Entrepôts de données et OLAP, analyse et décision dans l'entreprise

Jérôme Darmont (Université de Lyon, Lyon 2, ERIC EA 3083)

et Patrick Marcel (Université de Tours, LI EA 6300)

Les Big Data à découvert, CNRS Editions, 2017, pp. 132-133

D'après le cabinet Gartner, le marché de l'informatique décisionnelle (*business intelligence*) croît de 8,4 % par an et atteindra un chiffre d'affaire mondial de 27 milliards de dollars en 2019. L'informatique décisionnelle vise à améliorer la décision en entreprise sur la base de faits établis et offre à des décideurs non-informaticiens une vision transversale des informations stratégiques de l'entreprise. Elle peut par exemple répondre à la question : « Quelle entreprise du CAC 40 connaît la meilleure évolution de son cours de bourse ? »

Les problèmes qui se posent dans ce contexte sont le volume souvent important des données ; leur hétérogénéité lorsqu'elles sont issues de différentes sources (cf. III.9), leur qualité (cf. II.8) et leur niveau de détail, trop fin dans les systèmes d'information opérationnels pour donner une vision globale de l'activité.

Ces problèmes sont traités grâce aux entrepôts de données (*data warehouses*), *via* un processus en trois phases (figure 1). Premièrement, il s'agit d'extraire automatiquement les données des sources, de les transformer pour les uniformiser (conversion à la même unité de mesure, par exemple) et de les charger dans l'entrepôt (*extract, transform, load* ou ETL). Deuxièmement, le stockage des données nettoyées est assuré par l'entrepôt, qui est parfois subdivisé en plus petites unités métier ou magasins de données (*datamarts*). Finalement, l'entrepôt est exploité à l'aide de rapports et de tableaux de bord, en fouillant les données entreposées (*data mining*) ou grâce à l'analyse en ligne (*online analytical processing* ou OLAP).

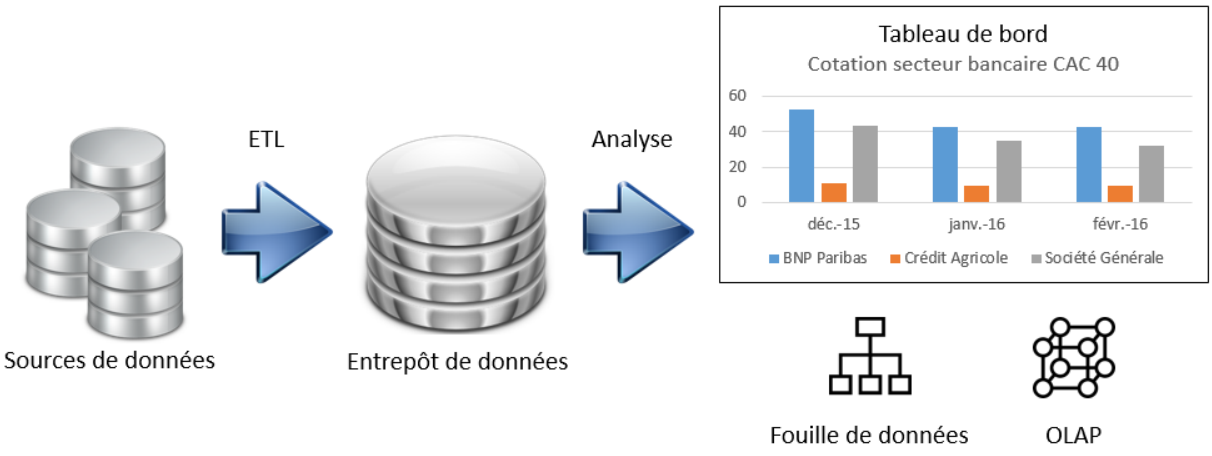

Figure 1 : Processus d'entreposage et d'analyse de données

OLAP *vs.* OLTP

Dans les systèmes opérationnels ou transactionnels (*OnLine Transaction Processing* ou OLTP) des entreprises, les données représentent l'état d'une information à un moment précis, par exemple la valeur des cours de bourse des entreprises du CAC 40. Ces données sont habituellement l'objet de nombreuses modifications (les cours de bourse évoluent rapidement) et les requêtes d'interrogation qui les ciblent sont relativement simples (afficher le cours de bourse d'une entreprise donnée). De plus, dans les systèmes OLTP, les schémas qui décrivent les données sont fortement normalisés, de manière à garantir la fiabilité des données.

Au contraire, dans le contexte OLAP, les données proviennent de différents systèmes OLTP et sont consolidées et historisées afin, par exemple, d'analyser l'évolution d'un cours de bourse. Les données décisionnelles sont donc rarement modifiées et au contraire enrichies au cours du temps (il est possible de stocker les cours de bourse successifs seconde par seconde). L'OLAP repose sur des requêtes d'interrogation complexes, comme les opérations effectuées avec un tableur (des moyennes de cours de bourse par mois, trimestre et année, par exemple), mais sur de gros volumes de données. C'est pourquoi les entrepôts présentent des schémas de description des données dénormalisés (ou combinant plusieurs tables), qui visent à optimiser le temps de réponse.

Cube de données et opérateurs OLAP

L'OLAP repose sur une représentation des données proche des intuitions de l'analyste, permettant l'expression aisée de requêtes complexes. La métaphore du cube (ou hypercube) représente les données à analyser (faits associés à des indicateurs numériques ou mesures) comme des points dans un espace multidimensionnel. Chaque dimension est un axe d'analyse et peut être organisée en hiérarchie. Les dimensions représentent les coordonnées du fait dans l'espace multidimensionnel. La figure 2 illustre un cube de donnée pour l'analyse de transactions boursières par place financière, type d'activité et date. Le fait de la cellule supérieure gauche indique que la mesure « volume d'actions » EDF échangé à Paris est 3,2 millions. La hiérarchie de la dimension date (en rouge sur la figure 2) permet par exemple de grouper les données quotidiennes pour les observer par mois ou par années.

L'expression d'analyses se fait *via* une succession d'opérations OLAP afin de comprendre un phénomène observé dans les données. Les opérations les plus populaires sont le *slice-and-dice* pour sélectionner des données, le *roll-up* pour les résumer (somme ou moyenne, par exemple), le *drill-down* pour les détailler, et le changement de la mesure étudiée. Par exemple, dans la figure 2, l'analyste applique un *slice-and-dice* pour se focaliser sur le secteur de l'énergie à Paris et Londres du 4 au 7 janvier 2016. Ensuite, il opère un *roll-up* pour avoir le total des volumes en Europe.

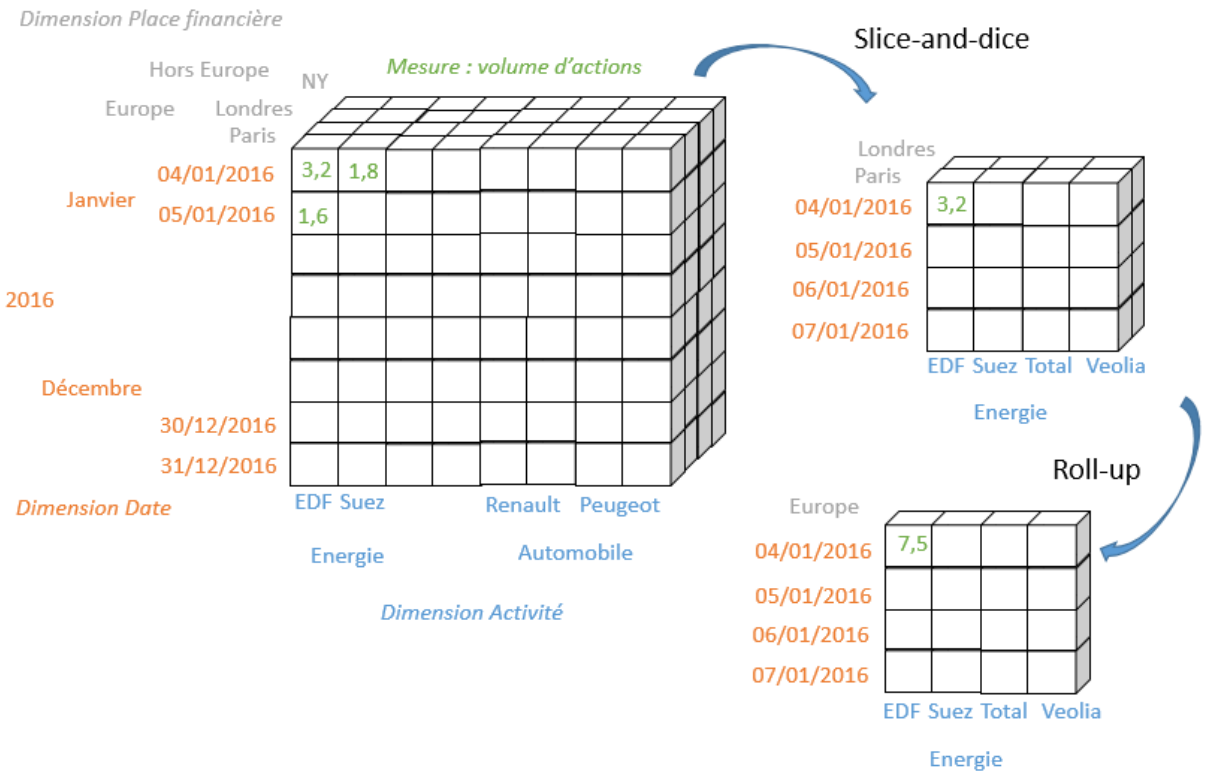

Figure 2 : Analyse d'un cube de données par Slice-and-dice et Rollup

Une source d'évolutions technologiques

Le succès de la technologie OLAP s'explique par l'évolution des dispositifs d'exécution et d'optimisation, en termes d'organisation logique et de stockage des données, pour le traitement efficace des requêtes. Dans les systèmes OLTP classiques, les données sont habituellement stockées sous forme de tables (cf. III.4), auxquelles un système d'index permet d'accéder rapidement. Dans les systèmes OLAP, les données résumées sont généralement pré-calculées pour accélérer les traitements lors de l'analyse. Des systèmes d'indexation propres aux données multidimensionnelles permettent également un accès rapide à un fait connaissant ses coordonnées. De plus, les données peuvent être partitionnées pour diminuer les volumes accédés lors des interrogations. Enfin, les faits peuvent être stockés sous forme de tableaux multidimensionnels (plus proches de la vue conceptuelle des données), notamment pour les données de faible dimensionnalité ; ou en colonnes, afin de regrouper physiquement les données de même nature.

L'OLAP a suscité de nombreux travaux de recherche et d'innovations menant à ces évolutions. Les principaux éditeurs de bases de données commerciales et le monde du logiciel libre offrent aujourd'hui des suites logicielles pour déployer entrepôts de données et outils OLAP. Actuellement, cette évolution se poursuit dans la mouvance des Big Data, en vue d'offrir au plus grand nombre l'analyse de données en temps réel, *via* des dispositifs mobiles ou à la demande (science des données).

Bibliographie

C. FAVRE, F. BENTAYEB, O. BOUSSAID, J. DARMONT, G. GAVIN, N. HARBI, N. KABACHI et S. LOUDCHER – Les entrepôts de données pour les nuls… ou pas ! *2e Atelier aide à la Décision à tous les Etages (AIDE), Toulouse*, 2013. *https://hal.archives-ouvertes.fr/hal-01273589*

R. KIMBALL et M. ROSS – *Entrepôts de données. Guide pratique de modélisation dimensionnelle (2e édition)*, Vuibert, 2003. *http://www.vuibert.fr/ouvrage-9782711748112-entrepots-de-donnees.html*

Glossaire

ANALYSE EN LIGNE (ou OLAP). Ensemble d'opérations permettant la navigation interactive dans un cube de données.

CUBE DE DONNÉES (ou hypercube). Représentation de données décisionnelles dans un espace multidimensionnel dont les dimensions décrivent les faits observés.

ENTREPÔT DE DONNÉES. Base de données à vocation décisionnelle supportant efficacement l'analyse en ligne.

EXTRACTION, TRANSFORMATION, CHARGEMENT (ou ETL). Intégration de données hétérogènes issues de diverses sources variées dans un entrepôt de données.

MAGASIN DE DONNÉES. Entrepôt de données spécifique à un métier de l'entreprise.