Avis d'expert : « Une mauvaise gestion des données peut vous conduire au chaos »

Où que l'on regarde ces derniers temps (Twitter, blogs sur les technologies, newsletters reçues par e-mail, etc.), on ne peut échapper au Big Data. L'expression fait le buzz au point que les professionnels et férus de technologie ne peuvent y couper. Les experts en bases de données ont l'habitude de traiter de grandes quantités de données depuis des années. Mais les kilo-octets ont laissé place aux méga-octets, puis aux giga-octets, et maintenant nous en sommes aux téra-octets de données.

Tant que ces volumes de données étaient encore « gérables », différentes méthodes nous permettaient de les traiter : partitionnement de tables, archivage avec purges régulières, et création de datawarehouses situés à distance des bases de données transactionnelles régulièrement utilisées. Nous avions le temps d'analyser les flux qui venaient alimenter nos bases de données pour réfléchir au moyen de les transformer en renseignements utiles.

Cette dernière vague, celle du « Big Data », nous oblige à abandonner certaines de ces approches pour plusieurs raisons, la vitesse et le volume en particulier.

A présent, les données nous arrivent en trop grandes quantités et avec une telle vitesse que nos systèmes peinent à suivre. Nous voici dans le monde « merveilleux » des données non structurées. Dans ce monde, peu importe la nature ou le format des données, nous nous contentons de les stocker. Un jour ou l'autre, nous en ferons quelque chose ! Cette approche est-elle réaliste ?

En tant que professionnel des bases de données, j'aspire à une certaine qualité des données. Si vous injectez des données non structurées dans mon monde, ma capacité à assurer la qualité des données en prend un sacré coup. Je peux toujours les stocker. Je peux éventuellement en interroger une bonne partie et en extraire des informations utiles. Mais, avec le temps, ces données seront de plus en plus difficiles à gérer.

Imaginons que je crée un tableau de bord à partir des fichiers qui répertorient le nombre de fois où nos clients se sont identifiés sur nos sites ces deux dernières années. Grâce à ce tableau de bord, je pourrais connaître la fréquence de visite de chacune de nos pages Web. Est-ce que je conserve le détail des informations d'analyse au cas où ? Dans l'espoir d'en déduire de nouvelles connaissances business ? Si oui, est-ce que je réintègre mes nouvelles connaissances business aux lignes de données non structurées pour reprendre l'exploration ? Dans certaines entreprises, c'est impossible ! L'archivage est peut-être la seule vraie option. En effet, pendant que j'analyserai ces stocks de données non structurées, mes clients vont continuer à produire rapidement des sommes de nouvelles données, dont il faudra que je fasse quelque chose un jour ou l'autre.

Des éditeurs et des solutions

- Il faut reconnaître que les éditeurs commencent à commercialiser des solutions aux problèmes que posent ces données. Les récentes technologies de base de données Open Source, comme

NoSQL et CouchDB (dérivé de NoSQL), sont des solutions à base de documents. Le système de fichiers Hadoop File System (HFS) est une solution de stockage de fichiers, simple d'accès, en théorie, et conçue pour stocker de gros volumes de données en vrac. Les développeurs complètent ces systèmes HFS avec des interfaces de type SQL, comme Hive, pour faciliter l'accès aux données par ceux qui maîtrisent le SQL.

Cependant, de nouvelles questions se posent : si ces données sont aussi non structurées qu'on le dit, comment savoir ce qu'il faut rechercher ? Si les données de multiples sources sont simplement déversées dans un système de fichiers ouvert, comment en extraire des informations exploitables ?

Une question d'expertise, de RDBMS et d'ETL

C'est là que les experts des bases de données reprennent du service. Et c'est aussi ce qui m'amène à penser que la fin du système de gestion de base de données relationnelle (RDBMS) n'est pas pour tout de suite. Il faudra toujours programmer des techniques ETL, d'extraction, de transformation et de chargement de ces énormes sources de données non structurées pour préparer ces données et leur donner une forme lisible et exploitable. Il faudra aussi les associer à des entités valides (par exemple utilisateurs ou clients) ou à des ressources physiques (par exemple serveurs et/ou datacenters). Si on ignore à quoi renvoie telle ou telle partie de données non structurées, il est difficile, voire impossible, d'en extraire une quelconque valeur.

N'oublions pas non plus que des acteurs des RDBMS (Relational Database Management System), ajoutent à leurs systèmes des fonctions d'analyse du Big Data qu'ils développent eux-mêmes ; ou ils intègrent à leurs produits phares les outils d'entreprises qu'ils rachètent. Comme par exemple l'intégration de moteurs qui permettent d'exécuter des requêtes sur les données non structurées et les données relationnelles ; ou bien une base de données NoSQL et une appliance Big Data configurées, prêtes à collecter les données de votre entreprise. Toutes ces approches sont bonnes et pertinentes… mais sans moyen d'ordonner les données, on ne sort pas du chaos.

Comment s'en sortir ?

Pour sortir de cette impasse, il faut commencer par adopter une approche systématique vis-à-vis des données que vous collectez. Interrogez-vous ensuite sur la valeur de ces données pour votre entreprise : chaque donnée est liée à une ou plusieurs divisions de votre organisation. Une fois que vous avez classé les données par catégorie, il vous reste à définir des règles de gouvernance. Vous ne pouvez pas conserver toutes ces données indéfiniment. Ce n'est pas réaliste et risque de vous conduire au chaos. Imposez-vous des règles de rétention, celles que vous déciderez d'appliquer ou celles que les réglementations vous concernant vous imposeront. Dans tous les cas, définissez des règles claires.

Il n'est pas toujours possible non plus de traiter les données non structurées en temps réel. Vous devez décider quelles données vous allez traiter et dans quel ordre. Ceci suppose aussi de définir des règles. Les données en lien direct avec les recettes de l'entreprise sont, bien entendu, plus importantes. Choisissez de les traiter avec ceux de vos équipements et applications les plus performants.

A ce stade, il est bon d'envisager plusieurs niveaux de stockage : des disques standard, bon marché, mêmes anciens et lents, peuvent suffire pour les données qui n'ont pas besoin d'être accessibles instantanément ; par contre, pour celles dont vous aurez besoin en temps réel ou quasi réel, préférez des disques SSD (solid state device). Les technologies les plus avancées permettent de mixer plusieurs niveaux dans des baies uniques et de laisser l'intelligence intégrée gérer le placement des données en fonction des règles que vous avez définies.

Voir aussi

Silicon.fr étend son site dédié à l'emploi IT