

Informatique décisionnelle

BI, Business Intelligence Système interprétant des données complexes permettant aux dirigeants d'entreprise de prendre des décisions en connaissance de cause. Les données sont analysées selon plusieurs dimensions (type de produits, régions et saisons par exemple). De plus en plus, l'informatique décisionnelle se rapproche de l'intelligence d'affaires, où un système informatique permet la recherche active et l'exploitation, sur le plan décisionnel, de l'ensemble des renseignements stratégiques essentiels qu'une entreprise doit posséder, si elle veut faire face à la concurrence et occuper la première place, dans son secteur d'activités.

L'architecture d'un système décisionnel peut être décomposée en un ensemble de composants. Cette décomposition permet de répartir les tâches entre différents responsables du système et de comprendre la synergie de leurs différents métiers.

1. Les sources de données (ou bases de production)

C'est l'ensemble des sources de données qui contiennent les informations qui serviront à alimenter un entrepôt de données. Ces sources sont dans la grande majorité internes à l'entreprise (capitalisées dans les bases de données opérationnelles), mais peuvent également être externes à l'entreprise (données du marché mondial, audits, documents diffusés sur le WEB, etc.).

2. Un ensemble d'outils ETL (ou ETL)

Les sources de données utilisées pour constituer un entrepôt de données sont hétérogènes et diffuses. Elles contiennent également des données qui ne seront pas utilisées par l'entrepôt de données. Il est donc nécessaire de disposer d'outils performants et rapides permettant d'extraire (**ETTL**) les données utiles de ces sources. Ces données extraites doivent être par la suite transformées (**ETTL**) pour les rendre globalement homogènes (standards). Si ces données sont réparties sur différents ordinateurs dans un réseau diffus, elles doivent initialement être transportées (**ETTL**) vers le (ou les) ordinateur gérant l'entrepôt de données. Pour finir ces données normalisées sont chargées (**ETTL**) dans l'entrepôt de données.

ETL

Extraction, Transformation and Loading

Outil d'extraction Outil informatique destiné à extraire des données de diverses sources (bases de données de production, fichiers, Internet, etc.), à les transformer et à les charger dans un entrepôt de données.

3.Un entrepôt de données

L'entrepôt de données correspond à une représentation multidimensionnelle de l'ensemble des données extraites des sources de données initiales, ainsi qu'à un ensemble de méta-données contenant les informations relatives à l'environnement de cet entrepôt. La partie essentielle de l'entrepôt est constituée d'une ou plusieurs bases de faits qui traduisent la vie d'une entreprise. Ces bases de faits sont reliées à la description de plusieurs dimensions intervenant dans la représentation multidimensionnelle. Un entrepôt peut être physiquement réparti dans plusieurs magasins métiers (datamarts) regroupés sur un même ordinateur ou sur plusieurs.

DATA MART(entrepot metier)

Sous-partie d'un entrepôts de données orienté vers un métier particulier de l'entreprise (marketing, finance, gestion de stock, etc.). Le datamart peut être interne ou élément satellite de l'entrepôt de données.

4.Des représentations OLAP de l'entrepôt de données

Ce sont des cubes de données multidimensionnelles extrait dynamiquement de l'entrepôts de données (ou des datamarts). Ils sont assujettis aux rapports, requêtes et analyses devant être fournis aux différents fournisseurs de l'entreprise.

5.Des outils de visualisation et d'analyse

Ensemble des outils permettant :
d'obtenir les différents indicateurs et rapports de synthèse sur la vie d'une entreprise, de réaliser de façon interactive à travers des interfaces graphiques utilisateurs des requêtes exploratoires dans les cubes dimensionnelles,
d'exploiter des techniques de data-mining sur ces données dimensionnelles.

Datamining

Exploration des données Technique d'analyse utilisant un logiciel (TANAGRA, SIPINA, R-PROJECT,...) pour dénicher des tendances ou des corrélations cachées parmi des masses de données, ou encore pour détecter des informations stratégiques ou découvrir de nouvelles connaissances, en s'appuyant sur des méthodes de traitement statistique.

OLAP

OnLine Analytical Processing

nous avons vu que la structure en hyper-cube des données était la façon la plus naturelle pour visualiser une entreprise comme un tout, car elle est fondée sur les structures réelles de l'activité d'une entreprise et non sur des constructions artificielles dictées par l'informatique. L'exploitation des données d'un entrepôt doit donc correspondre à la manipulation aisée de l'hyper-cube. C'est ce que propose

l'approche OLAP. L'objectif est de permettre aux décideurs de naviguer simplement dans les informations, via la manipulation des diverses dimensions, en vue de dégager plus rapidement les tendances intéressantes de l'entreprise. La figure ci-dessous, déjà utilisée dans la séquence 1, permet de situer OLAP dans l'architecture globale d'un système décisionnelle.

Technique d'analyse, élaborée en 1993 par E.F. Codd, un des créateurs des bases de données relationnelles, à la demande de la firme Arbor Software (devenue aujourd'hui Hyperion). L'objectif était de pouvoir sélectionner des données selon des critères multiples. Aujourd'hui, OLAP permet aux décideurs, en entreprise, d'avoir accès rapidement et de manière interactive à une information pertinente présentée sous des angles divers et multiples, selon leurs besoins particuliers. Très utilisés dans les secteurs de la banque, des télécommunications et de la grande distribution, les serveurs OLAP sont des outils opérationnels, qui permettent de valider une stratégie mise en oeuvre ou de vérifier des tendances. Ainsi, on pourra souhaiter examiner l'évolution des ventes d'un produit donné, dans une région géographique précise, au cours d'une saison donnée. Il suffira de préciser ces trois dimensions d'analyse au moteur OLAP. Les

valeurs trouvées dans la base pourront être représentées sous la forme d'un cube. Si l'on avait souhaité examiner plus de trois critères ou dimensions, on parlerait alors d'hypercube. Exemple : on mesurera l'évolution sur trois ans (axe 1) du chiffre d'affaires (axe 2) lié aux ventes d'une gamme de produits (axe 3) réalisées en direction d'un profil client particulier (axe 4) sur une zone géographique précise (axe 5).

- Technique d'analyse, élaborée en 1993 par E.F. Codd, un des créateurs des bases de données relationnelles
- Les bases de données relationnelles (SGBDR) ne sont pas adaptées aux traitements décisionnels
- L'objectif était de pouvoir sélectionner des données selon des critères multiples (multidimensionnelles).
- Aujourd'hui, OLAP permet aux décideurs, en entreprise, d'avoir accès rapidement et de manière interactive à une information pertinente présentée sous des angles divers et multiples, selon leurs besoins particuliers.
- OLAP désigne les bases de données multidimensionnelles (appelées aussi cubes ou hypercubes) destinées a des analyses complexes sur des données .
- **Les 12 règles de base**
- Pour répondre à cet objectif, E.F. Codd définit 12 règles de base permettant de qualifier le concept global nommé OLAP (On Line Analytical Processing).
- **Transparence** : le système doit être transparent pour l'utilisateur qui doit pouvoir accéder à la base par l'intermédiaire d'outils standards
- tels que des tableurs ou des formulaires HTML.
- **Accessibilité** : Les sources multiples servant à alimenter l'entrepôt doivent être facilement accessibles à travers la structure logique del'entrepôt.

- **Manipulation des données** : La navigation doit pouvoir s'effectuer simplement et intuitivement à travers des interfaces graphiques à forte
 - ergonomie.
- **Souplesse d'affichage** : le système doit pouvoir retourner les résultats des requêtes sous une forme graphique. Ces résultats doivent
 - pouvoir être manipulés et servir de base à d'autres requêtes.
- **Multidimensionalité** : le système doit permettre une manipulation multidimensionnelle de l'ensemble des données de l'entrepôt.
- **Dimensionnalité générique** : toutes les dimensions d'un hypercube doivent être accessibles à travers un même protocole.
- **Client/serveur** : l'approche doit respecter l'architecture client-serveur afin que plusieurs utilisateurs puissent manipuler l'entrepôt sans
 - difficulté.
- **Multi-utilisateur** : Des requêtes sur l'entrepôt de données doivent pouvoir être réalisées en simultanées par plusieurs utilisateurs.
- **Accès stable** : le nombre de dimensions et/ou le nombre de niveaux d'agrégation doivent pouvoir changer sans perturber le fonctionnement
 - de l'entrepôt.
- **Gestion des matrices creuses** : Les cellules " vides " d'un hypercube doivent être gérées de façon efficace afin de limiter les capacités de
 - stockage nécessaire et les temps d'accès aux données de l'hypercube.
- **Croisement des dimensions** : Les dimensions dans les différents entrepôts métiers (datamarts) doivent être accessibles et croisées.
- Toutes les règles de gestion doivent pouvoir s'appliquer à l'ensemble des tranches des hyper-cubes correspondant.
- **Nombre illimité de dimension et de niveaux d'agrégation** : il ne doit pas y avoir de limites imposées au nombre de dimension et de niveaux d'agrégation manipulable.

Selon le groupe Valoris, un système OLAP est un système d'analyse rapide d'information multidimensionnelle partagé.

Cette définition à l'avantage d'être concise et elle est en adéquation avec les 12 règles proposées par Codd. Elle insiste également sur le fait qu'un système OLAP doit être rapide.

Le groupe Valoris a également proposé les 5 caractéristiques principales d'un système OLAP de la façon suivante :

Multidimension :

C'est la caractéristique essentielle d'un tel système. Il doit fournir une vue multidimensionnelle des données, incluant le support des

hiérarchies simples et multiples des différentes dimensions.

Analyse :

Le système doit permettre tout type d'analyses statistiques sur les données multidimensionnelles sans que l'utilisateur ait besoin de connaissances avancées en programmation. La création de scripts d'analyse doit être la plus ergonomique possible (via interface graphique par exemple).

Rapidité :

Le système doit être conçu pour fournir une réponse en quelques secondes aux requêtes les plus courantes afin que les utilisateurs puissent réaliser une exploration et une analyse interactive des données de l'entrepôt. Pour les requêtes les plus complexes, le système doit pouvoir prévenir l'utilisateur de leurs durées importantes et si possible fournir une estimation.

Information :

Le système doit pouvoir manipuler des quantités d'information importantes. Il doit également capitaliser l'ensemble des manipulations qu'il réalise à travers la constitution de métadonnées. Ces capitalisations doivent permettre au système d'être plus performant par la suite.

Partage :

Le système doit permettre le partage des données multidimensionnelles à un grand nombre d'utilisateur. Toutes les conditions de sécurité et de confidentialité doivent être satisfaites, avec possibilité de descendre jusqu'au niveau de la cellule.

MOLAP :

- Les données détaillées de base ainsi que les données agrégées de l'entrepôt sont stockées dans une base de données multidimensionnelle (souvent appelée cube ou hypercube)
- Une base de données multidimensionnelle utilise une structure propriétaire au logiciel utilisé (≈ matrice)
- Le serveur MOLAP extrait les données de l'hypercube et les présente directement au module client

□

ROLAP :

- Les données détaillées de base ainsi que les données agrégées de l'entrepôt sont stockées sous forme de tables dans une base de données relationnelle
- La base de données relationnelle doit être structurée selon un modèle particulier (étoile, flocon, ...)
- Le serveur extrait les données par des requêtes SQL et interprète les données selon une **vue multidimensionnelle** avant de les présenter au module client

HOLAP

- Architecture qui consiste en un croisement des architectures MOLAP et ROLAP
- Les données détaillées de base de l'entrepôt sont stockées dans une base de données relationnelle et les données agrégées sont stockées dans une base de données multidimensionnelle
- Le serveur HOLAP accède deux bases de données et les présente au module client, selon une vue multidimensionnelle dans le cas des données de la BD relationnelle

SOLAP

Une plate-forme visuelle supportant l'exploration et l'analyse spatio-temporelle faciles et rapides des données selon une approche multidimensionnelle à plusieurs niveaux

d'agrégation via un affichage cartographique, tabulaire ou en diagramme statistique.

PLATE FORMES ET OUTILS

Dans cette séquence, nous nous intéressons à la plateforme décisionnelle **Pentaho**. Cette plateforme a l'avantage d'aborder tous les éléments de la chaîne décisionnelle :

- Extraction de données sources dans des supports de stockages hétérogènes,
- Manipulation et navigation des données multi-dimensionnelles,
- Création de rapports synthétiques pour la visualisation et l'analyse des données,
- Fouille des données dans un entrepôt à partir de différentes catégories d'algorithmes issus du datamining,
- Gestion des métadonnées à travers l'exploitation de référentiel,
- Communication sécurisée à travers les différents réseaux informatiques d'une entreprise,
- Plateforme de développement avec un fort potentiel évolutif.

Pentaho, est une suite logiciel qui intègre de façon homogène des composants open source existants pour constituer une plateforme décisionnelle performante :

Kettle pour l'extraction des données sources et l'alimentation des datamarts,

Mondrian, **JPivot**, **JRubik** pour la gestion et l'analyse des données multi-dimensionnelles (OLAP),

Weka pour la fouille des données,

[BIRT](#), [JfreeReport](#), [JaspertReports](#), Pentaho Reporting pour la réalisation de rapport graphiques et/ou textuelles,
[Enhydra Shark](#) pour la modélisation et la gestion des flux d'informations au sein des entités d'une entreprise,
[Hibernate](#) pour la persistance des objets,
[IDE Eclipse](#) comme environnement de développement,
[Java](#) comme langage principale de développement,
[Jboss](#) comme serveur d'applications et d'intégration,
[PHP](#) et [JSP](#) comme langage et outils web pour la création de rapport dans un navigateur Internet,
[HSQLDB](#), [MYSQL](#) comme base de données relationnelle,
[Quartz](#) pour la réalisation de la planification des taches.

GLOSSAIRE

La technologie OLAP regroupe un certain nombre de termes précis désignant des éléments de la structure multidimensionnelle. De plus, plusieurs abréviations spécifiques sont souvent rencontrées dans le monde OLAP. Un glossaire s'impose donc.

Ce glossaire n'a qu'une vocation didactique et ne constitue donc pas une référence stricte, il est appelé à évoluer suivant vos critiques.

Agrégation

Action de calculer les valeurs associées aux positions parents des [dimensions hiérarchiques](#). Cette agrégation peut être une somme, une moyenne, ou tout autre processus plus complexe comme la deuxième plus forte valeur.

Attribut

Un fait décrivant chaque [position](#) d'une dimension.

Axe

Correspond à une [dimension](#).

Cellule

Une donnée définie par une position de chaque dimension. Les cellules d'un [hypercube](#) peuvent être vides ou remplies. Lorsqu'un grand nombre de cellules sont vides, on parle de données éparses.

Cube

Le plus souvent, synonyme d'[hypercube](#).

Datamart

L'ensemble des données se rapportant à un des métiers de l'entreprise. Plusieurs datamart forment le datawarehouse de l'entreprise.

Datawarehouse

Entrepôt de données. Ce terme anglais est utilisé pour désigner l'ensemble des informations d'une entreprise, enregistrées sous un format informatique.

Dimension

Un ensemble de données du même type, permettant de structurer la base multidimensionnelle. Une dimension est parfois appelée un axe. Chaque cellule d'une [mesure](#) est associée à une seule [position](#) de chaque dimension. Temps, pays, produit sont des dimensions classiques.

DOLAP

Desktop OLAP. Ce terme désigne un petit produit OLAP faisant de l'analyse multidimensionnelle en local. Il peut y avoir une mini base multidimensionnelle (façon Personal Express), ou bien de l'extraction de cube (façon Business Objects).

DSS

Decision Support System, ou système d'information décisionnel. C'est un système d'interrogation et de présentation des données adapté pour l'aide à la décision. Le terme français équivalent est

SIAD, ou Système d'Information d'Aide à la Décision. Un autre terme anglais est EIS, ou Executive Information System.

EIS

Executive Information System. Le terme anglais plus couramment utilisé est DSS, ou Decision Support System.

FASMI

Fast Analysis of Shared Multidimensional Information, ou analyse rapide d'information multidimensionnelle partagée. Ces cinq termes ont tous leur importance dans la définition de la technologie [OLAP](#).

Formule

C'est un hypercube virtuel, c'est à dire que les valeurs obtenues sont le plus souvent calculées à la volée mais non stockée dans la base de données.

Hiérarchie

Les positions d'une dimension organisées selon une série de relations 1-n en cascade. Cette organisation de données est comparable à un arbre logique, ou chaque membre n'a pas plus d'un père mais un nombre quelconque d'enfants.

HOLAP

Hybrid OLAP. Désigne les outils d'analyse multidimensionnelle qui récupèrent les données dans des bases relationnelles ou multidimensionnelles, de manière transparente pour l'utilisateur.

Hypercube

Une construction multidimensionnelle formée de la conjonction de plusieurs dimensions. Chaque [cellule](#) est définie par un seul membre de chaque [dimension](#).

MDB

Multidimensional DataBase. Permet le stockage, le traitement et la restitution de données multidimensionnelles.

Mesure

Un hypercube, le plus souvent de type entier ou décimal, structuré par des dimensions. Salaire, Prix, Quantité, Coût sont des mesures classiques.

MOLAP

Multidimensional OLAP. Ce terme désigne plus spécifiquement une technologie de stockage cartésien. MOLAP s'oppose à [ROLAP](#). Pour le premier, les jointures sont déjà faites, ce qui explique les performances. Dans le second, les jointures entre les tables de dimension et de fait sont effectuées au moment de la requête.

Multicube

Une construction multidimensionnelle formée de plusieurs hypercubes partageant certaines dimensions.

Multidimensionnel

Structure de données ayant au moins trois dimensions indépendantes.

Niveau hiérarchique

Au sein d'une hiérarchie, les positions sont en général organisées en niveaux. Les positions d'un même niveau correspondent à une classification précise. Par exemple, on peut concevoir une dimension "temps", pour laquelle les jours sont au niveau 1, les mois au niveau 2 et les années au niveau 3.

OLAP

Littéralement, On-Line Analytical Processing. Désigne une catégorie d'applications et de technologies permettant de collecter, stocker, traiter et restituer des données multidimensionnelles, à des fins d'analyse. Une autre définition est résumée dans l'acronyme [FASMI](#)

(Fast Analysis of Shared Multidimensional Information), ou analyse rapide d'information multidimensionnelle partagée. Les outils OLAP doivent respecter 12 règles précises que vous pouvez découvrir à [cette page](#).

Position

Une valeur d'une [dimension](#).

RDBMS

Relational DataBase Management System. Permet le stockage, le traitement et la restitution de données stockées dans des tables relationnelles. Son équivalent français est [SGBDR](#), ou Système de Gestion de Base de Données Relationnelle.

Relation

Une relation entre les positions de deux dimensions distinctes permet d'effectuer facilement des calculs à la volée pour définir de nouvelles formules.

ROLAP

Relational OLAP. Il s'agit d'un ou plusieurs schémas en étoile stockés dans une base relationnelle. Cette technique permet de faire de l'analyse multidimensionnelle à partir de données stockées dans des bases relationnelles.

SGBDR

Système de Gestion de Base de Données Relationnelle. Equivalent de [RDBMS](#).

SIAD

Système d'Information d'Aide à la Décision. Equivalent de EIS.

Schéma en étoile

Arrangement de tables dans une base de données relationnelle. Au centre, on trouve la table de faits, dont les colonnes constituent les

mesures du multidimensionnel. Les branches de l'étoile qui rayonnent à partir de la table de fait correspondent aux dimensions. Le modèle conceptuel de données permet de retrouver cette forme en étoile.

Variable

En général synonyme de [mesure](#).

BIBLIOGRAPHIE

• <http://pagesperso-orange.fr/bernard.lupin/>

• http://www.journaldunet.com/solutions/0301/030108_olap.shtml

• <http://www.piloter.org/business-intelligence/olap.htm>

• <http://fr.wikipedia.org/wiki/OLAP>

• Cours informatique décisionnelle (NFE115) Cnam de basse-Normandie

Gilles LEBRUN et Christophe CHARRIER