Big Data et Statistique publique

Lettre d'information
N°3 juin 2016

Travailler avec des Big Data, qu'est-ce que ça change ?

Le Big Data recouvre un champ technique vaste qui évolue rapidement. On peut cependant dégager des axes structurants qui permettront d'orienter les investissements à venir et de tirer pleinement profit de cette révolution technologique et méthodologique. Au sein de l'écosystème Hadoop, qui apparaît aujourd'hui comme incontournable s'inscrivent des outils qui sont des références dans le monde informatique et qui pour certains sont déjà bien connus à l'Insee, en particulier Java, SQL, Python ou encore R. Alors que ces langages sont déjà utilisés à l'institut pour des usages « classiques », les acteurs du Big Data nous montrent comment une collaboration poussée entre les statisticiens et développeurs permettra à l'Insee de tirer parti de ces évolutions en innovant pour mieux répondre à ses utilisateurs. Le projet « données de caisse » (lettre n°2) est un exemple de cette collaboration. Ce numéro propose d'éclairer les aspects plus techniques du Big Data.
Benoît Rouppert, chef du Département de la production et de l'infrastructure informatiques

De nouveaux outils - infrastructures et logiciels- pour les données massives

Hadoop, Spark, MapReduce, de quoi parle-t-on ?

Il existe aujourd'hui un grand nombre de solutions pour le traitement des très grands volumes de données. Les performances de systèmes comme Hadoop ou MapReduce voire Spark font actuellement consensus. En grande majorité open-source, ces logiciels sont parfois mieux connus sous le nom des entreprises comme Cloudera ou Hortonworks qui diffusent ces logiciels et proposent des services payants d'assistance ou de formation.

Le principe central consiste à utiliser la puissance et les capacités de plusieurs ordinateurs connectés entre eux (regroupés en ***cluster***), et répartir (distribuer) sur ces machines à la fois des données et des traitements. Le **stockage** des données s'appuie le plus souvent sur un système de fichiers en réseau : en pratique, les fichiers sont découpés en blocs de taille constante, répartis et dupliqués de manière homogène sur l'ensemble des machines dédiées au stockage. La **répartition** permettra de distribuer les traitements sur les machines stockant les données, suivant l'idée que « déplacer les traitements est moins cher que déplacer les données ». Les données seront par ailleurs **répliquées pour offrir une sécurité** en cas de panne d'une ou plusieurs machines (d'autant plus probable que le nombre de machines est grand). La solution la plus utilisée pour le stockage des fichiers est HDFS, (Hadoop Distributed Files System), qui est une composante de la suite logicielle Hadoop.

Pour le **traitement,** le modèle de programmation le plus utilisé est **MapReduce (composant de la même suite logicielle Hadoop)**. Comme son nom l'indique, le principe de MapReduce comprend deux fonctions principales, *map* et *reduce.* La première consiste à décomposer le traitement à effectuer en traitements élémentaires qui seront appliqués en parallèle à tous les paquets de données, tandis que la seconde permet d'agréger les résultats partiels obtenus sur ces derniers. En pratique, il n'est pas forcément nécessaire de programmer ces différentes étapes, plusieurs langages permettent de soumettre des instructions qui seront retranscrites en MapReduce de façon invisible pour l'utilisateur. Par exemple, **Hive** est un outil qui permet d'écrire des programmes en SQL, comme on le ferait pour manipuler les données d'une base relationnelle classique. Pour des traitements statistiques plus avancés, des librairies de fonctions se développent et il est vraisemblable qu'à terme la plupart soient également transparentes pour le statisticien. En attendant, il est parfois nécessaire de procéder soi-même à l'implémentation des étapes *map* et *reduce* (par exemple en R avec Rhadoop, ensemble de librairies R permettant la parallélisation des traitements statistiques sous Hadoop), ce qui peut s'avérer complexe.

La programmation MapReduce peut se révéler très efficace pour de nombreux traitements sur de très grands volumes de données. Cependant, le principe de la répartition des calculs constitue un coût fixe du calcul qui a un intérêt sur des données massives, mais qui rend **inadaptée cette solution pour de plus petits volumes**. De plus, la performance de ce modèle de programmation dépend des choix faits en termes d'architecture. Ainsi, après chaque opération Map ou Reduce, Hadoop MapReduce écrit les résultats intermédiaires sur disque, ces résultats étant ensuite lus si nécessaire, ce qui entraîne d'importants temps d'exécution, en particulier pour des algorithmes itératifs.

Ces limites expliquent le succès d'une **solution alternative, Spark**, devenue la référence pour le traitement de données massives. Spark conserve un modèle d'exécution distribué comme Hadoop, mais présente de nombreux avantages : en particulier sa capacité à opérer bien plus souvent en mémoire vive lui permet de gagner en rapidité d'exécution, lui conférant des performances très nettement supérieures (traitements entre 10 à 100 fois plus rapides). En conséquence, il sera plus compétitif pour traiter des données en temps réel (traitement des flux de données au fur et à mesure de leur acquisition), ou pour des traitements complexes (d'importantes bibliothèques d'algorithmes de calcul se développent pour Spark). À noter que Spark offre des API (interface de programmation applicative) dans plusieurs langages, dont Python et R (à l'instar de Rhadoop pour Hadoop). En pratique, cela signifie qu'il est probable qu'à terme, un statisticien pourra écrire sans limitations en R un programme qui sera exécuté par Spark.

Pour en savoir plus :

http://eric.univ-lyon2.fr/~ricco/cours/slides/programmation%20mapreduce%20sous%20r.pdf

http://statoscope.wordpress.com/2016/06/08/clusters-big-data-vs-clusters-de-calcul-quelle-difference/

Une expérimentation sur la plateforme du CASD Teralab

Les différents outils standards pour le Big Data sont-ils aujourd'hui plus rapides, aussi fiables et faciles d'accès que les logiciels classiques pour les traitements économétriques ? Une première expérimentation a été menée en mai-août 2015 sur la plateforme Big Data Teralab du CASD, dédiée à la recherche, l'innovation et

l'enseignement et développée dans le cadre du Programme d'Investissements d'Avenir en partenariat avec l'Institut Mines-Télécom et l'Insee. Elle a porté sur l'analyse des observations quotidiennes de prix à la pompe pour environ dix mille stations services et 6 types de carburants depuis 2007. Cette base de données correspond à des millions d'enregistrements. D'une taille comparable à celle de nombreuses bases utilisées par exemple à l'Insee, elle peut être traitée sans trop de difficultés par des logiciels classiques tels que R, mais elle est aussi suffisamment importante pour justifier l'usage de techniques Big Data. L'expérimentation a montré que les outils ne sont pas encore d'un usage courant et qu'ils nécessitent pour l'instant un haut niveau d'expertise et donc un important coût d'entrée.

Cette étude a été l'occasion de s'approprier les principaux logiciels Big Data, d'en tester les performances et la « maniabilité ». L'ensemble des traitements a été effectué dans l'environnement Hadoop. Ainsi, **Hive** a été utilisé pour calculer des statistiques simples (avec des requêtes de type SQL), mais a nécessité une mise en forme des données au préalable. Cette phase de formatage fastidieuse des données brutes (format XML) a été effectuée à partir de l'outil **Pig**. Les traitements statistiques plus complexes (régressions linéaires et logistiques) ont été programmés via **RHadoop** et **Spark**. Ces **techniques évoluant très rapidement**, les constats tirés de cette utilisation peuvent être rapidement caducs. À ce jour, on peut en tirer deux conclusions principales.

En premier lieu, les solutions utilisées, bien qu'elles se démocratisent de plus en plus et existent pour certaines depuis presque dix ans, sont encore loin d'être actuellement simples et transparentes pour un statisticien. Un grand volume de données nécessite de repenser les traitements statistiques les plus simples, que ce soit le tracé d'un histogramme, le calcul d'un percentile ou d'une médiane. **RHadoop**, par exemple, impose la programmation manuelle des fonctions ***map*** et ***reduce*** même pour une régression linéaire ou logistique. Plus maniable, **Spark** dispose d'importantes librairies, mais elles sont spécialisées pour l'essentiel dans les méthodes d'apprentissage automatique (ou **machine learning,** ensemble de méthodes développées pour répondre par exemple à des problèmes de modélisation prédictive) et ne fournissent pas toujours, pour l'instant, les modèles économétriques classiquement utilisés dans la statistique publique.

Le second enseignement est que la parallélisation des traitements et donc leur rapidité peut se faire au prix d'une moindre précision des résultats (calculs approchés seulement). Il faut donc comprendre les limites des algorithmes. En outre, en développement actif, la pérennité et la qualité de ces algorithmes doivent être contrôlées.

Pour en savoir plus : rapport de stage de Stéphanie Himpens

Veille et formations : de nouvelles pratiques d'information et de partage des connaissances

Les nouvelles sources de données, la démocratisation des outils et techniques d'analyse de données avancées constituent des opportunités pour les statisticiens, mais il peut être difficile de s'y retrouver. Voici quelques repères pour s'orienter : les technologies évoluant vite, il est indispensable de procéder à une veille méthodologique continue. Cette veille peut en particulier se faire à partir de réseaux d'échanges, par exemple pour disposer de retour d'expériences, de conseils sur les outils. Ces réseaux ou forums sont un premier point d'entrée. En interne au SSP, un groupe Big Data sur le réseau social professionnel Yammer a été créé à cet effet pour une communauté d'utilisateurs potentiels (ouverte à toutes personnes intéressées). On peut aussi trouver des informations utiles sur certains blogs.

Pour développer des compétences plus pointues, par exemple dans le cadre d'un nouveau projet, une formation spécifique peut être nécessaire. Un grand nombre d'outils pédagogiques émergent aujourd'hui, au-delà des formations traditionnelles, permettant un apprentissage à la carte. On assiste notamment à l'explosion des MOOC (pour Massive Open Online Course, formation en ligne ouverte à tous généralement sous forme de vidéos avec des forums d'échange pour les étudiants et intervenants), des tutoriels avec captures de code... À titre d'illustration, il est possible d'apprendre un nouveau langage de programmation sur openclassroom, avec des simulateurs de code sur DataCamp ou de s'informer des derniers développements lors de conférences spécifiques.

La transversalité et le travail en équipe deviennent de plus en plus centraux dans un domaine mobilisant des compétences diverses mais pointues. Le recours plus systématique à des outils de travail collaboratif : gestion des versions de documents ou de codes partagés, via des outils internes (gforge) ou ouverts (github), peut permettre d'assurer un suivi et un partage de la compétence entre agents.

Des défis rapides pour innover

Pour stimuler le développement de solutions innovantes sur des problèmes concrets, il est devenu courant de recourir à des compétitions ouvertes. Cela peut prendre différentes formes ou dénominations. Les ***hackathons*** *ou* ***sprints****,* rassemblent des développeurs sur quelques jours pour programmer des utilisations innovantes des données. Etalab a déjà utilisé ce principe sur les données de la douane et sur les données de santé en s'adressant à des datajournalistes. Les compétitions plus larges ou ***challenges*** consistent à soumettre un sujet et des données et laisser les participants – non exclusivement des informaticiens – proposer des solutions innovantes à des problèmes ciblés. Le Genes a participé à la création d'un site de challenge datascience.net avec Bluestone. Eurostat a également lancé une « compétition » pour stimuler les propositions d'utilisation des Big Data pour la production d'indicateurs statistiques.

Pour en savoir plus :
https://www.etalab.gouv.fr/tag/hackathon
https://www.datascience.net/fr/home/
https://ec.europa.eu/eurostat/cros/content/big-data-official-statistics-competition-launched-please-register-10-january-2016_en

Agenda

Le lancement de l'ESSnet Big Data est officialisé. Il se déroulera de février 2016 à mai 2018. La première réunion de coordination a eu lieu à Tallinn du 13 au 15 juin.

Ont participé à ce numéro Stéphanie Himpens, Romain Tailhurat, Benoît Rouppert, Stéphanie Combes, Françoise Dupont et Pauline Givord. Cette lettre est une occasion d'informer largement et d'échanger. N'hésitez pas à transmettre vos réactions et suggestions d'articles à Stéphanie Combes : stephanie.combes@insee.fr
Les archives de cette lettre sont disponibles sur :
http://www.agora.insee.fr/jahia/Jahia/site/dmcsi/SiteDMCSI/DMSaccueil/*DMAEEaccueil*/BigData
Demande d'inscription individuelle à la lettre : dg75-l101@insee.fr