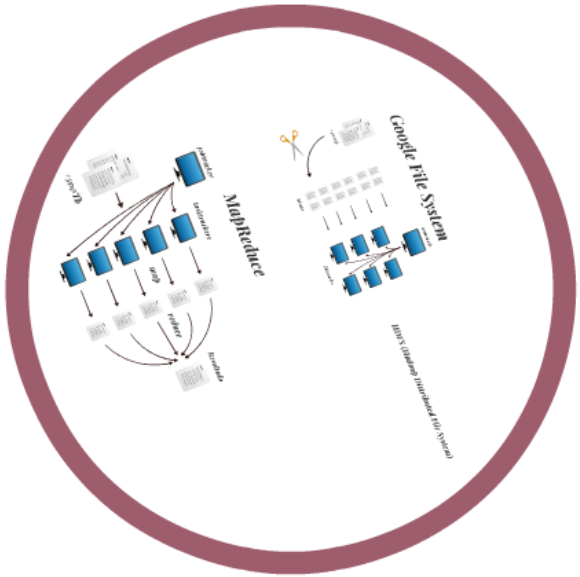
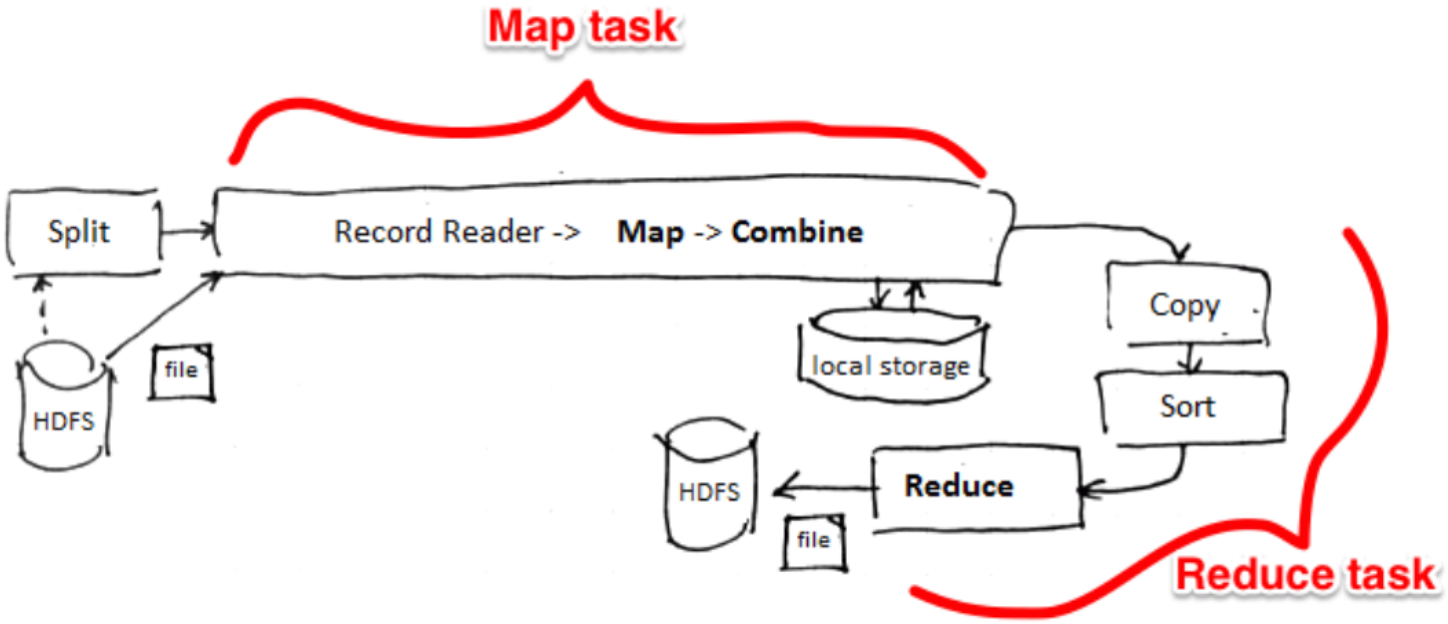
Introducción



HDFS y MapReduce

Ejemplo práctico

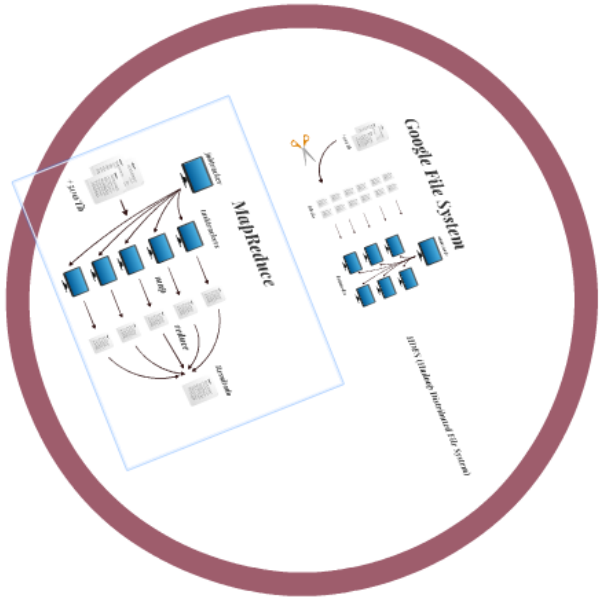
MapReduce, el algoritmo de Big Data



Alejandro Tejera Pérez

Jornadas Desmitificando Big Data - 2014

Introducción



HDFS y MapReduce

Ejemplo práctico

Introducción

Big Data

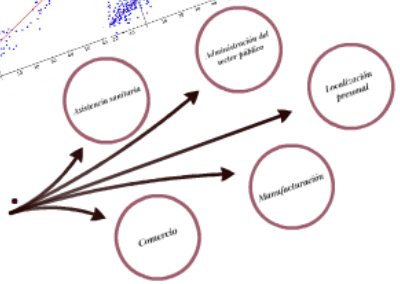
Big Data es información de gran volumen, gran velocidad de crecimiento, y gran variedad que requiere nuevas formas de procesar para explotar la toma de decisiones, la información oculta y los procesos de optimización.

Gartner Inc. 2001

Argumentos a favor

Fiabilidad

Áreas de utilización



Procesado de Big Data

Almacén

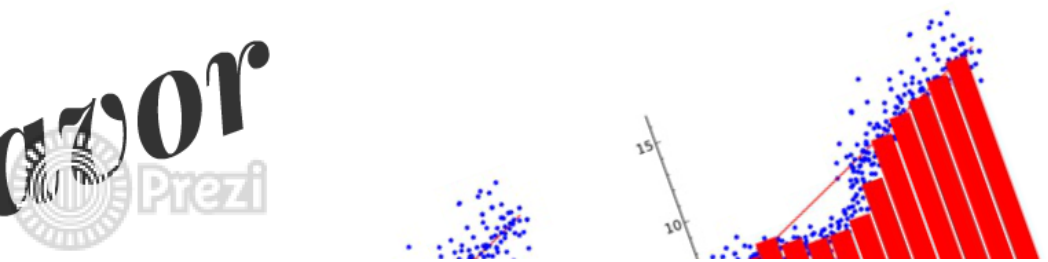
→ Google File System

Herramienta → MapReduce

Big Data

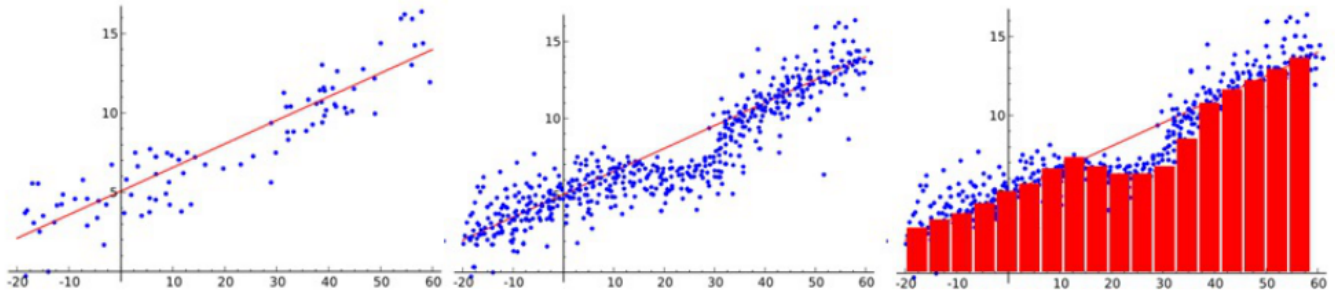
Big Data es información de gran volumen, gran velocidad de crecimiento, y gran variedad que requiere nuevas formas de procesar para explotar la toma de decisiones, la información oculta y los procesos de optimización.

Gartner Inc. 2001

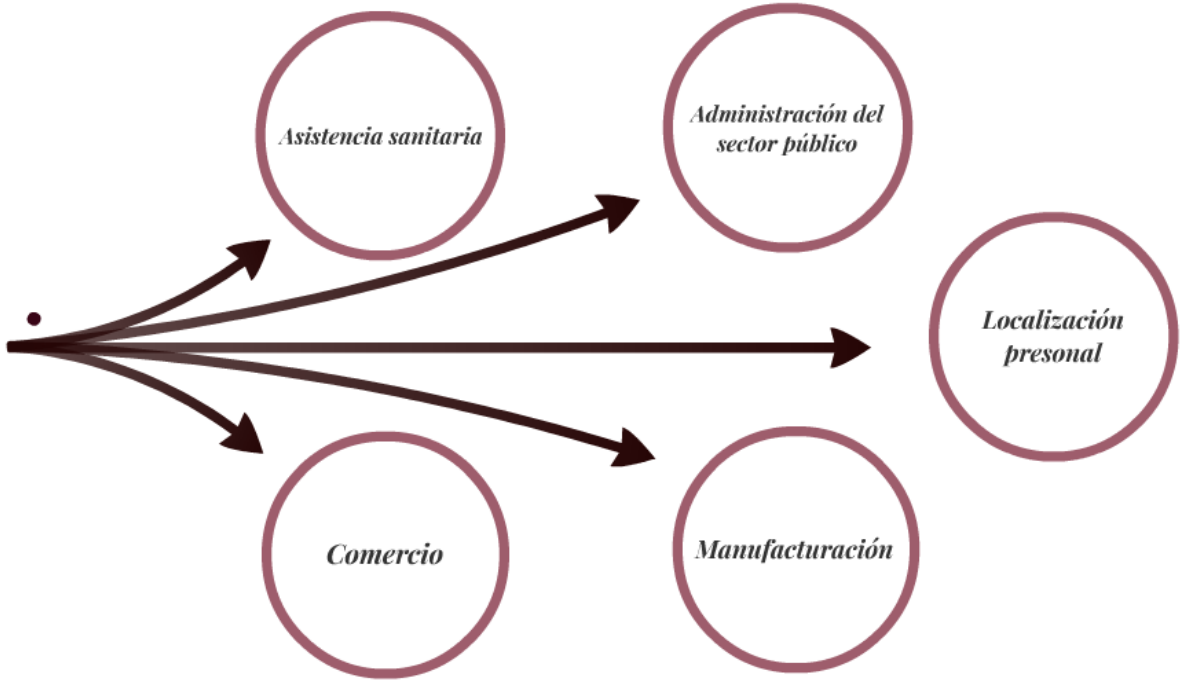


Argumentos a favor

Fiabilidad



Áreas de utilización



Procesado de Big Data

Almacén

+300Tb

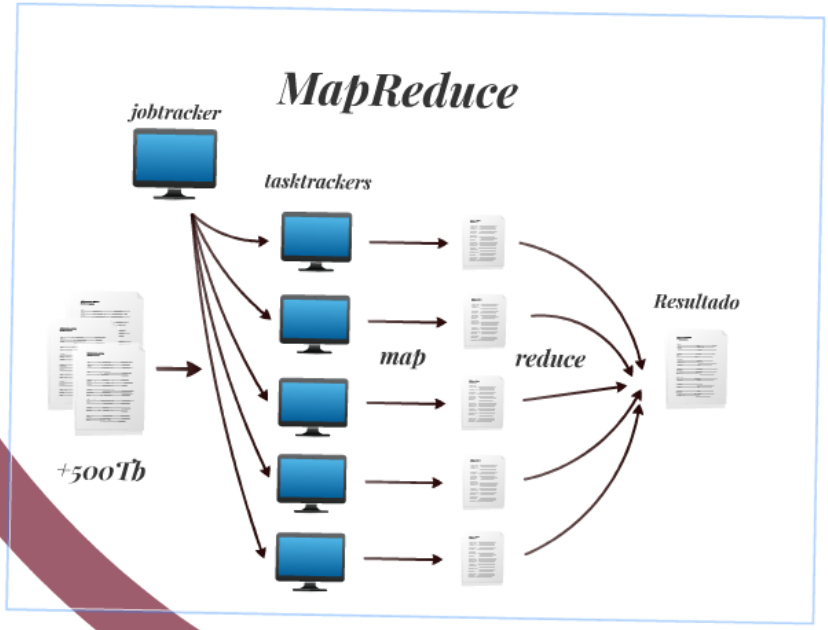
→ Google File System

Herramienta → **MapReduce**

Google File System

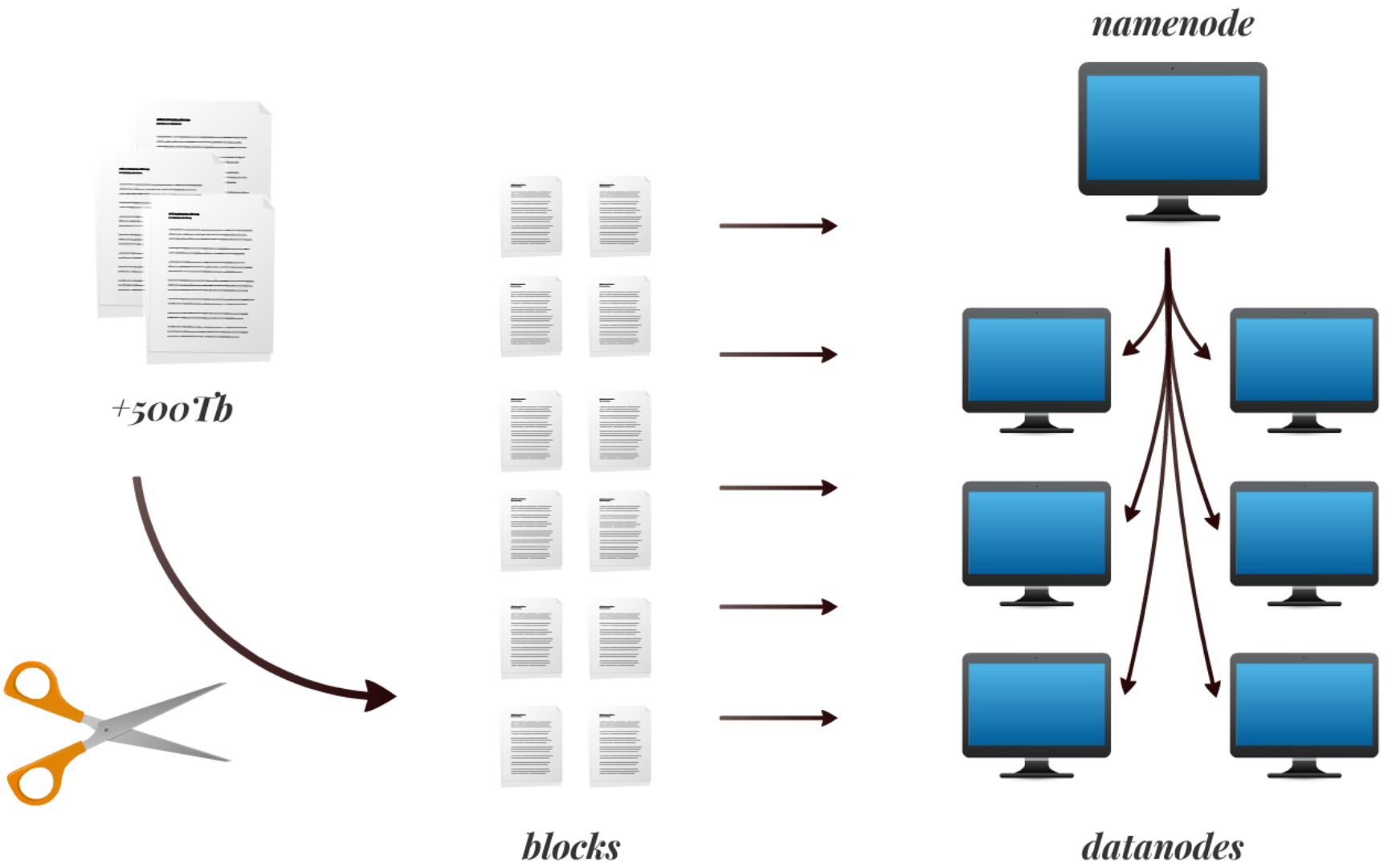
HDFS (Hadoop Distributed File System)

MapReduce



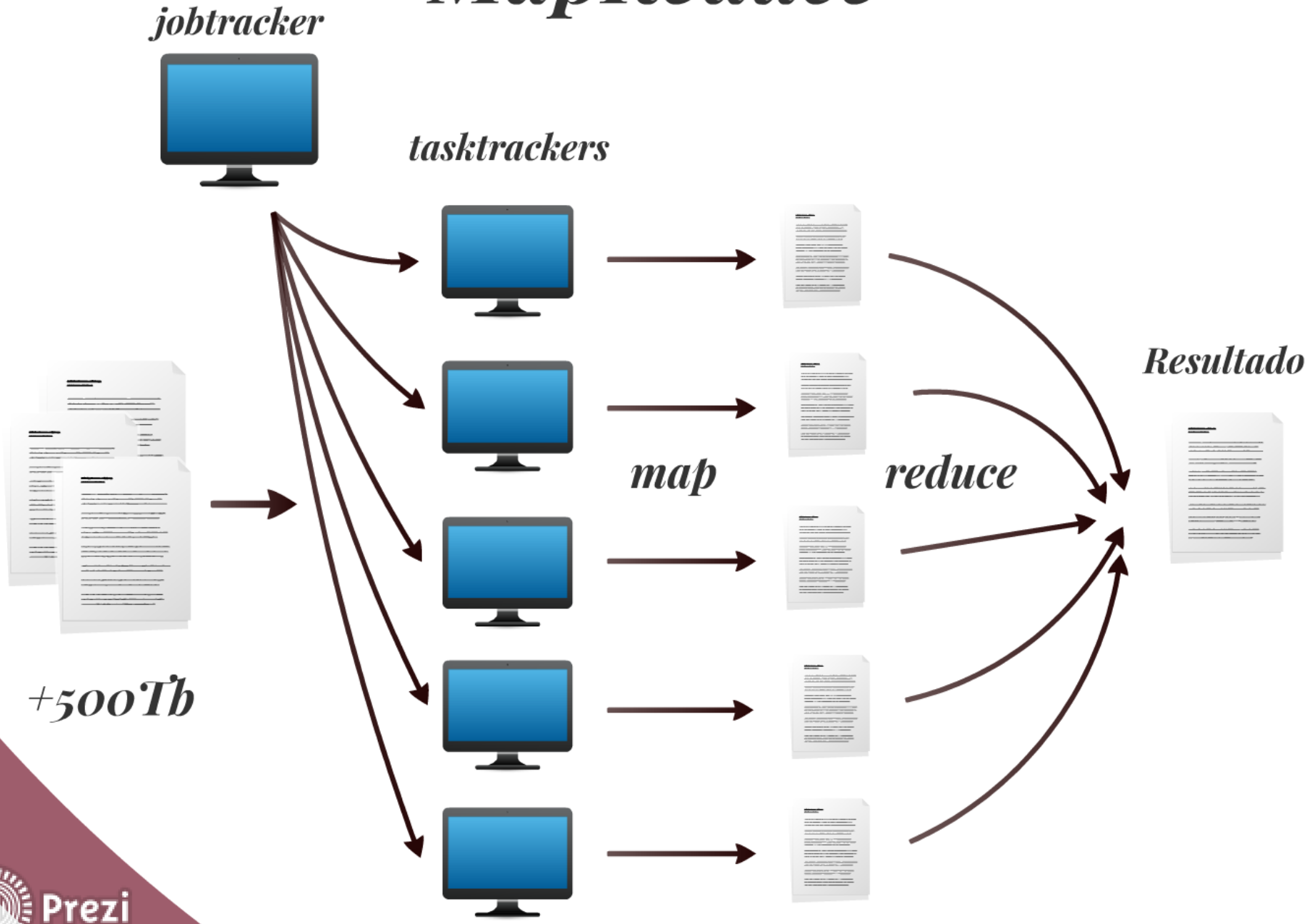
ES y MapRe

Google File System



HDFS (Hadoop Distributed File System)

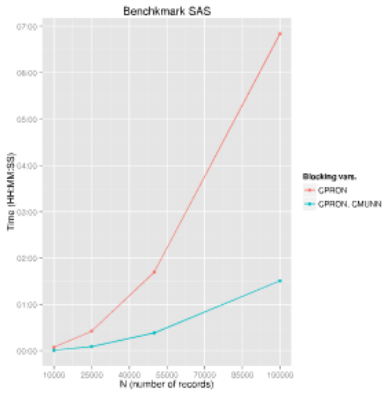
MapReduce



Ejemplo práctico

Deduplicación del Registro de padrón

SAS



70 Mb ?

500000 registros

(500000 x 499999) / 2

~=

125 000 000 000

comparaciones

x 2 x 0,00014

~=

17 Tb

Hadoop

Clúster MapReduce de 100 nodos en el Céntro de Cálculo de la ETSII

Hadoop job_201404252110_0010 on bigdata

User: hduser
Job Name: EstadisticosRegistro
Job File: hdfs://bigdata:54310/app/hadoop/tmp/mapred/staging/hduser/.staging/job_201404252110_0010/job.xml
Submit Host: bigdata
Submit Host Address: 192.168.101.228
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Running
Started at: Sat Apr 26 17:24:55 WEST 2014
Running for: 15mins, 24sec
Job Cleanup: Pending

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	31.34%	198	0	178	22	0	0 / 2
reduce	8.53%	1	0	1	0	0	0 / 0

Un

MapReduce, el algoritmo de Big Data

Map task

ación del Registro de padrón

500000 registros

(500000 x 499999) / 2

~=

125 000 000 000

comparaciones

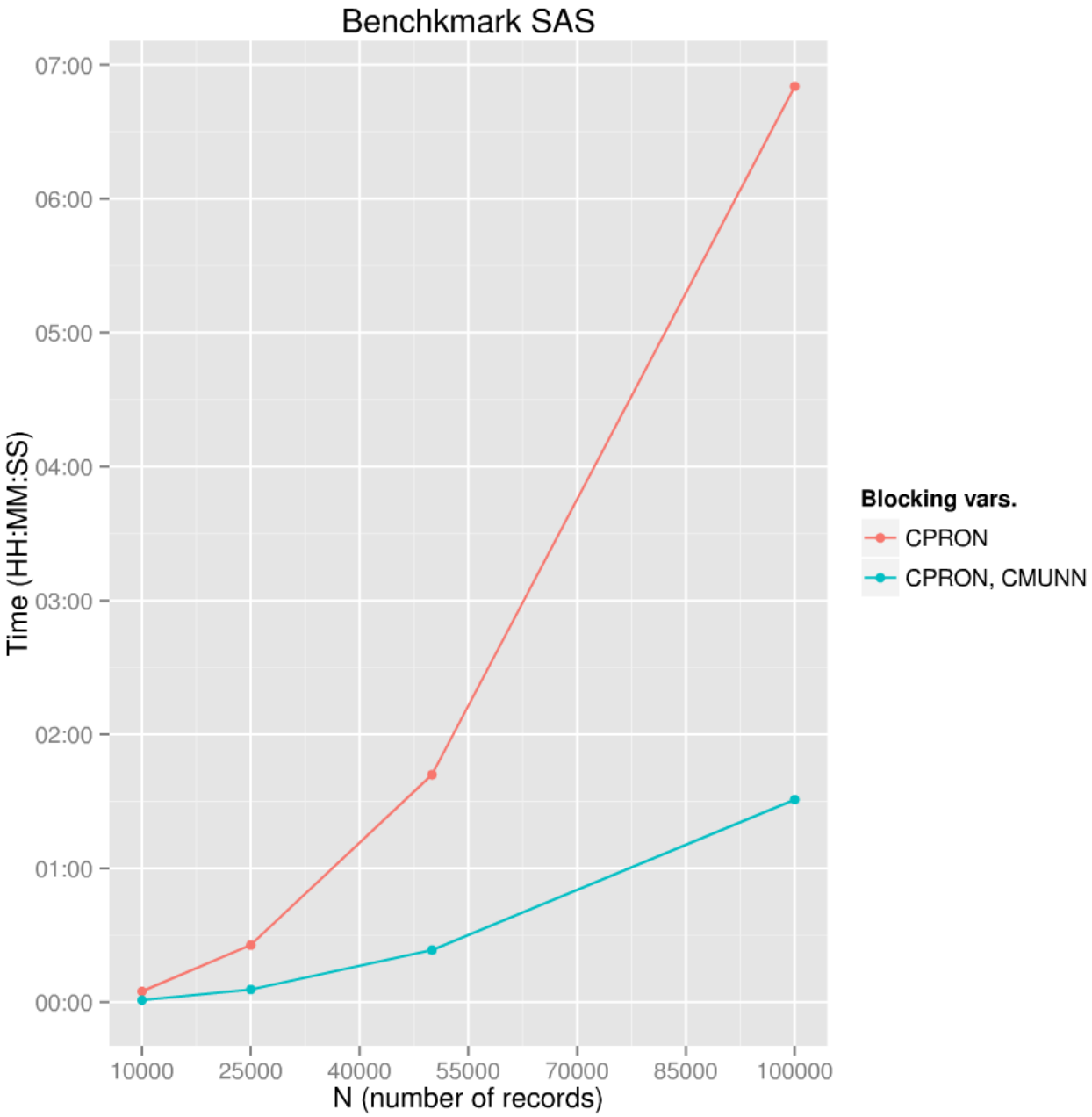
x 2 x 0,00014

~=

17 Tb

70 Mb ?

SAS


Benchkmark SAS
07:00
06:00
05:00
04:00
03:00
02:00
01:00
00:00
Time (HH:MM:SS)
10000
25000
40000
55000
70000
85000
100000
N (number of records)
Blocking vars.
CPRON
CPRON, CMUNN

Hadoop

Clúster MapReduce de 100 nodos en el Céntro de Cálculo de la ETSII

Hadoop job_201404252110_0010 on bigdata

User: hduser
Job Name: EstadisticasRegistro
Job File: hdfs://bigdata:54310/app/hadoop/tmp/mapred/staging/hduser/.staging/job_201404252110_0010/job.xml
Submit Host: bigdata
Submit Host Address: 192.168.101.228
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Running
Started at: Sat Apr 26 17:24:55 WEST 2014
Running for: 15mins, 24sec
Job Cleanup: Pending

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	31.34%	198	0	176	22	0	0 / 2
reduce	3.53%	1	0	1	0	0	0 / 0