

OLAP, les fondamentaux

par Yazid Grim ([Business Intelligen\(ce\)](#))

Date de publication : 07-05-2008

Dernière mise à jour : 09-05-2008

Suite à différentes discussions divergentes sur la notion de cube, de OLAP, de M-OLAP, etc., j'ai décidé de rédiger un article pour éclaircir ces termes et donner une vision consensuelle et universelle sur ces termes qui, aujourd'hui, veulent tout et rien dire. Bien que n'ayant pas la prétention de mettre tout le monde d'accord, ce qui est assez dur vu que la B.I est un domaine relativement jeune, cet article reviendra sur le pourquoi et le comment du OLAP, sa différence avec ce que l'on nomme OLTP, ses forces et ses faiblesses.

- I - Introduction : comment l'industrie du B.I brouille les pistes
- II - Là où OLAP est né : l'analyse
 - Un exemple
 - La solution : de l'analyse on-line
- III - Démystifions le OLAP
 - OLAP, Quézako ?
 - Ce que n'est pas OLAP
- IV - M-OLAP, R-OLAP, H-OLAP, D-OLAP ...
 - Définitions
 - R-OLAP
 - M-OLAP
 - H-OLAP
 - D-OLAP
- V - Conclusion
- VI - Remerciements

I - Introduction : comment l'industrie du B.I brouille les pistes

De nos jours, les définitions en BI fument selon l'inspiration ou la vision de celui qui les écrit. On trouvera des commerciaux qui veulent placer les mots OLAP, Data Warehouse parce que ça fait hi-tech. Des autodidactes qui, de par leur compréhension des choses, définissent des concepts " à leur manière ", avec tous les risques que cela comporte. Des entreprises qui " décident " que leurs outils de rapport sont des outils OLAP car ils permettent de faire du forage dans des tableaux Excel. Et enfin des leaders qui tentent d'imposer leur façon de voir les choses.

Il est important de clarifier ces termes et de les consensualiser pour les raisons suivantes :

- Parler de la même chose : il serait très gênant que vous parliez de OLAP alors que votre interlocuteur pense que ce sont des rapports dynamiques tirés d'une base de données. Les projets n'ont pas la même taille ...
- Apprécier les définitions : les définitions des termes de BI étant, en général, beaucoup plus profondes et complètes que leurs implémentations par les Microsoft, Oracle et compagnie. On a tendance à oublier qu'on adapte l'outil à la méthode et pas le contraire ...
- Maturité : imaginez si chacun se mettait à faire des petits dessins et appeler cela UML, ou si tout le monde se mettait à faire des bases de données comme ça l'arrange et appelait cela des schémas en troisième forme normale. La maturité d'un domaine passe forcément par une reconnaissance unanime d'un concept et de sa définition.

L'industrie de la BI est jeune, peut-être immature, mais c'est aux acteurs de cette industrie (producteurs et utilisateurs) de faire en sorte que cette industrie évolue dans le bon sens et dans la même direction surtout, l'union fait la force ;)

II - Là où OLAP est né : l'analyse

Pour mieux comprendre OLAP et les technologies gravitant autour, intéressons nous à la genèse de ce concept : l'analyse en entreprise.

L'analyse est un processus intellectuel qui, à partir d'hypothèses et de données, permet à une personne de générer de la connaissance. Cette connaissance peut se formaliser par l'explication d'un phénomène, la proposition d'une solution pour optimiser les ventes, des recommandations quant à la politique d'approvisionnement, etc. Bref, tout ce qui peut améliorer l'entreprise.

De part sa nature, l'analyse n'est pas un processus que l'on peut guider ou formater. La recherche de la cause de la baisse des profits, par exemple, peut impliquer des recherches du côté des succursales, des ventes, des prix des fournisseurs, des taxes locales, etc. Le travail de l'analyste consiste, à travers une série de question-réponse, à expliquer la raison d'un phénomène par les données qu'il possède, et ce processus n'est presque jamais linéaire.

Un exemple

Un exemple illustrera mieux le fond de ma pensée.

Imaginez que l'on demande à un analyste d'expliquer le fait que les profits de l'entreprise aient baissé durant les trois derniers mois. Voici un des cheminements que pourrait suivre un analyste :


- Voir les profits sur l'année pour constater la baisse.
- Voir le volume des ventes sur les trois derniers mois : les ventes n'ont pas baissé.
- Voir les coûts de revient des produits sur les trois derniers mois : les coûts ont sensiblement augmentés pour certains pays.
- Voir les coûts de revient des produits par zone de production sur les trois derniers mois : les produits ayant subi une augmentation sont tous produits en Asie.
- Pour les produits en question, pour les trois derniers mois, voir le coût de production par usine, ainsi que le coût moyen de la main d'oeuvre et la taxe et comparer ces valeurs avec celles des trois mois avant l'augmentation : les coûts sont à peu près les mêmes, la raison n'est pas là.
- Pour les produits en question, comparer le coût de la matière première avec les chiffres des trois mois avant la hausse des coûts : BINGO ! Le prix de la matière s'est envolé ces trois derniers mois. Il faut maintenant voir les raisons de cette augmentation, revoir les contrats avec les fournisseurs, réguler la production ... Il faut agir !

Observez bien ce processus. Nous avons commencé par un simple tableau des profits sur l'année et nous avons fini par une synthèse des prix des matières premières pour une liste de produits fabriqués en Asie, et pour les trois derniers mois ... Est-ce que des rapports peuvent supporter un tel processus ? Difficilement, les rapports sont parfaits pour distiller une information précise, structurée. Mais supporter un tel processus cognitif impliquerait trop de travail (solicitation perpétuelle du département informatique pour la génération de rapports, perte de temps et d'efficacité d'aller d'un rapport à un autre).

Remarquez aussi les différents niveaux d'agrégation par lesquels notre analyste est passé. L'analyse a commencé par un cumul annuel, et s'est vite transportée vers des cumuls par usine sur une période donnée, sur une zone géographique donnée. Cette charge de travail serait très difficilement supportable par une base de données de production OLTP classique (surtout quand elle implique plusieurs Go de données). Donc la solution impliquerait d'utiliser une structure orienté analyse à l'inverse des bases OLTP qui sont orientées production.

La solution : de l'analyse on-line

Après ce petit exemple, posons nous la question suivante : quelles sont les caractéristiques de la technologie utilisée pour suivre le processus cognitif de l'analyste ? La réponse se trouve dans un document qu'a écrit monsieur CODD à ce sujet et qui se nomme " *Providing On-Line Analytical Processing to Users Analysts* ". Dans ce document, monsieur CODD énumérait douze règles qui permettaient à une technologie de faire de l'analyse sur les données. Six règles furent ajoutées deux ans après la publication de ce document.

Ces règles, disponibles sur le Web, énoncent les caractéristiques qu'un outil doit avoir pour pouvoir se qualifier d'outil OLAP. On peut facilement voir que la plupart des outils proposés sur le marché se définissant comme tel ne respectent pas toutes les règles. Deux ans après l'établissement de ces règles, le *OLAP Council*, un organisme de normalisation qui a tenté d'unifier les définitions des concepts du OLAP, a publié un document avec une définition officielle du terme OLAP, en voici un extrait traduit par moi-même (pour la version originale,  [c'est ici que ça se passe](#)) :

"...Plus récemment, les fabricants de bases de données relationnelles se sont mis à vendre leurs bases de données comme étant des outils pour construire des entrepôts de données. Un entrepôt de données garde les informations tactiques (stratégiques) qui répondent aux questions du type "qui ?" et "quoi" à propos d'événements passés. Une requête type utilisée dans un entrepôt de données serait : "Quel a été le revenu total pour la région Est dans le troisième trimestre?"


Il est important de distinguer les capacités d'un entrepôt de données de celles d'un système OLAP (On-Line Analytical Processing). À l'inverse des entrepôts de données, qui se basent sur une technologie relationnelle, OLAP utilise une vue multidimensionnelle de données agrégées pour fournir un accès rapide à l'information stratégique pour des analyses plus poussées.

OLAP permet aux analystes, gestionnaires et au personnel exécutif de mieux comprendre les données via un accès rapide, consistant et interactif à une large variété de vues possibles de l'information. OLAP transforme les données du plus bas niveau afin de montrer leur véritable dimension dans l'entreprise, selon la compréhension de l'utilisateur.

Les systèmes OLAP permettent non seulement de répondre aux questions de type "qui" et "quoi", mais permettent aussi de répondre aux "que se passe t il si #?(what-if)" et aux "pourquoi?". Chose qui les distingue des entrepôts de données. OLAP permet la prise de décision sur des actions futures. Un calcul OLAP typique est plus complexe qu'une simple somme sur les données, par exemple : "Que serait l'effet sur les coûts des distributeurs de soft Drink si le prix du sirop augmentait de 0.10\$ le gallon et que le coût de transport baissait de 0.05\$/mile?"

OLAP et les entrepôts de données sont complémentaires. Un entrepôt de données stocke et gère les données. OLAP transforme les données de l'entrepôt en informations stratégiques. OLAP peut passer d'une navigation basique (connue aussi comme "slice and dice"), à des calculs ou des analyses plus sérieuses comme les séries temporelles ou la modélisation complexe. Ainsi, les décideurs expérimenteront les capacités avancées de OLAP, et pourront passer d'accès aux données à information, à connaissance. "

En résumé, cet organisme définit OLAP comme étant l'ensemble des technologies qui, se basant sur une représentation multi-dimensionnelle des données, permet aux analystes et décideurs de traiter leurs données de façon analytique, interactive (sessions), rapide et permettant de voir les données de l'entreprise sous plusieurs angles (dimensions).

Plus récemment, *The Olap Report*, un autre organisme normalisateur sur les notions OLAP a tenté de proposer une définition plus simple et digeste que celle du *OLAP Council* (qu'ils jugent has-been et, je cite, défunte). Cette définition tourne autour d'un acronyme : FASMI.  [L'article complet est disponible ici.](#)

Cet organisme définit les outils OLAP comme étant des outils basés sur une vue multidimensionnelle des données (donc une conception à base de dimensions et de faits), l'organisme précise que cette vue est totalement indépendante de la technologie utilisée pour le stockage. Pour peu que l'outil réponde aux critères qui suivent.

Un outil OLAP, selon The Olap Report doit être rapide (Fast), doit permettre de faire des analyses complexes (Analysis), répondre à une architecture Client/Serveur avec tout ce que cela implique en terme de sécurité et de gestion d'accès concurrent (Server). Un outil OLAP doit, comme dit plus haut, se baser sur une vue multidimensionnelle des données (Multidimensional) et finalement le volume d'information que peut prendre en charge ses outils (Information). Ces critères ont été simplifiés par l'acronyme FASMI.

Cette définition, plus simpliste que la première, n'entre pas en désaccord avec la précédente, ni avec la vision de CODD sur ce qu'est un outil OLAP. Je regrette cependant cette tentative de création d'une autre définition censée être universelle mais qui n'ajoute, en fin de compte, qu'une autre définition aux innombrables tentatives existantes sur le Web. Cette démarche a au moins le mérite d'être indépendante de tout producteur de solution B.I, chose qui n'est pas, selon The Olap Report, le cas de la définition de CODD.

III - Démystifions le OLAP


Nous allons essayer d'éliminer les légendes urbaines du B.I dans cette partie, en définissant strictement le mot OLAP et en disant " ce que n'est pas OLAP ".

OLAP, Quézako ?

Si l'on se réfère au chapitre précédent, ainsi qu'aux différents documents (non commerciaux) traitant de ce sujet. Nous nous apercevrons que le mot OLAP, acronyme de On-Line Analytical Processing, désigne l'ensemble des technologies permettant la prise de décision stratégique rapide et fiable sur des données modélisées en multi dimensionnel. Ces technologies, pour mériter le label OLAP, doivent obéir aux règles de CODD :

- Conception et vue multidimensionnelle : un outil OLAP doit se baser sur un modèle multidimensionnel pour faire de l'analyse.
- Transparence : la technologie utilisée, la conception ainsi que toutes les spécifications techniques doivent être invisibles à l'utilisateur final.
- Accessibilité : les outils OLAP doivent permettre d'accéder les données de façon à produire de la connaissance rapide. Une information pertinente et on-time doit être fournie en tout temps.
- Rapidité : les montées de charges ne doivent pas freiner l'analyste. L'outil doit pouvoir supporter de grosses requêtes (c'est la caractéristique la plus difficile à satisfaire).
- Architecture Client Serveur : pour un accès uniforme et des traitements plus rapides et plus sophistiqués.
- Dimensions génériques : essayer, autant que possible, d'avoir une unicité dans la définition des dimensions. Ne pas avoir deux dimensions client.
- Gestion des matrices creuses : en mathématiques, les matrices creuses sont des matrices qui contiennent beaucoup de zéros. En informatique, il existe des algorithmes qui utilisent cette spécificité pour optimiser le stockage de ce type de matrices. Les performances sont, en général, au rendez vous. Les outils OLAP doivent avoir cette capacité d'optimisation d'espace de stockage par la gestion des matrices creuses.
- Multi-utilisateurs : les outils OLAP sont, par définition, destinés à un accès concurrent.
- Croisement inter dimensions illimité : l'utilisateur ne doit avoir aucune restriction quand au nombre de croisements qu'il fait entre les dimensions.
- Intuitifs : les utilisateurs d'outils OLAP ne sont pas forcément informaticiens. Il est donc nécessaire d'offrir des solutions adaptées à leur style cognitif.
- Affichage flexible : l'utilisateur doit pouvoir aisément " arranger " son résultat au format désiré.
- Nombre illimité de dimensions et de niveaux d'agrégation.

En plus de ces règles, six autres conditions se sont ajoutées. Ces conditions traitent essentiellement de la navigation dans les solutions OLAP (drill-down, drill-through, etc.), de l'interopérabilité avec les solutions de l'entreprise et de la gestion des cas spéciaux (valeur manquante, données anormales, etc.).

 *Note : selon The Olap Report, la définition de monsieur CODD est souvent mise en doute et critiquée de par le fait que ce document, selon ses détracteurs, tient plus de la promotion commerciale que du véritable document de recherche. Il faut savoir que l'écriture de ces règles s'est fait pendant que monsieur CODD était en emploi chez Arbor Software, plus connue aujourd'hui comme Hyperion Solutions. Le conflit d'intérêt semble flagrant, d'autant plus que, selon The Olap Report, la majorité du travail pour l'établissement de ces règles fut émis par un chercheur temporaire chez Arbor ainsi que la compagnie de CODD, et que le document en question avait pour but, implicite, de promouvoir les produits de la compagnie. Mais n'entrons pas dans ces " peopleries " sans but ;)*

Ce que n'est pas OLAP

Je vais essayer de compiler la plupart des choses que j'ai entendu ici et là sur ce qu'est OLAP. Il faut noter que ces erreurs sont principalement dues aux messages commerciaux et à la " vulgarisation " de la BI dans les entreprises. Ça paraît tellement simple qu'on a tendance à croire très rapidement qu'on a tout compris# Je commence !

OLAP c'est un entrepôt de données : beaucoup de décideurs pensent cela. OLAP est une technologie d'analyse basée sur des données multidimensionnelles. Les DW sont des structures de données qui historisent les données de toute l'entreprise pour des fins d'analyse. C'est vrai que de nos jours l'un va toujours avec l'autre. Il est quasi improbable d'avoir un entrepôt de données sans outil d'analyse OLAP. Mais l'un n'est pas l'autre, ils sont complémentaires.

OLAP c'est des bases de données multidimensionnelles : l'analogie serait de dire qu'un SGBD est un ensemble de bases de données. Tout le monde sait qu'un SGBD c'est les bases de données avec tout l'aspect gestion, maintenance, optimisation et contrôle. Et bien pour OLAP c'est la même chose : les technologies OLAP se basent sur les bases de données multi dimensionnelles (conçues en dimensions et en faits) pour proposer des techniques d'analyse stratégique. Il est réducteur de dire que OLAP c'est des bases de données Multi dimensionnelles.

Excel fait de l'OLAP : cette délicate affirmation prêtera toujours à débat. Certains diront que ce n'est qu'un tableur avec des fonctions statistiques évoluées. Et d'autres diront que c'est un outil OLAP complet. Je dirais personnellement qu'à partir de la version 2003, avec la possibilité de se connecter à des cubes Analysis Services, Excel est devenu un outil qui satisfait la plupart des règles de CODD. La plus fondamentale étant la nécessité de se baser sur des données multi dimensionnelles. Le seul bémol est le volume de données limité que peut traiter Excel. En effet, quelques dizaines de milliers de lignes suffisent à le mettre à genoux# Je n'ai pas encore testé les version 2007 et 2007 Server, mais je suis prêt à parier qu'ils feront un effort de ce côté :)

Un rapport avec Drill-Down c'est du OLAP : FAUX ! Un rapport c'est un rapport. Un rapport ne peut pas supporter le processus cognitif d'une analyse. Un rapport peut ne pas se baser sur des données multi dimensionnelles, et un rapport ne permet pas de basculements de dimensions, d'inversions de colonnes, de changement de vues sur les données. Il faut vraiment faire une séparation entre l'analyse et le Reporting. On peut faire du Reporting avec des outils OLAP, mais pas le contraire.

Un cube OLAP ??? : Ces deux mots viennent souvent naturellement les uns après les autres. Vous connaissez maintenant la signification du mot OLAP, mais qu'en est il du mot " cube ". Cube est un terme gentillet pour désigner des bases de données multi dimensionnelles ou hypercubes. En effet, il était plus simple pour les décideurs de visualiser leurs données en trois dimensions (deux dimensions et un fait). C'est pour cela que cette image est apparue. Reste que les " cubes " sont rarement en trois dimensions, il n'est pas rare de voir des cubes avec 4 ou 5 dimensions en plus de la table de faits. Mais comme c'est moins visualisable par nos amis patrons, on utilise le mot cube pour dire base de données multi dimensionnelles.

IV - M-OLAP, R-OLAP, H-OLAP, D-OLAP ...

Résumons ce que nous avons vu jusqu'à maintenant :

- Nous savons d'où vient la nécessité du OLAP.
- Nous avons défini formellement le mot OLAP.
- Nous avons vu les abus de langages pour ce mot.
- Et finalement, nous avons vu ce qu'était un cube.

Intéressons nous à ces mots qui envahissent notre vocabulaire B.I : MOLAP, ROLAP, DOLAP, et HOLAP. Qu'est ce que c'est ? Quel est le lien avec OLAP ? Est-ce que ça reste du OLAP ?

Définitions

Ces termes ne sont ni plus ni moins que des implémentations du principe OLAP. L'analogie parfaite serait de comparer ces notions avec une interface et des classes implémentant cette interface en programmation orientée objets. L'interface donne les spécificités des classes. Et les classes implémentent ces spécificités de manières différentes. En gros l'interface dit quoi faire et les classes l'implémentant font ce que dit de faire l'interface chacune à sa manière.

OLAP est le fait de faire des analyses sur des bases de données multi dimensionnelles. X-OLAP définit la façon dont seront stockées physiquement les données pour permettre des analyses multi dimensionnelles.

R-OLAP stocke les données multi dimensionnelles dans un format relationnel (tables, relations), M-OLAP les stockent dans un format multi dimensionnel réel, H-OLAP utilise les deux méthodes pour le stockage, D-OLAP stocke les données en local pour l'analyse. Nous allons expliquer un peu plus ces technologies, mais il est important de comprendre que c'est du OLAP, la lettre avant désigne la spécificité technique qui permet de faire du OLAP.

R-OLAP

Relational OLAP. Comme son nom l'indique, il utilise le concept relationnel pour stocker des données modélisées dans le format multi dimensionnel. Les analyses (drill-down, pivot, ajout de dimensions, etc.) sont transformées en requêtes SQL classiques qui sont exécutées sur les tables. R-OLAP utilise aussi la notion de tables d'agrégats, c'est-à-dire créer des tables contenant des données sommaires et les stocker en mémoire en cas d'utilisation.

Les outils modernes permettent aussi la gestion du cache, l'optimisation des requêtes et la création de tables d'agrégats à la demande.

La technologie R-OLAP perd beaucoup de terrain face à ces concurrents (qui suivent) car elle implique beaucoup de lourdeur et d'émulation pour son implémentation. On simule des opérations sur des matrices avec du SQL, et le fait de simuler deux conceptions apparemment différentes apporte son lot de gestion lourde et de manque de performances.

R-OLAP reste la solution de choix dans le cas de gros volumes de données avec un accès restreint.

M-OLAP

Multi dimensional OLAP. Contrairement à R-OLAP, M-OLAP permet de stocker les données directement en un format permettant des opérations matricielles. Selon le constructeur, on trouvera un mode de stockage à base de tableau

de données, de technologies propriétaires et même à base de fichiers plats ! L'avantage de ce mode de stockage est la capacité à effectuer des calculs très poussés en un temps record vu que tous les calculs sont pré-compilés ! Le mode de stockage permet de pré calculer les résultats afin d'avoir accès directement à toute donnée, quel que soit le niveau de détail.

M-OLAP reste la meilleure solution du moment en terme de performances et d'efficacité. Reste que cette solution, à double tranchant, montre très rapidement ses limites quand on commence à jouer avec de gros volumes de données. En effet, le " pré-calcul " des résultats devient très pénible quand il s'agit de gros volumes de données.

H-OLAP

Hybrid Olap. C'est la solution " en vogue " du moment, car elle permet de minimiser les défaillances des technologies R-OLAP et M-OLAP. Il s'agit en fait d'un mix des deux solutions.

On utilisera un mode de stockage propriétaire pour les tables d'agrégat et les tables intermédiaires (permettant de ne pas avoir les points faibles du R-OLAP). On conservera un mode relationnel pour les tables de bas niveau.

D-OLAP

Desktop OLAP. Le cas spécial. Il ne s'agit en fait pas d'une technologie particulière mais plutôt d'un mode de fonctionnement.

D-OLAP permet à l'utilisateur d'enregistrer une partie de la base de données multi dimensionnelle en local. On voit très vite l'utilité d'une telle solution pour les commerciaux et les " nomades " de l'entreprise. Cela permettrait à un commercial, par exemple, de faire des analyses sur les ventes, conserver ses résultats, et vérifier l'évolution de ses analyses, une fois revenu de son voyage d'affaire.

V - Conclusion

Comme vous pouvez le voir, ce n'est pas si compliqué que cela de parler BI " proprement ". Les concepts présentés plus haut sont la base et la condition sine qua non que toute personne doit posséder pour bien apprendre les concepts du BI et ne pas se laisser bercer par les douces sérénades commerciales des vendeurs d'outils BI. N'oublions pas que leur but est de créer des clients fidèles, même si cela doit passer par de la désinformation ou même de la fausse information !

VI - Remerciements

Encore une fois un grand merci à l'équipe de développez.com pour son soutien et sa passion pour les TI. Mention spéciale pour Antoun sans qui ce document ne verrait pas le jour, Adrien Artero (El Padrino), Fleur-Anne.Blain et QI130 pour son oeil correcteur :)

