Statistique et Big Data Analytics; Volumétrie, L'Attaque des Clones

Philippe Besse

Statistique et *Big Data Analytics*
Volumétrie – L'Attaque des *Clones*

Philippe Besse*

26 mai 2014

Résumé : Cette article suppose acquises les compétences et expertises d'un statisticien en apprentissage non supervisé (NMF, k-means, svd) et supervisé (régression, cart, random forest). Quelles compétences et savoir faire ce statisticien doit-il acquérir pour passer à l'échelle "*Volume*" des grandes masses de données ? Après un rapide tour d'horizon des différentes stratégies offertes et principalement celles imposées par l'environnement *Hadoop*, les algorithmes des quelques méthodes d'apprentissage disponibles sont brièvement décrits pour comprendre comment ils sont adaptés aux contraintes fortes des fonctionnalités *Map-Reduce*.

Mots-clefs : Statistique ; Fouille de Données ; Grande Dimension ; Apprentissage Statistique ; Datamasse ; algorithmes ; Hadoop, Map-Reduce ; Scalability.

Abstract : This article assumes acquired the skills and expertise of a statistician in unsupervised (NMF, k-means, SVD) and supervised learning (regression, CART, random forest). What skills and knowledge do the statistician must acquire it to reach the "Volume" scale of big data ? After a quick overview of the different strategies available and especially those imposed by Hadoop, algorithms of some available learning methods are outlined to understand how they are adapted to high stresses of Map-Reduce functionalities.

Keywords : Statistics ; Data Mining ; High DImension ; Statistical learning ; Big Data ; algorithms ; Hadoop ; Map-Reduce ; Scalability.

*Université de Toulouse – INSA, Institut de Mathématiques, UMR CNRS 5219

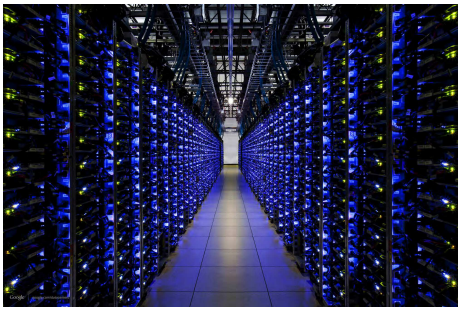


FIGURE 1 – *Armée de* clones*, baies de serveurs, alignés par milliers dans le hangar d'un centre de données de Google.*

1 Introduction

1.1 Motivations

L'historique récent du traitement des données est schématiquement relaté à travers une odyssée : *Retour vers le Futur III* (Besse et al. ; 2014)[2] montrant comment un statisticien est successivement devenu prospecteur de données, bio-informaticien et maintenant *data scientist*, à chaque fois que le volume qu'il avait à traiter était multiplié par un facteur mille. Cette présentation rappelle également les trois aspects : volume, variété, vélocité, qui définissent généralement le *Big Data* au sein d'un *écosystème*, ou plutôt une jungle, excessivement complexe, concurrentielle voire, conflictuelle, dans laquelle le mathématicien / statisticien peine à se retrouver ; ce sont ces difficultés de béotiens qui motivent la présentation d'une nouvelle saga : *Data War* en 5 volets dont seulement trois seront interprétées.

La Menace Fantôme est jouée avec la NSA dans le rôle de *Big Brother*,

L'Attaque des Clones contre la *Volumétrie* est l'objet du présent article avec les baies ou conteneurs d'empilements de serveurs dans le rôle des clones (cf. figure 1).

La Revanche des Maths intervient dans la vignette (à venir) consacrée à la *Variété* (courbes, signaux, images, trajectoires, chemins...) des données qui rend indispensable des compétences mathématiques en modélisation, notamment pour des données industrielles, de santé publique...

Le Nouvel Espoir est d'aider au transfert des technologies et méthodes efficaces issues du e-commerce vers d'autres domaines (santé, climat, énergie...) et à d'autres buts que ceux mercantiles des origines.

L'Empire Contre-Attaque avec GAFA [1] dans le rôle principal de la vignette (à

1. Google, Apple, Facebook, Amazon

venir) qui aborde les problèmes de flux ou *Vélocité* des données.

Une fois les données stockées, ou leur flux organisé, leur *Valorisation* nécessite une phase *Analytics*. L'objectif est de tenter de pénétrer cette jungle pour en comprendre les enjeux, y tracer des sentiers suivant les options possibles afin d'aider à y faire les bons choix. Le parti est pris d'utiliser, si possible au mieux, les ressources logicielles (langages, librairies) *open source* existantes en tentant de minimiser les temps de calcul tout en évitant la programmation, souvent reprogrammation, de méthodes au code par ailleurs efficace. Donc minimiser, certes les temps de calcul, mais également les coûts humains de développement pour réaliser des premiers prototypes d'analyse.

L'exploration aventureuse d'une littérature d'une littérature électronique particulièrement massive mais redondante fait émerger un ensemble de :

1.2 Questions

- Beaucoup des exemples traités, notamment ceux des besoins initiaux, se résument principalement à des dénombrements, d'occurrences de mots, d'événements (décès, retards...) de co-occurences (règles d'association). Pour ces objectifs, les architectures et algorithmes parallélisés avec *MapReduce* sont efficaces lorsqu'ils traitent l'ensemble des données même très volumineuses. Pour d'autres objectifs, on s'interroge sur la réelle nécessité, en terme de qualité de prévision d'un modèle, d'estimer celui-ci sur l'ensemble des données plutôt que sur un échantillon représentatif stockable en mémoire.
- Un accroissement de la mémoire ne résoudrait-il pas le problème ? Ou l'utilisation d'une librairie (R) simulant cet accroissement ?
- A partir de quel volume une architecture spécifique de stockage comme *Hadoop* s'avère-elle nécessaire ? Ou alors est elle donnée *a priori* ?
- Bien que dominant, *Hadoop* est-il bien le meilleur choix d'architecture en fonction du secteur d'activité et des objectifs poursuivis ?
- Quelles sont les procédures (quel langage ?) d'extraction disponibles pour l'architecture des données utilisée ? Quelle interface avec quel logiciel statistique ou d'apprentissage ?
- Quels algorithmes de modélisation et apprentissage sont-ils disponibles dans l'environnement utilisés ?

Questions dont les réponses dépendent évidemment de l'objectif. S'il semble pertinent en e-commerce de considérer tous les clients potentiels en première approche, ceci peut se discuter pas-à-pas en fonction du domaine et de l'objectif précis : *clustering*, score, système de recommandation..., visé.

1.3 Prépondérance de Hadoop ?

Le traitement de grandes masses de données impose une parallélisation des calculs pour obtenir des résultats en temps raisonnable et ce, d'autant plus, lors du traitement en temps réel d'un flux (*streaming*) de données. Les acteurs majeurs que sont Google, Amazon, Yahoo, Twitter... développent dans des centres de données (*data centers*) des architectures spécifiques pour stocker à moindre coût de grande masses de données (pages web, listes de requêtes, clients, articles à vendre, messages...) sous forme brute, sans format ni structure relationnelle. Ils alignent, dans de grands hangars, des empilements (conteneurs ou baies) de cartes mère standards de PC "bon marché" [2], reliées par Ethernet (cf. figure 1).

Le gestionnaire de fichier le plus populaire dans ce contexte est *Hadoop* qui inclut des fonctionnalités dites *MapReduce* (Dean et Ghemawat ; 2004)[4] de parallélisation des traitements sur l'ensemble des serveurs affectés chacun d'un espace disque. Initié par Google, il est maintenant développé dans le cadre de la fondation Apache. Le stockage intègre également des propriétés de duplication des données afin d'assurer une tolérance aux pannes. Lorsqu'un traitement se distribue en étapes *Map* et *Reduce*, celui-ci devient "échelonnable" ou *scalable* avec un temps de calcul en principe divisé par le nombre de nœuds ou serveurs impliqués. *Hadoop* est diffusé comme logiciel libre et bien qu'écrit en java, tout langage de programmation peut l'interroger et exécuter les étapes *MapReduce* prédéterminées.

Comparativement à d'autres solutions de stockage plus sophistiquées : cube, SGBDR (systèmes de gestion de base de données relationnelles), disposant d'un langage (SQL) complexe de requêtes, et comparativement aux architectures des machines massivement parallèles, *Hadoop* est perçu, sur le plan académique, comme une régression. Cette architecture et les fonctionnalités très restreintes de *MapReduce* ne peuvent rivaliser avec des programmes utilisant des langages et librairies spécifiques aux machines massivement parallèles connectant des centaines, voire milliers, de cœurs sur le même bus. Néanmoins, le poids des acteurs qui utilisent

2. La facture énergétique est le poste de dépense le plus important : l'électricité consommée par un serveur sur sa durée de vie coûte plus que le serveur lui-même. D'ou l'importance que Google apporte aux question environnementales dans sa communication.

Hadoop et le bon compromis qu'il semble réaliser entre coûts matériels et performances en font un système dominant pour les applications commerciales qui en découlent : les internautes sont des prospects à qui sont présentés des espaces publicitaires ciblés et vendus en temps réel aux annonceurs dans un système d'enchères automatiques

Hormis les applications du e-commerce qui dépendent du choix préalable, souvent *Hadoop*, *MongoDB*..., de structure, il est légitime de s'interroger sur le bon choix d'architecture (SGBD classique ou non) de la base de données en fonction des objectifs à réaliser. Ce sera notamment le cas dans bien d'autres secteurs : industrie (maintenance préventive en aéronautique, prospection pétrolière, usages de pneumatiques...), transports, santé (épidémiologie), climat, énergie (EDF, CEA...)... concernés par le stockage et le traitement de données massives.

1.4 Une R-éférence

Comme il existe de très nombreux systèmes de gestion de données, il existe de nombreux environnements logiciels *open source* à l'interface utilisateur plus ou moins *amicale* et acceptant ou non des fonctions écrites en R ou autre langage de programmation : KNIME, TANAGRA, Weka,...

Néanmoins, pour tout un tas de raisons dont celle d'un large consensus au sein d'une très grande communauté d'utilisateurs, le logiciel libre R[12] est une référence pour l'analyse ou l'apprentissage statistique de données conventionnelles, comme pour la recherche et le développement, la diffusion de nouvelles méthodes. Le principal problème de R (version de base) est que son exécution nécessite de charger toutes les données en mémoire vive. En conséquence, dès que le volume est important, l'exécution se bloque. Aussi, une première façon de procéder avec R, pour analyser de grandes masses, consiste simplement à extraire une table, un sous-ensemble ou plutôt un *échantillon représentatif* des données avec du code java, Perl, Python, Ruby... avant de les analyser dans R.

Une autre solution consiste à utiliser une librairie permettant une extension virtuelle sur disque de la mémoire vive. C'est le rôle des packages `ff, bigmemory`, mais qui ne gèrent pas le parallélisme, et également aussi de ceux interfaçant R et *Hadoop* qui seront plus précisément décrits ci-dessous.

Donc utiliser *Hadoop* s'il n'est pas possible de faire autrement :

– parti pris de vouloir tout traiter, sans échantillonnage, avec impossibilité d'ajouter plus de mémoire ou temps de calcul rédhibitoire avec de la mé-

moire virtuelle,
- volume trop important ou absence de structure des données,
- la méthodologie à déployer pour atteindre l'objectif se décompose en formulation *MapReduce*,
- l'architecture *Hadoop* est imposée, existe déjà ou est facile à faire mettre en place par les personnes compétentes.

Ensuite, utiliser R si cela s'avère utile, c'est-à-dire si des méthodes d'apprentissage ou de modélisation statistique sont prévues. Compter des occurrences de mots, calculer des moyennes... ne nécessitent pas la richesse méthodologique de R qui prendrait beaucoup plus de temps pour des calculs rudimentaires. Adler (2010)[1] compare trois solutions pour dénombrer le nombre de décès par sexe aux USA en 2009 : 15 minutes avec RHadoop sur un cluster de 4 serveurs, une heure avec un seul serveur, et 15 secondes, toujours sur un seul serveur, mais avec un programme Perl.

Bien entendu la richesse des outils graphiques de R (*i.e.* `ggplot2`) permet de visualiser de façon très élaborée les résultats obtenus sans changer d'environnement.

En résumé : *Hadoop* permet à R d'accéder à des gros volumes de données en des temps raisonnables mais rester conscient que, si c'est sans doute la stratégie la plus "simple" à mettre en œuvre, ce ne sera pas la solution la plus efficace en temps de calcul en comparaison d'autres comme *Mahout* ou *Spark*.

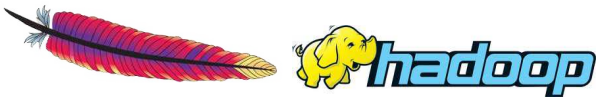
Dans le cas contraire, les compétences nécessaires à l'utilisation des méthodes décrites dans les parties Exploration et Apprentissage de wikistat suffisent. Seule la forte imbrication entre structures de données, algorithmes de calcul, méthodes de modélisation ou apprentissage, impose de poursuivre la lecture de cet article et l'acquisition des apprentissages qu'il développe.

Cet article se propose de décrire rapidement l'écosystème, surtout logiciel, des grandes masses de données avant d'aborder les algorithmes et méthodes de modélisation qui sont employés. L'objectif est d'apporter des éléments de réponse en conclusion notamment en terme de formation des étudiants. Les principes des méthodes d'apprentissage ne sont pas rappelés, ils sont décrits sur le site coopératif wikistat.

2 Environnements logiciels

2.1 *Hadoop*

Hadoop est un projet *open source* de la fondation Apache dont voici les principaux mots clefs :



HDFS (*hadoop distributed file system*) système de fichiers distribués sur un ensemble de disques (un disque par nœud ou serveur) mais manipulés et vus de l'utilisateur comme un seul fichier. Réplication des données sur plusieurs hôtes pour la fiabilité. HDFS est *fault tolerant*, sans faire appel à des duplications par système RAID, si un hôte ou un disque a des problèmes .

MapReduce au coeur de la parallélisation des traitements.

- *Map phase* : division en sous-ensembles et exécution en parallèle par chaque serveur sur son disque. Pour chaque "ligne" du fichier ou enregistrement, émission d'une *mapped* paire : *key* et *value*, comme sortie de la phase de *Map*.
- S'il est prévu plusieurs nœuds, une éventuelle phase intermédiaire, *Combine* ou *Shuffle*, est nécessaire pour attribuer les paires (clef, valeur) aux nœuds de la phase suivante.
- *Reduce phase* : le ou les nœuds maîtres collectent les réponses : paires (clef, valeur), de tous les sous-problèmes et les assemblent pour le résultat.

Hive langage développé par Facebook pour interroger une base *Hadoop* avec une syntaxe proche du SQL (*hql ou Hive query language*).

Hbase base de données orientée "colonne" et utilisant HDFS.

Pig langage développé par Yahoo similaire dans sa syntaxe à Perl et visant les objectifs de *Hive* en utilisant le langage *pig latin*.

Sqoop pour (Sq)l to Had(oop) permet le transfert entre bases SQL et Hadoop.

Plusieurs distributions *Hadoop* sont proposées : Apache, Cloudera, Hortonworks à installer de préférence sous Linux. Ces distributions proposent également des machines virtuelles pour faire du *Hadoop* sans *Hadoop* et donc au moins tester les scripts en local avant de les lancer en vraie grandeur. La mise en place d'un tel environnement nécessite des compétences qui ne sont pas du tout abordées ; il est également possible d'utiliser les services d'AWS (Amazon Web Services), qui offre un premier niveau d'utilisation gratuit tout en nécessitant une carte de crédit ! C'est souvent cette solution qui est utilisée par les *start-up* qui prolifèrent dans le e-commerce mais elle ne peut être retenue par une entreprise industrielle pour des contraintes évidentes de confidentialité, contraintes et suspicions qui freinent largement le déploiement de *clouds*.

2.2 *Hadoop streaming*

Bien qu'*hadoop* soit écrit en java, il n'est pas imposé d'utilisé ce langage pour analyser les données stockées dans une architecture *Hadoop*. Tout langage de programmation (Python, C, java... et même R) manipulant des chaînes de caractères peut servir à écrire les étapes *MapReduce* pour enchaîner les traitements. Bien que sans doute plus efficace, l'option de traitement en *streaming* n'est pour l'instant pas développée dans cet article pour favoriser celle nécessitant moins de compétences ou de temps de programmation.

2.3 Mahout

Mahout (Owen et al. 2011)[9] est également un projet de la fondation Apache. C'est une Collection de méthodes programmées en java qui s'exécute sur *Hadoop* mais *Hadoop* n'est pas indispensable.

Une communauté d'utilisateurs développent de façon collaborative les codes java permettant d'exécuter les méthodes les plus classiques de modélisation et apprentissage sur des bases *Hadoop* ou non, sans utiliser R ou tout autre logiciel dédié. Des compétences en java sont évidemment indispensables.

2.4 Spark

L'architecture *Hadoop* de fichiers distribuées est particulièrement adaptée au stockage et à la parallélisation de tâches élémentaires. La section suivante sur les algorithmes illustre les fortes contraintes imposées par les fonctionnalités *MapReduce* pour le déploiement de méthodes d'apprentissage. Dès qu'un algorithme est itératif, (*i.e.* k-means), chaque itération communique avec la suivante par une écriture puis lecture dans la base HDFS. C'est extrêmement contraignant et pénalisant pour les temps d'exécution, car c'est la seule façon de faire communiquer les nœuds d'un cluster en dehors des fonctions *MapReduce*. Les temps d'exécution sont généralement et principalement impactés par la complexité d'un algorithme et le temps de calcul. Avec l'architecture *Hadoop* et un algorithme itératif, ce sont les temps de lecture et écriture qui deviennent très pénalisants. Plusieurs solutions à l'état de prototype (HaLoop, Twister...) ont été proposées pour résoudre ce problème. *Spark* semble la plus prometteuse et la plus soutenue par la communauté.

Spark est devenu un projet *open source* de la fondation Apache dans la continuation des travaux du laboratoire amplab de l'Université Berkley. L'objectif est simple mais son application plus complexe surtout s'il s'agit de préserver les propriétés de tolérance aux pannes. Il s'agit de garder en mémoire les données entre deux itérations des étapes *MapReduce*. Ceci est fait selon un principe abstrait de mémoire distribuée : *Resilient Distributed Datasets* (Zaharia et al. 2012)[14]. Cet environnement est accessible en java, Scala [3], Python et bientôt R (librairie SparkR). Il est accompagné d'outils de requêtage (Shark), d'analyse de graphes (GraphX) et d'une bibliothèque en développement (MLbase) de méthodes d'apprentissage.

2.5 Librairies R

R parallèle

Le traitement de données massives oblige à la parallélisation des calculs et de nombreuses solutions sont offertes dont celles, sans doute les plus efficaces, utilisant des librairies ou extensions spécifiques des langages les plus courants (Fortran, java, C...) ou encore la puissance du GPU (Graphics Processing Unit) d'un simple PC. Néanmoins les coûts humains d'écriture, mise au point des programmes d'interface et d'analyse sont à prendre en compte. Aussi, de nombreuses librairies ont été développées pour permettre d'utiliser au mieux, à partir de R, les capacités d'une architecture (cluster de machines, machine multi-cœurs, GPU) ou d'une structure de données comme *Hadoop*. Ces librairies sont listées et décrites sur la page dédiée au High-Performance and Parallel Computing with R[12] du CRAN (Comprehensive R Archive Network). La consultation régulière de cette page est indispensable car, l'*écosystème* R est très mouvant, voire volatile ; des librairies plus maintenues disparaissent à l'occasion d'une mise à jour de R (elles sont fréquentes) tandis que d'autres peuvent devenir dominantes.

Le parti est ici pris de s'intéresser plus particulièrement aux librairies dédiées *simultanément* à la parallélisation des calculs et aux données massives. Celles permettant de gérer des données hors mémoire vive comme `bigmemory` d'une part ou celles spécifiques de parallélisation comme `parallel, snow`... d'autre part (cf. McCallum et Weston ; 2011)[8] sont volontairement laissées de côté.

3. Langage de programmation développé à l'EPFL (Zurich) et exécutable dans un environnement java. C'est un langage fonctionnel adapté à la programmation d'application parallélisable.

R & *Hadoop*

Prajapati (2013)[11] décrit différentes approches d'analyse d'une base *Hadoop* à partir de R.

`Rhipe` développée à l'université Purdue n'est toujours pas une librairie officielle de R et son développement semble en pause depuis fin 2012.

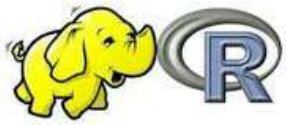
`HadoopStreaming` permet comme son nom l'indique de faire du *streaming hadoop* à partir de R ; les programmes sont plus complexes et, même si plus efficace, cette solution est momentanément laissée de côté.

`hive` Hadoop InteractiVE est une librairie récente (02-2014), dont il faut suivre le développement, évaluer les facilités, performances. Elle pourrait se substituer aux services de Rhadoop listé ci-dessous.

`segue` est dédiée à l'utilisation des services d'Amazon (AWS) à partir de R.

RHadoop

Revolution Analytics est une entreprise commerciale qui propose autour de R du support, des services, des modules d'interface (SAS, SPSS, Teradata, Web, AWS,...) ou d'environnement de développement sous Windows, de calcul intensif. Comme produit d'appel, elle soutient le développement de Rhadoop, ensemble de librairies R d'interface avec *Hadoop* pour la distribution *cloudera* et encore librement accessible :



`rhdfs` pour utiliser les commandes HDFS d'interrogation d'une base Hadoop.

`rhbase` pour utiliser les commandes HBase d'interrogation.

`rmr2` pour exécuter du *MapReduce* ; une option permet de tester des scripts (simuler un cluster) sans hadoop.

`plyrmr` utilise la précédente et introduit des facilités.

Quelques ressources pédagogiques sont disponibles notamment pour installer RHadoop sur une machine virtuelle *cloudera*.

Revolution Analytics promeut les librairies de RHadoop avec l'argument recevable que cette solution conduit à des programmes plus courts, comparativement à `HadoopStreaming, RHipe, hive`, donc plus simples et plus rapides à mettre en place mais pas nécessairement plus efficaces en temps de calcul. Une

expérimentation en vraie grandeur réalisée par EDF[4] montre des exécutions dix fois plus longues de l'algorithme k-means avec *RHadoop* qu'avec la librairie *Mahout*.

SparkR

Le site collaboratif de SparkR fournit les premiers éléments pour installer cette librairie. Celle-ci étant très récente (janvier 2014), il est difficile de se faire un point de vue sans expérimentation en vraie grandeur. Son installation nécessite celle préalable de *Scala*, *Hadoop* et de la librairie `rjava`.

3 Algorithmes

Le développement de cette section suit nécessairement celui des outils disponibles. Elle sera complétée dans les version à venir de cet article et accompagnée par un tuteuriel.

3.1 Choix en présence

Après avoir considéré le point de vue "logiciel", cette section s'intéresse aux méthodes programmées ou programmables dans un environnement de données massives géré principalement par *Hadoop*. Certains choix sont fait *a priori* :

- Si un autre SGBD est utilisé, par exemple MySQL, celui-ci est très généralement interfacé avec R ou un langage de requête permet d'en extraire les données utiles. Autrement dit, mis à part les questions de parallélisation spécifiques à *Hadoop* (MapReduce), les autres aspects ne posent pas de problèmes.
- Les livres et publications consacrées à *Hadoop* détaillent surtout des objectifs élémentaires de dénombrement. Comme déjà écrit, ceux-ci ne nécessitent pas de calculs complexes et donc de programmation qui justifient l'emploi de R. Nous nous focalisons sur les méthodes dites d'apprentissage statistique supervisé ou non.

La principale question est de savoir comment un algorithme s'articule avec les fonctionnalités *MapReduce* afin de passer à l'échelle du volume. La contrainte est forte, ce passage à l'échelle n'est pas toujours simple ou efficace et en conséquence l'architecture *Hadoop* induit une sélection brutale parmi les très nombreux algorithmes d'apprentissage. Se trouvent principalement décrits dans

4. Leeley D. P. dos Santos, Alzennyr G. da Silva, Bruno Jacquin, Marie-Luce Picard, David Worms,Charles Bernard (2012). Massive Smart Meter Data Storage and Processing on top of Hadoop. *Workshop Big Data*, Conférence VLDB (Very Large Data Bases), Istambul, Turquie.

– *Mahout* : système de recommandation, panier de la ménagère, SVD, k-means, régression logistique, classifieur bayésien naïf, random forest, SVM (séquentiel),
– *RHadoop* : k-means, régression, régression logistique, random forest.
– *MLbase* de *Spark* : k-means, régression linéaire et logistique, système de recommandation, classifieur bayésien naïf et à venir : NMF, CART, random forest.

Trois points sont à prendre en compte : la rareté des méthodes d'apprentissage dans les librairies et même l'absence de très connues comparativement à celles utilisables en R, les grandes difficultés quelque fois évoquées mais pas résolues concernant les réglages des paramètres par validation croisée ou bootstrap (complexité des modèles), la possibilité d'estimer un modèle sur un échantillon représentatif. Toutes ces raisons rendent *indispensable* la disponibilité d'une procédure d'échantillonnage simple ou équilibré pour anticiper les problèmes.

La plupart des nombreux documents disponibles insistent beaucoup sur l'installation et l'implémentation des outils, leur programmation, moins sur leurs propriétés, "statistiques". *Mahout* est développée depuis plus longtemps et les applications présentées, les stratégies développées, sont très fouillées pour des volumes importants de données. En revanche, programmée en java et très peu documentée, il est difficile de rentrer dans les algorithmes pour apprécier les choix réalisés. Programmés en R les algorithmes de *Rhadoop* sont plus abordables à un béotien pour s'initier.

3.2 Jeux de données

Des jeux de données reviennent de façon récurrente pour illustrer les fonctionnalités considérées. Cette section est à compléter notamment par des jeux de données hexagonaux. La politique de l'état et des collectivités locales d'ouvrir les accès aux données publiques peut y contribuer de même que le site de concours "data science" mais rien n'est moins sûr car la CNIL veille, à juste raison, pour rendre impossible les croisements de fichiers. Chacun de ceux-ci peut comporter beaucoup de lignes pour finalement très peu de variable ; il sera difficile de trouver et extraire des fichiers consistants.

– Données publiques de causes de mortalité en 2009 aux USA cité par Adler (2010)[1],
– historiques de cours d'actions fournis par Yahoo (Dowjones) cité par Prajapati (2013)[11],
– Prévision (cité par Prajapati ; 2013)[11] d'enchères de bulldozers sur le site kaggle qui propose régulièrement des concours. Random Forest l'a emporté sur celui-ci.
– ... à compléter

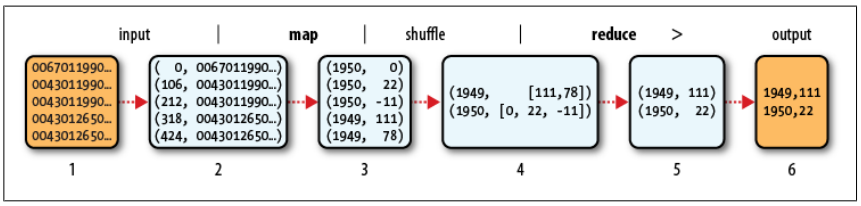


FIGURE 2 – *Flots des données dans les étapes* MapReduce *(Adler ; 2010)[1].*

3.3 MapReduce pour les nuls

Principe général

Les cinq étapes de parallélisation des calculs se déclinent selon le schéma ci-dessous. Selon l'algorithme concerné et l'architecture (nombre de serveurs) utilisée toutes ne sont pas nécessairement exécutées tandis que certaines méthodes nécessitent des itérations du processus jusqu'à convergence. Les fonctions *Map* et *Reduce* sont écrites dans un quelconque langage, par exemple R.

1. Préparer l'entrée de la phase *Map*. Chaque serveur lit sa part de la base de données ligne par ligne et construit les paires (clef, valeur) pour produire `Map input: list (k1, v1).`
2. Exécuter le programme utilisateur de la fonction `Map()` pour émettre `Map output: list (k2, v2).`
3. Répartir (*shuffle, combine* ou tri) la liste précédente vers les nœuds réducteurs en groupant les clefs similaires comme entrée d'un même nœud. Production de `Reduce input: (k2, list(v2))`. Il n'y a pas de code à fournir pour cette étape qui est implicitement prise en charge.
4. Pour chaque clef ou chaque pile, un serveur exécute le programme utilisateur de la fonction `Reduce()`. Émission de `Reduce output: (k3, v3).`
5. Le nœud maître collecte les sorties et produit, enregistre dans la base, les résultats finaux.

Adler (2010)[1] schématise avec la figure 2 le flot des données.

Exemple trivial

Considérons un fichier fictif de format texte où les champs, qui décrivent des appels téléphoniques, sont séparés par des virgules et contiennent, entre autres, les informations suivantes.

`{date},{numéro appelant},...,{numéro appelé},...,{durée}`

`Map`

- reçoit une ligne ou enregistrement du fichier,
- utilise un langage de programmation pour extraire la date et la durée,
- émet une paire : clef (date) et valeur (durée).

`Reduce`

- reçoit une paire : clef (date) et valeur (durée$_1$ ... durée$_n$),
- programme une boucle qui calcule la somme des durées et le nombre d'appels,
- calcule la moyenne,
- sort le résultat : (date) et (moyenne, nombre d'appels).

Si l'objectif est maintenant de compter le nombre d'appels par jour et par numéro d'appel, la clef de l'étape `Map` se complique tandis que l'étape `Reduce` se réduit à un simple comptage.

`Map`

- reçoit une ligne ou enregistrement du fichier,
- extrait la date et le numéro d'appel,
- émet une paire : clef (date, numéro) et valeur (1).

`Reduce`

- reçoit une clef : (date, numéro) et valeur (1...1),
- boucle qui calcule la somme des valeurs égale au nombre d'appels,
- sort le résultat : (date, numéro) et (nombre).

Toute l'astuce réside donc dans la façon de gérer les paires (clef, valeur) qui définissent les échanges entre les étapes. Ces paires peuvent contenir des objets complexes (vecteurs, matrices, listes).

3.4 Échantillonnage aléatoire simple

Des outils de base sont indispensables pour extraire un échantillon d'une base *Hadoop* et revenir à un environnement de travail classique pour disposer des outils de modélisation, choix et optimisation des modèles.

Reservoir Sampling

L'objectif est d'extraire, d'une base de taille N (pas nécessairement connu), un échantillon représentatif de taille n par tirage aléatoire simple en s'assurant que chaque observation ait la même probabilité n/N d'être retenue. Attention, la répartition des observations selon les serveurs peut-être biaisée. La contrainte est celle d'*Hadoop*, pas de communication entre les serveurs, et il faut minimiser le temps donc se limiter à une seule lecture de l'ensemble de la base.

L'algorithme dit de *reservoir sampling* (Vitter ; 1985) tient ces objectifs si l'échantillon retenu tient en mémoire dans un seul nœud. Malheureusement cette méthode est par principe séquentielle et les aménagements pour la paralléliser avec *MapReduce* peut poser des problèmes de représentativité de l'échantillon, notamment si les capacités de stockage des nœuds sont déséquilibrées.

1. $\mathbf{s}$ la ligne courante de la base qui est lue ligne à ligne,
2. $\mathbf{R}$ est la matrice réservoir de n lignes,
3. Les n premières lignes ou observations sont retenues :
4. `Pour` $i = 1, n$ $\mathbf{R}_i = \mathbf{s}(i)$
5. $i = n + 1$
6. `Tant que i lreste des observations` $\mathbf{s}$ `Faire`,
   - Tirer un nombre entier aléatoire j entre 1 et i inclus,
   - `Si` $j \leq k$ `alors` $\mathbf{R}_j = \mathbf{s}(i)$
   - `Fin Tant que`

La phase Map est triviale tandis que celle Reduce, tire un entier aléatoire et assure le remplacement conditionnel dans la partie "valeur" de la sortie. Les clefs sont sans importance. Vitter (1985)[13] montre que cet algorithme conçu pour échantillonner sur une bande magnétique assure le tirage avec équiprobabilité. Des variantes : pondérées, équilibrées, stratifiées, existent mais celle échelonnable pour *Hadoop* pose de problèmes si les donnéees sont mal réparties sur les nœuds.

Par tri de l'échantillon

Le principe de l'algorithme est très élémentaire ; N nombres aléatoires uniformes sur $[0, 1]$ sont tirés et associés : une clef à chaque observation ou valeur. Les paires (clef, observation) sont triées selon cette clef et les n premières ou n dernières sont retenues.

L'étape *Map* est facilement définie de même que l'étape *Reduce* de sélection mais l'étape intermédiaire de tri peut être très lourde car elle porte sur toute la base.

ScanSRS

Xiangrui (2013)[7] propose de mixer les deux principes en sélectionnant un sous-ensemble des observations les plus raisonnablement probables pour réduire le volume de tri. L'inégalité de Bernstein est utilisée pour retenir ou non des observations dans un réservoir de taille en $0(n)$ car nécessairement plus grand que dans l'exemple précédent ; observations qui sont ensuite triées sur leur clef avant de ne conserver que les n plus petites clefs. Un paramètre règle le risque de ne pas obtenir au moins n observations tout en contrôlant la taille du réservoir.

3.5 Factorisation d'une matrice

Tous les domaines de datamasse produisent, entre autres, de très grandes matrices creuses : clients réalisant des achats, notant des films, relevés de capteurs sur des avions, voitures, textes et occurrences de mots... objets avec des capteurs, des gènes avec des expressions. Alors que ces méthodes sont beaucoup utilisés en e-commerce, les acteurs se montrent discrets sur leurs usages et les codes exécutés, de même que la fondation Apache sur la NMF. Heureusement, la recherche en Biologie sur fonds publics permet d'y remédier.

SVD

Les algorithmes de décomposition en valeurs singulières (SVD) d'une grande matrice creuse, sont connus de longue date mais exécutés sur des machines massivement parallèles de calcul intensif pour la résolution de grands systèmes linéaires en analyse numérique. Son adaptation au cadre *MapReduce* et son application à l'analyse en composantes principales ou l'analyse des correspondances, ne peut se faire de la même façon et conduit à d'autres algorithmes.

Mahout propose deux implémentations *MapReduce* de la SVD et une autre version basée sur un algorithme stochastique est en préparation. En juin 2014, *RHadoop* ne propose pas encore d'algorithme, ni *Spark*.

NMF

La factorisation de matrices non négatives (Paatero et Tapper ; 1994[10], Lee et Seung ; 1999[5]) est plus récente. Sûrement très utilisée par les entreprises commerciales, elle n'est cependant pas implémentée dans les librairies ouvertes (*RHadoop, Mahout*) basées sur *Hadoop*. Il existe une librairie R (`NMF`) qui exécute au choix plusieurs algorithmes selon deux critères possibles de la factorisation non négative et utilisant les fonctionnalités de parallélisation d'une machine (multi cœur) grâce à la librairie `parallel` de R. Elle n'est pas interfacée avec *Hadoop* alors que celle plus rudimentaire développée en java par Ruiqi et al. (2014)[6] est associée à *Hadoop*.

3.6 k-means

L'algorithme k-means et d'autres de classification non supervisée sont adaptés à *Hadoop* dans la plupart des librairies. Le principe des algorithmes utilisés consiste à itérer un nombre de fois fixé *a priori* ou jusqu'à convergence les étapes *Map-Reduce*. Le principal problème lors de l'exécution est que chaque itération

provoque une réécriture des données pour ensuite les relire pour l'itération suivante. C'est une contrainte imposée par les fonctionnalités *MapReduce* de *Hadoop* (à l'exception de *Spark*) pour toute exécution d'un algorithme itératif qui évidemment pénalise fortement le temps de calcul.

Bien entendu, le choix de la fonction qui calcule la distance entre une matrice de k centres et une matrice d'individus est fondamental.

1. L'étape Map utilise cette fonction pour calculer un paquet de distances et retourner le centre le plus proche de chaque individu. Les individus sont bien stockés dans la base HDFS alors que les centres, initialisés aléatoirement restent en mémoire.
2. Pour chaque clef désignant un groupe, l'étape Reduce calcule les nouveaux barycentres, moyennes des colonnes des individus partageant la même classe / clef.
3. Une programme global exécute une boucle qui itère les étapes *MapReduce* en lisant / écrivant (sauf pour *Spark*) à chaque itération les individus dans la base et remettant à jour les centres.

3.7 Régression linéaire

Est-il pertinent ou réellement utile, en terme de qualité de prévision, d'estimer un modèle de régression sur un échantillon de très grande taille alors que les principales difficultés sont généralement soulevées par les questions de sélection de variables, sélection de modèle. De plus, les fonctionnalités offertes sont très limitées, sans aucune aide au diagnostic. Néanmoins, la façon d'estimer un modèle de régression illustre une autre façon d'utiliser les fonctionnalités *MapReduce* pour calculer notamment des produits matriciels.

Soit $\mathbf{X}$ la matrice $(n \times p)$ (*design matrix*) contenant en première colonne des 1 et les observations des p variables explicatives quantitatives sur les n observations ; $\mathbf{y}$ le vecteur des observations de la variable à expliquer.

On considère n trop grand pour que la matrice $\mathbf{X}$ soit chargée en mémoire mais p pas trop grand pour que le produit $\mathbf{X}'\mathbf{X}$ le soit.

Pour estimer le modèle

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

l'algorithme combine deux étapes de *MapReduce* afin de calculer chaque produit matriciel $\mathbf{X}'\mathbf{X}$ et $\mathbf{X}'\mathbf{y}$. Ensuite, un appel à la fonction `solve` de R résout les équations normales

$$\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{y}$$

et fournit les estimations des paramètres β_j.

Le produit matriciel est décomposé en la somme (étape Reduce) de la liste des matrices issues des produits (étape Map) $\mathbf{X}_i'\mathbf{X}_j (i, j = 1, n)$.

3.8 Régression logistique

Les mêmes réserves que ci-dessus sont faites pour la régression logistique qui est obtenue par un algorithme de descente du gradient arrêté après un nombre fixe d'itérations et dont chacune est une étape *MapReduce*. L'étape *Map* calcule la contribution de chaque individu au gradient puis l'écrit dans la base, l'étape *Reduce* est une somme pondérée. Chaque itération provoque donc n lectures et n écritures dans la base.

Des améliorations pourraient être apportées : adopter un critère de convergence pour arrêter l'algorithme plutôt que fixer le nombre d'itérations, utilisé un gradient conjugué. La faiblesse de l'algorithme reste le nombre d'entrées/sorties à chaque itération.

3.9 Random forest

Random Forest, cas particulier de bagging, s'adapte particulièrement bien à l'environnement d'*Hadoop* lorsqu'une grande taille de l'ensemble d'apprentissage fournit des échantillons indépendants pour estimer et agréger (moyenner) des ensembles de modèles. Comme en plus, *random forest* semble insensible au sur-apprentissage, cette méthode ne nécessite généralement pas de gros efforts d'optimisation de paramètres, elle évite donc l'un des principaux écueils des approches Big Data en apprentissage.

Principe

Soit n la taille totale (grande) de l'échantillon d'apprentissage, m le nombre d'arbres ou modèles de la forêt et k la taille de chaque échantillon sur lequel est estimé un modèle. Les paramètres m et k, ainsi que `mtry` sont en principe à optimiser mais si n est grand, on peut espérer qu'ils auront peu d'influence. Il y a en principe trois cas à considérer en fonction de la taille de n.

- $k \times m < n$ toutes les données ne sont pas utilisées et des échantillons indépendants sont obtenus par tirage aléatoire sans remise ,
- $k \times m = n$ des échantillons indépendants sont exactement obtenus par tirages sans remises,
- $k \times m > n$ correspond à une situation où un ré-échantillonnage avec remise est nécessaire ; $k = n$ correspond à la situation classique du bootstrap.

L'astuce de Uri Laserson est de traiter dans un même cadre les trois situations en considérant des tirages selon des lois de Poisson pour approcher celui multinomial correspondant au bootstrap. L'idée consiste donc à tirer indépendamment pour chaque observation selon une loi de Poisson.

Échantillonnage

L'objectif est d'éviter plusieurs lectures, m, de l'ensemble des données pour réaliser m tirages avec remise et également le tirage selon une multinomiale qui nécessite des échanges de messages, impossible dans *Hadoop*, entre serveurs.

Le principe de l'approximation est le suivant : pour chaque observation $x_i (i = 1, n)$ tirer m fois selon une loi de Poisson de paramètre (k/n), une valeur $p_{i,j}$pour chaque modèle $M_j (j = 1, m)$. L'étape Map est ainsi constituée : elle émet un total de $p_{i,j}$ fois la paire (clef, valeur) (j, x_i) ; $p_{i,j}$ pouvant être 0. Même si les observations ne sont pas aléatoirement réparties selon les serveurs, ce tirage garantit l'obtention de m échantillons aléatoires dont la taille est approximativement k. D'autre part, approximativement $\exp(-km/n)$ des données initiales ne seront présentes dans aucun des échantillons.

L'étape de tri ou *shuffle* redistribue les échantillons identifiés par leur clef à chaque serveur qui estime un arbre. Ces arbres sont stockés pour constituer la forêt nécessaire à la prévision.

Cette approche, citée par Prajapati (2010)[11], est testée sur les enchères de bulldozer de Kaggle par Uri Laserson qui en développe le code.

Conclusion

Quelques remarques pour conclure ce tour d'horizon des outils permettant d'exécuter des méthodes d'apprentissage supervisé ou non sur des données volumineuses.

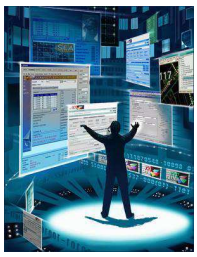
- Cet aperçu est un instantané à une date donnée, qui va évoluer rapidement en fonction des développements rapides du secteur et donc des librairies qui ne manqueront pas d'évoluer. C'est l'avantage d'un support de cours électronique car adaptable en temps réel.
- L'apport pédagogique important et difficile concerne la bonne connaissance des méthodes de modélisation, de leurs propriétés, de leurs limites. Pour cet objectif, R reste un outil à privilégier. Largement interfacé avec tous les systèmes de gestion de données massives ou non, il est également utilisé en situation "industrielle".
- Une remarque fondamentale concerne le réglage ou l'optimisation des paramètres des méthodes d'apprentissage : nombre de classes, sélection de variables, paramètre de complexité. Ces problèmes, qui sont de réelles difficultés pour les méthodes en question et qui occupent beaucoup les équipes travaillant en apprentissage (machine et/ou statistique), sont largement passés sous silence. Que deviennent les procédures de validation croisée, *bootstrap*, les échantillons de validation et test, bref toute la réflexion sur l'opti-

misation des modèles pour s'assurer de leur précision et donc de leur validité ?
- Si R, éventuellement précédé d'une phase d'échantillonnage (C, java, python, perl) n'est plus adapté au volume et
- Si *RHadoop* s'avère trop lent,
- *Mahout* et *Spark* sont à tester car semble-t-il nettement plus efficace. Un apprentissage élémentaire de java, s'avère donc fort utile pour répondre à cet objectif.
- Si ce n'est toujours pas suffisant, notamment pour prendre en compte les composantes Variété et Vélocité du *Big Data*, les compétences acquises permettent au moins la réalisation de prototypes avant de réaliser ou faire réaliser des développements logiciels plus conséquents. Néanmoins le passage à l'échelle est plus probant si les prototypes anticipent la parallélisation en utilisant les bons langages comme scala ou clojure.

Petit rappel pour conclure. La très volumineuse littérature notamment "grand public" sur le thème *Big Data* traduit une certaine confusion dans les objectifs poursuivis, par exemple entre les ceux ci-dessous.

- Le premier est d'ordre "clinique" ou d"enquête de police". Si une information est présente dans les données, sur le web, les algorithmes d'exploration exhaustifs vont la trouver ; c'est la NSA dans la *La Menace Fantôme*.
- Le deuxième est "prédictif". Apprendre des comportements, des occurrences d'évènements pour *prévoir* ceux à venir ; c'est GAFA dans *L'Empire Contre Attaque*. Il faut ensuite distinguer entre :
  - prévoir un comportement moyen, une tendance moyenne, comme celle souvent citée de l'évolution d'une épidémie de grippe à partir des effectifs des mots clefs associés et
  - prévoir un comportement individuel ou si M. Martin habitant telle adresse va ou non attraper la grippe !

Il est naïf de penser que l'efficacité obtenue pour l'atteinte du premier objectif peut se reporter sur le deuxième pour la seule raison que "toutes" les données sont traitées. La fouille de données (*data mining*), et depuis plus longtemps la Statistique, poursuivent l'objectif de prévision depuis de nombreuses années avec des taux de réussite, ou d'erreur, qui atteignent nécessairement des limites que ce soit avec 1000, 10 000 ou 10 millions... d'observations. L'estimation d'une moyenne, d'une variance, du taux d'erreur, s'affinent et convergent vers la "vraie" valeur (loi des grands nombres), mais l'erreur d'une *prévision individuelle*, et son taux reste inhérents à une variabilité intrinsèque et irréductible, comme celle du Vivant.



En résumé : *Back to the Futur* et *Star War*, mais pas *Minority Report*.

En revanche et de façon positive, l'accumulation de données d'origines différentes, sous contrôle de la CNIL et sous réserve d'une anonymisation rigoureuse empêchant une ré-identification (Bras ; 2013)[3] permet d'étendre le deuxième objectif de prévision à bien d'autres champs d'application que ceux commerciaux ; par exemple en Santé Publique, si le choix politique en est fait.

Références

[1] Joseph Adler, *R in a nutshell*, O'Reilly, 2010.

[2] Philippe Besse, Aurélien Garivier et Jean Michel Loubes, *Big Data Analytics - Retour vers le Futur 3 ; De Statisticien à Data Scientist*, http://hal.archives-ouvertes.fr/hal-00959267, 2014.

[3] Pierre Louis Bras, *Rapport sur la gouvernance et l'utilisation des données de santé*, http://www.social-sante.gouv.fr/IMG/pdf/Rapport_donnees_de_sante_2013.pdf, 2013.

[4] Jeffrey Dean et Sanjay Ghemawat, *MapReduce : Simplified Data Processing on Large Clusters*, Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6, OSDI'04, 2004.

[5] D. Lee et S. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature (1999).

[6] Ruiqi Liao, Yifan Zhang, Jihong Guan et Shuigeng Zhou, *CloudNMF : A MapReduce Implementation of Nonnegative Matrix Factorization for Large-scale Biological Datasets*, Genomics, Proteomics & Bioinformatics **12** (2014), n° 1, 48 – 51.

[7] Xiangrui M., *Scalable Simple Random Sampling and Statified Sampling*, Proceedings of the 30th International Conference on Machine Learning, 2013.

[8] E. McCallum et S. Weston, *Parallel R*, O'Reilly Media, 2011.

[9] Sean Owen, Robin Anil, Ted Dunning et Ellen Friedman, *Mahout in Action*, Manning Publications Co., 2011.

[10] Pentti Paatero et Unto Tapper, *Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics **5** (1994), n° 2, 111–126.

[11] Vignesh Prajapati, *Big Data Analytics with R and Hadoop*, Packt Publishing, 2013.

[12] R Core Team, *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2014, http://www.R-project.org/.

[13] Vitter J. S., *Random sampling with a reservoir*, ACM transaction on Mathematical Software **11 :1** (1985), 37–57.

[14] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauly, Michael J. Franklin, Scott Shenker et Ion Stoica, *Resilient Distributed Datasets : A Fault-Tolerant Abstraction for In-Memory Cluster Computing*, Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12) (San Jose, CA), USENIX, 2012, p. 15–28, ISBN 978-931971-92-8, https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/zaharia.