05/06/14 - BigData : Du fantasme adolescent à la pratique épanouie entre adultes consentants

«
?
?
?
? »

Cette blague de Dan Ariely a fait le tour des réseaux sociaux fin 2013. En dehors d'être assez drôle, elle permet d'illustrer une vérité indéniable : la plupart des gens ne comprennent rien au Big Data.

Nous vous proposons donc aujourd'hui une petite leçon de choses.

Lorsque la plupart des gens parlent de Big Data, cela se résume en général par :

? une rapide définition : « *s'tu veux, le Big Data ça désigne des volumes de données tellement importants qu'il n'est pas possible de les analyser par des outils « classiques »* (penser à mimer les guillemets avec deux doigts) »,

? un acronyme qui fait classe : « *Alors nous en Big Data on parle des 3V : Volume (de données), Vélocité (de la production de ces données) et Variété (des formats, supports et sources)* »,

? un chiffre impressionnant à l'appui : « *Vous savez le nombre d'informations produites chaque année dans le monde ? Non ? Et bien, c'est 912,5 exaoctets ! Oui Madame, un neuf suivi de 20 zéros ! On est quand même peu de choses*»,

? des perspectives vertigineuses : « *seulement 1% de l'information est analysée aujourd'hui ! Tu te rends compte du potentiel ?!* »

En étoffant un peu et en ajoutant un beau graphique ici ou là à l'appui, on pense avoir fait le tour de la question et on peut aller se coucher avec la satisfaction du devoir accompli.

...Seulement cette vision est un peu trop simpliste pour illustrer réellement ce qu'est le Big Data :

Tout d'abord la taille de la donnée n'est pas forcément proportionnelle à son contenu. Avec l'augmentation des débits réseaux et des capacités de stockage informatiques, on a assisté à une explosion du volume des données et des programmes. Or qui dit donnée ne dit pas forcément information. La moindre photographie ou vidéo prise aujourd'hui avec un mobile prend plusieurs méga octets de stockage, mais l'information exploitable en est extrêmement faible, voire nulle.

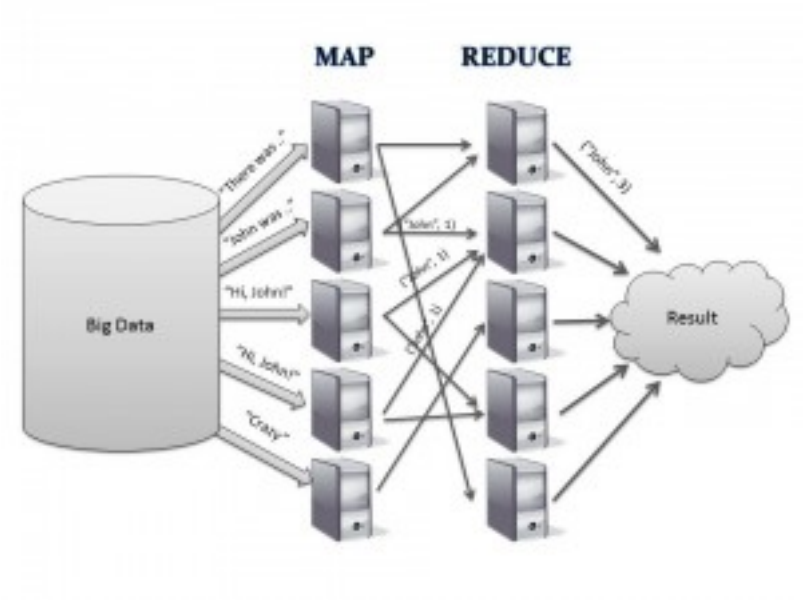
Au final, quelle est la part vraiment utile dans l'ensemble de l'information produite chaque jour sur Internet lorsque l'on enlève la pornographie (30% du trafic mondial, quand même !), les LOL/MDR et autre vidéos de chats-trop-mignons ?

Ensuite, la pertinence de l'analyse restera toujours et encore directement dépendante de la qualité de l'information manipulée. Alors que la plupart des sociétés peinent encore aujourd'hui à produire des données internes de gestion fiables, quelle confiance accorder à de l'information générée librement et sans aucun contrôle sur un réseau social ? Les algorithmes sont-ils capables d'interpréter toute la richesse du langage humain, comme par exemple un article sur le Big Data truffé d'allusions sexuelles ?

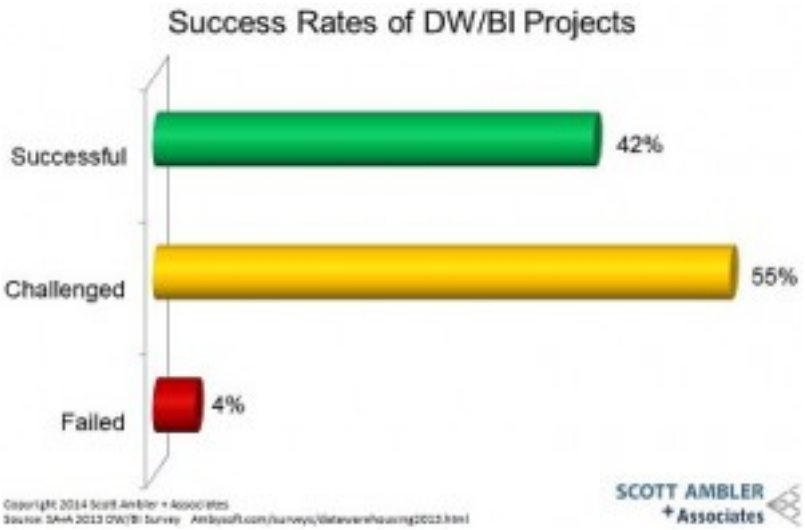
Enfin, La traduction de Big Data en « Gros volumes de données » est inadaptée ou en tout cas incomplète : La réelle problématique ne vient pas tant du nombre d'informations traitées mais plutôt du nombre de paramètres à prendre en compte pour dégager du sens à partir de cette information.

Si la plupart des systèmes décisionnels sont aujourd'hui à même d'agréger des téraoctets de données, ils deviennent rapidement inopérants lorsque le nombre d'axes d'analyses augmente.

L'intérêt des systèmes Big Data réside justement dans leur capacité à décomposer des calculs extrêmement complexes en une série d'opérations simples pouvant être distribuées sur des clusters et exécutées en parallèle, puis synthétisées pour obtenir un résultat final. Cette technologie appelée MapReduce (Décomposition/réduction) est un des constituants du framework Hadoop.



Les données peuvent également être stockées dans des bases de données non structurées dédiées (Hbase ou MongoDB par exemple). Ces systèmes sont également capables d'apprentissage, les arbres de corrélation évoluent en fur-et-à-mesure des traitements. La fiabilité et la précision des analyses ? et donc la complexité des algorithmes ? augmente donc au fur et mesure des traitements, par auto-évaluation et auto-correction.
Le risque étant que cette complexification ne finisse par rendre le système incompréhensible par un esprit humain : Sommes-nous prêts à faire aveuglément confiance à un programme informatique, uniquement parce qu'il est considéré comme statistiquement fiable ?Au final si le Big Data ouvre des champs d'exploration infinis dans des domaines extrêmement variés (recherche médical, sécurité, marketing et analyse de comportement client notamment) , il ne peut pas être LA réponse unique à toutes les problématiques décisionnelles rencontrées par les entreprises.
Comme toujours, il convient de garder à l'esprit qu'une technologie n'est pas forcément bonne en soi. Elle est bonne parce qu'elle est répond de manière adaptée à une problématique donnée. Les besoins réels de l'utilisateur final doivent rester au centre des préoccupations des services informatiques, au risque de dégrader encore les taux de succès déjà peu glorieux des projets décisionnels.



(Article publié le 15 mai 2014, révisé et enrichi le 20 mai 2014)

Références

Dan Ariely : "Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."

Un excellent article signé Yann Gourvennec sur la volumétrie des données

Un autre très bon article glané sur le site webgaga

Statistiques de succès des projets décisionnels en 2013