

Aspectos Estadísticos y Computacionales del análisis de Datos Masivos(Big Data)

Dr. Edgar Acuña

Departamento de Ciencias Matemáticas y Programa
Doctoral en Ciencias e Ingeniería de la Información y
Computación

Universidad de Puerto Rico-Mayaguez

E-mail: edgar.acuna@upr.edu , eacunaf@gmail.com

Website: academic.uprm.edu/eacuna

Contenido

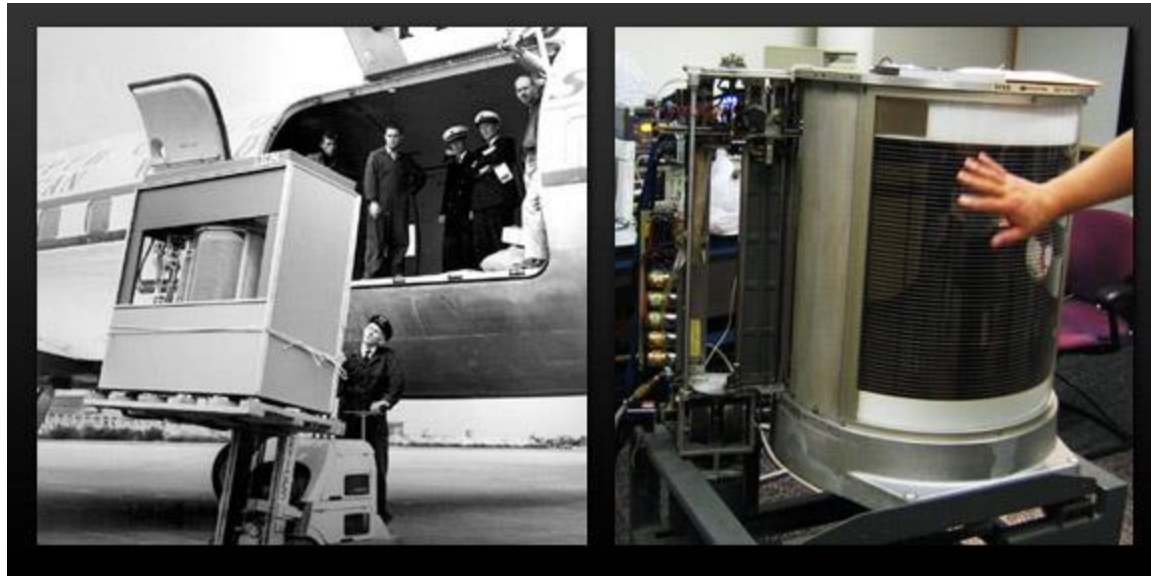
- I. Minería de Datos
- II. Datos Masivos (Big Data)
- III. Hadoop y MapReduce
- IV. Ejemplo: K-means
- V. Ciencias de datos

Mineria de Datos: Motivacion

Los mecanismos para la recolección automática de datos (por ejemplo, sensores) y el desarrollo de tecnología para bases de datos ha hecho posible que una gran cantidad de datos pueda estar disponibles en bases de datos, almacenes de datos y otros repositorios de información.

Hoy día, hay la necesidad de convertir estos datos en conocimiento e información.

El primer disco duro, 1956



IBM 350, tenía el tamaño de dos refrigeradoras y una capacidad de alrededor de 5MB. Costaba aprox 50,000 dolares. Mi laptop tiene 100,000 veces mas de esa capacidad

Mineria de Datos[3]: Tamaño (en Bytes) de bases de datos.

Descripcion	Tamaño	Medio de Almacenaje
Muy pequeño	10^2	Pedazo de Papel
Pequeno	10^4	Varios hojas de papel
Mediano	10^6 (megabyte)	Disco Floppy
Grande	10^9 (gigabite)	USB/Disco Duro
Masivo	10^{12} (Terabyte)	Disco Duro/USB
Super-masivo	10^{15} (Petabyte)	Archivo de datos distribuidos
Exabyte(10^{18}), Zettabytes(10^{21}), Yottabytes(10^{24})		

Ejemplos de grandes bases de datos

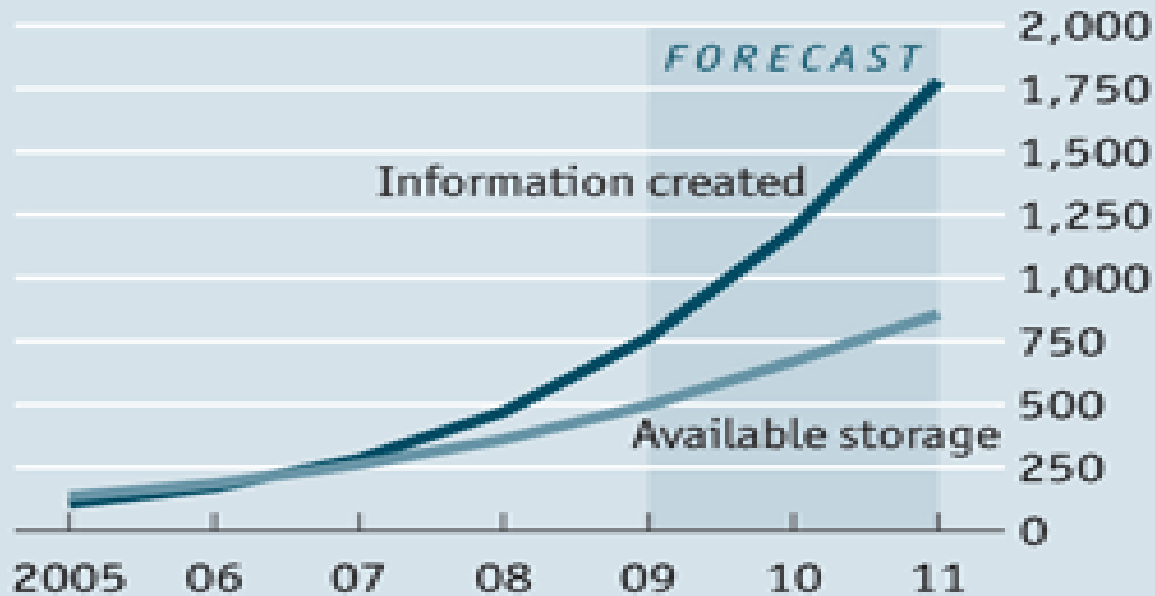
- El telescopio Large Hadron Collider (LCH) almacena al año cerca de 25 Petabytes de datos de sensores.
- En 2010, la base de datos de llamadas de ATT era de 323 Terabytes.
- El 2008, los almacenes Walmart almacenaba 10 Terabytes de transacciones por día.
- El 2011, Google buscaba en más de 20 billones de páginas, que representa más de 400 TB.
- Se estima que en el 2009 había 500 exabytes (1,000 petabytes) de información en la internet.
- El 2013, se anunció que el centro de datos de la NSA sería capaz de almacenar 5 zettabytes (1,000 exabytes).

The economist, February 2010

Overload

1

Global information created and available storage
Exabytes



Source: IDC

Mineria de datos: Que es Minería de Datos?

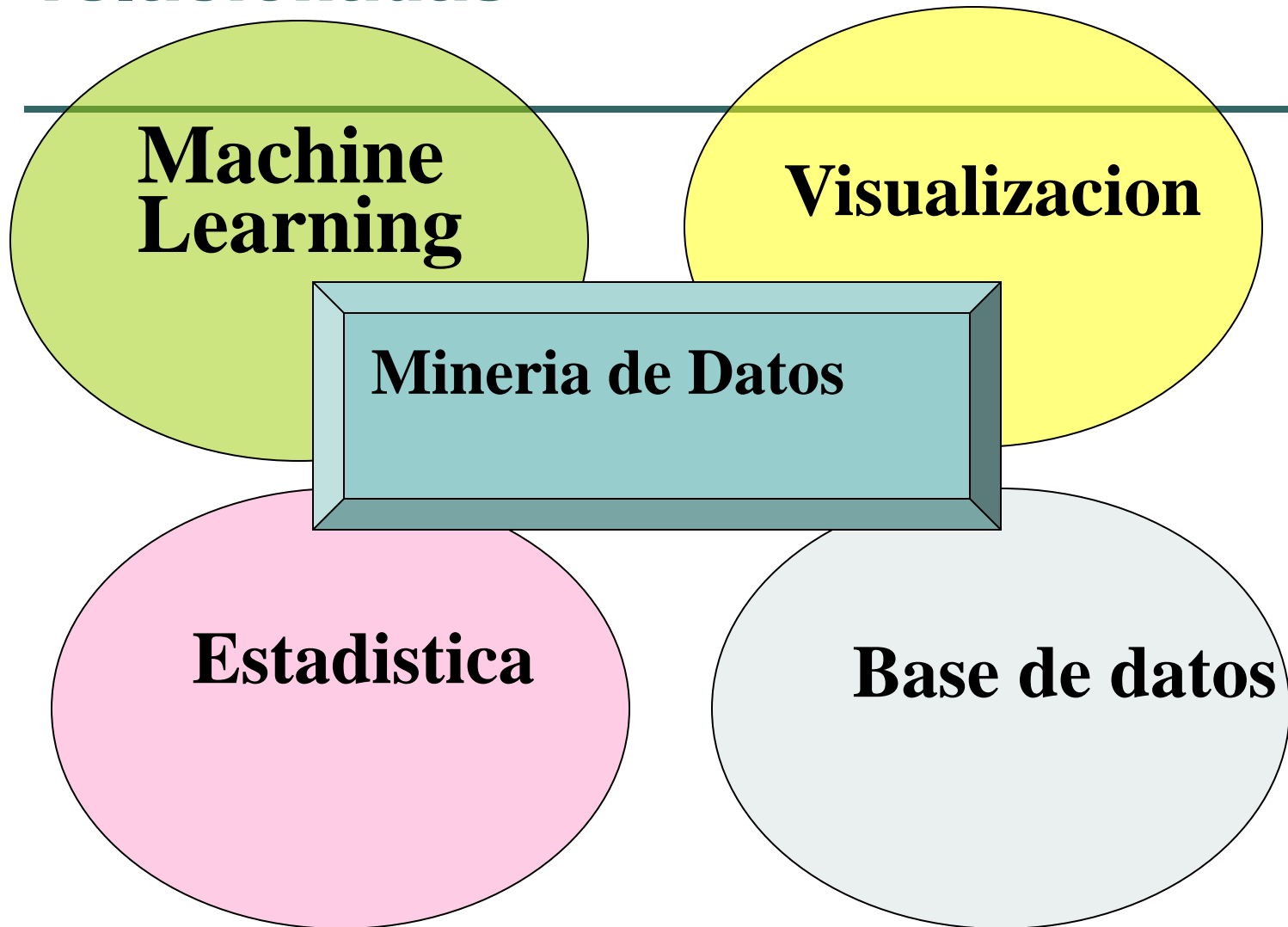
- Es el proceso de extraer conocimiento/información válida de una base de datos bien grande. El conocimiento es dado como características (“patrones”) y reglas que son no-triviales, previamente desconocidas, comprensibles y con alto potencial de ser útiles.
- Otro nombres: Descubrimiento de Conocimiento en bases de datos (KDD), Analisis Inteligente de datos, Ingenieria de datos, Inteligencia de negocios.
- El primer artículo en Minería de datos apareció en

1993

Data mining es ...

- Determinar las características de personas a las que se puede hacer un préstamo bancario.
- Detectar intrusos (casos anómalos) en un sistema.
- Determinar las características de los clientes de un banco que pueden cometer fraude.
- Determinar las características de los clientes que abandonan la suscripción a un servicio.
- Hallar ciudades de un país que tienen características similares (regionalización).

Mineria de datos[5]. Areas relacionadas



Estadística, Machine Learning

- Estadística (~35% de DM)
 - Se basa más en teoría. Asume propiedades distribucionales de las variables que están siendo consideradas.
 - Se enfoca más en probar hipótesis y en estimación de parámetros.
 - Se consideran eficientes estrategias de recolectar datos.
 - Estimación de modelos.
- “Machine learning” (Aprendizaje Automático) (~30 % de DM)
 - Parte de Inteligencia Artificial. Machine es equivalente a un modelo en estadística.
 - Más heurística que Estadística.
 - Incluye a: Redes Neuronales, árboles de decisión, algoritmos genéticos.

Base de datos, visualizacion, etc

- Base de datos relacionales (~25% de DM)
 - Conjunto de tablas conteniendo datos de una categoria predeterminada. Cada una de las tablas (llamada relacion) contiene un o mas columnas de datos las cuales representan ciertos atributos.
 - El interface entre el usuario y la base de datos relacional mas usado es SQL(structured query language).
- Visualizacion (~5 % de DM)
 - Se explora la estructura del conjunto de datos en forma visual.
 - Puede ser usado en la etapa de pre o post procesamiento del KDD.

Otras Areas: (~5%): Pattern Recognition, High Performance Computing, Expert Systems.

Aplicaciones de DM

Administración de negocios: Investigación de mercados, relación de los clientes con la gerencia, detección de Fraudes, Telecomunicaciones, etc.

Gobierno: detección de evasores de impuestos, terrorismo.

Ciencias: Astronomía, Bioinformática (Genomics, Proteonomics, Metabolomics), descubrimiento de medicinas.

Text Mining: Extraer información previamente desconocida de diversas fuentes escritas (e-mails)

Web mining: Recomendar lista de productos que un usuario puede comprar. (Amazon.com)

Tipos de tareas en data mining

- Descriptivas: Se encuentran las propiedades generales de la base de datos. Se descubren las características más importantes de la base de datos.
- Predictivas: Se entrena (estima) un modelo usando los datos recolectados para hacer predicciones futuras. Nunca es 100% precisa y lo que más importa es el rendimiento del modelo cuando es aplicado a nuevos datos.

Tareas en data mining

- Regresion (Predictiva)
- Classificacion (Predictiva)
- Classificacion No supervisada – Clustering (descriptiva)
- Reglas de Asociacion (descriptiva)
- Deteccion de Outliers (descriptiva)
- Visualizacion (descriptiva)

Clasificación Supervisada[1]

- Dado un conjunto de registros (records), llamado el conjunto de entrenamiento. Cada registro contiene un conjunto de atributos y usualmente el ultimo atributo identifica la clase del registro.
- *Objetivo: Encontrar un modelo para el atributo clase en funcion de los valores de los otros atributos. Asignar records que no se habian visto previamente (muestra de prueba) a una clase usando dicho modelo.*
- Usualmente el conjunto dado es dividido en muestra de entrenamiento (70%) y muestra de prueba (30%). La primera es usada para construir el modelo y la segunda es usada para validarlo. La precision del modelo es determinada en la muestra de prueba.
- Es un proceso de decision. La regla de decision es llamado el clasificador.

Ejemplos de clasificadores

Analisis Discriminante Lineal (LDA),

Naïve Bayes,

Regresion Logistica

k-nearest neighbor,

Decision trees,

Bayesian Networks

Neural Networks

Support Vector machine (SVM)

.....

Clasificación No-supervisada (Clustering)[1]

- Determinar grupos de objetos (clusters) de tal manera que los objetos dentro del mismo cluster sean bastante similar entre si, mientras que objetos en grupos distintos no sean tan similares entre si.
- Se necesita usar una medida de dissimilaridad (o de similaridad) para establecer si dos objetos pertenecen a un mismo cluster o a clusters distintos.
- Ejemplos de medidas de dissimilaridad: Distancia Euclideana, distancia Manhattan, distancia Hamming, etc.
- Ejemplos de medidas de similaridad: correlacion, medida de Jaccard-Tanimoto, etc.
- Problemas: Eleccion de la medida de similaridad, eleccion del numero de clusters, validacion de clusters.

Algoritmos de Clustering[2]

- Algoritmos de Particionamiento: K-means, PAM, SOM.
- Algoritmos Jerarquicos: Aglomerativo, Divisivo.
- Modelos de Mezcla Gausiana
- DBSCAN
- BIRCH

Reglas de asociacion[1]

- Dado un conjunto de registros cada uno de los cuales contiene algun numero de items de una coleccion dada. El objetivo es encontrar reglas de dependencia que permitan predecir la ocurrencia de un item basado en ocurrencia de otros items

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Reglas descubiertas:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Reglas de Asociacion[2]

- Las reglas ($X \rightarrow Y$) deben satisfacer un soporte minimo y una confianza impuesta por el usuario. X es llamado el antecedente Y es llamado el consecuente.
- $\text{Soporte} = (\# \text{ registros conteniendo } X \text{ y } Y) / (\# \text{ registros})$
- $\text{Confianza} = (\# \text{ registros conteniendo } X \text{ y } Y) / (\# \text{ de registros conteniendo } X)$

Ejemplo: El soporte de la Regla 1 es .6 y de la regla 2 es .4

La confianza de la Regla 1 es .75 y de la regla 2 es .67

Aplicacion: Mercadeo y Promocion de ventas

Deteccion de “outliers”[1]

- Los objetos que se comportan diferente o que son inconsistentes con la mayor parte de los datos son llamados “outliers”.
- Outliers pueden ser causados por un error de medicion o de ejecucion. Ellos pueden representar algun tipo de actividad fraudulenta.
- El objetivo de la deteccion de “outliers” es detectar las instancias que tienen un comportamiento fuera de lo comun.

Deteccion de “outliers”[2]

- Metodos:
 - Metodos basados en Estadisticos
 - Metodos basados en distancia
 - Metodos basados en densidad local.
- Aplicacion: Deteccion de fraude en tarjeta de creditos, Network intrusion

Retos de Data Mining

- Escalabilidad
- Dimensionalidad
- Datos complejos y Heterogeneos.
- Calidad de datos
- Preservacion de privacidad
- Analisis en tiempo real (streaming data)

Data Mining Software

- **Open source**
- R (cran.r-project.org). Relacionado a Estadística (38.5% de usuarios, Kdnuggets.com, agosto 2014).
- Weka (<http://www.cs.waikato.ac.nz/ml/weka/>) (17.0%): Relacionado a Machine Learning. En Java.
- RapidMiner (<http://rapidminer.com>). (44.2%) Relacionado a la comunidad de base de datos. Parcialmente gratis.
- Python y Hadoop están subiendo últimamente.
- **Comercial:** SAS Enterprise Miner(10.5%), KMINE(15.0%), XLMiner(25.8%), Microsoft SQL(10.5%), Oracle(2.2%).

Data Mining segun Google trends

En todo el mundo ▾

2004 - hoy ▾

Todas las categorías ▾

Búsqueda web ▾

Comparar Términos de búsqueda ▾

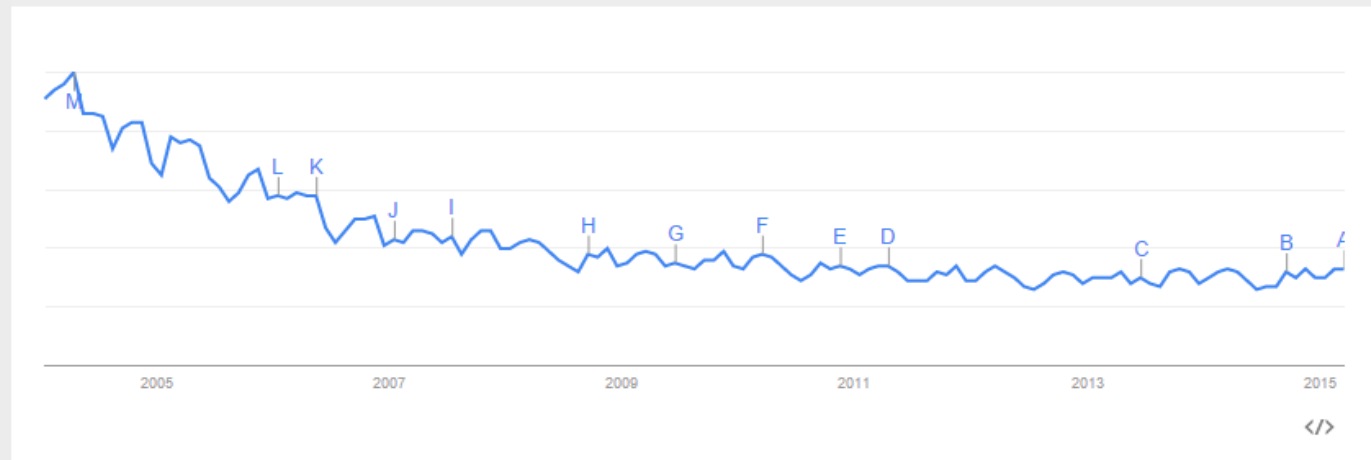
data mining

Término de búsqueda

+ Agregar término

Interés a lo largo del tiempo ?

Titulares de noticias Previsión ?



Big Data [1]: Antecedentes

Febrero 1977, L. Breiman organizo la conferencia “Analysis of large complex data sets”. Patrocinada por la ASA y IMS Dallas, USA.

1994 COMPSTAT, Proceedings of Computational Statistics. Part I. Treatment of “huge” Data sets.

Mayo 1997. The 29th Symposium on the Interface (Houston, TX) “ Data Mining and the analysis of large data sets”.

Octubre 1997, M. Cox y D. Ellsworth usaron el termino “big data” en una conferencia sobre visualization organizada por la la IEEE.

1998, Workshop on Massive data sets. Committee on Applied and Theoretical Statistics. NRC, USA.

Abril 1998, John Massey, jefe cientifico de SGI, presenta un paper “Big data... and the next wave of Infrastrass” en una reunion de la USENIX.

Big Data[2]: Antecedentes

En Febrero 2001, Doug Laney, analista para el Meta Group, menciona que los retos del análisis de datos van en tres direcciones: Volumen creciente (cantidad de datos), velocidad (rapidez de los datos que entran y salen), y variedad (una gran gama de tipos de datos y fuentes de datos).

Septiembre 2008, La revista Nature publica una edición especial acerca de “big data”.

Mayo del 2011, investigadores del McKinsey Global Institute publicaron: “Big data: the next frontier for innovation, competition and productivity”.

Big Data: Definicion

En 2012, Gartner actualizo la definicion de “Big data” como sigue: "Big data es informacion que se recibe en alto volumen, con alta velocidad, y/o alta variedad y que requiere nuevas formas de procesamiento para mejorar la toma de decisiones, entender mejor los descubrimientos y el proceso de optimizacion”.

Big data usualmente incluye bases de datos con tamanos mas alla de la capacidad del “software” comun para capturar, administrar, y procesar los datos dentro de un tiempo tolerable.

Los tamanos de Big data cambian constantemente. En 2002, era aprox 100GB, en 2012 de 10TB hasta petabytes para una sola base de datos.



Big data segun Google trends

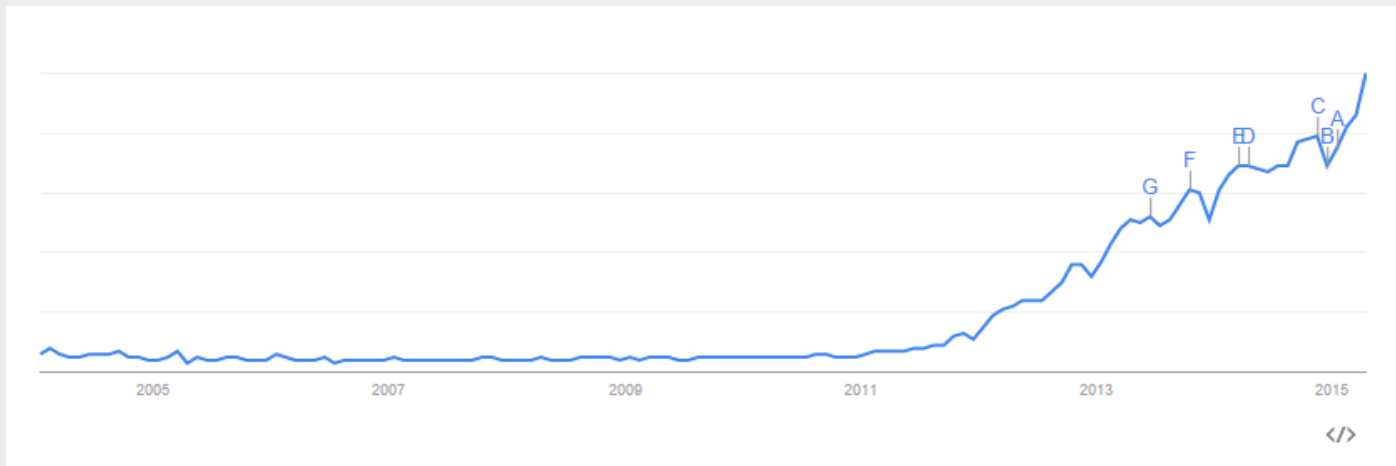
Comparar Términos de búsqueda ▾

big data
Término de búsqueda

+ Añadir término

Interés a lo largo del tiempo ?

Titulares de noticias Previsión ?



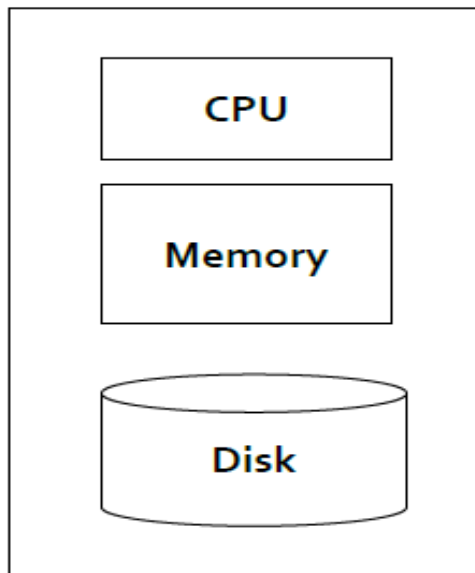
Big data es ...

Big data is like teenage sex:

everyone talks about it, nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it...

Dan Ariely (2013), Professor of Psychology. Duke University.

Arquitectura de un solo nodo



Machine Learning, Statistics

"Classical" Data Mining

Uso de R en computacion paralela y distribuida

RMPI: Paquete para correr R con MPI

Snow: Paquete para correr R con MPI

Gputools: Paquete para correr R con CUDA

Rmr: Interface entre R y Hadoop

Elio Lozano y Edgar Acuna: 2003-2006, varios articulos en aplicacion de R en computacion en paralelo y distribuida para calculo de metaclassifiers, metaclustering, boosting, bagging, kernel density estimation, ensembles and outlier detection.

Motivacion: Ejemplo de Google

Google busca en mas de 20 billion de paginas web x20KB= 400+ TB. Una computadora lee del disco con una velocidad de 30-35 MB/sec.

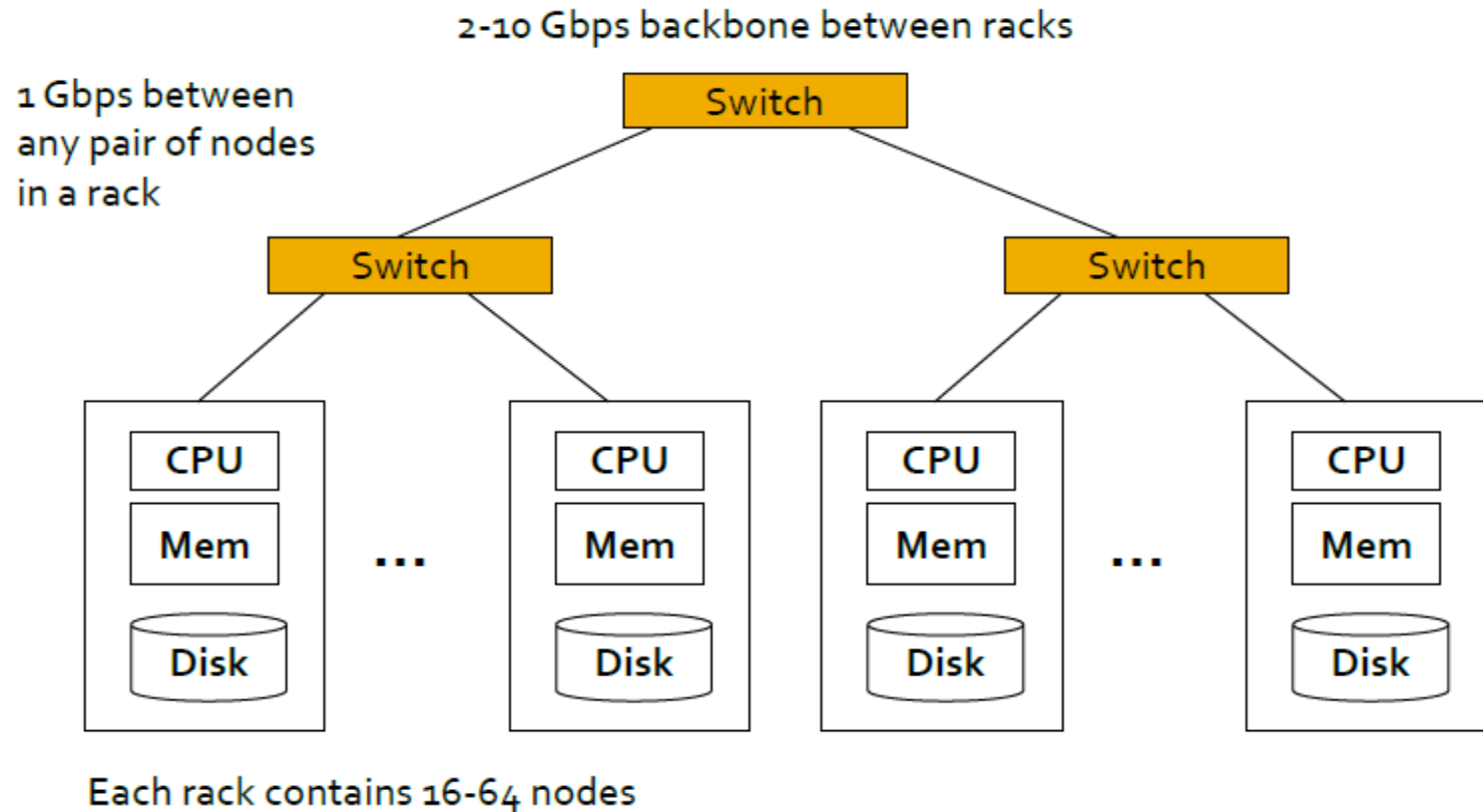
Le tomaria aprox 4 meses en leer la web.

Seria necesario aprox 1000 hard drives de 400GB para leer la web. Seria necesario aun mas disco duro para analizar los datos.

Hoy en dia, una arquitectura estandar esta siendo usada. Consiste de

- Un cluster de nodos Linux
- Una red de ethernet para conectar los nodos.

Arquitectura de clusters



Retos en computacion en gran- escala para mineria de datos

Como distribuir los computos?

Como escribir facilmente programas para computacion distribuida?

Las computadoras fallan!.

Una computadora puede funcionar sin problemas por tres anos (1000 dias)

Si hay 1000 servidores, se espera que en un dia cualquiera uno de ellos no funcione.

En 2011, se estimaba que Google tenia 1 million de computadoras, asi que en un dia cualquiera 1000 servidores podian fallar.

Que es Hadoop?

- En 2004, J. Dean y S. Ghemawat, empleados de Google escribieron un articulo explicando el modelo de programacion Mapreduce y una estructura para almacenamiento de grandes conjuntos de datos (sistema de archivos) llamado Google File System (GFS).
- MapReduce es open source pero GFS no lo es.
- En 2006, Doug Cutting de Yahoo! , creo un open source GFS y lo llamo Hadoop Distributed File System (HDFS). En el 2009, Cutting dejo Yahoo! y se fue a Cloudera.
- La estructura de software que apoya HDFS, MapReduce y otros programas relacionados es llamado el Proyecto Hadoop o simplemente Hadoop.
- Hadoop es distribuido por Apache Software Foundation.

Hadoop

Hadoop incluye:

Distributed Files System(HDFS) –que distribuye los datos.
Map/Reduce-que distribuye la aplicacion

Esta escrito en Java

Corre en

- Linux, MacOS/X, Windows, y Solaris
- Usa hardware comun

Distributed File System

Los servidores de Chunks:

El archivo es dividido en varios trozos (chunks) contiguos.

Tipicamente cada trozo es 16-128 MB

Cada pedazo es repetido (usualmente 2x y 3x)

El Nodo master:

Llamado el Namenode almacena metadata acerca de donde son almacenados los archivos.

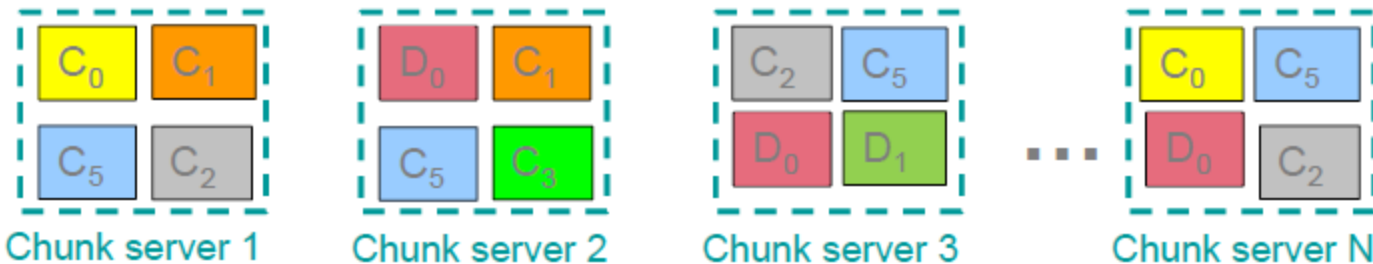
La libreria cliente para accesos de archivos.

Esta libreria trata de encontrar los servidores de chunks.

Se conecta directamente a los servidores de chunks para acceder a los datos.

Sistema de archivos distribuidos

- **Reliable distributed file system**
- Data kept in “chunks” spread across machines
- Each chunk **replicated** on different machines
 - Seamless recovery from disk or machine failure



Bring computation directly to the data!

Chunk servers also serve as compute servers

Arquitectura HDFS

- Provee
 - Paralelizacion y distribucion automatica
 - Tolerancia a Fallos
 - Planificacion de I/O
 - Monitoreo y status de los updates.

MapReduce

Map-Reduce es un modelo de programación para una computación distribuida eficiente.

Trabaja como un pipeline de Unix :

-cat input |grep | sort |unique -c |cat > output

- **Input | Map | Shuffle & Sort |Reduce |Output**

Es eficiente porque reduce la búsqueda y el uso de pipeline.

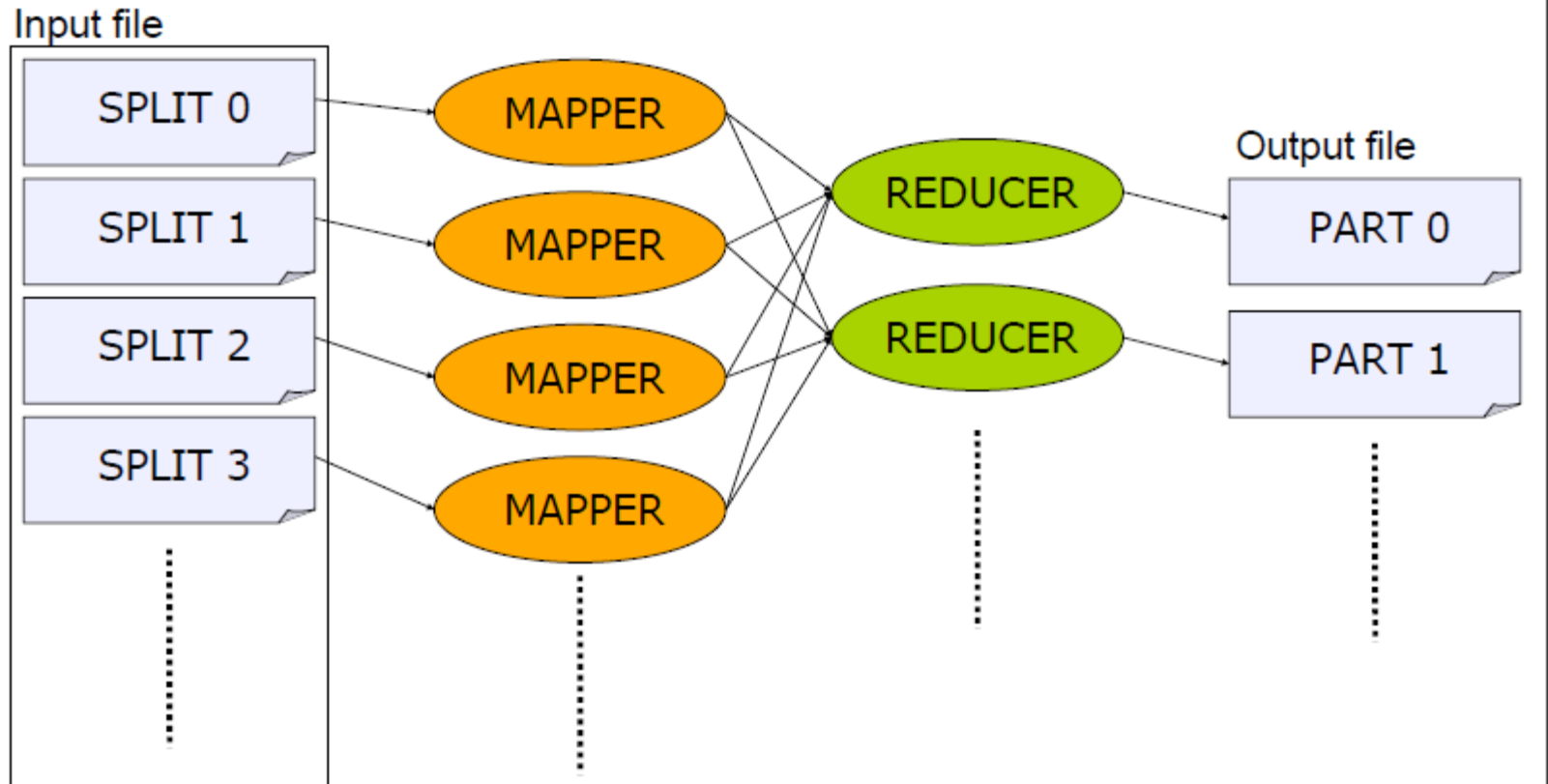
Modelo MapReduce

- Input & Output: Un conjunto de parejas key/value
- Dos operaciones basicas
 - **map:** $(k_1, v_1) \rightarrow \text{list}(k_2, v_2)$
 - **reduce:** $(k_2, \text{list}(v_2)) \rightarrow \text{list}(k_3, v_3)$
- Cada operacion map procesa una entrada de una pareja key/value pair y produce un conjunto de parejas key/value
- Cada operacion reduce
 - Junta todos los valores intermedios producidos por operaciones map para una key particular.
 - Produce al final un conjunto de parejas key/value

El flujo de trabajo del modelo

- Cargar los datos en el cluster (HDFS escribe)
- Analizar los datos (MapReduce)
- Almacenar los resultados en el cluster (HDFS)
- Leer los resultados del Cluster (HDFS lee)

Ejecucion del modelo:Flujo

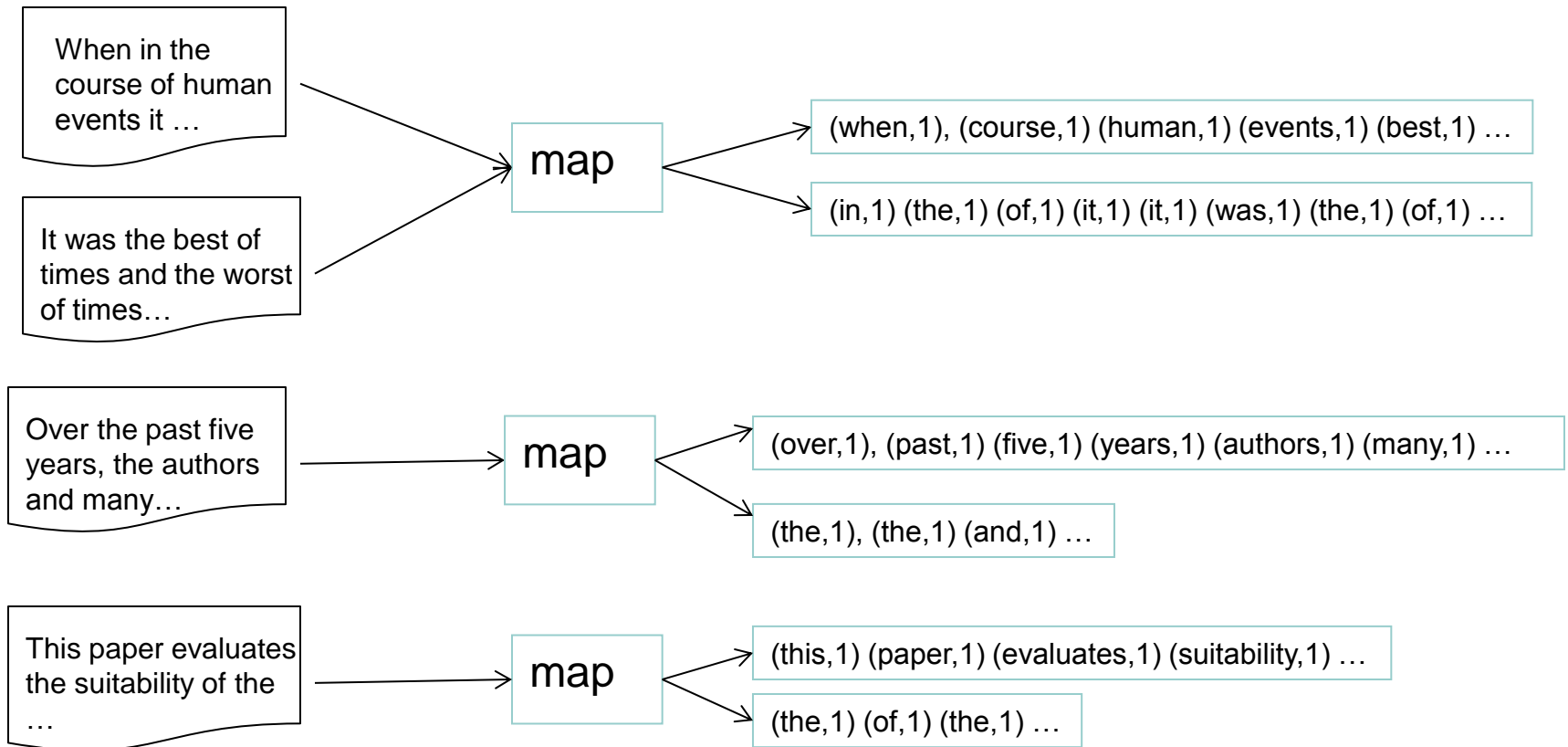


El Map (Ejemplo: Wordcount)

inputs

tasks (M=3)

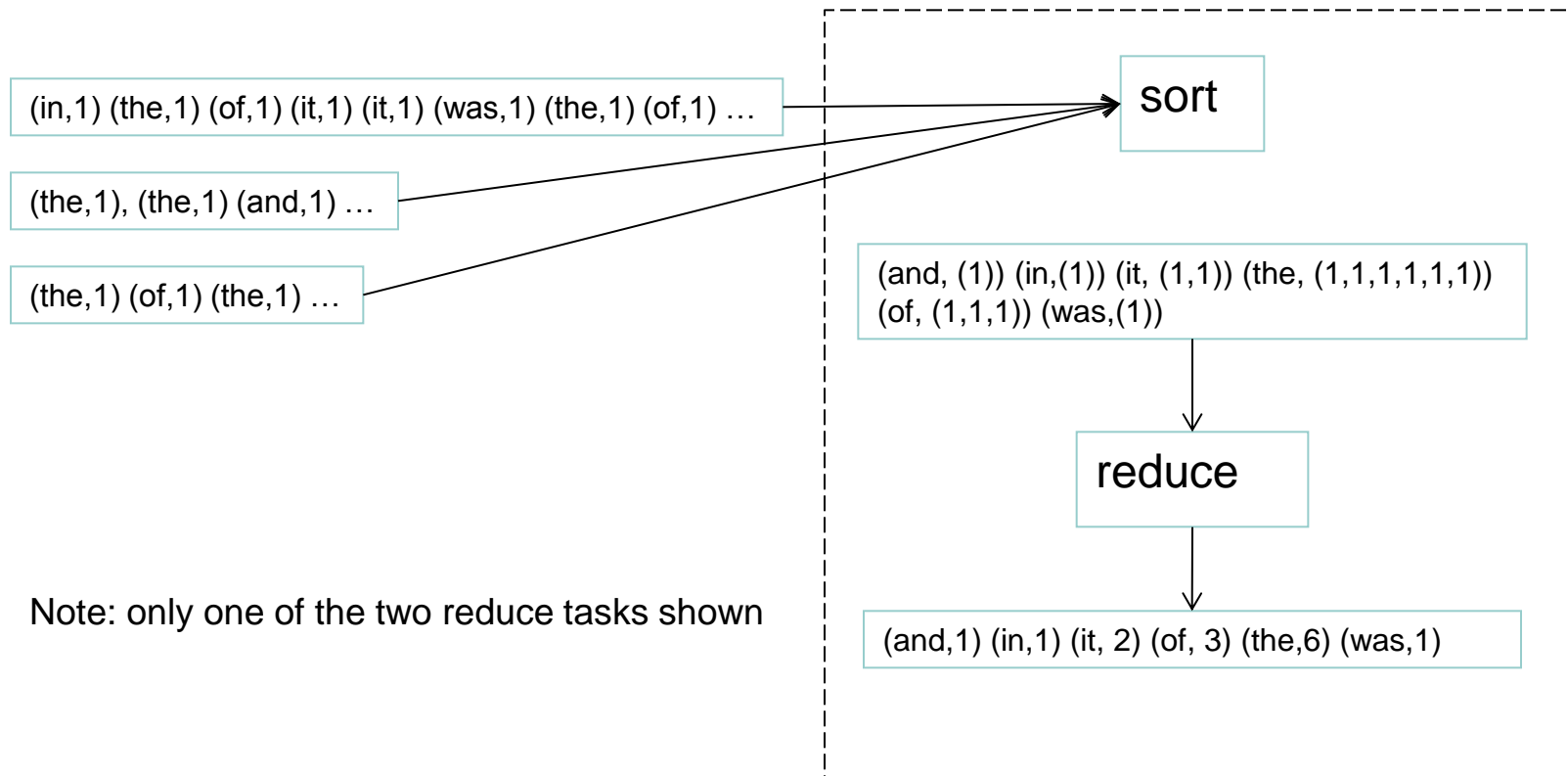
partitions (intermediate files) (R=2)



El Reduce (Ejemplo: Wordcount)

partition (intermediate files) (R=2)

reduce task



Note: only one of the two reduce tasks shown

Interfaces de otros programas con Hadoop

Hadoop streaming: C++,Python, perl,ruby,

Rhadoop (R and Hadoop), Weka(Mark Hall esta trabajando en esto), Radoop (Rapidminer and hadoop, comercial)

Hadoop Pipes: (C++) no es recomendable.

Donde se puede correr hadoop?

En su computadora personal usando hadoop para un cluster de un solo nodo. Si usa Windows instale a maquina virtual (Vmware Player) donde corra Ubuntu(Ver el website de Michael Noll)

Gratis:

En el cluster Gordon del San Diego Supercomputer Center (1024 nodes) a traves del proyecto XSEDE.

En varios clusters: India(128 nodes), Sierra(84 nodes) y Alamo del proyecto Futuregrid liderado por la Universidad de Indiana. Lamentablemente el proyecto se esta terminando.

No-Gratis, pero no muy costoso

Amazon Elastic Compute Cloud (EC2)

Quienes estan usando Hadoop?

Yahoo
Facebook
Amazon
Google
IBM
Netflix
Ebay
LinkedIn
Twitter

Ejemplo de Rendimiento: Hadoop in Yahoo!

- La base de datos para **Search Assist™** esta construida usando Hadoop.
- 3 anos de datos

	Sin Hadoop	Con Hadoop
Tiempo	26 dias	20 minutos
Lenguaje	C++	Python
Tiempo de programacion	2-3 semanas	2-3 dias

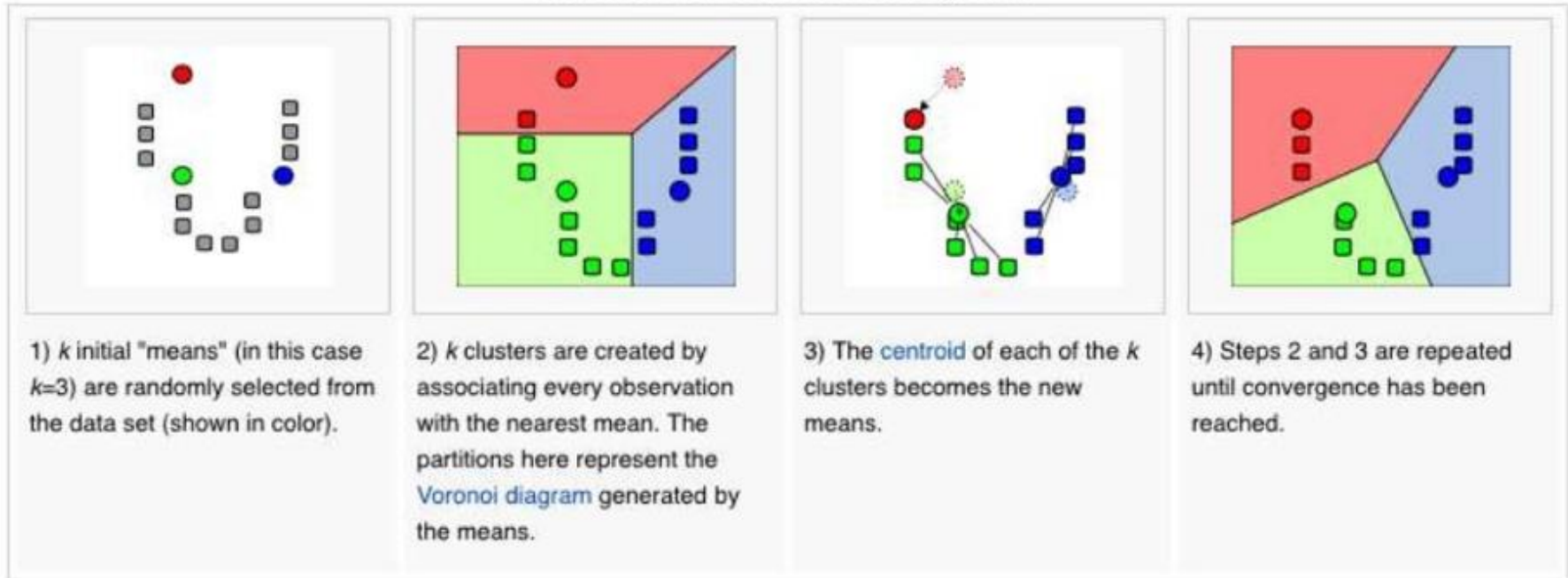
Escribiendo un programa usando MapReduce

Un programa en Hadoop necesita:

- Una clase en java con un main que configura y somete el trabajo
- Una clase que extiende la clase Hadoop Mapper (“The Mapper”)
- Una clase que extiende la clase Hadoop Reducer (“The Reducer”)
- Opcional: Una clase que extiende la clase Hadoop Reducer (“The Combiner”).

Ejemplo de Minería de datos: Clustering usando k-means

Demonstration of the standard algorithm



Iterative algorithm until converges

El Algoritmo K-Means

- **Paso 1:** Seleccionar K puntos al azar (centros)
- **Paso 2:** Asignar cada dato al centro mas cercano
 - Se han formen ahora K clusters
- **Paso 3:** Para cada cluster, re-calcular los centros
- **Paso 4:** Si los nuevos centros son distintos a los anteriores(iteracion previa) → Ir al paso 2

K-Means en MapReduce

- **Input**

- La base de datos
- Centroides iniciales (K puntos)

- **El paso Map**

- Cada map lee los K-centroides + un block de la base de datos
- Asignar cada punto al centroide mas cercano
- Output <centroide, punto>

K-Means en MapReduce (Cont.)

- **La etapa Reduce**
 - Consigue todos los puntos para un centroide dado
 - Re-calcula un nuevo centroide para este cluster
 - Output: <nuevo centroide>
- **Control de la Iteration**
 - Comparar el nuevo conjunto de centroide con el anterior
 - Si son similares → Stop
 - Else
 - Si se ha alcanzado el max iterations → Stop
 - Else → Empezar otra iteracion Map-Reduce

Kmeans applied to 20Newsgroup dataset (Jermaine, 2012)

Este es un conjunto de 20,000 postings a 20 diferentes newsgroups (estos son como los blogs de hoy en dia). El directorio "20_newsgroups" tendra 20 subdirectorios, cada uno correspondiente a un newsgroup. Hay 1000 archivos en cada directorio, cada uno de los cuales corresponde a un posting.

Primer los documentos son convertidos en vectores de canasta de palabras (bag-words). Usaremos un diccionario de 1000 words and 20000 documentos(1000 for cada uno de los 20 groups).

- `hadoop dfs -mkdir /data`
- `hadoop dfs -mkdir /clusters`
- `hadoop dfs -copyFromLocal vectors /data`
- `hadoop dfs -copyFromLocal clusters /clusters`

Falla de Big data

En Febrero del 2009 , investigadores de Google trataron de predecir la propagacion de la Influenza construyendo un modelo basado en las busquedas que hacia la gente en Google sobre 45 terminos relacionados a Influenza. (Nature, February 2009).

En los primeros dos anos el modelo tuvo un aceptable nivel de confiabilidad. Pero a partir del 2011 el modelo comenzo a fallar el principio en un 50% llegando hasta un 92% de error (Science, March 2014).

Otras herramientas computacionales

Spark del laboratorio AMP de UC Berkeley. Hay que programar en Scala, Java o Python. Gratis.

Watson Analytics de IBM (dueno de SPSS). Gratis para las universidades.

Azure de Microsoft Research, software hecho en Linux. Gratis para las universidades.

Antecedentes de Ciencias de Datos

John Tukey (Annals of Mathematical Statistics”, 1962). “The Future of Data Analysis”. Tukey se declara analista de datos y duda que sea un estadístico matemático.

En 1996, la International Federation of Classification Societies (IFCS) hacen una conferencia en “Data Science, Classification and Related Methods”.

J. Wu (1997). Propone cambiar el nombre de Estadística a Ciencias de Datos. Además de que debería haber un enfoque en el análisis de datos grandes y complejos. Estadístico=Científico de datos.

Antecedentes de Ciencias de Datos

J. H. Friedman (2000). “The role of the statistics in the data revolution”. Stanford University Tech Report. Propuso absorber Minería de Datos dentro de Estadística.

L. Breiman (Statistical Science, 2001). Statistical Modelling: The two cultures. Afirma que: “Si el objetivo de la estadística es usar datos para resolver problemas entonces deberíamos alejarnos de usar solamente “data models” y comenzar a usar “algorithmic models”.

Antecedentes de Ciencias de Datos

W. Cleveland (2001). "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics". Propone la creación de la carrera de Ciencias de Datos, similar a Estadística multidisciplinaria.

En el año 2002, comienzan a publicarse "Data Science Journal" y el "Journal of Data Science".

En el 2008, D.J.Patil (LinkedIn) y Jeff Hamerbacher (Facebook) introdujeron el término científico de datos.(????)

Ciencias de Datos como disciplina

En Enero del 2009. Hail Varian, el Jefe economista de Google afirma que el trabajo mas excitante por los proximos 10 anos sera ser estadistico!!!

Entre 2009 y 2010, mucha gente de ciencias de computacion y de administracion de empresas comenzaron a empujar la idea de la estadistica a la que se referia Hal Varian no era la tradicional sino un nuevo campo llamado “Data Science”.

T.H Davenport y J.H. Patil (2012, Harvard Business Review) inclinaron la balanza a favor de Ciencias de Datos

Estadística versus Ciencias de Datos

	Estadístico	Científico de Datos
Imagen	Basket/Futbol	Trabajo mas sexy del siglo
Trabaja	Solo	En equipo
Datos	Preparados, limpios	Distribuidos, Sucios, No estructurados.
Tamaño de datos	Kilobytes	Gigabytes
Tools	SAS, un CPU	R, Python, Hadoop, Linux
Focus	Inferencia	Predicción
tiempo	semanas	segundos

Data Science segun Google Trends

Todo el mundo ▾

2004 - hoy ▾

Todas las categorías ▾

Búsqueda en la Web ▾

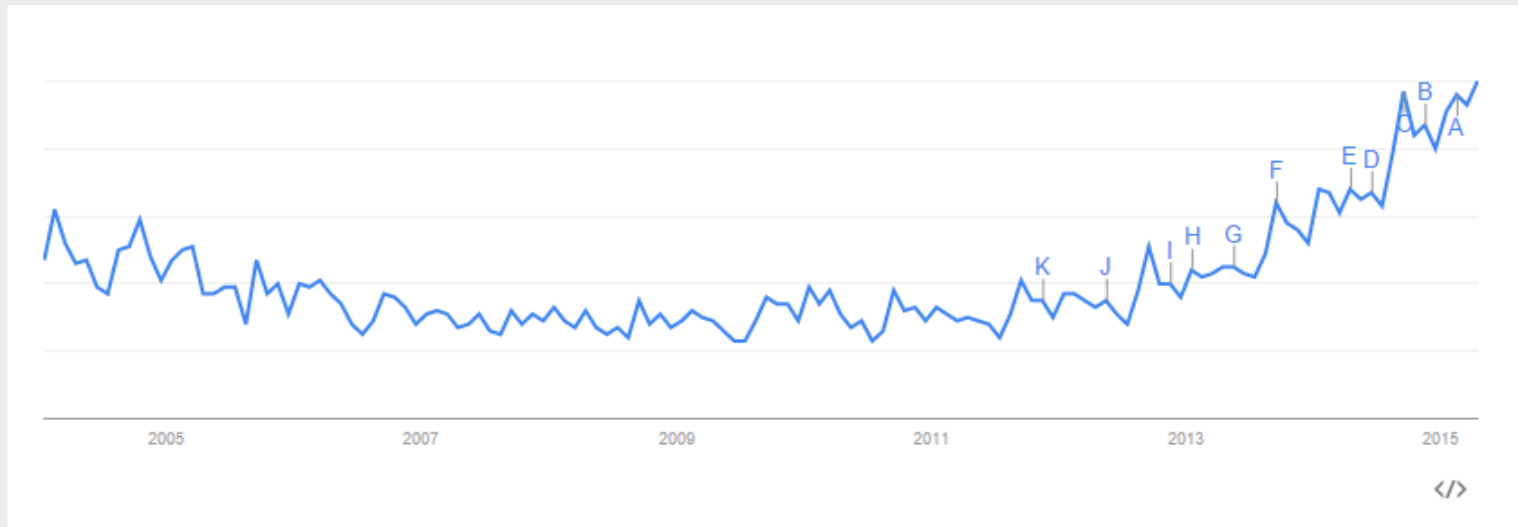
Comparar Términos de búsqueda ▾

data science
Término de búsqueda

+ Añadir término

Interés a lo largo del tiempo ?

Titulares de noticias Previsión ?



Programas en Ciencias de Datos

Bachillerato: University of Rochester, University of San Francisco, College of Charleston, Northern Kentucky University, Case Western University (mas de una decena en USA). Warwick University (UK).

Maestria: Carnegie Mellon University, Stanford University, Columbia University, Indiana University, NYU, SMU, Virginia. Mas de 30 en USA, Canada, UK, Espana y resto del mundo.

Doctorado: Edimburgh University (UK), Rochester, Columbia. No pasan la decena aun.

Latinoamerica: Solo Mexico (ITAM desde el 2014. Ni Brazil ni Chile.

Curriculum de un Bachillerato en Ciencias de Datos

Calculo Multivariado

Algebra lineal (matrices)

Dos cursos de probabilidad y estadística

Introducción a Ciencias de datos (estadística aplicada actualizada)

Data Mining

Machine Learning (aprendizaje automático)

Regresión

Muestreo

Estadística Bayesiana

Estadística No paramétrica

Visualización de datos

Bases de Datos

Algoritmos

Computación en Paralelo (computación de alto rendimiento)

Cursos de Economía/Administración de Empresas/Teoría de Juegos

Comparacion de Data Mining, Big Data, Bioinformatica y Data Science segun Google Trends

Todo el mundo ▾

2004 - hoy ▾

Todas las categorías ▾

Búsqueda en la Web ▾



Comparar Términos de búsqueda ▾

data sci...

Término de bús...

Big data

Término de bús...

data mi...

Término de bús...

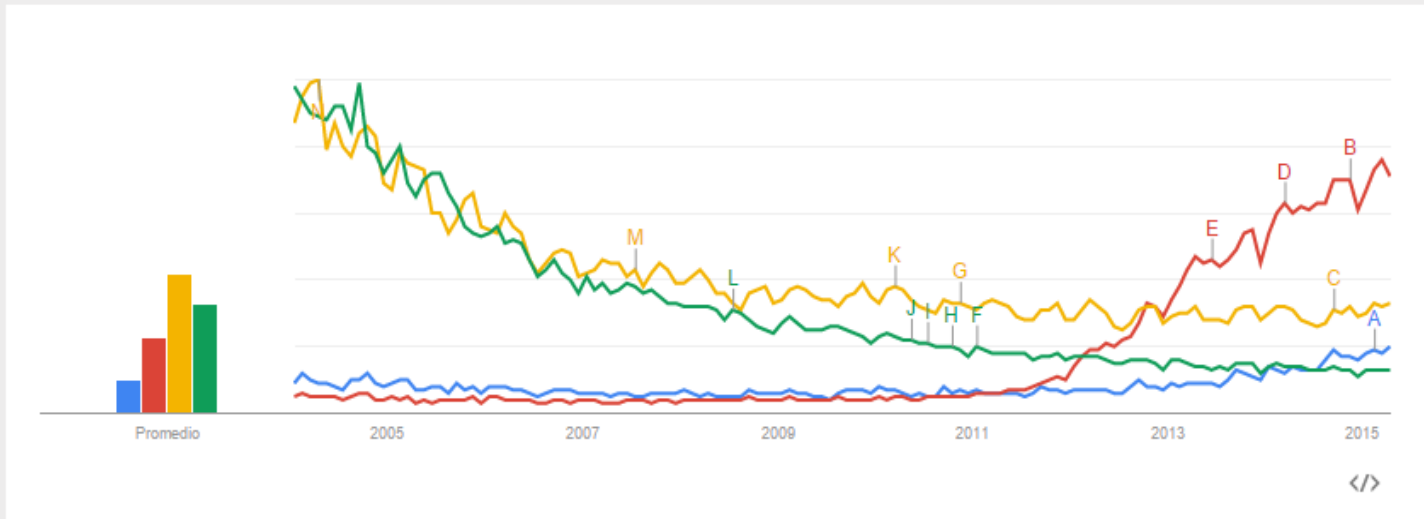
bioinfor...

Término de bús...

+ Añadir término

Interés a lo largo del tiempo ?

Titulares de noticias Previsión ?



Mi investigación en progreso

Objetivo; Evitar el GIGO(Basura que entra-Basura que sale)
Estamos implementando varias tecnicas de preprocesamiento de datos en Mapreduce y Hadoop.

- The ReliefF:Metodo de seleccion de atributos (Carlos Gomez)
- Deteccion de Outlier (tesis doctoral de Carlos Gomez)
- Reduccion de la dimensionalidad
- Tratamiento de valores faltantes
- Tratamiento de datos de distintos tipos (cuantitativos y nominales)
- Extendiendo R usando Hadoop