Enseigner la Statistique pour l'Analyse de Mégadonnées

Philippe Besse[1], Nathalie Villa-Vialaneix[2], Anne Ruiz-Gazen[3]

Journées de Statistique de la SFdS
Mercredi 3 Juin 2015

1. INSAT, IMT UMR CNRS 5219 – 2. INRA UR875 MIA-T – 3. TSE, GREMAQ UMR CNRS 5604, UMR INRA 1291

Contexte : le statisticien devient “Data Scientist” ?

Évolutions actuelles du stockage / analyse des données :

- accroissement important de la capacité de stockage des données ;
- évolution des systèmes de stockage : stockage en ligne, stockage partagé ;
- évolution des capacités de calcul des ordinateurs;
- explosion du commerce en ligne (GAFA).

Contexte : le statisticien devient “Data Scientist” ?

Évolutions actuelles du stockage / analyse des données :

- accroissement important de la capacité de stockage des données ;
- évolution des systèmes de stockage : stockage en ligne, stockage partagé ;
- évolution des capacités de calcul des ordinateurs;
- explosion du commerce en ligne (GAFA).

Nouvelles demandes pour les jeunes diplômés statisticiens

Dans les secteurs bancaires, marketing, systèmes de recommandation :

- volonté de tirer le maximum de profit de la masse de données stockées pour améliorer les performances de l’entreprise ;
- bientôt l’industrie ?
- besoins de connaissances en lien avec l’informatique.

$\Rightarrow$ évolution de l’offre de formation pour intégrer ces aspects

Un petit tour (très incomplet) des outils pour le statisticien des Mégadonnées

R : une porte d'entrée

Avantages :

- beaucoup de méthodes statistiques disponibles et nombreux outils pour aider à l'interprétation ;
- simple à prendre en main.

Un petit tour (très incomplet) des outils pour le statisticien des Mégadonnées

R : une porte d'entrée

Avantages :

- beaucoup de méthodes statistiques disponibles et nombreux outils pour aider à l'interprétation ;
- simple à prendre en main.

Mais : assez peu adapté à l'analyse de données volumineuses

- gestion des données en mémoire vive qui limite la taille des données pouvant être traitées ;
- lent...

Un petit tour (très incomplet) des outils pour le statisticien des Mégadonnées

Aller plus loin...

- Utiliser des langages de programmation orientés vers l'analyse de données qui n'ont pas les mêmes limites que R : python (pandas, Scikit-learn), julia, ..., ou des environnements dédiés à l'analyse de données massives hadoop (Mahout), Spark (MLib)... : méthodes pré-programmées limitées, plus difficiles à prendre en main

Un petit tour (très incomplet) des outils pour le statisticien des Mégadonnées

Aller plus loin...

- Utiliser des langages de programmation orientés vers l'analyse de données qui n'ont pas les mêmes limites que R : python (pandas, Scikit-learn), julia, ..., ou des environnements dédiés à l'analyse de données massives hadoop (Mahout), Spark (MLib)... : méthodes pré-programmées limitées, plus difficiles à prendre en main
- "Doper" R : CRAN TASK VIEW "High-Performance and Parallel Computing with R"
  - gestion de la mémoire : packages **bigmemory** (**biglm**, **bigrf**...)
  - calcul parallèle : **snow**, **snowfall**, **foreach**...

Un petit tour (très incomplet) des outils pour le statisticien des Mégadonnées

Aller plus loin...

- Utiliser des langages de programmation orientés vers l'analyse de données qui n'ont pas les mêmes limites que R : python (pandas, Scikit-learn), julia, ..., ou des environnements dédiés à l'analyse de données massives hadoop (Mahout), Spark (MLib)... : méthodes pré-programmées limitées, plus difficiles à prendre en main
- "Doper" R : CRAN TASK VIEW "High-Performance and Parallel Computing with R"
  - gestion de la mémoire : packages **bigmemory** (**biglm**, **bigrf**...)
  - calcul parallèle : **snow**, **snowfall**, **foreach**...
- Utiliser R comme une interface avec :
  - d'autres langages : python (**rJython**), C++ (**Rcpp**), Java (**rJava**)
  - des environnements comme hadoop : **RHadoop** est un ensemble de packages permettant l'utilisation de R dans un environnement Hadoop

Donner une compétence “données massives” à des formations existantes

Deux exemples de formation

- Spécialité Génie Mathématique et Modélisation (INSA Toulouse) du niveau L3 au niveau M2.
- Formation “Economics & Statistics” (Master, TSE): au niveau M1, cours de “Multivariate Data Analytics” (durée : 5 séances de 3 heures correspondant à la deuxième partie du cours ; 27 étudiants dans une salle de TP)

1. INSA: GMM-MMS

Contenu: Existant / Ajouté

L3
- Bases de données: SQL
- Sondages: *représentativité* des données

M1
- Optimisation: complétion de matrices
- Exploration: ACP, AFC, AFCM, MDS, NMF
- Logiciels Stat: R, SAS; Python (pandas)
- Classification données fonctionnelles ?
- Algorithmes stochastiques: Gradient stochastique et régression, SVD

M2
- Apprentissage statistique: régressions (KRLS, Lasso...), PLS, AD, NNet, Arbres, RF, Boosting, SVM, Imputation, Atypiques
- Atelier Science des Données: Études de cas avec RHadoop, Python, PySpark / Hadoop, Mesos...

Ressources pédagogiques: Site wikistat.fr

2. Master TSE: E & S

Contraintes et objectifs

Problèmes rencontrés :

- temps de formation limité avec des étudiants ayant peu ou pas du tout de notions de programmation R (ni du reste) ;
- infrastructure matérielle très limitée d'une école / université : pas de cluster de calcul, hadoop n'est pas installé, ...

2. Master TSE: E & S

Contraintes et objectifs

Problèmes rencontrés :

- temps de formation limité avec des étudiants ayant peu ou pas du tout de notions de programmation R (ni du reste) ;
- infrastructure matérielle très limitée d'une école / université : pas de cluster de calcul, hadoop n'est pas installé, ...

... et pourtant, il faudrait :

- faire comprendre la difficulté de traiter de gros volumes de données ;
- faire une introduction (sensibilisation ?) au calcul haute performance.

M1 E & S

Au niveau matériel

- utilisation des ordinateurs de bureau standards en utilisant les processeurs multi-cœurs pour illustrer le calcul parallèle
- utilisation du package **rmr2** de **rhadoop** pour présenter l'approche "MapReduce" sans avoir à installer Hadoop
- travail sur des données du package **mlbench** (pas des méga-données) et sur les données "adult" UCI repository ($\sim$ 6Mo)

M1 E & S

Au niveau matériel

- utilisation des ordinateurs de bureau standards en utilisant les processeurs multi-cœurs pour illustrer le calcul parallèle
- utilisation du package **rmr2** de **rhadoop** pour présenter l'approche "MapReduce" sans avoir à installer Hadoop
- travail sur des données du package **mlbench** (pas des méga-données) et sur les données "adult" UCI repository ($\sim$ 6Mo)

Au niveau pédagogique

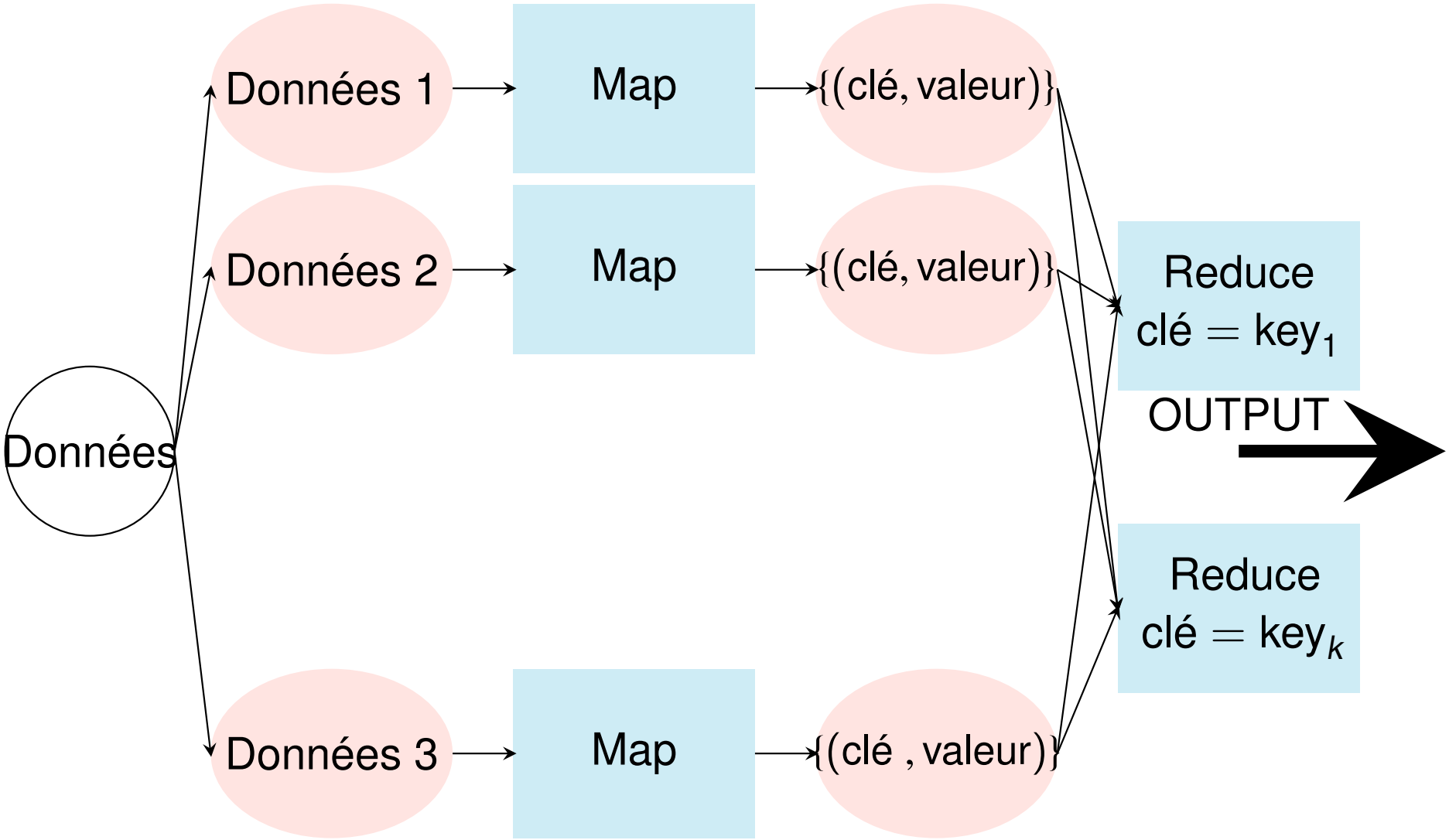
- partie théorique courte en début de séance (30/45 minutes)
- mise en œuvre pratique immédiate "immersion piscine"
- partie importante de travail personnel à la maison : chaque semaine, un exercice complexe sur des données réelles à traiter à la maison et à rendre (noté)

Notions abordées

5 séances :

1. boostrap + exercice : simulations et estimation bootstrap d'une moyenne et d'un IC avec une boucle `for`
   à la maison : reproduire cette application en utilisant le package **boot** ; comparaison du temps de calcul avec l'utilisation d'une boucle `for`
2. bagging + exercice : bagging d'arbres avec **rpart** et **boot** (CART a été étudié dans la première partie du cours)
   à la maison : bagging d'arbres avec le package **ipred** ; comparaison de résultats de bagging sur des arbres et sur k-plus proches voisins
3. forêts aléatoires + exercice : utilisation du package **randomForest** pour apprendre et interpréter RF sur des données réelles
   à la maison : comparaison bagging d'arbres et RF
4. parallélisation + exercice : paralléliser du bagging d'arbres et une forêt aléatoire avec **foreach** ; comparaison des temps de calcul avec l'approche séquentielle (+ packages **doMC** et **doParallel**)
5. MapReduce + exercice : utilisation de MapReduce (package **rmr2**) pour obtenir une table de contingence, une forêt aléatoire

MapReduce


Données 1
Map
{(clé, valeur)}
Données 2
Map
{(clé, valeur)}
Données
Reduce
clé = key$_1$
OUTPUT
Reduce
clé = key$_k$
Données 3
Map
{(clé , valeur)}

MapReduce

Cas d'étude stupide...

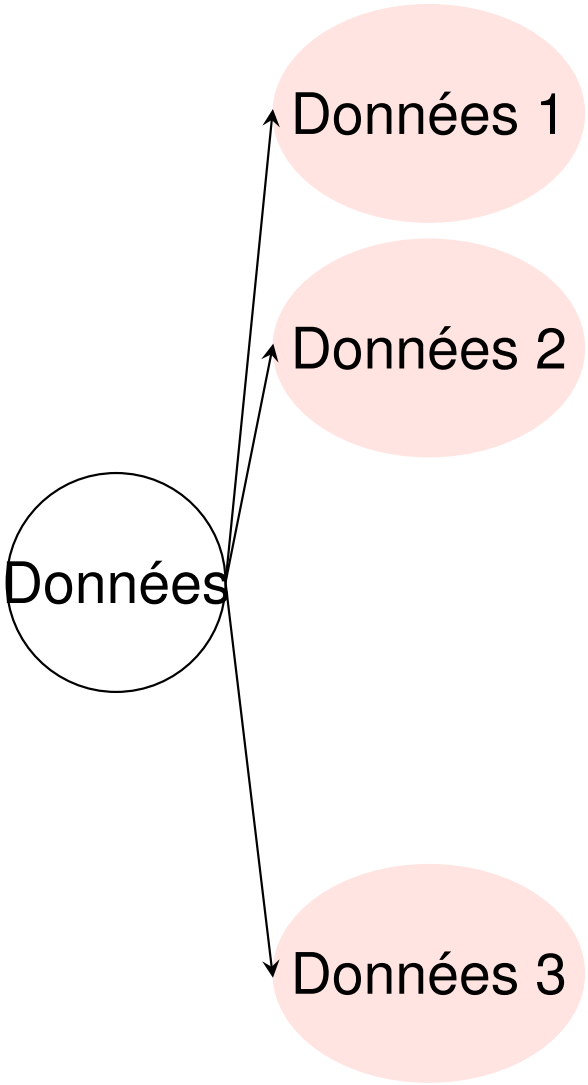
Fichier de ventes (très grand) dans lequel chaque vente contient des informations sur le nom de la boutique et le montant de la vente :

```
 shop1,25000
shop2,12
shop2,1500
shop4,47
shop1,358
...
```

Question : table de contingence du total des ventes par boutique

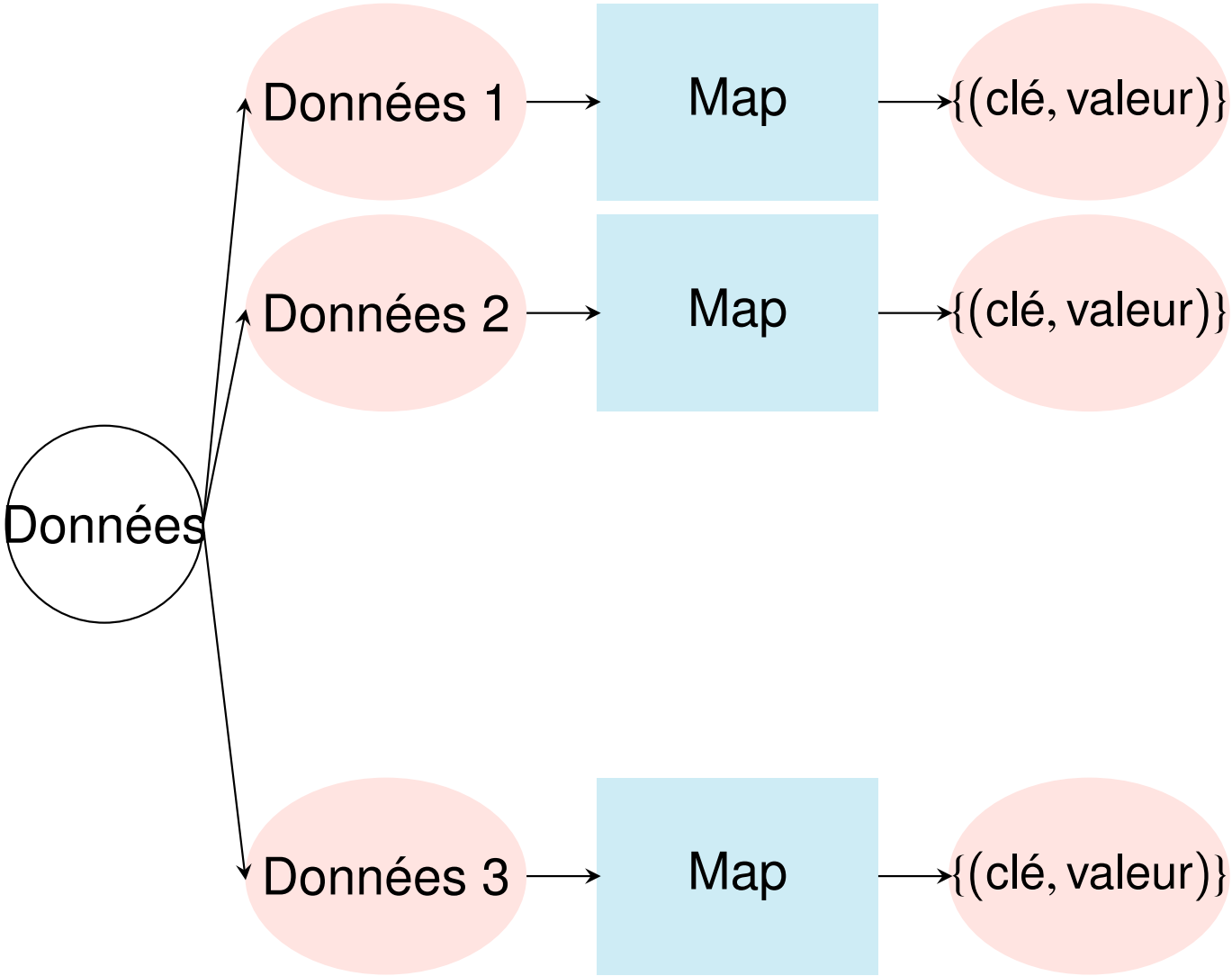
- Approche standard (séquentielle)
  - les données sont lues de manière séquentielle
  - un vecteur contenant la valeur courante de la somme des ventes pour chaque boutique est mis à jour à chaque nouvelle lecture
- Approche MR

MapReduce



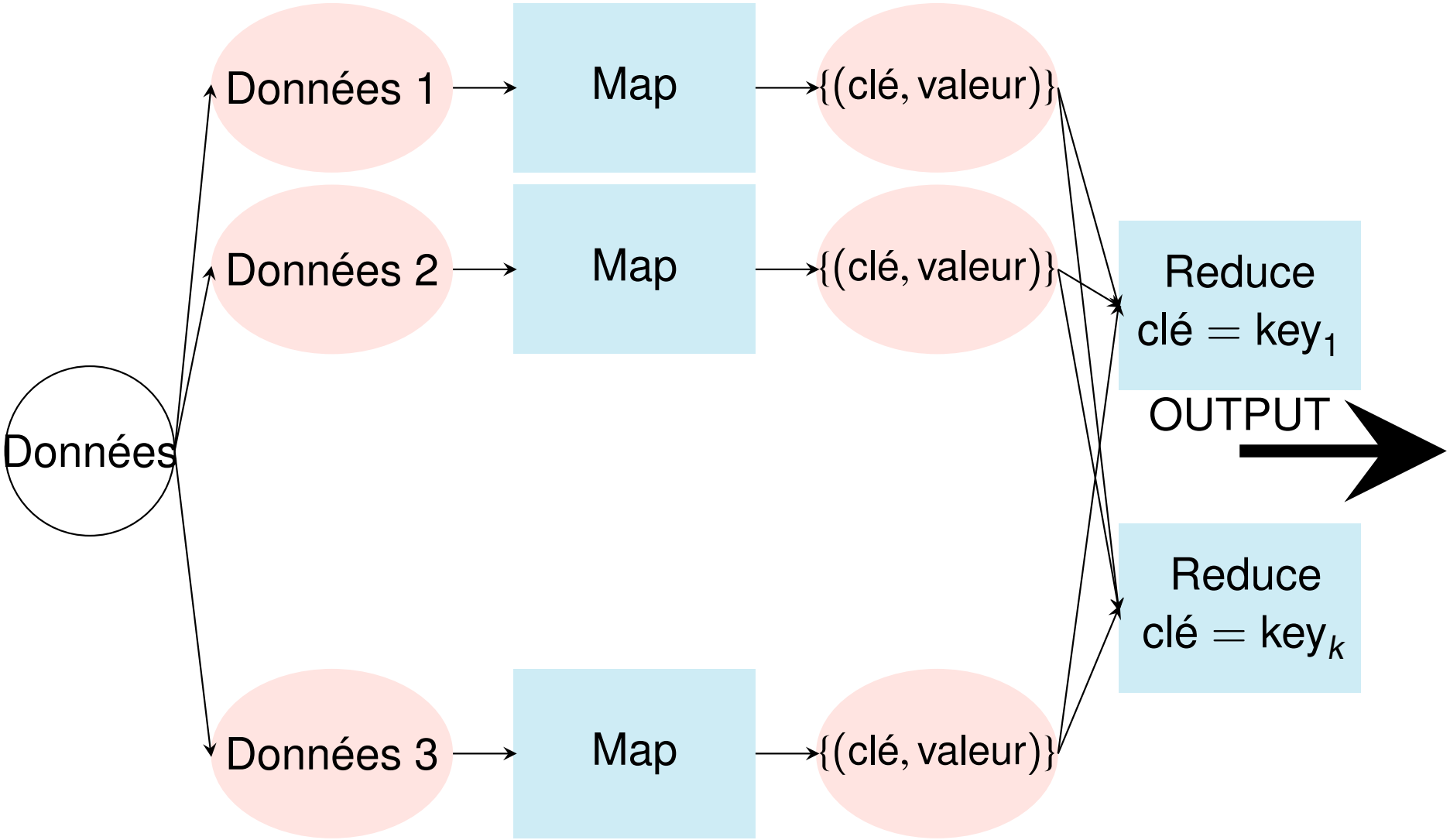
Les données sont séparées en plusieurs morceaux.

MapReduce



Map : lire la ligne et générer une paire clé=boutique et valeur=montant.

MapReduce



Reduce : pour chaque clé (*i.e.*, boutique), calculer la somme des valeurs.

Bilan et évaluation: M1 E & S

Déroulement du cours vision de l'enseignant

- conditions matérielles correctes pour illustrer le cours sauf la partie MapReduce (erreur pour l'utilisation de **rmr2** sur les ordinateurs de l'université)
- très bon gradient de progression des étudiants, notamment au niveau prise en main de R, qui semblaient très motivés

Bilan et évaluation: M1 E & S

Déroulement du cours vision de l'enseignant

- conditions matérielles correctes pour illustrer le cours sauf la partie MapReduce (erreur pour l'utilisation de **rmr2** sur les ordinateurs de l'université)
- très bon gradient de progression des étudiants, notamment au niveau prise en main de R, qui semblaient très motivés

Déroulement du cours vision de l'étudiant (recueillie via une enquête post-cours ; 15 répondants)

- ont eu le sentiment d'avoir "mis le pied à l'étrier" en calcul parallèle / données massives
- très satisfaits de la forme du cours (grande place à la mise en œuvre pratique).

Bilan et évaluation: M1 E & S

Déroulement du cours vision de l'enseignant

- conditions matérielles correctes pour illustrer le cours sauf la partie MapReduce (erreur pour l'utilisation de **rmr2** sur les ordinateurs de l'université)
- très bon gradient de progression des étudiants, notamment au niveau prise en main de R, qui semblaient très motivés

Déroulement du cours vision de l'étudiant (recueillie via une enquête post-cours ; 15 répondants)

- ont eu le sentiment d'avoir "mis le pied à l'étrier" en calcul parallèle / données massives
- très satisfaits de la forme du cours (grande place à la mise en œuvre pratique).

Perspectives

- ajouter une séance pour mieux appréhender MapReduce
- concepts uniquement introduits : prévoir une suite au niveau M2 avec un projet plus proche de la réalité... ?