Vers l'échantillonnage d'un entrepôt de données

Raphaël Féraud, Fabrice Clérot

France Télécom R&D, avenue Pierre Marzin, 22307 Lannion
Contact : raphael.feraud@orange-ftgroup.com

Résumé. L'afflux de données sur les usages des produits et services nécessite des traitements lourds pour les transformer en information. Or la capacité à traiter les données ne peut pas suivre l'augmentation exponentielle des volumes stockés. Avec les technologies actuelles, un difficile compromis doit être trouvé entre le coût de mise en œuvre et la qualité de l'information produite. Nous proposons une approche basée sur l'échantillonnage d'un entrepôt de données pour déployer à moindre coût un système d'information décisionnel utilisant tout notre potentiel d'information. La brique technologique essentielle pour construire ce système repose sur un opérateur d'échantillonnage des jointures.

1 Introduction

Une tendance lourde depuis la fin du siècle dernier est l'augmentation exponentielle du volume des données stockées. Les progrès de ces capacités de stockage ne se traduisent pas nécessairement par une meilleure compréhension de l'environnement. En effet, les données doivent être analysées afin de les transformer en connaissance or la capacité à traiter ces données n'a pas suivi cette augmentation exponentielle.

La première raison est simple : le processus d'analyse des données requiert une intervention humaine. Chaque augmentation de la couverture des domaines de données stockées induit une augmentation des équipes qui analysent ces données pour les transformer en information exploitable.

La seconde raison est liée à la capacité de traitement des données. Même si la puissance de calcul des ordinateurs suivait une pente équivalente à celle de la capacité de stockage (la loi de Moore), il existerait toujours des barrières algorithmiques bien plus infranchissables. Par exemple, pour un algorithme dont la complexité est en $O(n^2)$, le doublement d'une puissance de calcul ne permet d'augmenter que de $\sqrt{2}$ les possibilités de traitements.

Dès lors calculer des indicateurs pour alimenter un datamart devient problématique et demande une architecture technique coûteuse. De plus les requêtes ad hoc sont susceptibles de bloquer l'entrepôt de données. Elles sont en général prohibées ou fortement administrées de manière à prévenir l'engorgement du système d'information. Les techniques OLAP sont efficaces pour traiter des requêtes d'agrégation sur des cubes de données prédéfinis. Ces techniques manquent de souplesse et d'expressivité pour répondre à l'ensemble des requêtes ad hoc.

Avec nos technologies, il est difficile d'exploiter tout le potentiel de l'information contenu dans l'entrepôt.