Note numéro 5
juin 2014

Note du CREOGN

Centre de Recherche de l'Ecole des Officiers de la Gendarmerie Nationale

BIG DATA

Big data, ou « masse de données » en français, est la collecte et le traitement de données massives, grâce à des supercalculateurs, capables d'obtenir des résultats que les moyens classiques de gestion de bases de données ne permettent pas d'atteindre. Le *big data* traite en temps réel des milliards de données structurées (déterminées a priori) ou non structurées (texte, image, son, vidéo, etc.). L'analyse et l'interprétation dégagent des tendances (*reporting*) et favorisent la prospective (prédiction). Selon le *Boston Consulting Group*, le *big data* devrait représenter 8% du PIB européen en 2020. En France, le gouvernement a retenu le développement de l'économie numérique parmi les cinq priorités du programme Investissements d'Avenir. Il a confié à l'Association Française des Éditeurs de Logiciels et Solutions Internet (AFDEL) la mission de contribuer à la structuration d'une filière *big data* (projet *Big Data Launchpad*). Les applications de cette nouvelle « révolution numérique » vont concerner l'ensemble des activités : recherche scientifique, recherche environnementale, météo, médecine (recherche biomédicale, thérapies ciblées, maladies, épidémies,), défense, sécurité, cybersécurité, transports, logistique, régulation du trafic, énergie (*smart grids*, compteurs « intelligents »), enseignement, traçabilité des aliments et des médicaments, marketing, etc. Le *big data* est un des moteurs de la transformation numérique.

Le Big data à la confluence des progrès scientifiques et technologiques

Le *big data* est une des conséquences du développement exponentiel des machines connectées via Internet. En 1969, lors de la création d'Arpanet, 4 ordinateurs étaient connectés. En 1984, on en dénombrait un millier, puis un million en 1992, dix millions en 1996, cent millions en 2001, un milliard en 2010... En 2016, le cap des 10 milliards de machines connectées sera dépassé. Selon les experts, avec le développement de l'internet des objets (IDO), le nombre d'appareils connectés devrait connaître une nouvelle croissance exponentielle due à la multiplication des objets pouvant communiquer dans le cyberespace. Selon les sources, en 2020, leur nombre devrait atteindre 26 milliards (Cabinet Gartner), 50 milliards (Ericsson), 80 milliards (Idate), 212 milliards (Cabinet IDC). Il y a fort à parier que ces perspectives seront dépassées. Le développement fulgurant des connexions de toute nature (homme/homme, homme/machine, machine/machine) est la conséquence de la transformation numérique de l'Etat, des entreprises et, pour les particuliers, de la démocratisation de l'accès à internet par le biais notamment des tablettes, smartphones, etc. Le glissement du protocole IPv4 vers le protocole IPv6 autorise la multiplication à l'infini des connexions, puisque l'on passe d'une capacité d'environ 4,3 milliards d'adresses IPv4 à 340 milliards de milliards de milliards de milliards d'adresses IPv6 (667 millions de milliards par mm² de la surface de la terre). Tout objet peut donc être connecté. Le moindre grain de sable du désert pourrait ainsi communiquer dans un cyberespace dont les dimensions ne sont plus perceptibles à l'échelle du raisonnement humain.

Le *big data* bénéficie des progrès réalisés dans les domaines des mathématiques fondamentales, de l'informatique distribuée (avec des plates-formes open source d'algorithmes comme *Hadoop* autorisant un traitement à grande échelle sur des milliers de nœuds), de la sémantique, de la linguistique et des statistiques. Il s'appuie sur le *cloud computing* (informatique en nuage), « interconnexion et coopération des ressources informatiques, situées au sein d'une même entité ou dans diverses structures internes, externes ou mixtes, et dont le mode d'accès est basé sur les protocoles et standards internet » (Syntec Numérique). Le *cloud* permet de stocker les données et de les traiter à moindre coût dans des serveurs de données (*data centers)*.

Le *Big data* profite également des développements du web 3.0, dit « web sémantique », qui permet de donner, grâce à des métadonnées, du sens à des données non structurées, allant ainsi au-delà de la simple indexation de documents réalisée par le web 2.0. Enfin, le *machine learning*, branche de l'intelligence artificielle associée au *Big data,* permet un apprentissage statistique qui, grâce à un corpus d'algorithmes, établit des prévisions de fonctionnement ou de comportements à partir d'un calcul de probabilités fondé sur une quantité gigantesque de données.

Les 3 « V » caractérisant le Big data

Les spécialistes du cabinet Gartner (cabinet américain de conseil et de recherche dans les technologies numériques) caractérisent le *big data* par 3 « V », initiales de volumétrie, de vitesse (ou vélocité) et de variété.

La volumétrie

Le *big data* est caractérisé par le volume croissant des données qu'il traite. Entre les débuts de l'informatique dans les années 40 et 2010, l'humanité a produit 1 zettaoctet de données (10^{21}). Le tableau ci-après rappelle l'échelle qui sert de référence.

Kilo (Ko)	Mega (Mo)	Giga (Go)	Tera (To)	Peta (Po)	Exa (Eo)	Zetta (Zo)	Yotta (Yo)
10^3	10^6	10^9	10^{12}	10^{15}	10^{18}	10^{21}	10^{24}
1 page de texte	1 CD Rom	1 film	1 disque dur du PC le plus puissant	Production quotidienne de données	Trafic 2013 des smartphones	Production de données année 2015	Production de données 2035?
30 Ko	650 Mo	1 Go	1To	2200 Po	1Eo	8Zo	

Aujourd'hui l'humanité produit en deux jours autant d'informations qu'elle en a créées entre l'Antiquité et 2003 (Eric Schmidt, ancien PDG de Google). En 2015, le patrimoine des données sera porté à 8 zettaoctets (source cabinet Gartner). En 2020, on devrait atteindre les 40 zettaoctets, voire davantage, si l'on admet que le volume des données double chaque année. A ce rythme, le yottaoctet (1000 zettaoctets) pourrait devenir la référence avant 2030. Le plus grand *data center* existant est capable de traiter simultanément un Yo.

Le volume des données augmente du fait de la croissance du nombre des sources et du développement des usages. Les sources sont encore principalement des êtres humains. De quelques milliers en 1990, le nombre d'internautes dépassera sans doute le seuil des 3 milliards en 2015. 618 millions de chinois étaient présents sur la Toile en 2013 ; ils seront un milliard en 2020, tandis que la réduction de la fracture numérique permettra d'accroître le taux de pénétration en Afrique et en Amérique du Sud. Le web et les réseaux sociaux multiplient les usages et donc la création de données et de métadonnées (celles qui servent à décrire une autre donnée, comme un texte, une image, une photo ou une vidéo). Chaque minute, 350.000 tweets, 15 millions de SMS et 200 millions de mails sont émis à l'échelle de la planète. Facebook compte un milliard d'utilisateurs, Twitter, plus de 500 millions. Chaque jour, 200.000 nouvelles vidéos sont mises en ligne sur YouTube.

Le nombre d'internautes est croissant, mais les être humains ne sont plus, depuis longtemps, les seuls à produire des données. Toute machine connectée au cyberespace contribue à la création de données (capteurs, puces RFID, caméras, objets connectés, téléphones mobiles).

La quantité des données n'est cependant pas suffisante pour obtenir des résultats pertinents. Leur qualité doit être vérifiée au risque de produire des informations erronées. Cette évidence conduit certains spécialistes à ajouter le « V », initiale de « véracité » des données aux caractéristiques du *big data.*

La vitesse ou vélocité

Big data, par *data mining,* analyse en temps réel un nombre de données toujours croissant et arrivant en flux continus (*streaming*). Cela nécessite une architecture extensible s'appuyant sur des grappes de serveurs permettant d'exécuter simultanément par distribution plusieurs applications. La réponse doit parvenir dans la période du cycle concerné, sauf à perdre tout son intérêt (par exemple, le traitement des informations utiles aux *traders* pour la cotation boursière en continu). Les systèmes de bases de données relationnelles classiques (SGBDR) qui stockent l'information dans des tables indexées, dimensionnées et hiérarchisées ne peuvent répondre à cette exigence avec le langage SQL (*structured query Language* – langage de requête structurée). Le *big data* s'appuie sur des systèmes de fichiers distribués de type NoSQL (not only SQL), dont le plus connu est Hadoop. Ensemble logiciel structurel de licence libre, géré par la fondation Apache, il utilise des algorithmes de type MapReduce 2.0 qui, après avoir découpé les milliards d'enregistrements en blocs de même taille, répartit l'exécution du traitement « parallèle distribué » sur des serveurs distants, puis enchaîne les résultats reçus (concaténation) avant d'adresser le résultat sous une forme visuelle (data visualisation).

La variété

Les systèmes de gestion de bases de données relationnelles classiques ne sont opérants, on l'a dit précédemment, qu'avec des données structurées (nom, prénom, date de naissance, taille, poids, etc.). Or, la plupart des données produites aujourd'hui sont des données semi-structurées (issues de pages web par exemple) ou non structurées. *Big data* peut traiter des données exprimées dans toutes les langues. Il peut aussi exploiter d'autres données contenues dans les dessins vectoriels, les images matricielles (en pixels), la vidéo, le son, les logs, les données spatiales (géolocalisation), les sites web, les blogs, les mails, les échanges sur les réseaux sociaux (Facebook, Twitter, LinkedIn...), les éléments de biométrie, etc.

La variété des données est liée aussi à celle des sources qui peuvent être privées ou publiques. Le développement de l'*open data* favorise la mise à disposition des données créées ou recueillies par les administrations publiques (en France, www.data.gouv.fr).

Les applications du *big data*

Big data permet une amélioration de la connaissance mais surtout un essor sans précédent de l'analyse prédictive. C'est un outil d'anticipation favorisant une prise de décision rapide et fondée sur des données « fraîches » (sous réserve de la véracité des données exploitées). Tous les secteurs publics ou privés sont concernés et peuvent trouver dans *big data* une source d'innovation, un levier de gain de productivité, une amélioration du service rendu au client, à l'usager, au patient, à l'élève, etc. *Big data* peut aussi améliorer la vie quotidienne en luttant contre la pollution, l'insécurité, les risques de toute nature.

Le domaine de la santé est sans doute un de ceux qui va le plus bénéficier du *big data.* Il apporte une aide inédite au diagnostic, au traitement de maladies, à l'optimisation des soins via des protocoles ciblés. Le projet BrainSCANr , exploitant plusieurs millions d'articles scientifiques, a permis d'identifier un lien entre des maladies et certaines parties du cerveau. En 2014, au Memorial Sloan Kettering, centre de cancérologie de New York, le programme d'intelligence artificielle d'IBM, baptisé Watson, va entrer en action…Il a absorbé des millions de pages de publications spécialisées et de rapports cliniques. Les autorités médicales attendent non seulement une meilleure connaissance de la maladie en vue de son traitement, mais aussi des pistes pour mieux la prédire et donc la prévenir. La prévention est aussi une des voies ouvertes, notamment grâce à l'anticipation des

épidémies, au suivi des patients par le biais de leurs données biométriques.

Le marketing utilise le *big data* pour mieux connaître le comportement, la localisation, voire les sentiments des clients. Il s'ensuit une micro-segmentation de la publicité, un ajustement des prix, un approvisionnement des magasins adapté à la clientèle, etc.

Les compagnies d'assurance ont recours au *big data* pour tarifer et maîtriser les risques qui sont suivis en temps réel. Elles améliorent la détection des fraudes en captant les signaux faibles issus de l'analyse comportementale. Elles peuvent prévoir la sinistralité future probable d'un assuré en fonction de son profil.

La distribution d'énergie est de plus en plus bénéficiaire du *big data*. Les données sont fournies par les réseaux intelligents (smart grids) qui optimisent la production, la distribution, la consommation. Pour mieux coupler l'offre et la demande, des compteurs d'électricité « intelligents » sont déployés en France. Le projet « Linky » est expérimenté en Indre-et-Loire et à Lyon. 35 millions de compteurs sont prévus en 2020, avec des relevés pouvant être opérés toutes les minutes. On imagine l'impact sur la réduction de la pollution d'une consommation et d'une production mieux maîtrisées.

Les exemples peuvent être multipliés : constructeurs automobiles améliorant la qualité de leurs véhicules grâce à l'exploitation des avis des clients et des données émises par les voitures connectées, logisticiens optimisant les transports et la gestion des stocks, services après-vente connaissant l'état des équipements suivis et agissant de manière préventive, enseignants ajustant leur cours à la réceptivité des élèves, etc.

Les forces de sécurité peuvent aussi tirer un profit considérable du *big data*. Par exemple, les villes de Détroit ou de Memphis, aux Etats-Unis exploitent les données sur la criminalité, les croisent avec toutes les autres données pour identifier de manière prédictive les lieux et les périodes propices à la commission d'infractions. Cette capacité d'anticipation, basée sur l'exploitation instantanée de millions de données, permet d'ajuster dans le temps et dans l'espace, des moyens humains et matériels de plus en plus rares et coûteux. Le *Big data,* exploitant notamment les données émises par les voitures connectées, peut sensiblement améliorer la gestion du trafic, réduire les causes d'accident, etc. On ne peut négliger l'apport du *big data* dans le recueil, le traitement et l'exploitation du renseignement, notamment dans la lutte contre le terrorisme ou la criminalité organisée. La lutte contre la fraude est également renforcée au profit notamment de l'action des douanes, des services fiscaux, des services de police et de gendarmerie luttant contre les infractions économiques et financières.

Enfin, le *big data* peut apporter une aide à la cybersécurité pour mieux détecter certaines attaques (notamment les attaques persistantes avancées - APT) et permettre de les attribuer ou d'identifier leur origine avec une fort probabilité.

Le *big data* est, avec le *cloud computing* et l'internet des objets, l'un des moteurs de la transformation numérique. Il va modifier la création de valeur en faisant des données « l'or noir du XXIème siècle ». Déjà, des *data brookers* se placent sur un marché en plein expansion. Le Boston Consulting Group évalue à 315 milliards de dollars la valeur des données à caractère personnel collectées en 2011 auprès des consommateurs européens. 80% d'entre elles sont détenues par les GAFA (Google, Amazon, Facebook, Apple, etc.). L'enjeu du *big data* présente de multiples facettes: il faut rentrer en possession de nos données pour ne pas laisser à d'autres le soin d'en tirer profit, mais aussi se préserver de la « Voracité » du *big data* qui pourrait, sans que l'on y prenne garde, « dévorer » notre vie privée. Il convient également d'avoir une politique industrielle et d'entreprendre une démarche de formation pour pourvoir en personnes qualifiées (*data scientist, chief data officer*) un secteur en plein essor, porteur de croissance et créateur d'emplois.