Diffusez des informations dignes de confiance
Décembre 2006

IBM Information Management software

L'exploration des données, première étape pour assurer la qualité de vos données

Table des matières

Pour les entreprises, l'accès à l'information constitue un véritable défi qui consiste à localiser les données voulues, à les rendre disponibles au moment opportun et au format adéquat, puis à les analyser. Sans parler des problèmes que posent la validité et le contrôle de ces informations.

Et ce défi va s'avérer de plus en plus difficile à relever si les entreprises ne prennent pas les mesures adéquates pour garantir l'accès à des informations fiables, cohérentes, exhaustives et actualisées. De fait, bon nombre de projets d'entreprises sont voués à l'échec en raison de la piètre qualité des données. L'un des moyens de contrôler la qualité de l'information consiste à explorer les données sources dès qu'elles sont introduites dans le système.

Pourquoi explorer vos données ?

Selon des études de marché, plus de 75 % des projets d'intégration de données échouent ou dépassent le budget qui leur est imparti. Dans la plupart des cas, ces projets ne parviennent pas à délivrer les fonctionnalités requises ou sont annulés avant d'être menés à terme. Comment expliquer que ce taux d'échec soit aussi élevé ? La réponse réside peut-être dans l'approche traditionnelle de l'intégration de données, qui repose sur les étapes suivantes :

Etape 1. Analyse des besoins utilisateur et spécification d'une base de données cible. Les entretiens avec les utilisateurs aboutissent à la conception d'un modèle de base de données capable de répondre à l'ensemble de leurs attentes vis-à-vis de l'application cible.

Etape 2. Analyse des sources de données disponibles. La compilation et l'analyse des données provenant de systèmes hérités, de systèmes de production et autres sources d'informations disponibles permettent d'évaluer leur degré de pertinence vis-à-vis de la base de données cible. La documentation relative à ces données sources peut être inaccessible ou incomplète. Des échantillons de données sont alors collectés et analysés pour permettre l'identification des propriétés intrinsèques de ces informations.

Etape 3. Définition d'un jeu de correspondances entre les données sources et la base de données cible. Cette étape implique la mise en place d'une stratégie permettant de convertir des données sources disparates au format cible. Cette tâche est généralement exécutée au moyen d'un outil d'extraction, de transformation et de chargement (ETL) ou de programmes manuels.

Etape 4. Transfert des données. Le fait de charger les données sources dans une zone de transit permet de les épurer, de les nettoyer et de les manipuler au format requis par le magasin de données cible. Un logiciel de gestion de la qualité des données peut être déployé à ce stade pour normaliser et corréler les enregistrements.

Etape 5. Chargement des données. Le transfert des données de la zone de transit vers l'application cible implique de les formater à des fins de reporting. Si cette opération semble aller de soi, elle présente néanmoins un certain nombre de points faibles qui sont responsables de l'échec de nombreux projets d'intégration de données : en effet, elle repose dans une large mesure sur des procédures manuelles et sur l'impression trompeuse qu'ont les entreprises de « connaître » leurs données.

Ne présupposez jamais que « vous connaissez vos données »

La principale lacune de l'approche traditionnelle en matière d'intégration de données réside dans l'assertion selon laquelle les informations requises par l'application cible sont effectivement disponibles sur les systèmes sources. De grandes compagnies ont investi des millions de dollars dans des projets d'intégration de données pour finalement s'apercevoir que les données sources n'étaient pas compatibles avec le modèle cible. Et ce type de situation peut se produire aussi bien lorsque le modèle a été défini en interne que lorsqu'il a été développé par une société tierce. Dans la mesure où le processus se compose d'une série d'opérations discrètes, exécutées le plus souvent manuellement par des équipes de développement totalement indépendantes les unes des autres, la discontinuité entre les différentes étapes conduit généralement au désastre.

Les entreprises consacrent en moyenne 80 % du budget alloué à leur projet au transfert et au chargement des données. Malheureusement, l'opération qui consiste à définir un jeu de correspondances entre les données sources et la base de données cible ne représente qu'une étape mineure du processus d'intégration de données issues de sources disparates. Le plus gros du travail réside dans l'exercice indispensable qui consiste à chercher des réponses aux questions suivantes :

- *Que contiennent exactement les sources de données ?*
- *Comment sont structurées les données ?*
- *Quelle est la qualité des données ?*
- *Sont-elles adaptées au but recherché ?*

Bon nombre de projets d'intégration de données dépassent le budget qui leur est imparti ou échouent totalement faute d'une maîtrise suffisante des métadonnées. En l'absence de processus de rétroingénierie automatisés applicables aux métadonnées, les développeurs en sont réduits à analyser manuellement les données sources. Or, la documentation relative aux données sources sur les systèmes hérités est généralement incomplète, voire inexistante. Le personnel capable d'interpréter les données a souvent quitté la société, et les conjectures les plus hasardeuses se substituent aux analyses de contenu dignes de ce nom. En conséquence, le débogage du processus d'intégration des données sources dans le magasin de données cible intervient à un stade très avancé du cycle de développement. Les problèmes inhérents aux métadonnées sont pris en compte trop tardivement—au niveau des systèmes de production—au lieu d'être résolus durant la phase de conception.

Un défaut n'ayant pas été détecté en amont—lors de la phase de spécification ou de conception—peut s'avérer 10 à 100 fois plus coûteux à éliminer par la suite. En matière d'intégration de données, cela peut impliquer un coût financier non négligeable pour l'entreprise qui tente de développer manuellement des bases de données cibles et d'exploiter des informations sans maîtriser réellement les propriétés des données sources. Faute d'outils permettant d'identifier les incidents en amont, au niveau du processus ETL, les entreprises peuvent être amenées à sacrifier une part importante du budget consacré aux entrepôts de données.

L'exploration assure la réussite de vos projets d'intégration de données

La médiocre qualité des données est la principale cause d'échec de bon nombre de projets d'entreprise. L'exploration initiale des données offre les avantages suivants :

- *réduit les risques inhérents au projet ;*
- *contribue à optimiser le retour sur investissement d'un grand nombre de projets d'entreprise—veille économique, déploiement d'applications d'entreprise, consolidation des instances existantes, délivrance d'une vue client unifiée, gestion de données permanentes, mise en conformité, etc. ;*
- *permet de déterminer s'il est possible ou non de respecter le cahier des charges de l'entreprise ;*
- *permet de s'assurer que les données sources disparates sont compatibles avec les spécifications du système cible avant d'investir du temps et des ressources dans le développement des processus d'intégration de données ;*
- *permet de détecter les problèmes de données à un stade précoce du projet, et ainsi de s'affranchir de procédures de test et de résolution d'incidents particulièrement coûteuses ;*
- *permet de planifier de façon plus rigoureuse les ressources requises dans le cadre du projet, que ce soit en termes d'effectifs, de compétences ou de temps.*

IBM WebSphere Information Analyzer intègre des outils d'exploration

IBM® WebSphere® Information Analyzer, un module d'IBM Information Server, permet d'automatiser le processus fondamental que représente l'analyse des données sources, contribuant à accélérer le retour sur investissement, ainsi que le coût global et le nombre de ressources associés aux projets d'intégration de données stratégiques. WebSphere Information Analyzer est capable d'explorer des données issues de sources disparates, en analysant les colonnes, les tables, les clés primaires et externes, les relations et les redondances.

WebSphere Information Analyzer aide les utilisateurs à intégrer des sources d'informations disparates en délivrant une vue complète des métadonnées et en identifiant les relations de dépendance au sein des tables et bases de données, et entre ces dernières. Dans la mesure où elles reposent sur les données sources proprement dites, les métadonnées présentent généralement un taux d'exactitude de 100 %. WebSphere Information Analyzer contribue ainsi à limiter les risques inhérents au projet en permettant de détecter les problèmes d'intégration avant que le processus de développement ne commence. Ce puissant outil d'exploration peut aider les entreprises à mettre en place une implémentation robuste et fiable, en s'affranchissant des problèmes d'intégration de données à la fois complexes et coûteux. Sur un projet type d'une durée normale de six à huit mois, WebSphere Information Analyzer permet d'obtenir des résultats significatifs dans un délai de 30 à 60 jours—ce qui représente un gain de temps moyen de 70 %.

Analyse des données sources

WebSphere Information Analyzer n'émet aucune hypothèse quant au contenu des sources de données. Il lit et analyse automatiquement les données sources, puis génère leur profil complet de façon à ce que leurs propriétés intrinsèques —c'est-à-dire les métadonnées— soient exemptes d'erreur. Les propriétés englobent les tables, colonnes, clés et relations de dépendances entre les données. Une trentaine de rapports prêts à l'emploi permettent à l'utilisateur d'analyser rapidement et efficacement les résultats.

Le succès de l'exploration réside dans les processus stratégiques

Les fonctions d'exploration de données de WebSphere Information Analyzer reposent sur un certain nombre de processus stratégiques.

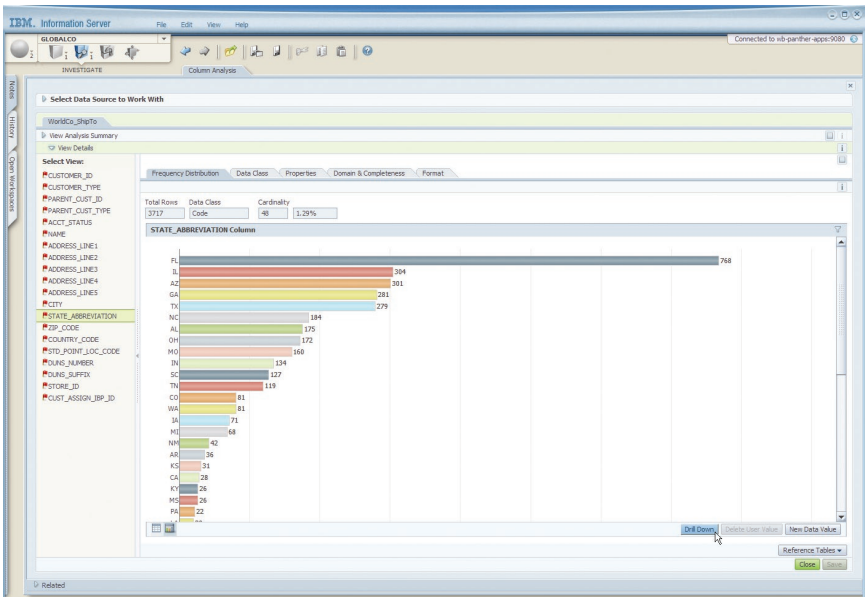
Analyse de colonnes

Ce processus consiste à examiner toutes les valeurs d'une colonne pour en déduire la définition et autres propriétés propres à cette colonne, telles que ses valeurs de domaine, ses valeurs minimale/maximale et les mesures statistiques qui s'y rapportent. Chacune des colonnes disponibles dans chaque table de données fait l'objet d'une analyse individuelle détaillée. Au cours de ce processus, un grand nombre de propriétés sont analysées et enregistrées, parmi lesquelles :

- *la longueur minimale, maximale et moyenne,*
- *l'échelle et le degré de précision des valeurs numériques,*
- *les principaux types de données rencontrés, dont les différents formats date/heure,*
- *les valeurs numériques minimale, maximale et moyenne,*
- *le nombre de valeurs vides, de valeurs NULL et de valeurs non NULL/vides*
- *le nombre de valeurs différentes, ou cardinalité,*
- *la répartition statistique des valeurs,*
- *la répartition statistique des formats de données,*
- *le nombre de formats de données.*

La Figure 1 fournit un exemple de répartition statistique.

Figure 1 : Répartition statistique



L'analyse de colonne permet en outre d'effectuer certaines déductions sur la nature des données contenues dans cette colonne. Elle permet notamment de déterminer :

- *le type de données, l'échelle et le degré de précision applicables à la colonne,*
- *si les valeurs NULL sont admises dans cette colonne,*
- *si la colonne contient une valeur constante,*
- *si les valeurs de cette colonne sont uniques.*

L'analyse de colonne permet par ailleurs d'évaluer la complétude et la validité des valeurs ou formats figurant dans cette colonne. En effectuant une sélection par valeur, plage de valeurs ou table de référence, les utilisateurs peuvent désigner certaines valeurs comme valeurs par défaut ou valeurs incorrectes, ou encore désigner certains formats comme non conformes aux standards. Ces informations peuvent être évaluées et faire l'objet de rapports périodiques afin de faciliter l'identification des problèmes potentiels au niveau de la qualité des données. Elles peuvent également être consignées dans des tables de référence à l'intention des développeurs.

Lors de l'analyse de colonnes, les utilisateurs génèrent des commentaires qui peuvent être partagés avec les processus d'intégration de données d'IBM Information Server, ce qui contribue à améliorer le retour sur investissement de façon significative.

Analyse de clés primaires

Ce processus consiste à identifier toutes les clés potentielles d'une ou plusieurs tables. L'objectif est d'identifier, pour chaque table, une colonne ou un groupe de colonnes susceptible de servir de clé primaire.

WebSphere Information Analyzer utilise immédiatement les résultats de cette analyse pour évaluer les clés primaires à colonne unique, sans nécessiter de traitement supplémentaire. L'utilisateur peut ensuite explorer les clés primaires multicolonnes. Il peut pour cela vérifier l'unicité d'un groupe de colonnes par rapport à un échantillon de données, en permutant jusqu'à neuf colonnes, ou par rapport à l'ensemble des données sources. L'utilisateur peut ensuite désigner l'une des clés potentielles comme la clé primaire.

Analyse de clés externes

Ce processus consiste à comparer toutes les colonnes des tables sélectionnées aux clés primaires de ces mêmes tables. L'objectif est de détecter la présence d'une relation de clé externe entre deux tables en recoupant les valeurs respectives de chacune des colonnes spécifiées avec celles de la clé primaire qui a été identifiée.

Sur la base de l'analyse des clés primaires et externes, WebSphere Information Analyzer commence par identifier la clé primaire associée à chaque table, puis recherche les colonnes contenant des données identiques à l'intérieur des tables ou des fichiers spécifiés. En cas de concordance, le processus d'analyse de clés externes identifie les données qui se recoupent. L'utilisateur peut alors définir la relation existant entre la clé externe et les colonnes correspondantes comme une relation de clé externe.

Analyse de relations interdomaine

Ce processus consiste à comparer toutes les colonnes de chacune des tables sélectionnées à celles des autres tables. L'objectif est de détecter les colonnes partageant un même domaine. Lorsque deux colonnes partagent un même domaine, cela peut indiquer qu'il existe une cohérence entre les données stockées dans chacune des deux tables—cohérence dans l'utilisation d'un code d'Etat ou de pays, par exemple—ou tout simplement que ces données sont redondantes.

Ce domaine partagé est accessible en mode bidirectionnel; c'est-à-dire que l'utilisateur peut visualiser la relation dans les deux sens, à partir de chacune des deux colonnes. S'il s'avère que les données sont redondantes, elles peuvent être marquées comme telles. Il est possible de réitérer périodiquement ce type d'analyse, que ce soit sur les sources de données existantes ou sur de nouvelles sources, à mesure que celles-ci sont ajoutées au projet, afin de contrôler les relations interdomaine de façon continue.

Evitez les pièges associés aux procédures manuelles traditionnelles

Le fait de consolider les différentes étapes du processus d'intégration de données à la lumière du résultat de l'analyse des métadonnées permet d'éviter les pièges associés aux procédures manuelles traditionnelles. L'utilisation de WebSphere Information Analyzer offre de nombreux avantages. Elle permet notamment :

- *de générer des métadonnées exactes, fondées sur le contenu réel des données, et non sur les conjectures des développeurs, qui ne sont souvent que de simples vues de l'esprit ;*
- *de détecter et de corriger les données erronées à un stade précoce du projet ;*
- *de générer automatiquement une documentation fiable concernant les données sources à partir des rapports système ; cette documentation, qui reflète les données réelles résidant sur le système source, peut ensuite être soumise à la vérification de l'utilisateur ;*
- *de s'affranchir de la tutelle des programmeurs qui ont développé les applications d'où sont issues les données sources ; l'accès aux données est la seule condition requise ;*
- *d'identifier des clés sur la base du contenu réel des données ;*
- *d'identifier des types de champ sur la base du contenu réel des données ;*
- *de générer et de mapper, dans les spécifications proprement dites, la plage réelle de valeurs de domaine associées aux champs de l'application cible ;*
- *d'identifier des relations de dépendance sur la base du contenu réel des données.*

Les gains de productivité résultant de l'utilisation de WebSphere Information Analyzer contribuent à réduire les besoins de main-d'œuvre associés à l'implémentation d'un projet d'intégration de données. Cela ne signifie pas pour autant que WebSphere Information Analyzer est capable d'éliminer l'ensemble des problèmes potentiels. Les analystes et les développeurs doivent toujours s'efforcer de prendre les bonnes décisions et d'employer leurs talents à résoudre les difficultés. Mais l'élimination des nombreux pièges traditionnellement associés aux projets d'intégration de données complexes contribue à réduire considérablement le temps et les efforts nécessaires à l'implémentation de ce type de projet.

Les exemples d'implémentation client dans de nombreux secteurs de l'industrie ont démontré que sur un projet type d'une durée normale de six à huit mois, WebSphere Information Analyzer permettait d'obtenir des résultats significatifs dans un délai de 30 à 60 jours. En identifiant les problèmes graves qui affectent les données sources à un stade précoce du processus, il contribue en effet à rendre la résolution de ces problèmes beaucoup moins coûteuse en termes de temps et de budget.

IBM Information Server délivre des informations dignes de confiance

WebSphere Information Analyzer fait partie intégrante de l'offre IBM Information Server dans la mesure où il permet de résoudre les problèmes stratégiques qui affectent les données sources dès le début de n'importe quel projet d'intégration de données.

IBM Information Server est une nouvelle plate-forme logicielle révolutionnaire qui vous aide à tirer parti des informations complexes et disparates disséminées sur l'ensemble de vos systèmes. IBM Information Server vous permet d'intégrer des données hétérogènes et de délivrer des informations fiables en tous lieux et à tout moment – en temps réel et en contexte – à des utilisateurs, des applications ou des processus spécifiques. Il permet au personnel informatique de collaborer avec les différents corps de métier afin de comprendre la signification, la structure et le contenu de n'importe quel type d'information résidant dans une source de données quelconque. Il délivre des niveaux de performances et de productivité exceptionnels en nettoyant, en transformant et en déplaçant ces informations de façon cohérente et sécurisée dans l'ensemble de l'entreprise, et peut par là même être mis à profit pour promouvoir l'innovation, renforcer l'efficacité opérationnelle et réduire les risques.

Pour en savoir plus

Pour plus d'informations sur WebSphere Information Analyzer ou IBM Information Server, veuillez contacter votre représentant marketing ou votre partenaire commercial IBM, ou consulter le site **ibm.com**/software/data/integration

IBM Software Group
Route 100
Somers, NY 10589

Imprimé aux Etats-Unis d'Amérique
Décembre 2006

TAKE BACK CONTROL WITH **Information Management**

IMW11808-FR-00