Tools for Mining Massive Datasets

Dr. Edgar Acuna
Departament of Mathematical Science
University of Puerto Rico-Mayaguez

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

Website: academic.uprm.edu/eacuna

In this talk, we have used several slides from the Salsahadoop group, Prof. Leskovec at Stanford U. and Prof Jermaine at Rice U.

Outline

I. Data Mining
II. Big Data
III. Hadoop and MapReduce
IV. Kmeans
V. Proposed Research

Data Mining [1]: Motivation

The mechanisms for automatic recollection of data and the development of databases technology has made possible that a large amount of data can be available in databases, data warehouses and other information repositories.

Nowdays, there is the need to convert this data in knowledge and information.

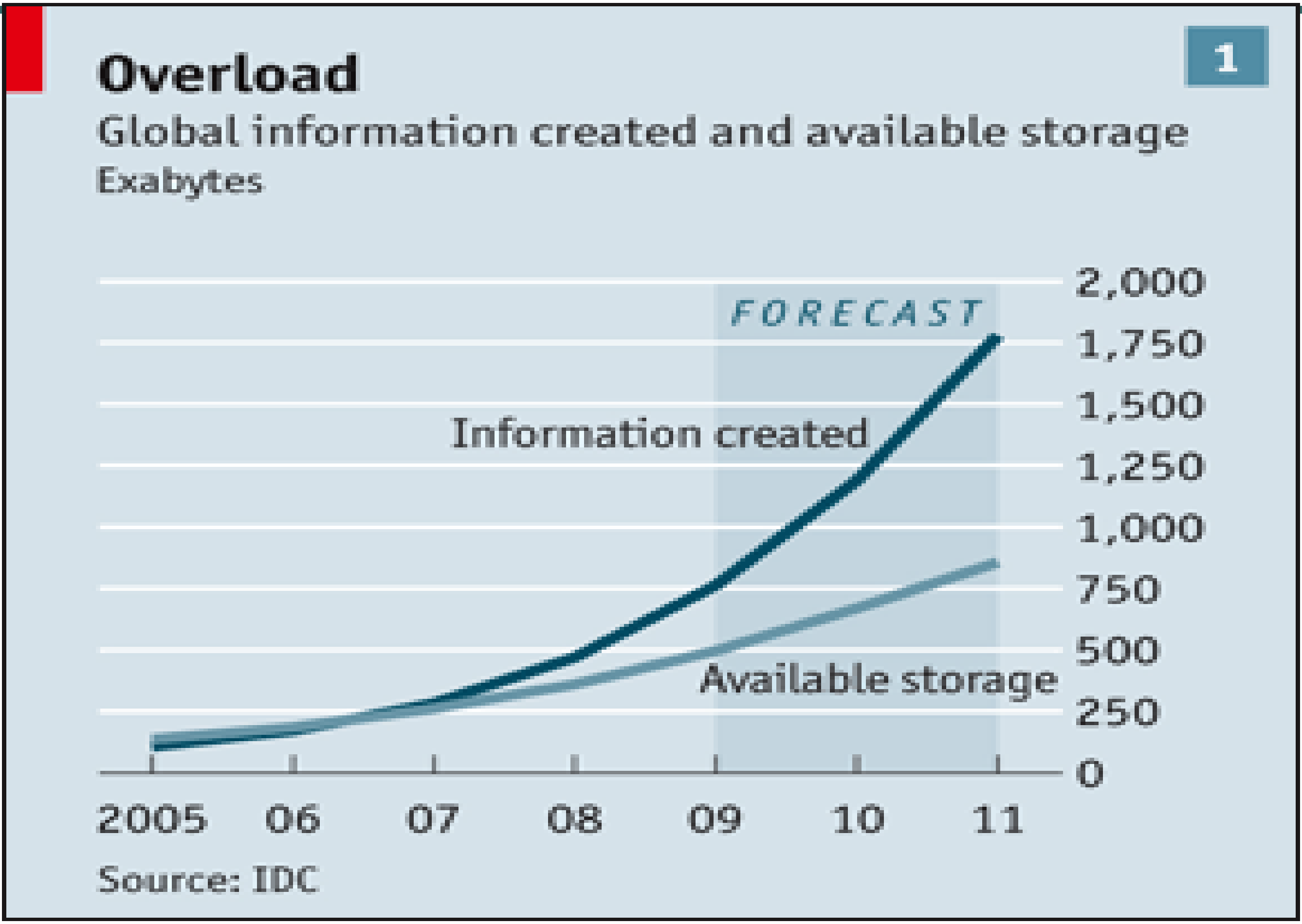
Data Mining[3]?What is Data Mining?

- It is the process of extracting valid knowledge/information from a very large dataset. The knowledge is given as patterns and rules that are non-trivial, previously unknown, understandable and with a high potential to be useful.
- Other names: Knowledge discovery in databases (KDD), Intelligent Data Analysis, Business Intelligence.
- The first paper in Data Mining appeared in 1993.

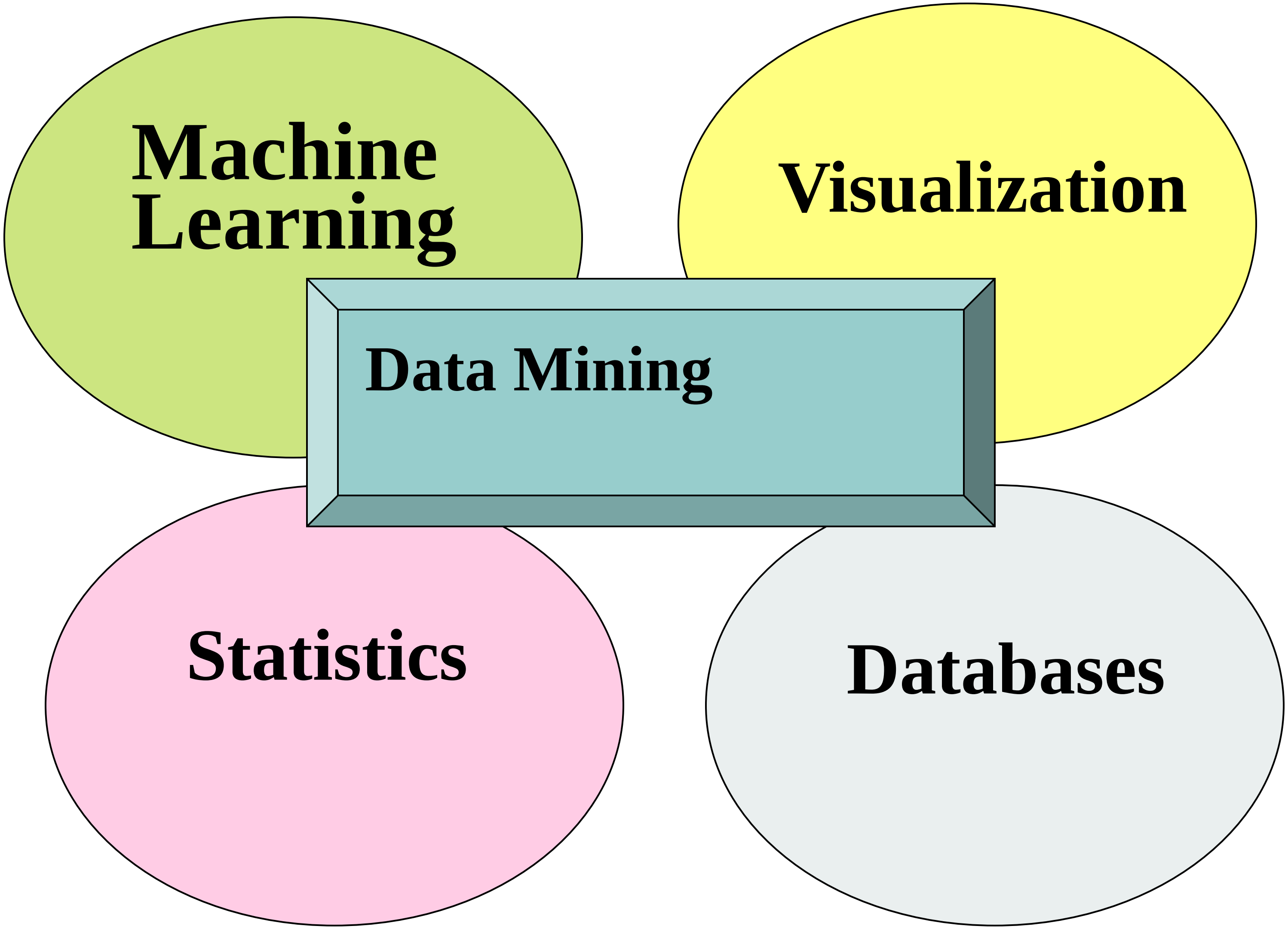
Data Mining[3]: Size (in Bytes) of datasets

Description	Size	Storage Media
Very small	10^2	Piece of papel
Small	10^4	Several sheets of paper
Medium	10^6 (megabyte)	Floppy Disk
Large	10^9(gigabite)	USB/Hard Disk
Massive	10^{12}(Terabyte)	Hard disk/USB
Super-massive	10^{15}(Petabyte)	File of distributed data
Exabyte(10^{18}), Zettabytes(10^{21}), Yottabytes(10^{24})		

The economist, February 2010



Data Mining[5]: Related Areas



Contribution of of each area to Data MIning

- Statistics (~35%): Estimation of prediction models. Assume distribution for the features used in the model. Use of sampling
- Machine learning: (~30 %): Part of Artificial Intelligence. More heuristic than Statistics. Small data and complex models
- Databases: (~25%):large Scale data, simple queries. The data is maintaned in tables that are accesed quickly.
- Visualization: (~ 5%).It can be used in either the pre-processing o post-processing step of the KDD process.
- Other Areas: (~5%): Pattern Recognition, Expert Systems, High Performance Computing.

Data Mining Applications

Science: Astronomy, Bioinformatics (Genomics, Proteonomics, Metabolomics), drug discovery.

Business: Marketing, credit risk, Security and Fraud detection,

Govermment: detection of tax cheaters, anti-terrorism.

Text Mining:

Discover distinct groups of potential buyers according to a user text based profile. Draw information from different written sources (e-mails).

Web mining: Identifying groups of competitors web pages. Recomemder systems(Netflix, Amazon.com)

Data Mining: Type of tasks

- Descriptive: General properties of the database are determined. The most important features of the databases are discovered.

- Predictive: The collected data is used to train a model for making future predictions. Never is 100% accurate and the most important matter is the performance of the model when is applied to future data.

Data Mining: Tasks

- Regression (Predictive)
- Classification (Predictive)
- Unsupervised Classification–Clustering (Descriptive)
- Association Rules (Descriptive)
- Outlier Detection (Descriptive)
- Visualization (Descriptive)

Regression

- The value of a response continuous variable is predicted based on the values of other variables called predictors assuming that there is a functional relationship among them
- It can be done using statistical models, decision trees, neural networks, etc.
- Example: car sales of dealers based on the experience of the sellers, advertisement, type of cars, etc..

Supervised Classification[1]

- The response variable is categorical.

Given a set of records, called the training set (each record contains a set of attributes and usually the last one is the class), a model for the attribute class as a function of the others attributes is constructed. The model is called the **classifier**.

- Goal: Assign records previously unseen (test set) to a class as accurately as possible.
- Usually a given data set is divided in a training set and a test set. The first data set is used to construct the model and the second one is used to validate. The precision of the model is determined in the test data set.
- It is a decision process.

Examples of classifiers

Linear Discriminant Analysis (LDA),

Naïve Bayes,

Logistic Regression,

k-nearest neighbor,

Decision trees,

Bayesian Networks

Neural Networks

Support vector machine (SVM)

………………..

Unsupervised Classification (Clustering)[1]

- Find out groups of objects (clusters) such as the objects within the same clustering are quite similar among them whereas objects in distinct groups are not similar.
- A similarity measure is needed to establish whether two objects belong to the same cluster or to distinct cluster.

- Examples of similarity measure: Euclidean distance, Manhattan distance, correlation, Grower distance, Hamming distance, Jaccard Distance, etc.
- Problems: Choice of the similarity measure, choice of the number of clusters, cluster validation.

Clustering Algorithms

- Partitioning algorithms: K-means, PAM, SOM.
- Hierarchical algorithms: Agglomerative, Divisive.
- Gaussian Mixtures Models
- DBSCAN
- BIRCH

Association rules[1]

- Given a set of records each of which contain some number of items from a given collection.
- The aim is to find out dependency rules which will predict occurrence of an item based on occurrences of other items

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules discovered:

{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Association rules[2]

- The rules (X->Y) must satisfy a minimum support and confidence set up by the user.

X is called the antecedent and Y is called the consequent.

- Support=(# records containing X and Y)/(# records)
- Confidence=(# records containing X and Y/(# de records containing X)

Example: The first rule has support .6 and the second rule has support .4.The confidence of rule 1 is .75 and for the rule 2 is .67

Applications: Marketing and sales promotion.

Algorithms: A-priori, PCY, SON, FP-Growth

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data.
- Quality of Data
- Privacy Data
- Streaming Data

Data Mining[3]:Software

- **Open source**
- R (cran.r-project.org). Related to Statistics (37.4% de usuarios, Kdnuggets 2013).
- Weka (http://www.cs.waikato.ac.nz/ml/weka/) (14.3%): Related to Machine Learning. Written in Java.
- RapidMiner (http://rapidminer.com). (39.2%) Related to the Database community.
- **Comercial:** SAS Enterprise Miner SPSS Clementine Statistica (22%) XLMiner(28%), Microsoft SQL(8%), Oracle.

Big Data [1]: Definition

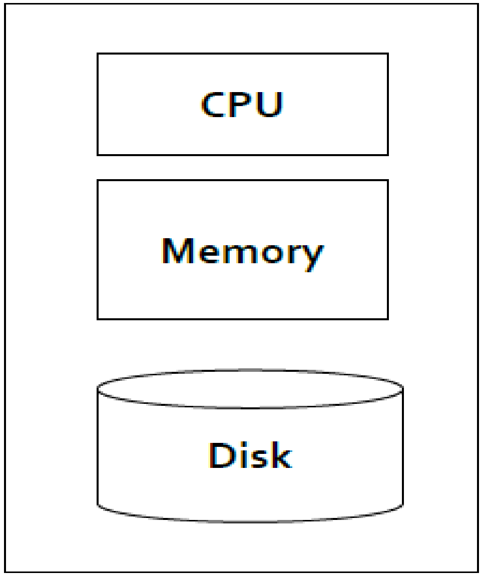
In a 2001, Doug Laney,analyst for the Gartner Group, defined data challenges and opportunities as being three-dimensional: increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources).

In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."
Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are constantly changing. In 2002, maybe 100GB, in 2012 perhaps from 10TB to many petabytes of data in a single data set.

Big Data[2]: Examples

- The Large Hadron Collider (LCH) storages around 25 Petabytes of sensor data per year.
- In 2010, the ATT's database of calling records was of 323 Terabytes.
- El 2010, Walmart handled 2.5 Petabytes of transactions hourly.
- In 2009, there was 500 exabytes of information on the internet.
- In 2011, Google searched in more than 20 billions of web pages. This represents aprox. 400 TB.
- In 2013, it was announced that the NSA's Data Center at Utah will storage up to 5 zettabytes (5,000 exabytes).

Single Node Architecture



Machine Learning, Statistics

“Classical” Data Mining

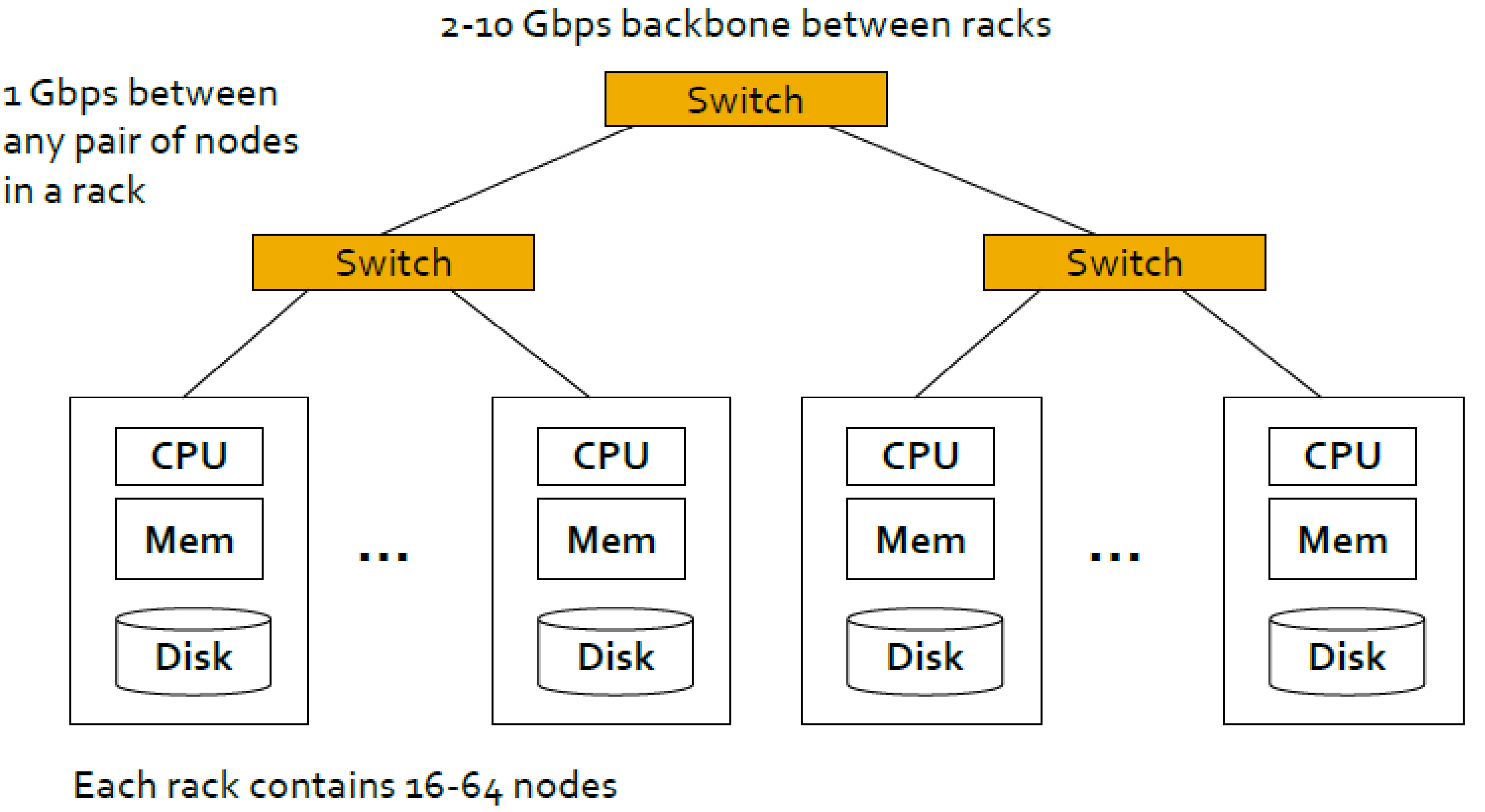
Motivation: Google Example

Google searches in more than 20 billion of webpages x20KB= 400+ TB. One computer reads with a speed of 30-35 MB/sec from disk. It will need aprox 4 months to read the web.

It will be necessary aprox 1000 hard drives to read the web
It will be necessary even more than that to analyze the data

Today a standard architecture for such problems is being used. It consists of
-a cluster of commodity Linux nodes
-commodity network (ethernet) to connect them

Cluster Architecture



2-10 Gbps backbone between racks

1 Gbps between
any pair of nodes
in a rack

Each rack contains 16-64 nodes

Challenges in large-scale computing for data mining

How to distribute the computation?
How to write down distributed program easily?
Machines fail!.
One computer may stay up three years (1000 days)
If you have 1000 servers , expect to loose 1 per day
In 2011, it was estimated that Google has 1 million of computers, so 1000 servers can fail everyday

What is Hadoop?

- In 2004, J. Dean and S. Ghemawhat wrote a paper explaining Google's MapReduce, a programming model and a associated infrastucture for storage of large data sets (file system) called Google File System (GFS).
- GFS is not open source.
- In 2006, Doug Cutting at Yahoo! , created a open source GFS and called it Hadoop Distributed File System (HDFS). In 2009, eh left to Cloudera.
- The software framework that supports HDFS, MapReduce and other related entities is called the project Hadoop or simply Hadoop.
- Hadoop is distributed by the Apache Software Foundation.

Hadoop

Hadoop includes:

Distributed Files System(HDFS) –distributes data
Map/Reduce-distributes application

It is written in Java

Runs on
-Linux, MacOS/X, Windows, and Solaris
-Uses commodity hardware

Distributed File System

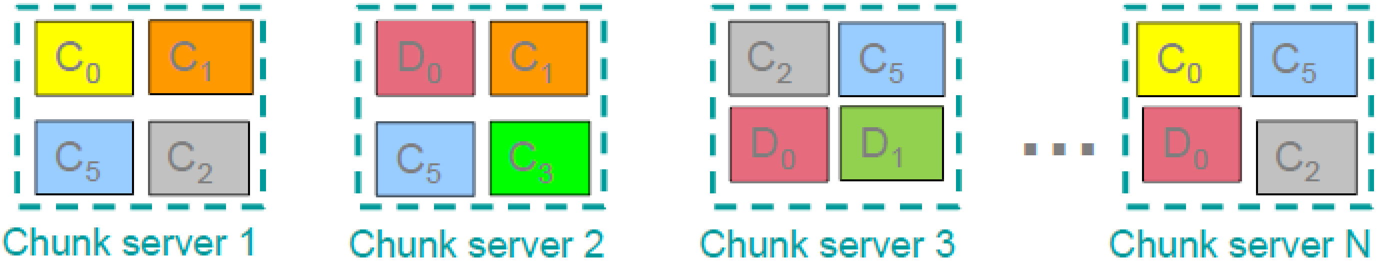
Chunk servers:
File is split into contiguous chunk
Typically each chunk is 16-128 MB
Each chunk is replicated (usually 2x and 3x)

The master node:
(It is called Namenode in Hadoop) stores metadata about where files are stored

The client library for file access.
This library talks to find chunk servers.
Connects directly to chunk servers to access data

Distributed File System

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



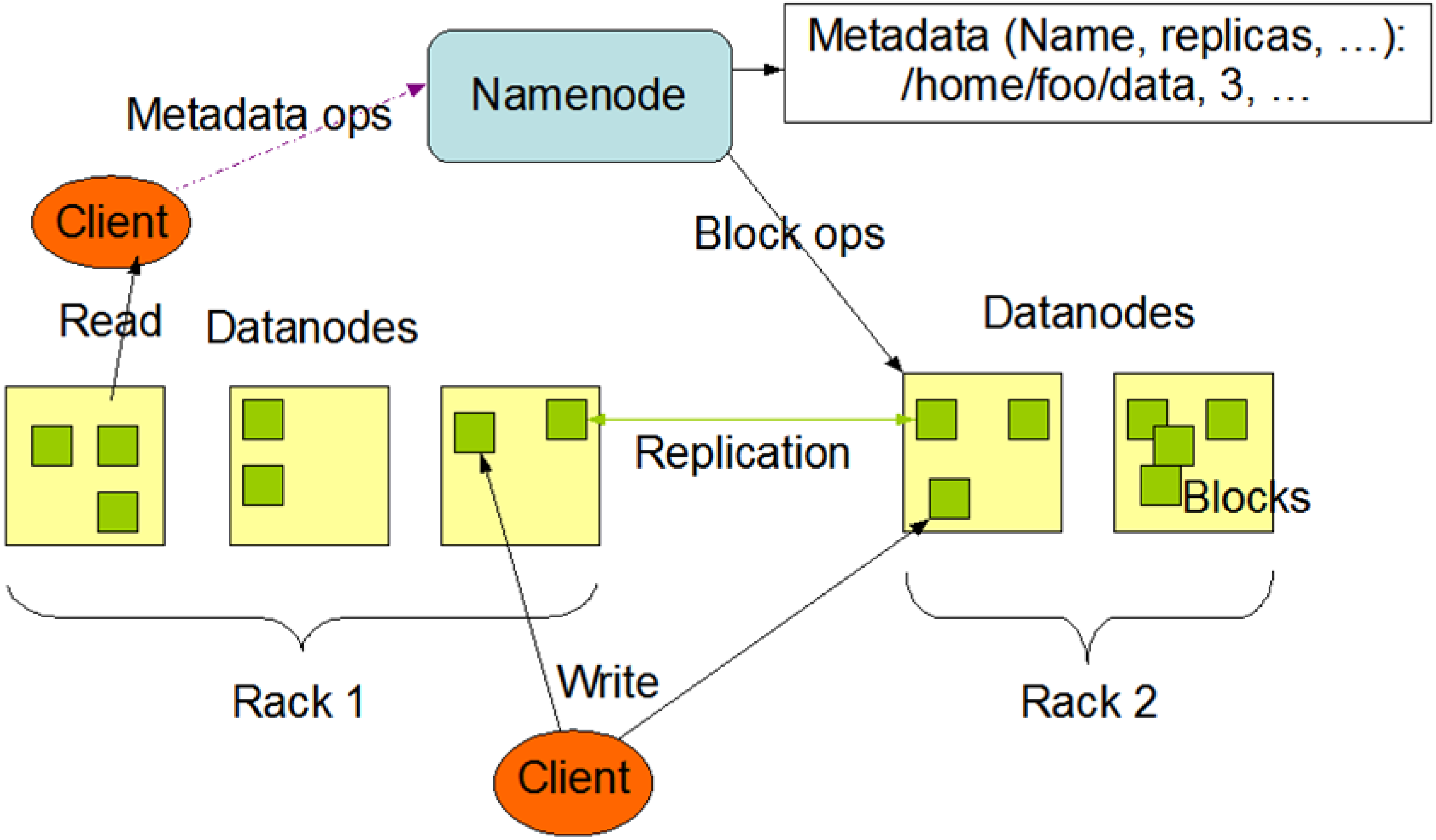
Bring computation directly to the data!

Chunk servers also serve as compute servers

HDFS Architecture

- Provides
  - Automatic Parallelization and Distribution
  - Fault Tolerance
  - I/O Scheduling
  - Monitoring and Status Updates

HDFS Architecture



MapReduce

Map-Reduce is a programming model for efficient distributed computing

It works like a Unix pipeline:

-cat imput |grep | sort |unique –c |cat > ouput

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

Efficient because reduces seeks and the use of pipeline

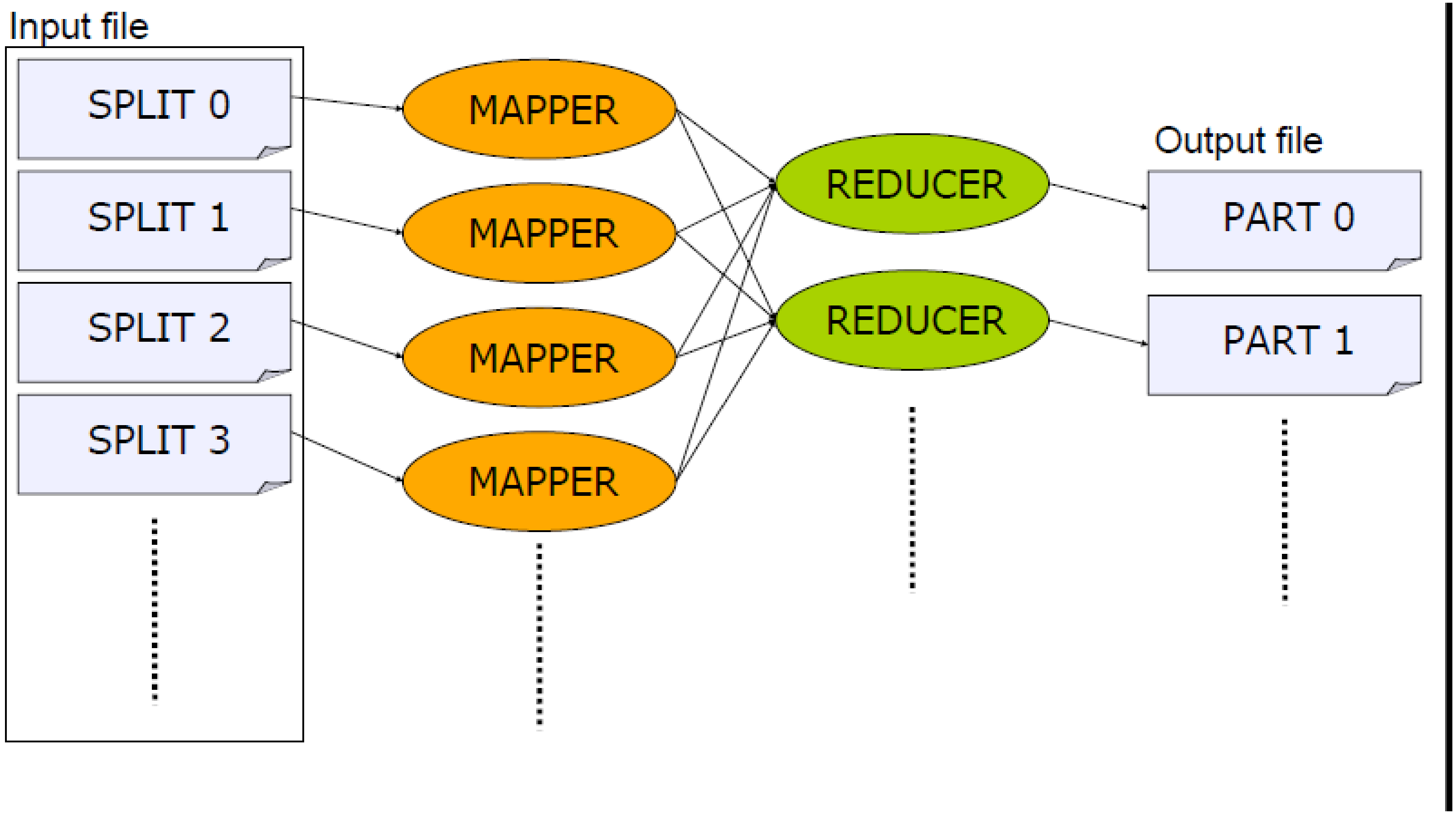
MapReduce Model

- Input & Output: a set of key/value pairs
- Two primitive operations
  - **map**: $(k_1, v_1) \rightarrow \text{list}(k_2, v_2)$
  - **reduce**: $(k_2, \text{list}(v_2)) \rightarrow \text{list}(k_3, v_3)$
- Each map operation processes one input key/value pair and produces a set of key/value pairs
- Each reduce operation
  - Merges all intermediate values (produced by map ops) for a particular key
  - Produce final key/value pairs

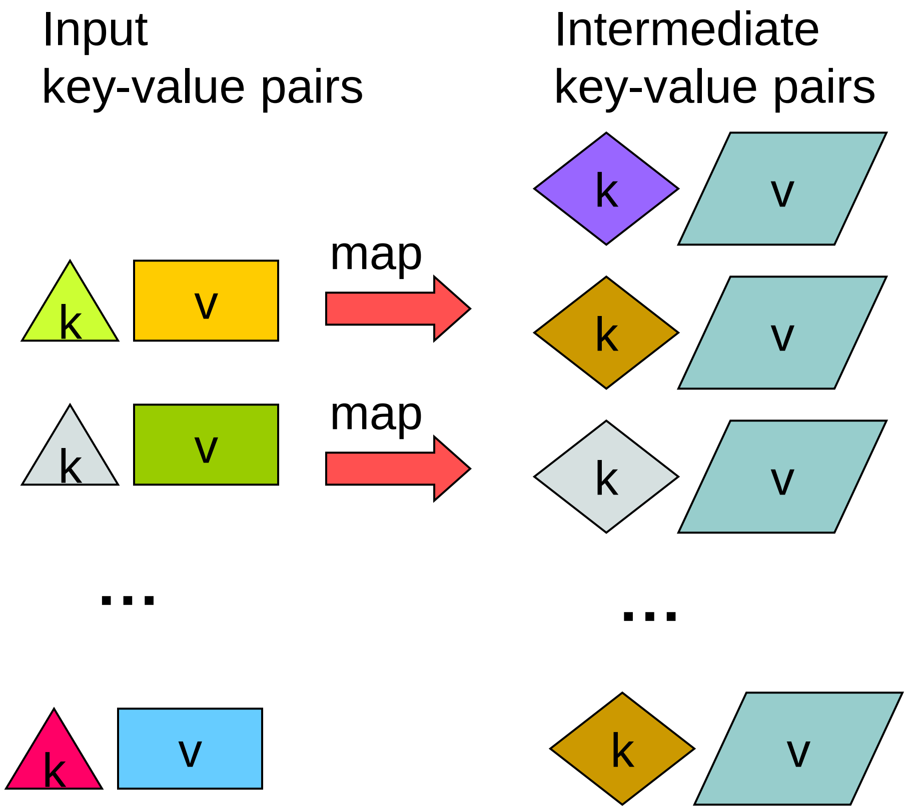
The WorkFlow

- Load data into the Cluster (HDFS writes)
- Analyze the data (MapReduce)
- Store results in the Cluster (HDFS)
- Read the results from the Cluster (HDFS reads)

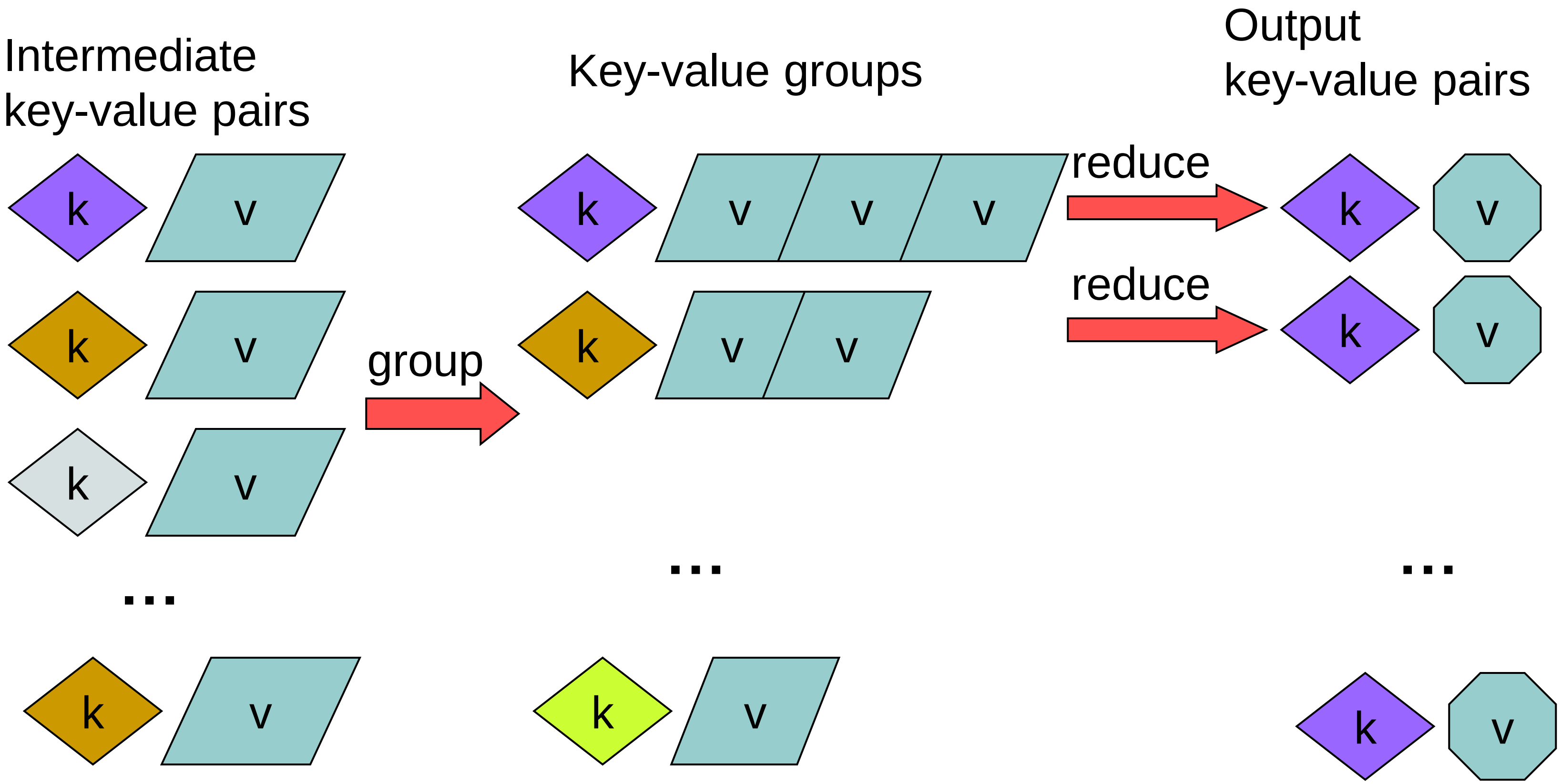
Execution model:Flow



MapReduce: The Map Step

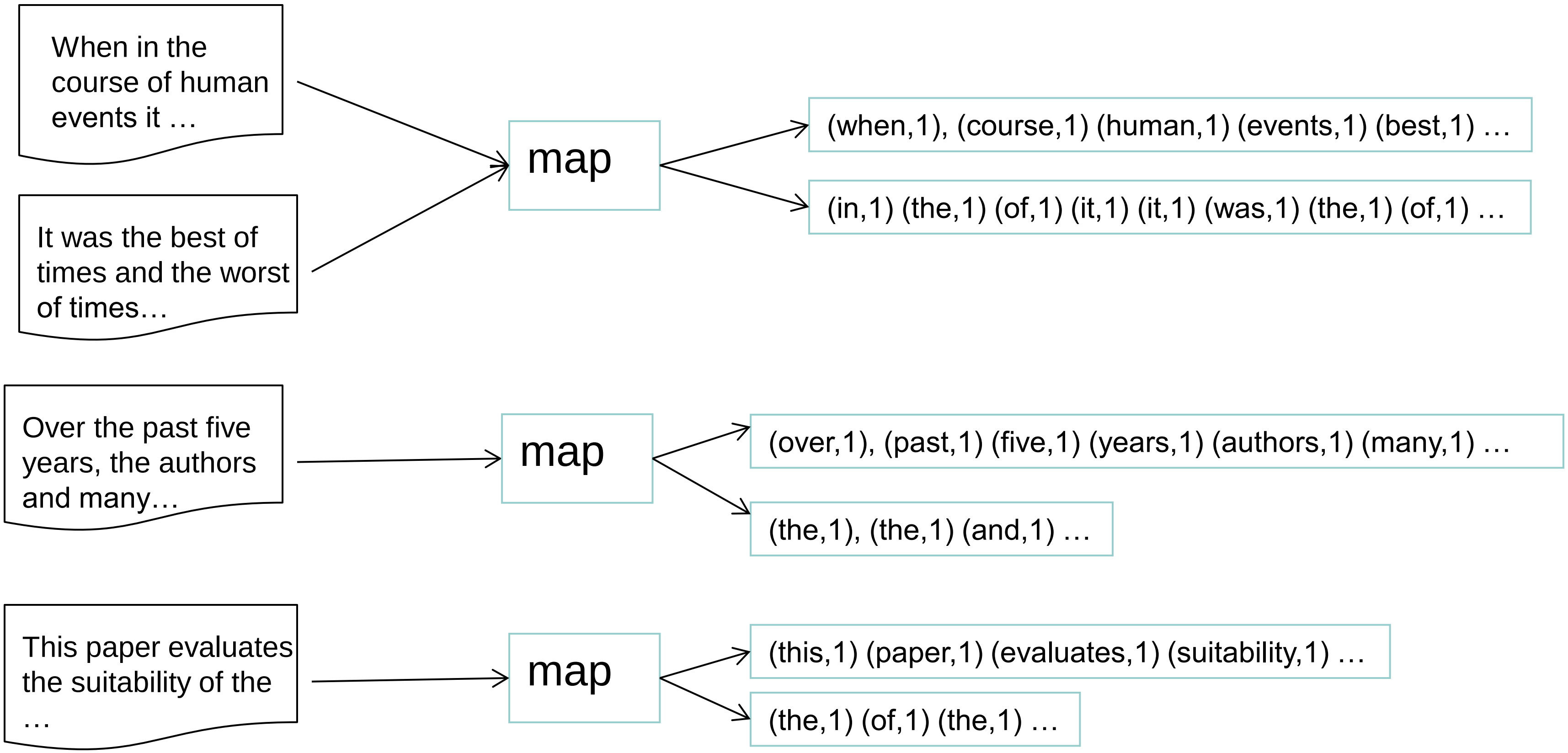


MapReduce: The Reduce Step



The Map (Example: Wordcount)

inputs **tasks (M=3)** **partitions (intermediate files) (R=2)**



When in the course of human events it …

It was the best of times and the worst of times…

map

(when,1), (course,1) (human,1) (events,1) (best,1) …

(in,1) (the,1) (of,1) (it,1) (it,1) (was,1) (the,1) (of,1) …

Over the past five years, the authors and many…

map

(over,1), (past,1) (five,1) (years,1) (authors,1) (many,1) …

(the,1), (the,1) (and,1) …

This paper evaluates the suitability of the …

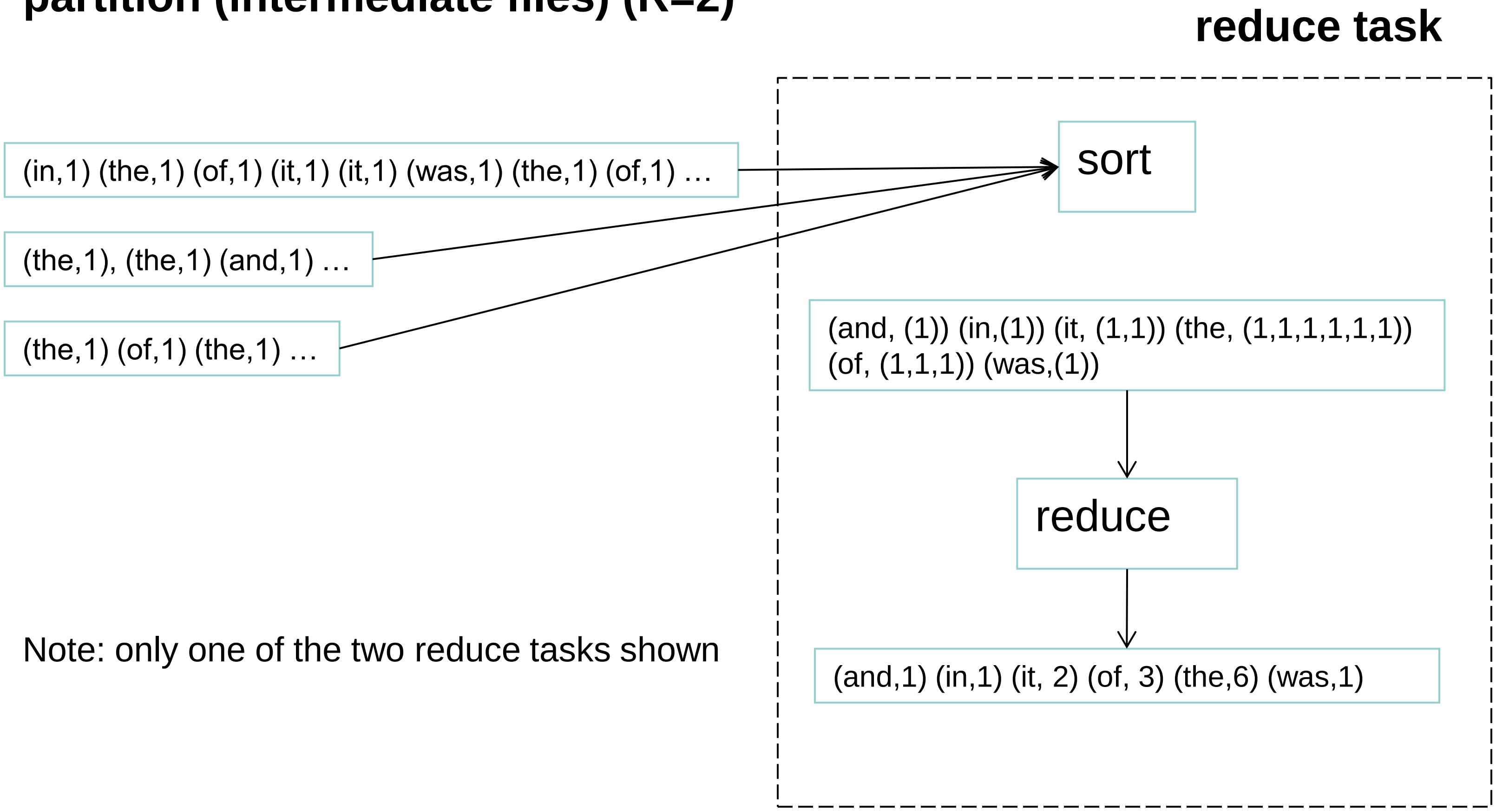
map

(this,1) (paper,1) (evaluates,1) (suitability,1) …

(the,1) (of,1) (the,1) …

The Reduce (Example:Wordcount)

partition (intermediate files) (R=2)



reduce task

(in,1) (the,1) (of,1) (it,1) (it,1) (was,1) (the,1) (of,1) …

(the,1), (the,1) (and,1) …

(the,1) (of,1) (the,1) …

sort

(and, (1)) (in,(1)) (it, (1,1)) (the, (1,1,1,1,1,1)) (of, (1,1,1)) (was,(1))

reduce

(and,1) (in,1) (it, 2) (of, 3) (the,6) (was,1)

Note: only one of the two reduce tasks shown

Hadoop Non-Java Interfaces

Hadoop streaming: C++,Python, perl,ruby,
Rhadoop (R and Hadoop),Weka(Mark Hall is working on that),Radoop (rapidminer and hadoop, comercial)
Hadoop Pipes: (C++) It not recommendable

Where can I run hadoop?

In your personal computer using the single node hadoop cluster. If you have windows install a virtual machine for running Ubuntu(See Michael Noll's website)

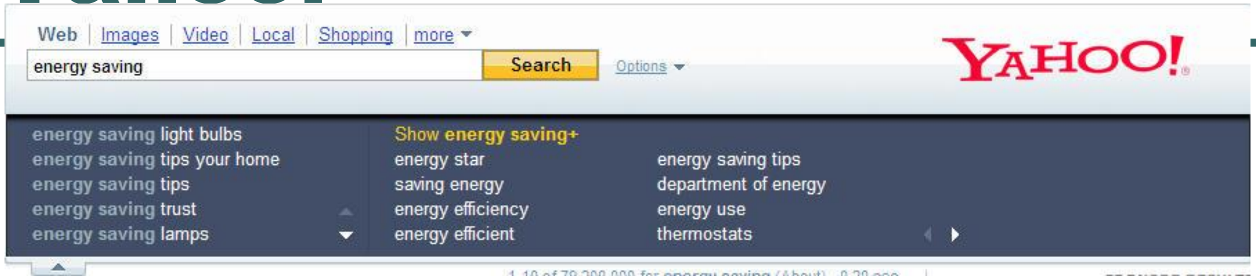
Free:
At the Gordon cluster of the SDSC (1024 nodes).
In several clusters (India(128 nodes), Sierra(84 nodes) and Alamo of the Futuregrid Project led by Indiana University

Non-Free, but too expensive
Amazon Elastic Compute Cloud (EC2)

Who is using Hadoop?

Yahoo
Facebook
Amazon
Google
IBM
Netflix
Ebay
LinkedIn
Twitter

Example of Performance:Hadoop in Yahoo!



energy saving

energy saving light bulbs
energy saving tips your home
energy saving tips
energy saving trust
energy saving lamps

Show energy saving+

energy star
saving energy
energy efficiency
energy efficient

energy saving tips
department of energy
energy use
thermostats

- Database for **Search Assist™** is built using Hadoop.
- 3 years of log-data

	Before Hadoop	After Hadoop
Time	26 days	20 minutes
Language	C++	Python
Development Time	2-3 weeks	2-3 days

Writing a MapReduce program

A Hadoop program needs:

-A java clas with a main that configures and submit the job

- A class that extends the Hadoop Mapper class ("The Mapper")
- A class that extends the Hadoop Reducer class("The Reduce")
- Optional: A class that extends Hadoop Reducer class ("The Combiner")

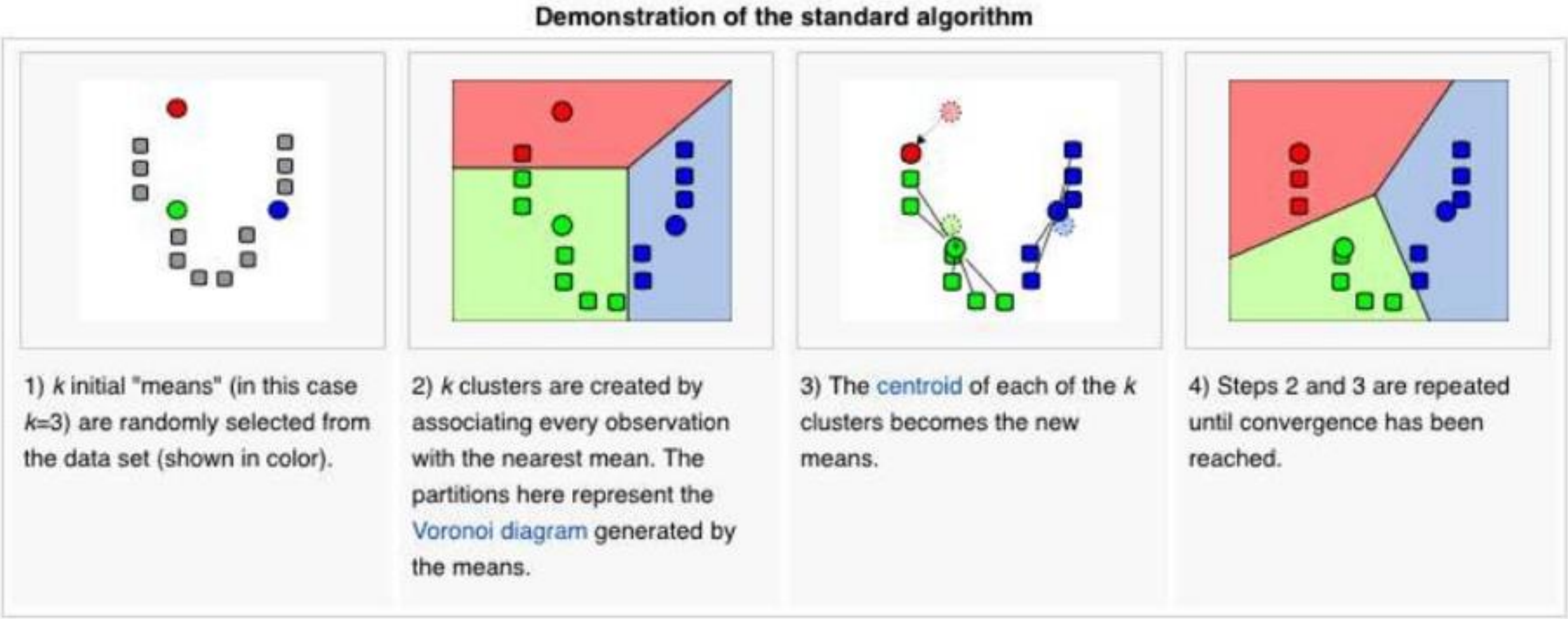
Mahout

Scalable Machine learning library that runs on Hadoop
Includes some algorithms for:

Recommendation mining
Clustering
Classification
Finding frequent itemsets

Some people thinks that the algorithms included in Mahaiut are not programmed in a optimal way.

Data Mining Example:K-Means Algorithm



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

K-Means Algorithm

- **Step 1:** Select K points at random (Centers)
- **Step 2**: For each data point, assign it to the closest center
  - Now we formed K clusters
- **Step 3:** For each cluster, re-compute the centers
- **Step 4:** If the new centers are different from the old centers (previous iteration) ➔ Go to Step 2

K-Means in MapReduce

- **Input**
  - Dataset --Large
  - Initial centroids (K points) --Small

- **Map phase**
  - Each map reads the K-centroids + one block from dataset
  - Assign each point to the closest centroid
  - Output <centroid, point>

K-Means in MapReduce (Cont’d)

- **Reduce Phase**
  - Gets all points for a given centroid
  - Re-compute a new centroid for this cluster
  - Output: <new centroid>

- **Iteration Control**
  - Compare the old and new set of K-centroids
    - If similar ➔ Stop
    - Else
      - If max iterations has reached ➔ Stop
      - Else ➔ Start another Map-Reduce Iteration

K-Means Optimizations

- **Use of Combiners**
  - Similar to the reducer
  - Computes for each centroid the local sums (and counts) of the assigned points
  - Sends to the reducer <centroid, <partial sums>>

- **Use of Single Reducer**
  - Amount of data to reducers is very small
  - Single reducer can tell whether any of the centers has changed or not
  - Creates a single output file

Kmeans applied to 20Newsgroup dataset (Jermaine, 2012)

This is a set of around 20,000 postings to 20 different newsgroups (these were sort of like the blogs of the 1990s). The "20_newsgroups" directory will have 20 subdirectories, each corresponding to a newsgroup. There are around 1000 files in each subdirectory, each of which is a posting.
First the documents are converted in a bag-words vectors. We will use a dictionary of 1000 words and 20000 documents(1000 for each of the 20 groups).

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

Current research

Avoiding the GIGO(Garbage in-Garbage Out)
Several preprocessing techniques will be implemented in MapReduce
The ReliefF feature selection method (Carlos Gomez)
Outlier Detection
Reduction of dimensionality
Treatment of Missing Values

Extending R using Hadoop

Twister: Improves performance of Hadoop in iterative algorithms