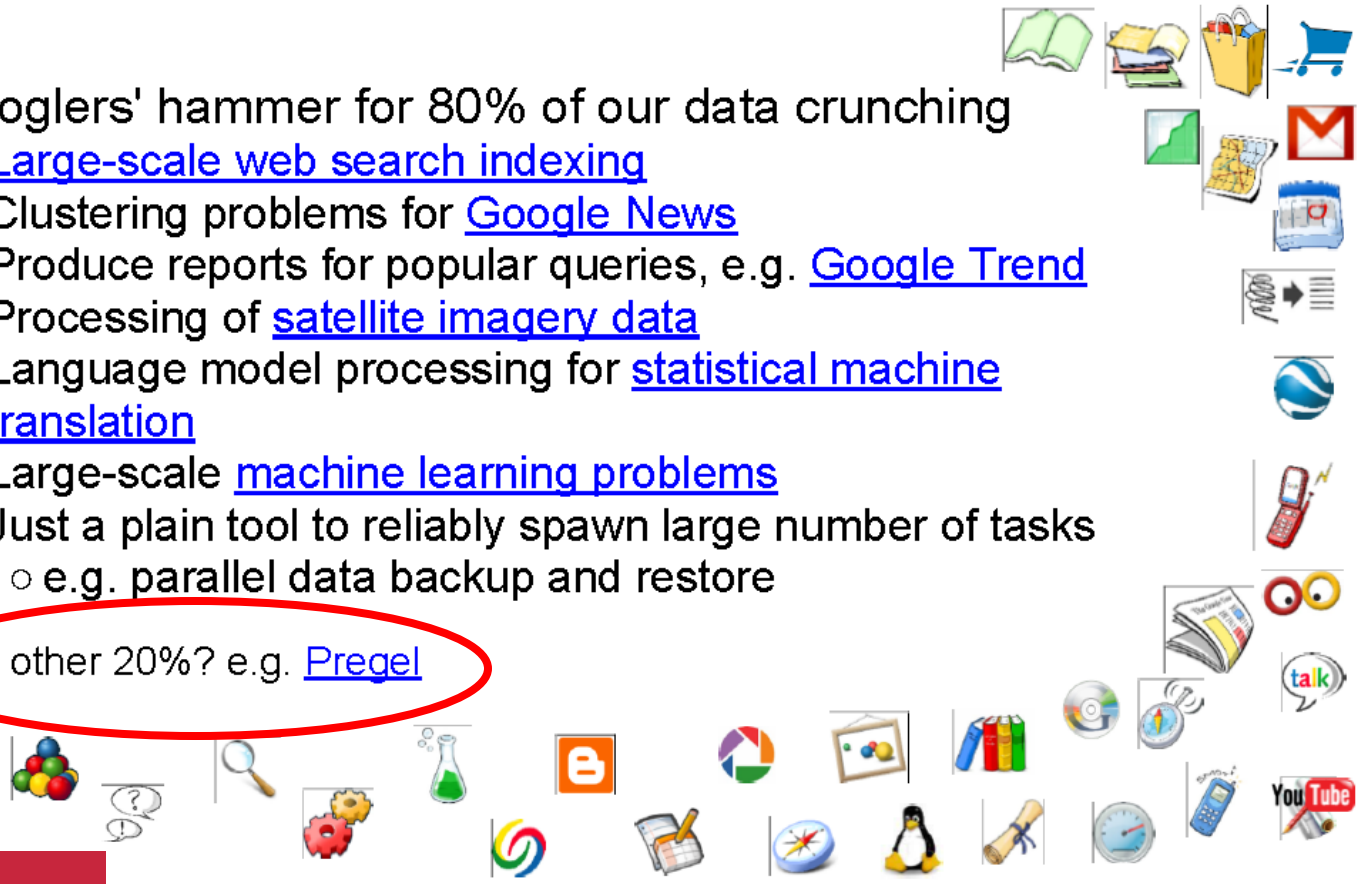
MapReduce inside Google

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

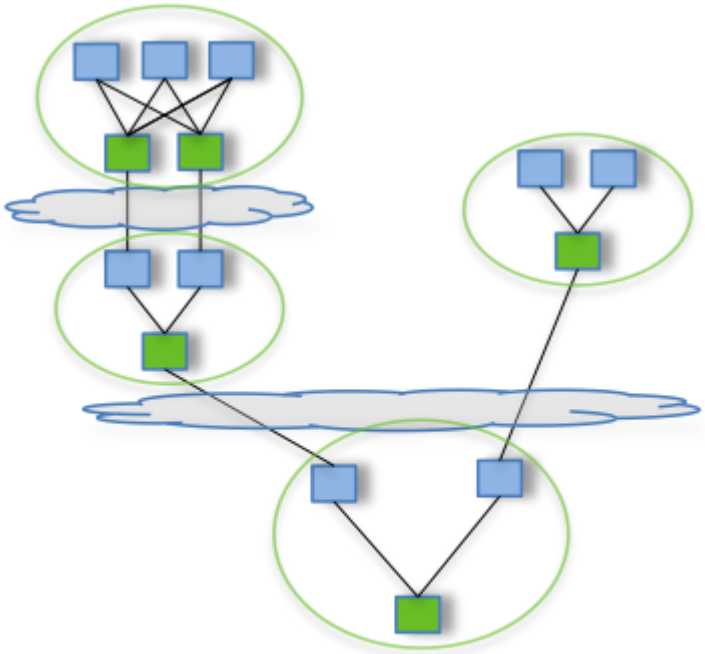
The other 20%? e.g. Pregel



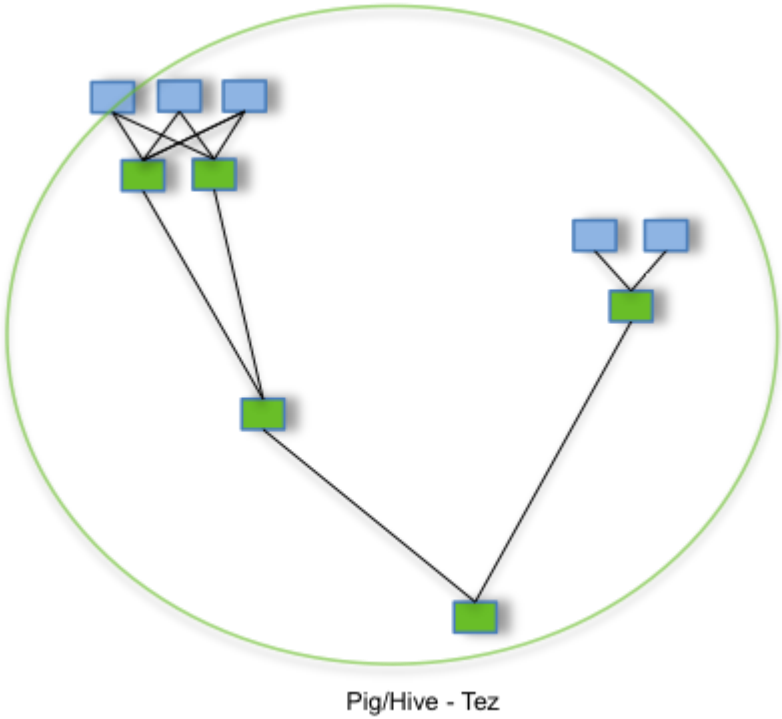
Enrique Alfonseca
Google Research Zurich

Limitaciones de MapReduce

Procesos con flujos acíclicos de procesamiento de datos



Pig/Hive - MR



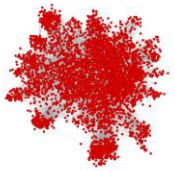
Pig/Hive - Tez

http://tez.apache.org/

Limitaciones de MapReduce

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

APACHE TEZ
(http://tez.apache.org/)
Procesos de datos con estructuras de grafos acíclicos complejas)



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
para 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.

2014 - Hadoop Spark

Big Data "in-memory"

ML Lib Spark

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

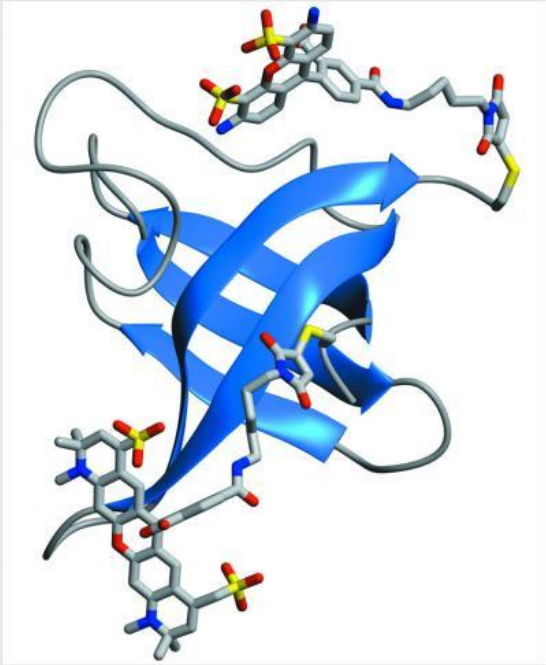
https://spark.apache.org/docs/latest/mllib-guide.html

Big Data

- ¿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

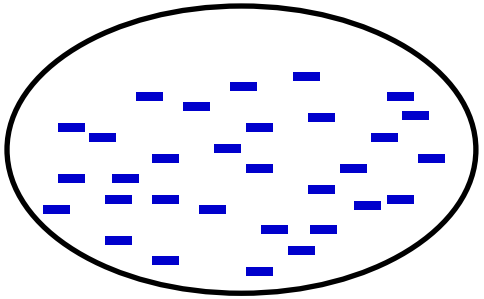
J. Bacardit et al, **Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features**, Bioinformatics 28 (19) (2012) 2441-2448

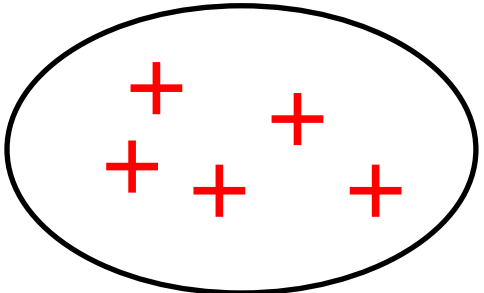
Evolutionary Computation for Big Data and Big Learning Workshop

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

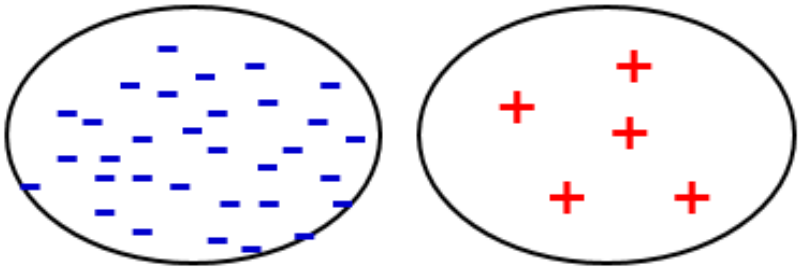
3. Modelo de aprendizaje
   - ❑ RandomForest

Clasificar el conjunto de test

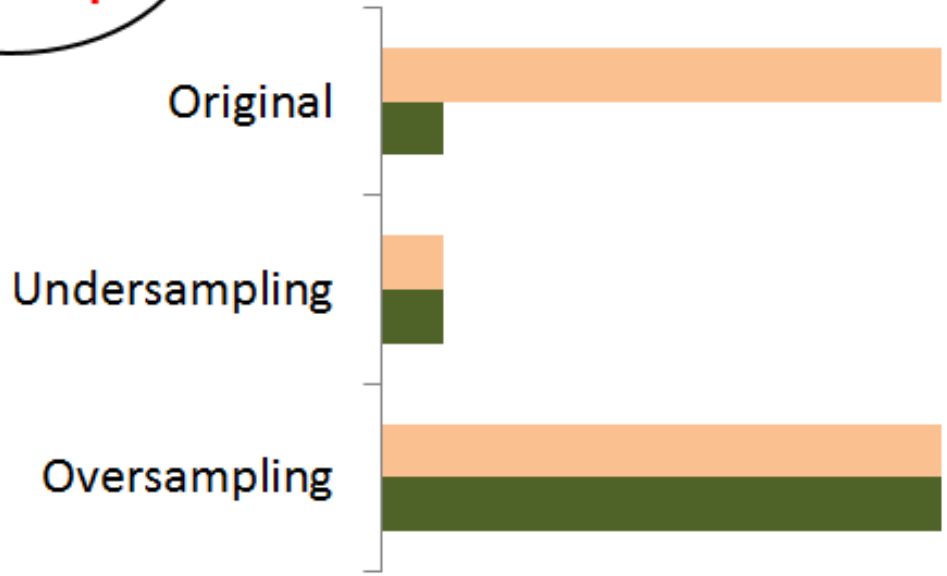
ECBDL'14 Competición Big Data

A MapReduce Approach for Random Oversampling

Low ratio of true contacts (<2%).
Imbalance rate: > 49. Unbalanced problem!



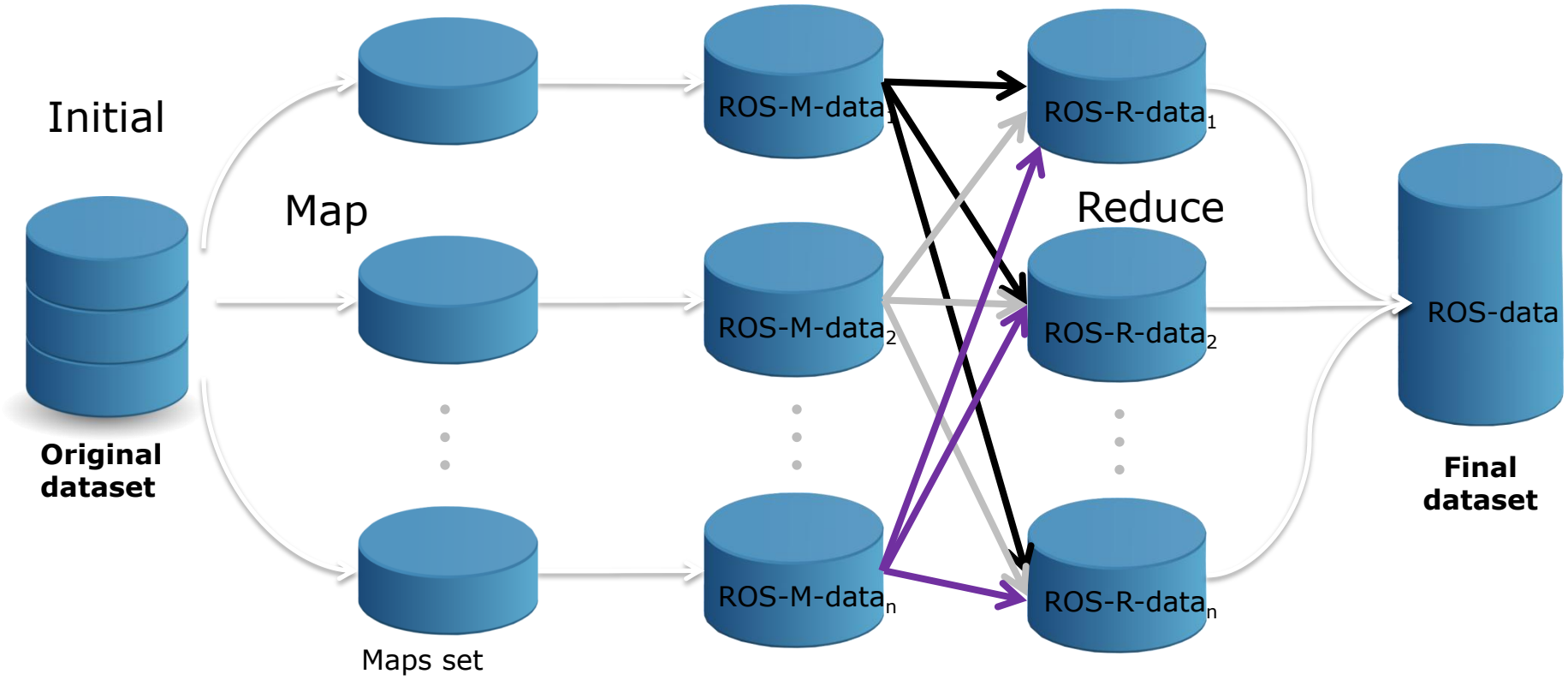
- Over-Sampling
  - Random
  - Focused
- Under-Sampling
  - Random
  - Focused
- Cost Modifying (cost-sensitive)
- Boosting/Bagging approaches (with preprocessing)



ECBDL'14 Competición Big Data

A MapReduce Approach for Random Oversampling

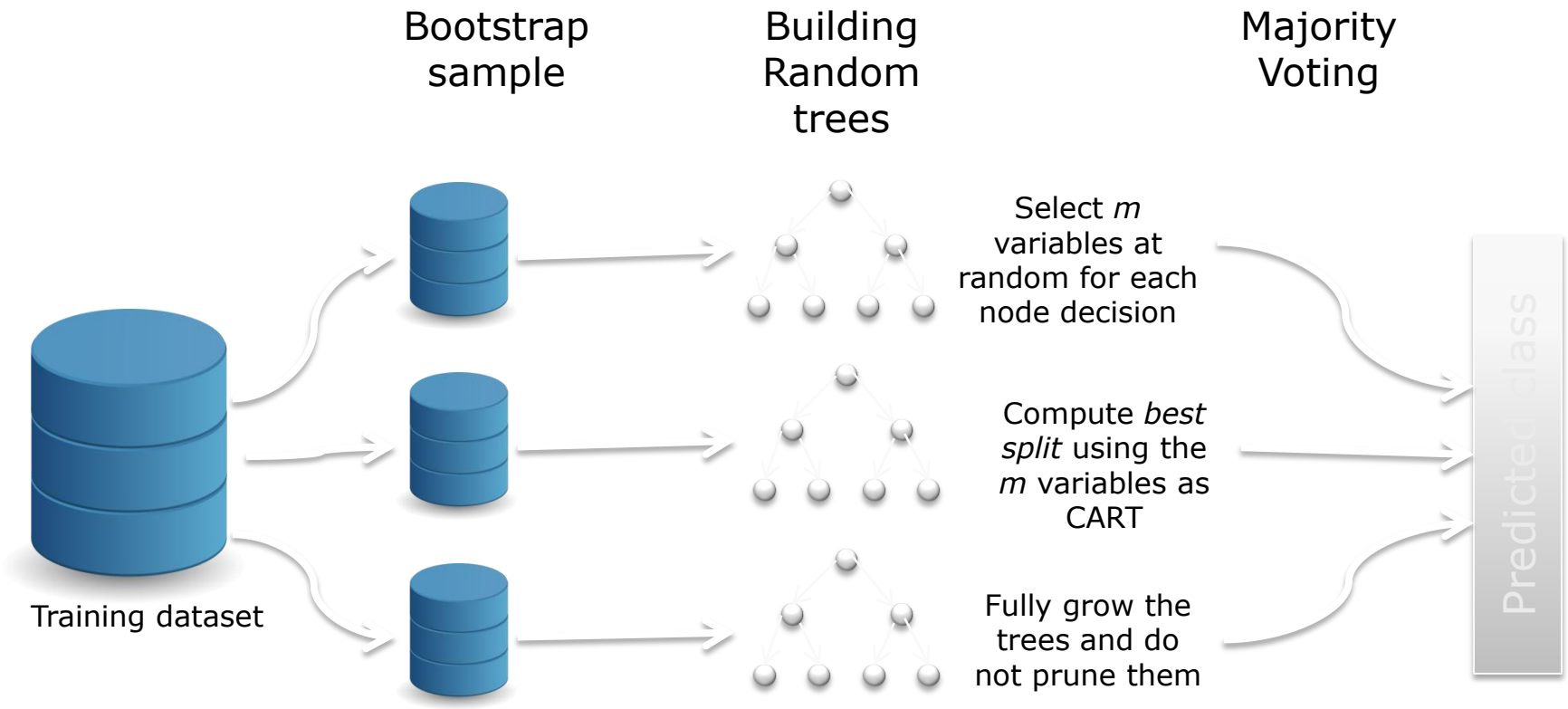
Low ratio of true contacts (<2%).
Imbalance rate: > 49. Unbalanced problem!



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
190	0,635175	0,773308	0,491186
1024	0,627896	0,756297	0,474876
2048	0,624648	0,759753	0,474578

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

190 mappers

Why we consider 190 maps?

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

Our cluster Atlas manages 192 cores (4 serves x 4 nodes x (2 x 6 cores) = 192 cores

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

ECBDL'14 Competición Big Data

We analyze again the number of mappers with ROS (130%)

- Oversampling rate: {130} Number of mappers: 128 to 2048

RandomForest:

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

ROS+RF (130%) and mappers

Num Mappers	TPR	TNR	TNR*TPR Test
128	**0.7450**	**0.6678**	**0.497624**
190	0.7472	0.6609	0.493913
256	0.7518	0.6531	0.491064
1024	0.7644	0.6121	0.467969
2048	0.7755	0.5860	0.454528

The equilibrium with the number of mappers is very important

The less number of maps, the less TPR and the high 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 Weigthing:

- 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

We investigate: The use of Evolutionary Feature Weighting. It allows us to construct several subset of features (changing the threshold).

	128 mappers			
Algorithms	**TNR*TPR Training**	**TPR**	**TNR**	**TNR*TPR Test**
ROS+RF (130% - Feature Weighting 19)	0.621638	0.684726	0.735272	0.503459
ROS+RF (115% - Feature Weighting 19)	0.628225	0.674569	0.750184	0.506051
ROS+RF (100% - Feature Weighting 19)	0.635029	0.629397	0.784132	0.493531
ROS+RF (130% - Feature Weighting 63)	**0.634843**	**0.683800**	**0.756926**	**0.517586**
ROS+RF (115% - Feature Weighting 63)	**0.639319**	**0.677015**	**0.764589**	**0.517638**
ROS+RF (100% - Feature Weighting 63)	0.648723	0.638567	0.794595	0.507402

	64 mappers			
Algorithms	**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

On the Random Forest: We decided to investigate the influence of the Random Forest's parameters (internal features and number of trees) as well as a higher number of features (90).

	190 mappers			
Algorithms	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
ROS+ RF (140%+ FW 90+30f+200t)	0.630104	0.721138	0.730323	0.526664

Correct decisions with FW 90 and RF with 25/30f and 200 trees.
Good trade off between TPR and TNR

ECBDL'14 Competición Big Data

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

	64 mappers			
Algorithms	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 increase the TPR mantaining a good TNR rate.

An equilibrium and good results

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**
HyperEns	0.640027	0.763378	0.761308	0.488583
PUC-Rio_ICA	0.657092	0.714599	0.713634	0.469558

EFDAMIS 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**

	64 mappers			
Algorithms	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

Al comienzo

Nº mappers	TPR_tst	TNR_tst	TNR*TPR Test
64	0,601723	0,806269	0,485151
190	0,635175	0,773308	0,491186
1024	0,627896	0,756297	0,474876
2048	0,624648	0,759753	0,474578

Al final (30 de Junio)

	64 mappers			
Algorithms	TNR*TPR Training	TPR	TNR	TNR*TPR Test
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

ECBDL'14 Competición Big Data

ECBDL'14: Evolutionary Computation for Big Data and Big Learning Workshop
July 13th, 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 ECBDL'14 big data competition

Jaume Bacardit, organizer
ECBDL'14 big data competition

Big Data

- ¿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

BIG DATA

- La paralelización 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.

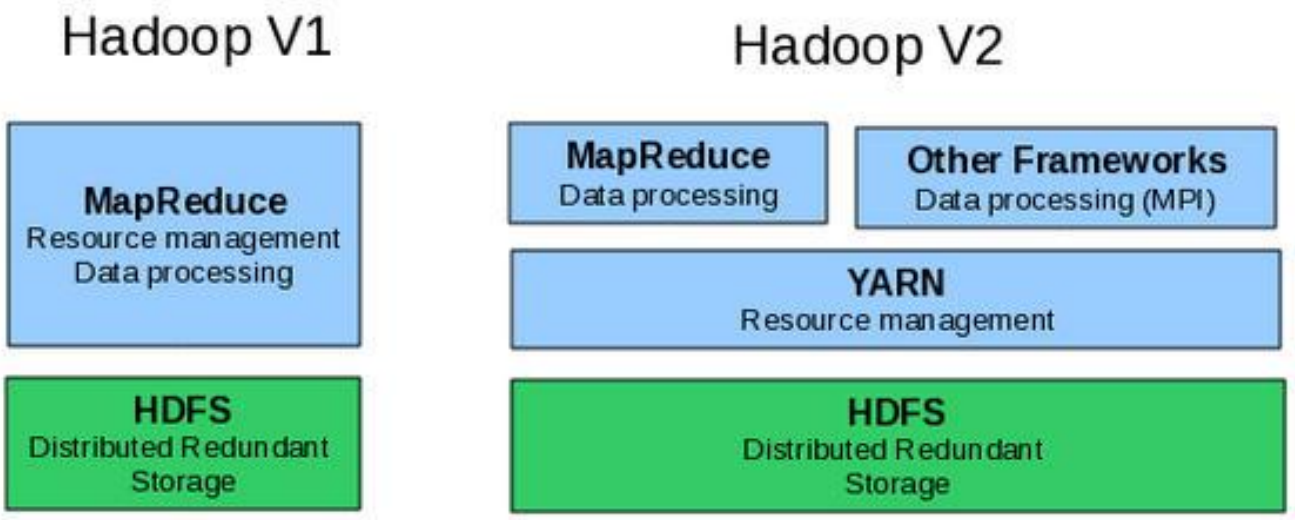
Bibliografía: A. Fernandez, S. Río, V. López, A. Bawakid, M.J. del Jesus, J.M. Benítez, F. Herrera, **Big Data with Cloud Computing: An Insight on the Computing Environment, MapReduce and Programming Frameworks**. *WIREs Data Mining and Knowledge Discovery, in press (2014).*

Generation	1st Generation	2nd Generation	3nd Generation
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 Descendent, 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 Descendent, 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

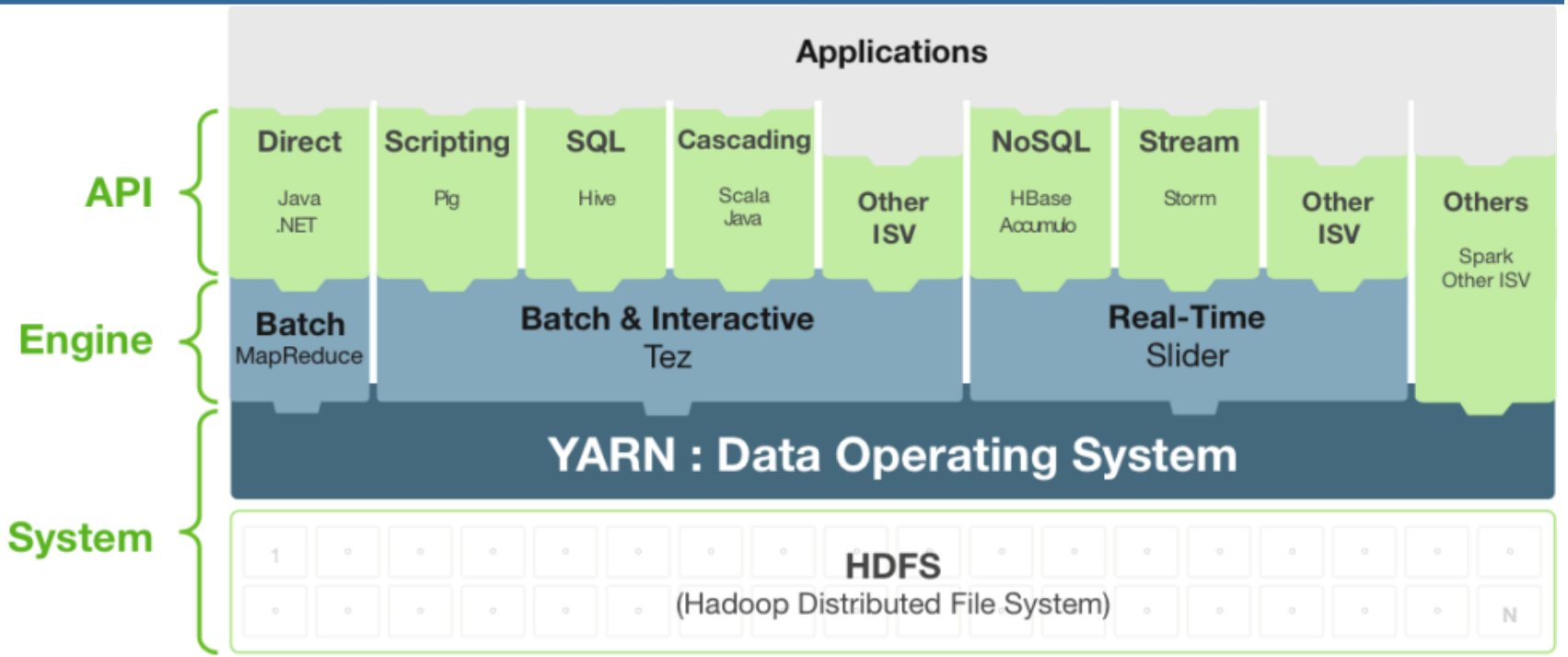
Comentarios Finales

BIG DATA

Evolución de Hadoop



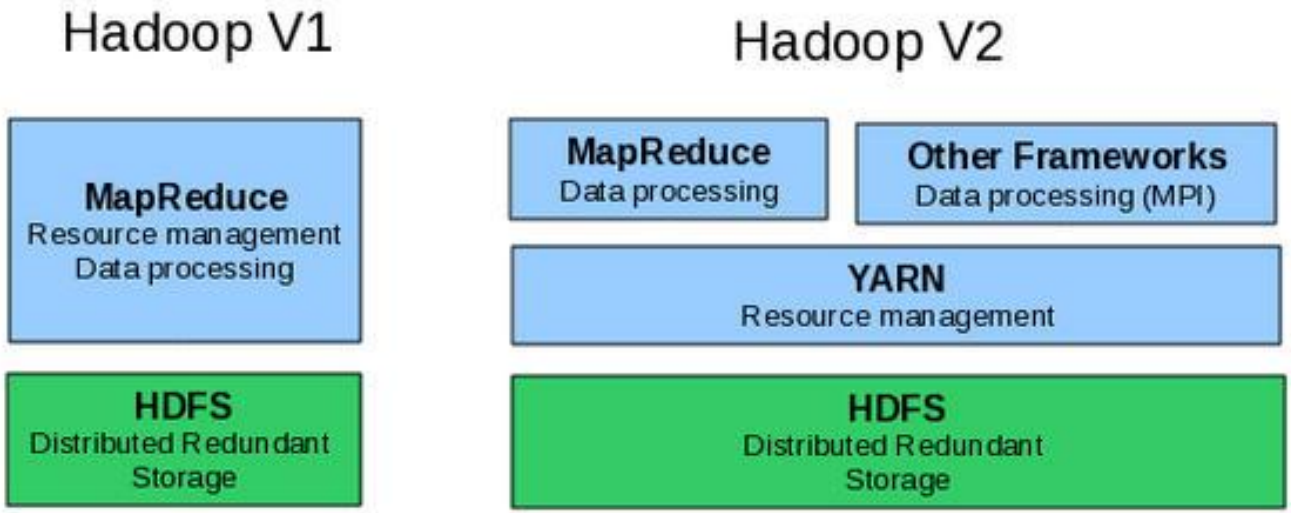
Comentarios Finales



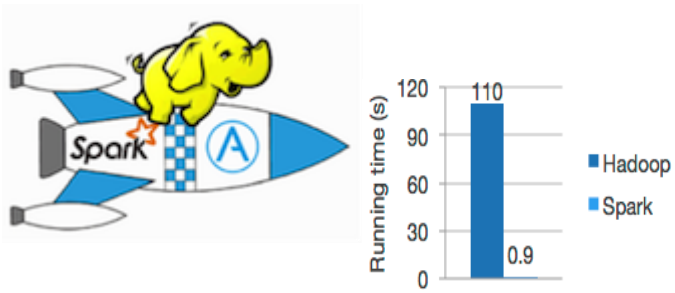
Apache Hadoop YARN es el sistema operativo de datos de Hadoop 2, responsable de la gestión del acceso a los recursos críticos de Hadoop. YARN permite al usuario interactuar con todos los datos de múltiples maneras al mismo tiempo, haciendo de Hadoop una verdadera plataforma de datos multi-uso y lo que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

Comentarios Finales

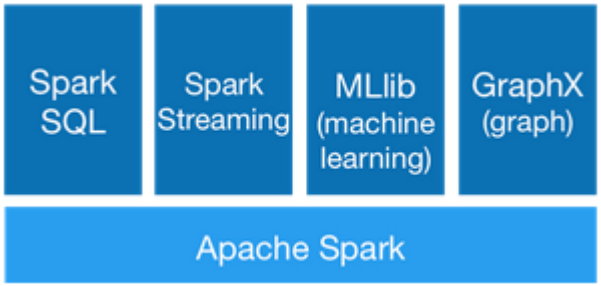
BIG DATA



Enfoque InMemory
HDFS Hadoop + SPARK



Ecosistema
Apache Spark



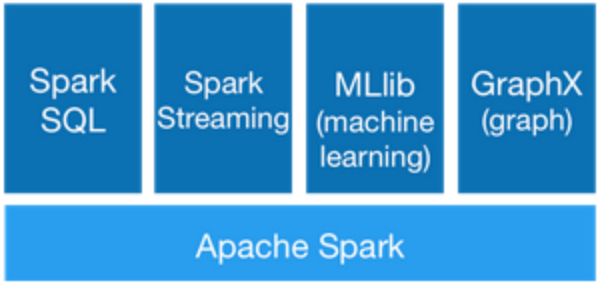
Futura versión de
Mahout con Spark

Comentarios Finales

BIG DATA

Spark
Lightning-Fast Cluster Computing

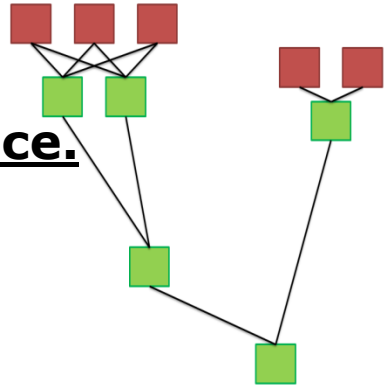
• Ecosistema Apache Spark



• **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**

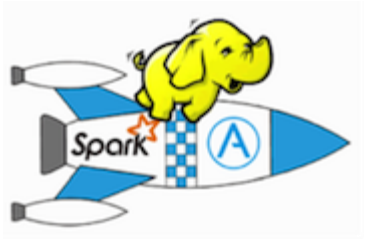
• **Esquema de computación más flexible que MapReduce.**
Permite la flujos acíclicos de procesamiento de datos

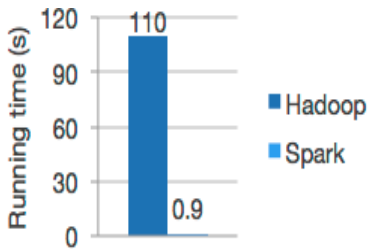


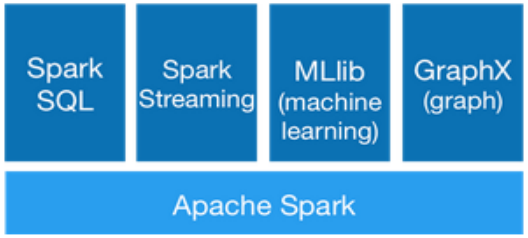
Comentarios Finales

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

BIG DATA

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

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

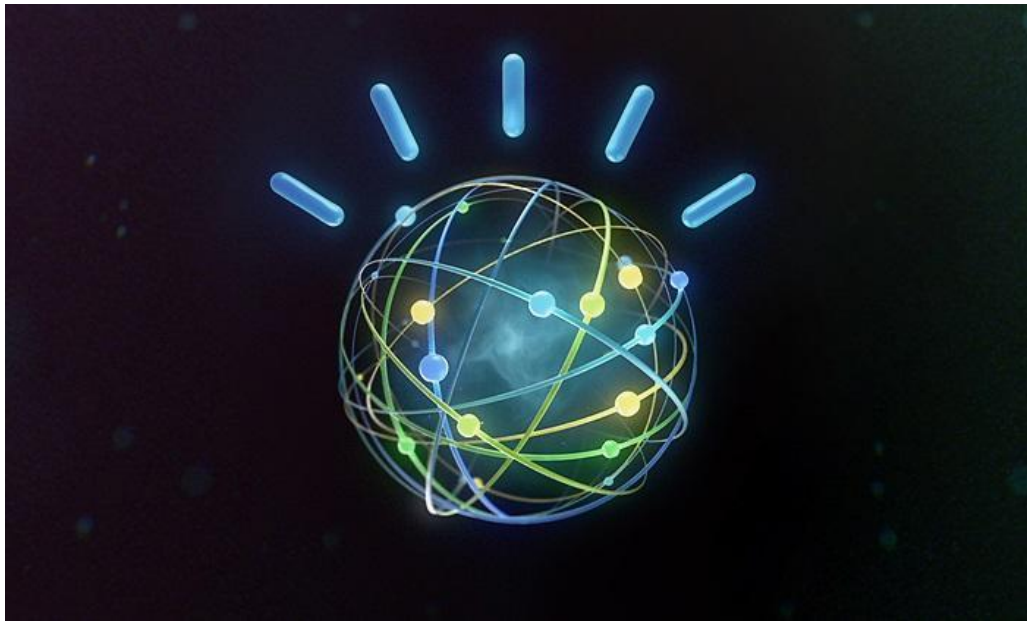
In many new applications - face recognition, speech understanding, recommendations, or fraud detection - bigger data does produces better results

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. 2. Watson the Journalist.
3. 3. Watson the Nurse.

Comentarios Finales

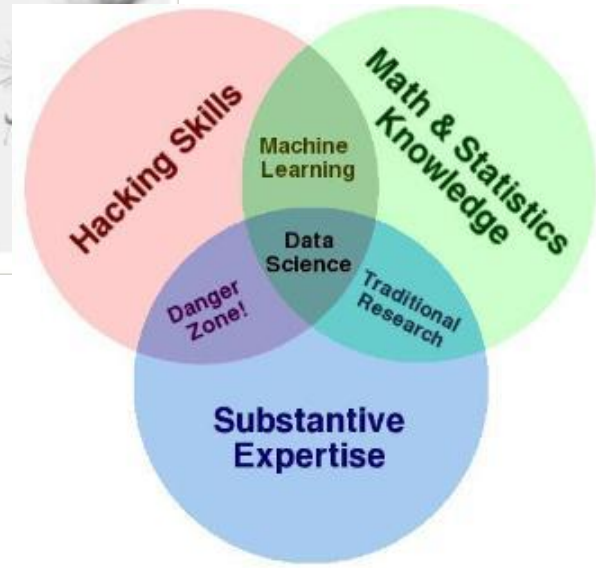
Oportunidades en Big Data

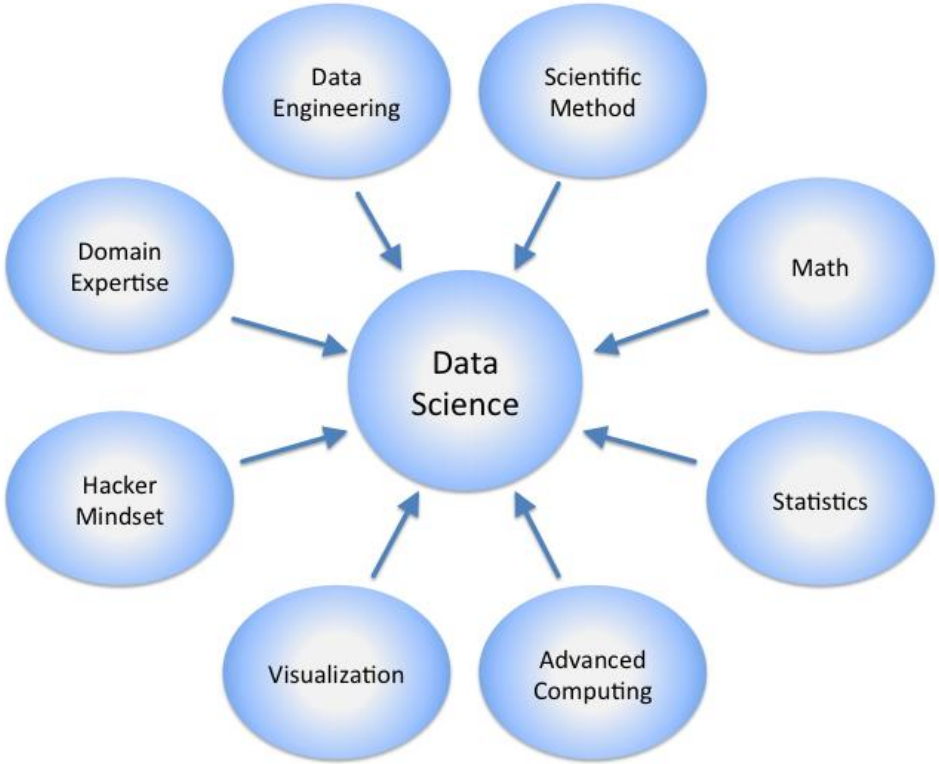
Big Data es un área emergente y en expansión.
Las posibilidades de desarrollo de algoritmos para nuevos datos, aplicaciones reales … es un nicho de investigación y desarrollo en los próximos años.

Comentarios Finales

Data Science

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





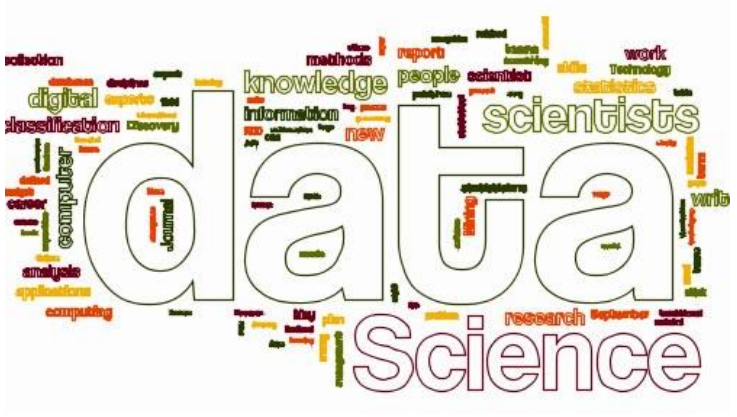
Comentarios Finales

Surge como profesión el "Científico de Datos"

Científico de Datos

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



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 PAÍS

PORTADA | INTERNACIONAL | POL

ECONOMÍA

ECONOMÍA EMPRESAS MERCADOS BOLSA FINANZAS PERSONALES VIVIENDA TECNOLOGÍ

ESTÁ PASANDO | Multa a la banca | Revuelo en Hacienda | Eléctricas y renovables | Paro

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

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#!

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 Toledo.

"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

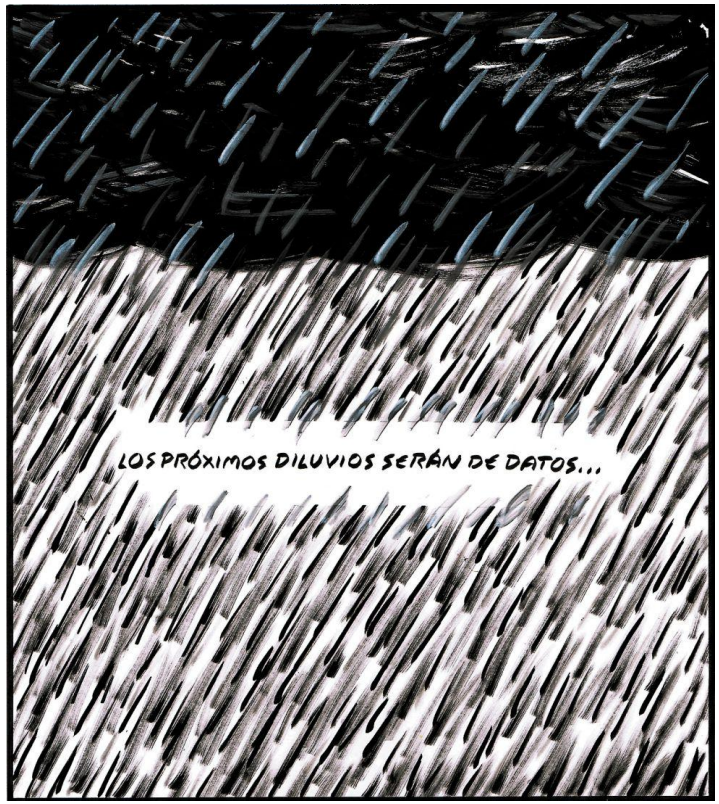
BIG DATA

http://elpais.com/elpais/2013/12/02/vinetas/1386011115_645213.html

El Roto

Viñeta de El Roto

3 de diciembre de 2013



Gracias!!!

BIG DATA