

Approches de gestion de données : des bases de données au big data

Mohand-Saïd Hacid



UNIVERSITÉ
LUMIÈRE
LYON 2



Historique (1/2)

- **60** : Fichiers, programmation (en cobol), rapports (*listings*).
- **65** : Beaucoup de fichiers (*synchronisation/cohérence, maintenance, développement, matériel...*)
- **70** : DAD
- **75** : Terminaux → OLTP
- **80** : OLTP + SAD
- **85** : OLTP + SAD + Extracteurs
- **90** : "attitude laissez-faire" (*crédibilité des données, productivité, difficulté à transformer les données en information*)

Historique (2/2)

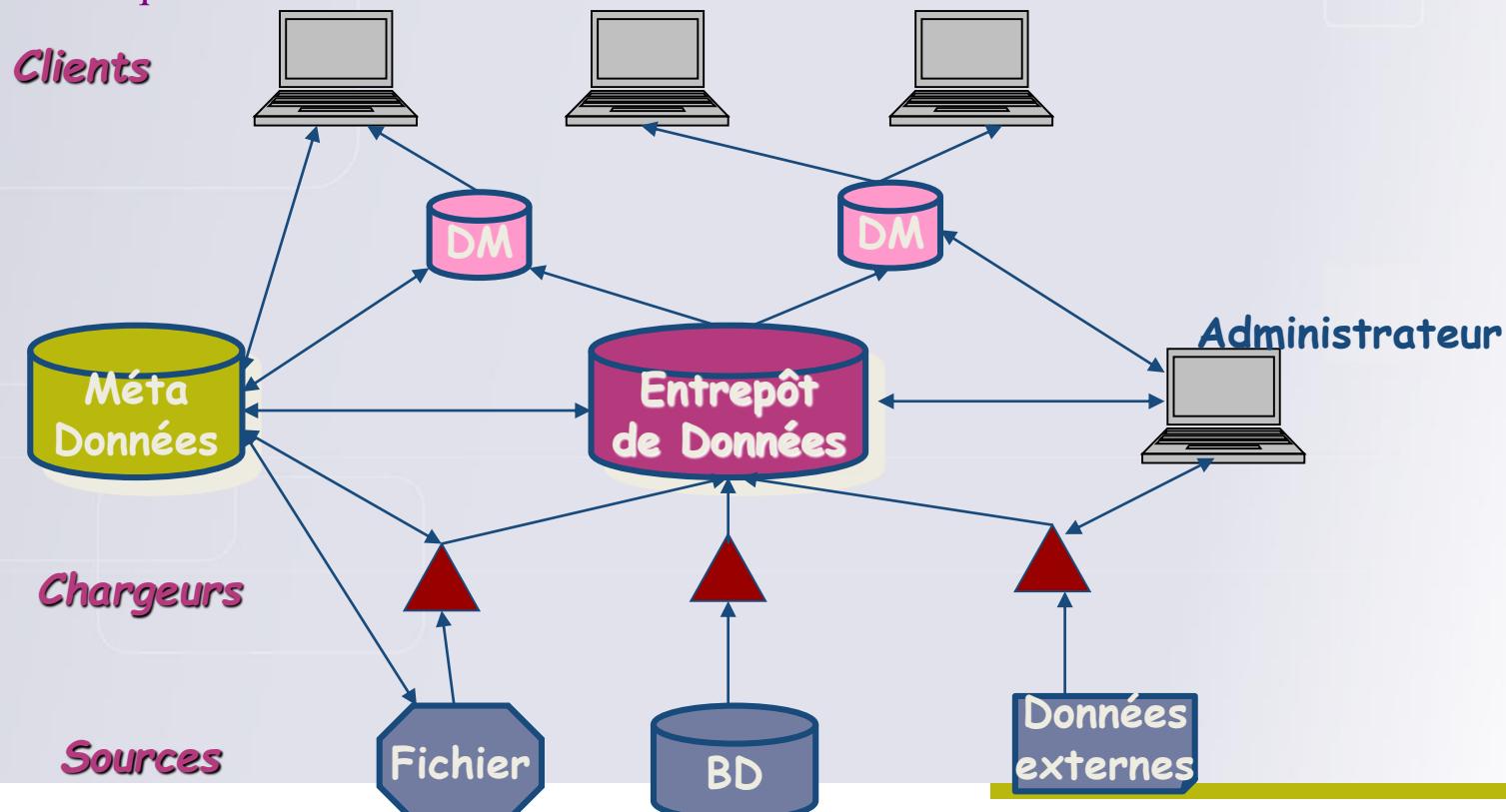
- **Les années 80 : puissance de calcul des ordinateurs**
- **Les années 90 : la révolution Internet (*mise en réseau des ordinateurs*) + Web 2.0 (*mise en réseau des humains*)**
- **Révolution de la donnée : intensification de nos pratiques en ligne et la massification des capteurs (ex.: téléphone portable)**

Passer des données à l'information

“Comparer les activités d'un compte d'une année sur l'autre pendant les cinq dernières années”

Les applications n'étaient pas construites dans un but d'intégration.

Pas assez de données historiques stockées



Gestion de données *avancée*

De la gestion de données traditionnelle à une gestion moderne

- Beaucoup de changements technologiques
- Réseaux
- Parallélisme (Clusters de calcul & Multicore)
- Systèmes de stockage
- Processeurs
- Nouvelles applications
 - Web
 - Données en flux
 - ...

Qu'est-ce que le Big Data ?

- Les big data représentent des ensembles d'informations volumineux et diversifiés qui se développent à un rythme *exponentiel*.
- Malheureusement, les volumes de données sont si importants qu'aucun des outils traditionnels de gestion de données ne peut les stocker ou les traiter efficacement.
- Plus que le volume de données, c'est la façon dont les organisations utilisent les données qui importe.
- Les big data peuvent être analysées pour en tirer des enseignements qui conduisent, pex, à de meilleures décisions et à des mouvements commerciaux stratégiques.

Exemples :

- La Bourse de New York crée environ un téraoctet de nouvelles données commerciales par jour.
- Les plateformes de médias sociaux contribuent également de manière importante à la multiplication des données.
- Les compagnies aériennes génèrent également de nombreux pétaoctets de données.
- ...

Big Data... mais pas seulement

Les dimensions « classiques » du BigData

● Volume

- Nombre de données, taille d'une donnée...
- Difficultés : acquisition, stockage, gestion, analyse...

● Vélocité

- Vitesse de production des données (flux de données).
- Difficultés : acquisition, traitement et analyse en ligne...

● Variété

- Hétérogénéité (type des données, format, sémantique)
- Difficultés : Intégration, analyse...

Mais aussi...

D'autres V pour le « Big Data »

● Validité des données

- Correction, complétude des données
- Difficultés : gestion, amélioration et analyse

● Vérité des réponses

- Correction des réponses (réponse exacte, incomplète, imprécise... précision/rappel).
- Difficultés : qu'est-ce qu'une réponse acceptable ?

● Variabilité

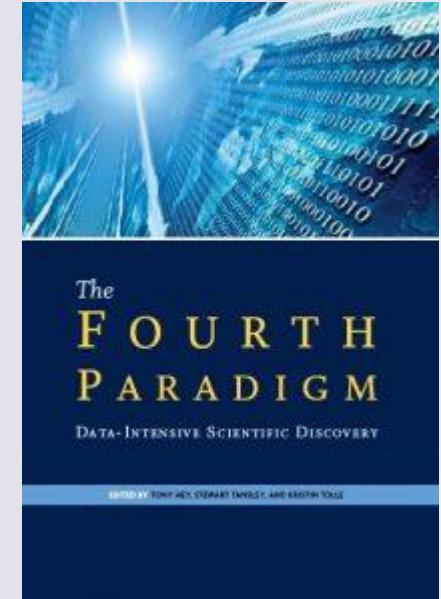
- Les données peuvent évoluer dans leurs format/schéma/sémantique, et les besoins aussi !
- Difficultés : solutions génériques, adaptables
- Vérification, Validation, Valeur, etc.

Plus de données:

- Nouvelles et meilleures solutions (pour des problèmes anciens!)
- Plus de précisions

Big Data

- Big data – Accélérateur d'innovation
- Big data – 4^{ème} paradigme
- Big data – une révolution technologique dans la capacité de collecte, de stockage et d'exploitation de données

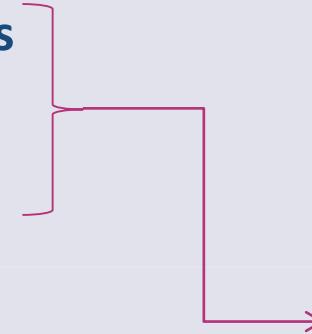


Big Data Prédicatif

Données générées par les objets connectés

+

Activités humaines



Volumes
Richesse

Analyse et corrélation : Poussées et Précises

Big Data

Applications :

Télécom, banques, assurances, distributeurs,
transporteurs, médical, sciences, ...

Contrôler les données :

Etre capable de les analyser (et de les posséder)

- Algorithmes/programmation
- Mathématiques/statistiques

Intégrer les données aux bilans des entreprises?

Big Data

Opportunité pour les internautes et les consommateurs

Découverte de services (nouveaux) → Innovation

Exemples:

- *Moteurs de comparaison des prix*
→ qui permettent d'acheter mieux et moins cher
- *Communication directe marques-consommateurs*
→ marques moins envahissantes

Big Data

Opportunité pour les entreprises

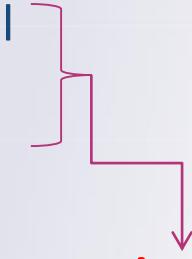
Exemples :

■ *Produits connectés*

Capter des informations sur l'utilisation de chaque produit  *améliorer la qualité, la durée de vie, anticiper les pannes, diagnostiquer rapidement, ...*

■ *Transformation des points de ventes et du rôle des vendeurs*

- Accéder à l'historique d'activités des clients en temps réel
- Accéder à des recommandations en temps réel

 Pour compléter un jugement

Technologie Big Data

- **Capteurs:** télescopes, caméras, IRM, puces ADN, individus, organisations, ...
- **Réseaux d'ordinateurs**
- **Supports de stockage**
 - Disque d'1 TB (~ 100 €)
 - Contenu des livres de la bibliothèque du congrès: 20 TB
- **Clusters d'ordinateurs (configuration matérielle choisie)**
 - Des milliers de nœuds (plusieurs disques et processeurs par machine)
 - Verrous algorithmiques

- Cloud
- **Algorithmes d'analyse de données**
 - Beaucoup de données → analyse (semi-)automatique
 - Difficulté: d'ordre algorithmique
 - Nouvelle forme de calcul
 - Combiner l'analyse statistique, optimisation et raisonnement

Défis!

- Infrastructure de gestion
- Distribution
- Optimisation de requêtes (cas des réseaux de capteurs) – traitement parallèle
- Indexation intelligente
- Traitement de flux de données
- Qualité des données (aspects *probabilistes, incomplétude ...*)
- Sémantique des données
- Visualisation et interaction intelligentes
- Outils d'analyse de données
- Intégration de calcul symbolique, de la fouille et de l'analyse
- ...

Complexité d'analyse à des échelles extrêmes

PB
↓

Générer des échantillons qui peuvent tenir en mémoire
plutôt que d'utiliser toutes les données

Outils statistiques

SAS, Excel, ...: utilisables seulement si des gros volumes de données sont réduits à des résumés pouvant tenir en mémoire

Conséquence: maintenir des schémas relationnels

normalisés et complexes peut s'avérer difficile et coûteux

Orientation ensembliste de SQL

Interfaces bas niveau ODBC/JDBC

Barrières pour les analystes

pour utiliser les bases de données

Nouveaux langages et modèles qui traduisent *naturellement* l'intention des analystes!

Analyses deviennent plus complexes

Reproductibilité de **workflows analytiques** et leurs résultats devient très important!



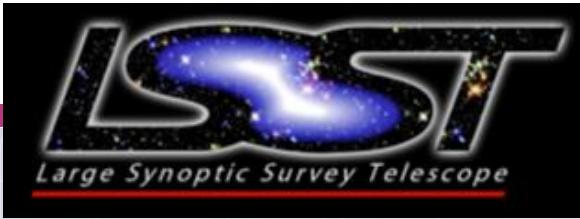
Exemple 1

Défis en gestion de données dans les relevés astronomiques

Le cas du LSST
<http://www.lsst.org/>

«La révolution du LSST réside dans le fait que cet instrument au design étonnant va révéler un univers exhaustif en mouvement. Des astéroïdes aux noyaux actifs des plus lointaines galaxies, en passant par les étoiles variables, les diverses occultations et les supernovae ou les sursauts gamma, tout ce qui changera dans le ciel sera enregistré tous les 3 jours... »

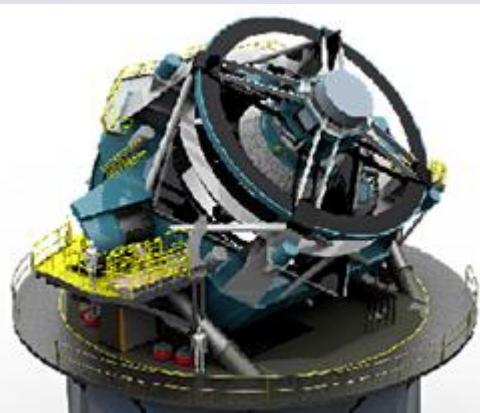
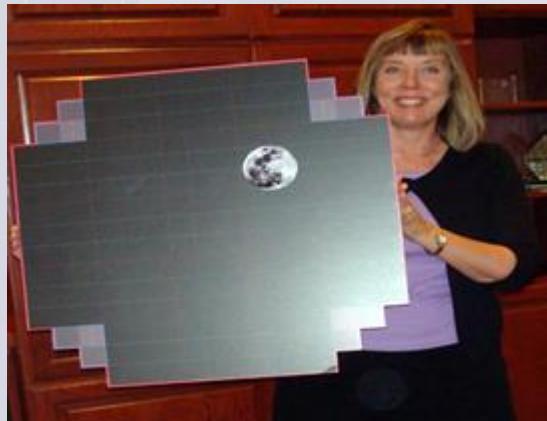
http://lsst.in2p3.fr/objectifs_scientifiques.html



en une diapo

■ Une nouvelle fenêtre sur le ciel :

- Télescope de 8,4 m
- Astronomie très grand champ :
 - caméra 9,6°
 - Cerro Pachon (Chili)



- 1 visite / 3 jours
- **10 ans, 60 PB de données**

Opportunités

Science

- Améliorer les découvertes
- Qualité des données
- Validation
- ...

Expérimentation -> Analyse->publication

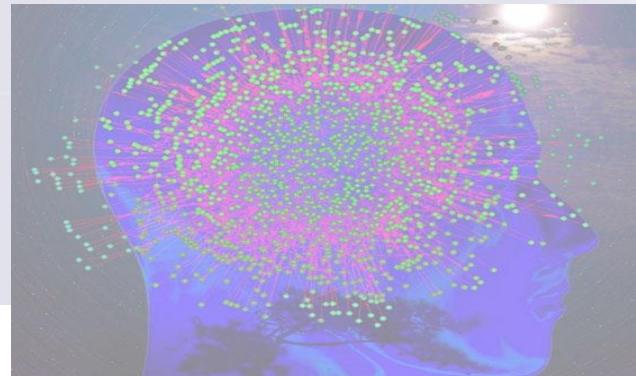


Expérimentation-> **organization des données**->Analyse->Publication

La science utilise
l'informatique pour
améliorer les
découvertes



L'informatique
permet de
découvrir



Les données LSST

Table	Taille	#enregistrements	#colonnes (arité)
Object	109 TB	38 B	470
Moving Object	5 GB	6 M	100
Source	3.6 PB	5 T	125
Forced Source	1.1 PB	32 T	7
Difference Image Source	71 TB	200 B	65
CCD Exposure	0.6 TB	17 B	45

Type de Requêtes LSST

1. Petites (quelques secondes)

Exemple: Récupérer tout type d'information sur un seul objet (identifié par un objectId).

```
SELECT * FROM Object WHERE objectId = 293848594;
```

2. Moyennes (environ 1 minute)

Exemple: Récupérer n'importe quel type d'information sur un groupe d'objets dans une petite zone spatiale

```
SELECT * FROM Object WHERE qserv_areaSpec_circle(1.0, 35.0, 5.0/60);
```

3. Coûteuses (environ 1 heure)

Exemple: Analyser tous les objets et appliquer un filtre sur un certain nombre d'attributs

```
SELECT MAX(scisql_fluxToAbMag(rFluxGaussian)) FROM Object WHERE  
rNumObs >= 5;
```

Type de Requêtes LSST

4. Très coûteuses (environ 1 jour)

Exemple: L'analyse des courbes de lumière à travers une grande zone spatiale

```
SELECT O.objectId, myFunction(S.taiMidPoint, S.psfFlux) FROM Object AS O  
JOIN Source AS S USING (objectId) WHERE O.varProb > 0.75 GROUP BY  
O.objectId;
```

5. Impossibles

Exemple: Une simple opération de tri sur tous les objets

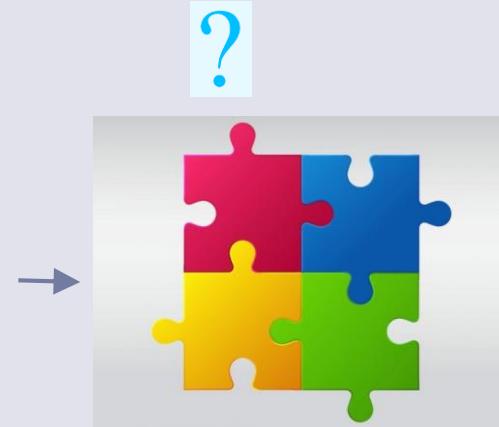
- 10 Peta => 6 h et 27 min avec 8000 machines (*google research*)
- LSST sera équipé de seulement 150 machines

```
SELECT * FROM Object ORDER BY rGaussianFlux DESC
```

- Liste complète des requêtes: <http://dev.lsstcorp.org/trac/wiki/dbQueries>
- Défis LSST :
 - ½ million de requêtes par jour
 - ~50 requêtes simples et ~20 requêtes complexes à n'importe quel moment

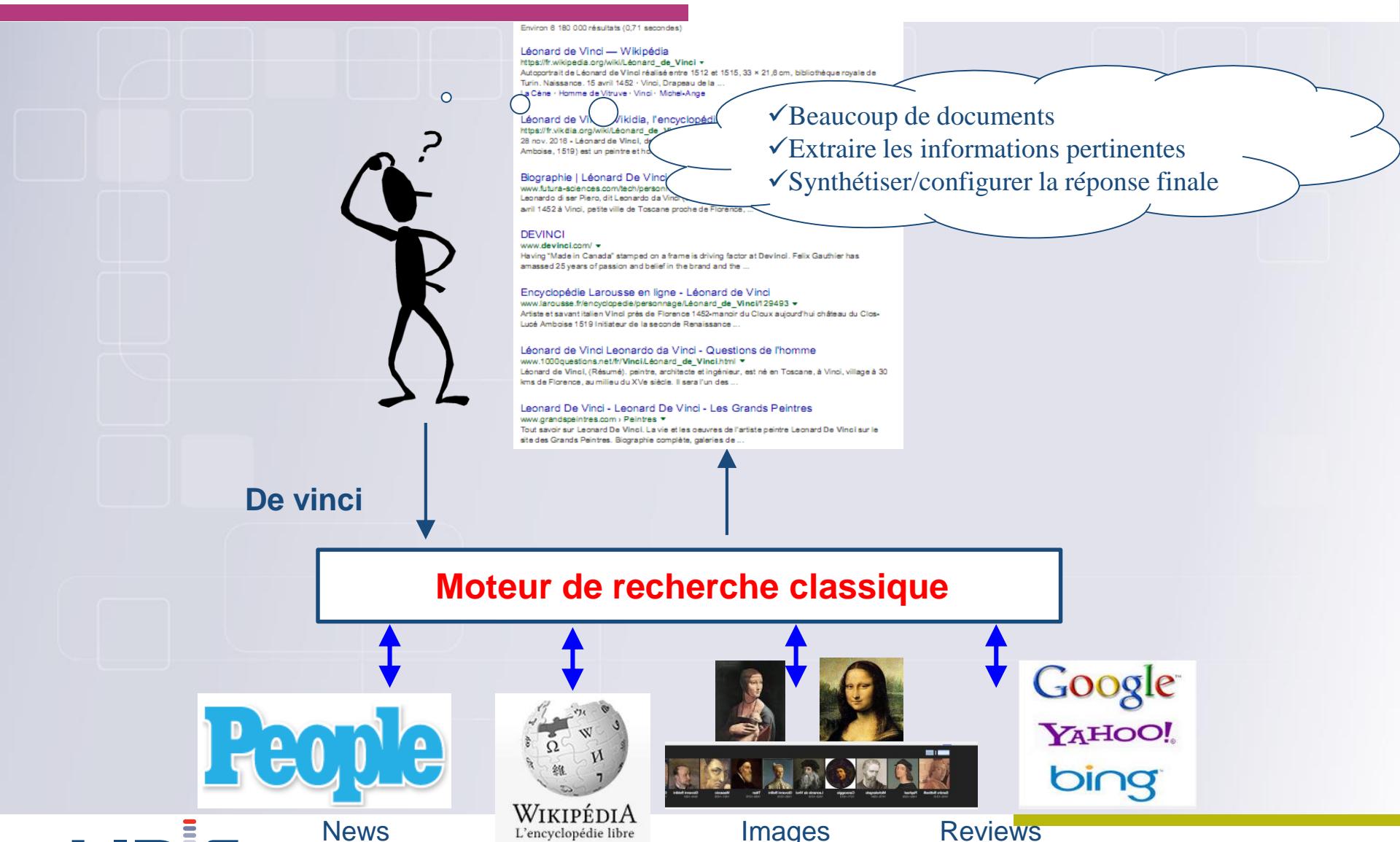
Exemple 2

Agrégation (Synthèse) d'information



Compléxité de la sélection et de l'assemblage des agrégats

Agrégation (Synthèse) d'information



Agrégation (Synthèse) d'information



Agrégation (Synthèse) d'information

Synthèse d'information autour d'un événement, d'une entité



Générer une synthèse

Microsoft

Microsoft Corporation est une entreprise américaine, fondée par Bill Gates et Paul Allen

... En 1975



Publié le 27-11-2015
Le 12 janvier 2016, Microsoft va arrêter complètement le support technique de toutes les versions d'Internet Explorer, à l'exception de la onzième et dernière.

Publié le 01-12-2015
A l'occasion du premier jour de Convergence 2015 EMEA, qui se déroule jusqu'au 2 décembre à Barcelone, Microsoft a dévoilé de nouveaux services et applications ainsi que la disponibilité d'une offre Premium d'Office 365.



27-11-2015 : Microsoft stoppera le support de tous les vieux Internet Explorer à partir du 12 janvier 2016.
01-12-2015 : A l'occasion du premier jour de Convergence 2015 EMEA, Microsoft présente un Office 365 enrichi.



Requête

?Date

Creation

5

Record

3
Type

?Album_title

?Album

3
Title

1



?Collection

1

hasPhotoCollection

creatorOf

?Composer

3

RDF Book Mashup

2



title

?title

1

1

RelatedBook

type

film

1

?book

?ISBN

2
identifier

4



?Writer

4
birthPlace

Occupation

4

Novelist

France

28

5



Agrégation (Synthèse) d'information



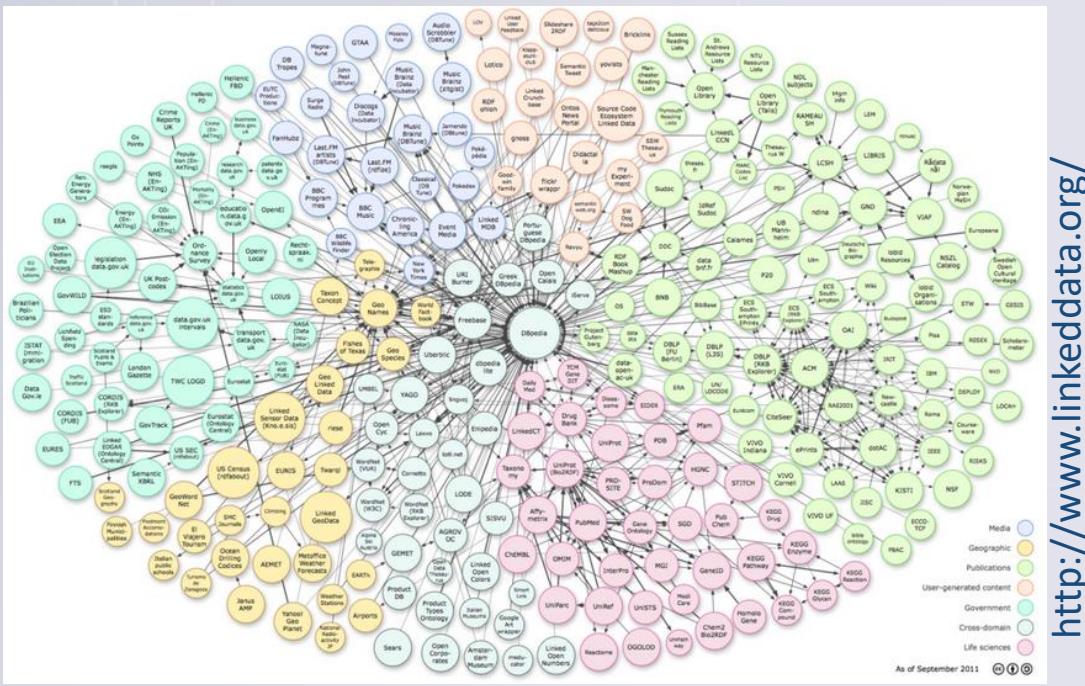
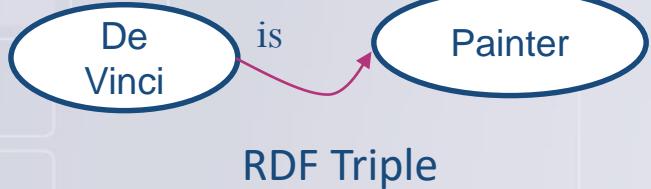
Défis

Sémantique

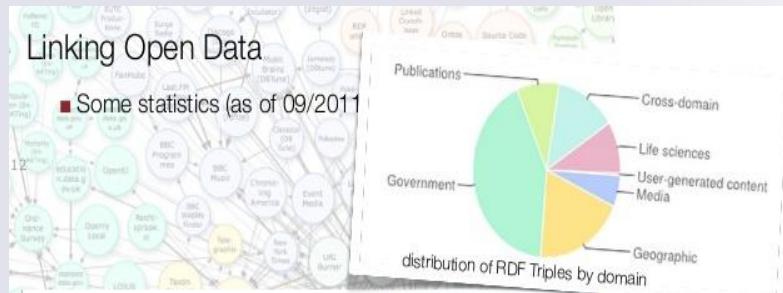
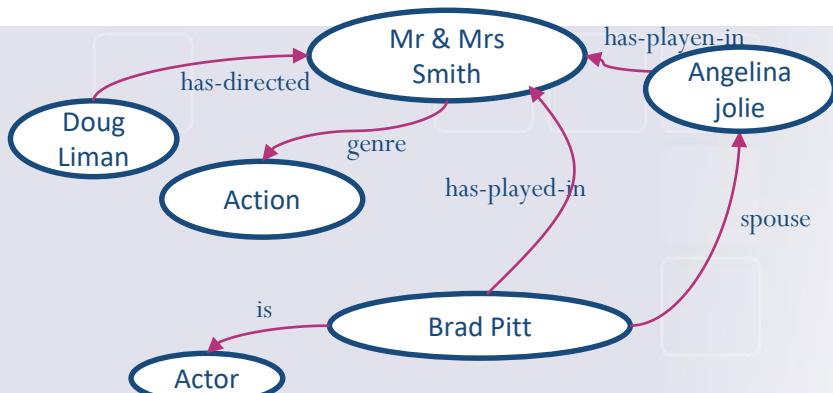
- *Interprétation* et refomulation de la requête
- Définition des *propriétés* (critères) permettant l'assemblage des objets
- Evaluation de la *qualité* des agrégats produits
- Calculatoire (problème lié à la combinatoire)
 - Choix des fragments à agréger
 - Choix des *combinations* à considérer (multiples façons de combiner les objets)

RDF - Resource Description Framework

(<Sujet>, <Prédicat>, <Objet>)



<http://www.linkeddata.org/>



Domain	Number of datasets	Triples	%	(Out-)Links	%
Media	25	1,841,852,061	5.82 %	50,440,705	10.01 %
Geographic	31	6,145,532,484	19.43 %	35,812,328	7.11 %
Government	49	13,315,009,400	42.09 %	19,343,519	3.84 %
Publications	87	2,950,720,693	9.33 %	139,925,218	27.76 %
Cross-domain	41	4,184,635,715	13.23 %	63,183,065	12.54 %
Life sciences	41	3,036,336,004	9.60 %	191,844,090	38.06 %
User-generated content	20	134,127,413	0.42 %	3,449,143	0.68 %
	295	31,634,213,770		503,998,829	

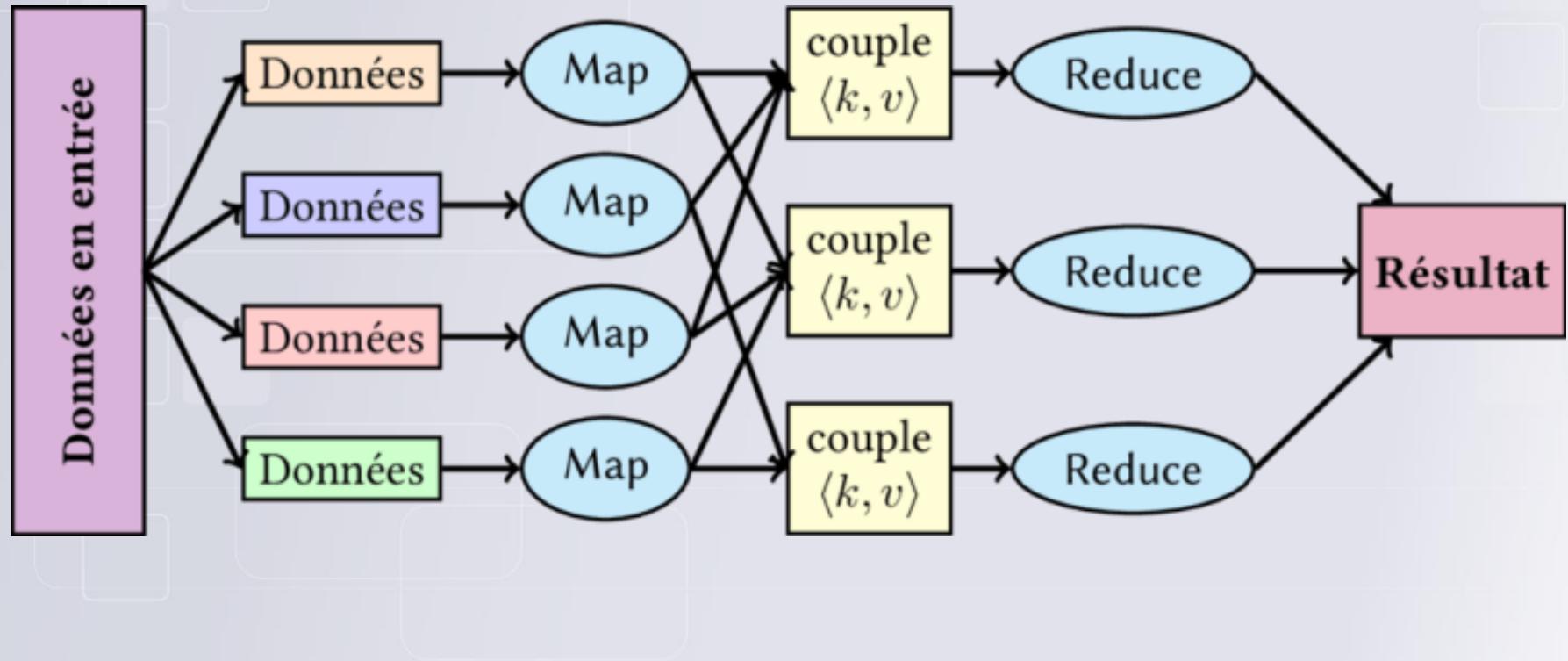
MapReduce

MapReduce ou Hadoop MapReduce est un modèle de programmation qui sert à calculer de gros volumes de données en parallélisant les calculs sur différents nœuds d'un cluster. On parle de calculs distribués.

Plus concrètement, MapReduce consiste, en deux fonctions :

- **Map()** pour distribuer le travail sur les nœuds du cluster
- **Reduce()** pour agréger le résultat de chaque nœud en un unique résultat

MapReduce



MapReduce (1/2)

On spécifie deux fonctions:

map (in_key, in_value) -> list(out_key, intermediate_value)

- Traiter les <clé, valeur> en entrée
- Produire un ensemble de <clé, valeur> comme résultat intermédiaire

reduce (out_key, list(intermediate_value)) -> list(out_value)

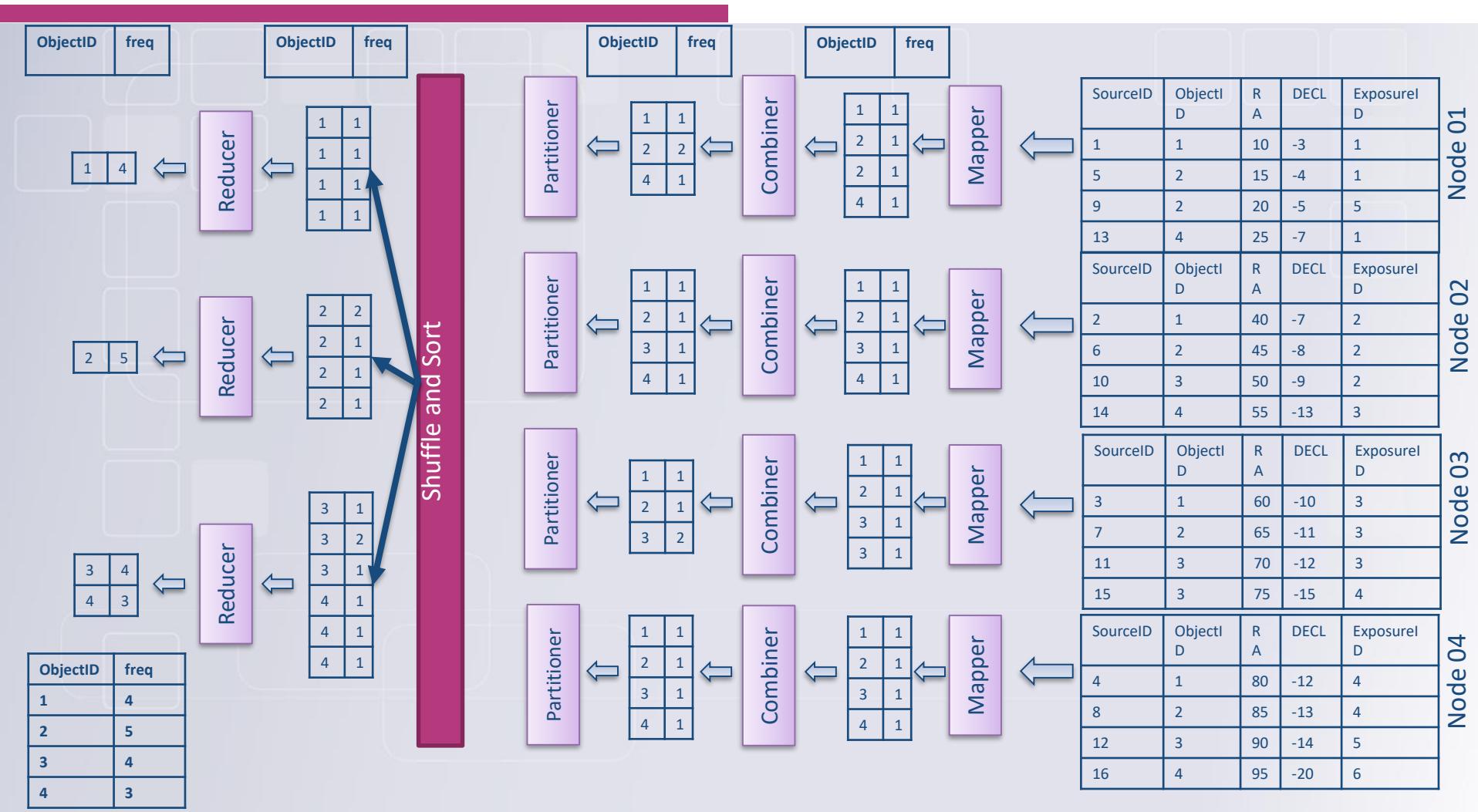
- Combiner toutes les valeurs intermédiaires avec une clé particulière
- Produire un ensemble de valeurs de sortie

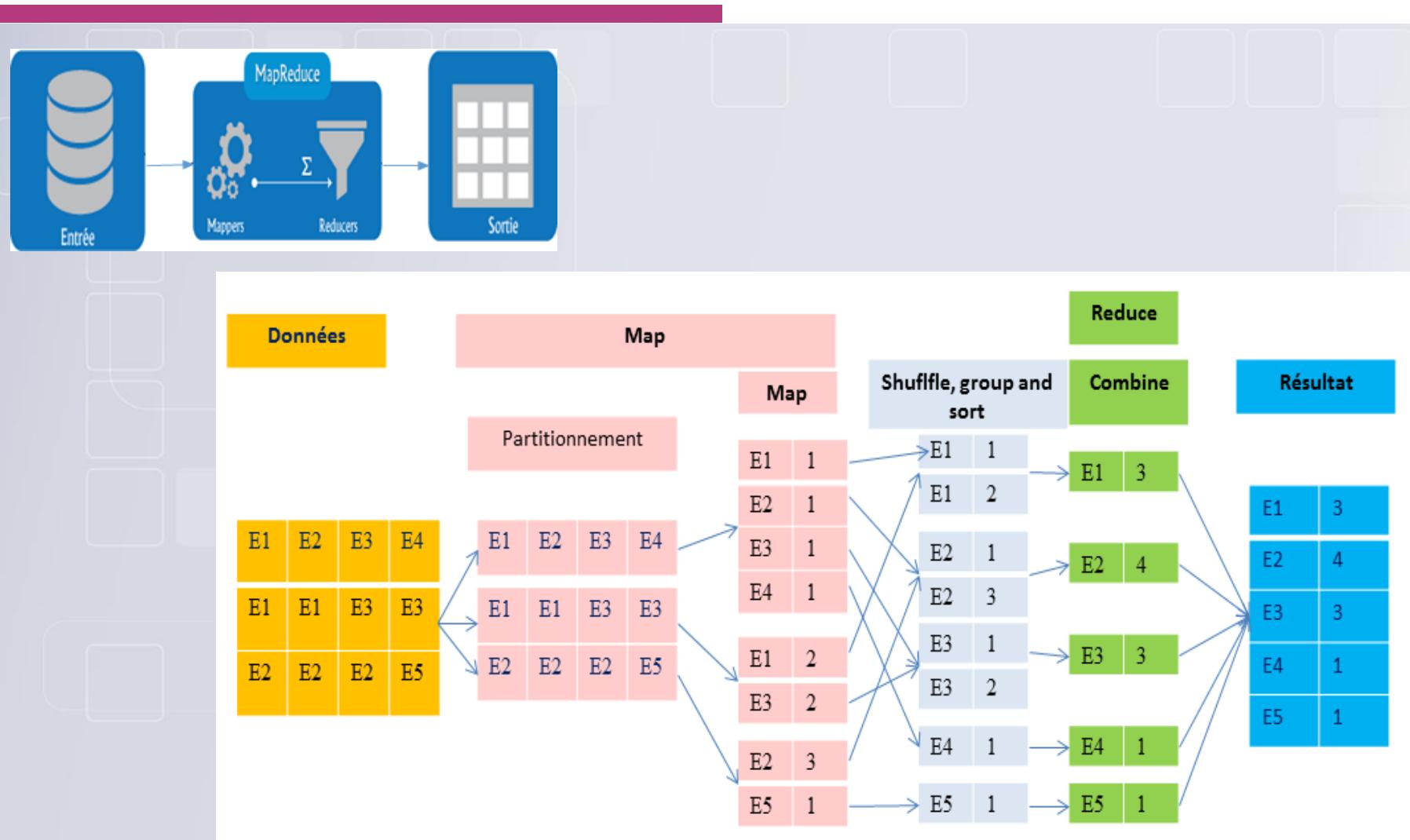
MapReduce (2/2)

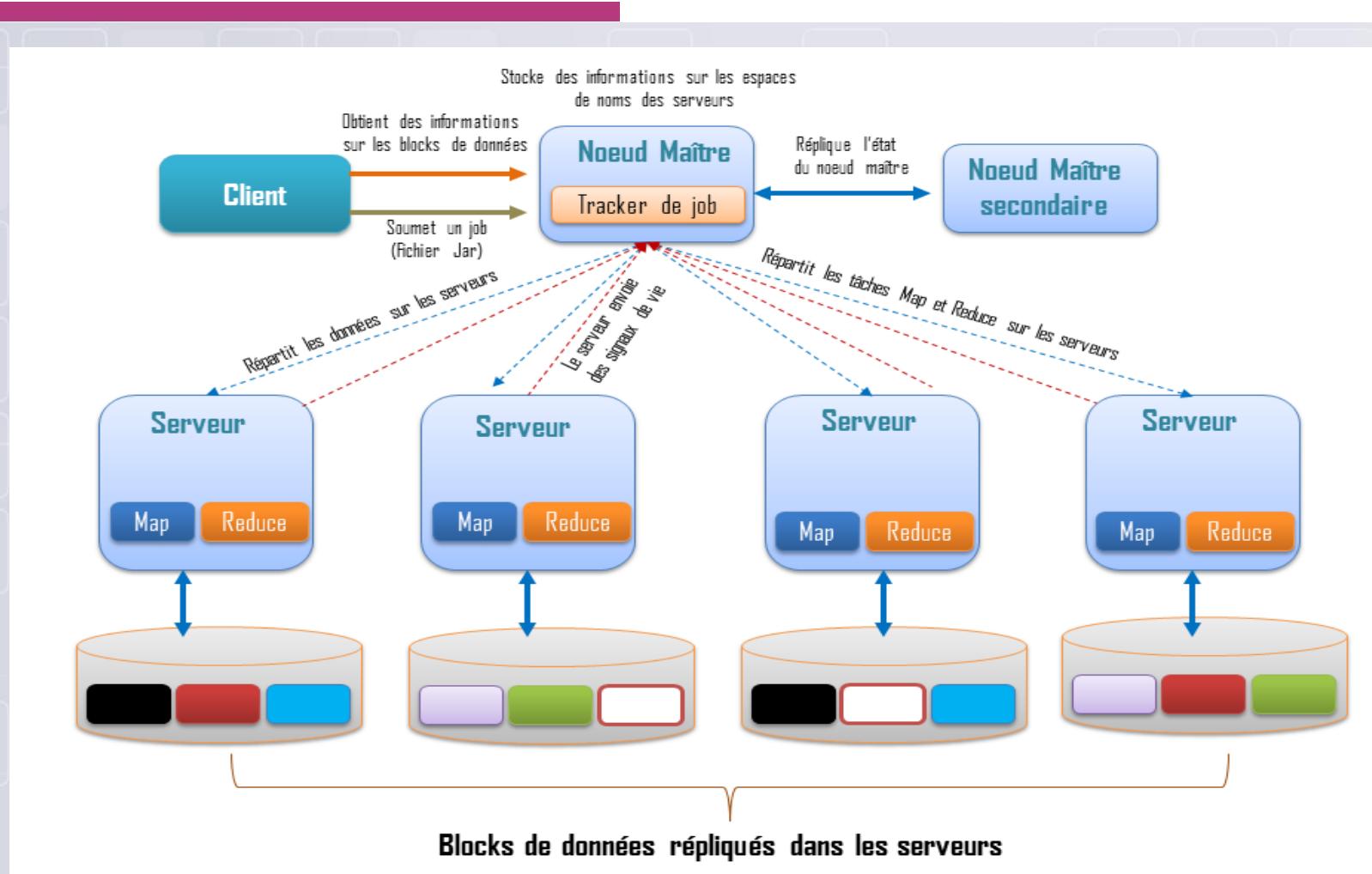
```
SELECT ObjectID,  
       count(SourceID) as freq  
  FROM Source  
 GROUP BY objectID
```

ObjectID	freq
1	4
2	5
3	4
4	3

SourceID	ObjectID	RA	DECL	ExposureID
1	1	10	-3	1
5	2	15	-4	1
9	2	20	-5	5
13	4	25	-7	1
2	1	40	-7	2
6	2	45	-8	2
10	3	50	-9	2
14	4	55	-13	3
3	1	60	-10	3
7	2	65	-11	3
11	3	70	-12	3
15	3	75	-15	4
4	1	80	-12	4
8	2	85	-13	4
12	3	90	-14	5
16	4	95	-20	6



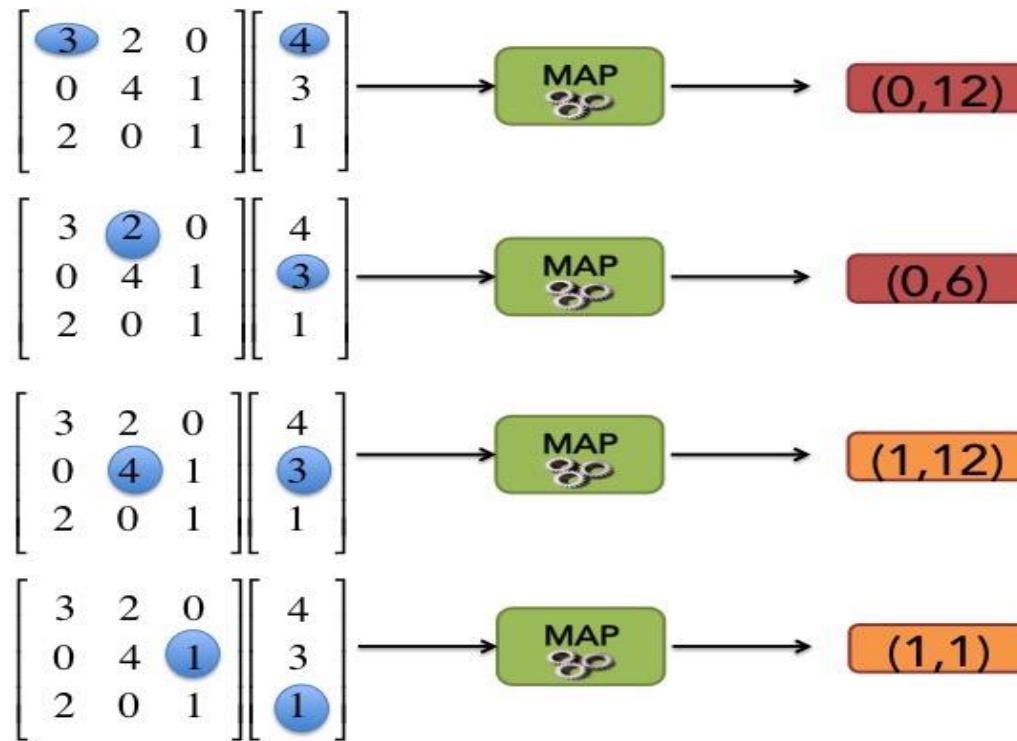


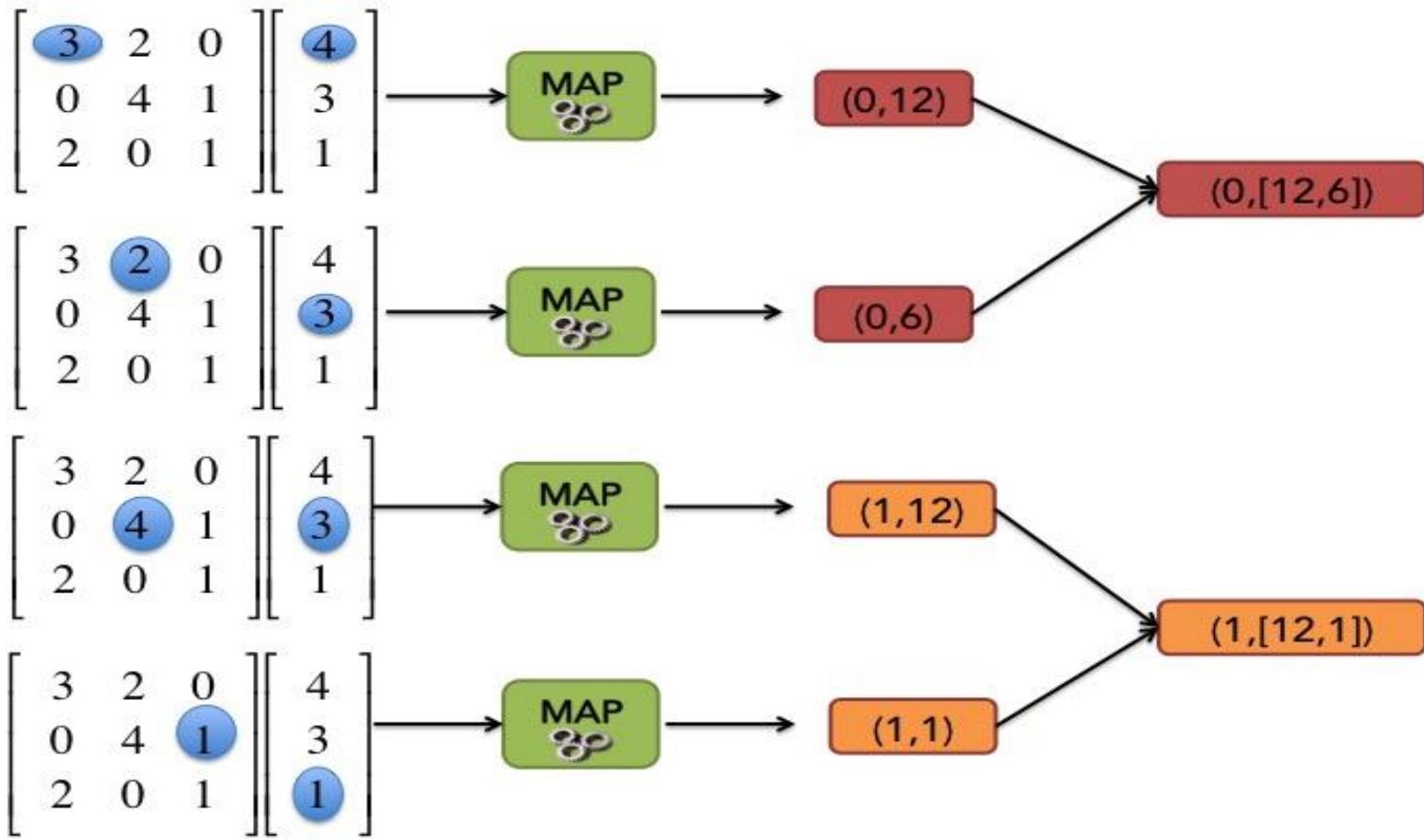


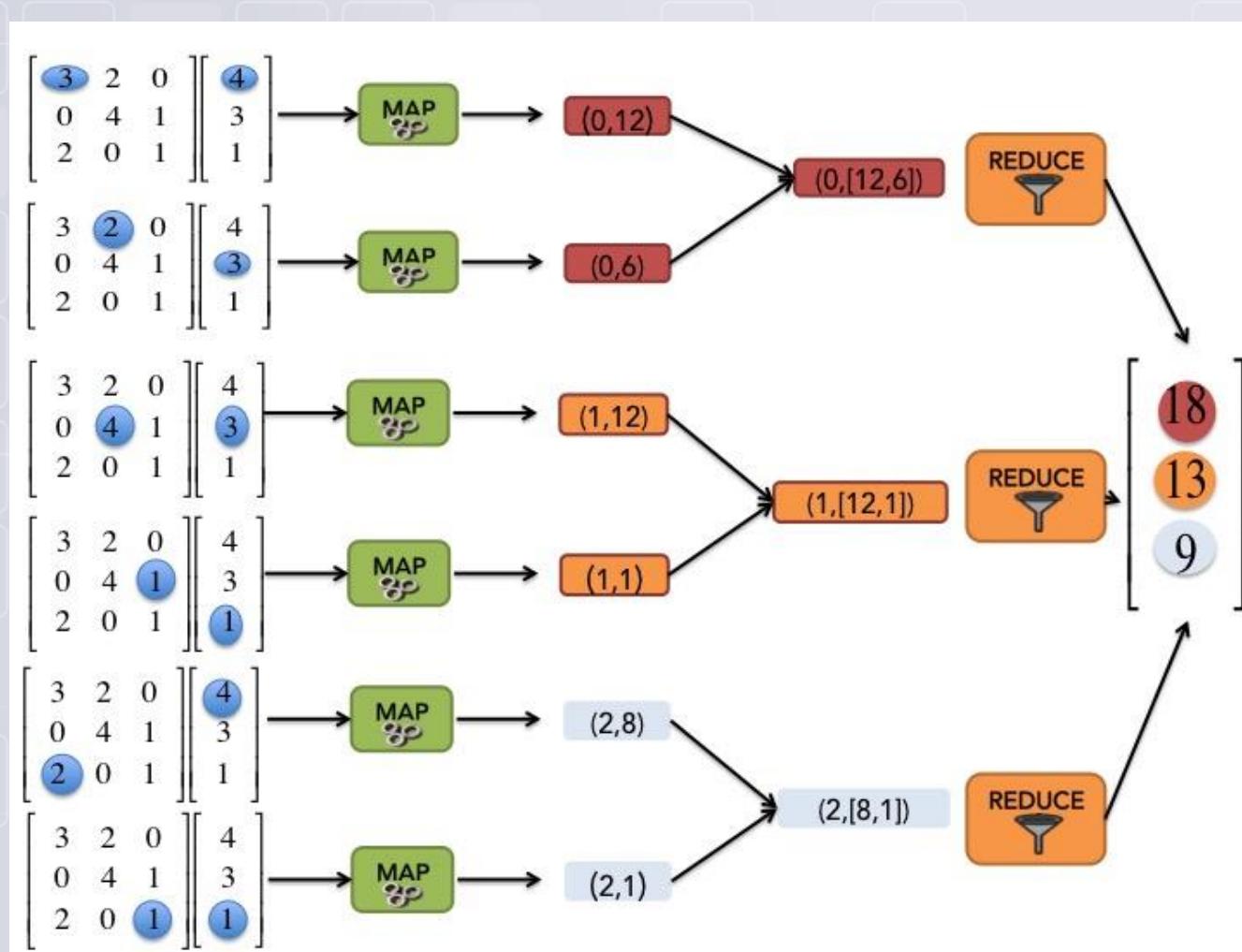
Exemple 1

Produit d'une matrice A par un vecteur V

Cas 1 : le vecteur V peut entièrement tenir en mémoire des nœuds MAP







Cas 2 : le vecteur **V** est trop gros pour tenir entièrement en mémoire des nœuds **MAP**. Il faut alors appliquer le principe diviser pour régner. Il faut découper le vecteur **V** en bandes horizontales (qui tiennent en mémoire) et faire de même mais verticalement pour la matrice **A**. Le problème initial est ainsi découpé en sous-tâches et on affecte à chaque nœud **MAP** un morceau de la matrice et la bande de vecteur correspondante.

Exemple 2

Jointure de deux tables

```
1 SELECT *  
2 FROM Films F JOIN Realisateurs R  
3 ON F.ID_realisateur=R.ID_realisateur
```

ID_film	Titre	ID_realisateur	ID_acteur	...
1111	Pulp Fiction	123	23	
1112	Le pianiste	4567	678	
1113	La leçon de piano	234	567	
...	

Films

jointure

ID_realisateur	Nom
123	Quentin Tarantino
4567	Roman Polanski
234	Jane Campion
...	...

Réaliseurs

ID_realisateur	Titre
123	Pulp Fiction
4567	Le pianiste
234	La leçon de piano
123	Reservoir dogs
...	...

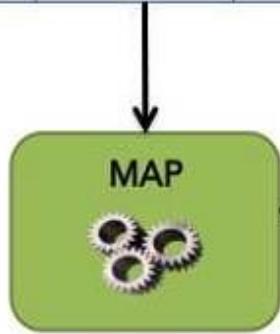
Films

ID_réalisateur	Nom
123	Quentin Tarentino
4567	Roman Polanski
234	Jane Campion
...	...

Réaliseurs

Films	123	Pulp Fiction
Films	4567	Le pianiste
Films	234	La leçon de piano
Films	123	Reservoir dogs
Réaliseurs	123	Q. Tarantino
Réaliseurs	4567	R. Polanski
Réaliseurs	234	J. Campion
...

Films	123	Pulp Fiction
Films	4567	Le pianiste
Films	234	La leçon de piano
Films	123	Reservoir dogs
Réaliseurs	123	Q. Tarantino
Réaliseurs	4567	R. Polanski
Réaliseurs	234	J. Campion
...



(123, (Films, Pulp Fiction))
 (4567, (Films, Le pianiste))
 (234, (Films, La leçon de piano))
 (123, (Films, Reservoir dogs))
 (123, (Réaliseurs, Q. Tarantino))
 (4567, (Réaliseurs, R. Polanski))
 (234, (Réaliseurs, J. Campion))

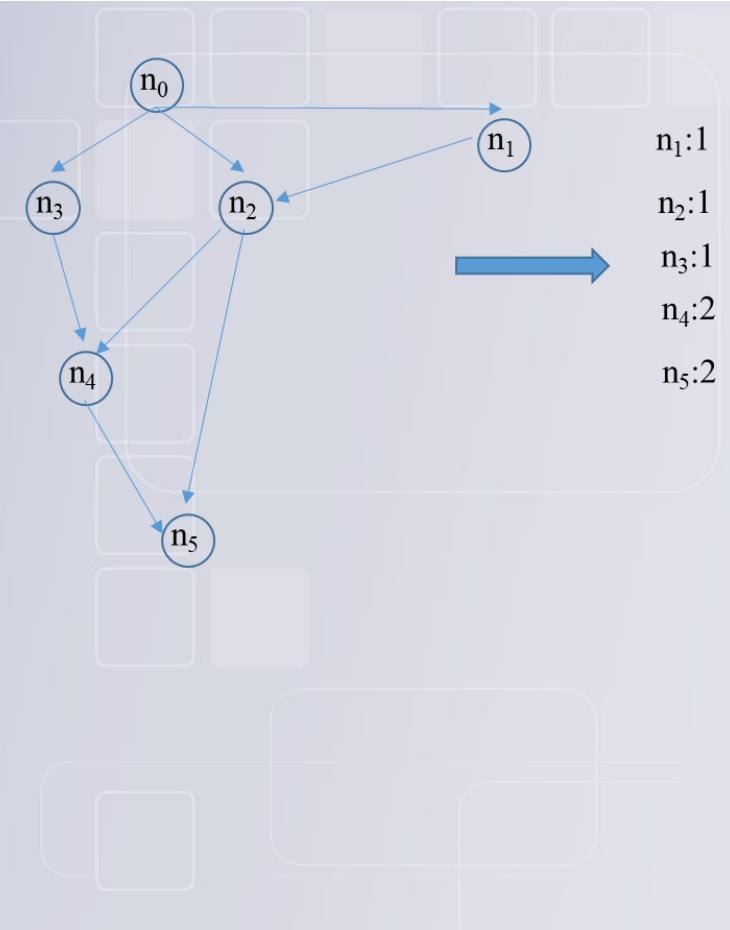
(123, [(Films, Pulp Fiction), (Films, Reservoir dogs),
(Réalisateur, Q. Tarantino)])



ID_réalisateur	Nom	Titre Films
123	Quentin Tarantino	Pulp Fiction
123	Quentin Tarantino	Reservoir dogs

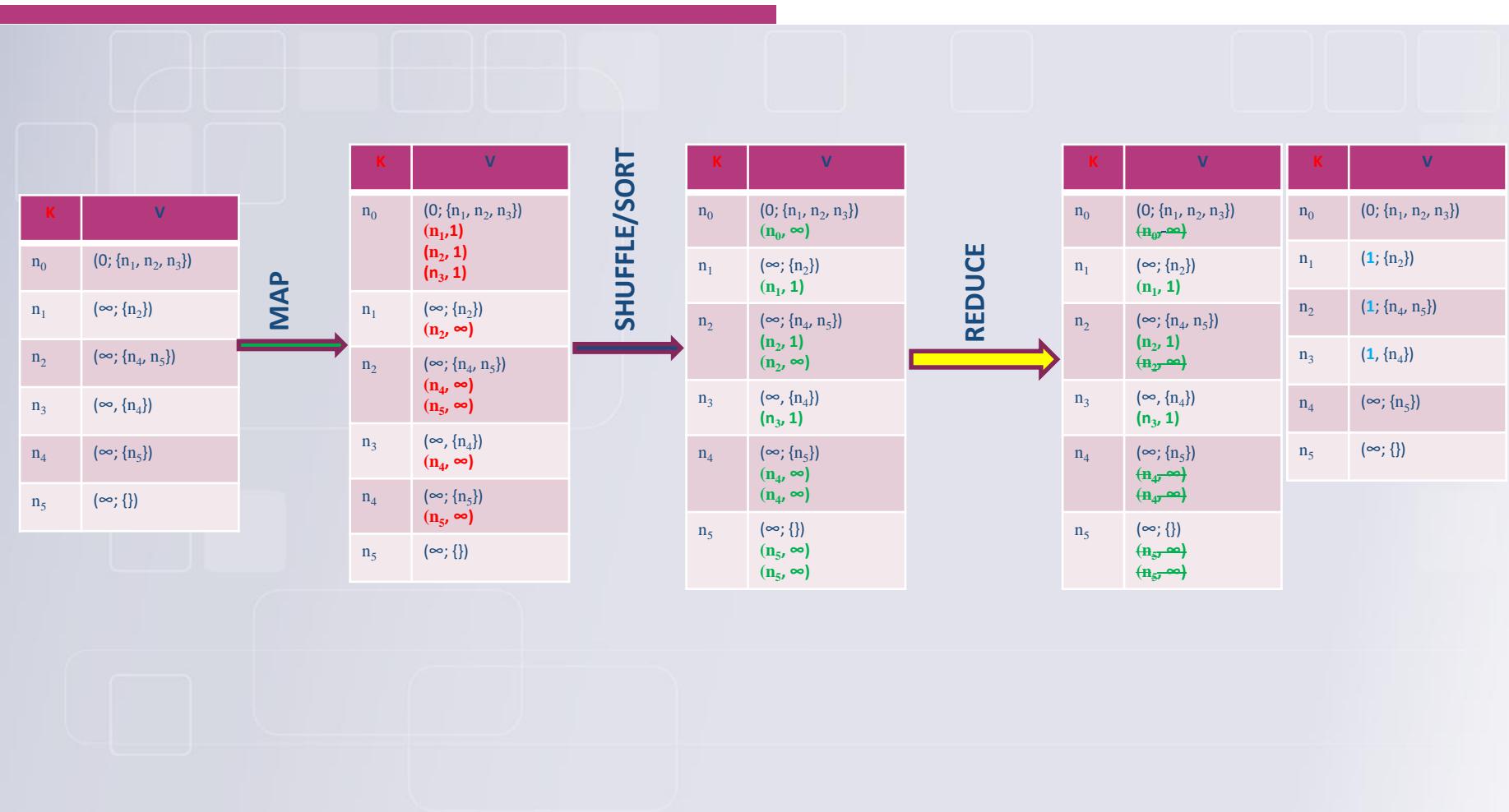
Exemple 3

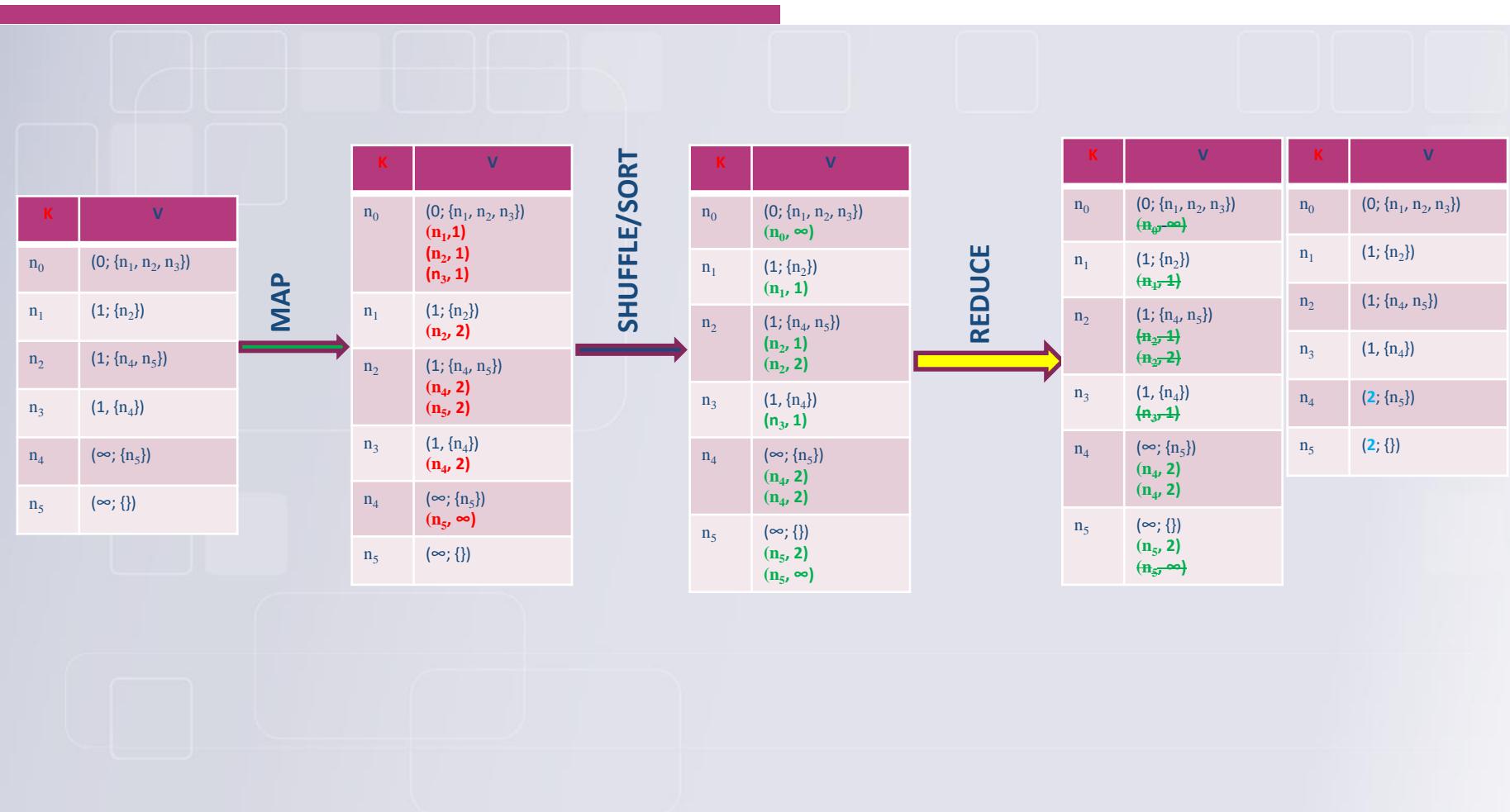
Calcul des plus courts chemins, par rapport à un nœud, dans un graphe



n₁:1
n₂:1
n₃:1
n₄:2
n₅:2

K	V
n ₀	(0; {n ₁ , n ₂ , n ₃ })
n ₁	(∞; {n ₂ })
n ₂	(∞; {n ₄ , n ₅ })
n ₃	(∞; {n ₄ })
n ₄	(∞; {n ₅ })
n ₅	(∞; {})





C'est tout pour aujourd'hui
MERCI