Le Big Data

Le big data, appelé littéralement « grosses données » ou encore données massives, désignent des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficile à travailler avec des outils classiques de gestion de base de données ou de l'information. Pour pouvoir travailler avec ce nouvel ordre de grandeur, la capture, le stockage, la recherche, le partage, l'analyse et la visualisation de toutes ces données doivent être revues. Les perspectives du big data sont énorme et encore insoupçonnées.

En effet, chaque jour, nous générons près de 2,5 trillions d'octets de données. A tel point que 90 % des données dans le monde ont été créées au cours des deux dernières années. Ces données proviennent de partout : de capteur utilisées dans les différents domaines de la recherche : astronomique, climatique, médicale,… mais également des messages sur des média sociaux notamment sur Facebook ou Twitter, et l'ensemble des fichiers numériques, images , vidéos , publiées en ligne. Il y a bien évidemment beaucoup d'autres sources de données, et c'est ce qui explique la masse de données produite chaque jour.

La notion du Big Data est donc une combinaison de progrès technologiques , d'innovations d'usage et d'évolutions sociales qui amène les entreprises à repenser leurs priorités. La première dimension fondamentale du Big Data c'est la composante technologique. En effet, le Big Data s'appuie sur un ensemble d'innovations qui transforment la façon dont les entreprises et les individus génèrent, transmettent, stockent et utilisent des données. Cela passe par la massification des échanges de données, la révolution des systèmes de stockage (cloud) et la structuration des données, progrès des techniques d'analyse,…
Cependant la montée en puissance du Big Data n'est pas essentiellement dû à la technologie. Il y a les évolution culturelles vis-à-vis de la génération et du partage d'information et les nouveaux usages et nouvelles possibilités de monétisation qui sont des éléments clés de l'augmentation du volume de données.

Le Big Data ce définis par 3 dimensions également appelés les 3V :

- Le Volume des données stockés aujourd'hui est en pleine expansion les données numériques créées dans le monde seraient passées de 1,8 zettaoctets en 2011 à 2,8 zettaoctets en 2012 et d'après des estimations ils s'élèveront à près de 40 zettaoctets en 2020.
- La Variété des données est un réel défi. Il ne s'agit généralement pas de données relationnelles traditionnelles, ces données sont généralement brutes et non structurés. Les systèmes doivent donc être capable d'analyser toutes ces données diverses et variées.
- La Vélocité représente à la fois la fréquence à laquelle les données sont générées, capturées et partagées. Les systèmes doivent être capable de répondre aux attentes des utilisateurs, par exemple les systèmes mis en place pour la bourse et les entreprises doivent être capables de traiter ces données avant qu'un nouveau cycle de génération ne commence, avec un risque de perte d'une partie des données. Les systèmes doivent être plus rapide pour permettre d'avoir toutes les informations nécessaire à temps et surtout ne pas avoir de pertes de données.

<u>Ces domaines d'application :</u>
Le Big Data trouve une application dans de nombreux domaines, dans de grand programmes scientifiques, dans des entreprises, dans des laboratoires,…

Les méthodes actuelles de modélisation de données ainsi que les systèmes de gestion de base de données ont été conçus pour des volumes de données très inférieur. Il ne s'agit pas ici de faire des requêtes mais plutôt une fouille de données. Dans le futur il faudrait des modélisation et des langages de requêtes permettant :

- une représentation des données en accord avec les besoins de plusieurs disciplines scientifiques ;
- de décrire des aspects spécifiques à une discipline (modèles de métadonnées) ;
- de représenter la provenance des données ;
- de représenter des informations contextuelles sur la donnée ;
- de représenter et supporter l'incertitude ;
- de représenter la qualité de la donnée

Les bases de données relationnelles classiques ne permettant pas de gérer les volumes de données du Big Data, de nouveaux modèles de représentation permettent de garantir les performances souhaités. Ces technologies dites de Business Analytics & Optimization (BAO) permettent de gérer des bases massivement parallèles. Des patrons d'architecture sont proposés par les acteurs du marché du Big Data comme MapReduce développé par Google et utilisé dans le framework Hadoop. Ce système permet de séparer les requêtes et de les distribuées à des nœuds parallélisés, puis les exécutées en parallèles. Les résultats sont ensuite rassemblés et récupérés.