Un déluge de données

Serge Abiteboul et Pierre Senellart

Dans de nombreux domaines, scientifiques ou non, les données s'accumulent en masse. Les gérer et les exploiter est le défi posé à l'informatique du *Big Data*.

Au CERN, près de Genève, le collisionneur LHC *(Large Hadron Collider)* est équipé d'énormes détecteurs capables d'enregistrer les traces de dizaines à centaines de millions de collisions proton-proton par seconde. Évaluons grossièrement le volume de données que cela représente, en faisant des hypothèses basses. L'information relative aux produits de chaque collision est représentée à l'aide de quatre octets (soit $(2^8)^4 = 256^4 = 4{,}3$ milliards de possibilités), l'accélérateur fonctionne dix heures par jour en moyenne, et 100 millions de collisions sont enregistrées chaque seconde. On calcule facilement qu'au bout d'un an, le LHC produit ainsi 5×10^{15} octets, soit cinq pétaoctets d'information (une évaluation plus précise fournit des valeurs supérieures, *voir l'article de F. Malek dans ce numéro*). Il faut 5 000 disques durs du commerce pour stocker une telle masse de données.

Cependant, ce n'est pas le stockage qui est la difficulté principale, mais l'exploitation : extraire de l'information utile – de la connaissance – de cette masse de données dépasse largement les capacités humaines. Pour relever de tels défis, une nouvelle science est apparue, la science des données. Il s'agit d'extraire des contenus intéressants à partir de masses de données gigantesques, changeantes, hétérogènes, incertaines, et ce à l'aide d'algorithmes efficaces. L'objet de cet article est de décrire les principaux enjeux de cette science des données et les difficultés qu'elle doit surmonter.

L'exemple du LHC par lequel nous avons commencé est extrême en termes de volumes d'octets, mais de nombreux autres domaines scientifiques et industriels doivent faire face à une avalanche similaire de données. En recherche médicale, par exemple, l'un des problèmes a longtemps été l'absence de données en nombre suffisant ; aujourd'hui, c'est plutôt l'inverse : les chercheurs sont confrontés à la difficulté de produire des connaissances utiles à partir des myriades de données accumulées par les différents essais cliniques, les services hospitaliers, etc. Citons la base de données MIMIC-II *(Multiparameter Intelligent Monitoring in Intensive Care)*, constituée aux États-Unis par une équipe de chercheurs du

ANALYSE DES DONNÉES : LES TÂCHES CLASSIQUES

– Rechercher des données et les acquérir.

– Homogénéiser les données provenant de plusieurs sources.

– Détecter et éliminer les doublons et les erreurs.

– Interagir avec des humains pour obtenir plus de données, résoudre les contradictions et combler les manques (« crowdsourcing »).

– Faciliter le travail des analystes en leur fournissant des outils de visualisation

– Réaliser des analyses statistiques automatiques des données.

– Développer des applications et de nouveaux services.

© D'après Matvienko Vladimir/shutterstock.com

MIT, de *Philips* et du Centre médical *Beth Israel Deaconess*. Elle regroupe des données médicales portant sur 32 000 personnes hospitalisées en unités de soins intensifs, soit plus de 40 000 séjours . Le volume de données est nettement plus modeste que celui produit par le LHC, mais son analyse n'en nécessite pas moins des compétences spécifiques et des algorithmes complexes.

Les exemples de domaines scientifiques où l'on doit recueillir et gérer des données massives ne manquent pas. On peut citer la génomique *(voir l'article de G. Perrière)*, les relevés astronomiques *(voir l'article de C. Reylé)*, les recensements de la flore et de la faune, la recherche pharmacologique, les études démographiques, etc.

On le comprend, des algorithmes d'analyse de données massives sont indispensables à la recherche scientifique d'aujourd'hui. Mais il en faut également pour des applications plus quotidiennes. En voici un exemple.

Vous voulez choisir un film pour votre soirée ; vous vous connectez à Internet et allez sur un site détaillant les possibilités. Son système de recommandation vous fera des propositions. Il peut procéder de manière très simpliste, par exemple en vous suggérant le film récent le plus regardé au cours de la semaine, ou celui ayant reçu les meilleures critiques. Mais la recommandation peut aussi résulter d'un algorithme complexe qui prend en compte les films que vous avez aimés, votre humeur, peut-être les goûts de celui ou celle qui partagera votre soirée. Pour vous aider à répondre à la simple question « Quel film pourrais-je regarder ? », l'algorithme tiendra peut-être compte d'un océan de données : des avis de centaines de millions de personnes sur des milliers de films.

Or les programmes qui permettent de réaliser des analyses statistiques, même très simples, sur de telles quantités de données sont d'une grande complexité. Dans cette masse, le logiciel de recommandation de films trouvera des personnes qui ont des goûts voisins des vôtres, révélera des proximités avec des internautes que vous ne connaissez pas. Il pourra découvrir vos goûts cinématographiques et vous proposera des films en moins d'une seconde !

Dans la jungle des chaînes de télévision de plus en plus nombreuses, des VoD (*Video on Demand*, vidéo à la demande) et SVoD (*Subscription Video on Demand*, vidéo à la demande avec abonnement), des innombrables films disponibles légalement ou pas sur la Toile, l'usager est perdu ; le but des systèmes de recommandation est de l'aider à s'y retrouver.

L'ingrédient principal qui alimente de tels systèmes est bien sûr constitué par les données numériques, lesquelles tiennent une place de plus en plus importante dans le monde d'aujourd'hui. Depuis les années 1960, les logiciels de bases de données se sont imposés pour partager des données au sein d'une entreprise ou d'une organisation. D'abord isolées dans des centres de calcul, ces données sont devenues accessibles partout dans le monde avec l'arrivée d'Internet, le réseau des réseaux de machines, puis du Web, le réseau des contenus, et finalement du Web 2.0 avec la participation de chacun, les réseaux d'individus.

Et il n'y a pas que les données stockées, il y a aussi les données échangées. Nous sommes entourés de milliards d'objets communicants. En 2008, le Web comptait déjà plus de 1 000 milliards de pages et, chaque mois, les internautes y réalisaient des dizaines de milliards de recherches. On estime que le monde numérique double en volume tous les 18 mois et le trafic sur Internet est déjà, chaque année, supérieur à tout ce que l'on pourrait stocker en utilisant tous les disques et autres supports disponibles.

Analyser les données pour les valoriser

L'ensemble de ces données disponibles sur le réseau mondial constitue un énorme gisement de connaissances à découvrir et à valoriser. L'analyse de données a été un domaine très actif presque depuis les débuts de l'informatique et a revêtu divers noms, tels que *fouille de données* ou *business intelligence*. En raison de l'accroissement des capacités des disques et des mémoires, ainsi que des puissances de calcul avec des grappes d'ordinateurs pouvant réunir jusqu'à des milliers de machines, en raison aussi de l'explosion du volume de données disponibles, l'analyse de données pour en extraire de la valeur est devenue une industrie florissante. Et c'est sous le nom de *Big Data* (données massives) qu'elle se développe aujourd'hui.

Le point de départ est de valoriser les gisements massifs de données. Le *Big Data* inclut en général deux aspects. D'une part, il sous-entend l'idée de croiser des données très structurées, par exemple celles d'une entreprise, avec des masses de données moins structurées et plus défectueuses disponibles sur le Web. D'autre part, il nécessite la mise en œuvre de calculs massivement parallèles en utilisant des techniques logicielles telles que *Hadoop*, issues des moteurs de recherche du Web *(voir l'encadré page ci-contre)*.

Des difficultés multiples

L'objectif étant de faire émerger de nouvelles connaissances à partir des données, les tâches sont celles, classiques, de l'analyse de données, qui vont de leur acquisition à leur exploitation *(voir l'encadré page 33)*. Les difficultés sont nombreuses et tiennent en premier lieu des quatre « V » du *Big Data* : le volume des données (il se compte typiquement en téraoctets ou pétaoctets, soit 10^{12} ou 10^{15} octets), leur variété ou hétérogénéité (sur le plan de la structure, de la langue, du format, etc.), leur vélocité (le rythme auquel elles sont modifiées), leur véracité (erreurs, incomplétude, confiance, provenance, fraîcheur, etc.).

D'autres difficultés proviennent de la répartition des données dans l'espace, des protections éventuelles (droit d'accès, restriction sur leur usage), etc. Par ailleurs, la nature des traitements auxquels ces données sont soumises a une importance considérable. Par exemple, un algorithme dont le temps d'exécution est proportionnel au cube du nombre de données reste inutilisable sur une base contenant un milliard d'enregistrements, même avec des milliers de machines qui fonctionneraient pendant des centaines d'années. Quant aux logiciels tels que *Hadoop*, développés spécifiquement pour traiter des données massives, ils sont encore relativement jeunes et compliqués à utiliser, en particulier parce qu'ils font travailler ensemble un grand nombre de machines.

L'analyse de grands volumes de données intervient dans d'innombrables domaines, afin de faire des prédictions de plus en plus fines, d'anticiper des épidémies, de mieux comprendre l'évolution du climat, d'aider à soigner les cancers, etc. Au niveau du

LES AUTEURS

Serge ABITEBOUL est directeur de recherche à l'INRIA-Saclay et à l'École normale supérieure de Cachan. Il est membre du Conseil national du numérique.

Pierre SENELLART est maître de conférences en informatique à *Télécom ParisTech*.

BIBLIOGRAPHIE

S. Abiteboul, **Sciences des données : de la logique du premier ordre à la Toile**, Leçon inaugurale au Collège de France, Fayard, 2012 : www.college-de-france.fr/site/serge-abiteboul/

S. Abiteboul *et al.*, **Web Data Management**, Cambridge University Press, 2011 : http://webdam.inria.fr/Jorge

D. Agrawal *et al.*, **Big data and cloud computing : current state and future opportunities**, EDBT/ICDT 2011, www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a50-agrawal.pdf

C. Lynch, **Big data : How do your data grow ?**, *Nature*, vol. 455, pp. 28-29, 2008.

G. Linden *et al.*, **Amazon.com recommendations : Item-to-item collaborative filtering**, *IEEE Internet Computing*, vol. 7(1), pp. 76-80, 2003.

grand public et pour l'instant, ce qui est le plus apparent du *Big Data*, c'est l'utilisation de données personnelles par de grandes entreprises commerciales, principalement pour des publicités ciblées. C'est le cas de *Google*, qui analyse les requêtes et les courriels des internautes pour mieux cibler sa publicité, ou d'*Amazon*, qui suggère aux usagers des livres à acheter.

Des armes nouvelles aux mains des dictatures... et des citoyens

Récemment, en particulier dans le cadre de l'affaire Edward Snowden aux États-Unis, la presse a souligné que divers gouvernements utilisent l'analyse de données privées à des niveaux surprenants. La principale raison invoquée pour ce type d'utilisations est la lutte contre le terrorisme. Mais le système PRISM de la NSA (*National Security Agency*, l'agence américaine de sécurité) et les systèmes équivalents des autres États sont aussi utilisés pour l'espionnage, notamment industriel. Ils peuvent l'être pour surveiller les opposants politiques. Le contrôle et l'analyse des données sont certainement les nouvelles armes parmi les plus inquiétantes des dictatures et des gouvernements totalitaires.

A contrario, la généralisation dans les pays démocratiques de l'*open data* (l'accès libre aux données du secteur public) devrait permettre aux « journalistes de données », et plus généralement aux citoyens concernés, de contrôler les actions de leurs gouvernants ainsi que celles des grandes entreprises. Les mêmes technologies rendent possibles la surveillance des personnes par l'État et celle de l'État par les citoyens. Les gouvernements de pays tels que les États-Unis, le Royaume-Uni, ou plus récemment la France se sont engagés dans le mouvement de l'ouverture des données. Cela ouvre la voie à de nouveaux services et permet aussi de mieux contrôler les actions des gouvernements et des grandes entreprises. On peut espérer que cela conduise à donner plus de responsabilité au citoyen et à refonder la démocratie.

Les technologies du *Big Data* semblent particulièrement adaptées pour prévoir des crises sanitaires, des problèmes environnementaux, des catastrophes naturelles, et pour y réagir. Elles devraient aider à résoudre les problèmes de santé, de transport, d'écologie, à lutter contre la pauvreté. Dans nombre de ces situations, il faut combiner des analyses de grandes masses de données stockées et des analyses en ligne sur un flux de données obtenues en temps réel. De telles combinaisons se retrouvent ainsi, par exemple, dans le suivi personnalisé de personnes en grande difficulté, de personnes très âgées, d'élèves en échec scolaire, etc. Il faut être capable d'analyser des évolutions sur de longues périodes, de prévenir si possible les problèmes, mais aussi savoir détecter l'urgence et réagir en temps réel à des crises.

Prenons l'exemple de la santé. Les données en rapport avec la santé d'un individu croissent sans cesse. Le cœur en est constitué par des informations telles que les examens médicaux, les diagnostics, les soins et les prises de médicaments. Mais on voit aussi se généraliser les données génomiques – la société californienne de biotechnologie *23andMe* propose ainsi aux individus le séquençage d'une partie importante de leur génome pour 99 dollars (moins de 80 euros). Il faut aussi prendre en compte les données sur la vie quotidienne de l'individu, son alimentation, son activité physique, son exposition à des pollutions particulières, etc. De plus en plus, de telles informations deviennent disponibles *via* des équipements de type téléphone intelligent ou *via* les réseaux sociaux.

Parallèlement, les chercheurs dans le domaine médical font un emploi croissant de l'analyse de données. Par exemple, ils espèrent découvrir des corrélations entre la prise de certaines combinaisons de médicaments et des pathologies particulières. Les patients, surtout, pourraient en profiter. Il s'agit d'abord de personnaliser les soins en adaptant les médicaments à chacun, en contrôlant les quantités administrées. On peut imaginer agir de manière préventive en proposant à chacun une hygiène de vie adaptée à ses risques, en accompagnant chaque personne dans sa nutrition, ses activités physiques, sa vie quotidienne. Mais on peut aussi espérer détecter les crises et aider à les gérer, par exemple en régulant la prise de médicaments.

Un problème : l'accès aux données personnelles

Le fait de disposer de toutes les données relatives à une personne ouvre ainsi considérablement le champ des possibles. Mais cela soulève le problème de l'accès aux données personnelles. Un assureur ou un employeur doit-il avoir accès à tout ou partie de l'information constituée par les données médicales et génomiques d'un client, ses données fiscales, ses achats, sa géolocalisation, ses courriels, ses échanges dans les réseaux sociaux ?

Ce sont les données personnelles du client. À ce titre, elles lui appartiennent, et il devrait être seul à pouvoir décider qui y a accès et comment elles sont utilisées. Mais ce n'est pas aussi simple. Par exemple, pour faire progresser la médecine, on devrait pouvoir faire des analyses sur les données médicales de tous. Il semble raisonnable que les résultats de ces statistiques soient publics. Mais il semble tout aussi clair que les données brutes, par exemple celles d'un hôpital, ne puissent pas elles-mêmes faire partie des données ouvertes, même anonymisées, puisqu'il serait impossible de garantir leur confidentialité. Il reste beaucoup à faire pour concilier ces différents aspects. ■

Hadoop pour gérer les données massives

Hadoop, logiciel libre de la fondation *Apache*, a été conçu en 2004 par l'Américain Doug Cutting. Ce logiciel est adapté à l'analyse de grandes masses de données. Il est fondé sur la technique *MapReduce*, que *Google* a utilisée pour son moteur de recherche.

Dans un calcul *MapReduce*, on commence par découper le problème en de nombreux sous-problèmes indépendants (étape *Map*) que l'on confie à des ordinateurs distincts. Ces machines résolvent les sous-problèmes et envoient leurs résultats à d'autres machines qui ont pour tâche de combiner les résultats (étape *Reduce*). L'objectif est de pouvoir travailler sur d'énormes volumes de données en parallélisant les calculs sur des grappes d'ordinateurs. *MapReduce* ne permet de traiter que des problèmes qui peuvent se décomposer en de multiples tâches parallèles.

Le logiciel de *MapReduce* le plus populaire est *Hadoop*, à la base des centres de données de géants du Web tels que *Amazon* et *Facebook*, et on le retrouve de plus en plus dans des offres de « Cloud computing ». Cependant, malgré ses atouts, *Hadoop* est un logiciel encore jeune et imparfait : il n'utilise pas un langage standardisé, ce qui rend sa gestion complexe, et ne réalise qu'un traitement différé, et non en temps réel, des données.

BULLETIN D'ABONNEMENT

POUR LA SCIENCE

À découper ou à photocopier et à retourner accompagné de votre règlement dans une enveloppe non affranchie à : Groupe Pour la Science • Service Abonnements • Libre réponse 90 382 • 75 281 Paris cedex 06

Oui, je m'abonne au magazine *Pour la Science* :

❑ **1 an • 12 numéros • 56 €** au lieu de ~~74,90 €~~

Tarif valable uniquement en France métropolitaine et d'outre-mer. Pour l'étranger, participation aux frais de port à ajouter : Europe 12 € – autres pays 25€.

❑ **2 ans • 24 numéros • 106 €** au lieu de ~~149,80 €~~

Tarif valable uniquement en France métropolitaine et d'outre-mer. Pour l'étranger, participation aux frais de port à ajouter : Europe 25 € – autres pays 51 €.

Pour 20 € de plus, recevez Hors-Série *Dossier Pour la Science* tous les trimestres !

Je préfère m'abonner à *Pour la Science* (12 nos/an) + le Hors-Série *Dossier Pour la Science* (4 nos/an)

❑ **1 an • 16 numéros • 76 €** au lieu de ~~102,70 €~~

Tarif valable uniquement en France métropolitaine et d'outre-mer. Pour l'étranger, participation aux frais de port à ajouter : Europe 16 € – autres pays 35 €.

❑ **2 ans • 32 numéros • 143 €** au lieu de ~~205,40 €~~

Tarif valable uniquement en France métropolitaine et d'outre-mer. Pour l'étranger, participation aux frais de port à ajouter : Europe 34 € – autres pays 72 €.

J'indique mes coordonnées :

Nom : ____________________

Prénom : ____________________

Adresse : ____________________

____

CP: ______ Ville : ____________________

Pays : ____________ Tél. : ____________

Pour le suivi client (facultatif)

Je choisis mon mode de règlement :

❑ Par chèque à l'ordre de *Pour la Science*

❑ Par carte bancaire

Numéro de carte ________________

Date d'expiration ________

Signature obligatoire

Mon e-mail pour recevoir la newsletter *Pour la Science* (à remplir en majuscule).

____

En application de l'article 27 de la loi du 6 janvier 1978, les informations ci-dessus sont indispensables au traitement de votre commande. Elles peuvent donner lieu à l'exercice du droit d'accès et de rectification auprès du groupe Pour la Science. Par notre intermédiaire, vous pouvez être amené à recevoir des propositions d'organismes partenaires. En cas de refus de votre part, merci de cocher cette case ❑.

DOS81 • Offre réservée aux nouveaux abonnés, valable jusqu'au 31.01.2014