PAROLE D'EXPERT

Big data, décodage et analyse des enjeux

Intégré aux investissements d'avenir, le Big data reste une notion floue. Décryptage par Philippe Rigaux, professeur au Centre d'études et de recherches en informatique et communications (Cédric) du Cnam.

Big data ! Tout le monde a entendu cette expression qui, depuis quelques temps, s'est diffusée à partir de quelques cercles d'experts jusqu'à atteindre les médias et, à travers eux, le grand public. Le mot est dans l'air, il impressionne, il semble désigner des enjeux importants et compliqués, mais il n'est pas toujours facile de savoir ce qu'il désigne précisément et pourquoi on devrait lui accorder tant d'importance. Tentons de répondre simplement à quelques questions générales.

Mais qu'est-ce que c'est ?

Sur ce point les choses sont claires : le **Big data** (on dira peut-être bientôt, la **datamasse**), désigne l'**explosion du volume des données numérisées,** collectées par des particuliers, des acteurs publics, des applications informatiques qui regroupent des communautés d'utilisateurs à l'échelle de la planète**.** Il suffira de citer quelques exemples plus ou moins connus de tous : **Google**, son moteur de recherche et ses services ; les réseaux dit sociaux : **Facebook**et son milliard d'utilisateurs qui déposent des images, des textes, des échanges ; les sites de **partage et de diffusion d'images et de photos** (FlickR) ; les sites communautaires (blogs, forums, wikis) ; les services administratifs et leurs échanges numérisés. Au centre de tous ces aspirateurs de **données**, on trouve Internet, le Web, et sa capacité à fédérer dans l'espace numérisé des milliards d'utilisateurs, mais également la profusion de capteurs de toutes sortes, accumulant des données scientifiques à un rythme inédit (images satellites par exemple). Pour en rester au Web, tous les messages, tous les documents, toutes les images et vidéos sont captés par des applications qui, en échange des services fournis, accumulent d'immenses **banques de données.** On parle en **millions de serveurs pour Google, Facebook ou Amazon,** stockés dans d'immenses hangars qui, par ailleurs, consomment une part non négligeable de l'électricité produite.

Et le mouvement semble aller s'accélérant, comme le suggèrent les innombrables études et rapports sur le sujet. Ainsi, **90 % de la masse d'information** créée depuis le début de l'humanité l'a été ces deux dernières années. Selon le cabinet McKinsey, la masse des données engendrée par les entreprises, les machines et les particuliers augmente chaque année de 40 %. Chaque minute, 150 000 tweets sont postés, 700 000 nouveaux commentaires sont écrits sur Facebook, plus de 30 heures de vidéo sont chargées sur YouTube... Arrêtons là : à ce stade, nous (c'est-à-dire, notamment, tous ceux connectés au Web) produisons des masses gigantesques de données, stockées dans des espaces informatiques d'une taille difficilement imaginable.

Mais pour quoi faire ?

C'est sans doute la question qui vient en premier lieu à l'esprit. **Pourquoi accumuler tant de données ?**Et à qui cela profite-t-il ? Pour tenter de répondre brièvement et (trop) simplement, la valeur du **Big data** vient de sa capacité à produire des informations **statistiques**de très grande ampleur, et d'en tirer des modèles relatifs aux **comportements**, aux tendances, aux opinions, aux corrélations de toutes sortes, à leur émergence ou disparition, etc… Les méthodes d'**analyse et fouille de données** sont ici mises à rude épreuve pour extraire ces informations et modèles de telles masses de données.

Concrètement, le **Big data** trouve des applications dans des domaines aussi divers que la **santé**(apparition, diffusion des épidémies, corrélations symptômes/maladies), les **services**collectifs (transports, enseignement), les études d'opinion, la rationalisation des consommations énergétiques, la recherche scientifique (astronomie, biologie). Enfin, et peut-être surtout, une motivation forte est la collecte d'information sur nous, utilisateurs de la toile, pour connaître nos intérêts, nos envies, nos habitudes, nos comportements, et converger vers un objectif idéal : connaître chaque personne assez intimement pour lui **proposer le bon produit, au bon prix, au bon moment**.

Dans une large mesure, nos actions sur la toile sont observées, agrégées, analysées, vendues pour alimenter notre **profil d'acheteur ou consommateur** potentiel. Faites simplement l'expérience d'envoyer un message ou d'effectuer une recherche mentionnant une ville, un pays, un voyage : il ne faudra pas longtemps pour qu'on vous présente des offres de billets d'avion, d'hôtels, de location de voiture. Un algorithme, une machine, produit automatiquement de la recommandation : vos données sont là quelque part, soumises à un processus d'analyse destiné à produire de la **valeur économique.**

Le**Big data** est donc bel et bien là, en voie de développement accéléré, avec des enjeux multiples, sociétaux, techniques, scientifiques, l'émergence d'une nouvelle activité économique et donc de nouveaux métiers.

Enjeux sociétaux

Un enjeu immédiat, évoqué ci-dessus, est la capture par des entités que vous ne connaissez pas (et qui ne vous demandent pas explicitement votre avis) d'**informations personnelles**sur lesquelles vous n'avez aucun contrôle. Souvent d'ailleurs, les données fournies semblent anodines : quelques éléments de contact (nom, adresse, date de naissance), associées à un achat. C'est l'agrégation et le cumul d'informations que vous fournissez par différentes sources (messages, navigation, recherche, discussion, achats) qui finit par constituer votre « profil », redoutablement complet, et très probablement ineffaçable. Le **respect de la vie privée**, le droit à l'oubli, et quelques autres droits fondamentaux sont potentiellement mis en cause par le processus de constitution de ces datamasses dont, rappelons-le, un des objectifs semble bien être de répertorier les utilisateurs du Web comme un catalogue de produits revendables.

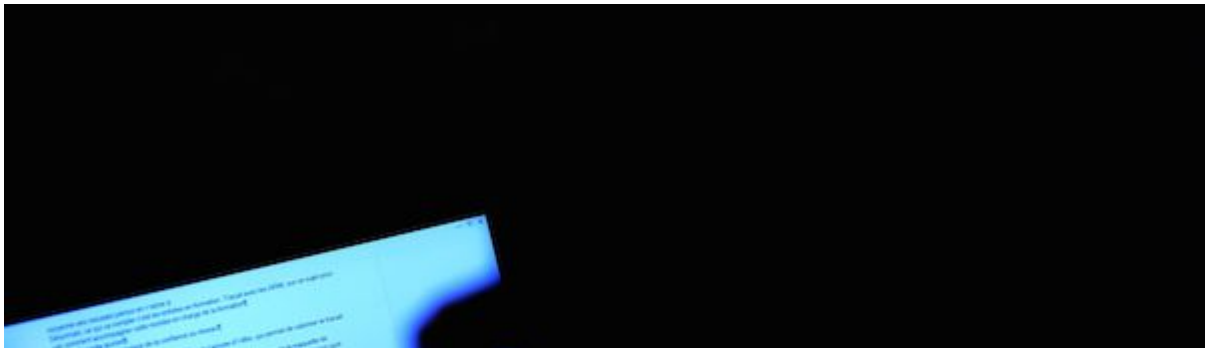
Enjeux : science et techniques

Le **Big data** est-il une rupture en termes techniques et scientifiques ? Le débat est ouvert. Il est certain que les problèmes soulevés par la**gestion de volumes** sans précédent a fortement stimulé des domaines comme le stockage et le calcul distribué, l'analyse automatique de texte, l'extraction de connaissances et, au sens large, la fouille de données. Mais dans certains cas (stockage et accès aux données par exemple) les solutions adoptées ressemblent à une régression : la « parallélisation » massive et la nécessité de tolérer les pannes multiples, qui apparaissent inévitablement dans des infrastructures conçues pour « passer à l'échelle » au moindre coût, ont pour contrepartie un appauvrissement des algorithmes ou de la finesse de gestion qui étaient devenus la norme dans les systèmes centralisés. Dans d'autres cas, il est clair que la motivation d'**analyser des données massives** et souvent hétérogènes est un encouragement à l'émergence de nouvelles recherches porteuses d'innovation. Quoi qu'il en soit, le **Big data** est maintenant un domaine privilégié par les financeurs de toutes sortes pour la création de projets, de structures, de débouchés technologiques, le tout considéré comme stratégique.

Enjeux : formation et métiers

Finalement, quel est l'impact sur la formation et les futurs métiers ? L'institut Mc Kinsey Global, estime que les besoins en analyse de masses de données induiront, aux Etats-Unis, le recrutement de 140 000 à 190 000 spécialistes d'ici à 2018. En termes de compétences, les technologies de la datamasse nécessitent la maîtrise d'outils mathématiques et statistiques de très haut niveau. Des compétences dans le domaine de l'informatique, et notamment en programmation, sont également requises.

Les gouvernements estiment donc qu'il existe une nécessité forte de former d'urgence à ce nouveau métier, dont la spécialité a d'ores et déjà été baptisée science des données. La création de nouvelles filières sur ce sujet est attendue et explicitement encouragée, comme l'indique le rapport intitulé « L'émergence d'une nouvelle filière de formation : data scientists » coordonné par Serge Abiteboul de l'Académie des sciences.



12 mai 2014