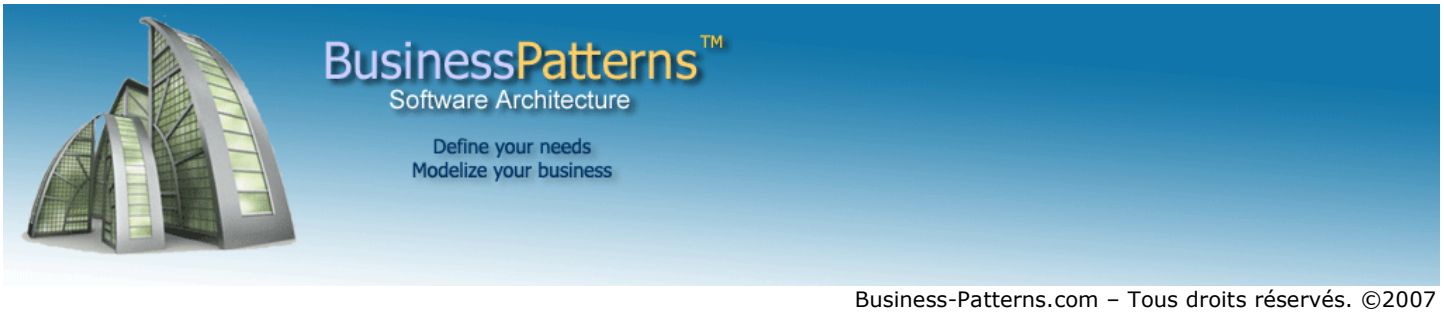


Acheter / développer / sélectionner son ETL (Extract Transform Load)

par Silver Nakache 12/05/2007

Dans les premières phases d'un projet d'entrepôt de données (Datawarehouse), la question de l'enrichissement (Feeding) est cruciale pour les étapes de restitutions.

Pour aider au choix décisionnel voici une liste non exhaustive des points à prendre en compte dans le choix de votre système d'enrichissement.

La multitude des sources à supporter :
Votre donnée peut provenir de plusieurs sources différentes, c'est très généralement le cas avec des fichiers formatées csv par exemple, mais cela peut provenir de source 'non conventionnelle' comme des logiciel propriétaire par exemple. Dans la majeure partie des cas, les ETL du marché supportent assez bien l'enrichissement à partir de fichier à plat, de base de donnée mais quand il faut inclure des données d'une 'Billing Telecom' ou d'une 'page Bloomberg' cela se corse un peu.

Il faut donc veiller à ce que l'ETL dispose d'un API suffisamment ouvert, qui dégage les contrats (les interfaces) nécessaires au bon enrichissement de l'entrepôt.

Vérifier que votre ETL supporte les sources standards du marché :

- Connecteur d'analyse des fichiers plats : .csv, .txt
- Connecteur d'analyse des fichiers hiérarchisés : XML
- Connecteur de base de données les plus connues : Oracle, Sql Server, Sybase, IBM DB2, Teradata, MS Access, MySQL, PostGres, Informix, FileMaker, Pervasive.
- Connecteur HTML
- Connecteur SOA/WebService
- Connecteur FTP
- Connecteur Mail IMAP/POP3
- Connecteur OLAP/MDX
- Connecteur aux ESB du marché : Tibco, MSMQ, MQSeries …
- Connecteur annexe : WMI, Standalone application

Vérifier que votre ETL supporte les outils de transformation flexibles et performants :
La donnée à importer nécessite parfois des transformations pour être mise au format cible (ou pivot) de l'entrepôt. C'est pourquoi, il est nécessaire de s'assurer que l'on dispose des leviers nécessaires pour transformer ou enrichir la donnée avant de la faire rentrer dans l'entrepôt.

Pour cela plusieurs solutions

- Le Scripting : l'ETL peut faire appel à un langage de Scripting tel que VBScript, WSH, PERL… qui permet d'appliquer une logique de transformation de la donnée.
- L'ouverture aux langages de programmation traditionnels tel que C#, Java, C++
- Les 'Fonctoid' : certains éditeurs exposent à travers des interfaces graphiques des éléments de transformation visuel qui permettent d'effectuer des transformations.

Support d'un Workflow customizable, Temps réel et multitâche:

- Votre donnée doit subir plusieurs étapes de transformations ?
- Votre donnée ne peut-pas être complètement réunie en une seule opération ?
- Votre donnée nécessite une intervention humaine ?

Alors vous avez besoin de modéliser le flux de votre donnée. Dans les outils les plus avancés la maitrise du flux est confiée à une interface de fabrication et visualisation du flux (basé sur WWF). L'aide au support est aussi considérable quand le flow est complexe en terme de diagnostique de la donnée disponible dans l'entrepôt. Visualiser l'enrichissement en multitâche et en temps réel est un plus non négligeable.

Support d'un outil de vérification sémantique :
Qu'est-ce qui garantit que l'enrichissement est correcte ?
L'intégrité référentielle de la base est un rempart contre une incohérence de la base dans ces relations entre tables et parfois au niveau de la colonne mais rien au niveau de la base ne peut appliquer un règle de vérification métier.
Pour cette raison il important de définir une *'short-live transaction'* ou une *'long-lived transaction'* couplée à des règles sémantiques spécialement codées pour s'assurer que la donnée entrée dans l'entrepôt est valide, dans le cas contraire, il faut annuler la transaction. C'est un problème trop largement sous estimé dans les entreprises où il reste à la charge du logiciel consommateur d'appliquer les règles de contrôle sémantique.
En définitive, il conviendra de s'assurer à la fin d'un cycle d'enrichissement, et avant d'entériner la transaction, de vérifier la cohérence de la donnée prête à entrer dans l'entrepôt.

Support d'un scheduler:

Il existe deux types d'insertion dans l'entrepôt, soit une insertion dite 'Temps Réel' soit une insertion en 'Mode Batch'. Dans cette dernière il est souvent nécessaire de programmer l'enrichissement soit sur un évènement (arrivée du fichier par FTP par exemple) soit sur un 'scheduling' particulier. Il est important de vérifier que ces tâches peuvent être accomplies sans encombre. Vérifiez la qualité des interfaces utilisateurs qui sont livrées et que les systèmes soient capables de vous notifier (email, Net send, …) quand l'enrichissement échoue.

Support d'un enrichissement multitâche :

Gestion des Données SCD (Slowly Changing Dimension)
Gestion des Données LAF (Late Arriving Fact)

Support d'outil de Monitoring :
Dans un souci de rationalisation les entreprises adoptent le plus souvent des systèmes de gestion d'alerte tel que MOM pour capter des métriques technique ou fonctionnel du système d'information. Si tel est votre cas vérifiez que votre ETLest des leviers de monitoring nécessaires : fichiers de log, connexion à des micro-agents, extension WMI ou compteur de performance NT par exemple.

Outil de Sécurisation des données :

Pour des raisons que l'on comprend aisément certaines données sont confidentielles et elles sont fournies sous forme cryptée. Vérifiez que votre ETL sera en mesure de décrypter la donnée juste avant son entrée en base.

Type de l'ETL :

Que l'ETL soit une solution du marché ou une solution maison, il se regroupe en trois catégories :

- **Engine-based** : Un moteur indépendant, généralement actif sur un serveur, réalisant le chargement et la transformation.
- **Database-embedded** : un moteur généralement livré sous forme d'utilitaire en ligne de commande pour charger la donnée.
- **Code-Generated** : Un moteur génère un package auto-exécutable pour mettre à jour la base de données.

Considérations liées à la performance :

Parmi les critères, il est important de vérifier que les temps d'enrichissement sont compatibles avec vos exigences métier. En pratique, vérifiez pour les volumétries attendues, avec une marge de sécurité bien-sûr, que vos 'batchs' aient finit avant le lendemain 8h par exemple.

Dans ce domaine, il n'y a pas trente-six mesures ! Soit vous avez la chance de tomber sur un benchmark officiel (les éditeurs s'engagent rarement sur les aspects performances) soit vous le testez vous-même.

Considérations liées à l'indépendance des éditeurs vis-à-vis de la base de données cible :

Faire le choix d'un ETL, c'est aussi s'assurer de son indépendance.

Vérifiez que l'outil est 'Agnostique' en terme de base de donnée cible. Sans quoi votre investissement risque d'être limité dans le temps et dans l'évolution de vos projets si vous décidez de changer de base de données (SQL Serveur vers Oracle par exemple).

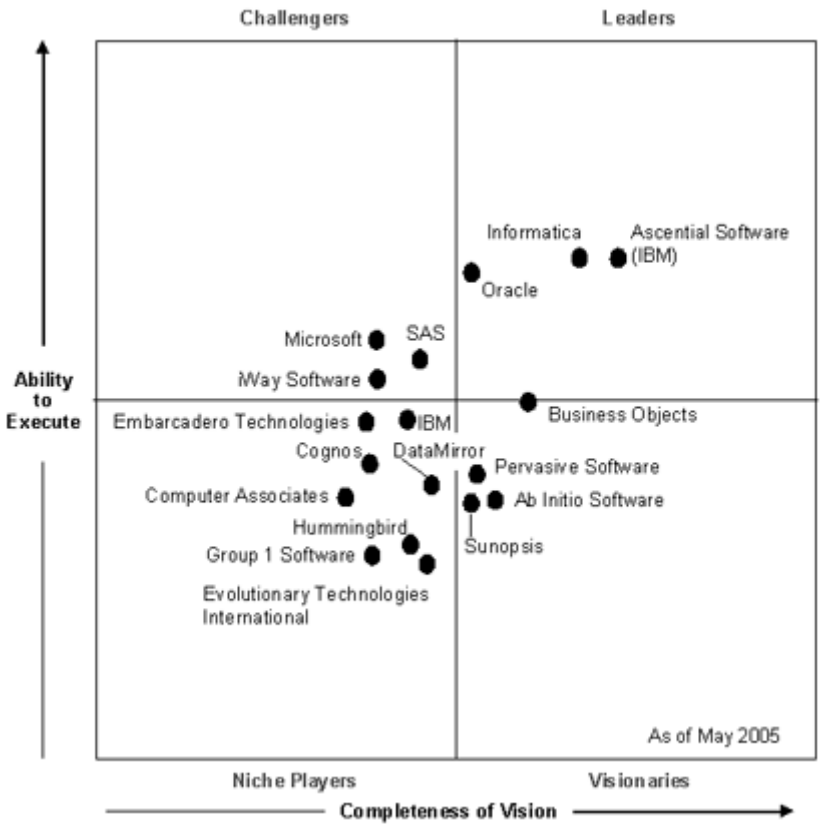
Considérations liées à l'indépendance de la / des plateforme(s) :

Dans les grosses entreprises, le parc informatique est très souvent hétérogène, les solutions ETL doivent clairement s'adapter aux différentes plateformes (AIX, Linux, Windows, …) pour en extraire la donnée. Dans ces cas seul les leaders savent s'adapter à de telles infrastructures. Mais le monde des ETL évolue rapidement, le développement du SOA et des protocoles d'échanges standardisés laissent encore un peut de place aux ETLs non 'cross-platefome', à suivre donc ….

Considérations liées aux risques/coûts :

- Vérifiez la simplicité de développement ou de l'achat des connecteurs spécifiques. Il n'arrive, pas si rarement que cela, que l'on se retrouve à faire des transformations d'un format à l'autre pour s'adapter à l'ETL, alors que la bonne marche veut que l'on développe un 'adapter' pour consommer la données avec le minimum d'étapes intermédiaires.
- Le coût de développement est bien sûr à mettre en balance avec la flexibilité et l'indépendance qu'il peut procurer.
- Vérifiez la pérennité des entreprises qui éditent l'ETL, prévoyez un contrat de maintenance/évolution qui couvre suffisament la durée de vie de votre projet.

Positionnement concurrentiel du Gartner (Magic Quadrant) :



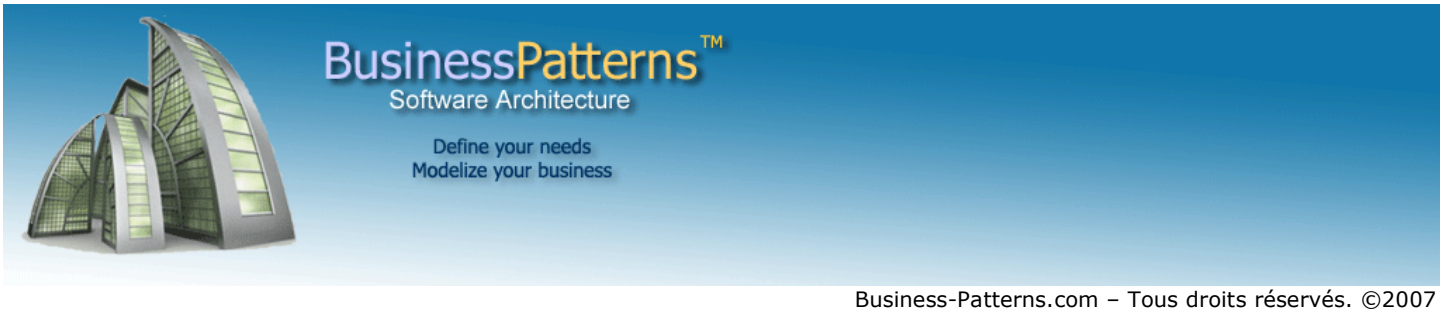
En conclusion:

Cet article ne peut pas adopter de position universelle concernant le choix d'un ETL, il vise surtout à vous faire pointer du doigt les aspects les plus cruciaux pour vous aider à sélectionner la solution la plus adaptée à vos besoins. Les ETL du marché ont beaucoup évolué ces dernières années et proposent de plus en plus des API permettant d'accomplir les tâches les plus compliquées. Si pendant plusieurs années beaucoup ont développé leur propre ETL pour des questions de limitations sur les produits du marché, la solution 'maison' est souvent consommatrice de ressource et donne des performances souvent moindres en termes de chargement « bulk ». Notre expérience est qu'il existe un positionnement par secteur de projets ; petits, moyens, gros, très gros, où il faut positionner chaque catégories d'ETL par groupe de projet. Il est aussi important de vérifier les 'ponts' qu'il existe entre chaque ETL pour permettre l'évolution d'un ETL à l'autre. N'hésitez pas à vous faire aider pour faire les bons choix, surtout qu'ils vous engagerons souvent sur plusieurs années…

Annexes :

Produits liés à des éditeurs de Bases de données parmi les plus connus :

Les éditeurs de bases de données livrent la plupart du temps des ETL, avec leurs bases de données. Ils sont généralement optimisés pour insérer de grande masse de données (bulk-load).



Références Editeurs :

1. Oracle Data Integrator (ODI): http://www.oracle.com/applications/oracle-data-integrator.html
2. IBM Information Server (IIS): http://www-306.ibm.com/software/info/ecatalog/fr_FR/products/J414133U71619Z27.html
3. Informatica Power Center (IPC) : http://www.informatica.com
4. Business Objects Data Integrator (BODI) : http://www.france.businessobjects.com/produits/dataintegration/dataintegrator/default.asp
5. Sunoposis : http://www.sunopsis.com
6. DataMirror Transformation Server: http://www.datamirror.com/fr/products/tserver/default.aspx
7. iWay Data Integration Solution : http://www.iwaysoftware.com
8. SQL Server Integration Services : http://msdn2.microsoft.com/fr-fr/library/ms141026.aspx

Références 'Open Source':

Produit non reconnus comme ayant une maturité suffisante pour les moyens est gros projets.

1. Talend Open Studio : http://www.talend.com
2. Pentaho Data Integration : http://www.pentaho.com/
3. BIRT Project : http://www.eclipse.org/birt/phoenix/
4. JetStream : http://sourceforge.net/projects/jetstream/
5. JasperETL (JasperSoft) : http://www.jaspersoft.com/JasperSoft_JasperETL.html

Références Editeurs de connecteurs spécifiques :

1. Oracle Data Provider for .NET (ODP.Net) :
2. Persistent : http://www.persistentsys.com/products/ssisoracleconn/ssisoracleconn.htm
3. FastReader http://www.wisdomforce.com/
4. ETI Pre-Built High Performance Connector: http://www.eti.com/products/connectors.html

Terminologie :

ETL : Extract Transform Load
Datawarehouse : Entrepôt de Données
Billing Telecom : Outil de Facturation et de CRM des opérateurs de téléphonie
SCD : Slowly Changing Dimension
LAF : Late Arriving Fact
WWF : Windows Workflow Fundation
MOM : Microsoft Operations Manager 2005
Feeding: Enrichissement de Données

Livres :
The Data Warehouse ETL Toolkit, Auteur(s): R. Kimball, J. Caserta, Wiley, ISBN: 978-0-7645-6757-5

Sources:
Magic Quadrant for Extraction, Transformation and Loading:
http://mediaproducts.gartner.com/reprints/oracle/127170.html