

1 Contexte et généralités

2 Principaux modèles de bases de données NoSQL

3 Fondements des systèmes NoSQL

- Partitionnement des données
- Réplication des données
- MapReduce
- Gestion des pannes

4 Travaux pratiques

MapReduce

- Conçu par Google labs
- Utilisé en interne par Google
- Se base sur des modèles de programmation parallèle
- Utilisé par Yahoo!, Amazon, ...
- Implémentations open source :
 - ▶ Hadoop, couchDB, mongoDB, ...

MapReduce est un paradigme de calcul distribué pour traiter des masses de données.

Il distribue les traitements sur les serveurs de données.

Traitement distribué

2 scénarios possibles

scénario 1

1 application dispose d'un jeu de données volumineux et souhaite appliquer un traitement disponible sur un site distant

Problème :

- acheminer les données sur le bon site

⇒ souvent architecture utilisant des web services

Client/Serveur

scénario 2

Le jeu de données est distribué sur plusieurs noeuds

⇒ le plus efficace est de déplacer les programmes

Data locality principle

Les meilleures performances sont obtenues quand chaque fragment de la collection est traité *localement*, minimisant ainsi les besoins d'échange entre les machines.

Caractéristiques

- permet de répartir la charge sur de nombreux serveurs
- abstrait l'infrastructure matérielle pour le programmeur

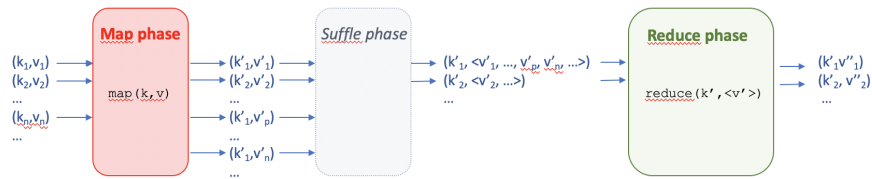
Travail du programmeur pour un calcul

se limite à l'écriture de 2 fonctions

- **Map**
 - ▶ fait des calculs élémentaires sur une paire (k, v) et retourne une liste de paires intermédiaires
- **Reduce**
 - ▶ prend toujours en entrée une paire intermédiaire $(k, list(v))$ et produit un résultat

Un programme complexe peut nécessiter plusieurs calculs.

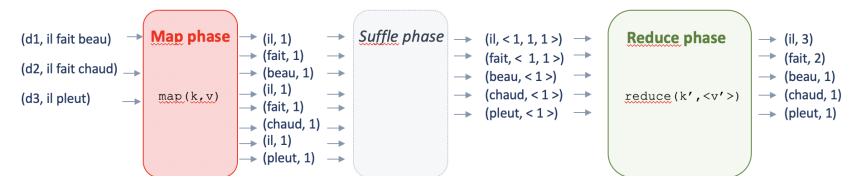
Principe MapReduce en ne tenant pas compte de l'aspect distribué



Exemple : Compter le nombre d'occurrence de chaque mot dans une liste de documents.

```
mapCptMot (String k, String v)
  for each w in v
    emitIntermediaire(w, 1)
```

```
reduceCptMot (String k, Iterator lv)
  result = 0
  for each v in lv
    result += v
  emitResult(result)
```



Exécution distribuée d'un traitement MapReduce

