Traitement du Big Data

MapReduce

Introduction de MapReduce

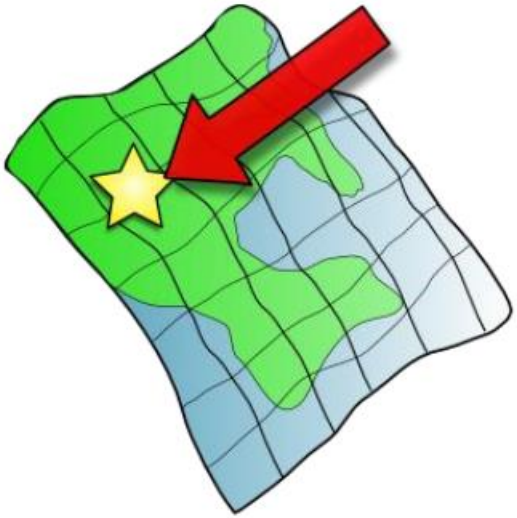
Cycle de vie des Big Data avec Hadoop

Collecte	Stockage	Traitement	Analyse	Exploration
Sqoop Flume	HDFS Hbase	**MapReduce** Hive, Pig	RHadoop Mahout Hive, Pig Spark Storm Solr	RHadoop Mahout Hive, Pig Spark Storm Solr

Objectifs de MapReduce

- Traiter de gros volumes de données en parallèle
- Equilibrer la charge sur le réseau
- Assurer la tolérance de panne
- Fonctionner sur des milliers de serveurs

Roles de Map et Reduce



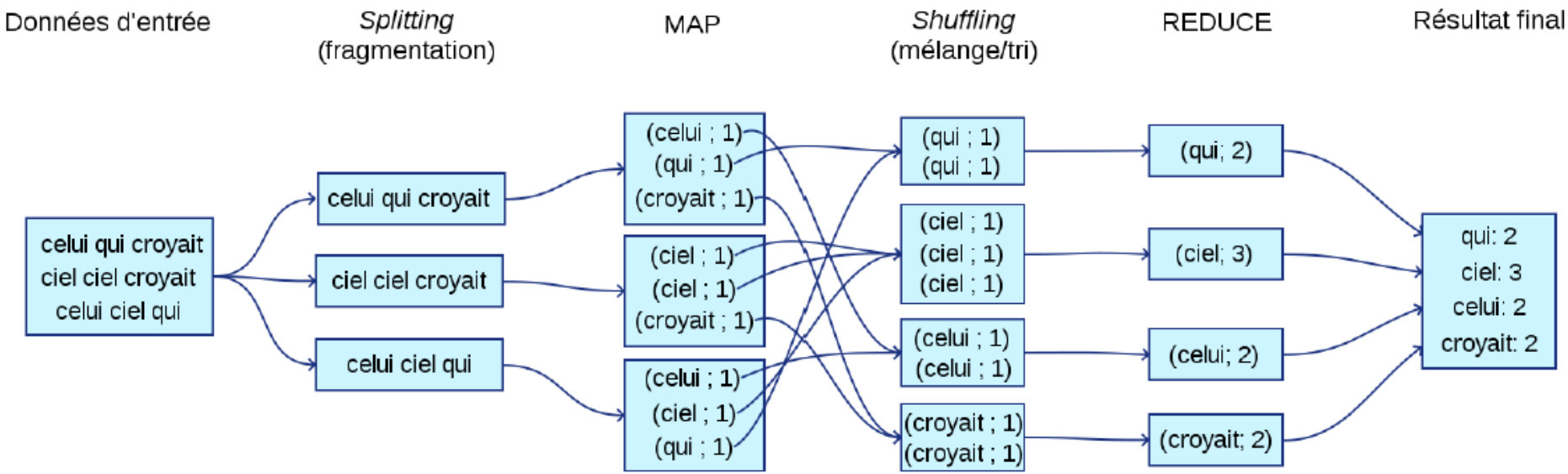
Map

Reduce

Extraire des informations sous forme clé / valeur

Agréger des informations ayant la même clé

Schéma de mise en œuvre de MapReduce



Exemple: Compter les mots d'un fichier texte

```
Celui qui croyait au ciel
Celui qui n'y croyait pas
[…]
Fou qui fait le délicat
Fou qui songe à ses querelles
```

(Louis Aragon, *La rose et le Réséda*, 1943, fragment)

Outils utilisés pour le comptage de mots

HDFS

1) Stocker dans HDFS

MapReduce

2) Split
3) Map
4) Shuffle
5) Reduce

1) Stocker dans HDFS

- Objectif: Compter les mots du fichier d'entrée ci-dessous

```
Celui qui croyait au ciel
Celui qui n'y croyait pas
[…]
Fou qui fait le délicat
Fou qui songe à ses querelles
```

(Louis Aragon, *La rose et le Réséda*, 1943, fragment)

- Stocker le fichier dans HDFS

2) Split

- Pour simplifier les choses, on va avant le découpage supprimer toute ponctuation et tous les caractères accentués. On va également passer l'intégralité du texte en minuscules.

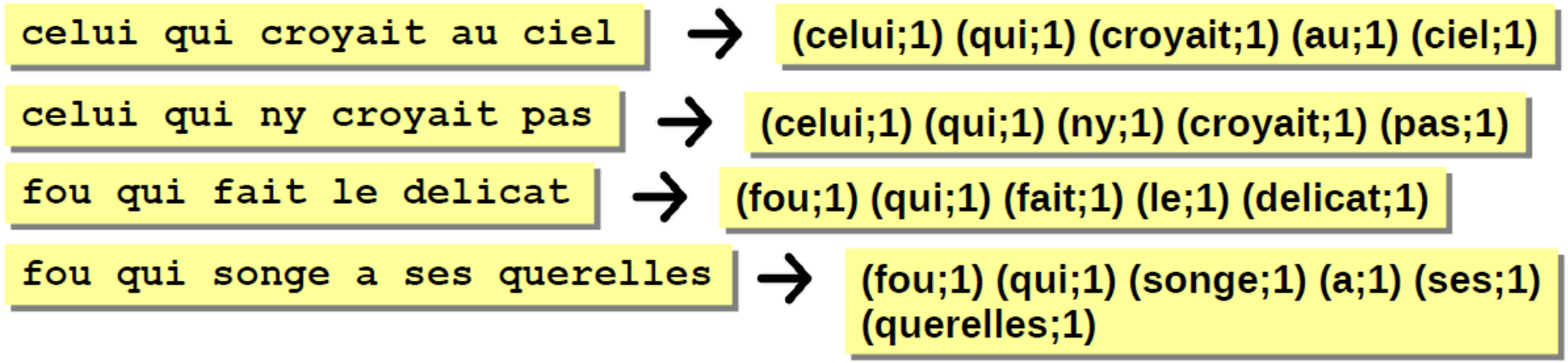
- Après découpage:

```
celui qui croyait au ciel
celui qui ny croyait pas
fou qui fait le delicat
fou qui songe a ses querelles
```

- On obtient 4 fragments avec nos données d’entrée.

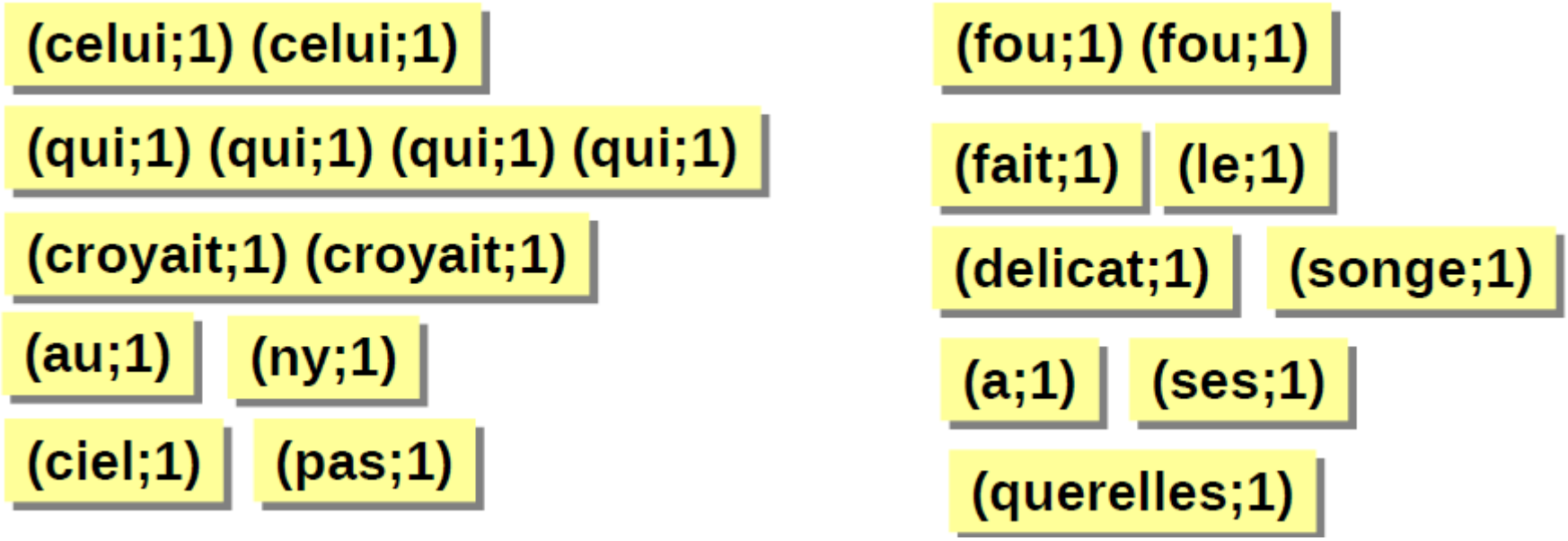
3) Map

- Pour chacun des fragments, l'opération MAP génère des couples (clef; valeur) :



4) Shuffle

- Grouper (shuffle) tous les couples par clef commune.
- Après son exécution, on obtient les 15 groupes suivants:



5) Reduce

- L’opération Reduce consiste à additionner toutes les valeurs liées à la clef spécifiée

- Une fois l'opération REDUCE effectuée, on obtiendra donc une valeur unique pour chaque clef distincte. En l’occurrence, notre résultat sera:

```
qui: 4
celui: 2
croyait: 2
fou: 2
au: 1
ciel: 1
ny: 1
pas: 1
fait: 1
[…]
```

Exercice de Map / Reduce

Comptage des mots

Exemple – Statistique Web

- On souhaite compter le nombre de visiteurs sur chacune des pages d'un site Internet.

- On dispose des fichiers de logs sous la forme suivante:

```
/index.html [19/Oct/2013:18:45:03 +0200]
/contact.html [19/Oct/2013:18:46:15 +0200]
/news.php?id=5 [24/Oct/2013:18:13:02 +0200]
/news.php?id=4 [24/Oct/2013:18:13:12 +0200]
/news.php?id=18 [24/Oct/2013:18:14:31 +0200]
...etc...
```

- Ici, notre clef sera par exemple l'URL d'accès à la page, et nos opérations MAP et REDUCE seront exactement les mêmes que celles qui viennent d'être présentées

- On obtiendra ainsi le nombre de vue pour chaque page distincte du site.

Merci