

Vient de Paraître

Christophe BRASSEUR

Enjeux et usages du Big Data : Technologies, méthodes et mise en œuvre

Hermès, Coll. Management & informatique, 2013, 203 pages, ISBN 978-2746245204, 49€



Le terme « Big Data » est entré rapidement dans le langage courant sans avoir reçu de traduction reconnue en français. Littéralement les « grosses données », il s’agit d’exploiter les énormes masses de données que produit continuellement le réseau internet dans le but d’en tirer des informations et des connaissances nouvelles. Le « tsunami de données » est tel que les techniques et méthodes traditionnelles de gestion de données sont dépassées. Les moteurs de recherche ont été les premiers à être confrontés au problème et sont souvent la source de certaines solutions technologiques actuelles. Alors que le traitement des données s’est développé dans les SI depuis les débuts de l’informatique, notamment dans les systèmes à vocation décisionnelle et de type *Business Intelligence*, la nouveauté réside dans l’extension fulgurante de ces applications avec des outils, des méthodes et des domaines d’applications nouveaux et très variés. Quelques ouvrages ont été publiés en français à ce sujet, à propos notamment de domaines d’application spécifiques (tel que le marketing) ou d’outils technologiques particulier (tel que les bases de données NoSQL ou le *data mining*) ; celui-ci est d’ordre général et donne un aperçu assez complet du phénomène Big Data.

Cet ouvrage contient sept chapitres organisés en trois parties. Les trois premiers chapitres introduisent le sujet et font le tour des enjeux et des applications. Le premier chapitre revient sur le phénomène Big Data et ses spécificités, et évoque brièvement l’impact potentiel sur l’entreprise et l’économie. Le deuxième chapitre discute les enjeux en termes de création de valeur, de progrès scientifiques et humain, et de développement technologique. On y aborde aussi le problème de la qualité des données (non redondance, incomplétude, etc.) et les questions de protection de la vie privée. Le troisième chapitre est peut être le plus intéressant dans cette partie du livre : un large échantillon d’applications y sont brièvement présentés dans des domaines tels que le marketing, le journalisme, la gestion de l’énergie dans les réseaux électriques intelligents, les services publics ou la santé.

Les deux chapitres de la deuxième partie concernent les technologies du Big Data. Le premier traite d’abord les problématiques de stockage et gestion des données, en analysant au préalable les limites des systèmes standards de gestion des bases de données. Les principales technologies, telles que MapReduce de Google ou Hadoop de Yahoo, sont ensuite brièvement présentées de manière synthétique, sans trop entrer dans les détails techniques. La plateforme R pour l’analyse et le traitement statistique est ensuite évoquée avec une présentation des principaux acteurs du marché dont la plupart des offres incluent des briques issues du logiciel libre (tel que le système R justement). Le chapitre suivant discute les méthodes d’analyse des données, dont les fondements sont à l’intersection des statistiques, de l’informatique et de l’intelligence artificielle. Sont ainsi évoquées les techniques de prédiction issues du domaine de la fouille des données (*data mining*) et de l’apprentissage automatique (*machine learning*) telles que les règles d’association, la régression linéaire, la classification automatique, la segmentation (*clustering*) ou encore les réseaux de neurones. Ce chapitre aborde aussi

des techniques et applications plus spécifiques telles que la fouille de texte (*text mining*), l'analyse des réseaux sociaux, « l'approvisionnement par la foule » (*crowdsourcing*) ou le géomarketing.

La troisième partie est à vocation plus managériale. Le chapitre 6 explore le champ des compétences que doit posséder un spécialiste de l'exploration des données (*data scientist*) en insistant un peu sur le caractère multidisciplinaire, et surtout, sur la pénurie de talents. Il évoque aussi quelques formations de niveau Master 2 ou école d'ingénieur qui préparent à ce type de profil. Le chapitre 7 est une brève discussion de ce qu'est un projet Big Data et quelles spécificités par rapport à un projet SI classique. Concernant les risques encourus et le retour sur investissement, l'auteur souligne l'importance de la maturité de l'entreprise en termes de gouvernance des données et d'expériences antérieures dans la mise en place de SI décisionnels.

Cet ouvrage n'a pas vocation à être un manuel détaillé des techniques et méthodes du Big Data. Bien au contraire, c'est une présentation synthétique qui fait le tour de la question et permet au lecteur d'avoir une compréhension globale du phénomène. Il constitue en quelques sortes un point d'entrée pour d'autres lectures et d'autres ouvrages plus pointus sur le sujet. Il s'adresse autant à des praticiens (consultants, ingénieurs, DSI) qu'à des étudiants et enseignants chercheurs qui souhaitent mieux comprendre le Big Data et en avoir une vision d'ensemble.

Par Saïd ASSAR, Télécom Ecole de Management