La technologie des mégadonnées (big data)

Nicolae Sfetcu

20.11.2019

Sfetcu, Nicolae, « La technologie des mégadonnées (big data) », SetThings (20 novembre 2019), URL = https://www.setthings.com/fr/la-technologie-des-megadonnees-big-data/

Email: nicolae@sfetcu.com

Une traduction partielle de :
Sfetcu, Nicolae, « Etica Big Data în cercetare », SetThings (6 iulie 2019), DOI: 10.13140/RG.2.2.27629.33761, MultiMedia Publishing (ed.), ISBN: 978-606-033-228-2, URL = https://www.setthings.com/ro/e-books/etica-big-data-in-cercetare/

Le terme *big data* désigne l'extraction, la manipulation et l'analyse des ensembles de données trop volumineux pour être traités de manière routinière. Pour cette raison, des logiciels spéciaux sont utilisés et, dans de nombreux cas, des ordinateurs et du matériel informatiques dédiés. Généralement, ces données sont analysées de manière statistique. Sur la base de l'analyse des données respectives, des prédictions de certains groupes de personnes ou d'autres entités sont généralement établies, en fonction de leur comportement dans diverses situations et à l'aide des techniques analytiques avancées. Ainsi, les tendances, les besoins et les évolutions comportementales de ces entités peuvent être identifiés. Les scientifiques utilisent ces données pour la recherche en météorologie, génomique, connectomique, (Nature 2008) simulations physiques complexes, biologie, protection de l'environnement, etc. (Reichman, Jones, and Schildhauer 2011)

Avec le volume croissant des données sur Internet, dans les médias sociaux, l'informatique en nuage, les appareils mobiles et les données gouvernementales, le big data constitue à la fois une menace et une opportunité pour les chercheurs de gérer et d'utiliser ces données, tout en maintenant les droits des personnes impliquées.

Définitions

Les mégadonnées comprennent généralement des ensembles de données dont les dimensions dépassent la capacité logicielle et matérielle habituelle, en utilisant des données non structurées, semi-structurées et structurées, l'accent étant mis sur les données non structurées. (Dedić and Stanier 2017) La taille du big data a augmenté depuis 2012, passant de dizaines de téraoctets à de nombreux exaoctets de données. (Everts 2016) Travailler efficacement avec le big data implique l'apprentissage des machines pour détecter les modèles, (Mayer-Schönberger and Cukier 2014) mais ces données sont souvent un sous-produit d'autres activités numériques.

Selon une définition de 2018, « les mégadonnées sont des données qui nécessitent des outils informatiques parallèles pour gérer les données », ce qui constitue un tournant dans le domaine de l'informatique, qui repose sur les théories de la programmation parallèle et le manque de certitude des modèles précédents. Le big data utilise des statistiques inductives et concepts d'identification des systèmes non linéaires pour déduire des lois (régressions, relations non linéaires et effets causals) à partir de grands ensembles de données avec une faible densité d'informations afin d'obtenir des relations et des dépendances ou de prédire des résultats et des comportements.

Au niveau de l'Union européenne, il n'y a pas de définition obligatoire mais, selon l'avis 3/2013 du groupe de travail européen sur la protection des données,

> « Le terme « **mégadonnées** » se rapporte à l'augmentation considérable de l'accès aux informations et de leur utilisation automatisée: il fait référence aux volumes gigantesques de données numériques contrôlées par des entreprises, des gouvernements et d'autres grandes organisations, qui sont ensuite analysés de façon approfondie en utilisant des

> algorithmes. Les mégadonnées peuvent servir à établir des tendances générales et des corrélations, mais leur utilisation peut également toucher directement les individus ». (European Economic and Social Committee 2017)

Le problème avec cette définition est qu'elle n'envisage pas de réutiliser des données personnelles.

Règlement no. 2016/679 définit les **données à caractère personnel** (article 4, paragraphe 1) comme

> « toute information se rapportant à une personne physique identifiée ou identifiable (ci-après dénommée "personne concernée"); est réputée être une "personne physique identifiable" une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale. »

La définition s'applique également au niveau de l'UE aux personnes non identifiées, mais qui peut être identifiée en mettant en corrélation des données anonymes avec d'autres informations supplémentaires. Les données personnelles, une fois anonymisées (ou pseudo-anonymisées), peuvent être traitées sans autorisation préalable, en tenant compte toutefois du risque de réidentification de la personne concernée.

Les dimensions du big data

Les données sont partagées et stockées sur des serveurs, via l'interaction entre l'entité impliquée et le système de stockage. Dans ce contexte, les big data peuvent être classés en systèmes actifs (interaction synchrone, les données d'entité sont envoyées directement au système de stockage) et passifs (interaction asynchrone, les données sont collectées par un intermédiaire, puis introduits dans le système.

De plus, les données peuvent être transmises directement consciemment ou non (si la personne dont les données sont transmises n'est pas notifiée de manière claire et en temps voulu). Les données sont ensuite traitées pour générer des statistiques.

Selon la cible des analyses statistiques respectives, les dimensions des données peuvent être : a) individuelles (une seule entité est analysée); social (nous analysons des groupes d'entités distincts au sein d'une population; et hybrides (lorsqu'une entité est analysée en fonction de son appartenance à un groupe déjà défini).

L'énorme production actuelle de données générées par les utilisateurs devrait augmenter de 2000% dans le monde d'ici 2020 et est souvent non structurée. (Cumbley and Church 2013) En général, le big data est caractérisé par:

- Volume (quantité de données) ;
- Variété (produits de différentes sources dans différents formats) ;
- Vitesse (rapidité d'analyse des données en ligne) ;
- Véracité (les données sont incertaines et doivent être vérifiées) ;
- Valeur (évaluée par analyse).

Le volume de données produites et stockées évolue actuellement de manière exponentielle, plus de 90% d'entre elles ayant été générées au cours des quatre dernières années. (European Economic and Social Committee 2017) Les volumes importants nécessitent une analyse rapide, avec un fort impact sur la véracité. Des données incorrectes peuvent causer des problèmes lorsqu'elles sont utilisées dans le processus de décision.

Un des problèmes majeurs du big data est de savoir si des données complètes sont nécessaires pour tirer certaines conclusions sur leurs propriétés, ou si un échantillon est suffisant. Big data contient dans son nom un terme lié à la taille, qui est une caractéristique importante du

big data. Cependant, l'échantillonnage (statistique) permet de sélectionner des points de collecte de données corrects parmi un ensemble plus large afin d'estimer les caractéristiques de la population entière. Le big data peut être échantillonné à travers différentes catégories de données lors du processus de sélection de l'échantillon à l'aide d'algorithmes d'échantillonnage pour le big data.

La technologie

Les données doivent être traitées avec des outils de collecte et d'analyse avancés, basés sur des algorithmes prédéterminés, afin d'obtenir des informations pertinentes. Les algorithmes doivent également prendre en compte les aspects invisibles pour les perceptions directes.

En 2004, Google a publié un article sur un processus appelé MapReduce, qui propose un modèle de traitement parallèle. (Dean and Ghemawat 2004) MIKE 2.0 est également une application open source pour la gestion de l'information. (MIKE2.0 2019) Plusieurs études de 2012 ont montré que l'architecture optimale pour traiter les problèmes liés au big data repose sur plusieurs couches. Une architecture parallèle distribuée distribue les données sur plusieurs serveurs (environnements d'exécution parallèle), ce qui améliore considérablement les vitesses de traitement des données.

Selon un rapport publié par le McKinsey Global Institute en 2011, les principaux composants et écosystèmes du big data sont : (Manyika et al. 2011) des techniques d'analyse des données (apprentissage automatique, traitement du langage naturel, etc.), des grandes technologies (informatique décisionnelle, informatique en nuage, bases de données) et vues (charts, graphiques, autres vues de données).

Les mégadonnées fournissent des informations en temps réel ou quasi réel, évitant ainsi le temps de latence autant que possible.

Applications

Les mégadonnées dans les processus gouvernementaux augmentent la rentabilité, la productivité et l'innovation. Les registres d'état civil sont une source pour le big data. Les données traitées aident dans des domaines critiques du développement, tels que les soins de santé, l'emploi, la productivité économique, la criminalité, la sécurité et la gestion des catastrophes naturelles et des ressources. (Kvochko 2012)

Big data fournit également une infrastructure permettant de mettre en évidence les incertitudes, les performances et la disponibilité des composants. Les tendances et les prévisions dans l'industrie nécessitent une grande quantité de données et des outils de prévision avancés.

Le big data contribue à l'amélioration des soins de santé en fournissant des médicaments personnalisés et des analyses prescriptives, des interventions cliniques avec évaluation du risque et analyse prédictive, etc. Le niveau de données générées dans les systèmes de santé est très élevé. Cependant, la génération de « données altérées » pose un problème urgent, qui augmente avec le volume de données, d'autant plus que la plupart d'entre elles sont non structurées et difficiles à utiliser. L'utilisation du big data dans les soins de santé a généré d'importants défis éthiques, ayant des implications pour les droits individuels, la vie privée et l'autonomie, la transparence et la confiance.

Dans les médias et la publicité, pour le big data sont utilisés de nombreux points d’information sur des millions de personnes pour servir ou transmettre des messages ou des contenus personnalisés.

Dans le domaine de l'assurance maladie, des données sont collectées sur les « déterminants de la santé », ce qui aide à élaborer des prévisions sur les coûts de la santé et à identifier les problèmes de santé des clients. Cette utilisation est controversée, en raison de la discrimination des clients ayant des problèmes de santé. (Allen 2018)

Les mégadonnées et les technologies de l'information se complètent, aidant ensemble à développer l'Internet des objets (IoT) pour interconnecter des appareils intelligents et collecter des données sensorielles utilisées dans différents domaines.

En sport, le big data peut aider à améliorer l'entraînement et la compréhension des compétiteurs à l'aide de capteurs spécifiques et à prédire les performances futures des athlètes. Les capteurs attachés aux voitures de Formule 1 collectent, entre autres choses, les données de pression des pneus pour rendre la consommation de carburant plus efficace.

En recherche

En science, les systèmes big data sont utilisés de manière intensive dans les accélérateurs de particules du CERN (150 millions de capteurs transmettent des données 40 millions fois par seconde, pour environ 600 millions de collisions par seconde, dont ils ne sont utilisés qu'après filtrage que 0,001% du total des données obtenues), (Brumfiel 2011) dans des radiotélescopes astrophysiques construits à partir de milliers d'antennes, décodant le génome humain (au départ, cela prenait quelques années, avec le big data peut être réalisé en moins d'une journée), des études climatiques, etc.

Les grandes entreprises informatiques utilisent des entrepôts de données de l'ordre de plusieurs dizaines de pétaoctets pour la recherche, les recommandations et le marchandisage. La plupart des données sont collectées par Facebook, avec plus de 2 milliards d'utilisateurs actifs mensuels, (Constine 2017) et Google, avec plus de 100 milliards de recherches par mois. (Sullivan 2015)

Dans la recherche on utilise beaucoup l'interrogation cryptée et la formation de grappes dans le big data. Les pays développés investissent actuellement beaucoup dans la recherche sur le

big data. Au sein de l'Union européenne, ces recherches sont incluses dans le programme Horizon 2020. (European Commission 2019)

Les programmes de recherche utilisent souvent des ressources API de Google et Twitter pour accéder à leurs systèmes big data, gratuitement ou avec paiement.

Les grands ensembles de données comportent des problèmes d'algorithme qui n'existaient pas auparavant et il est impératif de changer fondamentalement les méthodes de traitement. À cette fin, des ateliers spéciaux ont été créés. Ils réunissent des scientifiques, des statisticiens, des mathématiciens et des praticiens pour débattre les défis algorithmiques du big data.

Bibliographie

Allen, Marshall. 2018. "Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates." Text/html. ProPublica. July 17, 2018. https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates.

Brumfiel, Geoff. 2011. "High-Energy Physics: Down the Petabyte Highway." *Nature* 469 (7330): 282–83. https://doi.org/10.1038/469282a.

Constine, Josh. 2017. "Facebook Now Has 2 Billion Monthly Users… and Responsibility." *TechCrunch* (blog). 2017. http://social.techcrunch.com/2017/06/27/facebook-2-billion-users/.

Cumbley, Richard, and Peter Church. 2013. "Is Big Data Creepy." In . https://doi.org/10.1016/j.clsr.2013.07.007.

Dean, Jeffrey, and Sanjay Ghemawat. 2004. "MapReduce: Simplified Data Processing on Large Clusters." http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf.

Dedić, Nedim, and Clare Stanier. 2017. "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery." In *Innovations in Enterprise Information Systems Management and Engineering*, edited by Felix Piazolo, Verena Geist, Lars Brehm, and Rainer Schmidt, 114–22. Lecture Notes in Business Information Processing. Springer International Publishing.

European Commission. 2019. "Horizon 2020." Text. Horizon 2020 - European Commission. 2019. https://ec.europa.eu/programmes/horizon2020/en.

European Economic and Social Committee. 2017. "The Ethics of Big Data: Balancing Economic Benefits and Ethical Questions of Big Data in the EU Policy Context." European Economic and Social Committee. February 22, 2017. https://www.eesc.europa.eu/en/our-work/publications-other-work/publications/ethics-big-data.

Everts, Sarah. 2016. "Information Overload." Science History Institute. July 18, 2016. https://www.sciencehistory.org/distillations/magazine/information-overload.

Kvochko, Elena. 2012. "Four Ways to Talk About Big Data." Text. Information and Communications for Development. December 4, 2012. http://blogs.worldbank.org/ic4d/four-ways-to-talk-about-big-data.

Manyika, James, Michael Chui, Jaques Bughin, and Brad Brown. 2011. "Big Data: The next Frontier for Innovation, Competition, and Productivity." 2011. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Reprint edition. Boston: Eamon Dolan/Mariner Books.

MIKE2.0. 2019. "Big Data Solution Offering - MIKE2.0, the Open Source Methodology for Information Development." 2019. http://mike2.openmethodology.org/wiki/Big_Data_Solution_Offering.

Nature. 2008. "Community Cleverness Required." *Nature* 455 (7209): 1. https://doi.org/10.1038/455001a.

Reichman, O. J., Matthew B. Jones, and Mark P. Schildhauer. 2011. “Challenges and Opportunities of Open Data in Ecology.” *Science* 331 (February): 703. https://doi.org/10.1126/science.1197962.

Sullivan, Danny. 2015. “Google Still Doing At Least 1 Trillion Searches Per Year.” Search Engine Land. January 16, 2015. https://searchengineland.com/google-1-trillion-searches-per-year-212940.