

# UNIDAD 2: MODELO DE MAPEO Y REDUCCIÓN

## MODELO DE PROGRAMACIÓN

---

Blanca Vázquez

Febrero 2020

Mining of Massive Datasets

Jure Leskovec, Anand Rajaraman, Jeff Ullman

Stanford University

<http://www.mmds.org>

# ¿QUÉ ES MAPREDUCE?

Es un modelo de programación para el procesamiento de datos distribuidos a gran escala

- Fue inspirado en la programación funcional (LISP, 1960)
- Se caracteriza por ser simple y elegante
- Permite la construcción en bloques
- Está diseñado para ser ejecutado en clústers

## Características

- Toma ventaja del paralelismo
- Tolerante a fallas
- Es extensible para diferentes aplicaciones

- Grande cantidades de datos (Big Data)
  - Astronomía, finanzas, ciencias, sitios webs....
- Cuarto paradigma de la ciencia
  - Diseño de algoritmos capaces de procesar datos en tiempo real
- No es el algoritmo, ¡son los datos!
  - Más datos, mejor precisión

# CONTANDO PALABRAS CON MAPREDUCE

## Extracto de big\_file.txt

Armstrong joined the NASA Astronaut Corps in the second group, which was selected in 1962. He made his first spaceflight as command pilot of Gemini 8 in March 1966, becoming NASA's first civilian astronaut to fly in space. During this mission with pilot David Scott, he performed the first docking of two spacecraft; the mission was aborted after Armstrong used some of his re-entry control fuel to stabilize ...



```
cat big_file.txt | tr ' ' '\n' >> out_bigfile.txt
```



Armstrong  
joined  
the  
NASA  
Astronaut  
Corps  
in  
the  
second  
group,  
which  
was  
selected  
in  
1962.  
He  
made  
his  
first  
spaceflight  
as

# CONTANDO PALABRAS CON MAPREDUCE

## Salida de out\_bigfile.txt

Armstrong  
joined  
the  
NASA  
Astronaut  
Corps  
in  
the  
second  
group,  
which  
was  
selected  
in  
1962.  
He  
made  
his  
first  
spaceflight  
as  
command  
pilot



cat out\_bigfile.txt | sort | uniq -c

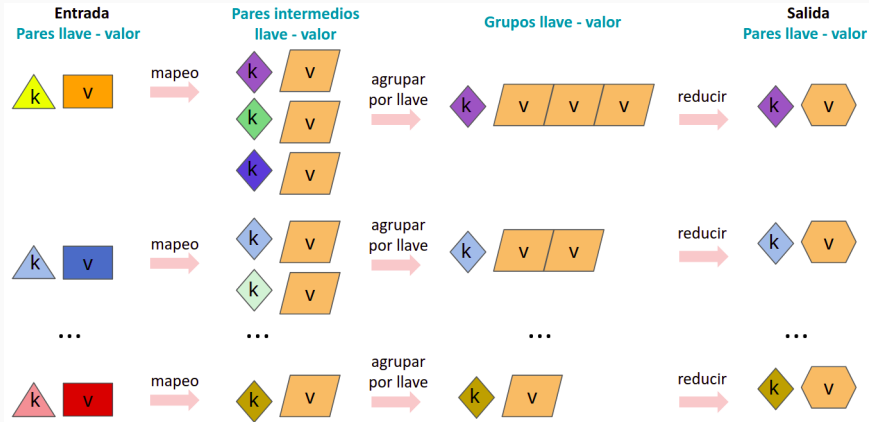


5 the  
4 in  
4 of  
3 a  
3 first  
3 to  
2 Armstrong  
2 as  
2 During  
2 he  
2 his  
2 mission  
2 pilot  
2 second  
2 spaceflight  
2 was  
1 aborted  
1 after  
1 and  
1 Apollo  
1 astronaut  
1 becoming  
1 before

## Contar frecuencias de palabras

- Función de mapa
  1. Escanea la entrada de un archivo, va un registro a la vez
  2. Por cada registro, extrae las palabras (llaves)
- Agrupación por llave
  1. Se agrupan las llaves con el mismo valor
- Función de reducción
  1. Realiza la función de suma (resumen, filtrado, agregación, transformación).
  2. Se escribe el resultado: palabra - frecuencia

# MAPREDUCE



# MAPREDUCE: EJERCICIO

**Mapeo**  
Lee una entrada  
y produce un  
conjunto de pares  
llave - valor

Hoy empecé la dieta verde:  
verde lejos la pizza,  
verde lejos los tamales,  
verde lejos las tortas,  
verde lejos el pan.

(hoy, 1)  
(empece, 1)  
(la, 1)  
(dieta, 1)  
(verde, 1)  
(verde, 1)  
(lejos, 1)  
(la, 1)  
(pizza, 1)  
(verde, 1)  
(lejos, 1)  
(los, 1)  
(tamales, 1)  
(verde, 1)  
(lejos, 1)  
(las, 1)  
(tortas, 1)  
(verde, 1)  
(lejos, 1)  
(el, 1)  
(pan, 1)

(llave, valor)

**Agrupar por  
llaves:**  
colecciona todos  
los pares con la  
misma llave

(hoy, 1)  
(empece, 1)  
(la, 1)  
(la, 1)  
(dieta, 1)  
(verde, 1)  
(verde, 1)  
(verde, 1)  
(verde, 1)  
(verde, 1)  
(lejos, 1)  
(lejos, 1)  
(lejos, 1)  
(lejos, 1)  
(lejos, 1)  
(lejos, 1)  
(tamales, 1)  
(las, 1)  
(tortas, 1)  
(el, 1)  
(pan, 1)

(llave, valor)

**Reducir:**  
colecciona todos  
los valores que  
pertenecen a la  
llave

(hoy, 1)  
(empece, 1)  
(la, 2)  
(dieta, 1)  
(verde, 5)  
(lejos, 4)  
(pizza, 1)  
(los, 1)  
(tamales, 1)  
(las, 1)  
(tortas, 1)  
(el, 1)  
(pan, 1)

(llave, valor)

Únicamente lecturas secuenciales



# MAPREDUCE: EJERCICIO

El programador indica cuántos nodos necesita para la tarea de Mapeo y cuántos para la tarea de reducción (5NM- 3NR)

Hoy empecé la dieta verde:  
verde lejos la pizza,  
verde lejos los tamales,  
verde lejos las tortas,  
verde lejos el pan.

**Mapeo**  
Lee una entrada y produce un conjunto de pares llave - valor

```
(hoy, 1)
(emepece, 1)
(la, 1)
(dieta, 1)
(verde, 1)
(verde, 1)
(lejos, 1)
(la, 1)
(pizza, 1)
(verde,1)
(lejos, 1)
(los,1)
(tamales, 1)
(verde,1)
(lejos,1)
(las, 1)
(tortas,1)
(verde,1)
(lejos,1)
(el,1)
(pan,1)
```

(llave, valor)

**Agrupar por llaves:**  
colecciona todos los pares con la misma llave

```
(hoy, 1)
(emepece, 1)
(la, 1)
(la, 1)
(dieta, 1)
(verde, 1)
(verde, 1)
(verde,1)
(verde,1)
(lejos, 1)
(lejos, 1)
(lejos,1)
(lejos,1)
(pizza, 1)
(pizza, 1)
(los, 1)
(tamales, 1)
(las,1)
(tortas,1)
(el,1)
(pan,1)
```

(llave, valor)

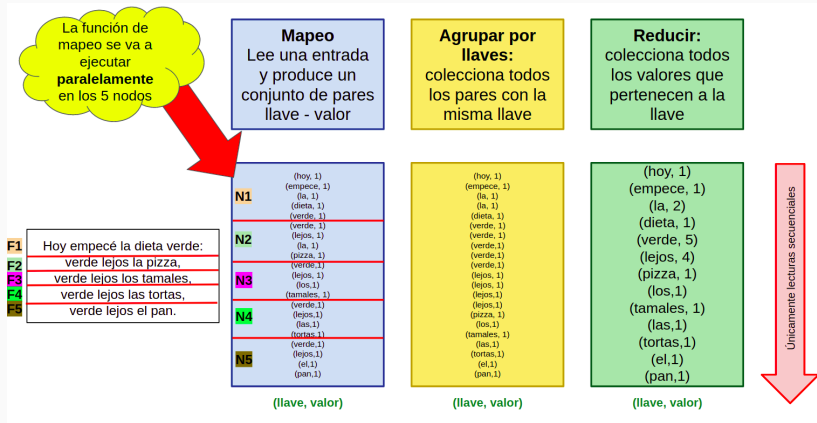
**Reducir:**  
colecciona todos los valores que pertenecen a la llave

```
(hoy, 1)
(emepece, 1)
(la, 2)
(dieta, 1)
(verde, 5)
(lejos, 4)
(pizza, 1)
(los,1)
(tamales, 1)
(las,1)
(tortas,1)
(el,1)
(pan,1)
```

(llave, valor)

Únicamente lecturas secuenciales

# MAPREDUCE: EJERCICIO



# MAPREDUCE: EJERCICIO

Indicamos que vamos a usar 3 nodos para reducción.

**Mapeo**  
Lee una entrada y produce un conjunto de pares llave - valor

**Agrupar por llaves:**  
colecciona todos los pares con la misma llave

**Reducir:**

F1 Hoy empecé la dieta verde:  
F2 verde lejos la pizza,  
F3 verde lejos los tamales,  
F4 verde lejos las tortas,  
F5 verde lejos el pan.

N1	(hoy, 1) (empece, 1) (la, 1) (dieta, 1) (verde, 1)	Out_m1
N2	(lejos, 1) (la, 1) (pizza, 1) (verde, 1)	Out_m2
N3	(lejos, 1) (los, 1) (tamales, 1) (verde, 1)	Out_m3
N4	(lejos, 1) (las, 1) (tortas, 1) (verde, 1)	Out_m4
N5	(lejos, 1) (el, 1) (pan, 1)	Out_m5

(llave, valor)

(hoy, 1) (empece, 1) (la, 1) (dieta, 1) (verde, 1) (verde, 1) (verde, 1) (verde, 1) (verde, 1) (lejos, 1) (lejos, 1) (lejos, 1) (lejos, 1) (pizza, 1) (los, 1) (tamales, 1) (las, 1) (tortas, 1) (el, 1) (pan, 1)
--

(llave, valor)

(hoy, 1) (empece, 1) (la, 2) (dieta, 1) (verde, 5) (lejos, 4) (pizza, 1) (los, 1) (tamales, 1) (las, 1) (tortas, 1) (el, 1) (pan, 1)
--

(llave, valor)

Unicamente lecturas secuenciales

# MAPREDUCE: EJERCICIO

Indicamos que vamos a usar 3 nodos para reducción.

**Mapeo**  
Lee una entrada y produce un conjunto de pares llave - valor

**Agrupar por llaves:**  
colecciona todos los pares con la

**Reducir:**  
colecciona todos los valores que pertenecen a la

## Funciones Hash

(enumera cada llave y determina un solo nodo)

F1 Hoy empecé la dieta verde:  
F2 verde lejos la pizza,  
F3 verde lejos los tamales,  
F4 verde lejos las tortas,  
F5 verde lejos el pan.

N1	(hoy, 1) (empece, 1) (la, 1) (dieta, 1) (verde, 1)	Out_m1
N2	(lejos, 1) (la, 1) (pizza, 1) (verde, 1)	Out_m2
N3	(lejos, 1) (los, 1) (tamales, 1) (verde, 1)	Out_m3
N4	(lejos, 1) (las, 1) (tortas, 1) (verde, 1)	Out_m4
N5	(lejos, 1) (el, 1) (pan, 1)	Out_m5

(llave, valor)

(hoy, 1) (empece, 1) (la, 1) (dieta, 1) (verde, 1) (lejos, 1) (verde, 1) (verde, 1) (verde, 1) (lejos, 1) (lejos, 1) (lejos, 1) (lejos, 1) (lejos, 1) (pizza, 1) (los, 1) (tamales, 1) (las, 1) (tortas, 1) (el, 1) (pan, 1)
--

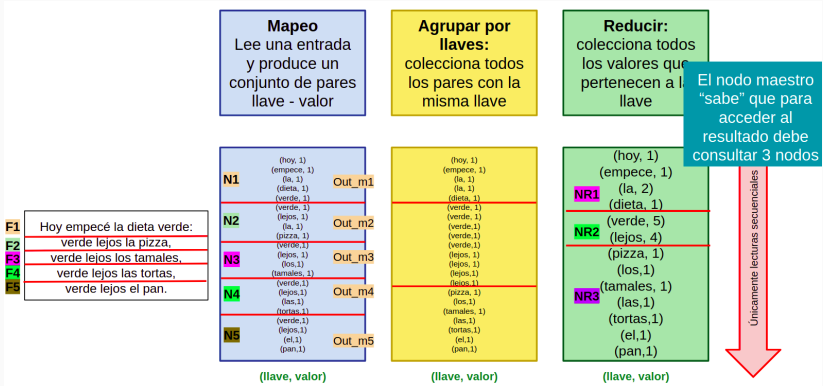
(llave, valor)

(hoy, 1) (empece, 1) (la, 2) (dieta, 1) (verde, 5) (lejos, 4) (lejos, 1) (los, 1) (tamales, 1) (las, 1) (tortas, 1) (el, 1) (pan, 1)
--

(llave, valor)

Unicamente lecturas secuenciales

# MAPREDUCE: EJERCICIO



# MAPREDUCE: EJERCICIO

Las lecturas  
secuenciales son  
mucho más eficientes  
que los accesos  
aleatorios

**Mapeo**  
Lee una entrada  
y produce un  
conjunto de pares  
llave - valor

**Agrupar por  
llaves:**  
colecciona todos  
los pares con la  
misma llave

**Reducir:**  
colecciona todos  
los valores que  
pertenecen a la  
llave

F1 Hoy empecé la dieta verde:  
F2 verde lejos la pizza,  
F3 verde lejos los tamales,  
F4 verde lejos las tortas,  
F5 verde lejos el pan.

N1	(hoy, 1) (empece, 1) (la, 1) (dieta, 1) (verde, 1)	Out_m1
N2	(lejos, 1) (la, 1) (pizza, 1)	Out_m2
N3	(verde, 1) (lejos, 1) (los, 1) (tamales, 1)	Out_m3
N4	(verde, 1) (lejos, 1) (las, 1) (tortas, 1)	Out_m4
N5	(verde, 1) (lejos, 1) (el, 1) (pan, 1)	Out_m5

(llave, valor)

(hoy, 1) (empece, 1) (la, 1) (dieta, 1) (verde, 1) (verde, 1) (verde, 1) (verde, 1) (lejos, 1) (lejos, 1) (lejos, 1) (pizza, 1) (las, 1) (tamales, 1) (los, 1) (tortas, 1) (el, 1) (pan, 1)
--

(llave, valor)

(hoy, 1) (empece, 1) (la, 2) (dieta, 1) (verde, 5) (lejos, 4) (pizza, 1) (los, 1) (tamales, 1) (las, 1) (tortas, 1) (el, 1) (pan, 1)
--

(llave, valor)

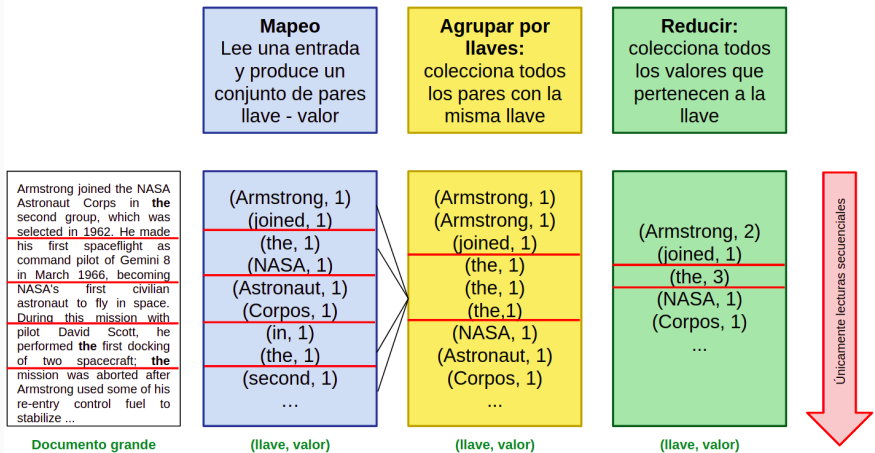
Únicamente lecturas secuenciales

MapReduce está construido sobre lecturas de archivos secuenciales y  
nunca sobre accesos aleatorios

# MAPREDUCE: MÁS FORMALMENTE

- Entrada: un conjunto de pares llave - valor
- El programador especifica dos métodos:
  - **Función de mapeo**
    - $\text{Mapeo}(k, v) \rightarrow \langle k', v' \rangle$
    - Se toma un par llave - valor y la salida es un conjunto de pares llave - valor
    - Existe un solo mapeo por cada par  $(k, v)$
  - **Función de reducción**
    - $\text{Reduccion}(k', \langle v' \rangle^*) \rightarrow \langle k', v'' \rangle^*$
    - Todos los valores  $v'$  con la misma llave  $k$  serán agrupados
    - Existe una sola función de reducción por cada llave única  $k'$
- Salida: un conjunto de llaves y su valor (resultado de una función)

# MAPREDUCE: CONTEO DE PALABRAS





# MAPREDUCE: CONTEO DE PALABRAS

