

La collecte des données est essentielle



Alain Fernandez

Consultant indépendant, il intervient depuis plus de 15 ans auprès des grands comptes et des PME sur la conception des systèmes d'information stratégiques

Capitale, la collecte des données d'un système décisionnel est une opération particulièrement délicate. Accessibilité, nettoyage, consolidation, qualité, traçabilité et coûts ne doivent jamais être perdus de vue. Les freins et les réticences des personnes non plus !

La collecte des données dans l'entreprise est une étape majeure du projet d'informatique décisionnelle. Lors des premières réalisations (projet EIS, Infocentre ou Data Warehouse), les concepteurs traitaient un peu rapidement cette question. Ils partaient du postulat qu'il "suffisait" de rapatrier un maximum de données en un point accessible. Le décideur serait alors à même de dénicher les informations pertinentes et essentielles cachées au sein des bases de données. Le mythe du décideur-orpailleur des bases informationnelles a perduré.

Ce mythe est à l'origine de nombreux systèmes décisionnels inutilisables. Pour éviter l'échec d'un tel projet, il ne

faut donc pas se limiter exclusivement aux aspects techniques. La question de la collecte des données doit être reconsidérée pour prendre en compte les besoins fondamentaux en terme d'utilisation des données.

Adopter une démarche projet

Par définition, toutes les données de l'entreprise n'ont pas la même valeur pour tout le monde. Cette notion de valeur est fortement dépendante de l'usage et de l'utilisateur. Certaines données seront importantes pour un utilisateur en fonction de ses préoccupations et insignifiantes pour d'autres. Avant de commencer les opérations de collecte, il est donc indispensable

Toutes les données de l'entreprise n'ont pas la même valeur pour tout le monde

Le vieil adage " Qui peut le plus peut le moins " est souvent faux en informatique

Les systèmes d'information contiennent de nombreuses données erronées et inutilisables sur le plan décisionnel

SUR LE TERRAIN

Il y a quelque temps, je suis intervenu comme " pompier " sur un projet de Data Warehouse en dérouté, censé répondre à des besoins de pilotage et d'études clientèle dans le service marketing d'un équipementier. La base construite comportait d'énormes quantités de données techniques relatives à la conception et à la fabrication des produits. Ces données, sûrement riches d'enseignements pour les hommes du bureau d'études, de la production ou des méthodes, demeuraient purement abstruses pour les gens du marketing. Pourquoi les avait-on placées dans la base décisionnelle ? Tout simplement parce que ces données étaient déjà structurées et aisées à collecter. Personne ne s'était préoccupé de l'enseignement que pouvaient en tirer les utilisateurs, en l'occurrence les gens du marketing, par nature peu versés dans les questions de mécanique.

d'identifier avec précision les besoins des utilisateurs et les limites du projet. Il ne sert à rien de mettre à la disposition des utilisateurs potentiels des quantités faramineuses de données fort éloignées de leurs préoccupations.

Une donnée devient une information lorsqu'elle est porteuse de sens. La perception du sens dépend des utilisateurs et des préoccupations du moment. Nous ne sommes pas égaux devant l'information et cette dernière n'est pas porteuse d'un sens universel. Seuls les utilisateurs, en fonction de leurs préoccupations, sont à même de transformer les données en informations. La simplicité de l'accessibilité technique ne doit pas être l'unique critère de collecte. Il faut adopter une démarche projet afin de bien cerner les objectifs et les frontières des différents domaines d'intervention en fonction du besoin présent. La base décisionnelle pourra par la suite être enrichie, au fur et à mesure des nouveaux projets. Ne perdons pas de vue que le vieil adage "Qui peut le plus peut le moins" est souvent faux en informatique.

Collecte, Consolidation et maîtrise des coûts

Une fois la problématique bien définie et les domaines d'intervention circonscrits, la question purement technique de la collecte peut alors être abordée. En règle générale, les concepteurs se heurtent à trois difficultés :

- l'accessibilité des données en raison de l'hétérogénéité du système d'information,
- le nettoyage des erreurs et aberrations contenues dans les bases,
- la consolidation nécessaire pour rendre les données utilisables.

Les systèmes d'information de nos entreprises n'ont pas été conçus en six jours. L'approche a toujours été parcelaire et étalée dans le temps. Chaque projet était lancé pour répondre à un besoin fonctionnel précis et ponctuel, limité le plus souvent à une unité, une division ou un service. Personne ne considérait à leur juste mesure les besoins futurs en terme d'accès aux données essentielles. Notons que les différents rapprochements, rachats et fusions d'entreprises viennent complexifier encore la question de la cohérence du système d'information.

Les opérations de nettoyage constituent la seconde difficulté rencontrée par les concepteurs. Les systèmes d'information contiennent de nombreuses données erronées et inutilisables sur le plan décisionnel (par exemple, des erreurs, des valeurs aberrantes, des omissions, etc.). Les applications de production pleinement opérationnelles et recueillant la satisfaction des utilisateurs n'en sont pas exemptes.

La présence d'erreurs au sein de systèmes d'information parfaitement rodés s'explique simplement. Les données qui n'influencent pas les résultats,

n'ont pas à être vérifiées trop soigneusement. Cependant, ces données peu significatives pour les tâches de production sont peut-être porteuses d'un sens informationnel beaucoup plus riche pour les décideurs. Lors de la collecte, il faudra contrôler la validité de toutes les données devant jouer un rôle décisionnel dans le projet. On ne peut se permettre de laisser les utilisateurs prendre des décisions à partir d'informations erronées car dans ce cas le système serait très rapidement mis au rebut.

Les travaux de consolidation constituent la troisième difficulté à considérer. Du fait des anciennes habitudes de cloisonnement et de découpage des entreprises en centres de profits autonomes, les règles de gestion élémentaires sont rarement standardisées au niveau du groupe. Ceci est d'autant plus vrai lors de fusion et de rachat, avec le rapprochement d'entreprises conservant leurs propres modes de calcul. Pensons simplement aux différentes façons de calculer un chiffre d'affaires. Chaque entreprise utilise sa propre méthode (avec ou sans les ristournes, les escomptes, les commissions, etc.).

Comment peut-on intégrer ou comparer les éléments d'activité lorsque ceux-ci sont calculés différemment ? Il est du ressort de l'architecte du projet de mettre à la disposition des décideurs des données cohérentes.

Ainsi, lors de la collecte, il faudra résoudre chacun de ces points. La tâche est conséquente et grèvera significativement les budgets "financier" et "temps" du projet. Pour éviter de se lancer dans des opérations trop complexes, il faudra toujours garder en ligne de mire le paramètre coût. Pour chaque point difficile de la collecte, nous nous poserons alors la question de la rentabilité en mettant en regard l'ap-

port sur le plan décisionnel des données à collecter et le coût de l'opération. Bien que fortement subjectif (il est difficile d'exprimer a priori l'apport décisionnel d'une information), ce questionnement permettra d'éviter les dépassements de coûts et de délais inconsidérés.

Une gestion centralisée

L'accroissement de la quantité des données en circulation dans les entreprises suit une loi exponentielle. Avec l'incertitude ambiante et la rapidité du changement, les hommes sont avides d'informations et chaque unité produit de plus en plus de données. Il est désormais temps de gérer les données non seulement en termes d'accessibilité mais aussi en termes de qualité et de traçabilité.

De plus en plus d'acteurs de l'entreprise sont concernés par la prise de décision. Le processus de banalisation de la fonction décisionnelle va rapidement s'accélérer avec l'essor attendu des portails informationnels d'entreprise (EIP). Déjà, les projets décisionnels se multiplient. Que ce soit pour des questions d'aide au pilotage, de gestion de relation client ou d'analyse qualité, de plus en plus d'applications voient le jour. Il existe une forte tendance à la mise en œuvre de Datamarts spécifiques pour répondre localement à un problème posé. Paradoxalement, cette solution satisfaisante quant aux résultats, contribue fortement à la dispersion des données dans l'entreprise. Ainsi, on peut retrouver dans des bases décisionnelles distinctes, des données redondantes à différents stades de transformation et de fraîcheur.

Pour limiter les erreurs et plus généralement pour mieux maîtriser les coûts de nettoyage et de mise en forme des données, il est temps de placer un observatoire centralisant non pas les

Il est du ressort de l'architecte du projet de mettre à la disposition des décideurs des données cohérentes

Il est temps de gérer les données en termes de qualité et de traçabilité

LES OUTILS DE COLLECTE

Les outils comme Datastage d'Ardent ou Genio de Hummingbird automatisent la collecte et le nettoyage des données provenant des bases de données et des ERP les plus courants. Une fois le schéma d'extraction modélisé pour les différentes sources, l'utilisateur définit les transformations nécessaires pour rendre les données disponibles à des fins décisionnelles (calculs d'agrégat, contrôle des valeurs, élimination des aberrations...). Ces outils gèrent aussi le référentiel de méta-données centralisé. La grande majorité des bases sont accessibles en mode natif et il est possible de développer sa propre passerelle spécifique pour les sources de données non supportées par le produit.

Il faut créer un observatoire centralisant non pas les données mais leur(s) définition(s), leur(s) parcours et leurs utilisations

données mais leur(s) définition(s), leur(s) parcours et leurs utilisations. Pour cela, Il faut au préalable construire un référentiel centralisé à même de répondre à des questions du genre : d'où provient cette donnée ? Quand est-elle mise à jour ? Comment est-elle calculée ? Quelles sont les précautions d'usages ?

On appelle "métadonnées" les données sur les données apportant des réponses à ces questions. Pour une parfaite gestion, tous les modules fonctionnels du système d'information assurant le stockage, l'extraction, le traitement et la présentation, devraient utiliser et mettre à jour le référentiel.

Jusqu'à récemment encore, la définition d'un format standard demeurait le dernier obstacle à la généralisation des métadonnées. Deux formats étaient en lice : l'OIM (Open Information Model) proposé par le Meta Data Coalition et soutenu par Microsoft, à l'origine de sa définition, et le CWM (Common Warehouse MetaData) supportée par l'OMG (Object Management Group), Oracle et IBM, entre autres. Les deux formats s'appuient sur UML (Unified Modeling Language) pour la phase de

modélisation et XML (Extensible Markup Language) pour les formats de description et d'échanges. Fin septembre 2000, l'OIM a choisi de baisser les bras et d'abandonner sa norme pour contribuer à la prochaine version de l'OMG, le futur format unifié.

Décloisonnement et pouvoirs "parallèles"

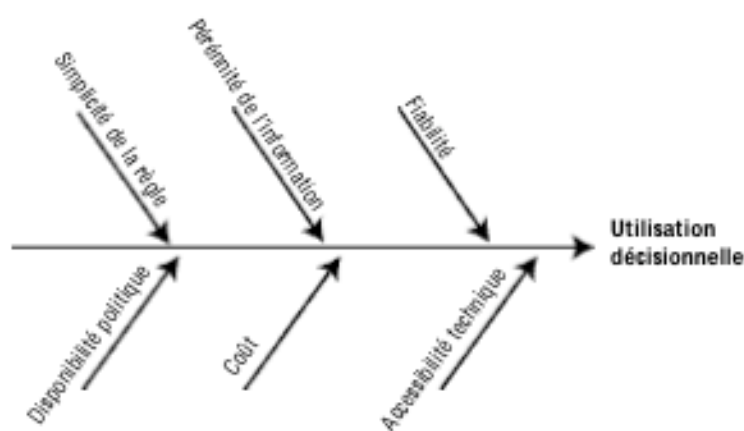
Les besoins en matière de collecte ne se limitent pas aux données présentes dans la sphère de prérogatives des utilisateurs du système décisionnel. Pour prendre les décisions, il faut pouvoir accéder à des données situées à l'extérieur, au sein de bases dépendantes d'autres activités et d'autres services. Là encore, il ne faudra pas limiter notre étude à l'aspect exclusivement technique mais bien s'informer auprès des producteurs et utilisateurs habituels des données concernées.

A ce stade, on aborde la difficulté majeure du projet. Prenons un exemple : récupérer les chiffres censés décrire l'activité d'une filiale peut présenter quelques difficultés techniques qui demeureront rarement insurmontables. Ces données seront alors dispo-

GERER LA DONNEE A PRIORI

Dans tous les cas, il faudra envisager la potentialité informationnelle des données gérées par les applications de production dès le lancement des projets système d'information. Nous perdons beaucoup trop de temps à récupérer des données peut-être secondaires pour une application de production mais importantes pour les besoins décisionnels. Rappelons-nous qu'une des principales récriminations des utilisateurs d'ERP concernait l'absence d'éléments décisionnels. Avec l'essor de l'e-business et plus généralement de l'importance stratégique liée à l'éclatement de l'entreprise et au développement de la supply-chain, l'information joue un rôle clé.

FIGURE 1 - LES CRITERES DE SELECTION DES INFORMATIONS



(D'après Fernandez, les Nouveaux Tableaux de bord des décideurs)

Le culte de la compétition individuelle encourage les hommes à thésauriser leurs informations

nibles. En revanche, nous ne disposons d'aucune garantie sur leur validité et surtout sur les précautions d'usage à respecter avant d'en tirer un enseignement.

Chaque filiale, chaque service, chaque activité définit ses propres règles de gestion. Avant d'utiliser une donnée et de chercher à en extraire une quelconque information, il est préférable au préalable d'en référer aux détenteurs de cette connaissance, les seuls à même de nous informer et de nous mettre en garde. Mais les détenteurs de cette connaissance seront-ils prêts à la communiquer à d'autres décideurs ? C'est toute la question du décloisonnement et du partage de la connaissance qui se pose. Le culte de la compétition individuelle, encore en vigueur dans de trop nombreuses entreprises, encourage les hommes à thésauriser leurs informations pour asseoir plus confortablement leurs pouvoirs. Ils ont, le plus souvent, tout intérêt à ne pas communiquer les recommandations nécessaires à l'usage des données... Surtout si personne ne le leur demande explicitement !

Cette question cruciale est à traiter avec précaution. D'autant plus que l'information n'est pas répartie uniformément dans l'entreprise. Certaines

personnes placées aux nœuds d'informations disposent d'un pouvoir de fait significatif. L'architecte du système décisionnel devra faire preuve de psychologie pour inciter les hommes à communiquer et éviter de heurter de front les sensibilités. Il prendra aussi garde à ne pas asseoir encore plus confortablement le pouvoir des hommes clés en entérinant techniquement les domaines informationnels.

Sur ce type de projet, l'engagement de la direction est indispensable ! Sans son réel appui, le projet est irréalisable. Seuls les dirigeants pourront expliquer les avantages de la coopération dans un esprit gagnant-gagnant au niveau de l'entreprise. Et pour les causes perdues, ils seront aussi les seuls en mesure de "faciliter" l'ouverture des portes des rebelles à l'entreprise communicante.

Sélectionner les données

Prenons le cas de la construction de tableaux de bord de pilotage. Pour chaque indicateur inclus⁽¹⁾, nous sélectionnerons les données nécessaires à sa construction en valorisant chacun des critères essentiels (voir figure 1). Il s'agit d'un travail de groupe, et chacun donnera son avis pour chaque critère à l'aide d'une note comprise entre 1 et 4.

L'architecte du système décisionnel devra faire preuve de psychologie pour inciter les hommes à communiquer

¹ - Pour la méthode de choix des indicateurs, se référer à l'article du même auteur dans le numéro 176 de l'Informatique Professionnelle.

