Bases d'information décisionnelles ou Data Warehouse

De la théorie à la pratique

Mouloud DEY
Directeur Stratégie et Nouvelles Technologies
SAS Institute

1 - Introduction

Le terme "Data Warehouse" s'est largement répandu au point de faire la une des magazines spécialisés. Les entreprises, après avoir informatisé les principales activités vitales pour leur survie au quotidien (comptabilité, gestion des stocks, facturation...) se trouvent aujourd'hui confrontées à une démultiplication et un gigantisme des systèmes d'information.

Cette notion de système d'information est pourtant impropre, car la réalité révèle avant tout de gigantesques "réservoirs de données" qui en dehors de leur fonction routinière (faire vivre l'entreprise au jour le jour) n'apportent aucune information propre à fournir une vision à plus long terme aux responsables de l'entreprise et susceptible d'assurer sa survie au-delà du quotidien (analyse des résultats, connaissance des clients, connaissance de l'environnement...).

Les experts s'accordent aujourd'hui à penser que la mise à disposition de véritables systèmes d'information utilisables pour permettre de meilleures prises de décision passe par la mise en œuvre de bases décisionnelles d'informations et de bases thématiques d'informations ("data warehouse" et "data marts") adaptées au contexte propre des utilisateurs, à leurs métiers et leur offrant une vision sur le long terme et une vision beaucoup plus large.

Ce principe de séparation nette entre les environnements de production et les environnements d'aide à la décision est aujourd'hui largement accepté à la fois par les responsables de département informatique et par les utilisateurs. Le concept est séduisant, certes, mais sa mise en œuvre mérite une attention particulière.

2 - Principes généraux

Les applications opérationnelles gèrent et produisent des données qui sont de toute évidence difficilement utilisables en dehors du contexte pour lequel elles ont été créées :

Les modes de représentation de ces données sont souvent complexes tant en structure (bases de données physiquement optimisées pour un seul et unique besoin), qu'en format. Il n'est pas rare en effet qu'une codification complexe ait été appliquée à une donnée et qu'historiquement cette même donnée ait pu être pourvue de codes complètement différents en fonction des contraintes de stockage, de temps de traitement ou simplement d'humeur ou d'habileté du programmeur.

Les sources d'information au sein d'une entreprise sont souvent multiples : données dispersées dans plusieurs systèmes, plusieurs bases de données. Certaines informations pouvant également provenir de sources externes à l'entreprise. Une même donnée peut aussi être gérée à deux endroits différents, pour des besoins différents.

Les données d'un environnement de production sont aussi très "volatiles" : une application de gestion de stock a pour objectif principal de fournir un état précis du stock mais ne renseigne pas forcément sur son écoulement ou sur la capacité des différents fournisseurs à respecter leurs engagements.

La nécessité de repenser le processus de création et de diffusion de l'information au sein de l'entreprise apparaît donc aujourd'hui comme un enjeu majeur pour les entreprises. Les informations devraient être structurées dans une base décisionnelle d'informations ("datawarehouse") possédant notamment les caractéristiques suivantes :

Elles doivent être systématiquement classées selon des thèmes (on parlera de "base de donnée orientée sujet"). Seules les données réellement nécessaires pour les utilisateurs seront extraites des environnements de production et organisées selon des thèmes prédéfinis correspondant aux métiers des utilisateurs.

Elles devront intégrer une dimension historique : les environnements opérationnels reflètent l'activité d'une entreprise à un instant donné mais intègrent mal la notion de temps. Une base décisionnelle d'informations pourra contenir des "clichés" périodiques de ces environnements pour permettre les comparaisons entre deux situations à deux périodes données et les analyses de tendances (projections et prévisions).

Elles devront être accessibles en consultation uniquement par tous les utilisateurs; chacun ayant toujours la même

source d'information. La mise à jour de ces informations n'étant ainsi que la mise à disposition d'un nouveau "cliché" ajouté aux informations existantes.

Enfin, elles devront être intégrées en un ensemble cohérent et homogène. Ceci signifie, en particulier, qu'on s'attachera à éliminer les redondances (données en double), les incohérences (une même donnée ayant deux noms, deux formats ou deux valeurs différentes pour une même période). On s'attachera également à respecter des principes d'exhaustivité (disposer de l'ensemble des informations utiles) et de structure permettant en particulier de restituer avec une performance suffisante aussi bien des informations agrégées que des informations de détail.

3. Mise en œuvre

La mise en œuvre d'une base décisionnelle d'informations implique un processus intégré en trois phases :

• une phase de "GESTION" qui permettra d'appréhender les problèmes liés à la collecte des données, au rapprochement entre les différentes sources de données et à leur transformation en informations,

• une phase d'"ORGANISATION" qui s'attachera à la recherche de la structure optimale de la base d'informations ainsi qu'à sa localisation (base centralisée ou répartie). La mise en œuvre de "métadonnées" couvrant les besoins de documentation des informations disponibles est une étape incontournable de cette phase d'organisation,

• une phase d'"EXPLOITATION" qui permettra de restituer les informations aux utilisateurs et facilitera leur efficacité à l'aide d'outils de productivité leur permettant de se consacrer à leurs métiers (et donc d'apporter à leur entreprise une réelle valeur ajoutée).

3.1. Gestion

Accès aux Données

La première chose à considérer dans le processus de mise à disposition de l'information reste l'accès aux différentes données des systèmes opérationnels (ordinateurs distincts, bases de données hétérogènes) et aux éventuelles données externes (banques de données externes, panels, statistiques officielles...) à intégrer dans la base décisionnelle d'informations.

Les outils de collecte doivent permettre une réplication des principales données par le moyen le plus efficace. Ceci se traduit par la mise en œuvre de méthodes d'accès, de passerelles ou d'interfaces de programmation (API ou Application Program Interface) vers les gestionnaires de bases de données et les progiciels permettant de puiser à la source les différentes données en respectant les contraintes d'un environnement de production (sécurité et intégrité des bases, consommation de ressources machines ou réseau...).

Transformation et Valorisation

Il est rare que les utilisateurs aient besoin de données très détaillées. En général, une base décisionnelle d'informations ne contiendra pas une réplication totale de la base de données opérationnelle mais seulement une partie des données de détail accompagnées d'informations composites issues d'opérations de transformation (basées sur des règles bien définies) et surtout de valorisation de la donnée brute.

Cette opération présente deux intérêts majeurs dans une organisation :

• D'une part, elle permet de s'assurer que tout le monde utilisera les mêmes règles de transformation ce qui permet de garantir l'unicité des résultats ("une seule et unique version de la vérité"). Il existe en effet aujourd'hui dans l'entreprise de multiples façons de représenter une même information.

• D'autre part, elle permet de décharger les utilisateurs de cette fastidieuse tâche de transformation des données pour les laisser se consacrer à leurs métiers.

La mise en œuvre de cette étape justifie l'implication des utilisateurs pour d'une part prendre en compte leurs besoins, d'autre part intégrer immédiatement les définitions précises (sémantique, règles) des informations telles que les utilisateurs les perçoivent et les analysent.

Ordonnancement et Chargement

La mise à jour de la base décisionnelle d'informations doit d'emblée prendre en compte le fait que les données vivent et évoluent au jour le jour au sein des bases opérationnelles. Un simple cliché des données, fussent-elles astucieusement valorisées, ne permet pas de mesurer l'évolution d'une tendance ou simplement de comparer deux situations. Il faut prendre en compte dynamiquement les changements dans les bases opérationnelles :

• soit de manière globale : la phase de conception de la base décisionnelle permettra de préciser la cadence de rafraîchissement de l'information. Dans la plupart des cas, ce rafraîchissement global ne sera pas réalisable de manière systématique compte tenu des volumes d'informations à prendre en compte et de leurs diversités,

• soit de manière incrémentale : on ne prendra en compte que les seules modifications intervenues dans les bases opérationnelles entre deux mises à jour de la base décisionnelle

d'information. La difficulté consistant dans ce cas dans l'identification de ces données. Plusieurs approches sont alors possibles, à titre d'exemple, on peut envisager de scruter les fichiers de journalisation (logs) alimentés par les gestionnaires de bases de données lors de toute opération de mise à jour; cette exploration permettant ensuite de répliquer de manière asynchrone toute mise à jour effectuée dans les systèmes opérationnels vers la base d'information décisionnelle.

3.2. Organisation

Référentiel

Une base d'information décisionnelle est avant tout un référentiel des connaissances et de l'information de l'entreprise. La constitution de ce référentiel doit être pensée dans le but de fédérer l'ensemble des informations d'une entreprise mais aussi la documentation de cette information. Il doit aussi être conçu pour une diffusion au plus grand nombre possible d'utilisateurs.

La documentation de l'information passe par la création d'un véritable dictionnaire de l'information ou "Métabase" qui au-delà des dictionnaires de données dont disposent habituellement les gestionnaires de bases de données permettra de donner du sens aux informations.

Métadonnées

Il faut concevoir la notion de Métabase sous trois angles :

• D'une part sous un angle technique (métadonnées techniques) comprenant l'ensemble des informations produites par le processus de transformation de la donnée brute en information, nous citerons en particulier sans vouloir être exhaustif :

- structure et contenu de la base décisionnelle : noms, formats, taille des données,

- dates de création des données,

- liens entre données opérationnelles et informations décisionnelles : origine des données,

- périodicité et ordonnancement des mises à jour,

- périodicité et ordonnancement des archivages...

• D'autre part sous un angle fonctionnel (métadonnées fonctionnelles) comprenant l'ensemble des caractéristiques différenciant la donnée brute de l'information finale, en particulier :

- règles et formules de calcul, d'agrégation et de consolidation,

- documentation explicite des informations : libellés, commentaires, notes d'utilisation,

- qualité et validité des données : dates de fraîcheur, contextes d'utilisation, précautions d'usage...

• Enfin sous un angle opérationnel (métadonnées opérationnelles) comprenant l'ensemble des caractéristiques liées à l'usage qui est fait de l'information :

- fréquence de consultation,

- caractéristiques des utilisateurs, notions de profils,

- sécurité de l'information...

Architectures : bases centralisées et bases de données thématiques.

La diversité des utilisations possibles de l'information, mais aussi les contraintes propres aux environnements d'exploitation (machines, réseaux, postes de travail, environnements logiciels) justifient qu'on s'intéresse aussi d'emblée aux aspects liés à l'architecture technique d'accueil des bases décisionnelles d'information.

Plusieurs aspects doivent être pris en compte :

• ouverture en amont vers les processus de collecte des données,

• ouverture en aval vers les postes de travail des utilisateurs et leurs différents outils,

• répartition éventuelle des bases décisionnelles sur plusieurs environnements.

En effet plusieurs architectures peuvent être mises en œuvre dans le cadre d'un tel projet. On distingue habituellement trois architectures :

• Archives Centralisées dans lesquelles les données sont habituellement rangées dans un espace de stockage de masse (éventuellement off-line) et accessibles ponctuellement par des requêtes de désarchivage. Ce mode s'applique au traitement de données historiques volumineuses.

• Données Centralisées : les données sont stockées et gérées sur un serveur central (usuellement de type mainframe) qui est à la fois un serveur de données et qui peut également être un serveur de calcul. Ce mode est particulièrement adapté pour une consultation par un nombre important de postes de travail accédant simultanément aux mêmes données.

• Données réparties (ou décentralisées, "data marts") : les données sont réparties entre plusieurs serveurs dédiés de moyenne puissance ou entre serveurs de réseaux locaux. Ce mode est particulièrement adapté lorsque les données peuvent être découpées selon des critères prédéfinis (nature des données, nature des utilisateurs, localisation des utilisateurs...). On l'utilisera principalement pour la mise en œuvre de bases de données thématiques.

Ces architectures expriment clairement qu'il ne suffit pas de recopier les données des systèmes de production dans une base décisionnelle, il faut aussi adapter l'architecture technique de cette base aux exigences de délivrance de l'information aux utilisateurs. Ceci peut impliquer de définir des niveaux de consultation et de stockage de l'information adaptés aux caractéristiques propres de la base de données décisionnelle parmi lesquels nous citerons :

• poste de travail utilisateur : localisation, nature,

• temps de réponse souhaité,

• fréquence de consultation,

• niveaux de consultation : données de détail, données agrégées...

Ces questions sont fondamentales et méritent une attention particulière, surtout lorsque la base de données est volumineuse et est susceptible de croître rapidement alors que le niveau de consultation varie d'un département à l'autre, d'un utilisateur à l'autre. D'où la relative nécessité de définir correctement des bases décisionnelles thématiques (ou data marts) considérées comme optimales par rapport à une utilisation spécifique.

3.3. Exploitation

Les modes d'utilisation des informations contenues dans une base décisionnelle vont être différents selon la nature des utilisateurs, ils vont aussi évoluer dans le temps. Les principales utilisations d'un tel environnement sont les suivantes :

• des besoins d'interrogation classiques aux fins de production de rapports sur un thème donné,

• des besoins de consultation évolués (analyses OLAP multidimensionnelles) dans le cadre d'applications d'aide à la décision (SIAD) ou de systèmes de pilotage (EIS ou Executive Information Systems) ou encore de systèmes d'information géographiques (GIS ou Geographical Information Systems),

• des besoins plus complexes d'étude et d'analyse : études statistiques simples ou complexes, modélisation, simulation,

• des besoins d'analyse systématique des bases décisionnelles aux fins de détection d'information (data mining) ou si on apprécie les néologismes d'"infodétection".

4. Méthodologie

L'implémentation et la maintenance d'une base décisionnelle d'informations doit procéder d'une méthodologie rigoureuse concernant les trois aspects évoqués plus haut de gestion, d'organisation et d'exploitation. Cette méthodologie doit être pensée comme un processus itératif et incrémental. Nous déclinerons ci-dessous les principaux aspects méthodologiques de ce processus.

Justification

• Création de l'équipe de projet constituée d'utilisateurs et d'informaticiens.

• Etude d'opportunité liée à la création d'une base décisionnelle d'informations.

• Mesure de l'impact d'une telle solution sur le métier des utilisateurs.

• Justification de l'investissement.

Etude des besoins

• Interview des utilisateurs potentiels de la solution.

• Bilan de la situation existante : richesse et/ou pauvreté des outils disponibles.

• Maquettage de la solution pour obtenir l'adhésion des utilisateurs.

• Choix des outils permettant de couvrir l'ensemble des besoins.

Conception

• Modèle logique de la solution : identification des sources de données adaptées à la satisfaction du besoin.

• Modèle physique de la solution : architecture physique de la base décisionnelle d'information.

• Modèle de transformation : agrégation, consolidation, règles de transformation, valeur ajoutée apportée à la donnée brute pour satisfaire les besoins exprimés.

• Prototypage de la solution : mise en situation de la phase de conception.

Implémentation

• Accès à l'ensemble des données : volumétrie réelle.

• Moteurs de transformation et de valorisation.

• Ordonnancement des accès et des transformations.

• Chargement des bases décisionnelles.

• Administration des métadonnées.

• Optimisation des accès.

• Définition des conditions d'exploitation : interfaces, profils d'utilisation...

Audit

Dans la mesure où il s'agit d'un processus incrémental et cyclique, une phase d'audit permettra de reboucler sur les phases précédentes en effectuant des bilans à court, et moyen terme sur la mise en œuvre de la base décisionnelle d'informations, les problèmes rencontrés et les améliorations à apporter à brève échéance : nouveaux besoins, nouveaux utilisateurs, nouveaux thèmes à traiter...

SAS Institute

La société SAS Institute est le neuvième éditeur indépendant de logiciels dans le monde, avec un chiffre d'affaires de 653 millions de dollars en 1996.

SAS Institute, créée en 1976, compte aujourd'hui plus de 4.700 salariés, dont près du quart dans les services de Recherche et Développement. La société réinvestit chaque année entre 30 et 35% de son chiffre d'affaires en Recherche et Développement, soit plus du double de la moyenne constatée dans ce secteur.

La filiale française, créée il y a quinze ans, emploie actuellement 145 personnes. Au cours de son exercice 1996, la société a réalisé un chiffre d'affaires de 149 millions de Francs. A ce jour, 7.000 produits SAS ont été installés en France chez 1.400 clients, parmi lesquels figurent les grandes entreprises et administrations.

Le Système SAS, implanté sur plus de 30.000 sites à travers le monde, est actuellement utilisé par plus de 3,5 millions d'utilisateurs. Ses multiples fonctionnalités en font le produit stratégique de la plupart des grandes entreprises et administrations pour les solutions de type Data Warehouse :

- Applications d'entreprise (Data Warehouse - EIS - OLAP)

- Applications de type analyses (contrôle qualité - prévisions - simulations - gestion de projets - Data Mining)

- Applications verticales (suivi des essais cliniques - pharmacocinétique - ingénierie des performances - consolidation et reporting financier).

Le Système SAS est le seul logiciel complet et ouvert du marché. Il présente toutes les caractéristiques indispensables aux décideurs pour accroître la compétitivité de leur entreprise. En assurant la fiabilité, la personnalisation, l'évolution et le développement de solutions adaptées, SAS Institute permet à ses clients une plus grande réactivité face aux mouvements des marchés.