# Big data

## Description

Expression anglophone dÃ©signant lâ€™exploitation des masses de donnÃ©es numÃ©riques, quelle que soit leur nature (textes, chiffres, photos,Â sons, vidÃ©os, graphiques, clics, liens hypertextes, pages vues…), transmises grÃ¢ce Ã  Internet. Leur volume ne cesse dâ€™augmenter avec le dÃ©veloppement des usages numÃ©riques. La multiplication des terminaux connectÃ©s, la popularitÃ© des rÃ©seaux sociaux et lâ€™informatique en nuage (*cloud computing*, voir *REM* nÂ°9, p.43) en ont accÃ©lÃ©rÃ© la croissance depuis 2010. Les ordinateurs, les tÃ©lÃ©phones portables et les tablettes sont autant dâ€™Ã©metteurs de donnÃ©es numÃ©riques en provenance des particuliers, des entreprises, des institutions ou des Etats, auxquels sâ€™ajoutent les objets connectÃ©s (ampoules, rÃ©frigÃ©rateurs, compteurs Ã©lectriques, voitures, jouets, plantes vertes, pour ne citer que des applications domestiques), dont le nombre devrait atteindre 50 milliards en 2020 selon Cisco, ainsi que les camÃ©ras de surveillance, les sondes mÃ©tÃ©o et autres tÃ©lescopes. De grandes capacitÃ©s de stockage, des algorithmes et une puissance de calcul qui augmente de faÃ§on exponentielle constituent les ingrÃ©dients nÃ©cessaires Ã  lâ€™Ã©mergence de ce nouveau dÃ©fi consistant Ã  produire de la connaissance Ã  partir de lâ€™analyse du volume incommensurable de donnÃ©es gÃ©nÃ©rÃ©es par les techniques de lâ€™information et de la communication.

Le champ dâ€™application du *big data* est infini, des sciences dures aux sciences sociales, de lâ€™Ã©tude des phÃ©nomÃ¨nes climatiques ou astronomiques Ã  celle des mouvements sociaux ou des politiques Ã©conomiques, avec pour finalitÃ©, toujours, de mieux comprendre le monde. Plus prosaÃ¯quement, derriÃ¨re le concept de *big data* se cache aussi une activitÃ©Â lucrative liÃ©e Ã  Internet, celle de prÃ©dire les comportements des consommateurs en particulier et de dÃ©busquer ceux des citoyens en gÃ©nÃ©ral. Le marchÃ© du *big data* est estimÃ© Ã  50 milliards de dollars dans cinq ans. Sur les 4,4 millions dâ€™emplois crÃ©Ã©s dans le secteur de lâ€™informatique dâ€™ici Ã  2015, dont un peu moins de la moitiÃ© aux Etats-Unis, grÃ¢ce en partie au *big data*, deux tiers dâ€™entre eux pourraient ne pas trouver preneur faute de qualifications adÃ©quates. DotÃ© de compÃ©tences en informatique, en statistiques et en marketing, le *data* scientist est un profil dÃ©jÃ  recherchÃ© par un tiers des entreprises aux Etats-Unis.

En 2011, la somme des donnÃ©es enregistrÃ©es sur Internet atteint 1,8 zettaoctet (Zo), soit 1 800 milliards de gigaoctets. Chaque minute circulent sur le rÃ©seau mondial 2,3 millions de requÃªtes sur Google, 1,7 million de contenus partagÃ©s sur Facebook, 300 000 tweets, 200 millions de courriels, 72 heures de vidÃ©os mises en ligne, 13 400 nouveaux sites web, 12 000 applications tÃ©lÃ©chargÃ©es sur lâ€™App Store, etc. En 2013, 5 exaoctets de donnÃ©es sont engendrÃ©es en dix minutes seulement contre deux jours en 2011. Avec plus dâ€™un tiers de la population mondiale connectÃ©e aux rÃ©seaux, soit plus de 2

milliards dâ€™internautes, la somme dâ€™informations produites en 48 heures est Ã©gale Ã  celle produite depuis le dÃ©but de lâ€™humanitÃ©. En 2020, elle atteindra 40 zettaoctets, soit 40 000 milliards de gigaoctets.

Lâ€™exploitation de ce flux grossissant de donnÃ©es stockÃ©es dans dâ€™immenses fermes de serveurs Ã  travers le monde constitue un enjeu majeur pour la connaissance et la comprÃ©hension des activitÃ©s humaines, aussi bien dâ€™ordre Ã©conomique, social que politique. Le tÃ©lÃ©chargement de films ou de musiques, les Ã©changes de messages ou de photosÂ sur les rÃ©seaux sociaux, la gÃ©olocalisation, la navigation sur le Web, ainsi que les requÃªtes adressÃ©es aux moteurs de recherche constituent des donnÃ©es dont lâ€™analyse se rÃ©vÃ¨le riche dâ€™enseignements pour de nombreux secteurs dâ€™activitÃ©s scientifiques ou commerciales. Les capacitÃ©s de stockage informatique supÃ©rieures au tÃ©raoctet (To = 1 000 Go) et les puissances de calcul utilisÃ©es communÃ©ment par les gÃ©ants dâ€™Internet comme Google dÃ©passant le pÃ©taoctet (1 000 To) rendent dâ€™ores et dÃ©jÃ  possible le brassage dâ€™un nombre infini de donnÃ©es mises en ligne afin dâ€™en tirer des renseignements utiles aux stratÃ©gies des entreprises. InstallÃ©es dans le monde entier, les fermes de serveurs dÃ©tenues par le groupe Amazon hÃ©bergent plus de 1 000 milliards de documents, et plus de 500 millions de particuliers Ã  travers le monde avaient recours Ã  un service dâ€™informatique en nuage en 2012.

## L'échelle des octets

| | |
|---|---|
| Octet (1 o) | Unité de base (série de huit chiffres 0 ou1) |
| Kilooctet (1 Ko = 1 000 o) | Une page de texte = 30 Ko |
| Mégaoctet (1 Mo = 1 000 Ko) | Un morceau de musique = 5 Mo |
| Gigaoctet (1 Go = 1 000 Mo) | Un film de deux heures = 1 Go |
| Téraoctet (1 To = 1 000 Go) | Six millions de livres = 1 To |
| Pétaoctet (1 Po = 1 000 To) | Une pile de DVD de la hauteur de la tour Montparnasse = 1 Po |
| Exaoctet (1 Eo = 1 000 Po) | Toutes les informations produites depuis les premières mesures jusqu'en 2003 = 5 Eo |
| Zettaoctet (1 Zo = 1 000 Eo) | La totalité des données enregistrées en 2011 = 1,8 Zo |
| Yottaoctet (1 Yo = 1 000 Zo) | La capacité du nouveau data center de la NSA prévu en 2013 =1 Yo |

Source : CNRS, d'après *Les Echos*, 10 décembre 2012.

Le *big data* constitue un nouveau marchÃ© de lâ€™information sur lequel opÃ¨rent des experts en algorithmes, ainsi que des entreprises qui se sont faites comme spÃ©cialitÃ© de confÃ©rer du sens au traitement de milliards de fichiers numÃ©riques. Selon ces professionnels des *datas*, lâ€™exploitation des bases de donnÃ©es internet, assortie de lâ€™*open data* promu par les pouvoirs publics (data.gov depuis

2009 aux Etats-Unis et data.gouv.fr depuis fin 2011 enÂ France), est un facteur majeur dâ€™innovation, contribuant Ã  amÃ©liorer la productivitÃ© des entreprises. Selon le cabinet dâ€™Ã©tudes Gartner, *le big data* pourrait amÃ©liorer de 20 % leur performance dâ€™ici Ã  2015, autour de 6 % estime le MIT, ou encore permettre aux administrations europÃ©ennes dâ€™Ã©conomiser 250 milliards dâ€™euros par an selon McKinsey qui Ã©value le gain de productivitÃ© Ã  0,7 % dans le secteur de la santÃ©, 0,5 % dans lâ€™administration et de 0,5 % Ã  1 % dans le commerce. Au-delÃ  des applications marketing permettant un profilage de plus en plus prÃ©cis des consommateurs, le *big data* est dÃ©jÃ  utile Ã  de nombreux secteurs comme la santÃ©, la sÃ©curitÃ© ou les transports. Il est dÃ©jÃ  possible dâ€™utiliser les requÃªtes effectuÃ©es Ã  partir de Google pour suivre la progression dâ€™une Ã©pidÃ©mie de grippe ou encore de prÃ©voir les embouteillages routiers en se basant sur les donnÃ©es fournies parÂ les GPS des automobilistes. La police de Memphis dans lâ€™Etat du Tennessee, aux Etats-Unis, se sert dâ€™un programme dÃ©veloppÃ© par IBM qui analyse les interactions entre les jours de paye, la population par quartier et le calendrier des manifestations sportives, afin de dÃ©terminer les zones potentiellement Ã  risque en termes de criminalitÃ©. En septembre 2013, la National Security Agency (NSA), chargÃ©e du renseignement informatique aux Etats-Unis, inaugurera le plus grand centre de traitement de donnÃ©es au monde, capable de traiter plus dâ€™un yottaoctet (Yo = 1 000 Zo) de donnÃ©es, dans le but dâ€™archiver toutes les communications Ã  lâ€™Ã©chelle de la planÃ¨te.

Au rythme oÃ¹ vont les choses, les bases de donnÃ©es classiques construites sur lâ€™indexation de textes ou de chiffres pour rÃ©pondre Ã  des questions prÃ©dÃ©terminÃ©es appartiendront bientÃ´t Ã  lâ€™histoire de lâ€™informatique. Des modÃ¨les de programmation informatique permettent dÃ©sormais dâ€™assurer le traitement de trÃ¨s grandes quantitÃ©s de donnÃ©es (vidÃ©os, sons, images, voix…) rÃ©parties sur des grappes de serveurs (*clusters*), tel MapReduce de Google ou Hadoop, projet open source dÃ©veloppÃ© par la Fondation Apache et utilisÃ© par la plupart des grands groupes internet, notamment Facebook, Yahoo!, Twitter et Microsoft. Afin de vÃ©rifier sa e-rÃ©putation, le groupe Disney a passÃ© au crible tous les blogs, forums et rÃ©seaux sociaux de la Toile avec Hadoop. Louant des capacitÃ©s de mÃ©moire et de calcul auprÃ¨s des spÃ©cialistes de lâ€™informatique enÂ nuage, comme IBM ou Amazon, de nombreuses start-up prestataires de services spÃ©cialisÃ©s dans lâ€™analyse des donnÃ©es naissent sur le marchÃ© du *big data*. Mesagraph propose aux chaÃ®nes de tÃ©lÃ©vision des Ã©tudes dâ€™audience dâ€™un genre nouveau, basÃ©es sur les Ã©changes effectuÃ©s sur Twitter, restituant sous forme de graphiques le comportement des tÃ©lÃ©spectateurs sur le rÃ©seau social liÃ© Ã  la diffusion dâ€™un programme. Lâ€™algorithme inventÃ© par Tinyclues sert Ã  dÃ©terminer le sexe et lâ€™Ã¢ge dâ€™un internaute Ã  partir de son adresse Ã©lectronique, avec une marge dâ€™erreur ne dÃ©passant pas les 15 %. MFG Labs scrute les sites de partage de photos afin de dÃ©terminer les lieux de vacances privilÃ©giÃ©s des touristes. Semiocast surveille lâ€™image des marques Ã  partir des conversations en temps rÃ©el sur Twitter.

Pour les sciences sociales, le *big data* renouvelle les mÃ©thodes dâ€™enquÃªte, permettant de procÃ©der par induction, mettant en lumiÃ¨re des causalitÃ©s Ã  partir de lâ€™analyse des corrÃ©lations existantes entre des masses de donnÃ©es. Ignorant lâ€™individu, le *big data* dresse des profils statistiques Ã  son

insu. Selon les chercheurs, le risque existe alors de privilÃ©gier les corrÃ©lations crÃ©Ã©s de toute piÃ¨ce par les algorithmes au dÃ©triment du travail par vÃ©rification dâ€™hypothÃ¨ses prÃ©alablement dÃ©finies.

Au fur et Ã  mesure que nos vies se numÃ©risent, se pose la question de la protection des donnÃ©es privÃ©es contre le piratage, dont furent victimes, en 2011, plus de 100 millions de titulaires dâ€™un compte Sony (PlayStation Network, Sony Online Entertainment et Qriocity). Mais se pose aussi la question de la commercialisation lÃ©gale des donnÃ©es privÃ©es. La loi franÃ§aise LOPPSI 2 sur la sÃ©curitÃ© intÃ©rieure autorise le ministÃ¨re de lâ€™intÃ©rieur Ã  revendre Ã  des tiers les informations personnelles des dÃ©tenteurs dâ€™une carte grise obtenue depuis aoÃ»t 2011, notamment Ã  des fins de prospection commerciale. Les rÃ¨gles applicables au respect de la vie privÃ©e, Ã  la protection des libertÃ©s individuelles et Ã  lâ€™Ã©thique en matiÃ¨re dâ€™exploitation des donnÃ©es personnelles, ainsi que les normes de sÃ©curitÃ© contre le piratage ou la perte, se trouvent rÃ©guliÃ¨rement dÃ©passÃ©es par des techniques nÃ©es avec Internet en perpÃ©tuelle Ã©volution et imprÃ©visibles. En outre, les donnÃ©es archivÃ©es dans des fermes de serveurs rÃ©parties dans le monde entier posent le problÃ¨me crucial de la Â« *souverainetÃ© numÃ©rique* Â»,Â notamment au regard du *Patriot Act* amÃ©ricain. Cette loi antiterroriste, votÃ©e Ã  la suite des attentats du 11 septembre 2001, autorise les autoritÃ©s amÃ©ricaines Ã  accÃ©der aux donnÃ©es numÃ©riques europÃ©ennes hÃ©bergÃ©es par des infrastructures appartenant Ã  des entreprises amÃ©ricaines, mÃªme si celles-ci sont installÃ©es en Europe. Face Ã  la domination des entreprises amÃ©ricaines sur le marchÃ© de lâ€™informatique en nuage, les Etats europÃ©ens, notamment lâ€™Allemagne, le Royaume-Uni et la France (projet AndromÃ¨de), ont entrepris de construire leur nuage souverain pour hÃ©berger leurs donnÃ©es nationales (administrations, dÃ©fense, nuclÃ©aire…).

Les donnÃ©es liÃ©es aux traces laissÃ©es par les internautes sont le nouvel or noir de la sociÃ©tÃ© de lâ€™information, et la valorisation de cette matiÃ¨re premiÃ¨re numÃ©rique est principalement lâ€™affaire des groupes amÃ©ricains, Google, Apple, Facebook ou Amazon, qui rÃ¨gnent sur Internet. AprÃ¨s un an et demi dâ€™enquÃªte, la Federal Trade Commission a dâ€™ailleurs annoncÃ©, le 3 janvier 2013, quâ€™elle abandonnait les poursuites pour abus de position dominante Ã  lâ€™encontre de Google, accusÃ© de privilÃ©gier ses propres services aux dÃ©pens de ceux de ses concurrents dans les rÃ©ponses apportÃ©es aux requÃªtes des internautes sur son moteur de recherche. Les conclusions de lâ€™enquÃªte similaire en cours menÃ©e par les services de la concurrence de la Commission europÃ©enne seront-elles moins clÃ©mentes, au regard des amendes substantielles infligÃ©es antÃ©rieurement aux gÃ©ants Microsoft et Intel ? En janvier 2013, une mission lancÃ©e par le gouvernement franÃ§ais propose de soumettre les gÃ©ants amÃ©ricains dâ€™Internet Ã  une taxe fiscale portant sur la collecte, la gestion et lâ€™exploitation commerciale de donnÃ©es personnelles des internautes localisÃ©s en France, en espÃ©rant les inciter ainsi Ã  plus de transparence dans leurs pratiques (voir *supra*).

Â« *DÃ©sormais, nous devrons faire confiance aux machines. Notre histoire appartient Ã  Google, Facebook et Apple. Sans les armÃ©es de serveurs de ces nouveaux maÃ®tres du monde numÃ©risÃ©, que restera- t-il de notre mÃ©moire individuelle, collective, de nos rÃªves et rÃ©alisations dans dix ans, un siÃ¨cle, un millÃ©naire ? Rien.*

Â» Ã©crit Jean-Christophe FÃ©raud dans *LibÃ©ration* du 3 dÃ©cembre 2012.

**Categorie**

1. A retenir
2. RepÃ¨res & tendances

**date crÃ©Ã©e**
21 dÃ©cembre 2012
**Auteur**
francoise