Veille technologique : Big Data

En cette deuxième année de préparation du BTS SIO, il nous a été demandé de choisir un sujet de veille technologique. Ayant été l'un des derniers élèves à choisir, j'ai eu accès à peu de choix de sujet. J'ai donc choisi de faire des recherches sur le Big Data. Compte tenu de l'augmentation du volume et de l'hétérogénéité des données échangées, ce sujet me semblait être le plus intéressant parmi ceux m'étant proposés.

Définition du Big Data

Le Big Data ou données massives désigne des ensembles de données qui deviennent tellement volumineux et en provenance multiple, qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données.

Problèmes rencontrés par les bases de données actuelles

Le volume des données stockées aujourd'hui est en pleine expansion. A tel point que 90% des données dans le monde ont été créées au cours des 2 dernières années seulement. À titre d'exemple, Twitter générait, en janvier 2013, 7 teraOctets (valant 10^{12} octets) de données chaque jour, devancé par Facebook, qui génèrait 10 teraOctets par jour. Les bases de données actuelles n'ont bien sûr pas de limite concernant la quantité de données à stocker, mais le problème principal est la rapidité pour exécuter des requêtes et afficher le résultat.

La solution Big Data

Les grandes entreprises se lancent sur le mouvement Big Data comme l'un des grands défis informatiques de la décennie 2010-2020 et en on fait une de leurs nouvelles priorités de recherches et développement. Le Big Data couvre en effet 4 dimensions : Volume, Vélocité, Variété et Véracité.

Volume : les entreprises sont submergées de volumes de données croissants de tous types. Les données numériques créées dans le monde seraient passées de 1,2 zettaoctets (valant 10^{21} octets) par an en 2010 à 1,8 zettaoctets en 2011, puis 2,8 zettaoctets en 2012 et s'élèveront à 40 zettaoctets en 2020.

Vélocité : elle représente la fréquence à laquelle les données sont à la fois générées, capturées et partagées. Pour répondre aux besoins des processus chrono-sensibles, tels que la détection de fraudes, le Big Data doit être utilisé à mesure que les données sont collectées par l'entreprise, c'est à dire que les flux croissants de données doivent être analysés en temps réel afin d'en tirer le meilleur
résultat.

Variété : le Big Data se présente sous la forme de données brutes, complexes et parfois non-structurées. Quelques exemples de sources : les messages sur les sites de médias sociaux, les images et vidéos publiées en ligne, les signaux GPS de téléphones mobiles. Les analyses sont

d’autant plus complexes qu’elles portent sur des liens entre des données de natures différentes.

Véracité : c’est la fiabilité de l’information pour permettre aux entreprises de faire confiance aux données sur lesquelles elles se basent pour prendre leurs décisions.

Applications du Big Data

Le Big Data trouve une application dans de nombreux domaines : dans la recherche scientifique par exemple concernant la rapidité à décoder le génome humain, en politique dans l’analyse des opinions de la population pendant les campagnes d’élection, dans le secteur privé avec l’exploitation rapide des données clients compte tenu de la forte augmentation de leur volume des dernières années.

Le Big Data peut aider les entreprises à réduire les risques et faciliter la prise de décision grâce à l’analyse prédictive et une expérience client plus personnalisée, à la différence avec l’informatique décisionnelle traditionnelle qui utilise des statistiques descriptives.

Évolutions du Big Data

Afin de pouvoir exploiter au maximum le Big Data, de nombreuses évolutions doivent être effectuées afin de garantir que l’information pertinente arrive au bon endroit et au bon moment. Il faut reconsidérer les concepts de bases de la gestion de données qui ont été déterminés dans le passé.

Pour la recherche scientifique par exemple, il s’agit de donner des réponses rapides et peu coûteuses même approximatives pour permettre au scientifique de prendre des décisions dans sa recherche plutôt que de fournir une réponse complète et correcte qui prendrait du temps et des ressources. Dans le secteur privé pour les entreprises, le Big Data permet d’accélérer le temps d’analyse de l’ensemble des données clients et non seulement un échantillon de celles-ci. Il permet aussi aux entreprises de récupérer et de centraliser de nouvelles sources de données clients.

Il est nécessaire de concevoir des outils permettant de mieux visualiser, analyser et cataloguer les ensembles de données afin de permettre une recherche rapide guidée par la donnée.