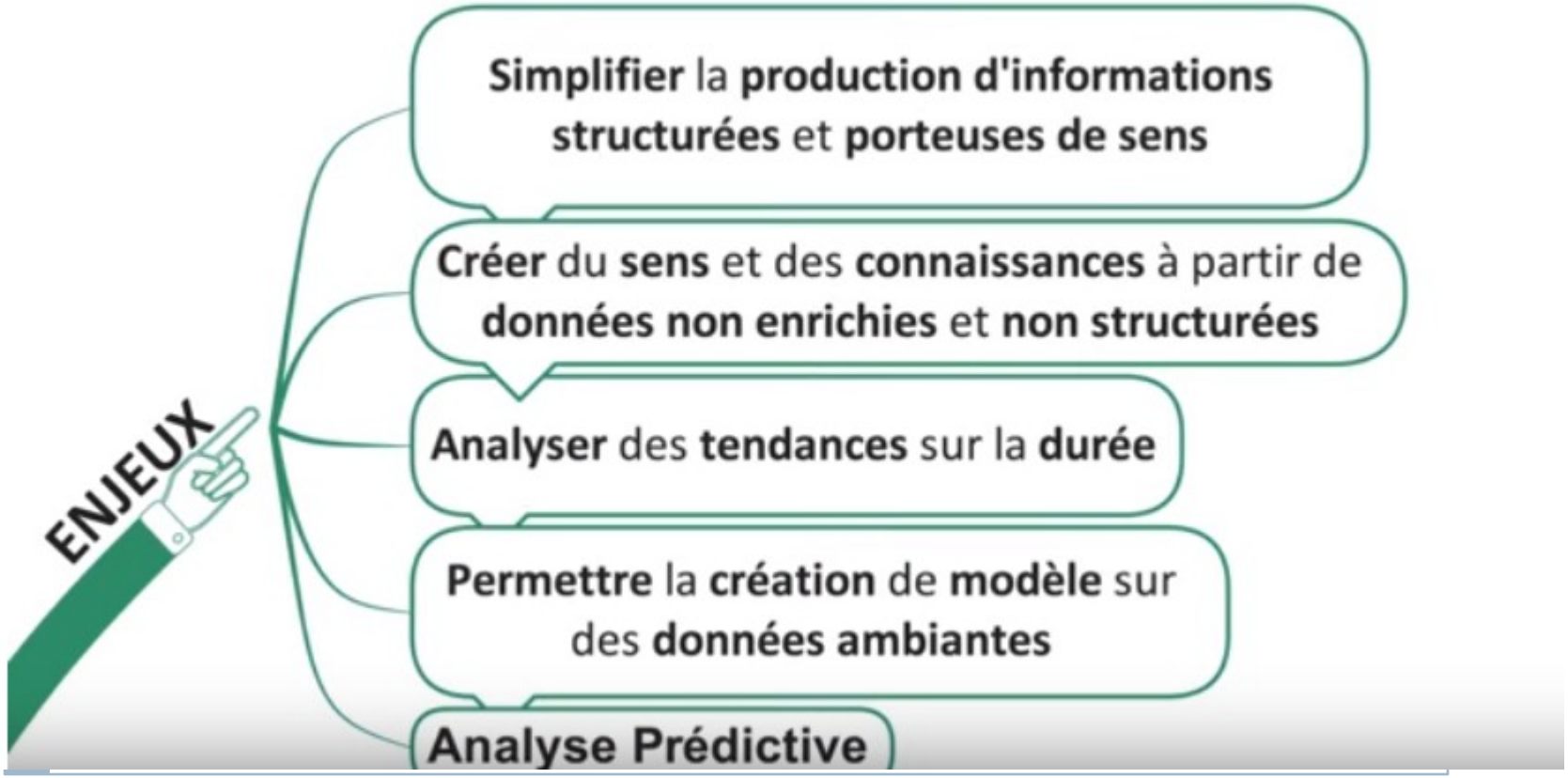
# Swarm Intelligence en Big Data

## Master M2: GADM

Cours Présenté par : Mme Mohamed Ben Ali

# Big Data

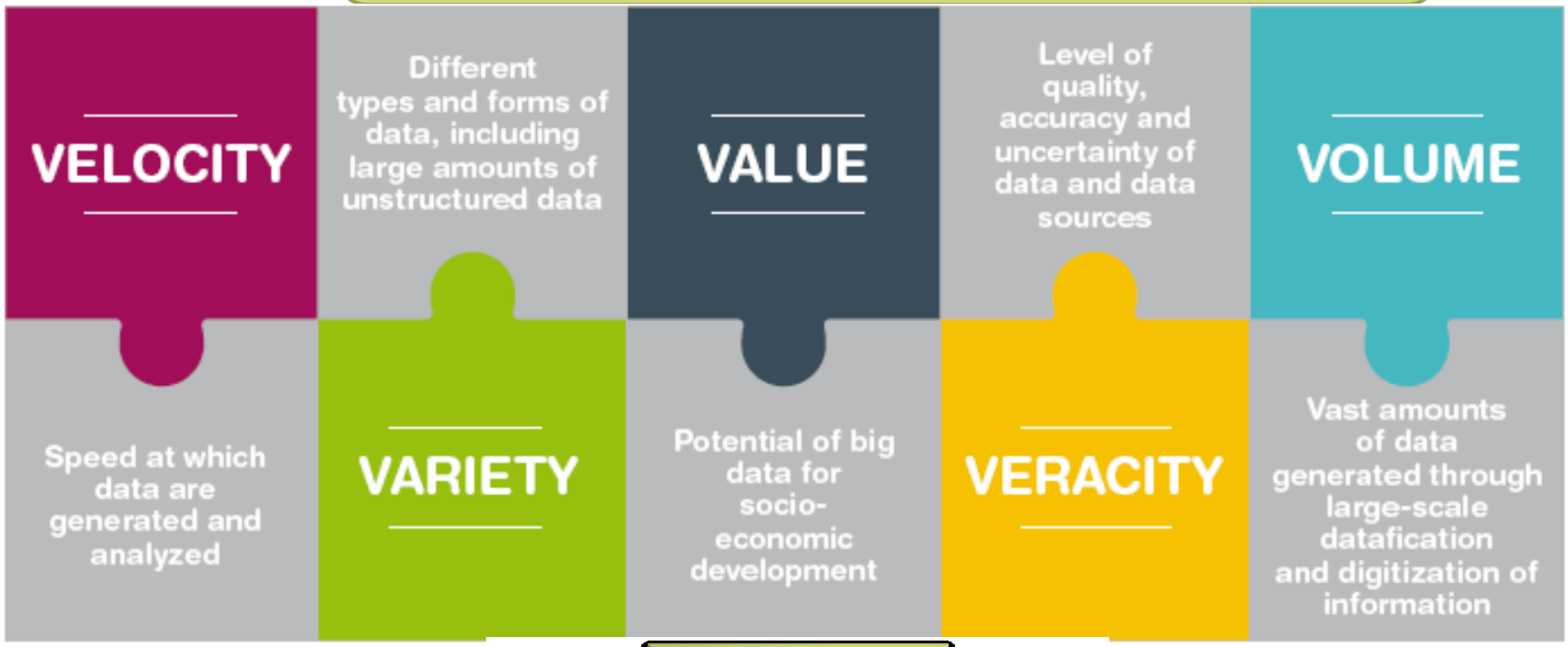
## Enjeux du Big data



# Big Data

Big Data fait référence aux vastes quantités d'informations qui proviennent de différentes sources.

Ce n'est pas qu'une question de masse d'informations: Mais, de différents types de data délivrées à différentes vitesses et fréquences.




Dimensions

# Qualité des Données

- **Le volume**: lié aux multiples sources de production des données. Qu'il s'agisse de données d'entreprises, de données publiques, de données issues de transactions, de données produites par des capteurs automatisés, des objets connectés ou publiées sur les médias sociaux, ces informations sont toujours collectées et stockées sur des supports numériques sous forme de fichiers binaires. Leur volume est donc facilement calculable.
- **La variété** résulte des sources de données hétérogènes, souvent non ou peu structurées (données de capteurs, données de géolocalisation, sons, vidéos, textes,...). Cette variété a motivé la construction de systèmes capables de « gérer » la non structuration (NoSql, Hadoop,...) tout en assurant une meilleure répartition de la charge des volumes sur l'infrastructure de calcul.
- **La vélocité** intervient dans les contextes de données en mouvement, de « data streaming » et de traitement temps réel de ces données. Elle est liée à la vitesse de production de la source, au flux, au débit et à la vitesse de collecte du système. Ici encore, la vélocité est une grandeur facilement mesurable.
- **La visibilité** des données dépend fortement du support de stockage, et de l'efficacité des algorithmes de collecte et autres crawler . On pourrait compléter ces quatre premiers V par celui de la variabilité de la donnée dans certains contextes. Cette variabilité s'exprime pour des données dont le contenu évolue dans le temps et l'espace. Ces évolutions produisent alors de nouvelles données indicées par le temps.
- Les deux derniers V désignent la valeur et la véracité d'une donnée, des qualités beaucoup plus complexes à définir et à mesurer que les quatre premières.
- **La valeur** recouvre en effet plusieurs spectres nécessitant chacun une analyse spécifique. On parlera ainsi de valeur d'impact sur un contexte, de valeur de modélisation, de valeur de prédiction, de valeur de management, de valeur économique ou de revente.
- **La véracité** conditionne quant à elle directement la pertinence de la donnée. Si des données incertaines peuvent être traitées au même titre que des données « certifiées », leur interprétation dans le cadre de fausses données peut engendrer de fortes turbulences sur l'ensemble des systèmes associés et provoquer des sinistres conséquents lorsque des décisions sont prises sur la base de cette interprétation. En fait, il n'existe pas de valeur « absolue » d'une donnée mais plutôt des valeurs relatives à un contexte d'interprétation, à un instant donné.

# Nouveaux Métiers du Big Data

## BIG DATA ARCHITECT


Expert des infrastructures IT permettant le stockage, la manipulation et la restitution des « méga données », il conçoit et administre des Data Centers, en hybride ou dans le cloud sur des plateformes comme Amazon AWS ou Microsoft Azure. Il travaille en amont dans la chaîne de traitement de la donnée et est le pilier de tout projet Big Data.

## DATA ANALYST

Plutôt en fin de chaîne des projets Data et avec l'appui du Data Scientist sur les dimensions technico-scientifiques, il se concentre sur l'exploration et l'exploitation des données métier, dont il extrait des KPI pertinents. Il peut ainsi vulgariser et restituer les résultats aux décideurs, notamment avec des Data Visualisations.


# Nouveaux Métiers du Big Data

## DATA SCIENTIST


Au cœur des projets Data, il s'appuie sur ses compétences techniques et scientifiques avancées, contextualisées par des connaissances métier indispensables. Il élabore des algorithmes complexes, utilise des outils mathématiques, statistiques et du marché (SAS, SPSS, R, etc.) pour extraire, analyser et transformer des données (massives ou non) en information pour répondre à un besoin métier.

## DATA CONSULTANT

Interagissant avec les divers acteurs de la chaîne des projets Data (du Big Data Architect au CDO), il aide les entreprises à définir et à implémenter leurs stratégies Data. Sa connaissance générale des outils du marché, sa créativité et sa compréhension des enjeux métier lui permettent de leur proposer des solutions innovantes.

# Nouveaux Métiers du Big Data

## CHIEF TECHNOLOGY OFFICER


Manager de haut niveau expérimenté rattaché à la Direction Générale, il est en charge des outils et des solutions technologiques innovantes, dont il est l'instigateur au sein de l'entreprise. Il pilote leur conception, leur mise en œuvre et les fait évoluer.

## CHIEF DATA OFFICER

Manager de haut niveau expérimenté rattaché à la Direction Générale, il est responsable de toute la gouvernance des données et de leur valorisation. Il est le garant des données, de leur agrégation et de leur exploitation pour répondre aux enjeux décisionnels de l'entreprise.

# Big Data

## Typologies

**SMART DATA**

**Processus** qui **permet d'enrichir** des **données brutes** pour qu'elles **deviennent qualifiées** dans un **contexte particulier**

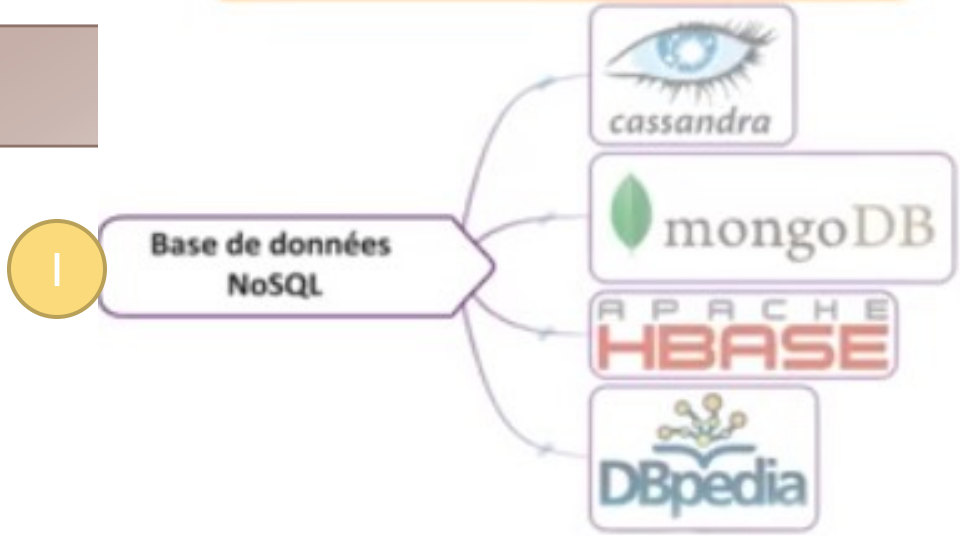
1. Définir une **stratégie**
2. **Cartographier** les **sources**
3. **Sélectionner** les **données pertinentes**
4. **Donner** du **sens** aux **données brutes**
5. **Modéliser** les **ontologies**

**OPEN DATA**

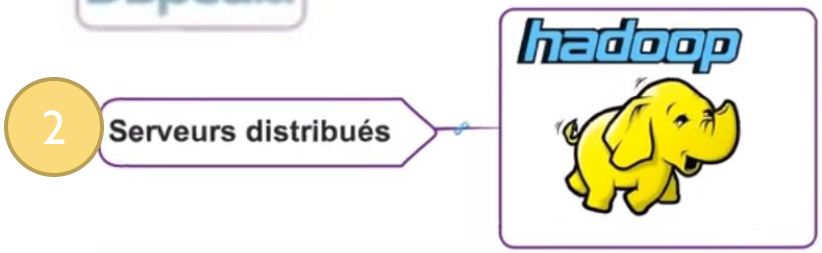
**Désigne l'effort** et la **volonté** des **organisations** à **mettre** à **disposition gratuitement** des **données librement utilisables** par **chacun** d'entre-nous

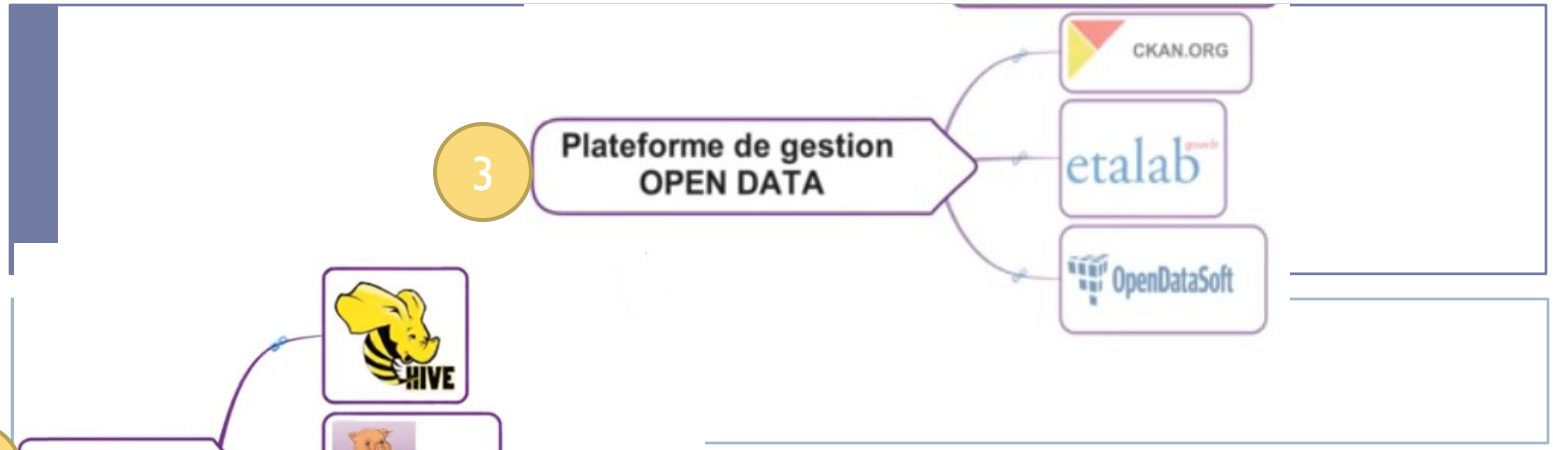
etalab gouv.fr

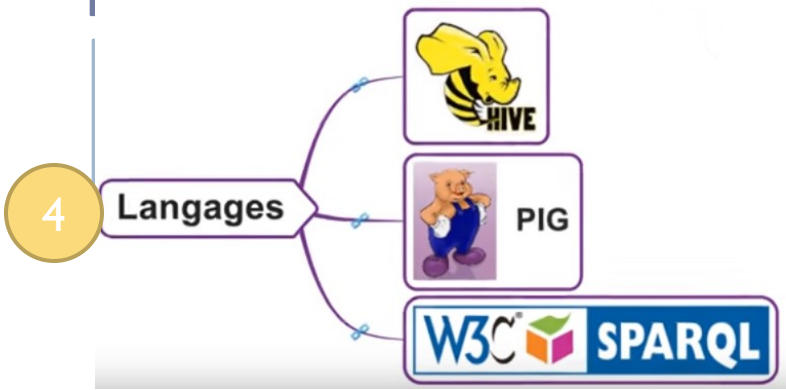
# Big Data



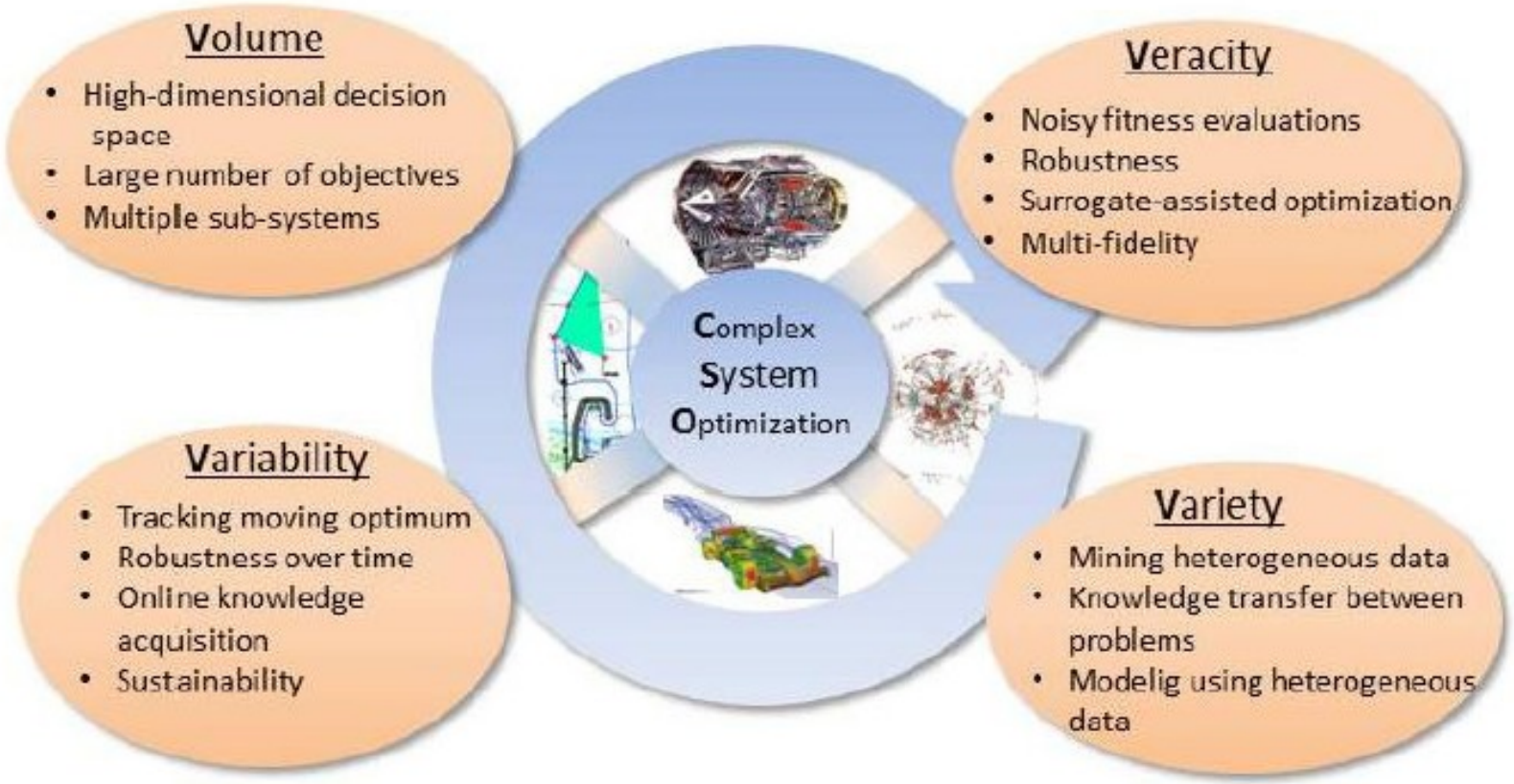

## MOYENS TECHNIQUES



# Relation entre les challenges en optimisation d'ingénierie Complexe et la nature du Big data



# Challenges Big Data


**Génération des Données et Acquisition**

Ce n'est pas un problème dû aux nombreuses sources : Réseau de Capteurs, achat de clients, données astronomiques, messages texte, etc. Challenge = identifier les données intéressantes à générer

**Stockage des Données**

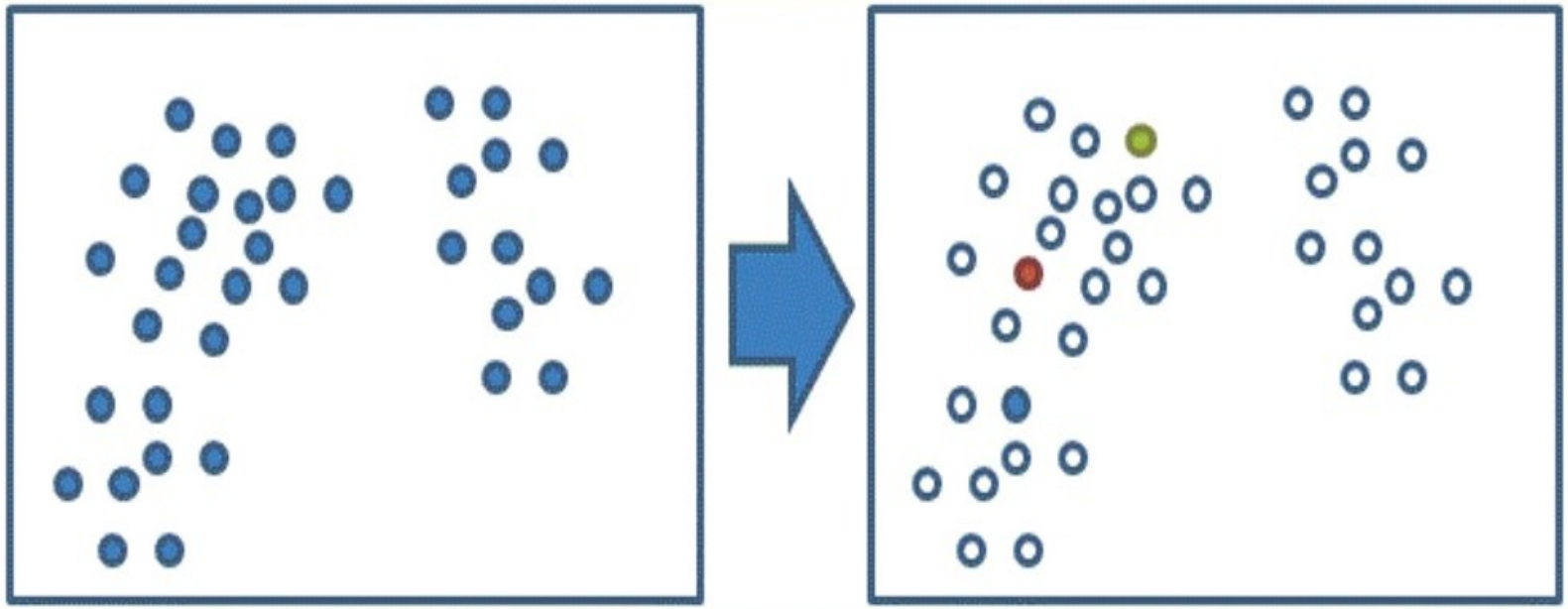
Développement de technologies allant avec les spécificités du Big data (oublier SGBD relationnelles) Spécificités: Scalabilité (adaptation à l'échelle des infrastructures), données non structurées, vélocité incluant l'acquisition asynchrone des données. Résultat: BD non-relationnelles / NoSQL et le modèle à paires de (clé, valeur).

**Analyse des Données**

Créer ou Adapter les outils de data mining au profit du Big Data. Analyse complexe pour découvrir de nouveaux patterns.

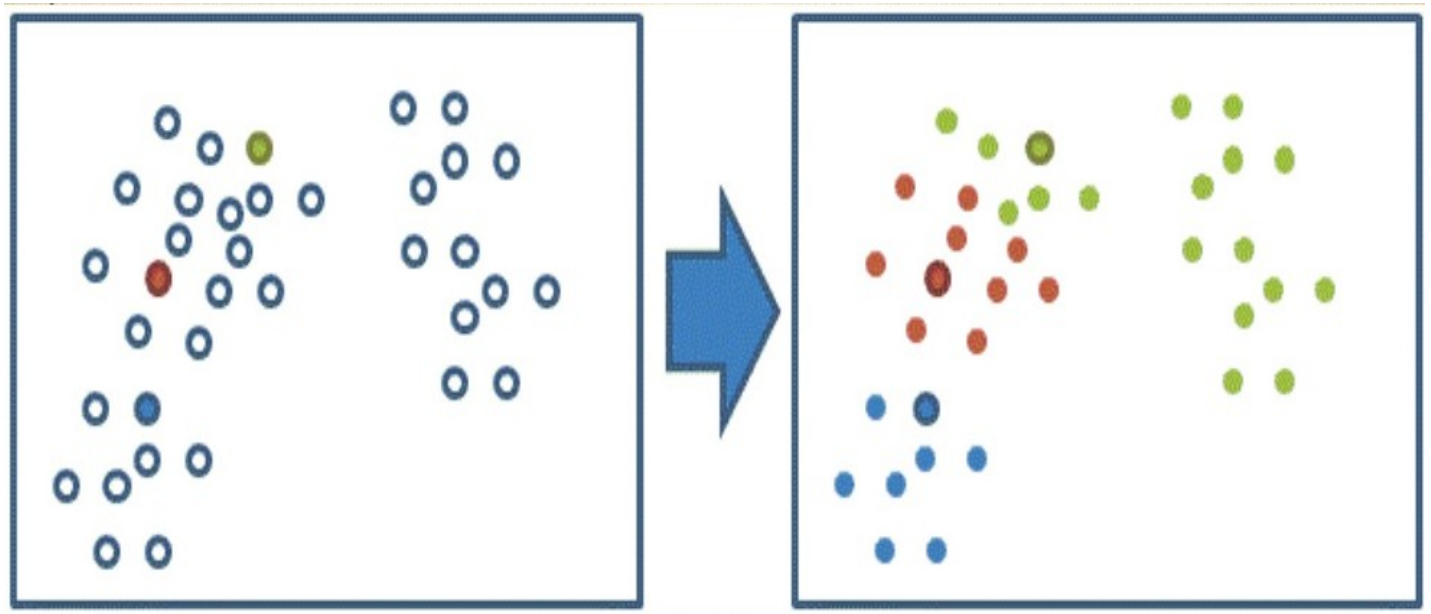
# Parallélisasion de l' Algorithme *K*-Means Basé MapReduce

# 1. Algorithme K-Means



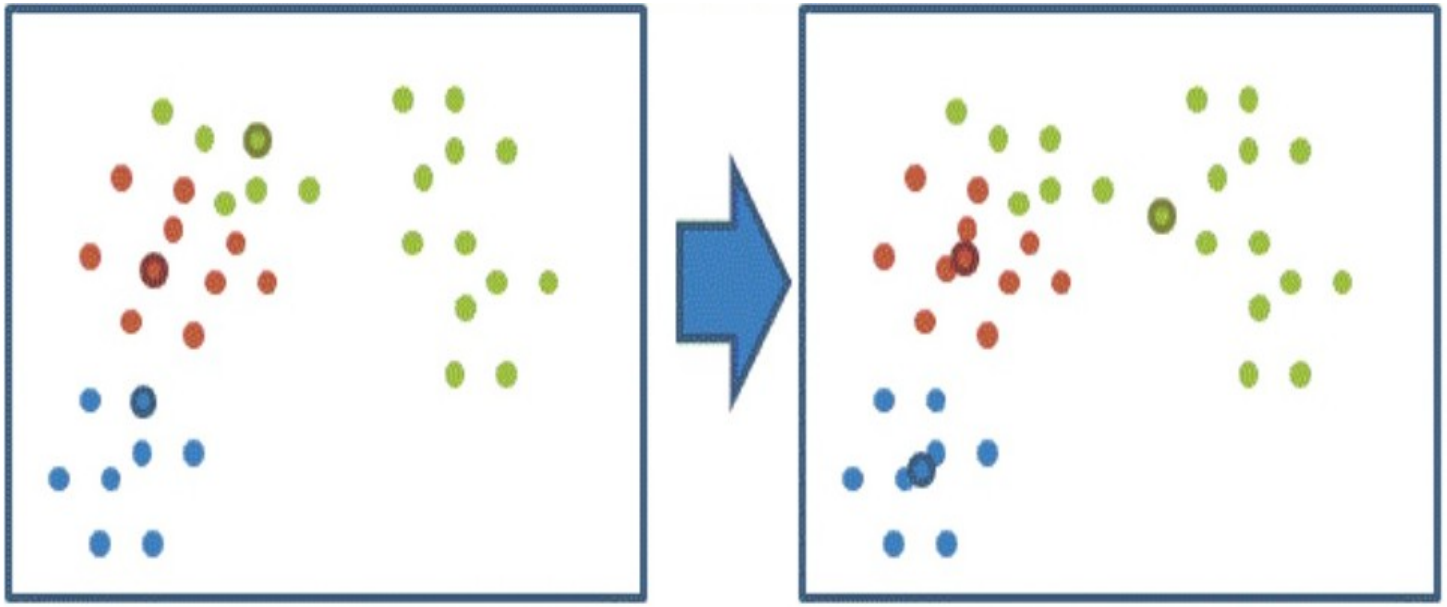
- Tout d'abord, il sélectionne aléatoirement k objets parmi les objets entiers qui représentent les centres de cluster initiaux.

# Algorithme K-Means



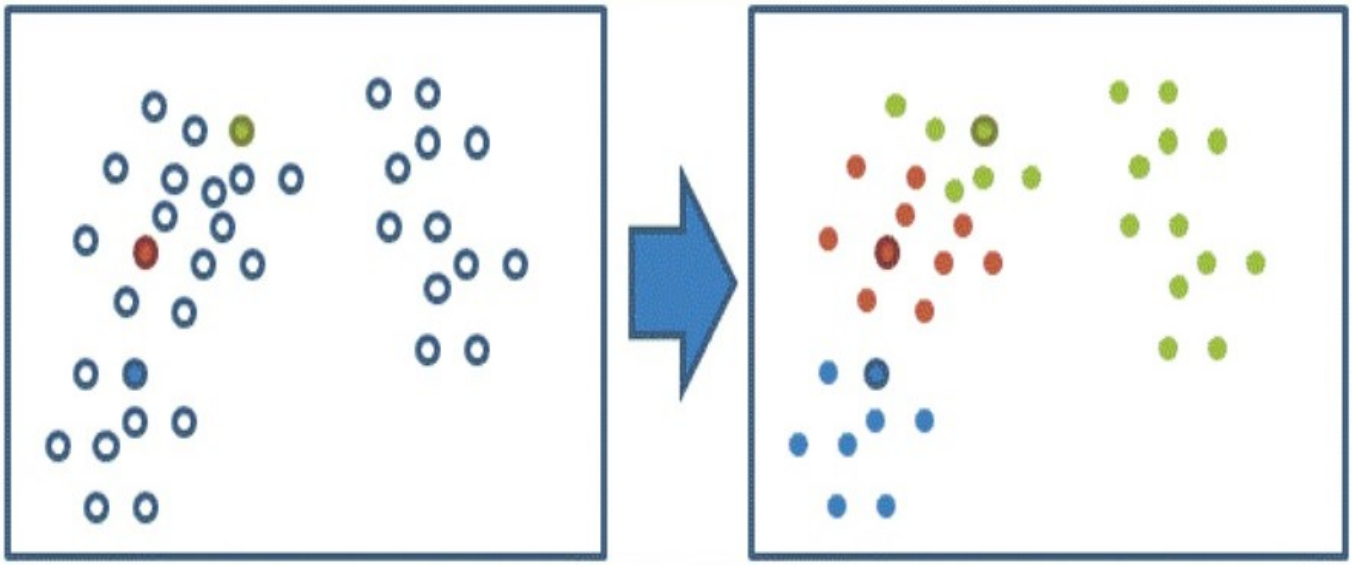
- Chaque objet restant est affecté au cluster auquel il se rapproche le plus, en fonction de la distance entre l'objet et le centre du cluster.

# 1. Algorithme K-Means



- La nouvelle moyenne pour chaque cluster est ensuite calculée. Ce processus itère jusqu'à ce que la fonction critère converge.

# 2. Algorithme K-Means parallèle basés MapReduce



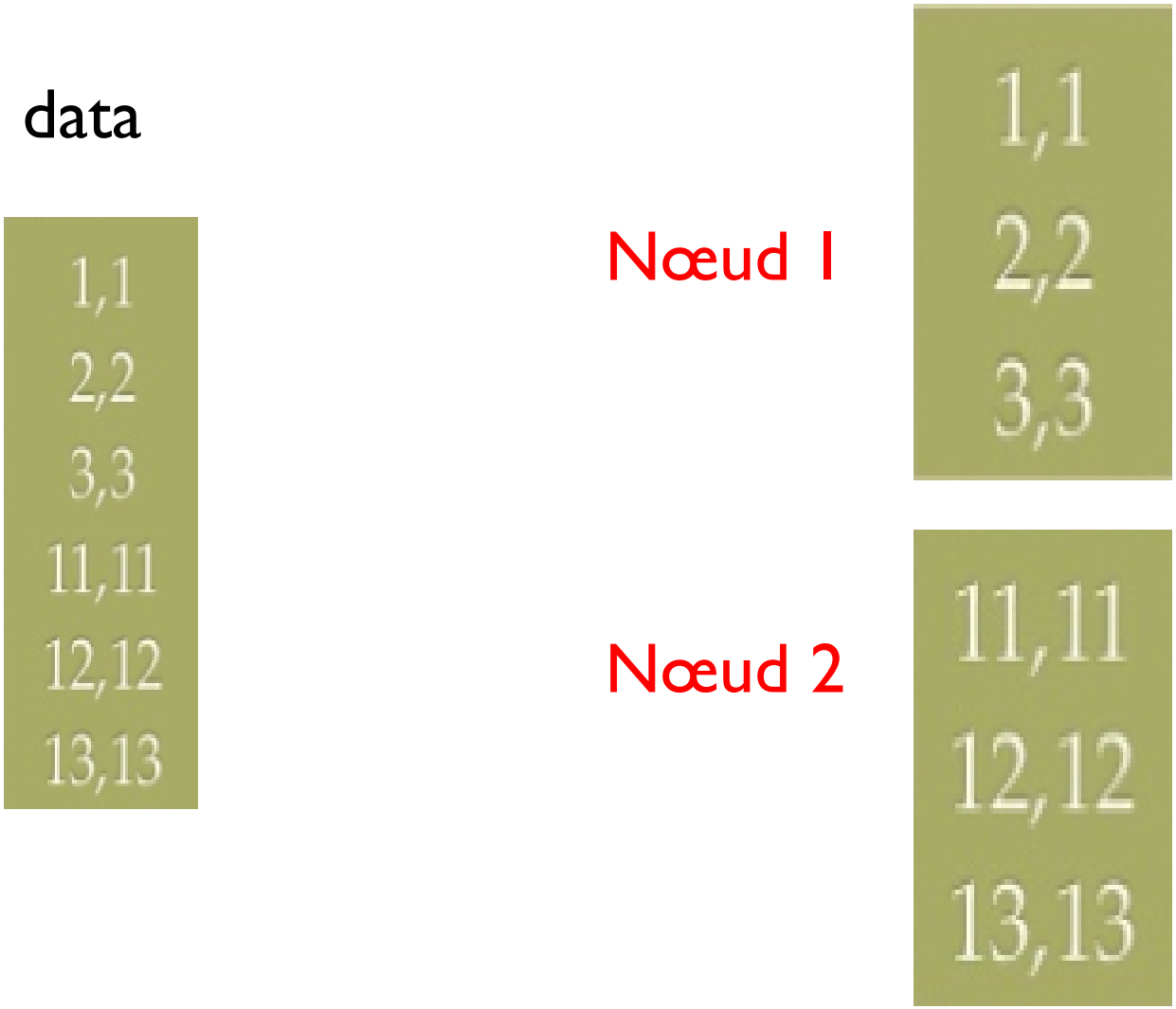
- Basé sur MapReduce, le calcul le plus intensif est le calcul des distances.
- chaque itération nécessite une distance nk.

# 2. Algorithme K-Means parallèle basés MapReduce

- les calculs de distance entre un objet avec les centres sont sans importance pour les calculs de distance entre d'autres objets avec les centres correspondants.

- les calculs de distance entre différents objets avec des centres peuvent être exécutés en parallèle.

# 2. Algorithme K-Means parallèle basés MapReduce



# 2. Algorithme K-Means parallèle basés MapReduce

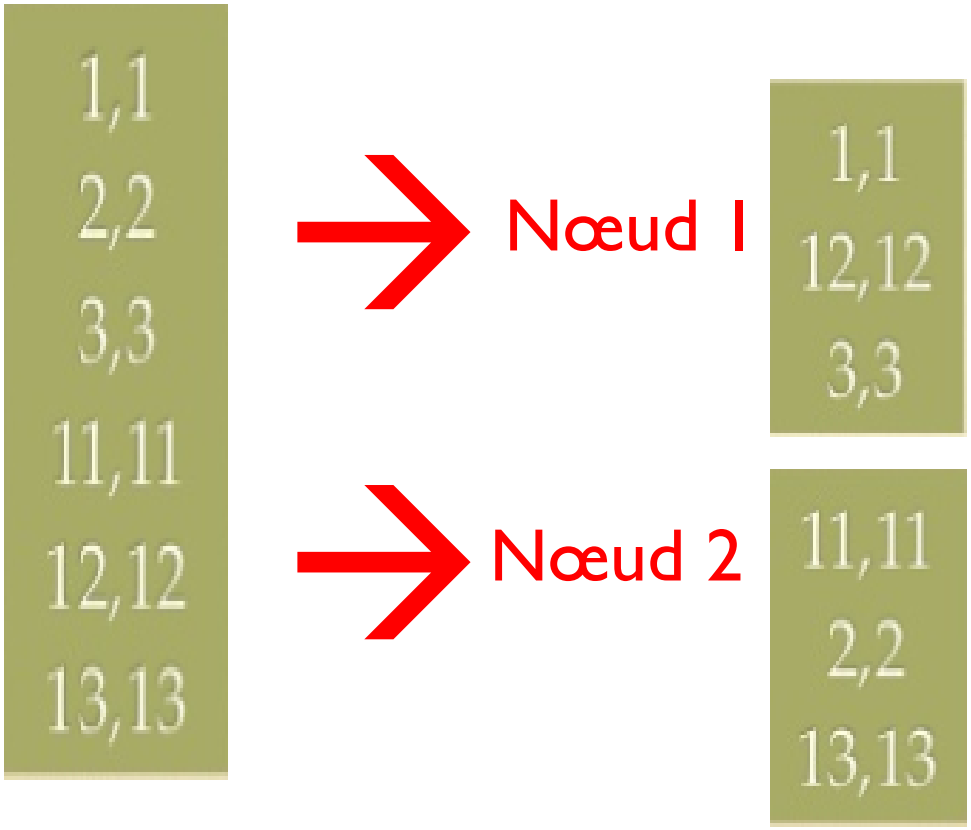
1,1
2,2
3,3
11,11
12,12
13,13

Choix de
2 centres aléatoires

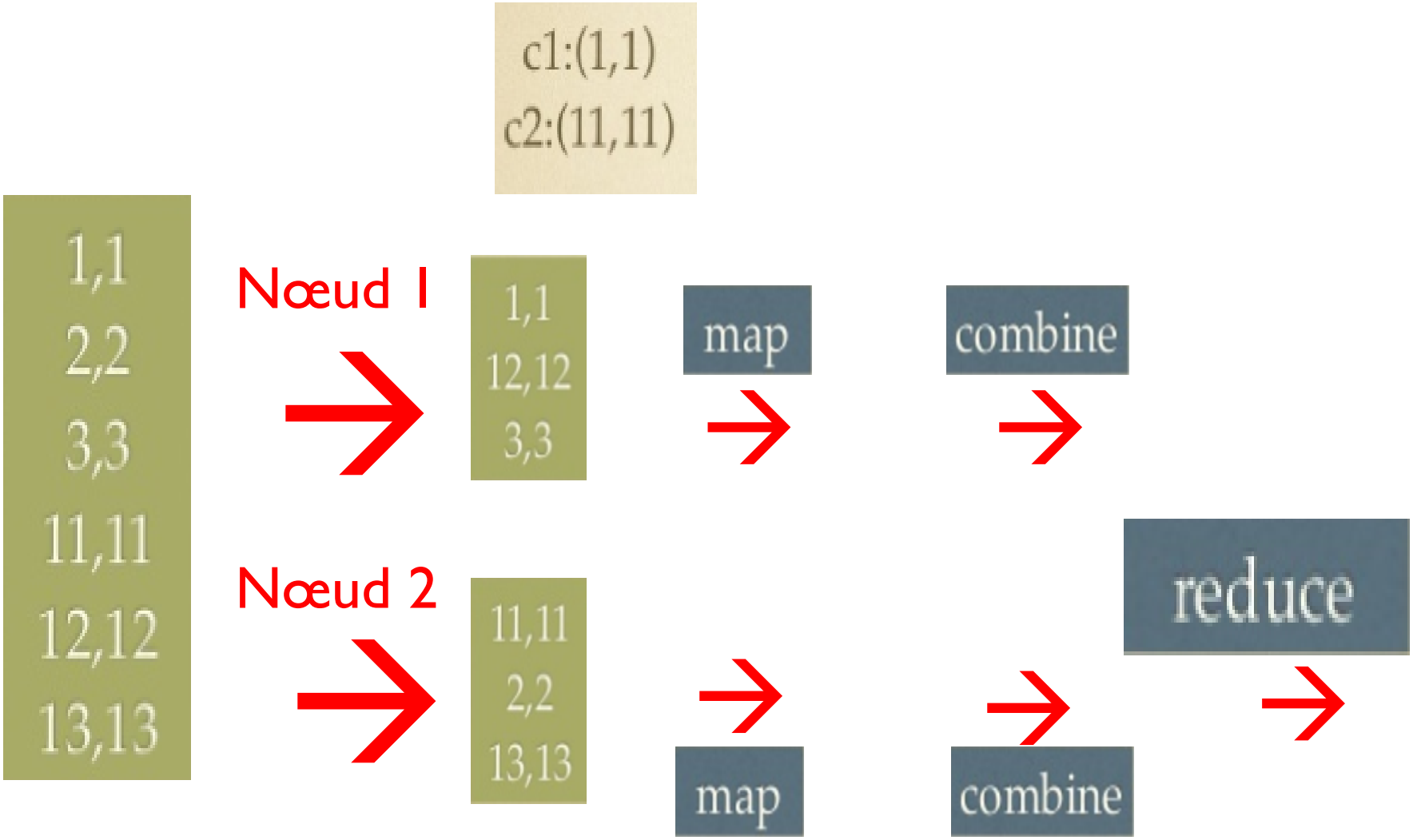
c1:(1,1)
c2:(11,11)

# 2. Algorithme K-Means parallèle basés MapReduce

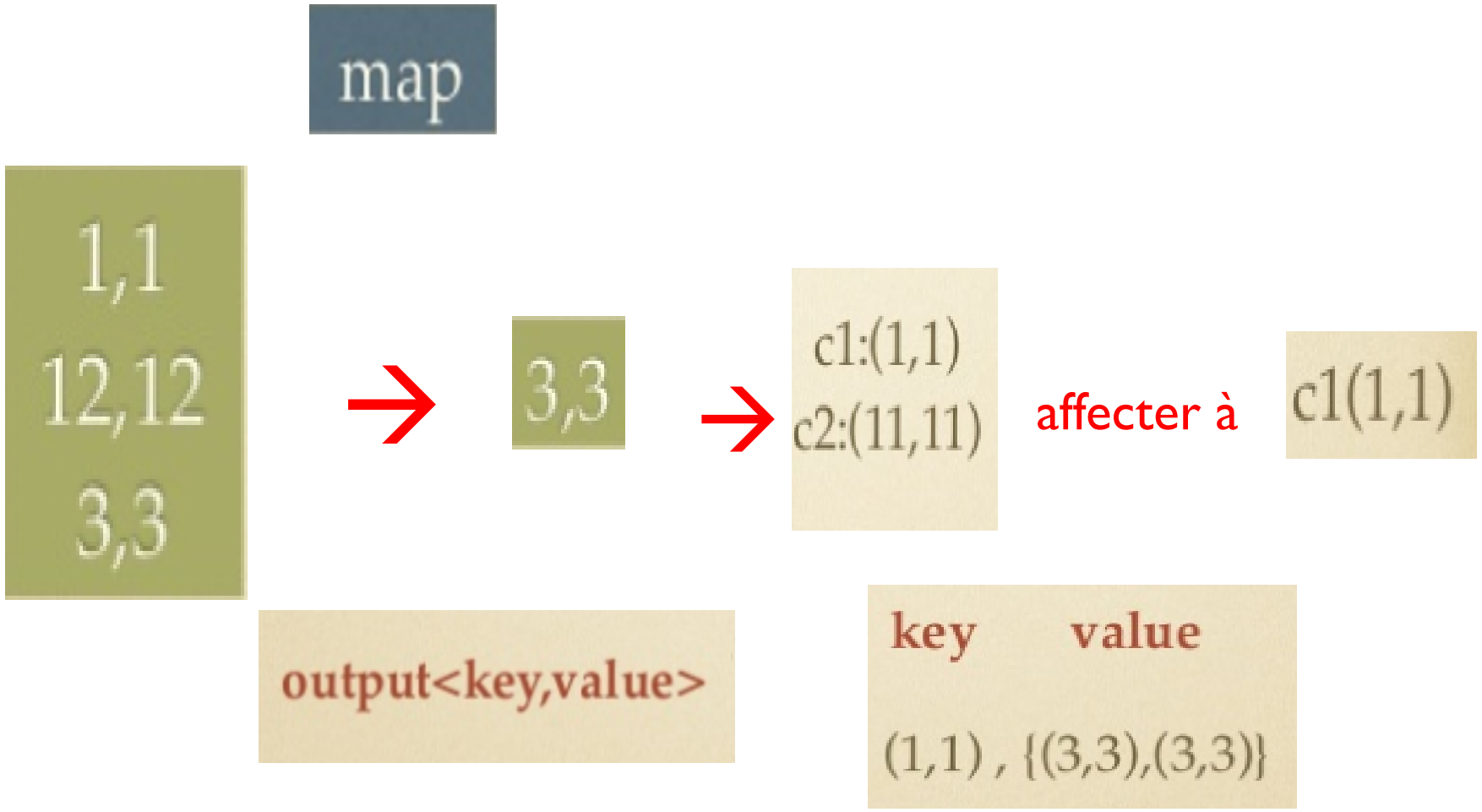
c1:(1,1)
c2:(11,11)



# 2. Algorithme K-Means parallèle basés MapReduce



# 2. Algorithme K-Means parallèle basés MapReduce



# 2. Algorithme K-Means parallèle basés MapReduce

output<key,value>

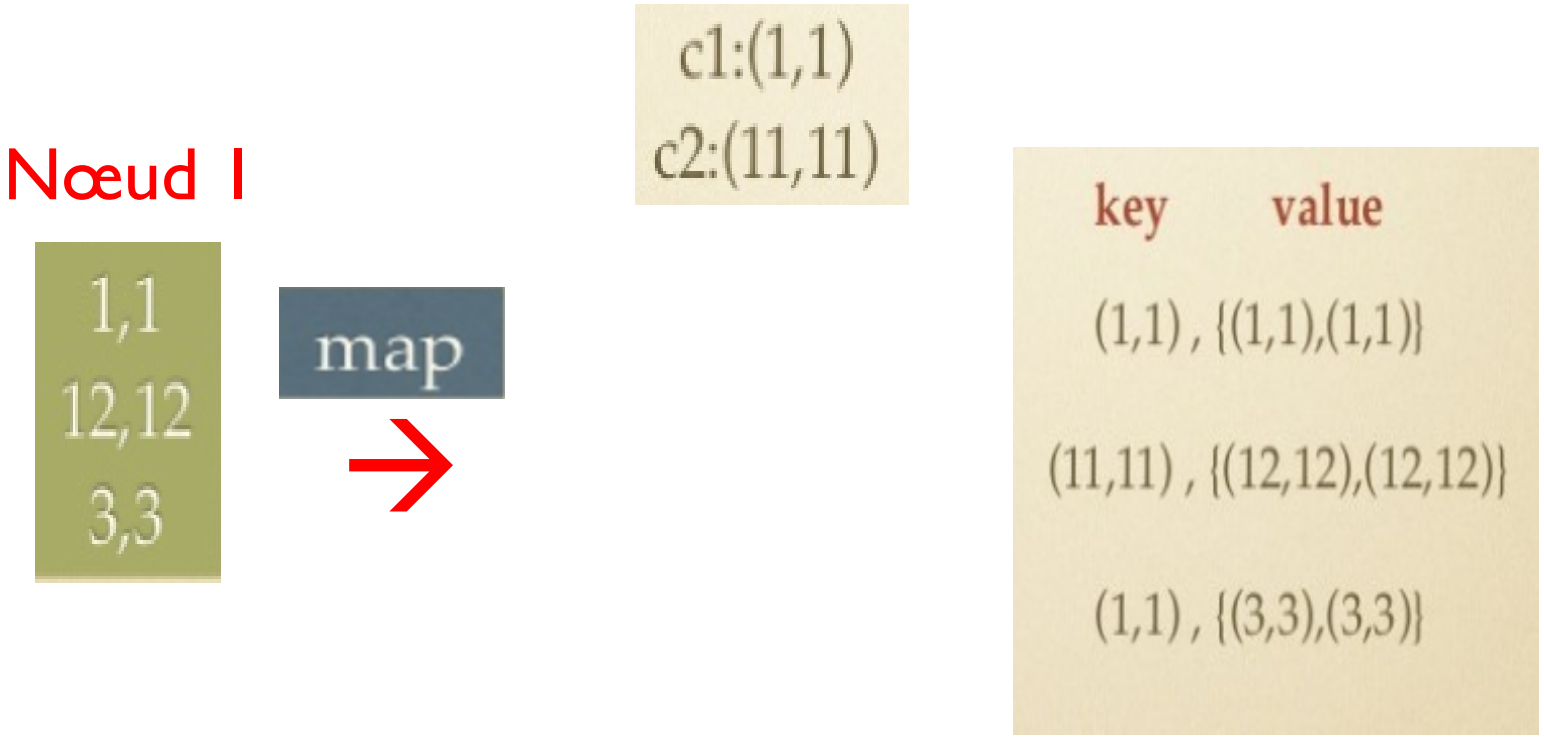
| key | value |
| --- | --- |
| (1,1) | {(3,3),(3,3)} |

centroid

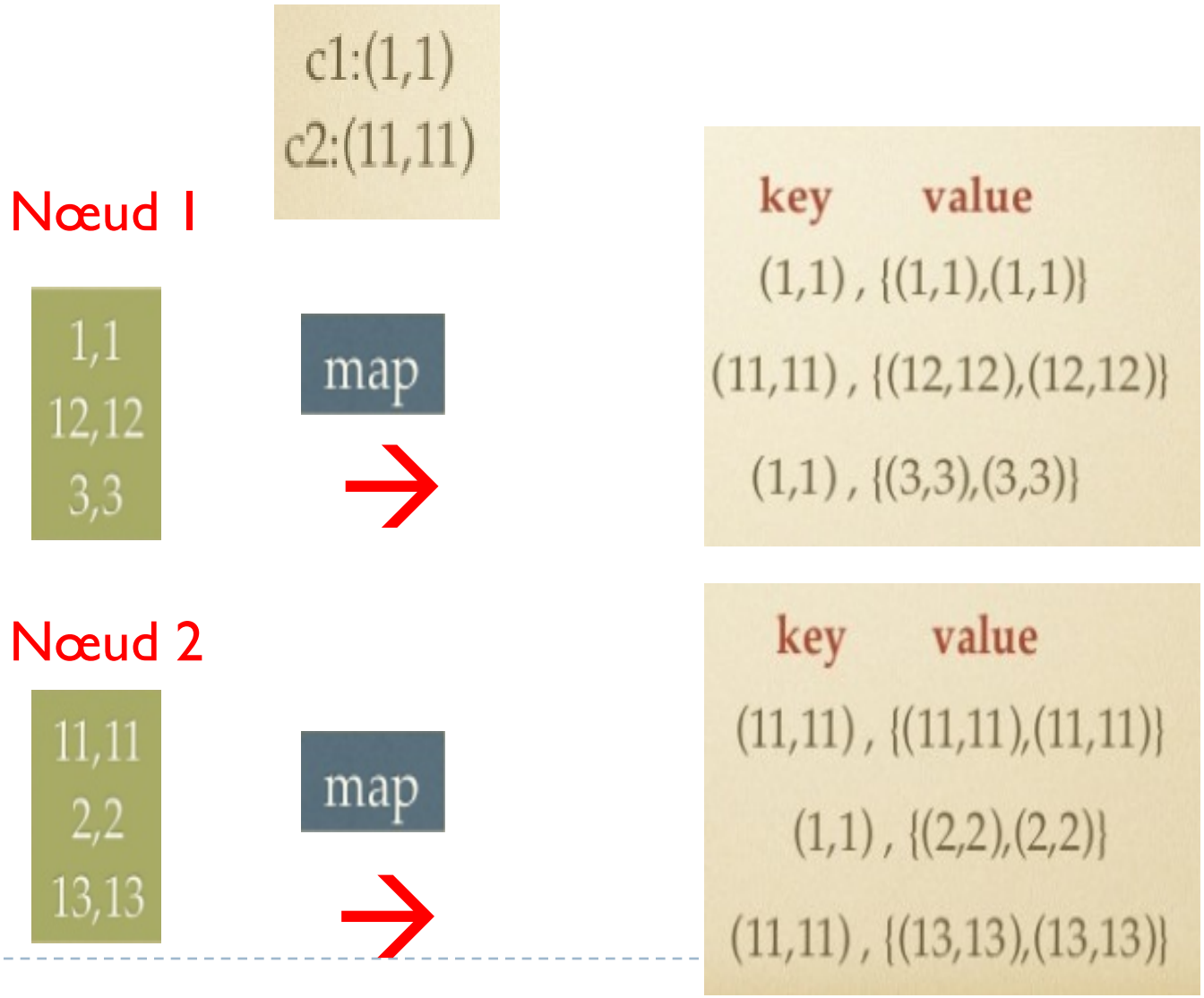
| |
| --- |
| (1,1) |
| {(3,3),(3,3)} |

temporaire pour calculer le nouveau centre

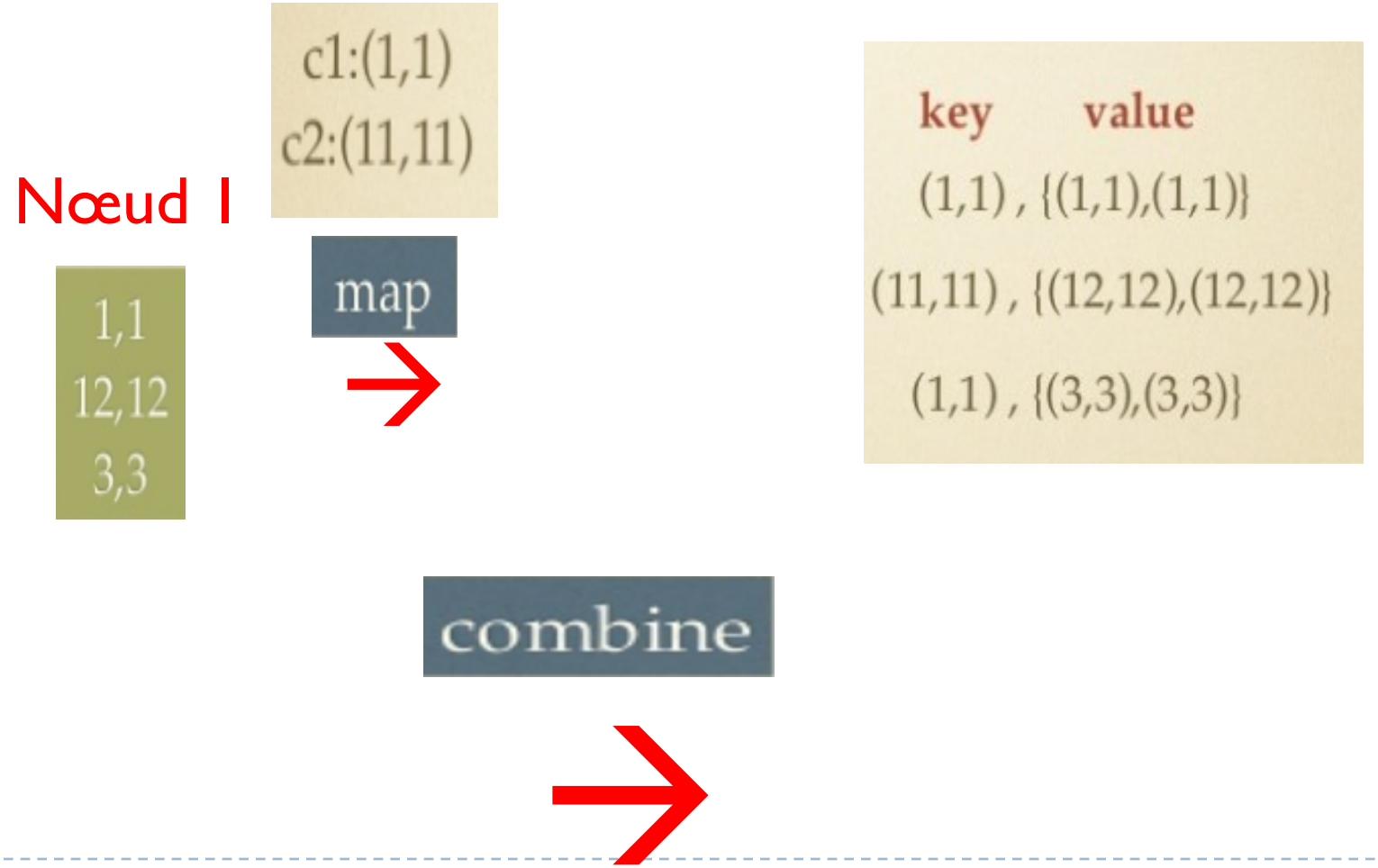
# 2. Algorithme K-Means parallèle basés MapReduce



# 2. Algorithme K-Means parallèle basés MapReduce



# 2. Algorithme K-Means parallèle basés MapReduce



# 2. Algorithme K-Means parallèle basés MapReduce

| key | value |
| --- | --- |
| (1,1) | {(1,1),(1,1)} |
| (11,11) | {(12,12),(12,12)} |
| (1,1) | {(3,3),(3,3)} |

same key combine

combine

| key | value |
| --- | --- |
| (1,1) | {(4,4),{(1,1),(3,3),2} |
| (11,11) | {(12,12),(12,12),1} |

# 2. Algorithme K-Means parallèle basés MapReduce

output<key,value>

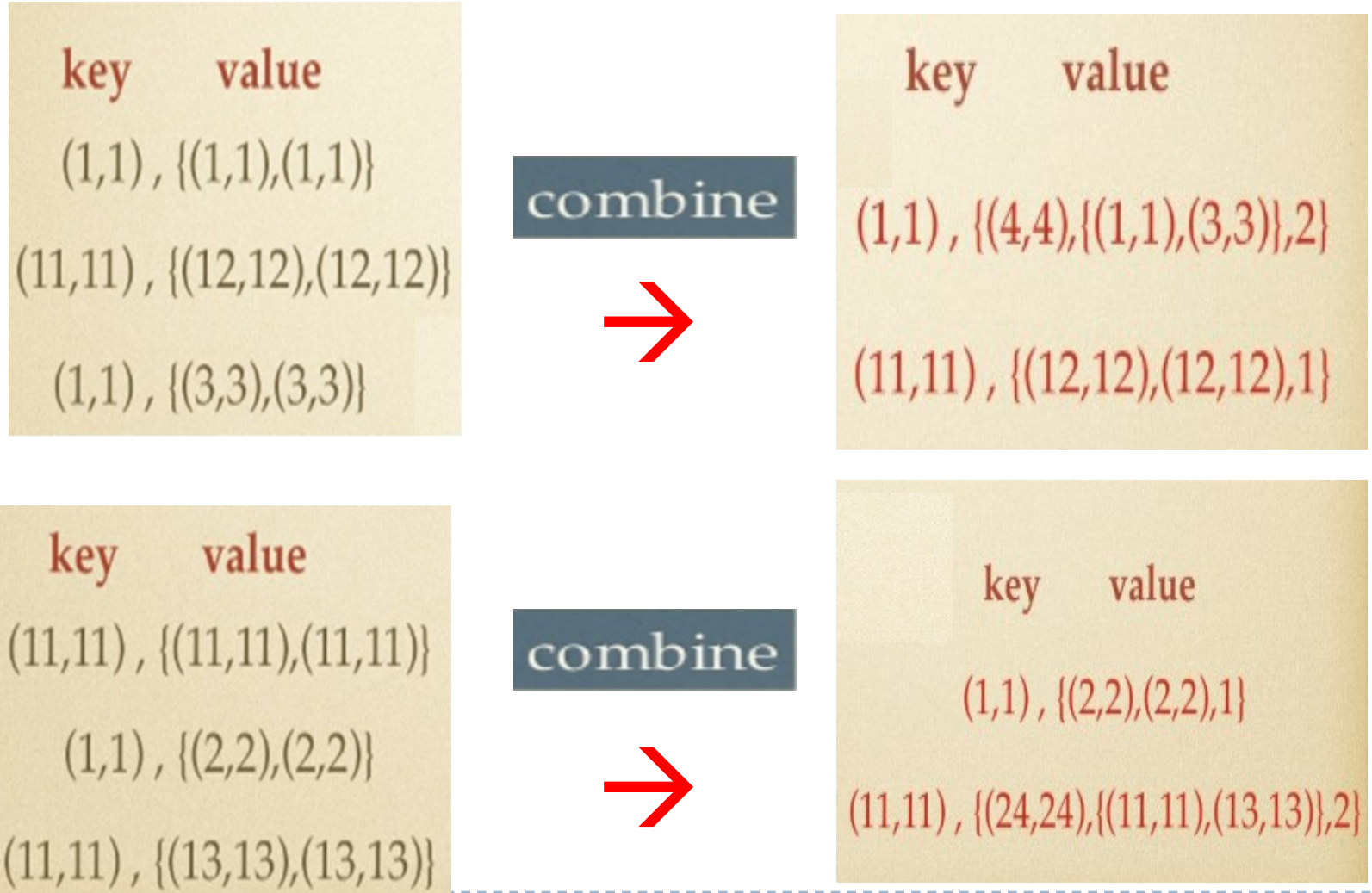
| key | value |
| --- | --- |
| (1,1) | {(4,4),{(1,1),(3,3)},2} |
| (11,11) | {(12,12),(12,12),1} |

centroid

| |
| --- |
| (1,1) |
| {(4,4),{(1,1),(3,3)},2} |

temporaire pour calculer le nouveau centre de gravité, les objets, le nombre d'objets

# 2. Algorithme K-Means parallèle basés MapReduce



# 2. Algorithme K-Means parallèle basés MapReduce

| key | value |
| --- | --- |
| (1,1) , | {(4,4),{(1,1),(3,3)},2} |
| (11,11) , | {(12,12),(12,12),1} |

same key reduce

reduce

→

| key | value |
| --- | --- |
| (1,1) , | {(2,2),(2,2),1} |
| (11,11) , | {(24,24),{(11,11),(13,13)},2} |

# 2. Algorithme K-Means parallèle basés MapReduce

same key reduce

reduce

(1,1) , {(4,4),{(1,1),(3,3)},2}

(1,1) , {(2,2),(2,2),1}

(1,1) , {(2,2),{(1,1),(2,2),(3,3)}

# 2. Algorithme K-Means parallèle basés MapReduce

(1,1) , {(4,4),{(1,1),(3,3)},2}

(1,1) , {(2,2),(2,2),1}

(4+2)/(2+1) ,(4+2)/(2+1) = 2,2

**2,2 = new centroid**

(1,1) , {(2,2),{(1,1),(2,2),(3,3)}

1,1
2,2
3,3

**centroid is 2,2**

# 2. Algorithme K-Means parallèle basés MapReduce

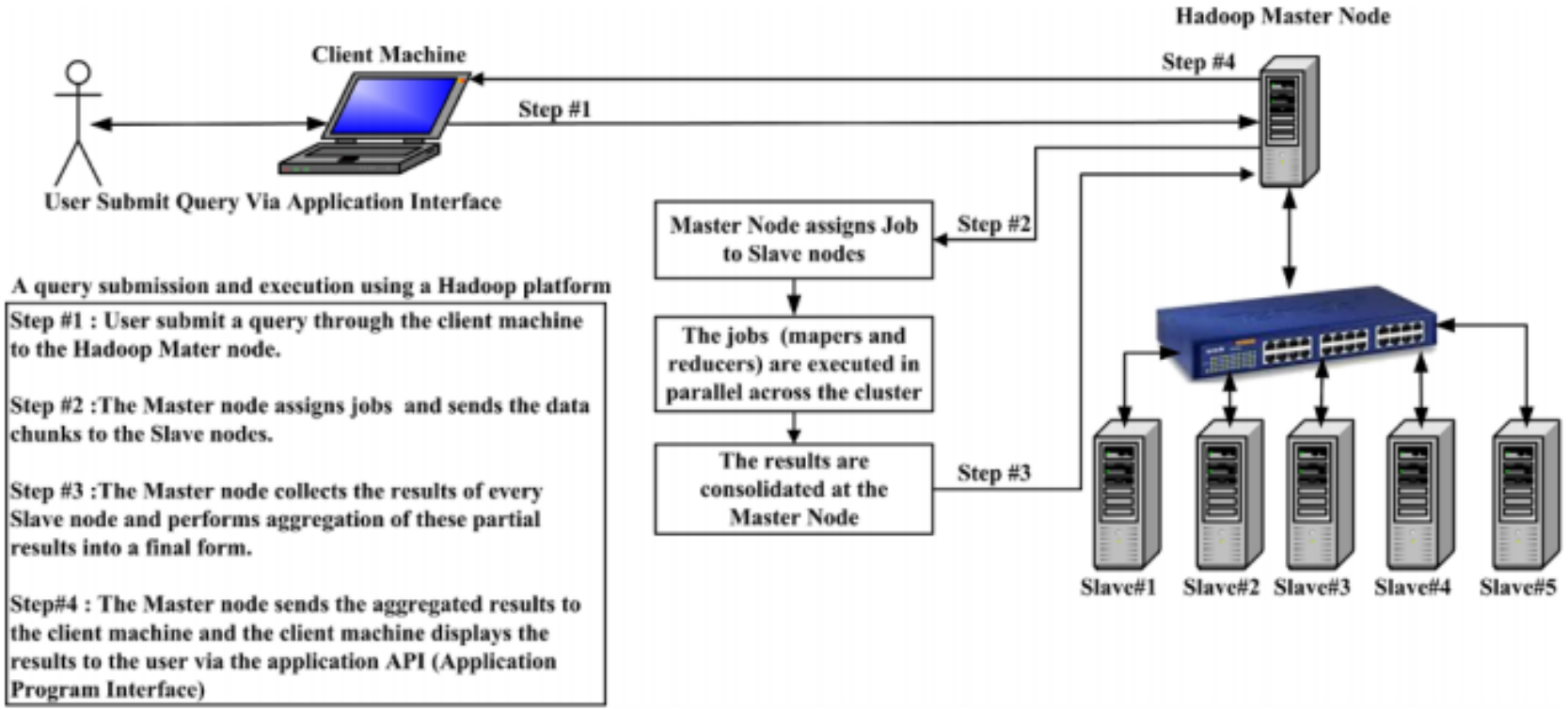
reduce

(1,1) , {(2,2),{(1,1),(2,2),(3,3)}

(11,11) , {(12,12),{(11,11),(12,12),(13,13)}

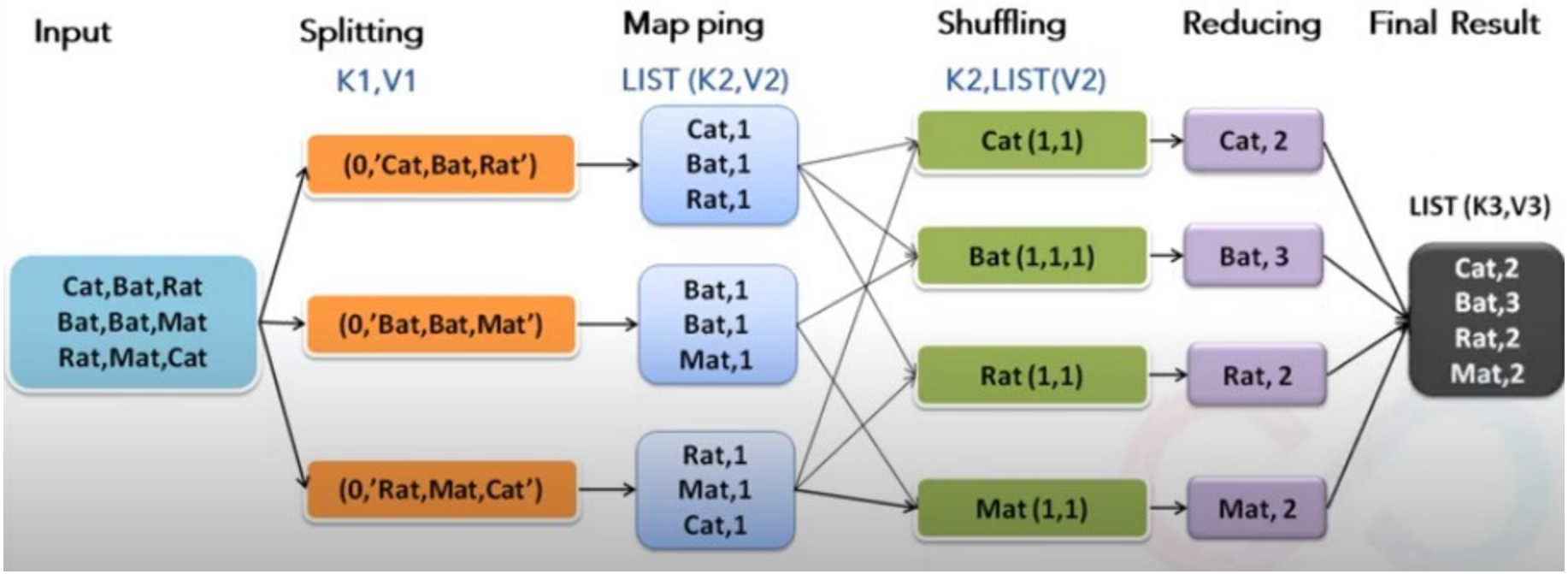
Mettre à jour le nouveau centre et la prochaine itération jusqu'à ce qu'ils convergent ou on parvienne au nombre d'itérations.

# Calcul Distribué utilisant Hadoop




Architecture du cluster Hadoop montrant les nœuds de calcul distribués qui sont Master Node, (NameNode), Slaves Nodes (DataNode), et le switch Ethernet.

# Map Reduce Word Count



# Métaheuristiques Parallèles mais encore!...

**Question**: Comment peut-on exploiter réellement les caractéristiques parallèles d'une métaheuristique telle que un Algorithme génétique?

1. Utilisation d'une infrastructure de calcul distribué et traitement de données à large échelle tel que le Système Cloud.

2. Utilisation Apache Hadoop plateforme : Framework elephant56.

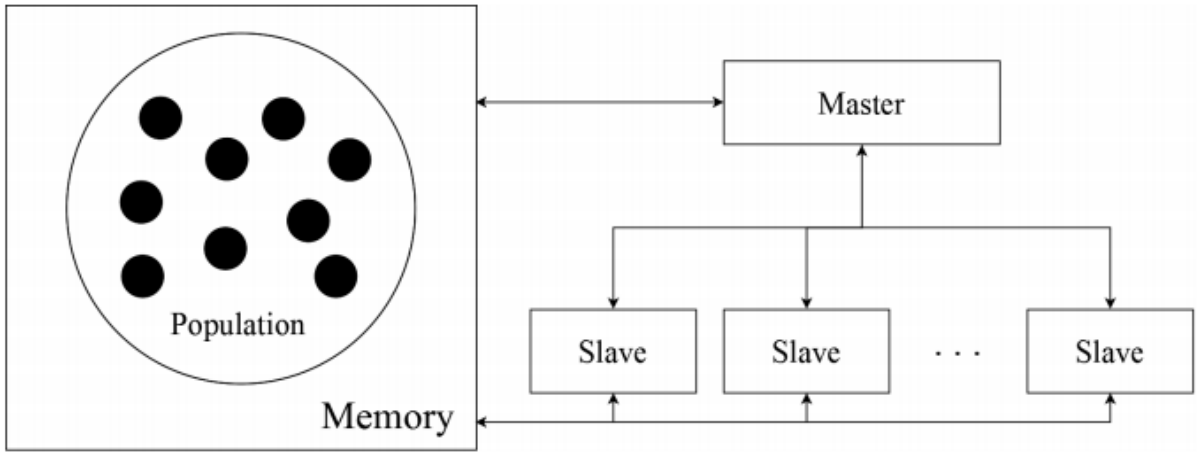
Cette appélation provient de 2 idées:

- ✓ elephant. pour la plateforme Hadoop
- ✓ 56 pour le nombre de chromosomes dans le génotype d'un éléphant.

# Modèle d'Inspiration AG Parallèle pour MapReduce (1)

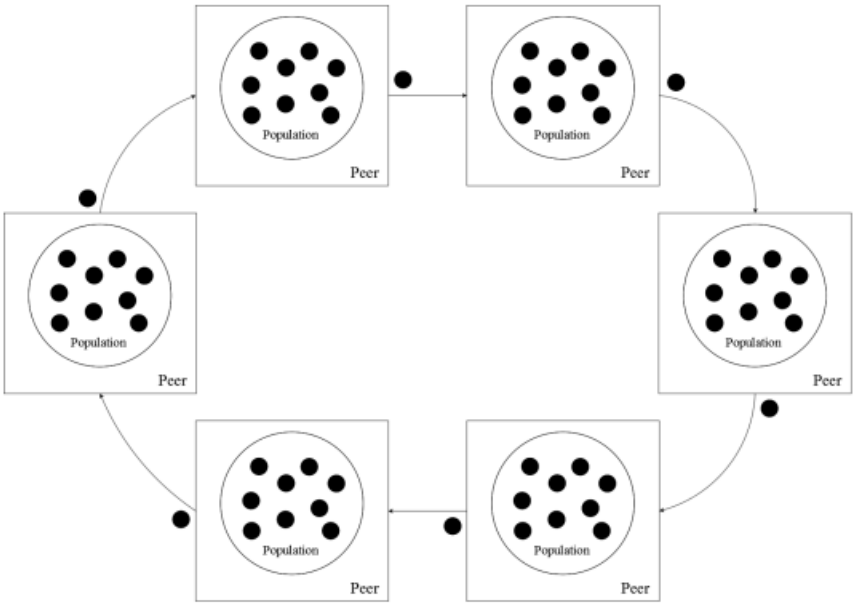
Il existe dans ce modèle trois niveaux de parallélisme.

1. **Niveau Evaluation Fitness** (modèle de parallélisasion global). Le nœud master gère la population et calcule toutes le fonctions AG. Evaluation fitness est calculée par les nœuds esclaves.



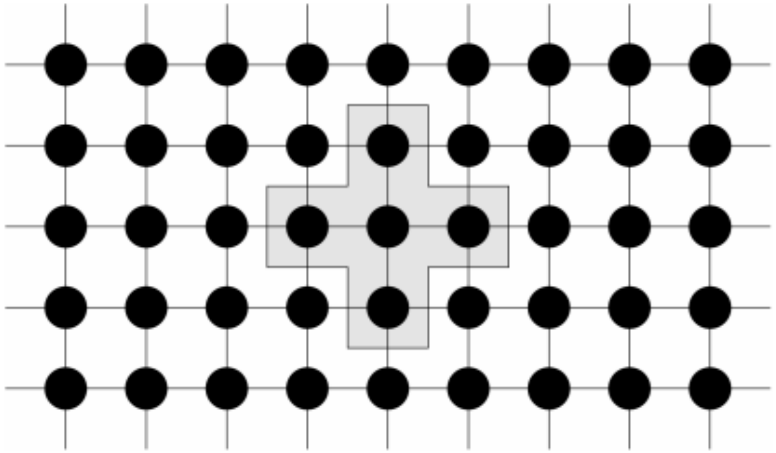
# Modèle d'Inspiration AG Parallèle pour MapReduce (2)

2. **Niveau Population** (coarse-grained parallelization model ou Island model). La population est divisée en îlots et l'AG est exécuté indépendamment pour chacun. Périodiquement, les îlots échangent des informations par migration d'individus.



# Modèle d'Inspiration AG Parallèle pour MapReduce (3)

3. **Niveau Individu** (fine-grained parallelization model ou Grid model). Chaque individu est placée sur la grille et les opérations sont exécutées en parallèle en évaluant simultanément la fitness et en appliquant la sélection est uniquement limitée au voisinage adjacent le plus petit.



# Hadoop MapReduce Framework
## AG Architecture Sans MapReduce