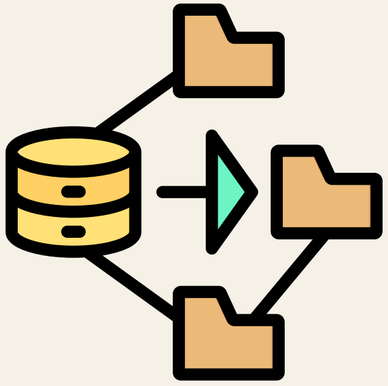


SUJET

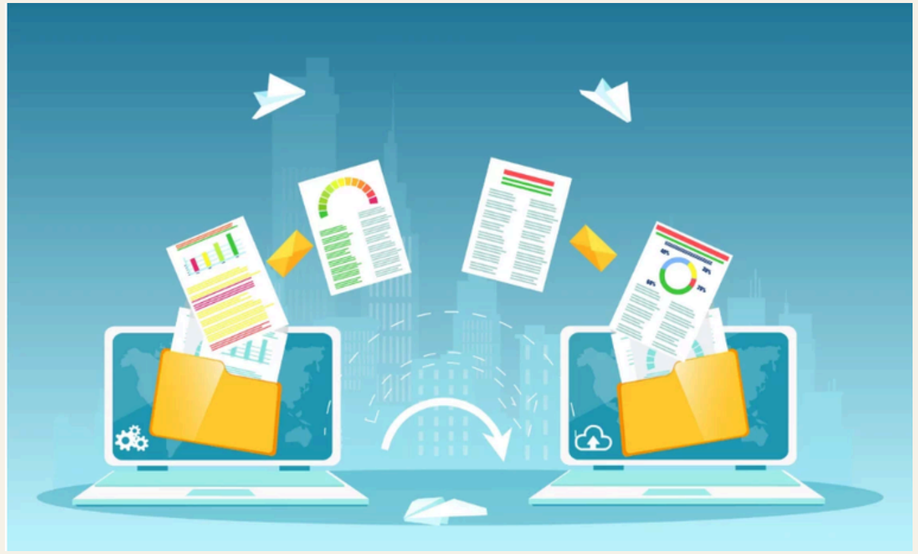
Optimisation du traitement des données massives avec MapReduce et le rôle du Combiner

BY Marwan / Yas / Moh / Leo

M.E

POURQUOI MAPREDUCE ?

- Aujourd'hui, on traite des données massives (big data) : ex. recherches Google, vidéos YouTube, transactions bancaires...
- Problème : un seul ordinateur ne suffit pas.
- Solution : répartir le travail entre plusieurs machines → c’est le but de MapReduce.



M.E

LE MODÈLE MAPREDUCE EN 2 PHASES

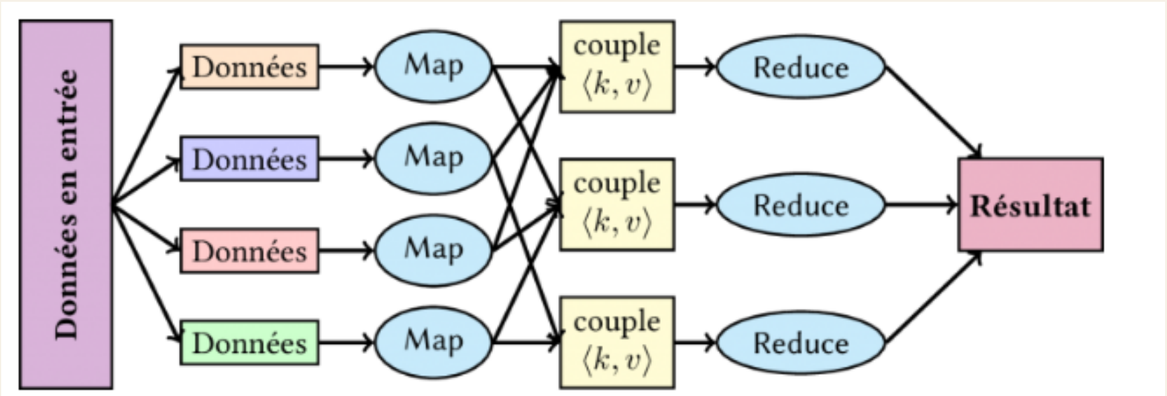
1. Map :
- Lit les données.
- Transforme chaque morceau en paires clé/valeur (ex : "chat", 1).
- Fonctionne en parallèle sur chaque machine.
2. Reduce :
- Regroupe toutes les valeurs ayant la même clé.
- Calcule un résultat final (ex : compter les "chat").

M.E

ET ENTRE MAP ET REDUCE ?

1. Entre Map et Reduce, il y a une phase Shuffle & Sort.
2. Toutes les paires sont envoyées sur le réseau pour être regroupées par clé.
3. Cela peut créer un goulot d'étranglement, surtout avec des milliards de données.

Texte avec "chat chien chat" → devient → ("chat", 1), ("chien", 1), ("chat", 1)



M.y

UNE SOLUTION : LE COMBINER

- Le Combiner est une sorte de mini-Reduce local, exécuté avant le Shuffle.
- Il regroupe déjà les paires similaires sur chaque machine.
- Moins de données à envoyer ➜ réseau plus fluide ➜ traitement plus rapide.

Analogie simple :

Avant de livrer tous les colis individuellement, un entrepôt regroupe les colis par ville.

M.y

ILLUSTRATION CONCRÈTE

Étape	Sans Combiner	Avec Combiner
Map	("chat",1) × 1000	("chat",1) × 1000
Combiner	—	("chat",1000)
Shuffle	1000 éléments	1 élément
Reduce	Calcule total	Calcule (déjà résumé)

✓ **Moins de données →** ✓ **Moins de trafic →** ✓ **Plus rapide**

L.S

CE QUE LE COMBINER NE PEUT PAS FAIRE

- N’est pas toujours exécuté (Hadoop décide).
- Ne doit être utilisé que si la fonction est :
  - Associative : `(a + b) + c = a + (b + c)`
  - Commutative : `a + b = b + a`
- Ne convient pas pour des fonctions comme la moyenne ou médiane.

Exemple :
Somme : oui
Moyenne : non

L.S

BIEN UTILISER LE COMBINER

Recommandations :

- Utilisez-le si les résultats peuvent être agrégés localement sans perte de sens.
- Évitez-le pour des calculs sensibles à l’ordre ou au contenu exact.
- Testez sur un petit jeu de données avant de généraliser.

M.L

À RETENIR SUR MAPREDUCE ET LE COMBINER

En résumé :

- MapReduce = traitement massif distribué.
- Combiner = optimiseur de performances.
- Bien utilisé, il accélère le traitement en réduisant les données intermédiaires.
- À manipuler avec précaution selon les cas.

"Moins de données" → "Moins de trafic" → "Plus de rapidité"

M.L

CONCLUSION

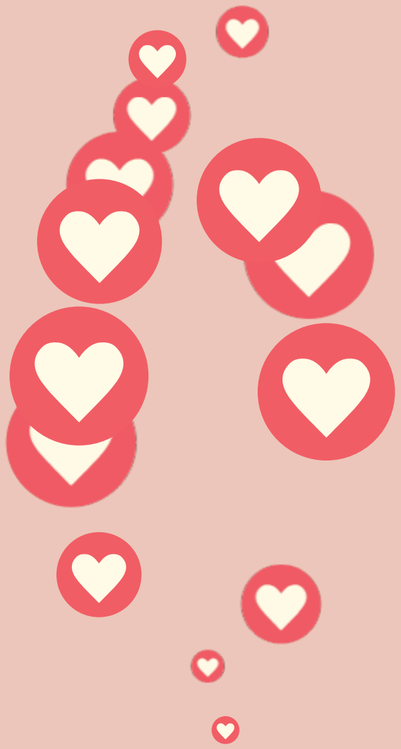
MapReduce, un modèle puissant pour le traitement distribué de grandes quantités de données, peut rencontrer un goulot d’étranglement lors du Shuffle & Sort entre les phases Map et Reduce.

Le Combiner, un mini-Reduce local, réduit le trafic réseau et le temps d’exécution, améliorant ainsi l’efficacité du traitement. Utilisez-le judicieusement pour le traitement des données, mais uniquement dans des cas appropriés.

Une bonne gestion du Combiner améliore la performance du traitement des données massives.

M.L

Supinfo



MERCI !