BIG Data et R: opportunités et perspectives

Guati Rizlane [1] & Hicham Hajji [2]

[1] *Ecole Nationale de Commerce et de Gestion de Casablanca, Maroc, rguati@gmail.com*
[2] *Ecole des Sciences Géomatiques, IAV Rabat, Maroc, h.hajji@iav.ac.ma*

Résumé. Récemment, de nouvelles techniques ont été proposées pour améliorer le stockage et le traitement de données massives, en s'appuyant sur le projet Hadoop et sur la distribution des calculs et de stockage sur un cluster de serveurs. Bien que ces solutions ont permis de résoudre un certain nombre de problèmes, ce n'est que le couplage récent entre Hadoop avec le logiciel R qui a permis de donner un nouvel élan pour une meilleure exploitation des données massives. L'article explique l'architecture de ce couplage, et montre à travers quelques exemples comment les commandes sont exécutées à l'intérieur de Hadoop.

Mots-clés. BIG Data, Logiciel R, Hadoop, Statistiques, Modélisation

Abstract. Recently, new techniques have been proposed to improve the storage and processing of BIG Data, based on the Hadoop project and the distribution of computation and storage on a servers cluster. Although these solutions have solved a number of problems, it is only the recent coupling between Hadoop with R software that has given new impetus to a better exploration of massive data. The paper describes the architecture of this coupling, and shows some examples of how the commands are executed within Hadoop.

KeyWords BIG Data, R software, Hadoop, Statistics, Modelling

1 Introduction

1.1 BIG Data

L'importance prise par le numérique dans le quotidien, a permis de générer des volumes d'informations sans cesse plus importants. Selon IBM (Jacobs et Dinsmore(2013)), chaque heure sont générés 2,5 trillions d'octets de données, soit 2,5 Po. Il est prévu également que pendant l'année 2020 seront générés 35 Zettaoctets, alors que seulement 1 Zettaoctets de données numériques ont été générés par l'humanité entre le début de l'informatique (1940) et 2010.

Formellement, une définition du BIG Data a été proposée pour mieux comprendre ce phénomène de données massives (Davenport et Barth (2012)). Elle est basée sur les trois dimensions: Volume, Vélocité et Variété (ou **3V**).

- Volume : Le volume des données stockées aujourd'hui est en pleine expansion. Ainsi, Twitter génère quotidiennement à l'heure actuelle 7 teraoctets de données et Facebook 10 teraoctets.

- Vélocité: Elle représente à la fois la fréquence à laquelle les données sont générées, capturées et partagées. Par exemple, chaque seconde sont émis plus de 500 tweets, ce qui fait 50 millions chaque jour.

- Variété : En plus des données du types bases de données relationnelles, le Big Data s'intéresse également à un nombre important de formats et variétés de données: image, texte, réseaux sociaux, capteurs.

Les techniques Big Data doivent répondre à un certain nombre de problématiques, comme l'extraction de sens à partir de ces données, et permettre la possibilité de passer d'une analyse reporting (du passé), vers une analyse prédictive (futur).

1.1.1 Hadoop comme framework pour le BIG Data

Etant donné que les bases de données traditionnelles ne sont pas capable de gérer correctement ce volume important de données, de nouvelles techniques ont été proposées. La plus connue est celle autour du système Hadoop (Anand (2008)). Hadoop est un environnement d'exécution distribué, performant et scalable, dont la vocation est de traiter des volumes de données considérables. Il repose sur un principe simple: distribuer les données sur un cluster de serveurs (nœuds de données), et migrer les traitements vers ces nœuds. Cette approche assure une mise à l'échelle des algorithmes et des traitements à développer.

1.2 Les techniques statistiques et les données massives

Les techniques statistiques et économétriques classiques fonctionnent souvent bien avec des volumes de données moins importants, cependant dès que le volume de données devient important, il y a un certain nombre de problèmes qui apparaissent (Fan et Han (2013)); d'abord le volume des données peut exiger des outils de manipulation de données plus sophistiqués. Ensuite, il peut y avoir des estimateurs potentiels plus appropriés pour l'estimation, d'où le besoin d'effectuer une sélection de variables. Également, les données massives peuvent permettre la découverte de relations plus flexibles que de simples modèles linéaires. Des techniques d'apprentissage, comme les arbres de décision, SVM, les réseaux neuronaux peuvent permettre des modélisations plus efficaces des relations complexes. Pour un aperçu plus complet des contraintes et défis statistiques liés au BIG Data, voir Fan et Han(2013).

2 Outils R pour l'analyse des données massives

Comme expliqué auparavant, Hadoop représente actuellement le système de référence pour le stockage et le traitement des données massives. En donnant la possibilité aux concepteurs de distribuer leurs algorithmes sur des noeuds de données, Hadoop offre une scalabilité sans précédant aux algorithmes qui demandaient des ressources de plus en plus importantes pour l'exécution des applications sur un large volume de données. Néanmoins les développements sur Hadoop nécessitent la réécriture des algorithmes en des modules MapReduce[1] pour pouvoir les partager sur les noeuds de données. Le même problème se pose pour exécuter les algorithmes statistiques sur de grands volumes de données. Il est nécessaire de réécrire les algorithmes statistiques selon le modèle MapReduce pour pouvoir les exécuter avec Hadoop. Récemment, **Revolution Analytics**[2] a proposé, une extension de Hadoop et R pour permettre aux utilisateurs d'utiliser R sur des grands volumes de données, sans être obligé de les réécrire en format MapReduce. Une approche qui permet de cacher la complexité de Hadoop et MapReduce aux utilisateurs de R, et leur permette de se concentrer sur leurs algorithmes et méthodes statistiques.

2.1 Composantes clés de l'architecture R avec Hadoop

L'architecture proposée repose sur l'abstraction des difficultés liées aux BIG Data et sur l'utilisation de modèles de distributions des algorithmes comme MapReduce. L'architecture ainsi proposée (fig. 2) est composée principalement de trois éléments:

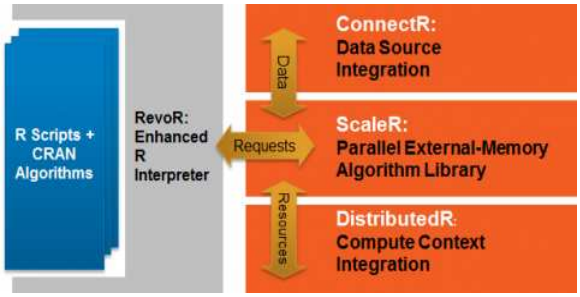

Figure 1: Composantes clés de l'architecture R avec Hadoop

DistributedR C'est un Framework d'exécution parallèle adaptable qui fournit des services incluant les communications, le stockage, la gestion de la mémoire pour permettre aux algorithmes statistiques de ScaleR d'analyser rapidement et d'être scalable.

ScaleR C'est ce module qui fournit les algorithmes statistiques optimisés pour une exécution parallèle sur les données massives.

[1] le langage de développement de Hadoop

[2] http://Fournisseur commercial de logiciels et services pour le projet open source R www.revolutionanalytics.com/

ConnectR Permet d'établir des connections à plusieurs sources de données, allant d'une base de données aux clusters d'ordinateurs comme dans le cas de Hadoop.

2.2 Exemples d'algorithmes R dans Hadoop

Cet article illustre l'utilisation de R dans Hadoop à travers un large ensemble de données de presque 120 millions de lignes au total(de 120 gigaoctets)[3]. La série de données représente des données des lignes aériennes et se composent des détails d'arrivée et des départ de tous les vols commerciaux au sein des Etats-Unis, du 10/1987 à 04/2008.

2.2.1 Fonctionnement des commandes R dans Hadoop

Les données sont stockées dans le système Hadoop[4], et distribuées à travers les nœuds de données. L'utilisateur en exécutant une commande R, déclenche une distribution de la commande sur l'ensemble des nœuds en utilisant le langage MapReduce. Une fois que la commande est exécutée sur l'ensemble des nœuds (fig.2), l'extension Hadoop se charge pour agréger les réponses et les présenter à l'utilisateur de R. Le grand avantage de ce Framework c'est qu'il permet à l'utilisateur d'exécuter ses commandes R sans se soucier de la manière avec laquelle elles seront distribuées à travers l'ensemble des données.

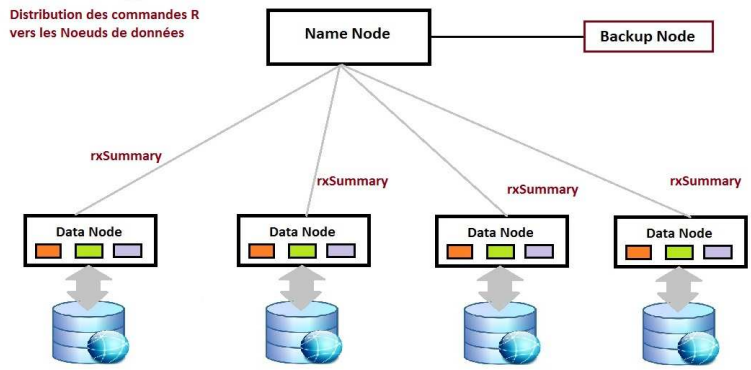

Figure 2: Distribution de la commande rxSummary à travers les nœuds de données

2.2.2 Description des données avec R

Pour afficher le résumé des neuf variables qui constituent les données exemples, la commande *rxSummary* peut être utilisée, avec MyData est la référence aux données distribuées à travers les nœuds de données de Hadoop:

[3]Fichier exemple du site de Revolution Analytics www.revolutionanalytics.com
[4]Principalement dans le HDFS Hadoop Distributed File System

```
rxSummary(~., myData)
```

```
Name Mean StdDev Min Max ValidObs MissingObs
Year 2000.6216670 7.2197464 1987 2012 10399755 0
CRSDepTime 13.5039649 4.7525385 0 24 10399755 0
DepDelay 8.1574627 29.2068325 -1199 2467 10211531 188224
TaxiOut 15.5218607 10.6178418 0 1439 7632487 2767268
TaxiIn 6.4172542 8.2726772 0 1439 7618449 2781306
ArrDelay 6.5668612 31.6039828 -1233 2453 10186272 213483
ArrDel15 0.2000864 0.4000648 0 1 10186272 213483
...
Number of valid observations: 10399755
Number of missing observations: 0
DayOfWeek Counts
Mon 1531219
Tues 1518605
```

L'exécution de la commande *rxSummary* permet d'afficher un résumé sur les différentes variables rencontrées dans les données. Ainsi le nombre d'observations valide est de *10399755*, et le retard à l'arrivée moyen sur toute la période est de presque 6,5 minutes.

2.2.3 Estimation d'un modèle linéaire d'un large ensemble de données

Pour créer un modèle linéaire, le même principe est à appliquer. L'utilisateur construit à travers sa commande les différentes variables qu'il compte utiliser pour composer son modèle, et laisse au Framework le soin de distribuer la commande sur les noeuds de données, et l'agrégation des résultats pour les présenter à l'utilisateur.

```
formula <- ArrDel15 ~ DayOfWeek + UniqueCarrier + Origin + Dest + CRSDepTime + DepDelay
+ TaxiOut + CRSElapsedTime + Distance
model <- rxLinMod(formula, trainingData)
summary(model)
```

```
Estimate Std. Error t value Pr(>|t|)
DayOfWeek=Mon 9.093e-03 1.685e-03 5.396 6.84e-08 ***
DayOfWeek=Tues 4.721e-03 1.696e-03 2.784 0.005373 **
...
DayOfWeek=Sat -1.872e-03 1.763e-03 -1.062 0.288298
DayOfWeek=Sun Dropped Dropped Dropped Dropped
CRSDepTime 5.388e-03 1.001e-04 53.855 2.22e-16 ***
DepDelay 6.953e-03 1.413e-05 492.024 2.22e-16 ***
TaxiOut 1.175e-02 5.671e-05 207.276 2.22e-16 ***
CRSElapsedTime -3.392e-03 6.918e-05 -49.032 2.22e-16 ***
...
Distance 4.080e-04 8.370e-48.739 2.22e----
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

L'exécution du modèle montre que les variables sont significatives.

3 Conclusion et discussions

Les techniques Big Data apportent de nouvelles opportunités intéressantes pour les organisations, et de nouveaux challenges aux statisticiens et gestionnaires de données. Bien que le framework présenté dans cet article essaye de montrer une approche qui répond au problème de mise à l'échelle des algorithmes statistiques. Il reste néanmoins un certain nombre de contraintes et d'améliorations à développer.

D'un coté, la taille et le volume ainsi que la grande dimensionnalité de données introduisent des défis statistiques qu'il reste à résoudre comme l'accumulation du bruit, les fausses corrélations, l'hétérogénéité et les erreurs de mesures. D'autre coté, parallèlement à Hadoop, s'est développé un autre framework largement plus performant: le projet Spark (Matei et al (2012)). Bien que l'idée de distribuer les calculs sur un ensemble de nœuds de données, a été maintenu comme Hadoop, le framework Spark réalise les calculs distribués directement en mémoire, ce qui accélère nettement les calculs itératives qui sont souvent rencontrés dans les algorithmes statistiques. [5]

Bibliographie

[1] Matei, Z. Mosharaf, C. Tathagata, D. Ankur, D. Justin, M. Murphy M. Franklin, M. Shenker, S. Stoica, I. (2012), Fast and Interactive Analytics over Hadoop Data with Spark, Number 4 Volume 37, USENIX.
[2] Anand, A. (2008), Scaling Hadoop to 4000 nodes at Yahoo! http://goo.gl/8dRMq
[3] Davenport et Barth, A. (2012), How "Big Data" Is Different, In MIT Sloan Management Review
[4] Jacobs, B. Dinsmore, T.(2013), Delivering Value from Big Data with Revolution R Enterprise and Hadoop, http://www.revolutionanalytics.com/sites/default/files/driving-value-from-big-data-rre-hadoop.pdf
[5] Fan, J. et Han, F.(2013), Challenges of Big Data Analysis, http://arxiv.org/pdf/1308.1479.pdf

[5] Le blog de *revolution analytics* prévoit une version augmentée de spark avec R au courant de l'année 2014 http://blog.revolutionanalytics.com/2013/12/apache-spark.html