Herramientas para Mineria de Datos Masivos(Big Data)

Dr. Edgar Acuna
Departamento de Ciencias Mathematicas y Programa Doctoral en Ciencias e Ingenieria de la Informacion y Computacion
Universidad de Puerto Rico-Mayaguez

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

Website: academic.uprm.edu/eacuna

Contenido

I. Mineria de Datos
II. Datos Masivos (Big Data)
III. Hadoop y MapReduce
IV. Ejemplo: K-means
V. Investigacion en progreso

Mineria de Datos [1]: Motivacion

Los mecanismos pra la recoleccion automatica de datos (por ejemplo, sensores) y el desarrollo de tecnologia 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 informacion.

Hoy dia, hay la necesidad de convertir estos datos en conocimiento e informacion.

Mineria de datos [2]Que es Mineria de Datos?

- Es el proceso de extraer conocimiento/informacion valida de una base de datos bien grande. El conocimiento es dado como caracteristicas (“patrones”)y reglas que son notriviales, previamente desconocidas, comprensibles y con alto potencial de ser utiles.
- Otro nombres: Descubrimiento de Conocimiento en bases de datos (KDD), Analisis Inteligente de datos, Ingenieria de datos, Inteligencia de negocios.
- El primer articulo en Mineria de datos aparecio en 1993.

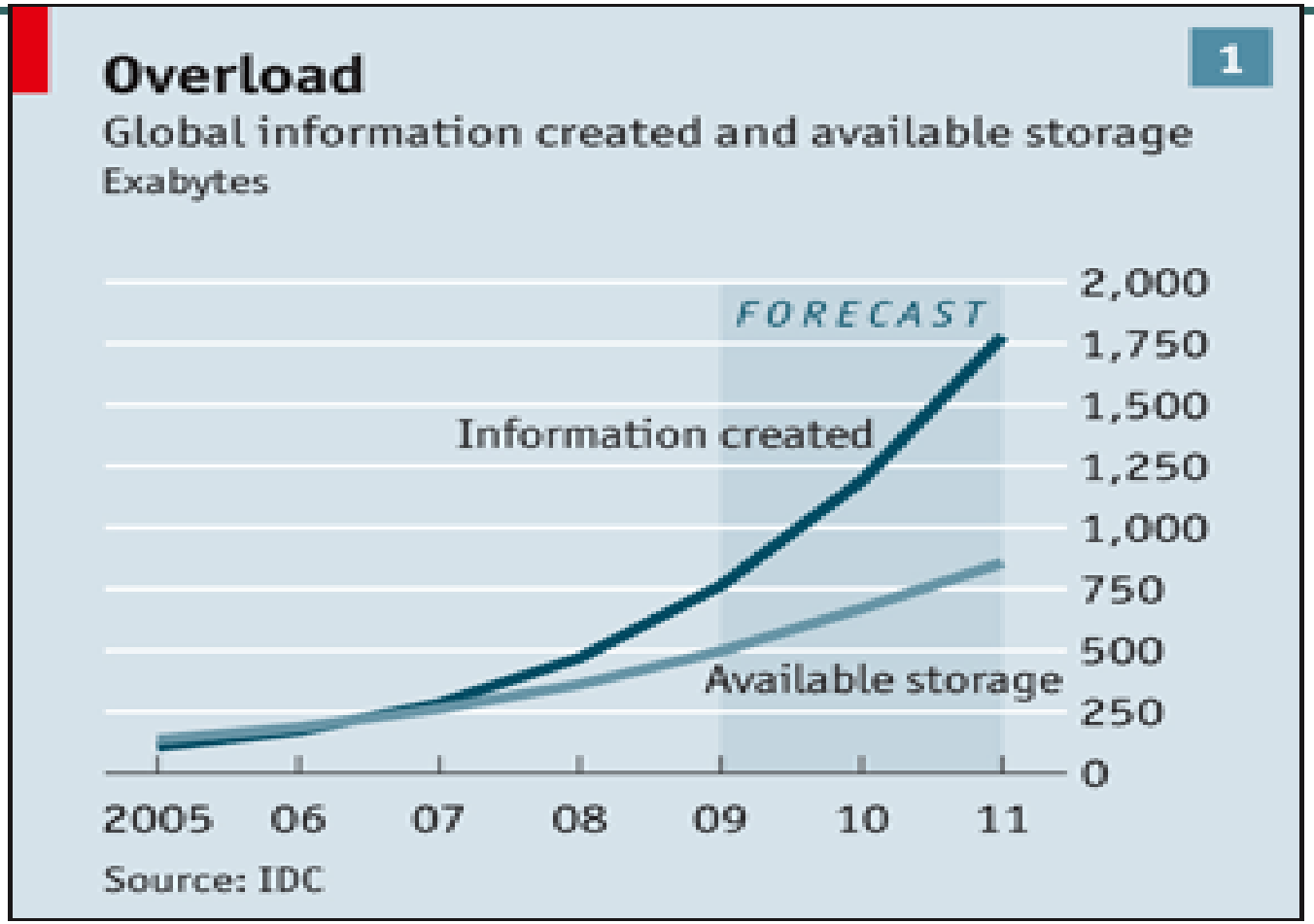
Mineria de Datos[3]: Tamano (en Bytes) de bases de datos.

Descripcion	Tamano	Medio de Almacenaje
Muy pequeno	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})		

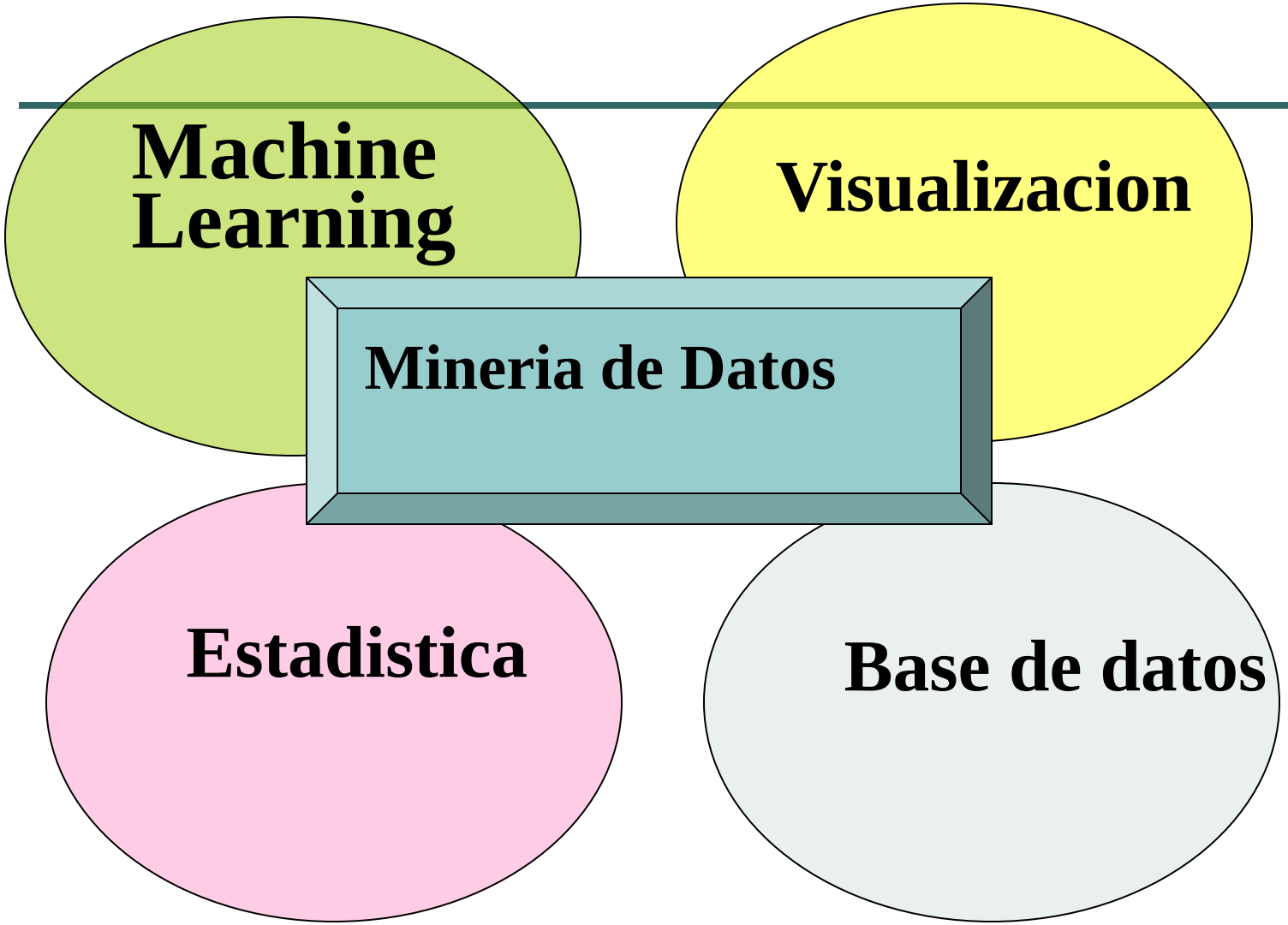
Ejemplos de grandes bases de datos

- Un telescopio puede generar hasta 1 gigabyte de datos astronomicos por segundo.
- ATT almacena hasta 26 Terabytes de informacion en llamadas telefonicas.
- El 2008, Walmart almacenaba 10 Terabytes de transacciones por dia.
- El 2011, Google buscaba en mas de 20 billones de paginas, que representa mas de 400 TB.
- Se estima que en el 2009 habia 500 exabytes (1,000 petabytes) de informacion en la internet.
- El 2013, se anuncio que el centro de datos de la NSA seria capaz de almacenar 5 zettabytes (1,000 exabytes).

The economist, February 2010



Mineria de datos[5]. Areas relacionadas



Estadistica, Machine Learning

- Estadistica (~35% de DM)
  - Se basa mas en teoria. Asume propiedades distribucionales de las variables que estan siendo consideradas.
  - Se enfoca mas en probar hipotesis y en estimacion de parametros.
  - Se consideran efiicientes estrategias de recolectar datos.
  - Estimacion de modelos.
- “Machine learning” (Aprendizaje Automatico) (~30 % de DM)
  - Parte de Inteligencia Artificial. Machine es equivalente a un modelo en estadistica.
  - Mas heuristica que Estadistica.
  - Se enfoca en mejorar el rendimiento de un clasificador basado en sus experiencias pasadas.
  - Tambien considera el tiempo que dura el proceso de aprendizaje.
  - Incluye a: Redes Neurales, arboles de decision, algoritmos geneticos.

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.
  - Fue introducida por E. F. Codd de IBM en 1970.
  - El interface entre el usuario y la base de datos relacional mas usado es SQL(structured query language).
  - Una base de datos relacional puede ser agrandada facilmente
- 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, Expert Systems, High Performance Computing.

Data Mining no es …

- Buscar un numero en una guia telefonica
- Buscar una definicion en Google.
- Generar histogramas de salarios por grupos de edad.
- Hacer un SQL query y leer la respuesta del query.

Data mining es ...

- Determinar las caracteristicas de personas a las que se puede hacer un prestamo bancario.
- Detectar intrusos (casos anomalos) en un sistema.
- Determinar las caracteristicas de los clientes de un banco que pueden cometer fraude.
- Determinar las caracteristicas de los clientes que abandonan la subscripcion a un servicio.
- Hallar ciudades de un pais que tienen caracteristicas similares (regionalizacion).

Aplicaciones de DM

Administracion de negocios: Investigacion de mercados, relacion de los clientes con la gerencia, deteccion de Fraudes, Telecomunicaciones, etc.

Gobierno: deteccion de evasores de impuestos, terrorismo.

Ciencias: Astronomia, Bioinformatica (Genomics, Proteonomics, Metabolomics), decubrimiento de medicinas.

Text Mining: Extraer informacion 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 encuentra las propiedades generales de la base de datos. Se descubre las caracteristicas mas 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 mas 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)

Regresion

- Se predice el valor de una variable de respuesta continua basado en los valores de otras variables (predictoras) asumiendo que hay una relacion funcional entre ellas.
- Se puede usar modelos estadisticos, arboles de decision o redes neurales.
- Ejemplo: ventas de carros basados en las experiencia de los vendedores, publicidad, tipo de carros, etc.

Clasificacion 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.

Examplos de clasificadores

Analisis Discriminante Lineal (LDA),

Naïve Bayes,

Regresion Logistica

k-nearest neighbor,

Decision trees,

Bayesian Networks

Neural Networks

Support Vector machine (SVM)

…….

Clasification 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 encontar reglas de dependencia que permitan predecir la ocurrencia de un item basado en ocurrencia de otros items

TID	*Items*
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->Y) deben satisfacer un soporte minimo y una confianza impuesta por el usuario. X es llamado el antecedente Y es llamado el consecuente.
- Soporte=(# registros conteniendo X y Y)/(# registros)
- Confianza=(# registros conteniendo X y Y/(# 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 MiningSoftware

- **Open source**
- R (cran.r-project.org). Relacionado a Estadistica (38.5% de usuarios, Kdnuggets.com 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.
- **Comercial:** SAS Enterprise Miner(10.5%), XLMiner(25.8%), Microsoft SQL(10.5%), Oracle(2.2%).

Big Data [1]: Definicion

En 2001, Doug Laney, analista para el Gartner Group, menciono que los retos del analisis de datos va 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).

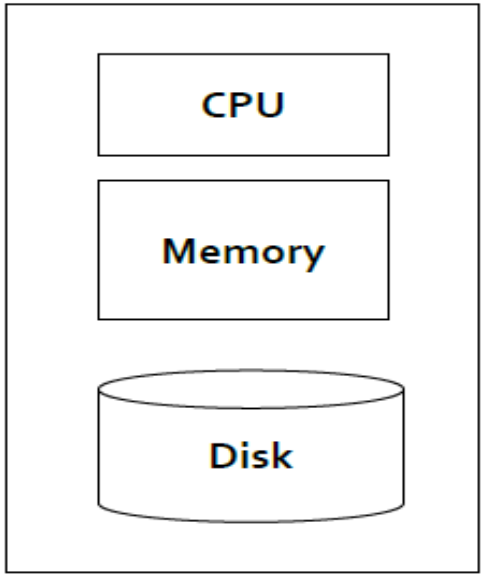
En 2012, Gartner actualizo su informacion como sigue: "Big data es informacion que se recibe en alto volumen, con alta velocidad, y/o alta variedad y que requiere nuevas formas de procesamiento para mejorar la toma de decisiones, entender mejor los descubrimientos y el proceso de optimizacion".

Big data usualmente incluye bases de datos con tamanos mas alla de la capacidad del "software" comunmente 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[2]: Ejemplos

- El telescopio Large Hadron Collider (LCH) almacena al ano cerca de 25 Petabytes de datos de sensores.
- En 2010, la base de datos de llamadas de ATT era de 323 Terabytes.
- El 2010, la tienda Walmart manipulaba 2.5 Petabytes of transacciones.
- En 2009, habia 500 exabytes de informacion en la internet.
- En 2011, Google buscaba en mas de 20 billions de paginas web. Esto representa aprox. 400 TB.
- El 2013, se anuncio que el centro de datos de la NSA en Utah almacenara 5 zettabytes (5,000 exabytes) de datos.

Arquitectura de un solo nodo



Machine Learning, Statistics

"Classical" Data Mining

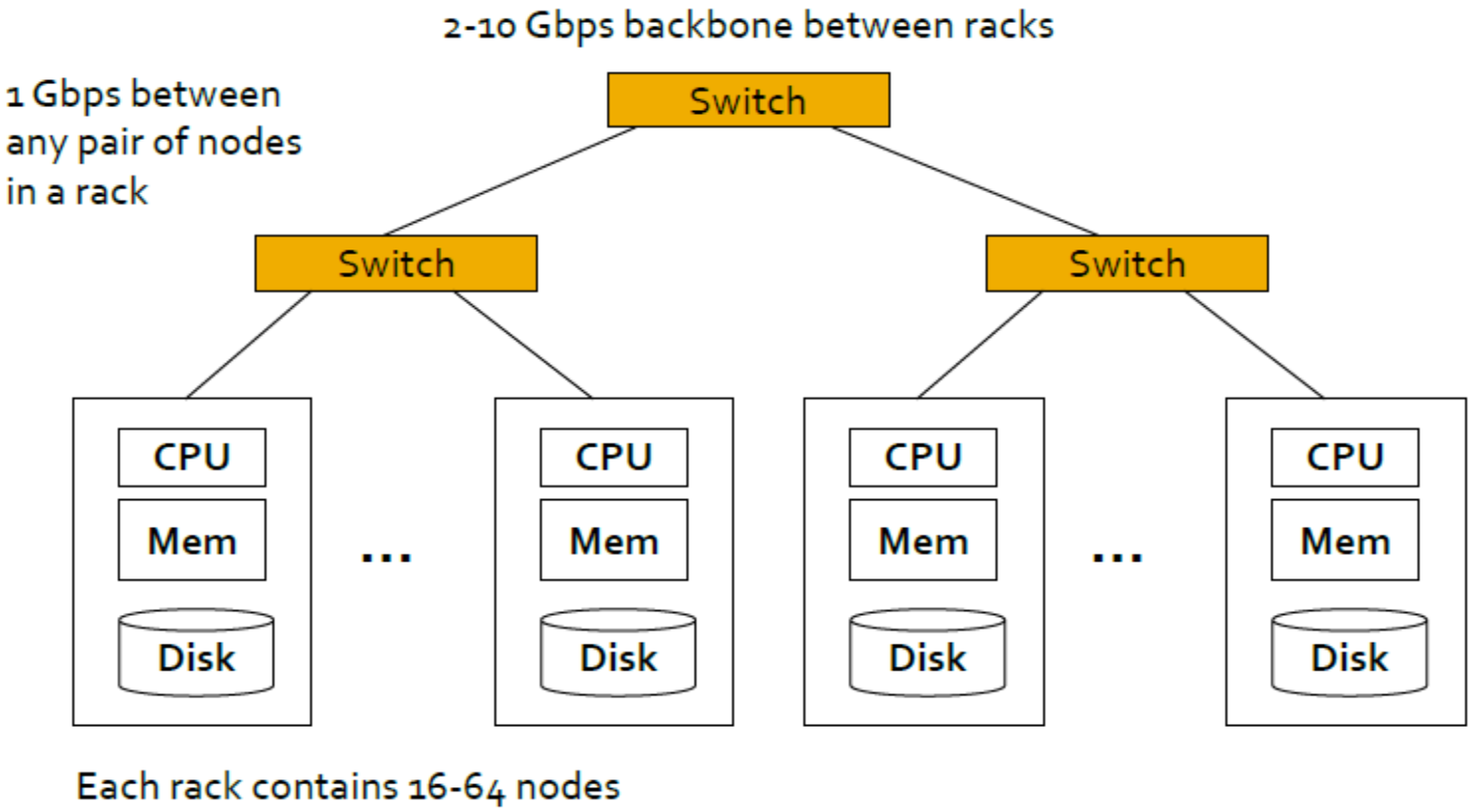
Motivacion: Ejemplo de Google

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?
Ls 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. Ghemawhat, 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, el 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

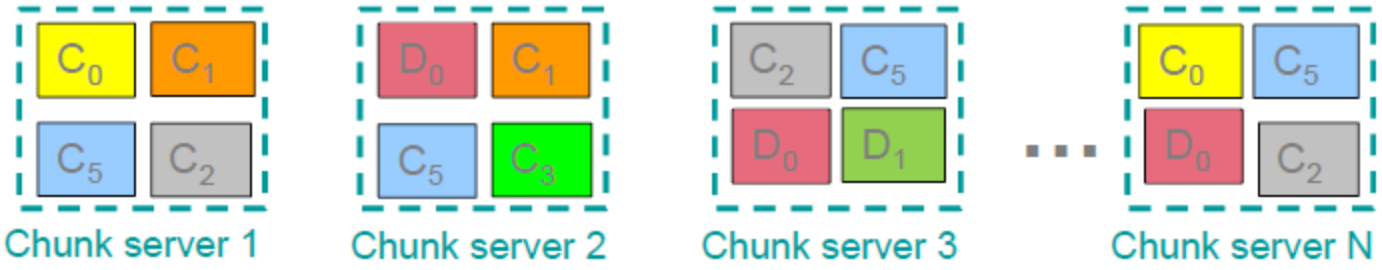
Los servidores de Chunks:
El archivo es dividido en varios trozos (chunks)contiguos.
Tipicamente cada trozo es 16-128 MB
Cada pedazo es repetido (usualmente 2x y 3x)

El Nodo master:
Llamado el Namenode alamcena 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 accesar 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



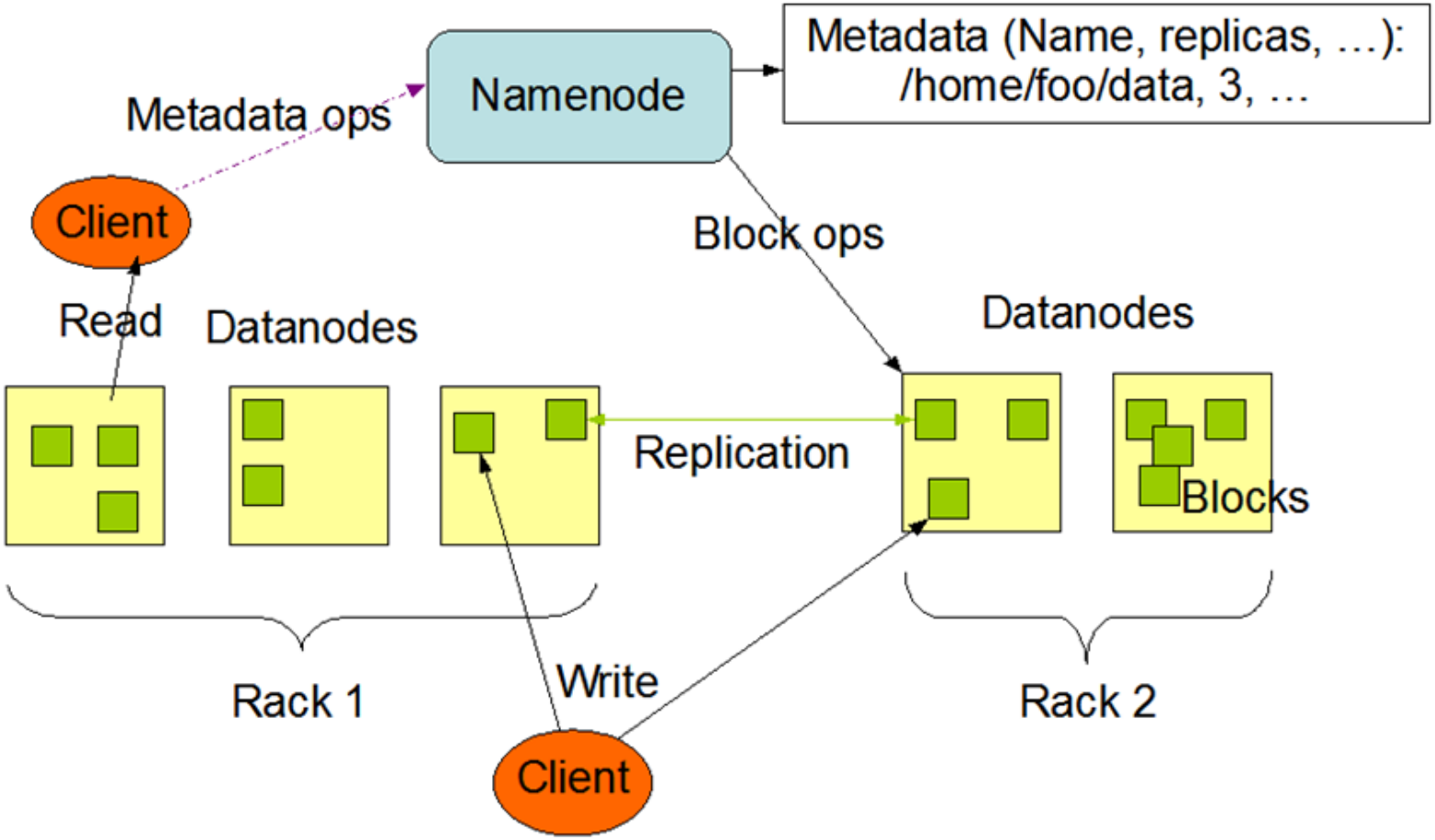
Bring computation directly to the data!

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.

Arquitectura HDFS [2]



MapReduce

Map-Reduce es un modelo de programacion para una computacion distribuida eficiente.

Trabaja como un pipeline de Unix :
-cat imput |grep | sort |unique –c |cat > ouput
- **Input | Map | Shuffle & Sort |Reduce |Output**

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

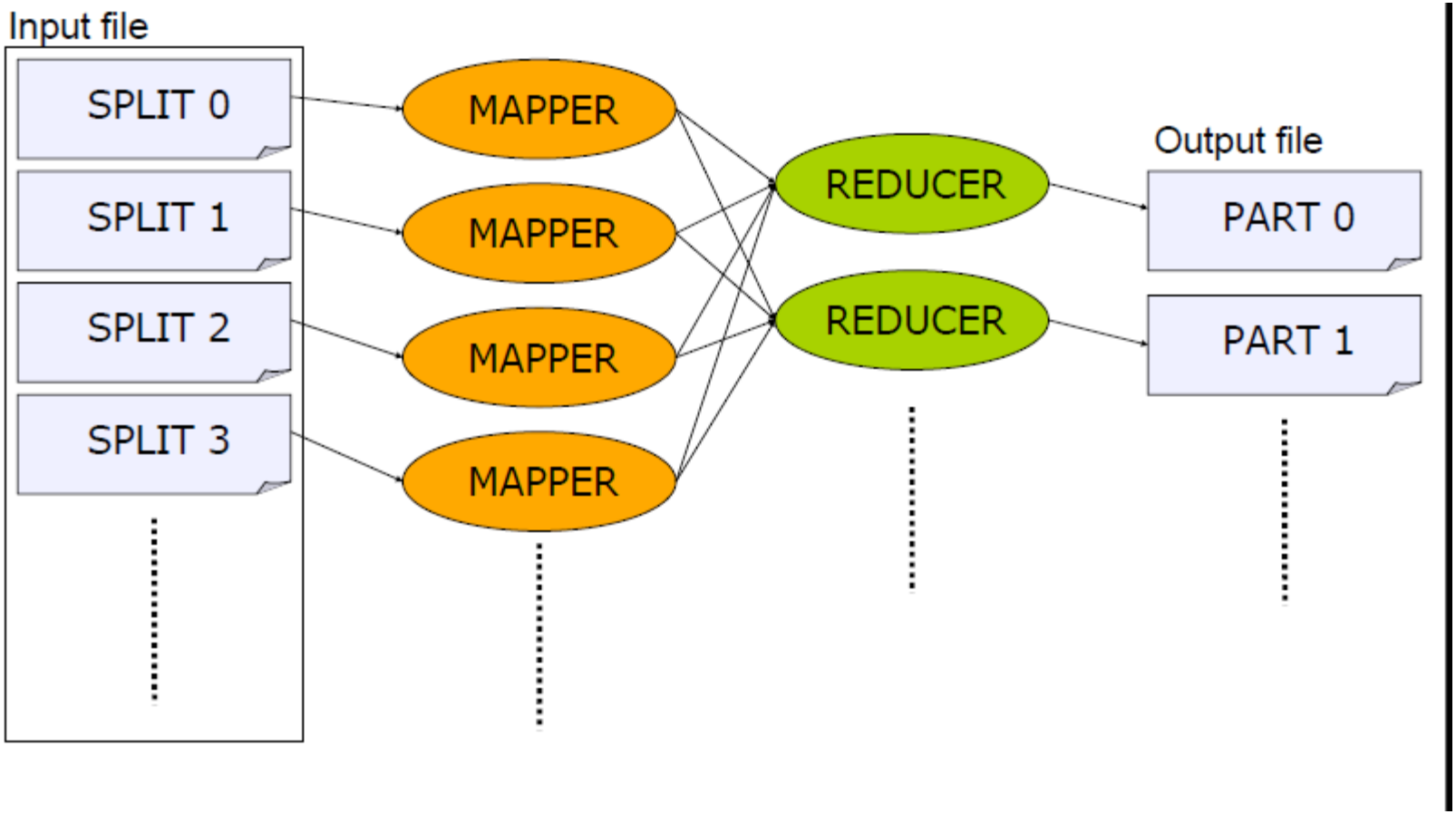
Modelo MapReduce

- Input & Output: Un conjunto de parejas key/value
- Dos operaciones basicas
  - **map**: $(k_1, v_1) \rightarrow list(k_2, v_2)$
  - **reduce**: $(k_2, list(v_2)) \rightarrow list(k_3, v_3)$
- 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

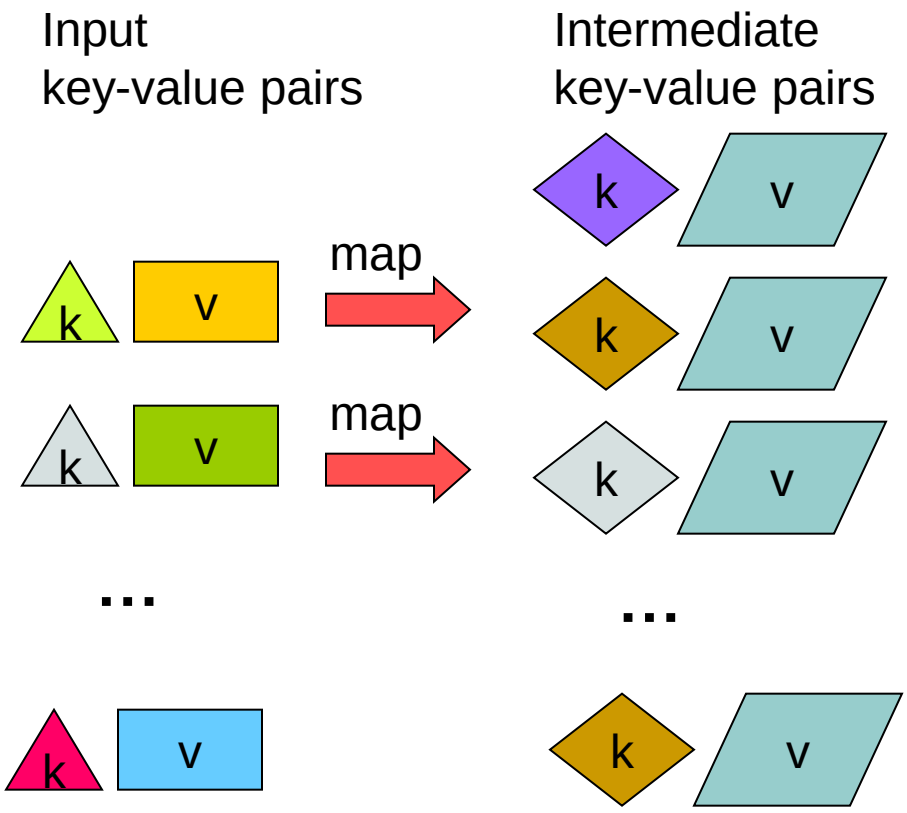
El flujo de trabajo del modelo

- 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)

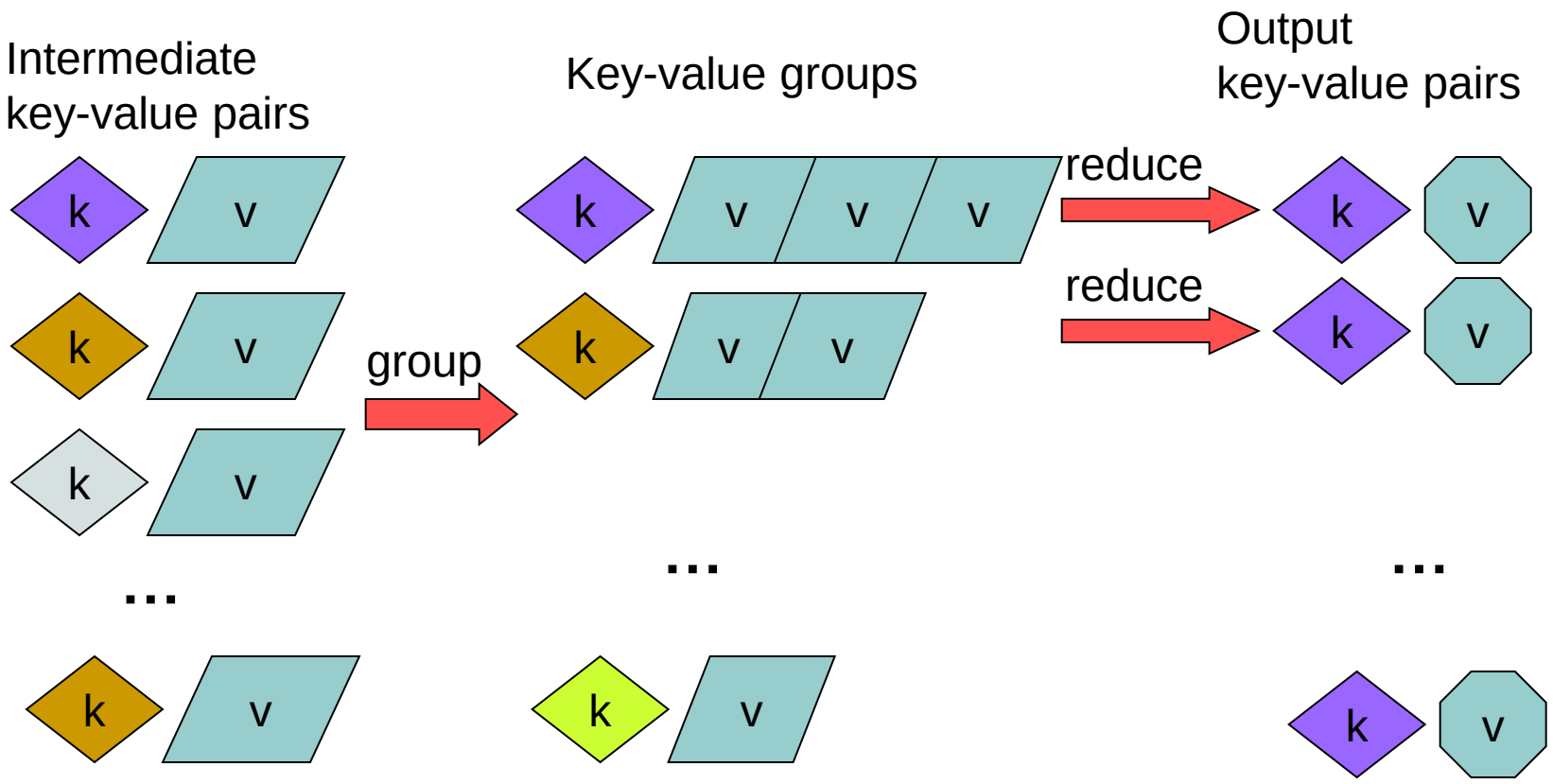
Ejecucion del modelo:Flujo



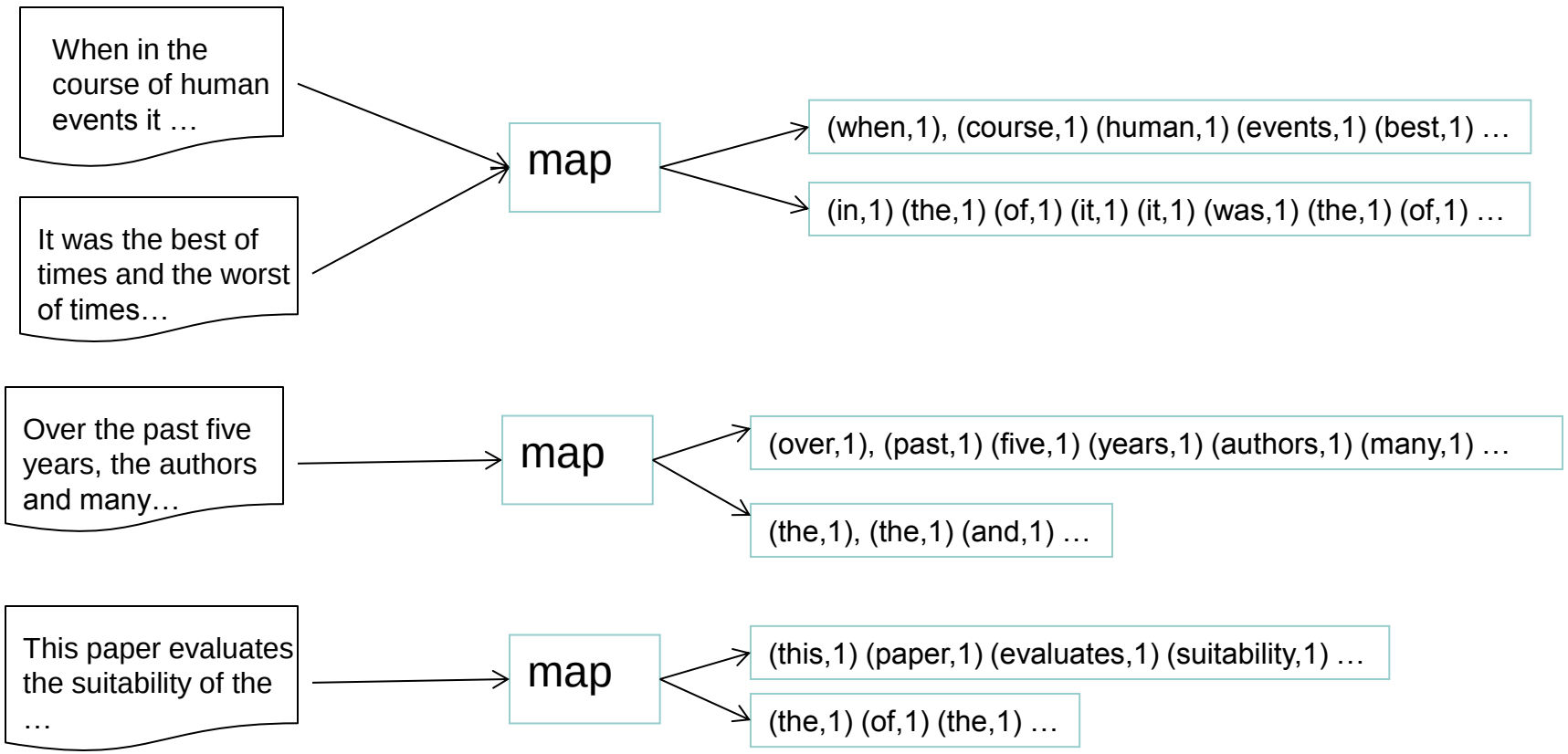
MapReduce: El paso Map



MapReduce: El paso Reduce

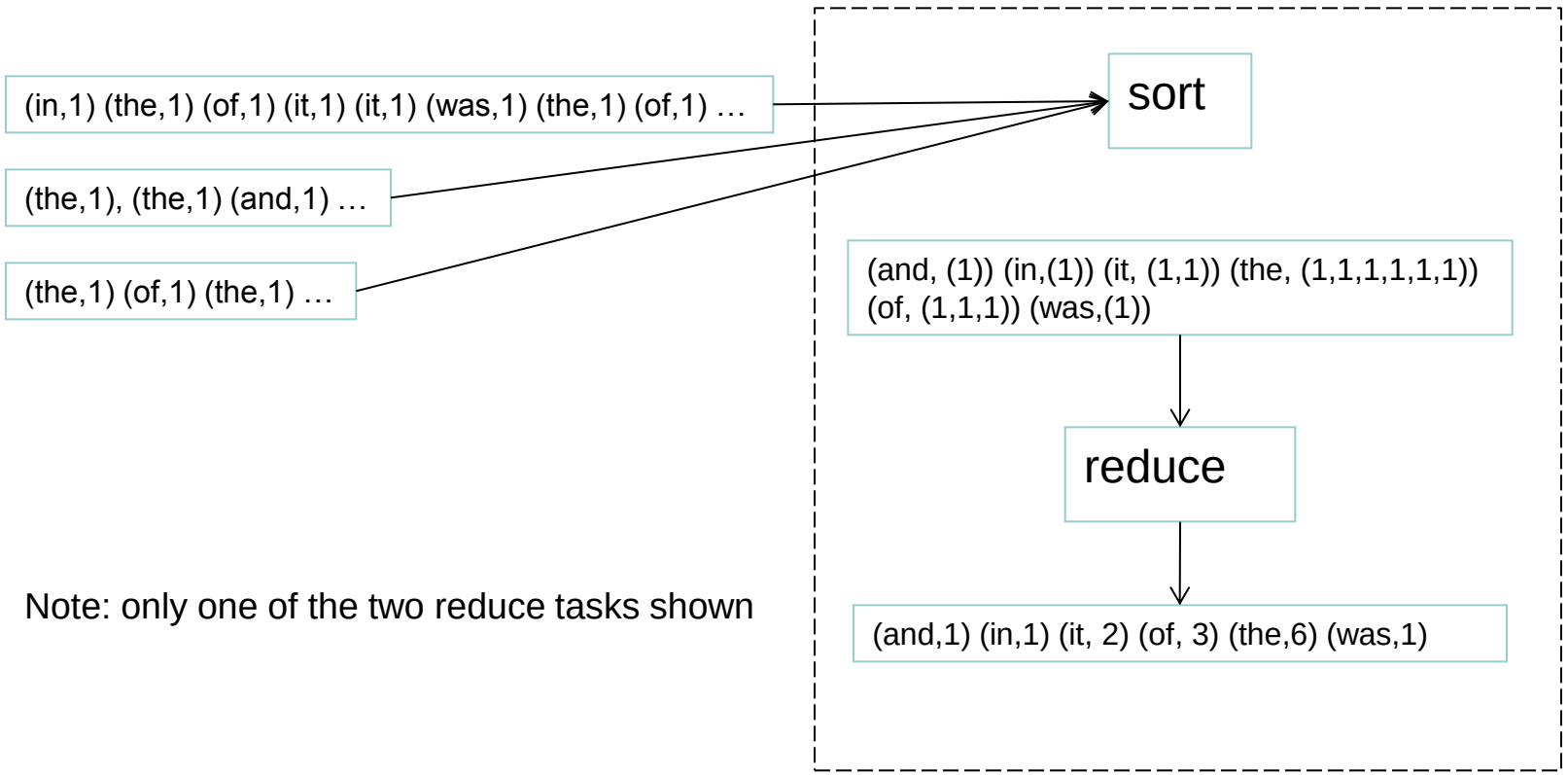


El Map (Ejemplo: Wordcount)



El Reduce (Ejemplo:Wordcount)

partition (intermediate files) (R=2)



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 Sandiego 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”).

Mahout

Una libreria para realizar escalable Machine learning que corre en Hadoop

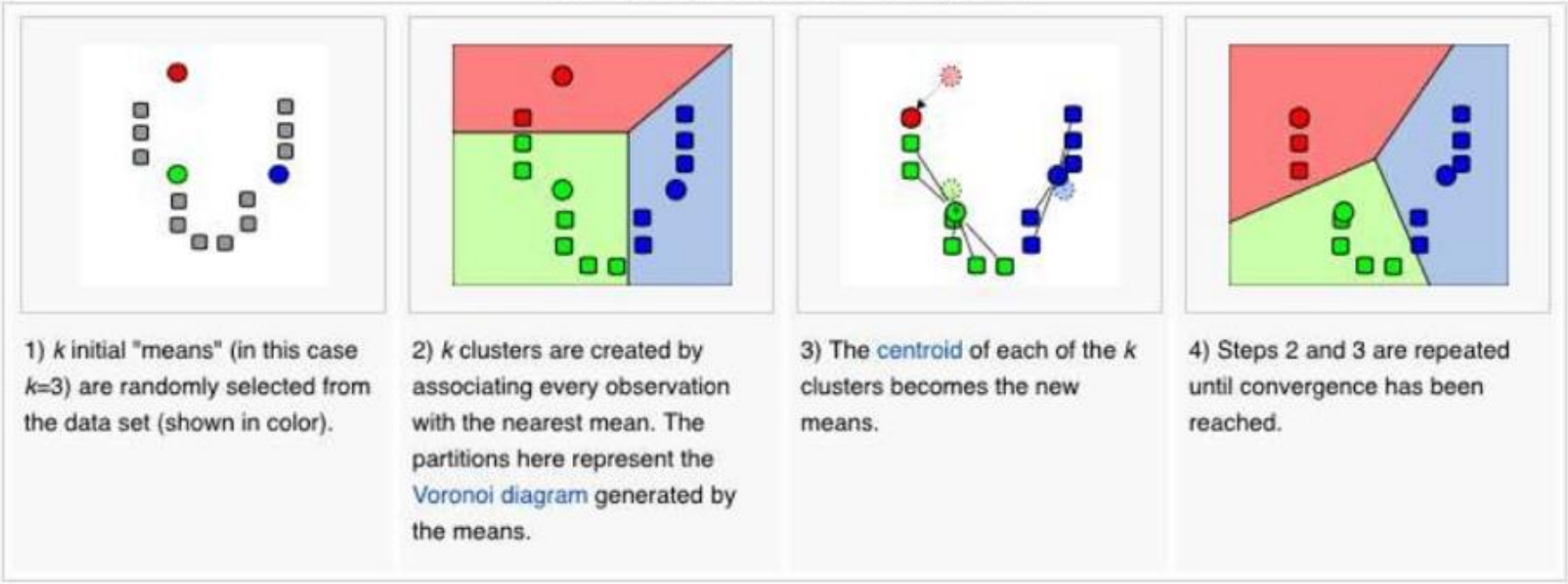
Incluye algunos algoritmos para:

Clustering
Clasificacion
Busqueda de itemsets frecuentes
Sistemas de recomendaciones a consumidores.

Algunas personas piensan que los algoritmos de Mahaout no han sido programados de manera optima.

Ejemplo de Mineria de datos: Clustering usando k-means

Demonstration of the standard algorithm



1) *k* initial "means" (in this case *k*=3) are randomly selected from the data set (shown in color).

2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3) The centroid of each of the *k* clusters becomes the new means.

4) Steps 2 and 3 are repeated until convergence has been reached.

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 forman 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 simulares ➔ Stop
    - Else
      - Si se ha alcanzado el max iterations ➔ Stop
      - Else ➔ Empezar otra iteracion Map-Reduce

K-Means Optimizationes

- **Uso de Combiners**
  - Similar al reducer
  - Calcula para cada centroide las sumas (y conteos) locales de los puntos asignados
  - Envia al reducer <centroide, <sumas parciales>>

- **Uso de un solo Reducer**
  - Se usa cuando la cantidad de datos que se envia a los reducers es bien pequena.
  - Un solo reducer puede decir si algunos de los centros ha cambiado o no
  - Crea un solo archivo de salida.

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 cuals corresponde a un posting.
Primeor 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

Investigacion 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
- Uso de Twister para mejorar el rendimiento de Hadoop en algoritmos iterativos.