MEMOIRE DE FIN D'ETUDE

Présenté par

MATTHIEU GRIMAUD

Pour obtenir le Diplôme de

L'INSTITUT SUPERIEUR DU COMMERCE DE PARIS

Spécialisation :

Marketing et Management des Technologies de l'Information

Comment assurer l'intégration des données structurées dans l'entrepôt de données

Soutenu en Juillet 2007

Consultant de mémoire : Mr Liottier Miguel

Mémoire réalisé dans le cadre d'un stage pour la société CGI auprès du client GEFCO

Sommaire

Synthèse

Les données du Data Warehouse proviennent de progiciels spécialisés (Best of Breed) ou de progiciels intégrés (ERP). Les ERP permettent de couvrir plusieurs métiers de l'entreprise (comptabilité, commercial,…) et les données correspondantes sont présentes dans une même base de données. En revanche, les Best of Breed sont spécialisés sur un métier mais doivent échanger des données avec les autres progiciels. Dans le cadre de l'intégration des données, il est plus simple d'importer l'ensemble des données depuis une même base de données (cas de l'ERP). Toutefois, cela est rarement le cas et il est donc nécessaire de spécifier plusieurs bases de données sources.

Les données stockées dans le Data Warehouse sont ensuite utilisées pour diverses applications. Ce mémoire étant focalisé sur les données, seuls le reporting de la Business Intelligence et le Data Mining sont concernés. La Business Intelligence permet, entre autres, de présenter les données de l'entrepôt, de même, la navigation OLAP permet à l'utilisateur de modifier ses critères de recherche sans avoir de compétences spécifiques. Dans le cadre du Data Mining, c'est l'application qui se charge d'identifier les relations entre les données afin de déterminer des tendances : pour un supermarché un exemple de tendance est d'avoir identifié que les clients achetant de la bière achetaient aussi des couches. Rapprocher ces deux produits a permis d'augmenter les ventes.

Pour intégrer les données, on peut avoir recours à deux approches différentes : l'approche médiateur qui permet de mettre en relation des données sources avec les requêtes d'analyse, et l'approche basée sur le Data Warehouse où les données sont extraites des systèmes sources, stockées physiquement sur un support puis mises à disposition des requêtes d'analyse (Business Intelligence et Data Mining). Pour mettre en place une analyse des données, la structure comportant un entrepôt de données est mieux adaptée car elle permet d'obtenir des analyses transversales à plusieurs métiers [1] dont on peut suivre l'évolution au cours du temps. Cette solution permet aussi de préserver les ressources du système source.

[1] Un métier regroupe par exemple les informations de la comptabilité et du commercial

Le Data Warehouse entrepose les données issues de l'intégration selon une structure multidimensionnelle, basée sur une table de faits et des tables de dimension. La table de dimension contient les données liées à un thème et la table de fait désigne, pour une combinaison de dimensions, les données chiffrées correspondantes. Certaines dimensions comme le temps sont obligatoires pour permettre l'analyse. Le Data Mart est une base de stockage qui diffère du Data Warehouse dans la mesure où les données peuvent y être dupliquées et que leur objectif est de faciliter l'analyse et non de recenser l'ensemble des données.

Dès lors, intégrer les données dans le Data Warehouse se résume à prélever des données hétérogènes et à leur appliquer les modifications nécessaires pour qu'elles puissent être chargées dans l'entrepôt de données.

Il faut dans le système source identifier les données concernées par l'extraction (toutes les données ne sont pas nécessaires pour réaliser les analyses en décisionnel). Pour éviter de consommer les ressources du système source et pour simplifier l'extraction, on modélise les données que l'on a précédemment sélectionnées. Cela revient à constituer une base de données supplémentaire qui recense toutes les informations qui devront être présentes dans le Data Warehouse. Ensuite, on marque les données afin d'identifier celles qui seront concernées par la prochaine extraction. Différentes techniques de marquage existent, le choix dépend de la nature ainsi que du contexte des données à marquer[2]

La phase de transformation permet de nettoyer les données afin de les débarrasser de leurs doublons, de leurs problèmes de synonymie et de réaliser des jointures ou des répartitions. Cette phase permettra au Data Warehouse d'avoir une information transformée et de meilleure qualité par rapport aux données sources. Dans le cas de GEFCO, cette phase permet de répartir la valeur des amortissements en fonction des unités opérationnelles.

La phase de chargement permet d'envoyer les données transformées vers le Data Warehouse en mode différé (mode batch). Lors du premier chargement des données, on envoie la totalité des informations sélectionnées : il s'agit du mode « Full ». Ensuite, il est nécessaire d'actualiser les données de l'entrepôt : on utilise le mode « Reconstruction» pour remplacer la totalité des données précédentes par les nouvelles. Cette technique consomme beaucoup de ressources système, c'est pourquoi on peut intégrer les données en mode « Delta ». Dès lors,

[2] W. H. Inmon, *Loading data into the Warehouse*, 2000, pp.8-17

on ne collecte que les données qui ont été modifiées ou créées depuis la dernière actualisation de l'entrepôt de données.

Les ETL se rapprochent de plus en plus des EAI (Entreprise Application Integration) et proposent maintenant d'intégrer les données en temps réel. Ceci est généralement utilisé dans le cadre du reporting opérationnel où les données manipulées ne sont pas trop agrégées et limitent par conséquent l'utilisation des systèmes sources. D'autre part, les ETL permettent actuellement d'alimenter plusieurs sources en même temps, ce qui permet de répondre aux problématiques d'entreprise en terme par exemple de délocalisation des équipes de travail ou de logique d'intégration avec les clients ou les fournisseurs. C'est ainsi que GEFCO intègre certaines données dans les Data Warehouse du groupe PSA.
Cependant, les ETL restent centrées sur la manipulation des données et ne s'occupent pas des processus. On peut dès lors affirmer que l'ETL et l'EAI sont encore actuellement deux outils différents.

L'intégration des données selon le processus général, précédemment décrit, diverge de celui d'un système décisionnel intégré de type SAP BW. Dans ce cas, les bases de données sources et cibles sont modélisées par des structures virtuelles et confrontées lors du Mapping. Cette étape permet de « traduire » les données sources en données cibles.
Ce modèle diffère aussi au niveau des processus : sous SAP BW, la phase de chargement est présente tout au long du processus et la phase de transformation intervient à plusieurs moments que se soit lors du Mapping (correspondance entre les champs sources et cibles), lors des règles de transformation (agrégation de plusieurs sources, application d'un algorithme pour répartir les données,…) et éventuellement au sein des cubes où les données peuvent être agrégées selon des règles définies pour chaque cas (regroupement des données par jour,…).

Introduction

Actuellement les entreprises doivent faire face à un marché très concurrentiel qui évolue extrêmement rapidement. Pour acquérir de nouvelles parts de marché, l'entreprise doit mieux comprendre ses clients et être réactive pour identifier les nouveaux relais de croissance. Dès lors, l'entreprise doit davantage prendre en compte l'évolution du marché afin d'en détecter les opportunités et les menaces.

Pour ce faire, l'entreprise a besoin d'outils qui lui permettent de déceler tous ces éléments. Les outils traditionnels actualisés par les salariés ne sont pas fiables, du fait du manque d'intégrité des données, des difficultés de consolidation et de l'absence de prévision de tendances.

Dès lors, l'entreprise se doit d'acquérir un système décisionnel qui puisse répondre à ces contraintes, c'est le cas des outils de Business Intelligence. Ceci va permettre de collecter les informations nécessaires à la mise en place de tableaux de bord adaptés au pouvoir de décision des salariés : opérationnel, décisionnel ou stratégique. L'entreprise pourra, par le biais du Data Mining, identifier les points communs entre les habitudes de consommation de ses clients afin d'améliorer sa relation avec ces derniers.

Les applications de Business Intelligence permettent à l'entreprise d'avoir une image parfaite de son activité et de faciliter la prise de décisions éclairées.

Cependant, la Business Intelligence peut se révéler être un dangereux outil pouvant mener l'entreprise à sa perte si les informations dont elle dispose ne sont pas correctes ou si elles ne sont pas mises à jour. Alors l'entreprise prendra des décisions en fonction de données fausses et la situation pourra se révéler catastrophique.

C'est pourquoi l'intégration des données est capitale dans un projet de Business Intelligence. En effet, les données jouent un rôle central par rapport à leur analyse et à leur présentation, beaucoup plus simple à mettre en œuvre. Selon Ralph Kimball, un spécialiste dans le domaine, l'intégration des données représente près des trois quart du temps et du budget accordés pour un projet de ce type. Dès lors, on peut considérer que l'intégration des données est l'étape critique d'un projet de Business Intelligence.

Les données à intégrer peuvent avoir plusieurs formes : issues d'un système d'information ou de fichiers plats, on parle alors de données structurées. En revanche, pour celles provenant d'Internet, elles sont présentées sans réelle logique commune, on parle de données semi ou non structurées.

Nous limiterons notre analyse aux données structurées dans la mesure où elles fournissent des informations stratégiques et représentent une part considérable des données intégrées.

Le support de destination de ces données est le Data Warehouse ou entrepôt de données. Il s'agit d'une base de données dont la structure dimensionnelle permet de faciliter le stockage et la disposition des informations afin de les analyser ultérieurement.

Ce mémoire s'articule en deux parties principales : la première présentera l'état de l'art permettant d'expliquer le contexte du sujet et la seconde s'intéressera à l'intégration des données.
Dans la partie de l'état de l'art, le premier chapitre présentera les systèmes d'informations qui génèrent l'information en amont du Data Warehouse.
Le second chapitre analysera les applications qui utilisent les données du Data Warehouse et indiquera les différences entre la présentation, l'analyse et l'interprétation des données.
Le chapitre suivant présentera les architectures destinées à exporter les données sources. On abordera une architecture qui « concurrence » le Data Warehouse et on verra en quoi ce dernier est adapté pour intégrer les données structurées.
Le dernier chapitre de cette première partie examinera le Data Warehouse, son fonctionnement, son architecture et ses caractéristiques.

La seconde partie sera centrée sur l'intégration des données structurées dans le Data Warehouse. Le premier chapitre abordera d'un point de vue général l'extraction, la transformation et le chargement des données.
Le second chapitre présentera ce processus par le biais de la solution SAP BW au travers de l'exemple du groupe GEFCO. Ce chapitre expliquera comment assurer la répartition de la valeur des immobilisations lors de leur intégration dans le Data Warehouse.

1ère partie : Etat de l'art

chp 1. En amont du Data Warehouse

Les données liées à l'activité de l'entreprise naissent au sein de deux principaux types de systèmes d'informations : les Progiciels de Gestion Intégrés (ERP) ou les progiciels spécialisés (Best Of Breed).

Les Best of Breed sont des progiciels spécialisés dont le domaine d'application est centré sur un métier [3] de l'entreprise. Les informations contenues dans un même progiciel spécialisé circulent très facilement entre les différents postes du métier couvert. En revanche, sans outil complémentaire, ce progiciel ne peut communiquer ces informations aux autres applications de l'entreprise.

Les ERP sont au contraire généralistes et couvrent un ensemble de métiers de l'entreprise. La gestion des processus est centralisée et l'ensemble des métiers concernés communique facilement.

I. L'ERP

Schéma d'un ERP – Source Microsoft

A. Son utilité

L'ERP a pour objectif d'améliorer la performance des processus de l'entreprise. Pour ce faire, il se compose d'une base à laquelle on peut ajouter des modules qui couvrent les métiers de l'entreprise. On trouve par exemple des modules financiers (comptabilité générale et analytique, finance), logistiques, ressources humaines, relation client,…

[3] Métier : désigne l'activité des salariés : comptabilité, ressources humaines, logistique,…

Ainsi, l'entreprise peut s'équiper au fur et à mesure et faciliter la formation de ses salariés. Cependant, l'atout majeur de l'implémentation d'un ERP reste l'homogénéisation du système d'information. Les opérations enregistrées par les salariés sont immédiatement prises en compte et selon des règles définies, auront des répercutions sur d'autres métiers de l'entreprise. Ces tâches étaient auparavant exécutées par les salariés : ils sont alors déchargés de ces tâches répétitives et sans valeur ajoutée et l'entreprise gagne du temps de main d'œuvre et augmente la fiabilité de ces processus.

B. Une base de données commune

L'ensemble des données liées à la gestion des métiers est stocké dans une base de données relationnelle appelée base métier. L'information est classée selon une logique métier afin d'optimiser leur traitement.

C. Mise en place d'un reporting opérationnel

L'ERP conserve l'ensemble des informations dans une base de données. Dès lors, effectuer leur analyse pourrait sembler assez simple.
En réalité, les ERP (logique métier) et les systèmes de reporting (logique analytique) ont une approche différente vis à vis des données. Par exemple, une information comme le chiffre d'affaires est composée d'une multitude de données disséminées dans les nombreuses tables de l'ERP.
Pour pallier à ce problème, on met en place une vue métier qui permet de regrouper et d'organiser les données en fonction des métiers de l'entreprise. Ainsi, en fonction de son métier, l'utilisateur accèdera aux informations dont il pourrait avoir besoin pour construire un reporting opérationnel.

D. Avantages de ce type de reporting

Un grand débat est mené pour déterminer le support le plus adéquat pour le reporting opérationnel. Les partisans du support ERP se basent sur divers arguments que voici[4].
Les applications métiers ne rencontrent pas de problème d'actualisation. En effet, le système génère des données qui sont immédiatement stockées dans des bases de données et instantanément disponibles pour l'analyse.

---

[4] J.M. FRANCO, J. De LIGNEROLES., *Piloter l'entreprise grâce au Data Warehouse*, Editions Eyrolles 2000

Dans le cadre de ce reporting, utiliser un Data Warehouse n'apporte pas de valeur ajoutée à l'utilisateur. De plus, le nombre de salariés concernés par ce type d'analyse est très important, les coûts de formation pour utiliser ce nouveau support seraient donc conséquents. Dès lors, utiliser un Data Warehouse entraînerait des coûts supplémentaires sans apporter de valeur ajoutée. Il vaut donc mieux utiliser l'ERP qui évite à l'entreprise d'être confrontée à ce contexte.

E. Limites du reporting opérationnel

Ce système de reporting est seulement adapté aux opérationnels et rencontre donc des limites sur plusieurs plans :

Du point de vue de l'évolutivité : plus on souhaite se détacher des données opérationnelles et avoir une vision synthétique des données, plus le système aura des difficultés à les produire. En effet, l'agrégation [5] des données nécessite le rapprochement de différents métiers de l'entreprise pour avoir une vision globale. L'ERP n'est pas capable de réaliser cela car il raisonne par métier (il sera difficile de rapprocher les données financières et client par exemple)

Du point de vue technique : Les ERP sont capables de gérer de nombreuses situations au sein de l'entreprise. Pour ce faire, ils sont dotés d'une structure complexe qui leur permet d'assurer un service fiable. Cela implique que les données soient décomposées et éparpillées dans les bases de données du système. En vue de constituer le report, extraire l'ensemble de ces données dans de nombreux endroits de la base de données peut occuper une part importante des ressources du système.

Du point de vue analytique : L'ERP ne permet pas de fournir des données historisées : seul le Data Warehouse est capable de le faire du fait de sa modélisation multidimensionnelle.

Du point de vue de l'intégration : si l'entreprise utilise d'autres systèmes d'information, il devient impossible d'intégrer ces données dans les bases de l'ERP. Il en est de même pour l'intégration des données externes à l'entreprise.

Ainsi, le reporting basé sur l'ERP ne peut répondre aux attentes de l'entreprise que dans de très rares cas. Parmi ces exceptions, on recense la situation d'une entreprise dont la création du système décisionnel et du Data Warehouse est en cours. En attendant que l'application d'aide à la décision fonctionne pour ses utilisateurs, l'entreprise a toujours besoin d'un

[5] Agrégation : regroupement des données

reporting pour assurer son fonctionnement. C'est dans ce cas de figure que l'entreprise peut baser temporairement son analyse décisionnelle sur son ERP.

II. Le Best of Breed

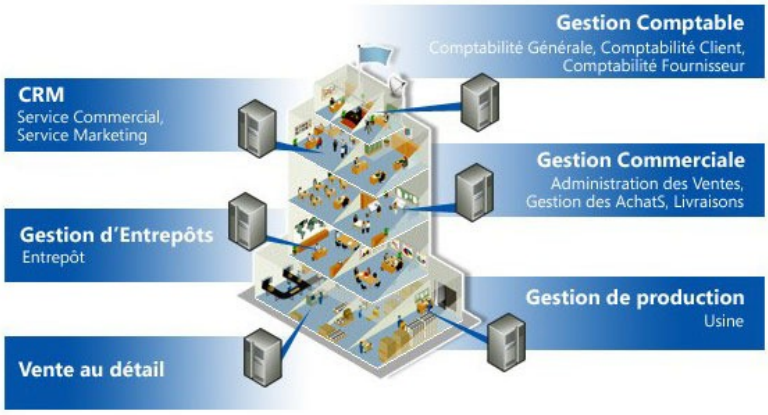


Schéma Best of Breed – Source Microsoft

A. Définition

Best of Breed se traduit littéralement par : « meilleur de sa catégorie ». Il qualifie une solution logicielle prétendant offrir des fonctions avancées sur un segment de marché bien délimité. Cette solution est souvent utilisée pour des problématiques que les ERP ne peuvent résoudre du fait de leur polyvalence. En effet, certains secteurs ont des process originaux qui nécessitent un progiciel spécialisé pour y répondre.

Ils sont aussi utilisés par les entreprises qui reprochent aux ERP leur manque d'intuitivité et de flexibilité ; en effet, ces progiciels intégrés se basent sur le fait que c'est à l'utilisateur de s'adapter à l'ERP et non l'inverse.

Le champ d'application de ce type de solution est limité à un métier de l'entreprise : il devra donc cohabiter avec les autres systèmes d'information. En effet, pour fonctionner, ce logiciel doit avoir à sa disposition les données générées par les autres applications (données client, fournisseur,…) mais doit aussi communiquer les données qu'il génère. Cela soulève le problème de son intégration avec les autres progiciels de l'entreprise que l'EAI permet de résoudre.

B. Outils d'EAI

Avant l'existence des EAI (Intégration d'Applications d'Entreprise), les applications métiers communiquaient par le biais d'interfaces spécifiques point à point. Ce modèle était très lourd à mettre en place et très peu évolutif[6].

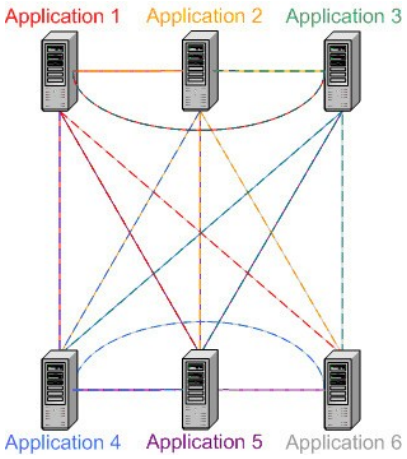


Schéma d'une infrastructure point à point – source CNRS

L'outil d'EAI a permis de pallier à ces problèmes en permettant au système d'information de gagner en souplesse et en réactivité. Ainsi, L'EAI permet de synchroniser et de faire communiquer des applications hétérogènes (les Best of Breed entre autre) par échange de flux (informations et processus) en temps réel, indépendamment des plates-formes et du format des données.

Son fonctionnement repose généralement sur l'utilisation d'un bus de communication qui permet aux applications de communiquer entre elles.

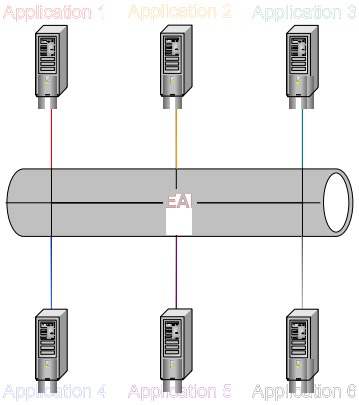


Schéma d'une interface EAI – source CNRS

Les Progiciels de l'entreprise peuvent se brancher sur le bus grâce à des interfaces génériques selon le mode « Publish & Subscribe ». Ainsi, dès qu'une information présente dans une

[6] ROUSSE D., *Panorama d'une infrastructure EAI*, CNRS, ref NOT03K012DSI 2003

application doit être transmise à d'autres applications, l'application source « publie » l'information dans le bus d'échange. Les applications destinataires peuvent alors « s'abonner » pour recevoir cette information via le bus[7]

L'EAI assure l'intégration à différents niveaux :

Au niveau des données : utilisation de connecteurs légers : c'est le mode message. Il est orienté batch[8] et les données sont transmises en temps réel (au fil de l'eau)

Au niveau des applications : utilisation de connecteurs lourds de type API. Les triggers permettent d'identifier certaines informations et de déclencher des actions sur les autres progiciels en présence.

Au niveau des processus métiers : il s'agit d'un cas de workflow où l'EAI se charge d'orchestrer l'exécution des tâches entre les différentes applications et les éventuelles interventions humaines.

Le rôle d'un EAI s'arrête à la gestion des interfaces et des échanges inter applicatifs. Il n'intervient pas dans le fonctionnement interne des Best of Breed qui restent indépendants entre eux et conservent leurs fonctions et processus métier.

GEFCO a eu recours à des Best of Breed avant de s'équiper en ERP. Le passage sur SAP s'est effectué sur les modules de finance, comptabilité, contrôle de gestion, achat et vente. Cependant, le groupe a conservé quelques outils métiers pour coller à la spécificité de l'activité de l'entreprise.

Ainsi, en amont du Data Warehouse, le système d'information est constitué de progiciels dont le type est déterminé par la complexité des processus métiers de l'entreprise. Pour ceux qui présentent une spécificité particulière, seul le Best of Breed sera en mesure d'apporter une solution efficace. En revanche pour les processus plus standardisés, l'entreprise aura plus intérêt à recourir à un ERP.

Dans le cadre de l'intégration des données au sein du Data Warehouse, le progiciel de gestion intégré présente l'avantage de disposer ses données dans une même base de données. De plus, l'architecture de celle-ci est connue des spécialistes de la construction de Data Warehouse, ce qui facilite leur manipulation ultérieure.

---

[7] Octo Technologies, *Le Livre Blanc de l'EAI – Intégration des applications d'entreprise*, 1999

[8] Batch : un message est envoyé dans le système, cela génère des actions et éventuellement un fichier de réponse

chp 2. Aval du Data Warehouse – le Décisionnel

Cette partie a pour objectif de présenter les applications principales qui permettent d'exploiter les données du Data Warehouse dans une logique décisionnelle. On trouvera alors le traitement analytique en ligne (OLAP) et l'interprétation des données : le Data Mining.

I. OLAP et les tableaux de Bord

A. Définition

OLAP signifie « On Line Analytical Processus » et repose sur une base de données multidimensionnelle, destiné à exploiter rapidement les dimensions d'une base de données. Le modèle OLAP est utilisé au sein des Data Warehouses, il permet de sélectionner et de croiser plusieurs données provenant des sources diverses.
Selon Carron, il s'agit d'une « catégorie de logiciels axés sur l'exploration et l'analyse rapide des données selon une approche multidimensionnelle à plusieurs niveaux d'agrégation ».
Ainsi, les systèmes OLAP permettent de rassembler, de gérer, de traiter et de présenter des données multidimensionnelles à des fins d'analyse et de décision. Un outil OLAP est capable de fournir une information multidimensionnelle partagée pour l'analyse rapide.
L'outil OLAP repose sur la restructuration et le stockage des données dans un format multidimensionnel. Ce format multidimensionnel connu sous le nom d'hyper cube, organise les données de long de dimensions. Ainsi, les utilisateurs analysent les données suivant les axes propres à leur métiers

B. Les travaux des chercheurs

En 1993, Edgard F. Codd publie un livre blanc "Providing OLAP to User Analysts" dans le cadre de son activité professionnelle chez Arbor Software. Il est le premier à formaliser cette notion et devient un des maîtres des bases de données multidimensionnelles. Cette réalisation défini les 12 règles qu'un système de pilotage multidimensionnel doit respecter[9].

"Ce qu'il y a d'agréable avec ces outils OLAP", explique Eric Klusman, de Cantor Fitzgerald LP, "c'est que je suis en mesure de distribuer les données aux utilisateurs sans les obliger à apprendre des complexes formules de programmation, d'interrogation ou même à ce

[9] Codd, E. F. et Salley C.T., *Providing OLAP to User-Analysts: An IT mandate*, Hyperion Livre Blanc 1993

qu'ils aient à programmer leurs tableurs". D'une façon générale, tous affirment que l'on peut interfacer de nombreux outils d'utilisateurs avec des bases de données multidimensionnelles sans qu'il soit nécessaire de consentir à de lourds efforts de formation ou des interventions importantes du service informatique.

Les opposants au modèle OLAP

Thomesen et Watson & Gray font remarquer l'absence de débat sur l'utilisation de technologies de bases de données relationnelles et multidimensionnelles pour le traitement automatique en ligne[10]

On recense cependant les propos de Nigel Pendse qui estime que le travail d'Edgard F. Codd ne serait pas totalement objectif. En effet, l'ouvrage de référence a été publié sur demande d'une société ; dès lors le terme d'OLAP lui semble biaisé et controversé.

Mr Pendse propose alors le terme de FASMI (Analyse Rapide d'Information Multidimensionnel Partagée) qui comporte cinq règles et permet ainsi de simplifier les douze règles de Codd et de faciliter l'évaluation des outils d'OLAP[11]

C. Les tableaux de bord

Selon le degré d'importance des décisions des utilisateurs, les tableaux de bord sont différents : l'outil de reporting est adapté aux décisions opérationnelles, l'outil d'analyse aux décisions managériales et l'outil ouvert donne accès à l'ensemble des données pour les décisions stratégiques.

Les outils de reporting présentent des données opérationnelles de type vente par client, par article,… Les états sont prédéfinis et donc très rapides à générer. Ils sont utilisés par le plus grand nombre d'utilisateurs de l'entreprise.

Les outils d'analyse présentent des données plus agrégées afin d'avoir une vision plus globale du périmètre considéré. L'utilisateur pourra donc constater des évolutions de la performance de l'entreprise (total chiffre d'affaires, marge,…) puis « naviguer » au sein des données (outil OLAP) pour comprendre les facteurs explicatifs. Par exemple, si l'on constate que les ventes pour la France sont en baisse, en ajoutant le critère région, on constatera que seule une région

---

[10] THOMESEN E, *OLAP Solutions : Building Multidimensional Information Systems*, New York, 1997.

[11] <http://www.olapreport.com/fasmi.htm>

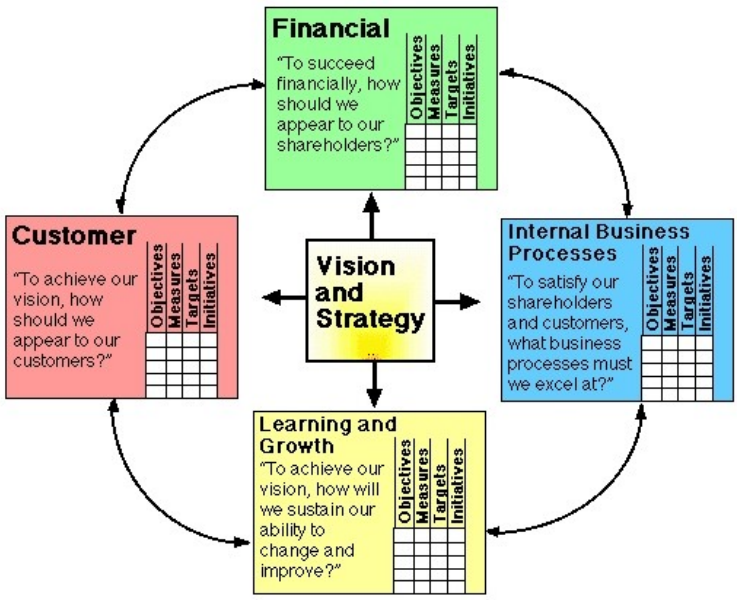
a fortement baissé ses ventes alors que les autres ont maintenu leur activité. En revanche, chaque utilisateur ne peut accéder qu'à un périmètre d'analyse limité (ventes, capacité financière, ressources humaines de l'entreprise). De même, les attributs navigationnels (région dans l'exemple) sont déterminés à l'avance.

Les outils de libre accès reprennent le principe des outils d'analyse mais n'ont aucune limitation et disposent d'une vue beaucoup plus agrégée. Cette transversalité leur permet de rassembler des éléments dont le périmètre d'analyse est différent.

Selon la nature des outils de tableaux de bord, les données peuvent être présentées sous Excel, ce qui simplifie l'apprentissage des utilisateurs : c'est le cas de SAP BW ou d'Essbase. En revanche les solutions de Business Objects s'opèrent sur des interfaces spécifiques.

II. Le Balanced Score Card

Le Balances Score Card (BSC) ou tableau de bord prospectif diffère des tableaux de bord évoqués dans la partie précédente. Le BSC s'adresse uniquement aux dirigeants de l'entreprise et procurent une vision très synthétique de l'activité de l'entreprise. Le BSC est un indicateur de la performance de l'entreprise dont D. Norton et R. Kaplan ont déterminé les axes d'analyse : finance, client, processus interne (les processus clés de l'entreprise) et apprentissage organisationnel (pilotage du changement et de l'organisation).
Les axes d'analyse sont évalués selon différents critères dont voici un exemple ci-dessous.



Source <http://www.balancedscorecard.org

Le BSC ne permet que de constater la performance de l'entreprise selon certains axes qui ne prennent pas en compte la complexité de l'entreprise. De plus, on ne peut déterminer les raisons qui expliquent les résultats présentés contrairement aux tableaux de bord issus de la navigation OLAP.

III. Le Data Mining

A. Définition

Selon Jean Michel Franco, le Data Mining permet aux utilisateurs d'accéder aux données de l'entreprise sans faire appel à un informaticien. Ces informations permettent de découvrir des corrélations entre les données [12] et de définir des tendances. Ainsi, le Data Mining permet de transformer les données en connaissance.
Les systèmes d'information décisionnels exécutent les requêtes demandées par les utilisateurs et déterminent les chiffres correspondants. En revanche, dans le cadre du Data Mining, c'est le système qui met en relation les données afin de donner à l'utilisateur des tendances. C'est pourquoi on considère que le Data Mining permet de tirer une richesse supplémentaire au sein du Data Warehouse[13].

La définition du Data Mining diffère selon les auteurs :
« L'extraction d'informations originales, auparavant inconnues, potentiellement utiles à partir des données » Frawley et Piateski Shapiro
« La découverte de nouvelles corrélations, tendances et modèles par le tamisage d'un large volume de données » John Page
Dimitri Chorafas va même jusqu'à annoncer que le Data Mining consiste à « torturer l'information disponible jusqu'à ce qu'elle avoue ».

[12] Witten et Frank, *Data Mining*, Editions Morgan Kaufmann, 2005, pp.5

[13] FRANCO J.M. et EDS institut Prométhéus*, Le Data Warehouse, le Data Mining*, Eyrolles 1997

B. Distinction entre données et connaissance

Gio Wiederhold du Stanford Institute a très bien défini la différence entre ces notions :
« Une donnée décrit des exemples ou des événements précis. Elle peut être recueillie de manière automatique ou par écrit. Son exactitude peut être vérifiée par référence au monde réel. » Les données décrivent par exemple le détail des articles d'une commande.
« Une connaissance décrit une catégorie abstraite. Chaque catégorie peut couvrir plusieurs exemples. Des experts sont nécessaires pour recueillir et formaliser la connaissance. » Dans le même exemple que précédemment (les commandes client), les connaissances indiqueront quels sont les bons clients et ceux qui présentent des risques de se tourner vers la concurrence.

C. ECD et Data Mining

La Data Mining est à l'origine une technique de fouille au sein des données tandis que l'ECD permet le nettoyage et la récupération des données.
L'ECD (Knowledge Discovery in Databases en anglais) a été défini par Fayyaed en 1996 comme "un processus non-trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données". Cette définition est la première en la matière, on ne compte pas de tentative ultérieure pour mieux définir la notion.
L'ECD est avant tout un cadre précisant la démarche à suivre pour exploiter les données, en vue d'en extraire de la connaissance.
Comme évoqué au début de cette partie, le Data Mining correspondait initialement au forage des données, il s'agissait d'une étape du processus que l'on évoque. Les spécialistes du domaine ont utilisé ce terme afin de décrire l'ensemble du processus : c'est pourquoi on parle de Data Mining pour désigner l'ensemble des actions à mener pour transformer une donnée en connaissance.

D. Opérations techniques de Data Mining

1 Les Règles d'Association

Le problème de recherche de règles d'association a fait l'objet de nombreux travaux [14]. Une publication[15], a pris l'exemple du panier de la ménagère pour illustrer le propos suivant: une

[14] H. Toivonen - *Sampling Large Databases for Association Rules* - International Conference on Very Large Databases 1996

base de données de transactions (les paniers) est composée d'items (les produits achetés). La découverte des associations consiste à chercher des ensembles d'items fréquemment liés dans une même transaction ainsi que des règles qui les combinent. Un exemple d'association pourrait révéler que « 75% des gens qui achètent de la bière, achètent également des couches».

2 Les Motifs Séquentiels

Introduits dans un ouvrage[16], les motifs séquentiels peuvent être vus comme une extension de la notion de règles d'association intégrant des contraintes temporelles. Cette recherche met en évidence des associations entre les transactions alors que les règles d'association déterminent les liens au sein d'une même transaction.

Dans ce contexte, et contrairement aux règles d'association, l'identification des individus ou objets au cours du temps est indispensable afin de pouvoir suivre leur comportement. Par exemple, des motifs séquentiels peuvent montrer que « 60% des gens qui achètent une télévision, achètent un magnétoscope dans les deux ans qui suivent ».

3 Les Dépendances Fonctionnelles

L'extraction de dépendances fonctionnelles à partir de données est abordée sous l'angle de la fouille de données [17]. La découverte de dépendances fonctionnelles est un outil d'aide à la décision à la fois pour l'administrateur de la base, les développeurs d'application, les concepteurs et intégrateurs de systèmes d'information.

4 La Classification

Elle consiste à analyser de nouvelles données et à les affecter à une classe prédéfinie en fonction de leurs caractéristiques ou attributs. On part de la supposition: « plus les volumes de données traités sont importants, meilleure devrait être la précision du modèle de classification »

---

[15] R. Agrawal, T. Imielinski, et A. Swami, Mining *Association Rules between Sets of Items in Large Databases* , 1993, pp. 207-216

[16] R. Agrawal et R. Srikant. Mining *Sequential Patterns, 11th International Conference on Data Engineering* , 1995, pp3-14

[17] S. Lopez, J.-M. Petit, et L. Lakhal, *Discovery of Functional Dependencies and Armstrong Relations* – 7ème conférence internationale sur les technologies d'extension des bases de données, 2000, p 350-364

Les techniques de classification sont par exemple utilisées pour cibler la population d'un mailing ou pour accorder un prêt en fonction du profil de l'emprunteur.

5 La Segmentation

La technique de segmentation reprend celle de la classification mais ne dispose pas de classes prédéfinies[18] : l'objectif est de grouper des enregistrements qui semblent similaires dans une même classe. De nombreux algorithmes efficaces ont été proposés pour optimiser les performances et la qualité des classes obtenues dans de grandes bases de données.
Les applications concernées incluent notamment la segmentation de marché ou encore la segmentation démographique en identifiant par exemple des caractéristiques communes entre populations.

6 Les Séries Chronologiques

L'objectif est de trouver des portions de données (ou séquences) similaires à une portion de données précise, ou encore de trouver des groupes de portions similaires issues de différentes applications.
Cette technique permet par exemple d'identifier des sociétés qui présentent des séquences similaires d'expansion, ou encore de découvrir des stocks qui fluctuent de la même manière.

[18] A.K. Jain et R.C. Dubes - *Algorithms for Clustering Data*, 1988.

chp 3. Faire le lien entre l'amont et l'aval

Les données des systèmes sources sont modélisées et stockées sur diverses sources de données hétérogènes entre elles. Le décisionnel nécessite d'avoir à sa disposition ces données afin de pouvoir générer des analyses qui reflètent l'
activité de l'entreprise.
Pour mettre à disposition ces données, deux approches sont possibles : une approche virtuelle (système médiateur) et une approche matérialisée (Data Warehouse)

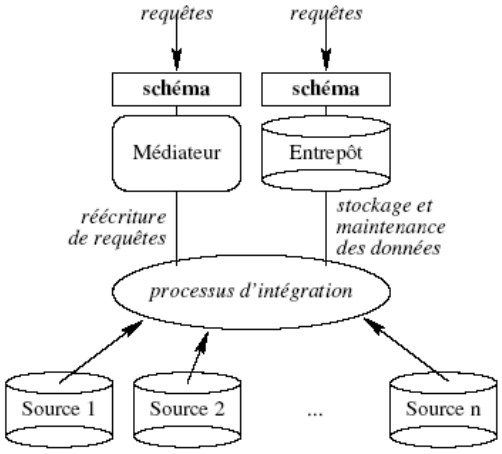


Schéma architecture d'un système d'intégration

Gio Wiederhold a défini une architecture basée sur trois niveaux :

- Sources de données : ce niveau comprend les données des systèmes sources (cf partie 1 Chapitre 1 du mémoire)
- Médiation : ce niveau permet de fournir les données au système cible
- Application : ce niveau désigne les outils d'analyse (cf partie 1 chapitre 3 du mémoire)

I. Approche virtuelle

L'approche virtuelle est apparue pour intégrer les sources de données hétérogènes. Elle propose un schéma de représentation unique qui permet de visualiser l'ensemble des données sources.

Gio Wiederhold a défini un médiateur comme "un module logiciel qui exploite l a connaissance de certains ensembles ou sous-ensembles de données pour créer de l'information pour des applications à un niveau supérieur »[19]

[19] W. Gio - *Mediators in the Architecture of Future Information Systems* -.1992.

Cette approche est qualifiée de virtuelle dans la mesure où elle ne stocke aucune donnée : elle se charge de faire le lien entre les requêtes du décisionnel et les sources de données. Ensuite, ces dernières fournissent directement les données correspondantes aux demandes du décisionnel.

Les requêtes des applications sont interprétées par les médiateurs qui, en fonction des règles de réécriture des requêtes, formulent des sous requêtes. Ces dernières seront transmises aux données métier avec une sémantique qu'elles peuvent comprendre.

Cette approche a été l'objet de nombreuses expérimentations dont le projet TSIMMIS réalisé par l'Université de Stanford. L'objectif de ce projet est d'intégrer des données non structurées et évolutives comme celles présentes sur Internet [20].. On recense aussi des projets comme DISCO[21] et YAT[22] qui permet l'intégration de données hétérogènes via XML.

Pour répondre aux requêtes des applications décisionnelles, l'approche virtuelle mobilise une partie des ressources des systèmes sources. En pratique, le nombre de requêtes ainsi effectuées est important, la performance des systèmes sources est alors affectée et par conséquent l'ensemble des utilisateurs aussi.

II. Approche matérialisée

A. Répondre au problème du médiateur

Pour résoudre ce problème, l'approche matérialisée permet d'exporter les données des systèmes sources vers un entrepôt de données appelé Data Warehouse.
Cette approche est dite matérialisée dans la mesure où les données sont physiquement présentes sur un support à part entière[23]

---

[20] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman et J. Widom, *The STIMMIS approach to mediation : Data Models and Languages*, 1995.

[21] Valduriez P, *Bases de données et Web – Enjeux, problèmes et directions de recherché*, INRIA, France

[22] G. Gardarin, *Architecture de médiation tout XML*, Laboratoire Prism

[23] J. Widom, *Research Problems in Data Warehousing*, CIKM, 1995, pp 25-30

B. Les avantages de l'approche matérialisée

Le Data Warehouse permet d'organiser l'information différemment pour permettre des analyses transversales. Prenons l'exemple des attributs d'un client : coté progiciel, ce dernier disposera d'une adresse de livraison alors que cette information ne sera pas forcément utile pour l'analyse décisionnelle. Il faudra donc se passer de cette information au risque de complexifier l'analyse

Au contraire, certains attributs d'une même analyse peuvent être présents sur plusieurs bases de données sources. Pour l'exemple de l'analyse client : d'autres attributs nécessaires peuvent être présents sur des bases sources différentes : le potentiel estimé, la codification SIREN,…

Le Data Warehouse permet de réorganiser l'information selon une logique d'analyse et non plus d'exploitation.

	Points forts	Points faibles
Virtuelle	- Très bonnes performances en terme de volume, les données sont directement manipulées dans les sources. - Mises à jour rapides	- Performances, toute requête doit être traduite pour être interprétée par les différentes sources de données. - Gestion difficile de l'historique.
Matérialisée	- Performances, les actions sont directement effectuées, sans traduction, dans le référentiel. - Possibilité d'historisation des données au sein du référentiel - Systèmes de stockages efficaces (arbres…)	- Volume, les données sont à la fois dans le référentiel et dans les sources de données - Mise à jour nécessitant la copie des données du référentiel vers les sources de données ou inversement

C. Le reporting opérationnel

La première partie annonçait les avantages et inconvénients du reporting opérationnel reposant directement sur les bases d'un ERP, voici le cas d'un modèle avec un Data Warehouse.

1 Avantages d'un reporting opérationnel basé sur un Data Warehouse

Selon les spécialistes en faveur du Data Warehouse, il est nécessaire d'intégrer les données opérationnelles en son sein car elles sont utilisées dans le cadre d'analyses transversales. En cas de phénomène inattendu au niveau de ces résultats, les dirigeants peuvent accéder aux

données détaillées afin de mieux comprendre le contexte opérationnel dans lequel ces résultats se sont formés.
L'architecture et la disposition des données selon une logique métier permettent d'optimiser le Data Warehouse et de consommer ainsi moins de ressources. On obtient des requêtes plus rapides à exécuter et les ressources du système sont moins mises à contribution.

2 Limites d'un reporting opérationnel base sur un Data Warehouse

L'actualisation des données ne se fait pas en temps réel. Cela pose des problèmes pour l'analyse des données opérationnelles comme les ventes en cours, l'état des stocks dans un contexte de flux tendus,… Or ces données sont importantes pour les fonctions opérationnelles de l'entreprise.
On retrouve ce problème d'actualisation lorsque des erreurs ont été inscrites au cours de la saisie : l'actualisation de la correction ne sera pas immédiate. Par exemple, cela peut poser de sérieux problèmes pour évaluer la performance des commerciaux. Il suffit qu'en fin de période, ces derniers passent des commandes avec des erreurs de quantité, leurs statistiques seront rehaussées.
Le Data Warehouse prête une attention particulière à la qualité de ses données : celles qui ne sont pas complètes ne sont pas intégrées. Une vente peut être considérée comme une information complète dans la mesure où elle est constituée de données liées à la commande et à la livraison. Dans ce cas, tant que le produit n'est pas livré, la vente n'est pas intégrée au sein du Data Warehouse. Une profonde étude doit déterminer les éléments contenus dans chaque information liée à l'activité de l'entreprise.

3 Une solution hybride pour le reporting opérationnel

Le leader des ERP, SAP a mis en place une solution hybride pour pallier aux problèmes liés au reporting opérationnel. Celui-ci est intercalé entre les applications opérationnelles et le Data Warehouse.
Cette solution comporte la notion d'historique des données et celles-ci sont alimentées en temps réel par les applications transactionnelles.
Le Data Warehouse est quant à lui alimenté par cette structure ainsi que par des données externes. Il dispose d'une analyse transversale et se trouve donc réservé aux dirigeants qui souhaitent des analyses.

chp 4. Le Data Warehouse

I. Les définitions

Voici les définitions du Data Warehouse ou entrepôt de données par les principaux théoriciens du domaine :

W.H. Inmon, souvent considéré comme le "père de l'entreposage de données", estime qu'un entrepôt de données est « un ensemble de données thématiques, cohérentes, évoluant dans le temps, fiables, sur lequel les dirigeants fondent leur processus de décision »[24]

Ralph Kimball, qui est probablement le "gourou" le plus connu après W.H. Inmon dans le domaine des entrepôts de données, définit ce concept comme "un exemplaire de données relatives à des transactions structuré spécifiquement à des fins de consultation et d'analyse"[25] .

Barry Devlin décrit un entrepôt de données comme étant "une mémoire unique, complète et cohérente de données provenant de sources diverses et mises à la disposition des utilisateurs finals sous une forme compréhensible et utilisable dans un contexte commercial" [26]

Même si les théoriciens sont d'accord sur la plupart des caractéristiques, leurs points de vue peuvent toutefois diverger considérablement dans les détails. Ainsi, l'entrepôt de données de W.H. Inmon ne correspond pas en tous points à celui de R. Kimball.

II. La modélisation de l'entrepôt de données

Ralph Kimball est un farouche défenseur de la modélisation dimensionnelle. Selon lui, un entrepôt de données devrait être constitué de plusieurs schémas en étoile avec des cubes de données thématiques reliés entre eux et formant un "dépôt de données". Des liaisons transversales sont établies entre les différents dépôts au moyen de dimensions homogènes (par exemple, le "client" ou le "produit"). Les données relatives aux dimensions provenant des

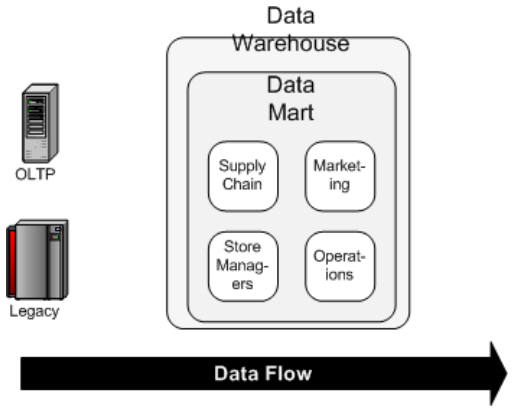
---

[24] W.H. Inmon, "What is a Data Warehouse?", 2000 publié sur le Web à l'adresse suivante: <http://www.cait.wustl.edu/cait/papers/prism/vol1_no1>

[25] R. Kimball, *The Data Warehouse Toolkit*, Editions John Wiley & Sons, 2002

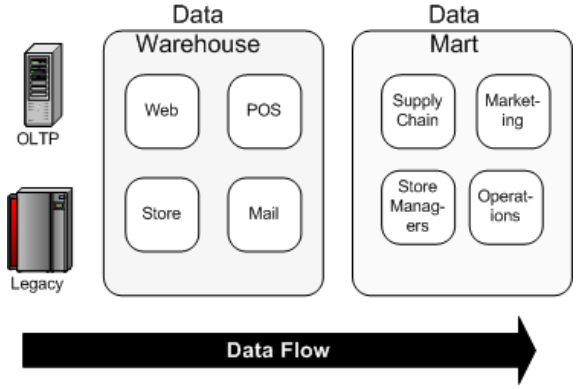
[26] B. Devlin, *Data Warehouse: from architecture to implementation*, Editions Addison-Wesley, 1997

différents systèmes en amont sont consolidées et intégrées dans une "zone de transfert" (qui ne doit pas nécessairement être relationnelle mais peut se composer de fichiers plats).



source DMReview.com

D'autres auteurs en revanche, comme W.H. Inmon, définissent un entrepôt de données comme un répertoire standardisé à l'échelle de l'entreprise auxquels les utilisateurs finaux peuvent accéder directement, seulement dans des cas exceptionnels. Des quantités fragmentaires de données circulent de ce lieu de stockage central vers des dépôts de données spécifiques à des services et fonctions dotées d'une structure multidimensionnelle. Cette architecture à plusieurs niveaux requiert l'élaboration d'un modèle de données à l'échelle de l'entreprise. Dans la pratique, l'échec des projets sur des entrepôts de données est souvent imputable à la complexité de cette tâche.



source DMReview.com

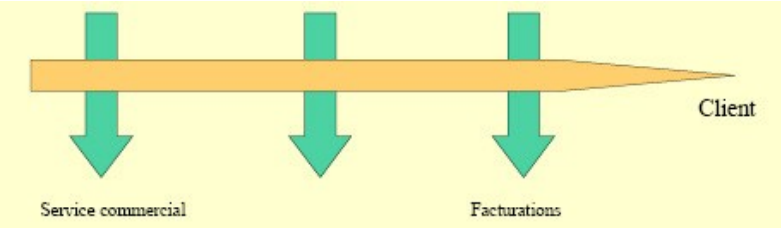
Bill Inmon, père fondateur du Data Warehouse en donne une définition[27]

« Le Data Warehouse est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision »

[27] B. Inmon, *Using the Data Warehouse*, 1994

A. Données thématiques

La vocation du Data Warehouse est de prendre des décisions autour des activités majeures de l'entreprise. Les données sont ainsi structurées autour de thèmes ce qui facilite l'analyse transversale. Pour éviter le doublonnage des données, on regroupe les différents sujets dans une structure commune. Ainsi, si le sujet client contient des informations dans les sujets marketing, ventes, analyse financière, on regroupera ces trois sujets au sein du thème client. Dès lors, chaque donnée n'est présente qu'à un endroit et le Data Warehouse joue bien un rôle de point focal.



Données orientées sujet

B. Données intégrées

Les données sont mises en forme selon un standard afin d'obtenir la transversalité recherchée. Cela nécessite une forte normalisation, une bonne gestion des référentiels et de la cohérence, une parfaite maîtrise de la sémantique et des règles de gestion des données manipulées.
Lors de l'alimentation des données, ces dernières sont hétérogènes et proviennent de systèmes opérationnels différents (ETL, Best of Breed). Il faut doter ces données d'une codification unique et pertinente afin qu'elles puissent aisément s'intégrer dans le Data Warehouse. Il faudra donc faire appel à des conventions de nomage, des structures de codage, qualifier les mesures et réaliser l'intégration de la sémantique.

Exemple d'unification de codage : la donnée Homme ou Femme a diverses codifications selon les sources. Cette notion soulève le problème de la granularité des données : le niveau de détail des bases de données de production n'est pas celui du Data Warehouse. Ce sujet sera développé dans le premier chapitre de la seconde partie : La Transformation des données.

C. Données non volatiles

Les données intégrées dans l'entrepôt le sont « Ad Vitam Eternam » et ne peuvent subir aucune altération. Cela se justifie pour assurer la fiabilité des résultats des requêtes. Ainsi, une même requête lancée à plusieurs mois d'intervalle donnera toujours les mêmes résultats. Cela

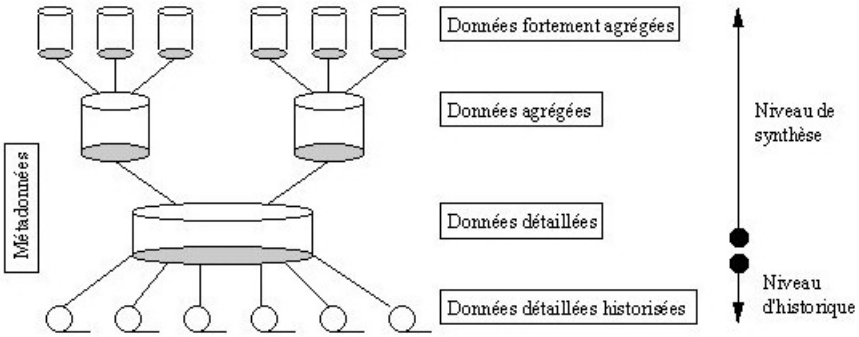
permet au Data Warehouse d'acquérir au cours du temps un historique détaillé de l'activité de l'entreprise. Ceci s'oppose fondamentalement à la logique des systèmes de production qui remettent à jour les données qui sont de nature volatiles.

D. Données historisées

L'ensemble des données qui sont intégrées dans l'entrepôt contient un ensemble de caractéristiques qui sont datées. L'historisation est nécessaire pour suivre dans le temps l'évolution des différentes valeurs des indicateurs à analyser. Ainsi, un référentiel temps doit être associé aux données afin de permettre l'identification dans la durée de valeurs précises.
Si tel n'était pas le cas, l'analyse ne serait pas possible, le suivi de l'évolution non plus. Cela rejoint la notion de non volatilité expliquée précédemment.

III. Structure des données

Au sein du Data Warehouse, les données sont organisées en quatre classes selon un axe historique et un axe synthétique : les données détaillées, agrégées, historisées et les métadonnées



Structure des données – source Université de Rennes1

A. Données détaillées

Elles constituent le socle de l'entrepôt de données et sont directement issues des systèmes de production. Elles reflètent les évènements les plus récents.
Même si elles sont détaillées, les données ne sont pas forcément identiques à celles des bases de production. En effet, le niveau d'agrégation entre ces deux bases de données n'est pas le même : les progiciels fonctionnement avec des données exhaustives et éparpillés. L'entrepôt n'a pas besoin d'avoir un niveau de détail si élevé : c'est pourquoi il faudra définir des règles d'agrégation. Par exemple en grande distribution, les bases de production raisonnent en référence produit alors qu'en décisionnel, la référence de base est le ticket de caisse.

Malgré cette agrégation, le volume de données peut être supérieur à celui du système transactionnel du fait de l'historisation (cf paragraphe précédent sur les données historisées) et du positionnement transversal (cf paragraphe précédent).
Ce niveau d'analyse permet de réaliser des comparaisons avec les périodes antérieures.

B. Données agrégées

Elles correspondent à des éléments d'analyse représentatifs des besoins utilisateurs.
Elles constituent déjà un résultat d'analyse et une synthèse de l'information contenue dans le système décisionnel, et doivent être facilement accessibles et compréhensibles. Les structures multidimensionnelles facilitent l'accès et la navigation au sein des données.
La définition complète de l'information doit être mise à la disposition de l'utilisateur pour une bonne compréhension. Dans le cas d'un agrégat, l'information est composée du contenu présenté (moyenne des ventes, …) et de l'unité (par mois, par produit,…).

C. Les métadonnées

Un suivi administratif est nécessaire pour contrôler le contenu du Data Warehouse lors de son alimentation, sa mise à jour et son utilisation. En effet, l'alimentation est réalisée depuis des sources de nature diverses et les processus de création des informations décisionnelles sont complexes. Pour pallier à ces problèmes, il est nécessaire de se doter d'un dictionnaire des données unique composé des métadonnées.
Les métadonnées sont des « données sur les données » et définissent les informations relatives à l'entrepôt et aux processus associés.
Le dictionnaire des données ou référentiel est au cœur du Data Warehouse et décrit les informations nécessaires à l'administration et à la gestion de l'entrepôt de données.
Les métadonnées gèrent le contrôle de l'information en assurant sa fiabilité, sa cohérence sa réplication, sa distribution, leur historisation et la détermination du périmètre de calcul des données.

	Sémantique des données de l'entrepôt
	Localisation de la donnée dans les systèmes de production
Back room	Procédures de chargement
	Historique des mises à jour
	Règles de calcul et processus de transformation des données
Front room	Utilisation de la donnée dans différentes applications
	Profil / rôle des utilisateurs de l'entrepôt

Les informations que fournissent les métadonnées

D. Données historisées

Comme évoqué précédemment, les données au sein du Data Warehouse comportent toutes une date à laquelle se rattache une combinaison d'attributs et des données chiffrées : les ratios.

Sachant que l'ensemble des données sont conservées, on peut pour les données anciennes les rassembler afin qu'elles soient moins volumineuses. On pourra stocker ces données sur des supports moins coûteux et consultables sur demande.

IV. Les Data Marts ou magasin de données

Le risque d'échec du Data Warehouse est sa difficulté d'utilisation ; l'attention est portée sur sa problématique technique. En revanche, le Data Mart minimise la complexité informatique et se concentre sur les besoins utilisateurs.

Le Data Mart est une base de données qui se greffe au Data Warehouse et qui reprend les données qu'il contient. Son approche est plus simple car elle est centrée sur un objectif commun à un groupe relativement limité d'utilisateurs.

Les données du Data Warehouse peuvent être dupliquées dans plusieurs Data Marts, ce qui représente une différence fondamentale avec l'entrepôt. Le magasin de données ne contient que les données dont les utilisateurs ont besoin ; il devient alors très simple d'exploiter son contenu.

	Data Warehouse	Data Mart
Cible utilisateur	Toute l'entreprise	Département
Implication service informatique	Elevée	faible ou moyen
Base de données d'entreprise	SQL type serveur	SQL milieu de gamme, bases multidimensionnelles
Modèles de données	A l'échelle de l'entreprise	Département
Champ applicatif	molti sujets, neutre	quelques sujets, spécifique
sources de données	multiples	quelques unes
stockage	Bases de données	Plusieures bases distribuées
Taille	De centaines de Go à des dixaines de To	Centaines de Go

V. Les cubes

Le cube est une modélisation d'une base de données multidimensionnelle. Cela permet aux utilisateurs de l'entrepôt de données de pouvoir trouver, extraire et évaluer les données dont ils ont besoin sans l'aide des informaticiens.

Chaque coté du cube dispose d'une dimension (exemple tiré du schéma ci-dessous : pays, temps, produits) et les caractéristiques correspondantes (pour pays : France, Espagne, Allemagne, de même pour chaque dimension).

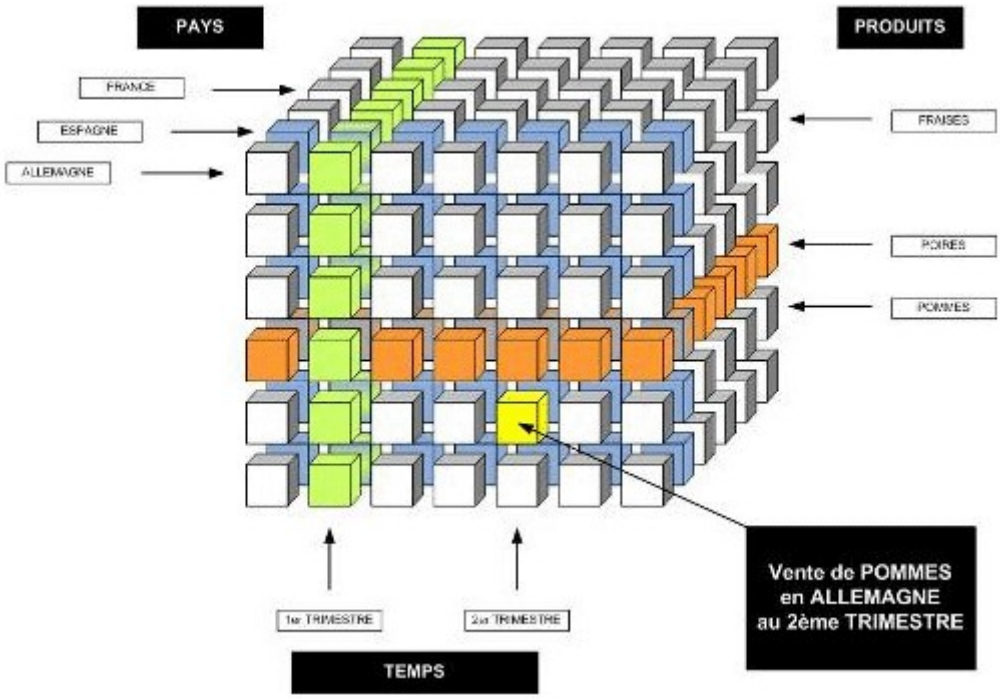


Schéma cube – source Supinfo

Au sein du cube, l'intersection des différentes dimensions (ici : intersection du produit pomme, pays Allemagne et temps second trimestre) détermine les données chiffrées correspondantes.

Au sein d'un dimension, les éléments sont hiérarchisés, c'est ce que l'on appelle la granularité : par exemple, la dimension temps peut comporter : année, mois, jour, heure.

L'utilisateur peut régler l'intersection qu'il recherche en modifiant le degré de précision des dimensions analysées. Il s'agit de la navigation OLAP (On Line Analytical Processing) que l'on peut définir par l'assistance offerte aux usagers dans leur analyse en leur facilitant l'exploration des données et en leur donnant la possibilité de le faire rapidement

VI. Les agrégats

Un agrégat est obtenu en effectuant une opération (somme, maximum, minimum, le compte) de plusieurs données détaillées[28] [29]

Le pré-calcul et le stockage du résultat de l'agrégation permettent d'améliorer la rapidité du traitement des requêtes[30].

La grandeur du cube est représentée par la quantité d'agrégations pré-calculées et stockées dans la base de données [29] Ces agrégations étant volumineuses à stocker dans l'entrepôt de données, le cube peut rapidement devenir très volumineux.

Agrégations calculées « à la volée »

Des spécialistes estiment qu'il n'est pas nécessaire de calculer à l'avance et stocker les agrégations[31] [32]. Il y a un débat sur l'espace de stockage que nécessitent les agrégations pré-calculées par rapport aux données détaillées. Il y aurait un rapport de deux cent à cinq cent selon Shukla alors que le CRG de l'Université Laval (Canada) annonce un rapport de trente[33].

---

[28] Lenz, H. et Shoshani, *A Summarizability in OLAP and Statistical Data Bases*, 1997, , pp.39-48

[29] R. Kimball et M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, 2002

[30] R. Kimball et M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, 2002

[31] Pedersen, T. B. et Nektaria, T., *Pre-aggregation in Spatial Data Warehouses* , Advances in Spatial and Temporal Databases, 7th International Symposium, 2001, pp.460-478

[32] Hung, E., Cheung, D. W. et B.Kao, *Optimization in Data Cube System Design* , Journal of Intelligent Information Systems Vol.23 No.1, 2004, pp.17-45

[33] The OLAP Report, Database explosion publié le 11 Février 2005, <http://www.olapreport.com>, Mai 2007

2eme partie L'intégration des données structurées dans le Data Warehouse

chp 1. Principe général

Comme vu jusqu'ici, les entreprises utilisent actuellement divers systèmes sources pour gérer leur activité et leur processus. Les systèmes décisionnels permettent d'avoir la meilleure vision possible de leur activité, cela permet aux dirigeants d'avoir les données adéquates pour prendre les bonnes décisions.
Ces systèmes décisionnels puisent leurs informations au sein du Data Warehouse.
Il est donc nécessaire d'alimenter ces entrepôts de données avec deux types de données : les structurées issues des progiciels et les non structurées issues d'Internet : c'est la phase d'intégration des données à destination du Data Warehouse.

I. Définition

« Un système d'intégration de données fournit une vue unifiée des données provenant de sources multiples et hétérogènes. Il permet d'accéder à des données au travers d'une interface uniforme, sans se soucier de leur structure ni de leur localisation. »

Les données sources sont disséminées dans des systèmes divers qui disposent de leurs propres structures format et définition. De plus, ces données ne sont pas totalement fiables et peuvent être en partie erronées, incomplètes, contradictoires, incohérentes,…. [34] Dès lors, ces données issues de sources hétérogènes doivent subir un traitement avant d'être intégrées dans l'entrepôt de données [35]
Les données structurées sont des informations organisées selon une logique « compréhensible » par les systèmes d'information. On désigne ainsi les bases de données, les fichiers plats,…

[34] Moss, *Data Cleaning : A dichotomy of data warehousing ?,* Data Management Review, <http://www.dmreview.com>, mai 2007

[35] B. Kathy, *Converting data for warehouses*, DBMS On Lin, <http;//www.dbmsmag.com/9706d15.html, 1997>

II. Processus

Pour aborder en détail les différents processus de l'intégration des données, nous nous appuierons sur le projet WHIPS (Warehouse Information Project at Stanford) de l'Université de Stanford[36]

A. Extraction

1 Découverte des données

Il s'agit de localiser dans le système opérationnel les données qu'il est nécessaire de prélever. Cette étape est importante dans la mesure où elle va déterminer le niveau de finesse des analyses du Data Warehouse : il s'agit de la granularité

Prendre un trop grand nombre de données complexifiera les étapes d'intégration des données, mobilisera d'autant plus de capacités système et d'espace de stockage pour l'entrepôt de données.

A l'inverse, diminuer le nombre d'informations peut limiter, voire fausser les analyses de l'entrepôt. Cette problématique a été abordée précédemment dans cette étude dans la partie ….

Ainsi, cette étape est déterminante et résulte d'un équilibre entre ces deux éléments.

2 Extraction des données

Les données sont en général hétérogènes du fait de la diversité des systèmes de données sources : il peut s'agir de fichiers plats, de bases de données relationnelles (cf partie 1 du mémoire) . Les données sont donc complexes (organisation transactionnelle) et diffuses (les systèmes sources sont multiples, et géographiquement éloignés).

Toutes ces données doivent être prélevées sans pour autant perturber les systèmes sources.

3 Rôle de l'adaptateur

Les données sources doivent être prélevées sans pour autant perturber les systèmes sources. Pour ce faire, l'adaptateur transforme les données sources en une représentation intermédiaire. Ainsi, les données sont dupliquées et les opérations décrites dans la suite de cette partie n'auront pas d'effet sur les informations de base.

---

[36] J. Widom, *Research problems in data warehousing*, 4th International conference on Information and Knowledge Management, 1995, pp. 25-30

4 Rôle du moniteur

Lorsque les données sont chargées pour la première fois, la démarche est globale (cf partie 2 du mémoire : l'intégration en mode Full), les actualisations sont ensuite incrémentales (cf partie 2 : le mode Delta). Il faut alors déterminer quelles sont les données apparues après la dernière intégration, c'est le moniteur qui s'en charge.

Ce sujet a suscité beaucoup de débats chez les spécialistes du sujet et plusieurs théories sont défendues :
Jennifer Widom [37] a défini des clases pour les sources de données:

Les sources coopératives sont dotées de règles actives ou notifications qui permettent de spécifier les changements de manière automatique.

Les sources avec journal décrivent leur activité, ce qui permet de facilement identifier les modifications à intégrer.

Les sources permettant des requêtes à la demande : le moniteur interroge cette source régulièrement afin de détecter les modifications concernées.

Les sources permettant les photographies : on constate les changements au travers des différences entre leurs photographies consécutives.

Il est nécessaire de paramétrer pour chaque type de source un ensemble adaptateur et moniteur. Pour limiter l'investissement, on peut utiliser des générateurs automatisés comme l'illustre une publication sur le sujet[38]
Bill Inmon rajoutera d'autres éléments facilitant la détection des changements et précisera que le choix de l'indicateur dépend de la nature des informations à « marquer »[39]

B. Transformation

Le système décisionnel se doit de fournir des informations fiables car celles-ci serviront de base pour prendre les décisions stratégiques de l'entreprise. Cela repose sur la qualité des données au sein de l'entrepôt et c'est l'étape de transformation qui se charge de la garantir.

---

[37] J. Widom, *Research problems in data warehousing*, 4th International conference on Information and Knowledge Management, 1995, pp. 25-30

[38] A. Sahuguet, F. Azavant, *Building light weight wrappers for legacy web Data Sources using W4F* , 1999, pp 738– 741

[39] B. Inmon, *Loading Data into Warehouse*, 2000, pp 6-17

Le transformateur se charge de nettoyer les données afin de les rendre compatibles avec la granularité et le schéma du Data Warehouse[40] [41].
Ce processus a été étudié dans un contexte multi base de données dont voici les principales caractéristiques [42] :
Il s'agit tout d'abord de recenser les entités dont la sémantique est similaire. Ainsi, une même entité peut être représentée de manière différente dans les diverses sources et réciproquement. Cette tâche est assez délicate à gérer, c'est pourquoi on fait appel à l'intervention humaine[43].

Ensuite, il faut identifier et résoudre les éventuels conflits entre les entités. La cardinalité des champs sources et cibles doit être respectée : une adresse complète peut être décomposée en rue, ville et code postal par exemple.
Les données synonymes doivent être rassemblées : Par exemple, la désignation femme peut dans les systèmes sources être désignée par les données suivantes : « F », « Femme », « 1 » en cas de codification numérique. L'utilisateur définit que ces appellations correspondent à la donnée « femme ». Le précurseur dans ce domaine est ETI (Evolutionaru Technologies International), première application d'ETL.

Enfin, il faut s'assurer de la cohérence des données en supprimant les doublons grâce à des filtres prédéfinis : les données manquantes ou incohérentes sont ainsi corrigées. L'exemple le plus connu est celui d'une agence de location de véhicules de Boston : du fait des contraintes techniques, les salariés considéraient les clients étrangers comme résidents de Boston. Les analyses des données ont révélé une proportion anormale de clients de Boston, la société a dû revoir le process source.

Tout comme lors de l'extraction, la transformation nécessite au système de disposer d'intégrateurs spécifiques. Les règles définies (cf paragraphe précédent sur le synonymie)

---

[40] R. Kimball, *The Data Warehouse Toolkit*, 1996

[41] H. Galhardas, D. Florescu, D. Shasha et E. Simon, *Ajax: an extensible data cleaning tool* , International conference on management of data, 2000

[42] M. Bright et A. Hursen, *A taxonomy and current issues in multidatabase system*, 1992

[43] Mauricio A. Hernandez et Salvatore J. Stolfo, *Real world data is dirty : data cleansing and the merge/purge problem*, Data Mining and Knowledge discovery : an international journal, p 9-37 - 1998

permettent de générer automatiquement des intégrateurs, ce qui diminuera le coût de l'opération.
Cet aspect a été développé dans le cadre du projet H20 d'une université Américaine [44].

C. Chargement ou rafraîchissement

Une fois que l'intégrateur a fusionné les données issues des adaptateurs, les données intégrées peuvent être chargées dans le Data Warehouse. Cette étape s'appelle le chargement ou le rafraîchissement.
Lorsque le chargement des données s'opère dans le cadre de la création de l'entrepôt, les moniteurs envoient une copie intégrale des données sources et l'intégrateur se chargera de charger l'entrepôt.
En revanche, l'actualisation des données peut être réalisée selon deux techniques différentes : la reconstruction et l'incrémental.
La reconstruction permet périodiquement de remplacer l'ensemble des données de l'entrepôt. Ce procédé est équivalent au premier évoqué : on écrase les données précédentes par les nouvelles.
Cela pose des problèmes de ressources système, c'est pourquoi on utilise plus habituellement le chargement incrémental. Le moniteur décèle les données qui ont évolué depuis la précédente actualisation de l'entrepôt. Le chargement peut alors s'effectuer en temps réel ou périodiquement. Des travaux comme le WHIPS [45] [46] ou le SIRIUS.(Supporting Incremental Refreshment of information Warehouse) analysent le chargement incrémental [47].
Le chargement incrémental pour les bases relationnelles a fait l'objet de nombreuses études, ce n'est pas le cas pour les systèmes multidimensionnels.
Le chargement incrémental dans les systèmes multidimensionnels a été abordé par Hyperion [48] mais le sujet n'a pas été beaucoup étudié par les spécialistes. Dans les bases relationnelles,

---

[44] O. Zaiane, J. Han, Z. Li, S. H. Chee et J. Chiang, *Multimedia-Miner : A system prototype for multimedia Data Mining*, Conference on Management of Data, pp 581-583, 1998

[45] J. Hammer, H. Garcia-Molina, J. Widom, W. Labio et Y. Zhuge, *The Stanford Data Warehousing Project* , Data Engineering Bulletin, pp 41-48, 1995

[46] W. Labio, Y. Zhuge, J. L. Wiener, H. Gupta, H. Garcia Molina et J. Widom – *The WHIPS prototype for Data Warehouse creation and maintenance* – International conference on management of data, pp 557-559, 1997

[47] A. Vavouras, S. Gatziu et K. R. Ditrich, *The SIRUIS approach for refreshing Data Warehouses incrementally*, pp. 80-96, 1999

[48] Livre blanc Hyperion, *Essbase Partitioning option*, 1999

les spécialistes ont confirmé que les changements des relations de base ne modifiaient qu'une partie de la vue des données [49] [50] [51]

Cette étape mobilise l'entrepôt qui ne peut plus restituer les données aux applications auxquelles il est associé. Il est donc important que les phases de chargement soient de courte durée, quitte à répartir la tâche en plusieurs parties. C'est l'objet du projet Européen DWQ (Foundation of Data Warehouse Quality) [52] [53]

D. Conclusion

Cette partie a permis de présenter les principes de base du phénomène d'intégration des données dans le Data Warehouse. Chaque source de donnée est donc associée à un adaptateur/ moniteur qui se charge de découvrir les modifications et de transformer ces données dans un format homogène. Les données sont ensuite exploitées par l'intégrateur qui les fusionne, puis chargées dans l'entrepôt.

III. Les différentes générations d'ETL

Le livre blanc d'Acta recense cinq générations d'outils d'ETL.
La première génération a vu apparaître des ETL qui générait du code Cobol pour assurer la transition entre les données sources et celles du Data Warehouse. Même si ces outils ont permis de faciliter la tâche, ils péchaient dans l'automatisation des environnements d'exécution. L'intervention humaine était nécessaire pour assurer l'intégration des données.
La mise en place d'un ETL générateur de code nécessitait un budget prohibitif.

---

[49] A. Gupta et I. S. Mumick, Maintenance of *materialized views : problems, techniques and applications* , Data engineering bulletin, pp 3-18, 1995

[50] I. S. Mumick, D. Quass et B. S. Mumick, *Maintenance of data cubes and summary tables in a warehouse* , Proceedings of the ACM SIGMOD conference on management of data, pp 100-111, 1997

[51] T. Griffin, L. Libkin et H. Trickey, *An improved Algorithm for the incremental recomputation of active relational expressions*, Knowledge and data engineering bulletin, pp 3-18, 1997

[52] H. Garcia Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, J. D. Ullman, V. Vassalos et J. Widom, *The TSIMIS approach to mediation : data models and languages*, Journal of intelligent Information system, 1997

[53] D. Calvanese, G de Giacome, M Lenzerini, D Nardi et RE. Rosati, *Source integration in Data Warehousing* , Proceedings of the 9th international workshop on database and expert systems application, pp. 192-197, 1998

La seconde génération a bénéficié de l'avènement des technologies client / serveur pour les applications propriétaires pour proposer un nouveau langage : le SQL. Cependant, l'apparition de solutions intégrées comme les ERP a créé des difficultés pour extraire les données sources.

La version suivante a permis de pallier à ce problème en proposant des adaptateurs spécifiques. La diversité des données sources est mieux supportée : les données non structurées (ie extraites d'Internet) commencent à être intégrées et des adaptateurs spécifiques permettent d'intégrer les données issues des ERP.

Dans un contexte où les délais pour prendre des décisions se raccourcissent, le quatrième type permet d'intégrer les données au fil de l'eau (intégration instantanée) et en batch (en différé). Cela permet à l'entreprise d'avoir des résultats en temps réel et aux dirigeants de prendre plus rapidement les décisions adéquates. De plus, les messages brokers apparaissent, cela permet en cas de suractivité des ressources, de mettre les données à intégrer dans une file d'attente. Le langage XML fait aussi son apparition et permet de réaliser des modifications qui seront immédiatement prises en compte par l'outil.

La dernière génération dispose d'API (Application Programming Interface) qui permettent aux données échangées de pouvoir être dirigées vers des cibles multiples. Ensuite, on a réussi à standardiser l'utilisation des fonctionnalités de la précédente génération. On utilise une intégration au fil de l'eau pour les données opérationnelles car les contraintes de l'activité l'exigent et car ces données ne représentent pas un volume trop important. Les données analytiques sont quant à elles intégrées en batch car le contexte est opposé au précédent.

IV. Les différentes approches d'ETL

A. ETL « indépendants »

Il s'agit des applications dont nous avons parlé précédemment, ils sont indépendants des systèmes sources et des systèmes décisionnels. Cette indépendance permet d'assurer une capacité d'adaptation aux divers systèmes d'information en présence.
Des solutions open sources existent, Talend est le « dernier né » et semble, d'après les premiers retours, très fiable et très intuitif. Ceci permet d'assurer une véritable concurrence au sein des ETL, ce qui pousse les éditeurs d'applications à améliorer leurs solutions.

B. ETL intégrés

Ces ETL fonctionnement comme des générateurs de code : la solution de gestion des bases de données (SGBD) se charge d'effecteur les transformations et les agrégations. Tout cela est traduit en un code spécifique pour assurer l'alimentation de l'entrepôt de données.
Cependant, cette approche est quelque peu réductrice : l'ensemble des fonctionnalités d'un ETL classique n'y sont pas reprises. Le codage manuel est toujours nécessaire et il est difficile de faire évoluer ces solutions lors de l'apparition d'une nouvelle version du SGBD.
Le principal inconvénient de ce modèle est la dépendance du client : le code généré pour l'intégration ne peut être interprété que par la « marque » du Data Warehouse. On peut y voir une sorte d'assujettissement des entreprises vis à vis de l'éditeur de SGBD.

V. L'ETL est il un EAI ?

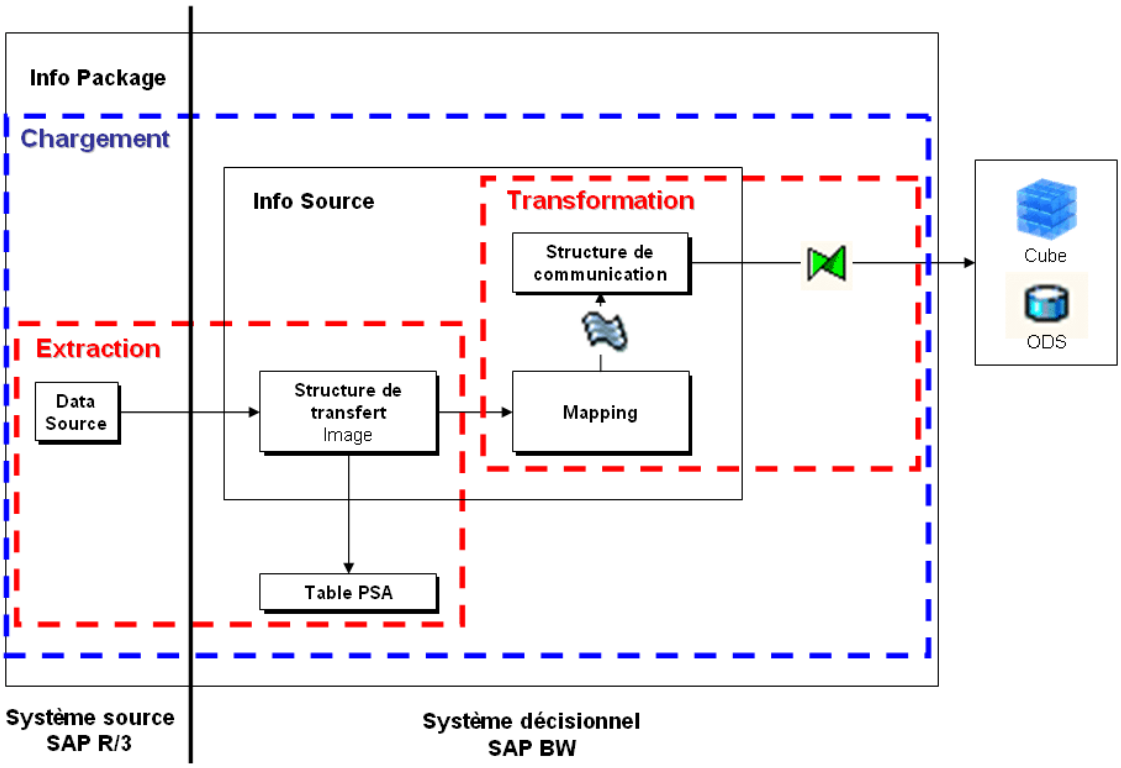
Dans l'absolu un ETL se charge de l'intégration des données suite à leur sélection et leur transformation. Ce processus permet d'échanger des grands volumes de données en mode batch (en différé).
L'EAI (Entreprise Application Integration) s'occupe de relier des applications entre elles en leur permettant d'échanger des flux .
Comme vu précédemment, les ETL ont évolué et proposent actuellement d'échanger les données en temps réel, disposent les données en files d'attente quand nécessaire et permettent aux données d'êtres intégrées dans plusieurs cibles en même temps. Dès lors, l'ETL dispose de certaines caractéristiques de l'EAI, on peut alors s'interroger sur les différences entre les deux mécanismes d'échange.
Il s'agit bien d'application différentes dans la mesure où l'ETL se limite à échanger des données alors que l'EAI permet d'échanger des processus entre les applications.

chp 2. L'intégration avec SAP BW : l'exemple de GEFCO

Pour illustrer ce mémoire, voici un projet sur lequel je travaille chez GEFCO : la répartition des immobilisations dans le cadre de la détermination du ROCE[54].

Cela illustre ce mémoire dans la mesure où les données sont prélevées dans le système source : l'ERP SAP/R3 et intégrées dans le système décisionnel SAP BW.

I. Fonctionnement général de l'extraction des données sous SAP BW



A. Data Source

La Data Source sélectionne, dans le système source SAP R/3, les données utiles à prélever. Cette compilation permet de modéliser les données que l'on veut extraire, ces dernières peuvent provenir d'une table, d'une vue (un ensemble de tables) ou d'un Infoset (une jointure de tables). Il n'y a pas de stockage physique des données au sein de la Data Source : celle-ci permet de modéliser la structure des données que l'on veut extraire.

[54] ROCE : rentabilité de l'actif économique, il correspond à la rentabilité des actifs quelque soit déduction faite de son financement

B. Structure de transfert

La structure de transfert se trouve dans la partie décisionnelle SAP BW et permet de recenser les données que l'on souhaite extraire du système source. Il est donc séparé de la Data Source par un mécanisme de réplication et les données sont modélisées par le biais d'images.

C. Table PSA

La modélisation des données issues de la structure de transfert peut être matérialisée dans les tables PSA. Cela permet de pallier aux éventuels problèmes qui pourraient surgir dans la suite du mécanisme d'intégration des données.

D. Cube ou ODS.

Il s'agit des organes de stockage cibles, c'est-à-dire le support de destination final de l'intégration des données étudiée par ce mémoire. Ces bases de données multidimensionnelles seront ensuite utilisées par des applications de Business Intelligence (cf partie 1 chapitre 2 l'aval du Data Warehouse)

E. Structure de communication

Tout comme les structures de transfert par rapport aux Data Sources, les structures de communication sont la modélisation des données du support de stockage cible (cube ou ODS). Plusieurs Info Sources peuvent alimenter une même base de données cible, les règles de mise à jour s'occupent de cela.

F. Mapping

Cette étape se trouve entre les structures de transfert et les structures de communication. Il permet de relier les données des deux structures et de convertir les données sources en données cibles grâce aux règles de transfert.

La phase d'extraction est réalisée par la Data Source ainsi que par la structure de transfert. La transformation s'opère au travers du mapping, de la structure de communication ainsi qu'avec les règles de mise à jour. Le chargement des données est réalisé par l'info package qui compte toutes les structures énumérées ci-dessus, à l'exception des structures de stockage cibles.

La phase de transformation peut avoir lieu à divers moments : lors de l'agrégation des données

II. Lien entre l'extraction sous SAP BW et le modèle ETL

Dans le cadre de la phase d'extraction, le rôle de l'adaptateur dans la phase de découverte des données est pris en charge par les Data Sources dans la mesure où elles sélectionnent les données à intégrer.

III. Intégration des données dans le cadre du calcul du ROCE

Le projet sur lequel j'interviens en stage consiste à répartir les amortissements des actifs immobilisés par Unité Opérationnelle, les autres données sont fournies par les dirigeants souhaitant réaliser le calcul.

Dès lors, il s'agit de répartir les immobilisations dans les unités opérationnelles en fonction d'un algorythme de répartition. Comme le présente le schéma ci dessous, les données sources liées aux immobilisations sont présentes dans le module FI-AA (Asset Acounting gestion des immobilisations) de SAP, les cycles de répartition sont eux présents dans le Module CO (Finance comptabilité). La mise en relation de ces deux types de données source permet de déterminer les résultats cibles.

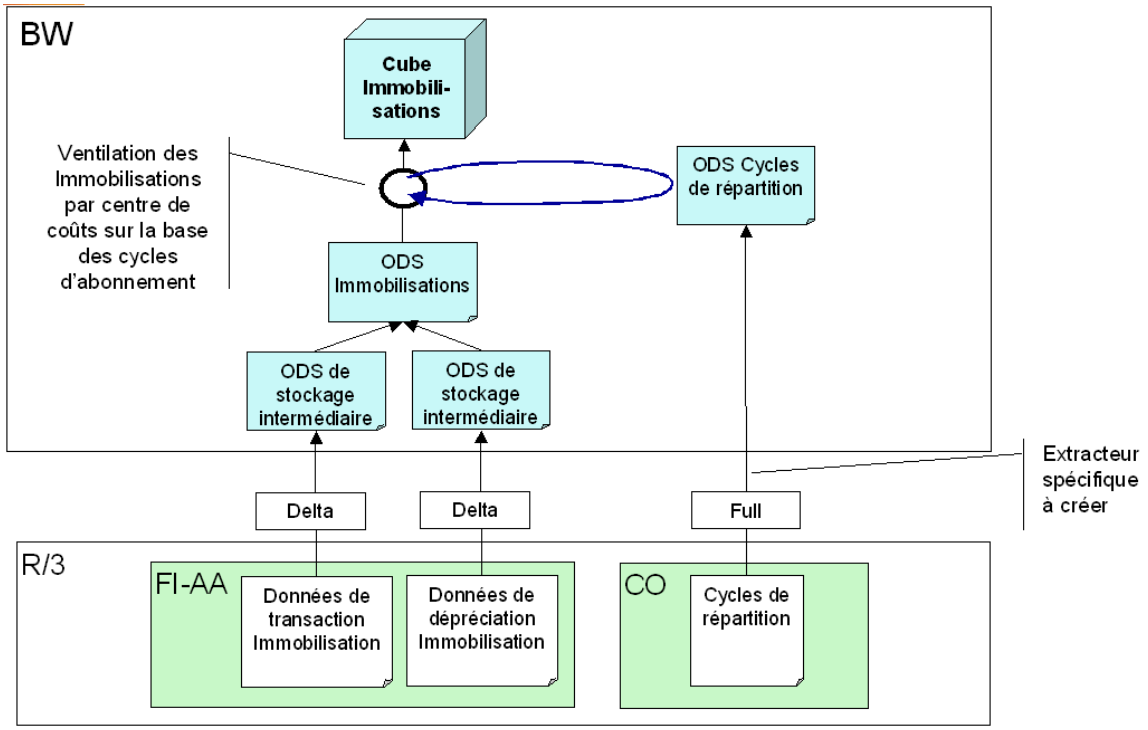


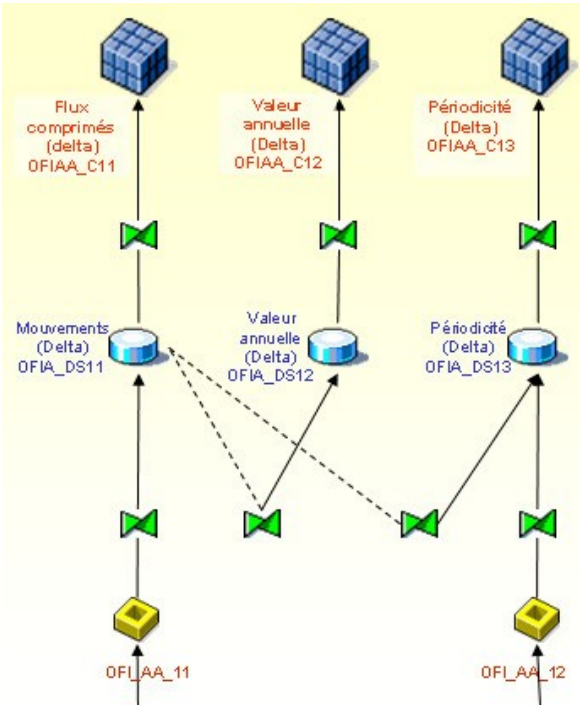
Schéma fonctionnel de l'intégration des données
et de leur répartition pour le calcul du ROCE.

En effet, le montant global des amortissements est calculé par les données issues des immobilisations, celles liées aux cycles de répartition permettront par le biais d'un algorithme de répartir les montants des amortissements en fonction des Unités Opérationnelles.

Le rafraîchissement des données se fait en delta pour les données liées aux immobilisations. Cela peut s'opérer de deux manières :
En mode Time Stamp, toutes les données sources disposent d'un marqueur qui indique leur date de modification. Lors du rafraîchissement, le système ajoute les données désirées dont la date est postérieure à celle de la dernière mise à jour. Les données sont ainsi envoyées par le bais des modélisations aux supports physiques : les bases PSA et les bases cibles.
En mode « LE Cockpit », les données désirées qui ont été modifiées depuis la dernière actualisation sont mises dans un « delta queue », c'est-à-dire une liste des données à intégrer. Lors de l'intégration, seules les données présentes dans cette base sont importées.

Les ODS «immobilisations » et « cycles de répartition » sont situés dans les structures de communication (cf schéma précédent) et subissent des règles de mise à jour, lesquelles sont régies par un algorithme.

Cet algorithme permet d'orienter les flux de données en fonction de critères spécifiques. Dans le cas du ROCE, les immobilisations (les données) sont selon leurs caractéristiques en terme d'Unité Opérationnelle, orienté vers un processus de traitement différent.



Dès lors, les amortissements sont répartis

Ainsi, on constate que le mécanisme d'intégration des données dans un système intégré (SAP BW) est différent de celui d'un système indépendant. SAP n'utilise pas le principe d'une table à part entière qui copie les tables des données sources : il utilise une Data Source qui ne fait que modéliser ces données. Dès lors, le système est confronté aux problèmes d'utilisations de ressources métier (cf partie 1….), c'est pourquoi l'intégration des données a lieu à des moments où l'activité est moindre : la nuit par exemple.
Les phases de transformation ont lieu à plusieurs moments : lors du mapping qui transforme le format des données sources en données cibles (cf partie 2 chapitre 1 : phase de transformation), lors des règles de mise à jour qui permettent de regrouper plusieurs Info Sources pour un même support de stockage cible. Enfin, les supports cibles peuvent eux-même réaliser une transformation dans certains cas : il peut s'agir d'opérations d'agrégation des données selon des critères prédéfinis.
La particularité de ce modèle est que la phase de chargement regroupe toutes les phases de l'intégration des données. En effet, le processus est une succession de modélisation de données, les données ne sont réellement stockées que dans les supports cibles, toutes les modifications antérieures sont effectuées par le biais de modélisation. C'est pourquoi cette phase de chargement est si globale.

Conclusion

L'intégration permet d'extraire des données dans des sources hétérogènes et de les rendre homogènes afin de pouvoir les intégrer dans un entrepôt de données. Ces deux supports de stockage étant différents, il est nécessaire d'appliquer à ces données des transformations conséquentes. Pour rester fiable, le Data Warehouse nécessite d'être régulièrement remis à jour avec des données récentes, c'est l'objectif de la phase de chargement qui l'alimente en temps réel ou en différé selon les besoins de l'entreprise.

Cependant, les principaux acteurs du secteur se font racheter au fur et à mesure par les applications intégrées du type Business Objects, Oracle,… Dès lors, on peut s'interroger sur l'avenir des solutions indépendantes d'intégration des données. L'attention sera alors portée sur les solutions décisionnelles intégrées (voire même les systèmes d'ERP intégrant une partie décisionnelle comme c'est le cas chez SAP). Ceci est dangereux dans la mesure où le code généré pour l'intégration des données est spécifique à la marque du système d'information. Il devient très difficile d'intégrer une solution d'analyse d'un autre éditeur. Les clients deviennent alors assujettis au système décisionnel choisi initialement.

Cet important et dangereux risque de dépendance des entreprises par rapport à leurs fournisseurs crée, d'ailleurs, depuis peu, l'apparition de nouveaux systèmes comme Talend. Leur objectif est en effet de simplifier l'intégration et donc de permettre à d'autres systèmes d'analyser les données de l'entrepôt.

On peut, cependant se poser la question de l'avenir d'une solution du monde du libre face aux éditeurs intégrés, à moins de déplacer le débat sur la Web Ontologie. En effet, les données non structurées issues d'Internet revêtent une importance grandissante pour les entrepôts de données et l'approche de l'Open Source paraît la plus à même d'être performante sur ce secteur.

Les sources

Toutes les documents au format numérique que j'ai utilisé dans le cadre de ce mémoire sont consultables sur <http://esnips.com/web/memoire-BI>

Outre les sources citées en bas de page, voici le liste des sources qui m'ont permis de mener à bien ce travail d'investigation :

Bibliographie :

J. F. Goglin, *La construction du Data Warehouse*, Editions Lavoisier, 2001

R. Kimball, L. Reeves, M. Ross et W. Thornthwaite, *Le Data Warehouse – Guide de conduite de projet*, Editions Eyrolles, 2007-06-11

J.M. Franco, *Le Data Warehouse, Le Data Mining*, Editions Eyrolles, 1997

J.F. Goglin, *La cohabitation électronique*, Editions Lavoisier, 2005

Caseau Yves, *Urbanisation et BPM: Le point de vue d'un DSI*, Editions Dunod, 2005

Articles de recherche Publiés

G. Zettl, Institut national de statistique Autrichien, *Entrepôt de données et INS* , 9e séminaire CEIES – Innovation dans la founiture et la production de statistiques, pp10-14

A. Cali, D. Lembo, M. Lenzerini, R. Rosati, *Multidimensional Databases : Problems And solutions*, chapitre 12, *Source integration for Data Warehousing* , Editions Rafanelli, 2003, pp.361-392

Laboratoire ERIC Lyon, Pôle Bases de Données Décisionnelles, *Plateforme d'entreposage XML de données complexes*, 2006

M. Alia, A. Lefebvre, C. Collet, France telecom R&D, *Système d'intégration des données : une approche à composants* , Revue des Sciences et Technologies de l'information, vol 9, 2004

J. Widom, *Research problems in Data Warehousing*, 1995

Revues

Solutions Magazine, *Dossier Business Intelligence*, pp10-24, 05/2005, <http://www.solutions-magazine.be>

F. Bentayeb, O. Boussaïd, J. Darmont, S. Loudcher, *Entrepôts de Données et l'Analyse en ligne*, Revues des Nouvelles Technologies de l'Information (RNTI), Editions Cépaduès 2005

Auteur collectif, *RNTI N° 1 Entreposage et fouille de données*, Editions Cépaduès, 2003

M. Noirhomme-Fraiture, G. Venturini , *RNTi E9 - Extraction et gestion des connaissances* , Editions Cépaduès, 2007-06-11

Articles de recherche sur Internet

M. Leitzelman, H Dou, Essai de typologie des Systèmes d'Informations, *Les systèmes d'information décisionnels*, 1998, <http://isdm.univ-tln.fr/PDF/isdm2/isdm2a15_leitzelman.pdf>

W.H. Inmon, *Loading Data into the Warehouse*, 2000, <http://www.inmoncif.com/registration/whitepapers/ttload-1.pdf>

W. H. Inmon, *Metadata in the Data Warehouse*, 2000, <http://www.inmoncif.com/registration/whitepapers/ttmeta-1.pdf>

W. H. Inmon, *Building the Data Warehouse: Getting started*, 2000, <http://www.inmoncif.com/registration/whitepapers/ttbuild-1.pdf>

Supports de formation

M. Fbone, *support de cours ETL*, Université de Vannes

G. Hébrial, Telecom Paris, *Systèmes d'information décisionnels*, 2005-2006

E. Grislin-Le Strugeon, Université de Valenciennes, ISTV, D. Donsez, Université Joseph Fourier, *Systèmes d'information décisionnels (Data Warehouse, Data Mining),* 1006, 2006

S. Oberlechner, IUT de Quimper, *Le décisionnel d'entreprise*

M. Raphalen, Université de Bretagne Sud, DESS ASIR, *Systèmes d'informations décisionnels*, 2002

M. Ester, *Data Warehousing*,

C. Friguet, M. Cousseau, Université de Bretagne Sud, *Etude comparative des différents outils d'ETL du marché,* 2005

Conservatoire National des Arts et Metiers de Lille, *Data Warehouse et Data Mining*, 1998

Travaux Universitaires

S Lebouche, Université de Nancy2, *De l'informatique décisionnelle à la multidimensionnalité en vue de présenter les fondements élémentaires du tableau de bord électronique* , Cahier de recherche n°1999-02

E. Serres, Thèse professionnelles, *Comment piloter l'entreprise grâce au datawarehouse* , 2004

O. Teste, Thèse, *Modélisation et manipulation d'entrepôts de données complexes et historisées*, 2000

M. Lambert, Thèse, Université de Laval, Canada, *Développement d'une approche pour l'analyse Solap en temps réel*, 2006

P. Bonnet, Université de Savoie, *Prise en compte des sources de données indisponibles dans les systèmes de médiation*, 1999

E. I. Benitez Guerrero, Thèse, Université de Grenoble, *Infrastructure adaptable pour l'évolution des entrepôts de données*, 2002

Autres supports :

Smile, Livre blanc, *Décisionnel – Solutions Open Source*, 2006

Sunopsis, Livre Blanc, *The future of Data Integration Technologies*, 2004

Webogrpahie

Business Intelligence :

Pierre Audoin Consultants, PAC online, <http://www.pac-online.com/pac/pac/live/pac_world/pac/hotnews_list/index.html>

Oxio, Data Intelligence, le 1er ETL 100% Web <http://www.oxio.fr/>

SupInfo, Tous les projets des élèves ingénieurs de Supinfo, <http://www.supinfo-projects.com>

Indexel, Les multiples usages du décisionnel, <http://www.indexel.net/1_6_3615__3_/6/23/1/Les_multiples_usages_du_decisionnel.htm>

Data Warehouse

Information Technology, Data Warehouse Glossary, <http://it.csumb.edu/departments/data/glossary.html>

Intégration de données

Data Warehousing Review, Data Cleansing for Data Warehousing <http://www.dwreview.com/Articles/Data_Cleansing.html>

CXP, Data Integration Management et ETL, <http://www.cxp.fr/domaine-expertise_etl.htm>

Progisphere, EntropySoft annonce la disponibilité de sa solution EntropySoft Content ETL, le premier ETL de contenu au niveau mondial, <http://www.progisphere.com/EntropySoft-annonce-la-disponibilite-de-sa-solution-EntropySoft-Content-ETL,-le-premier-ETL-de-contenu-au-niveau-mondial_a2299.html>

Learn Data Modeling, ETL Tools, <http://www.learndatamodeling.com/etl_tools.htm>

A. Elomari, TOUT SUR LES SYSTÈMES ETL, choix d'un outil d'ETL <http://systemeetl.blogspot.com/2005/09/choix-doutil-etl.html>

01Net, IBM s'offre un ETL à 1 milliard, <http://www.01net.com/article/274557.html>

Decideo, Talend passe son ETL open source en V2, <http://www.decideo.fr/Talend-passe-son-ETL-open-source-en-V2_a1998.html>

DW Facile, Fonctionnalités ETL, <http://www.systemeetl.com/fonctionnalites_ETL.htm>

Journal du net, L'Open Source fait une percée dans les offres ETL, <http://www.journaldunet.com/solutions/0702/070221-panorama-etl/1.shtml>

Newsplex, Relier nos bases de données, <http://www.campusxml.org/news/fullstory.php/aid/932/Relier_nos_bases_de_donn%E9es.html>

Journal du net, EAI versus ETL : que choisir ? <http://www.journaldunet.com/solutions/0203/020314_eai_vs_etl.shtml>

Business Intelligence Network, David Loshin, <http://www.b-eye-network.com/channels/index.php?filter_channel=1148>

DMReview, ETL Portal - ETL white papers, books and articles <http://www.dmreview.com/portals/portal.cfm?topicId=230206>

Executive Information Systems, Papers, Briefs, and Presentations On Data Warehousing,

OLAP

Data Mining, and OLAP, <http://www.dkms.com/dwdmolappapers.htm>

Data Warehousing Review, OLAP Analysis, <http://www.dwreview.com/OLAP/index.html>

BDAS – Labbate <http://perso.enst.fr/~saglio/bdas/97/Labbate.html#Heading2>

Best Of breed

Lefebvre Software, approche Best of Breed, <http://www.lswe.com/vdocportal/portal/template/CmSP/app/Lswe/pageId/051-000007-01d/OpenIris.htm>