Análisis de datos masivos

Modelo de programación MapReduce

Blanca Vázquez

27 de febrero de 2025

¿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

Motivación

- 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

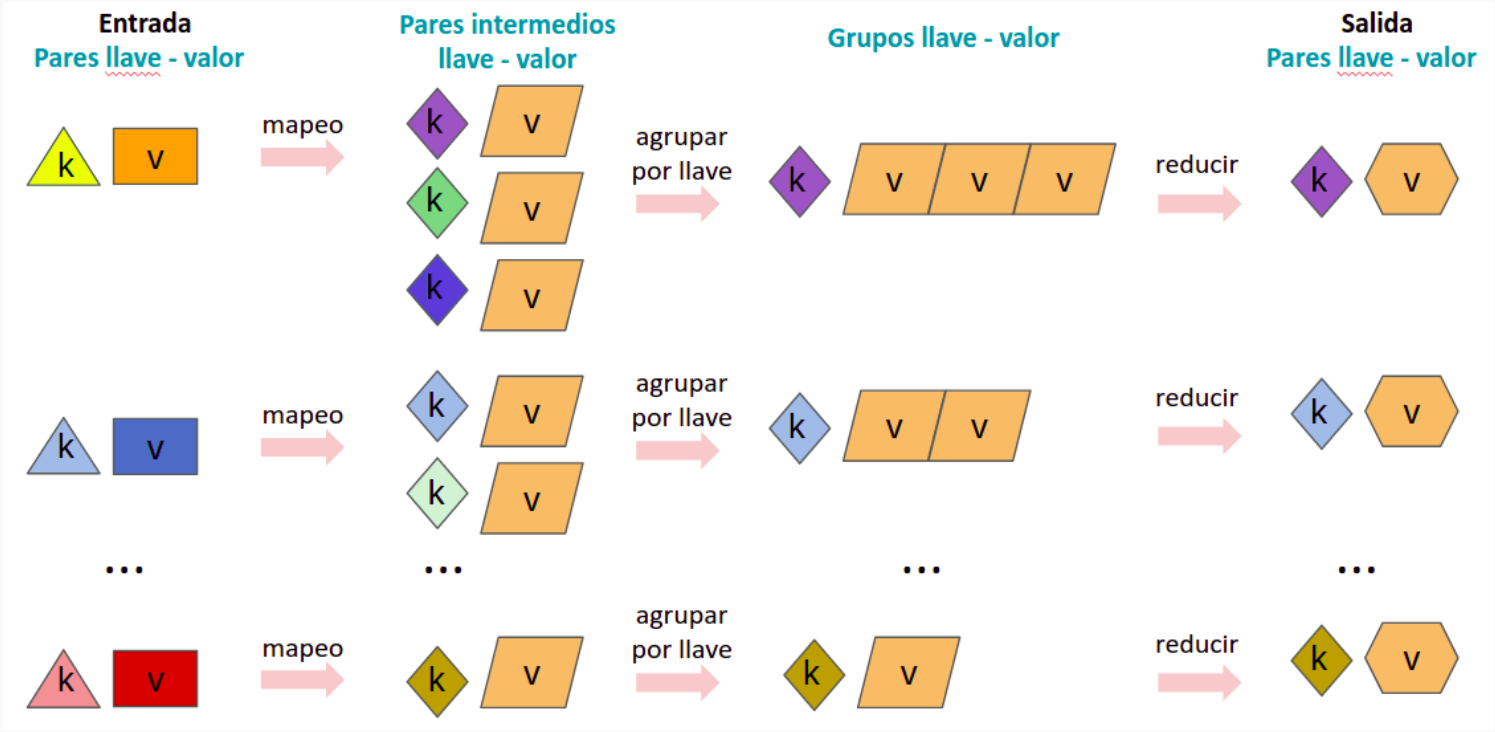
→ cat out_bigfile.txt | sort | uniq -c →

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

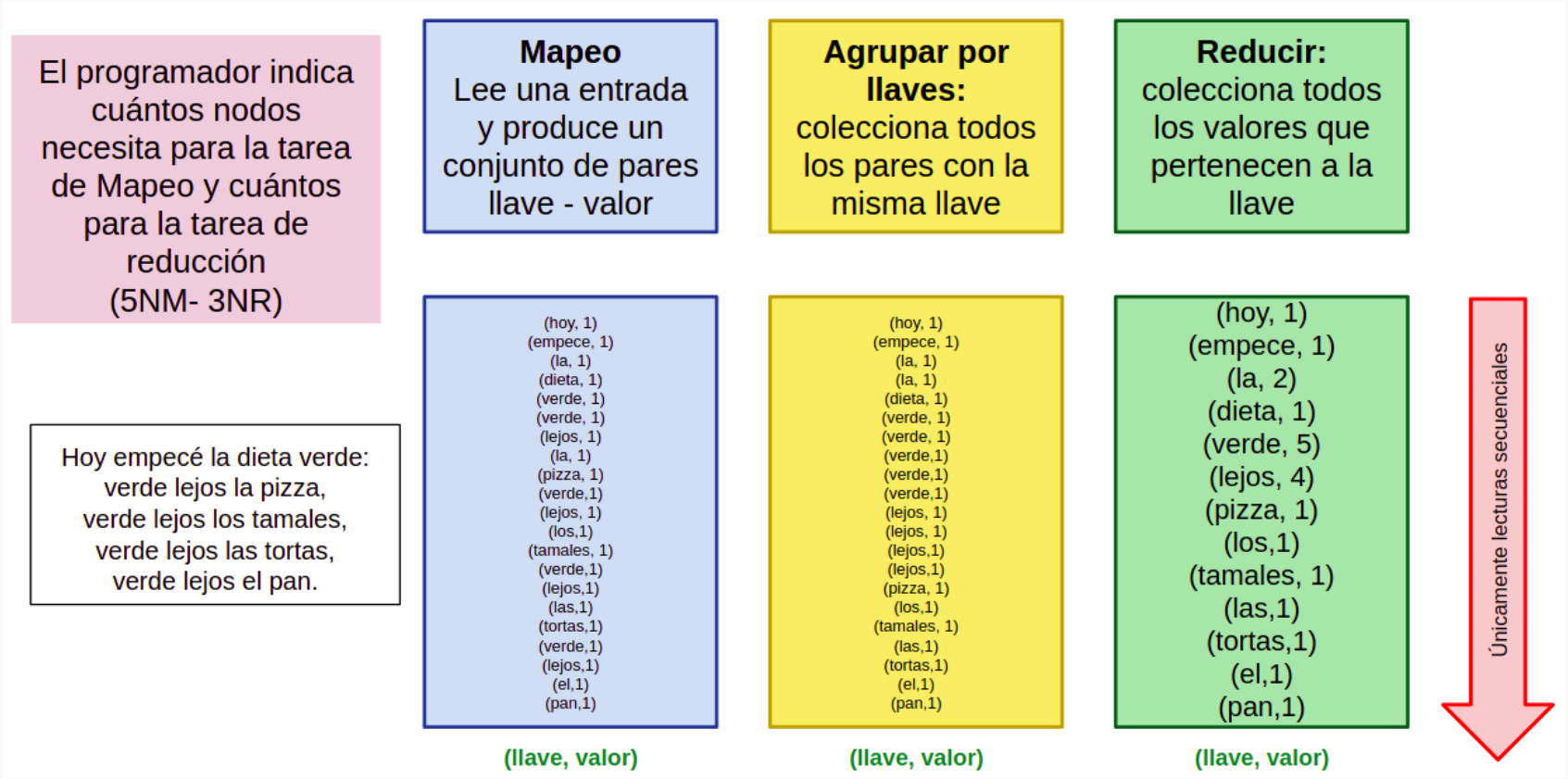
- Función de mapa
  1. Escanea la entrada de un archivo, va un registro a la vez
  2. Por cada registro, extrae las palabras (llaves)
- Agrupación por llave
  1. Se agrupan las llaves con el mismo valor
- Función de reducción
  1. Realiza la función de suma (resumen, filtrado, agregación, transformación).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

MapReduce

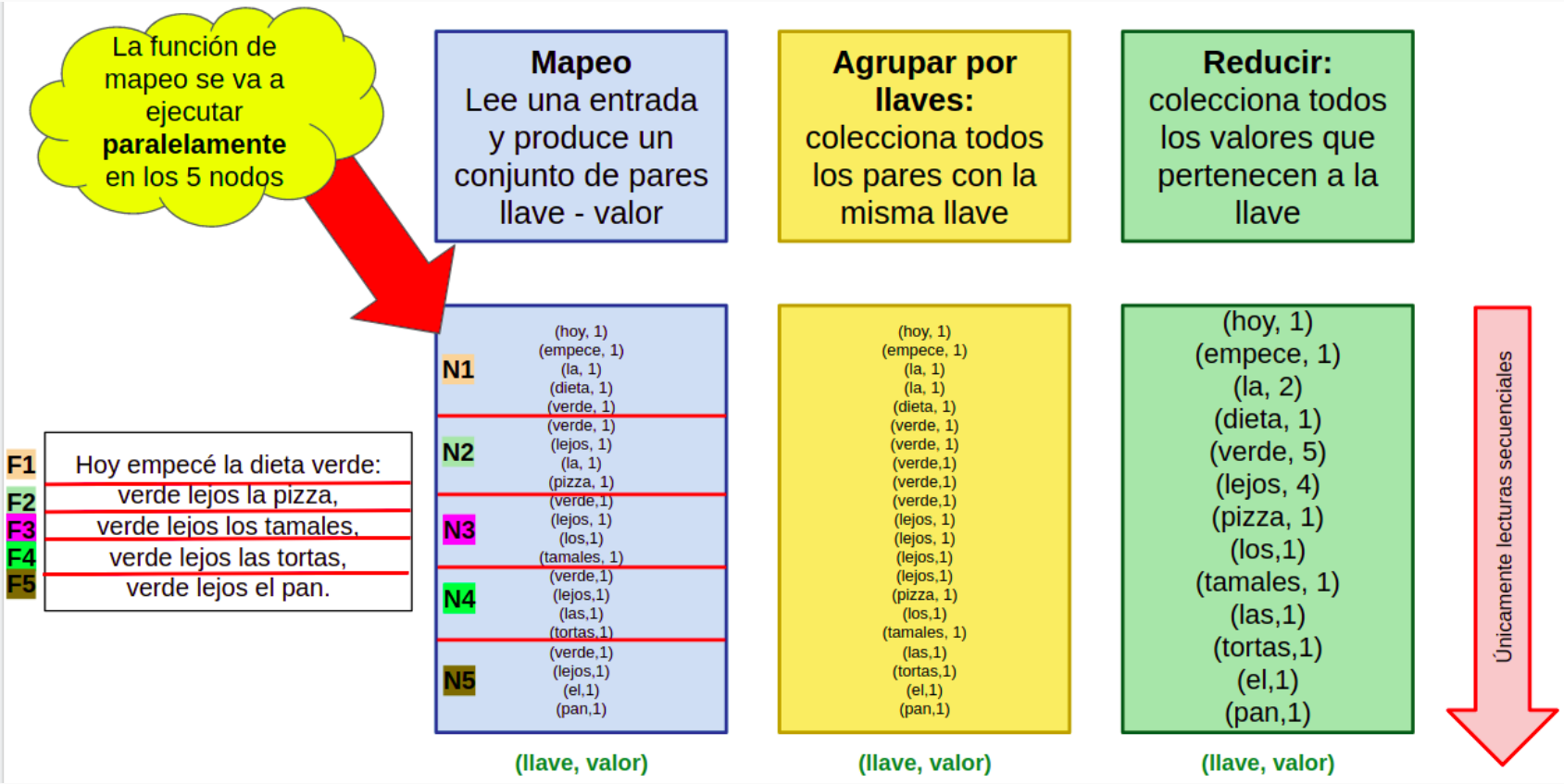

Entrada
Pares llave - valor
Pares intermedios
llave - valor
Grupos llave - valor
Salida
Pares llave - valor
mapeo
agrupar
por llave
reducir
k
v

MapReduce: ejercicio

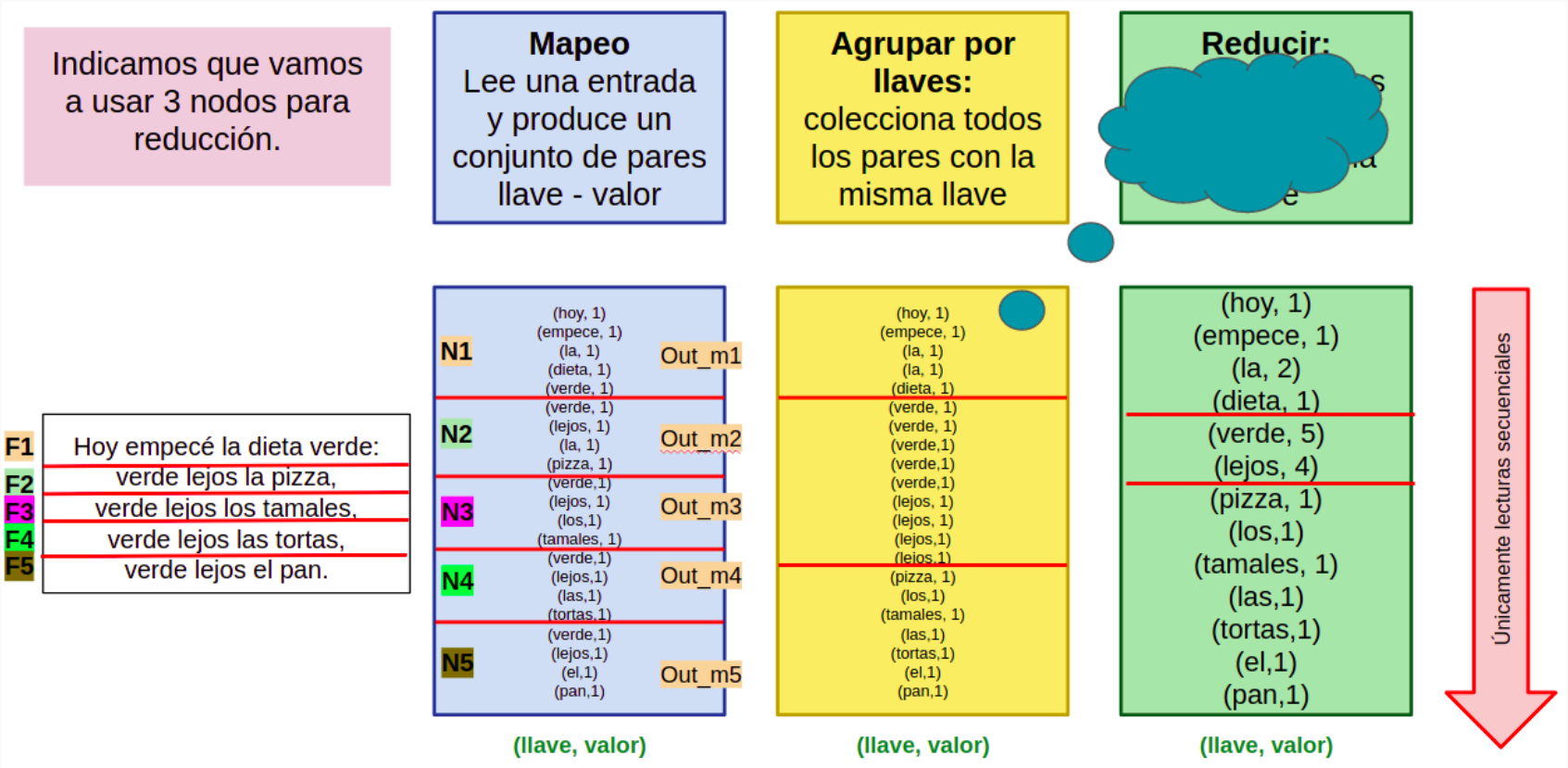
MapReduce: ejercicio



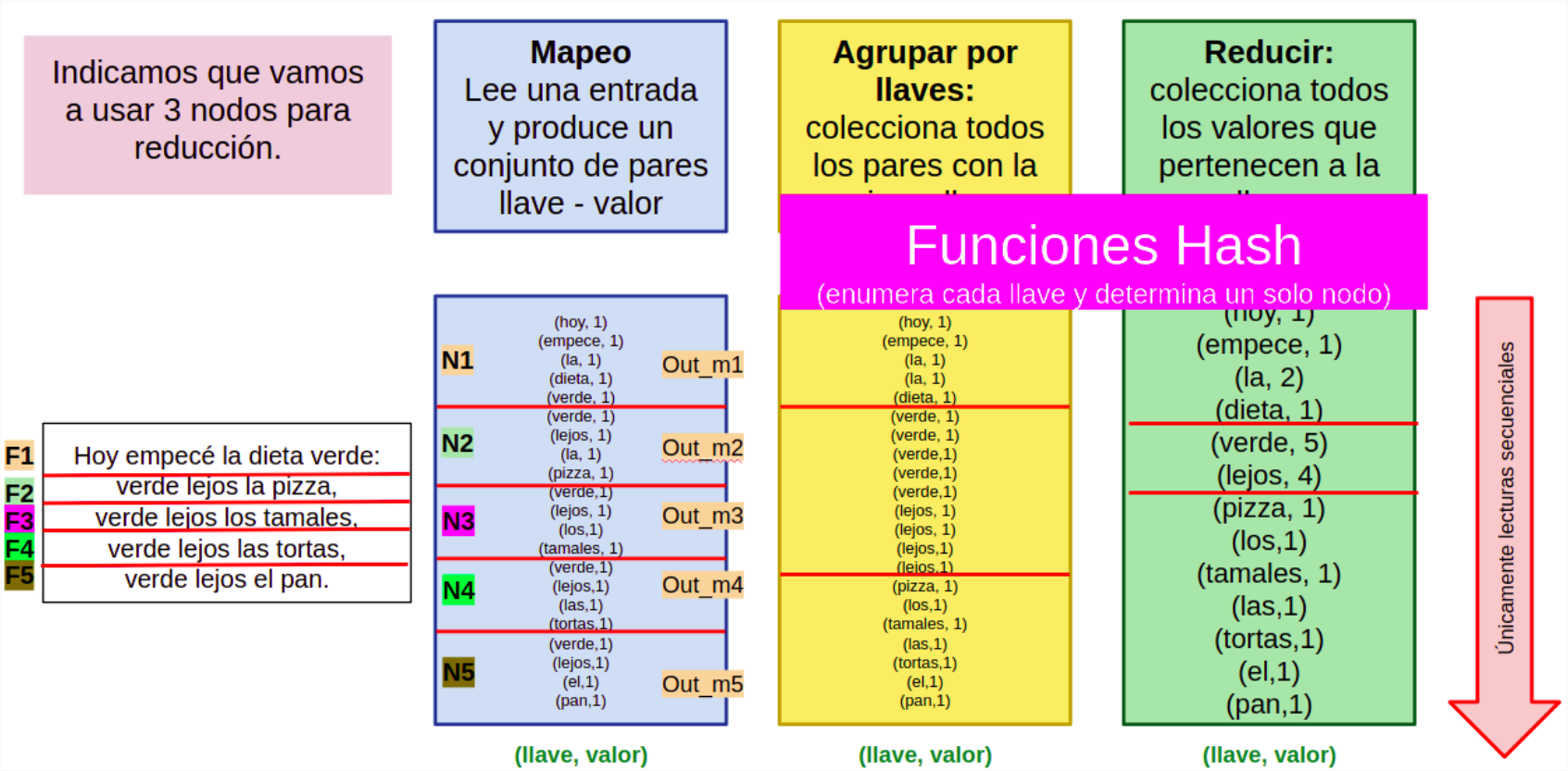
MapReduce: ejercicio



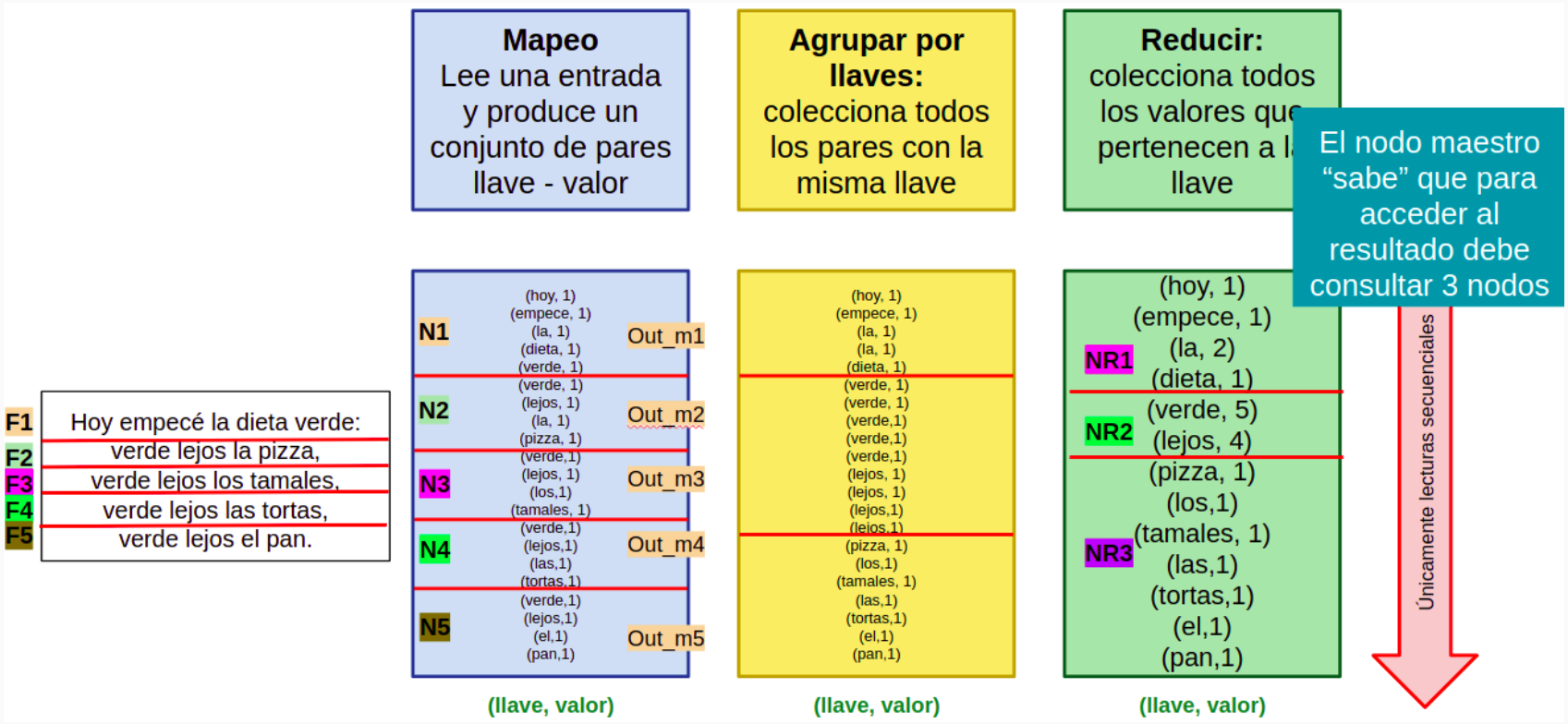
MapReduce: ejercicio



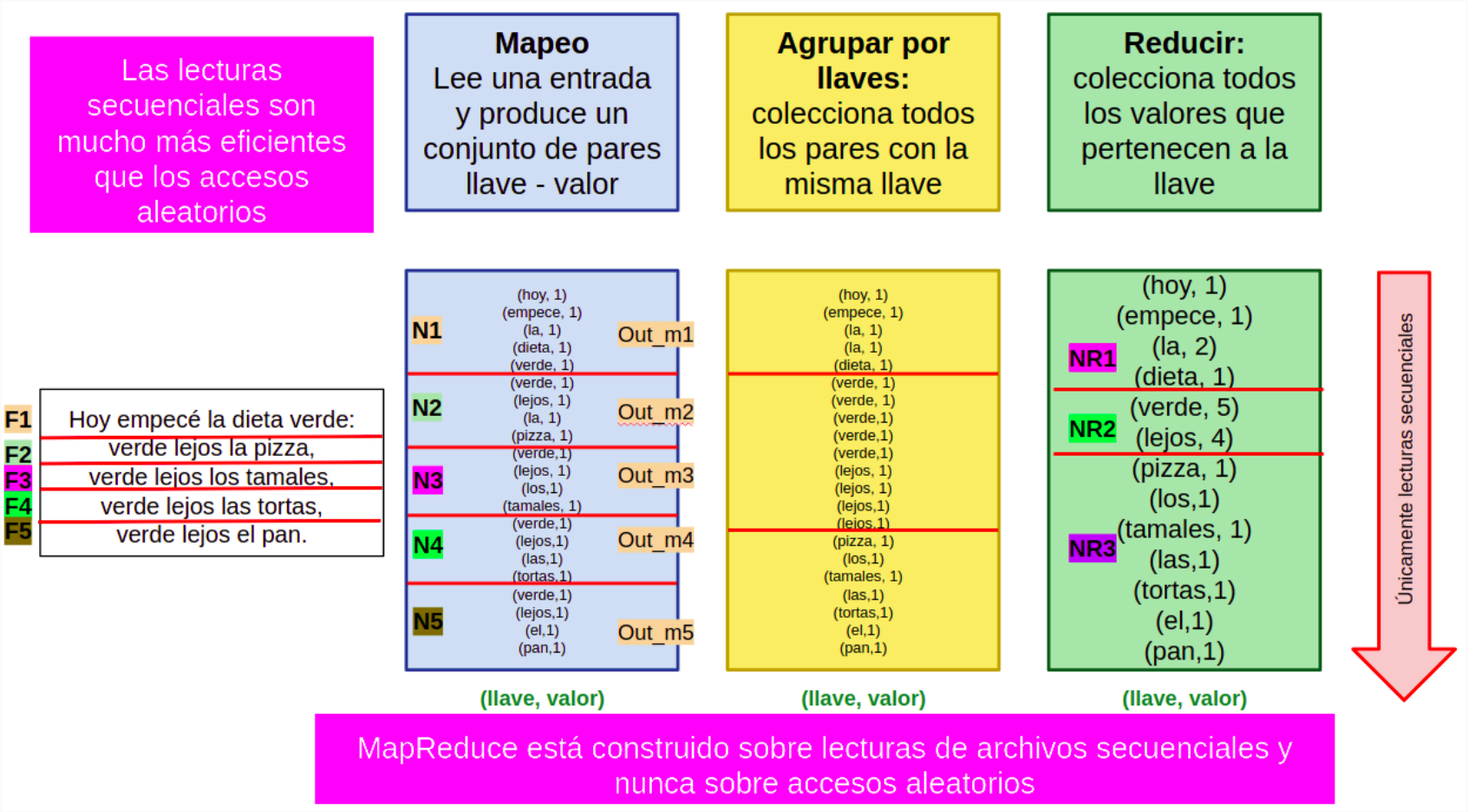
MapReduce: ejercicio



MapReduce: ejercicio



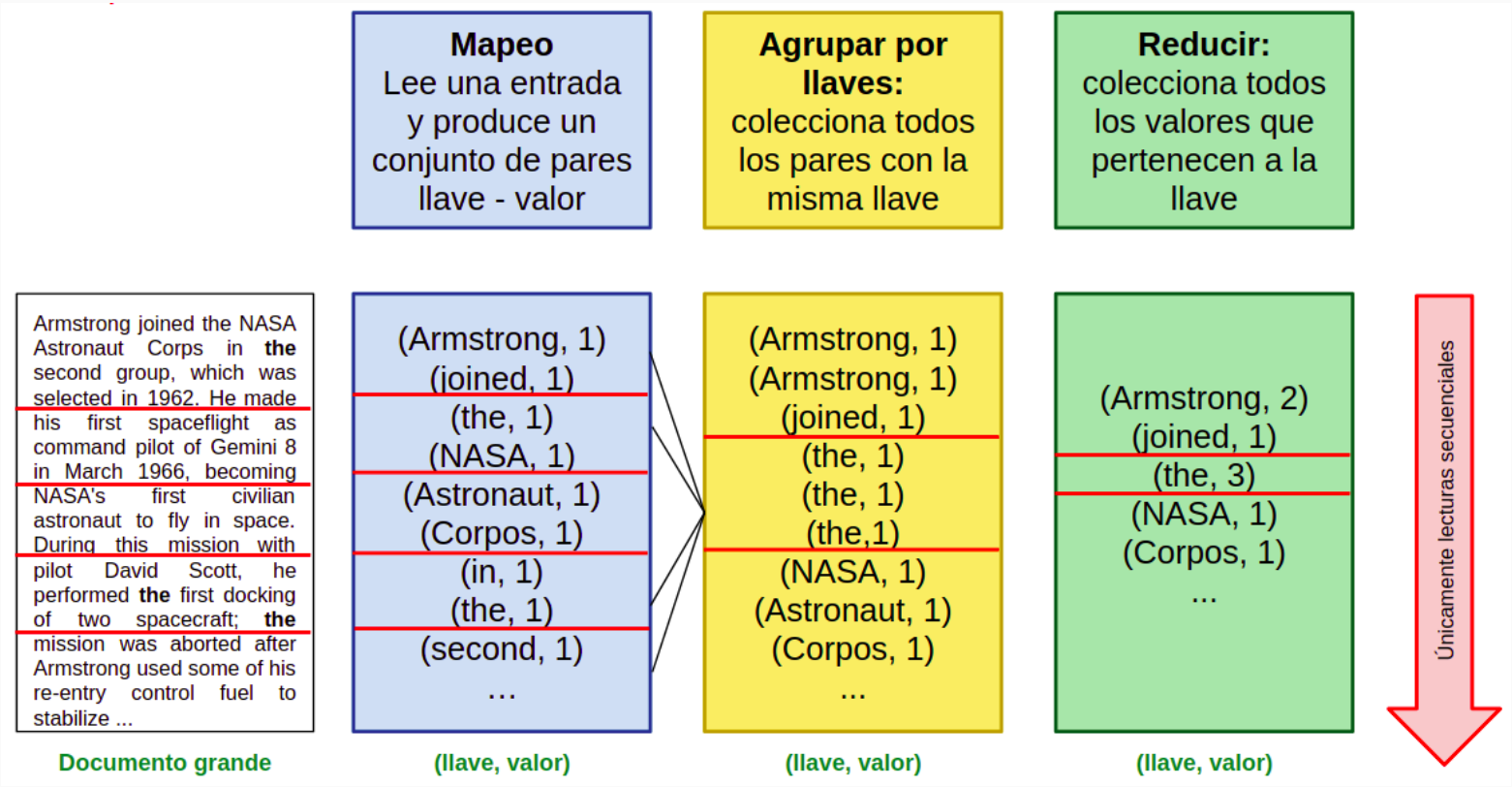
MapReduce: ejercicio



MapReduce: más formalmente

- Entrada: un conjunto de pares llave - valor
- El programador especifica dos métodos:
  - Función de mapeo
    - $Mapeo(k, v) \rightarrow < k', v' >$
    - 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
    - $Reduccion(k', < v' >^*) \rightarrow < k', v'' >^*$
    - 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