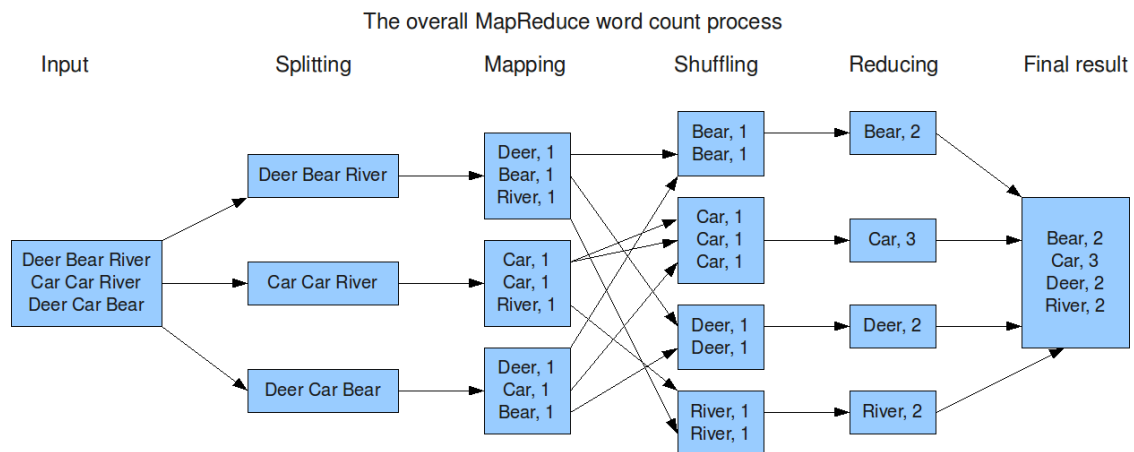
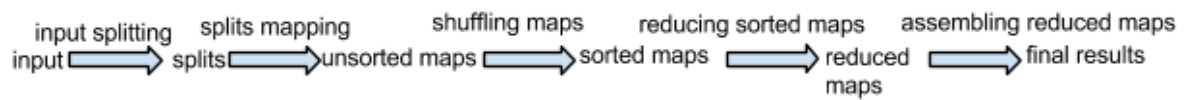


Voir le site

<http://www.guru99.com/introduction-to-mapreduce.html>



Définitions (à partir du schéma avec les blocs bleus):



Sx: Input Split x

Umx: Unsorted Map x

SMx: Sorted Map x

RMx: Reduced Map x

Pour votre propre implémentation vous avez besoin de:

Les dictionnaires du master - les clés ne sont pas dupliquées dans un dictionnaire, elles n'apparaissent qu'une seule fois :

- Dictionnaire "UMx - machines"
pour chaque UMx, le master connaît la machine slave qui a fait le traitement Sx -> UMx (c'est le master qui a effectivement lancé ce calcul).
- Dictionnaire "clés - UMx"
pour chaque clé, le master connaît les UMx correspondants (et donc en utilisant le dictionnaire "UMx - machines", il connaît la machine de chaque UMx).
- Dictionnaire "RMx - machines"
pour chaque RMx, le master connaît la machine qui a fait le traitement SMx -> RMx (c'est le master qui a effectivement lancé ce calcul).

L'algorithme principal du master - les dictionnaires sont sur le master:

1. **Prendre INPUT, faire l'INPUT splitting pour fabriquer les Sx (attention à la taille des splits pour éviter qu'un traitement ne dépasse les capacités d'une machine - en mémoire et en temps - ceci est très difficile à juger, il faut tester...).**
Envoyer les splits sur les slaves.
2. **Pour chaque Sx, lancer le traitement Split Mapping en parallèle (Sx -> UMx) sur les slaves. On peut donc construire le dictionnaire "UMx - machines (slaves)"**
3. **Recevoir les "clés-UMx" depuis les slaves vers le master, issus du traitement Split Mapping et construire le dictionnaire "clés - UMx"**
4. **Attendre la fin des traitements Split Mapping**
5. **Pour chaque clé du dictionnaire "clés - UMx", copier les UMx correspondants sur un slave et lancer les traitements en parallèle shuffling maps + reducing sorted maps (UMx -> SMx + SMx -> RMx). On peut alors construire le dictionnaire "RMx - machines"**
6. **Attendre la fin des traitements**
7. **Fusionner les RMx dans un fichier résultat final**
8. **Faire éventuellement une phase de tri sur les valeurs**