Quels filets pour attraper les gros poissons ?

Les outils du Big Data





Séminaire de méthodes statistiques
Insee
Big Data

Insee
Mesurer pour comprendre

How big are you ?

- Le big data suis-je concerné ?
- Première question :
  - la taille
  - et la croissance
- Des technologies pour y répondre
- Ce que ça complique pour le statisticien

Plan

- Les gros poissons : une approche nouvelle pour stocker et traiter les données
  - L'approche distribuée
  - L'exemple d'Hadoop
- Les gros filets : l'évolution pour le statisticien
  - Des nouveaux logiciels (régulièrement nouveaux)
  - Une adaptation des outils classiques
    - Exploration
    - Approximations

Les gros poissons

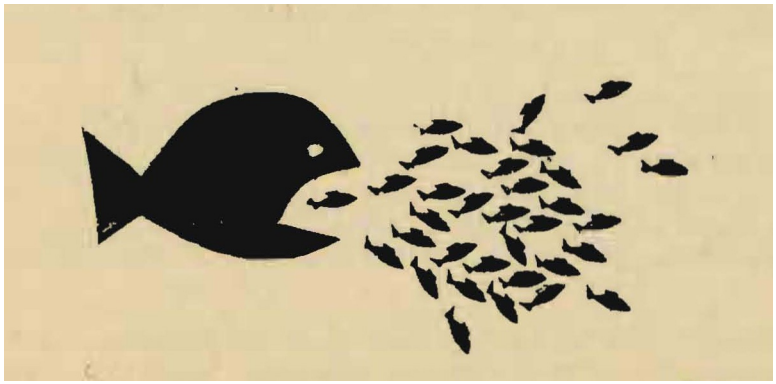
Nécessité de nouveaux types de BdD, un problème d'abord informatique

Limite de taille :

capacité du disque

Limite de performance :

capacité de calcul



Limite de fiabilité :

bon fonctionnement du serveur

Des ordres de grandeur

Données de caisse :

40Go / semaine

Données Orange de téléphonie :

3 milliards d'événements, 22 M utilisateurs / mois, en 2007 (sur le serveur d'Orange)

De quoi demain sera fait ?

Nécessité de pouvoir passer à l'échelle

Base de données distribuée

Limite de taille :

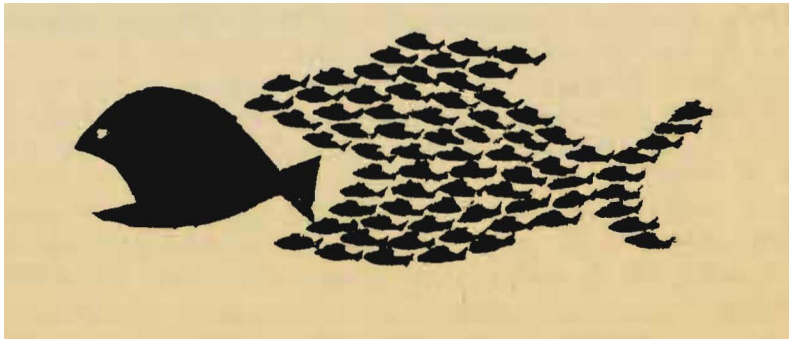
somme des capacités des éléments uniques

Limite de performance :

répartition des requêtes entre les différents nœuds du système distribué

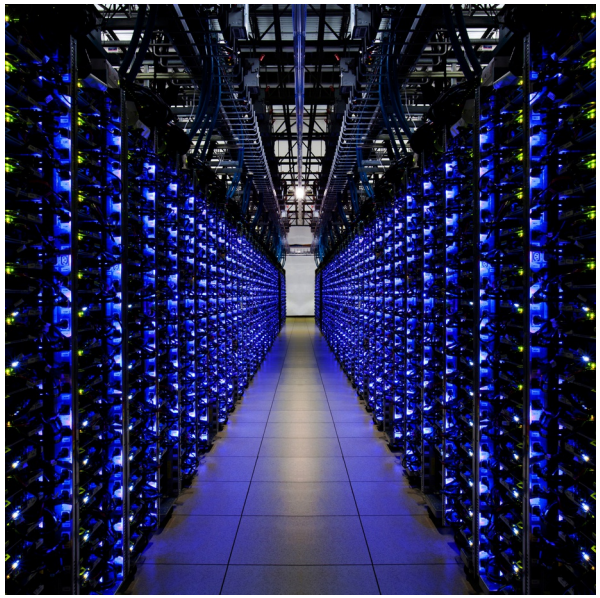
Limite de fiabilité :

des nœuds interchangeables



Système distribué

Un ensemble de calculateurs, communiquant à travers une connexion de réseau partagée, qui coopèrent pour résoudre un même problème en se présentant pour l'extérieur comme un unique calculateur



Approche distribuée

Avantages :

- **Scalability**
- **Temps de calcul**

Inconvénients :

- Compétences à acquérir
- Technologies développées d'abord pour le requêtage

Pas de critère strict aujourd'hui
pour savoir quand passer au distribué

Mise en œuvre par exemple
pour **données de caisse** et **données mobiles**

HDFS :

hadoop distributed file system

Données enregistrées dans une approche distribuée

Un environnement complet et performant :

map-reduce

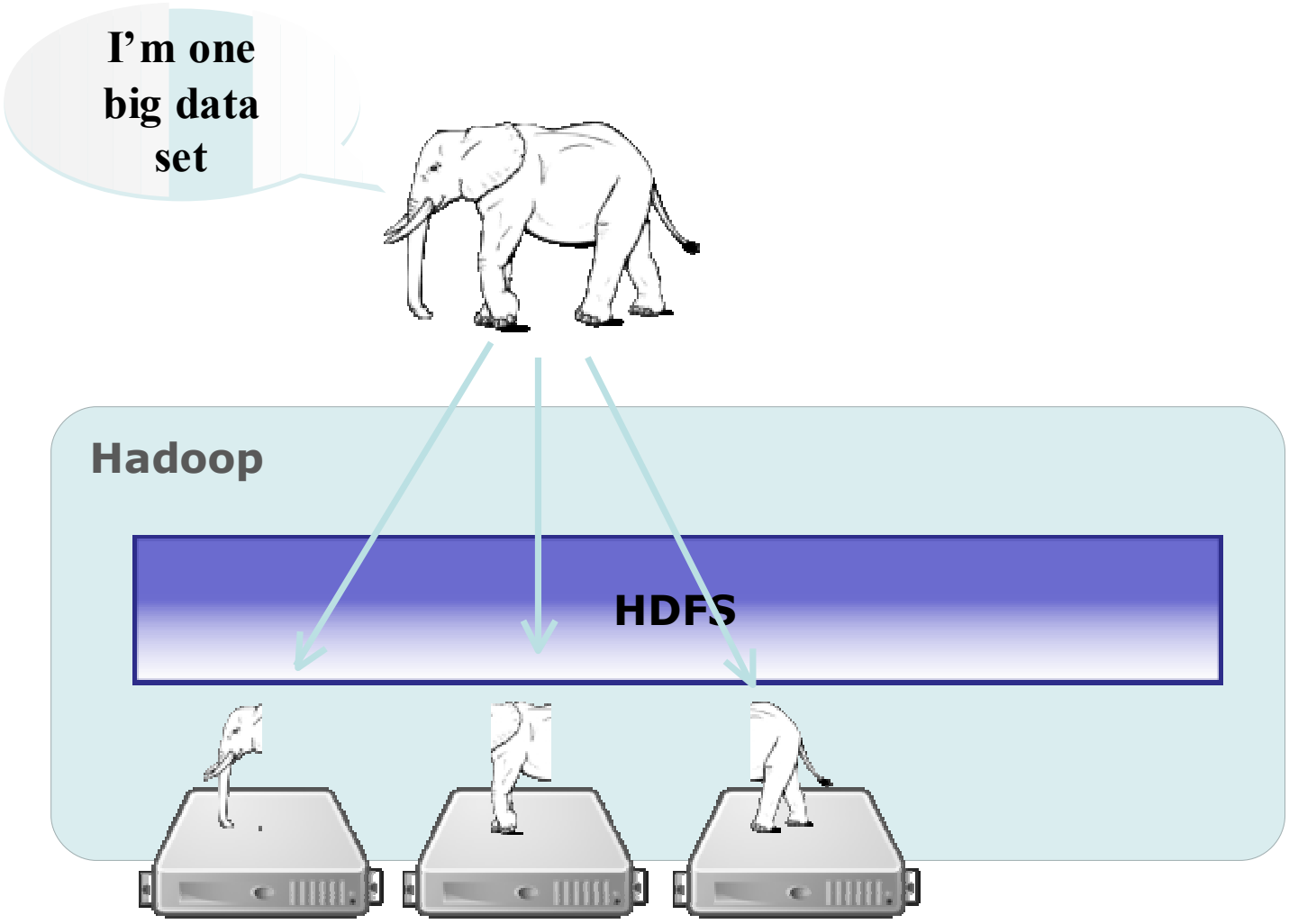
Lettre Big Data n°3

http://www.agora.insee.fr/jahia/Jahia/site/dmcsi/Actu/DMS/DMAEE/BigData

Blog : Statoscope

https://statoscope.wordpress.com/

Des données distribuées



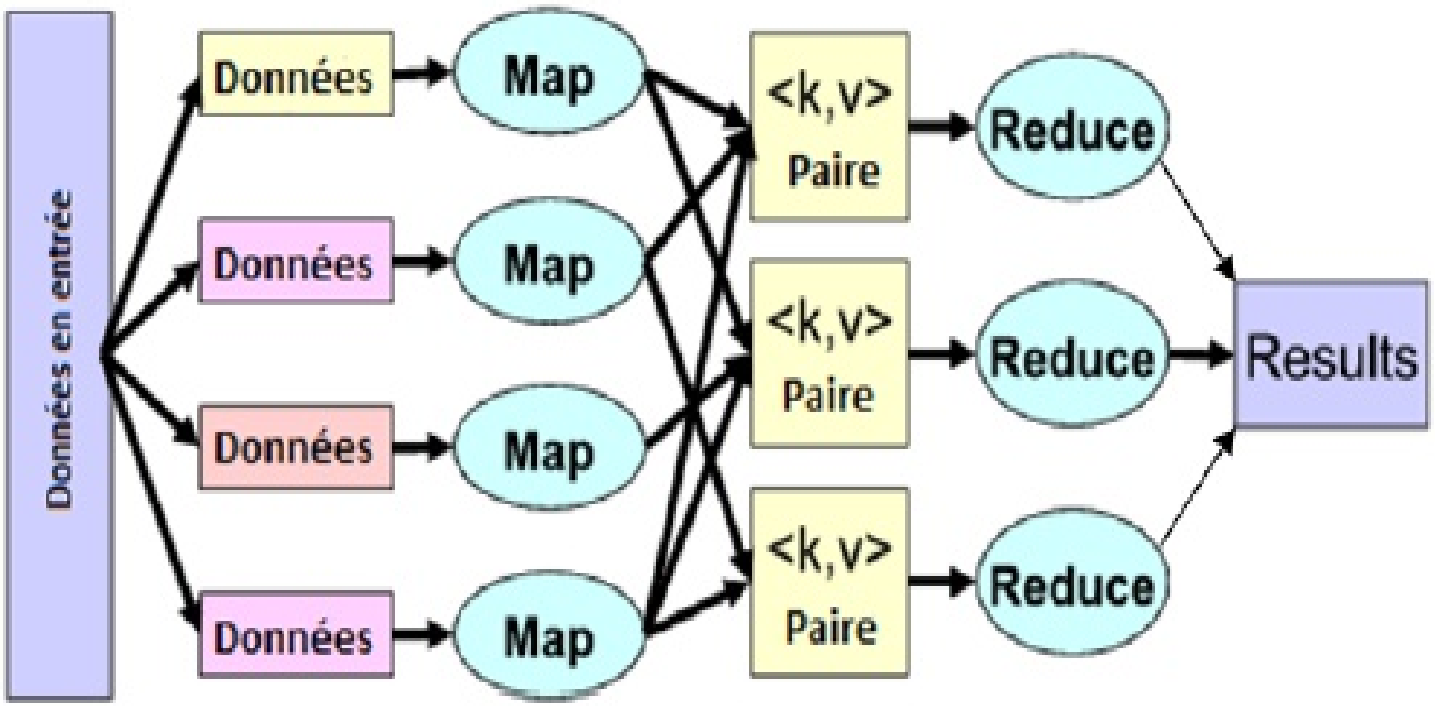
Le principe :

- Une phase de décomposition
- Une phase d'agrégation

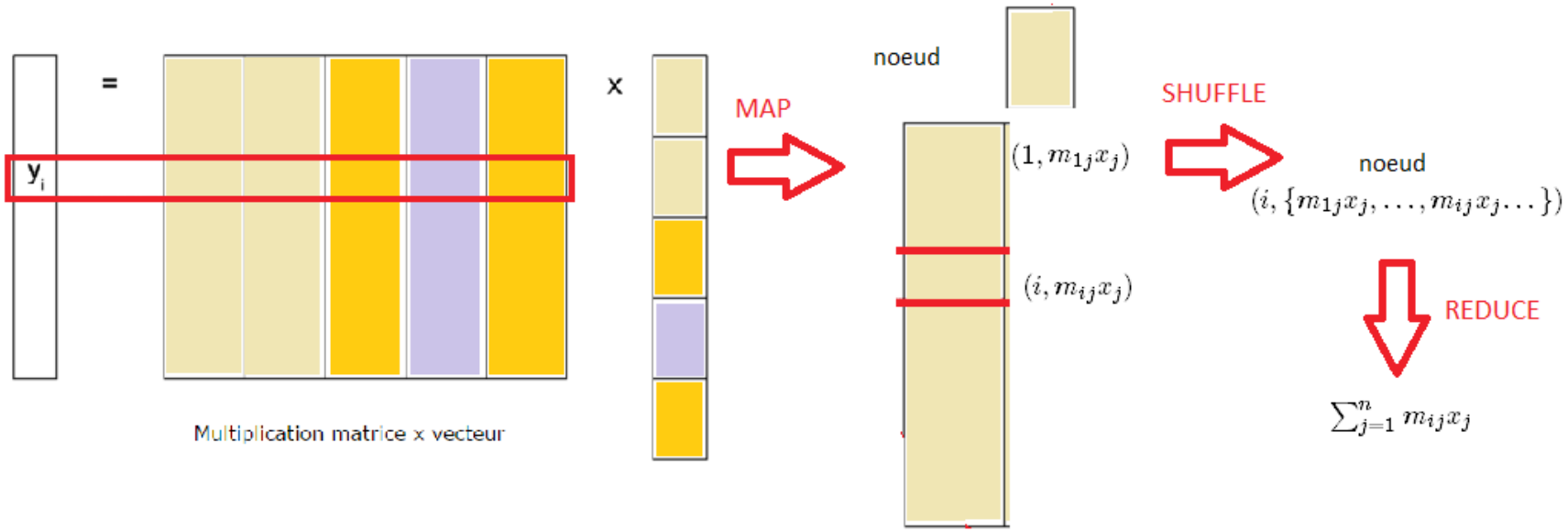
Mise à profit de l'infrastructure distribuée :

- Une parallélisation des tâches


MapReduce


Données en entrée
Données
Données
Données
Données
Map
Map
Map
Map
<k,v>
Paire
<k,v>
Paire
<k,v>
Paire
Reduce
Reduce
Reduce
Results

- Un exemple : multiplication de matrice



Une amélioration de map-reduce

Mémoire vive

Gain de temps jusqu'à x10

Les gros filets

Les implications pour le statisticien

D'autres architectures que Hadoop
connues des informaticiens

Des nouveaux outils logiciels
et tous les jours plus nouveaux

Des changements dans les outils d'analyse

Et dans la pratique : quels logiciels ?

Des requêtes dans des simili SQL

CypherQL, QL…

Des langages pour utiliser directement mapreduce :

Java, Python…

Des nouveaux outils :

Pig, Hive…

Mais heureusement...

Des outils nouveaux qui rentrent dans les anciens

RHadoop

PySpark

Spark'R

SAS in-Memory Statistics for Hadoop

C'est pas si compliqué !

- Une utilisation transparente
  de la structure distribuée des données
- Les langages sont apparus parce que
  le format de stockage a changé
- Plus besoin de s'en soucier
- Rapidement accessible
  avec un petit bagage de R ou Python

Une exploration coûteuse

Phase d'exploration plus sensible

- Temps des traitements
- Signal / bruit : avoir une idée de ce qu'on veut regarder

Représentation

- Tableau, peut se brancher au dessus de la pile Hadoop

Un problème pas inconnu mais amplifié

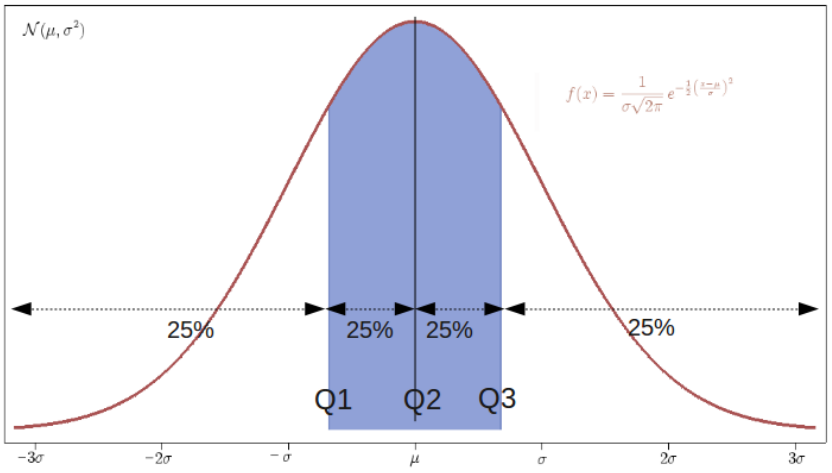
Des calculs approximés

Exemple :

```
df.approxQuantile("x", [0.5], 0.25)
```

calcul approximé pour les quantiles

on doit donner la précision en input !



Des régressions pas encore aussi fluides

Refaire les régressions à la main ? Inversion de la matrice à calculer…

Rapport de stage de Stéphanie Himpens

http://www.agora.insee.fr/jahia/Jahia/site/dmcsi/Actu/DMS/DMAEE/BigData

Outils se développent plus pour le ML

que l'économétrie

Des librairies demachine learning

Regression avec la librairie mllib de Spark

- préciser le nombre d'itérations et la taille des pas de la descente de gradient
- pas d'écart-type en sortie

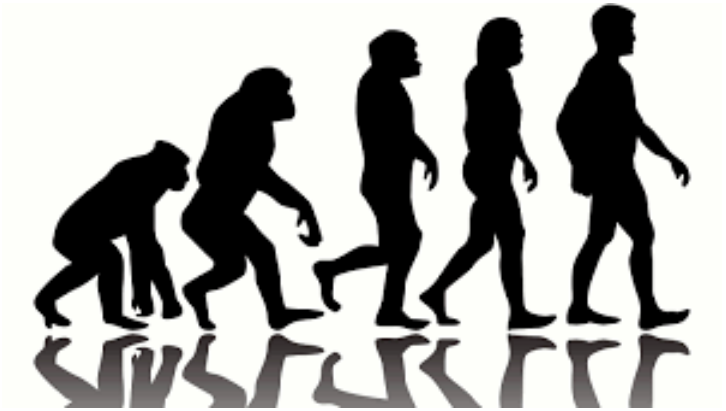
N trop grand :

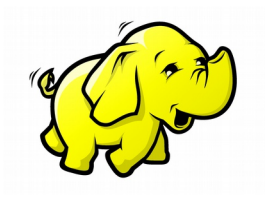
tout est significatif !

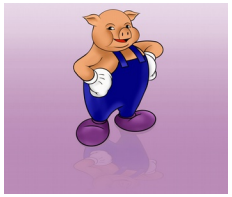
Une période de constante évolution

Des nouveaux packages qui apparaissent régulièrement

Évolution dans le sens d'un usage plus aisé







Une période de constante évolution

Adaptation rapide

Adaptation réciproque :

les outils s'adaptent aux usagers

Rythme très soutenu des évolutions :

celui des R&D de Google, Yahoo, Linkedin, Amazon…

Conclusions

- Le dialogue avec l'informatique
  - Une évolution qui vient de l'informatique
  - Hackathon de Eurostat à Bruxelles : une équipe mixte
- Flexibilité
  - Les outils changent rapidement mais dans le sens d'un usage plus aisé
  - Les jeunes arrivent formés à ces questions

Des questions ?