

Thème : Alimentation de l'entrepôt de données

LE QUANG Sylvain
BERNARD Valentin
FAURE Loïc
GÉRANTET Nicolas

Définition et objectif de l'alimentation À partir de la page 132 du SFD

Le rôle de l'alimentation est de **mettre à jour l'entrepôt de données** par rapport aux **BDD** existante afin de pouvoir faire des analyses.

Pour ce faire, il faudra concentrer les tables de la BDD afin d'obtenir un **cube de données multidimensionnel**, basé sur le système OLAP (On Line Analytical Processing).
C'est-à-dire que l'entrepôt de données permettra un **historique** des transactions effectuées.

Il y aura deux type d'alimentation : **mensuelle** et **annuelle**.
L'alimentation se lance par **batch** (cronjob), c'est-à-dire de façon **périodique** et **automatique**.
Il existe aussi une alimentation exceptionnelle au cas où l'utilisateur veuille rajouter tout de suite une donnée.

Les logiciels que nous pouvons utiliser pour l'entrepôt de données

Pentaho

Plates-formes : Windows, Linux

Licence : Open source (Mozilla Public License 1.1), payant (abonnement, version d'évaluation de 30 jours disponible)

Le projet Pentaho BI (Business Intelligence) est une application open source pour le reporting, l'analyse, le tableau de bord de gestion, le data mining, le workflow et l'ETL (Extract-Transform-Load) afin de satisfaire les besoins de l'informatique décisionnelle. L'entreprise a été créée à Orlando (Floride) en 2004 par des spécialistes de l'informatique décisionnelle.

Pentaho fonctionne par abonnement annuel.

La suite Pentaho est composée de cinq modules dont la description officielle est la suivante :

- **Pentaho Reporting** - accès aux données et délivrance des informations à l'entreprise
- **Pentaho Analysis** - exploration et analyse des données de manière interactive et rapide
- **Pentaho Dashboards** - donne une visibilité immédiates aux données et ICP (Indicateurs Clés de Performance)
- **Pentaho Data Integration** - nettoie et intègre des données (ETL)
- **Pentaho Data Mining** - découvre des motifs cachés et des indicateurs pour les performances futures

Pentaho possède une grande compatibilité avec les sources de données, avec un support comprenant plus de 30 plates-formes de bases de données (libres et propriétaires), les fichiers Excel, les fichiers bruts, etc.

Talend

Plates-formes : Windows, Linux

Licence : Open source, payant (abonnement)

Talend est un projet open source d'intégration de données, aux fonctionnalités à peu près similaires à celles de Pentaho. La société Talend a été fondée en France en 2005 ; leur première solution, Talend Open Studio, a été lancée en octobre 2006.

Talend propose une variété de solutions au sein d'une plate-forme unifiée afin de couvrir les besoins en management de données :

- Migration
- Synchronisation
- ETL pour le décisionnel
- Qualité de données du CRM (gestion de la relation client)
- Qualité de données de l'entrepôt
- Gouvernance des données
- Master Data Management (gestion des données de référence)

Les interrogations

Les alimentations exceptionnelles ne risquent-elle pas de mettre en danger l'intégrité des données ?

L'utilisateur pourrait avoir envie de mettre à jour directement dans l'entrepôt au lieu de passer à la BDD, et de ce fait, la BDD ne servirait plus à rien.

De plus, l'entrepôt de données est normalement en lecture seule pour l'utilisateur afin d'assurer une stabilité dans les données et la cohérence des analyses.

Est-ce que les alimentations exceptionnelles ont un impact sur la BDD ?

Car l'unicité des clés pourrait être compromise par exemple.

Talend et Pentaho sont des logiciels open source commercial ?

Y a-t-il des contrats avec l'ISTIL afin de pouvoir utiliser ces logiciels gratuitement sur la durée ?

Remarques

Pour l'entrepôt des données, il faudra une machine puissante et beaucoup d'espace car il y aura beaucoup de redondance.

Nous sommes un peu perdus sur le programme à tenir. Nous ne savons pas par où commencer etc...

Une réunion sur les choses à faire ne serait pas du luxe.

Nous n'avons pas beaucoup de questions précises compte tenu de notre visibilité limitée sur le projet pour l'instant.

Programme prévu du groupe

Choisir le modele de datawarehouse à modéliser : a priori, nous avons choisi par flocon étant donné qu'il y a beaucoup de tables, la table centrale sera la table *faits*.

Définir la source de données par le biais de l'ETL : choisir entre Pentaho et Talend.

Concevoir le cube ainsi que les dimensions : long travail d'analyse.

Utiliser un générateur d'états pour interpréter les résultats : il existe deux logiciels (Crystal reports ou Quickreports)

[Documentation](#)

À lire pour tout le monde, document sur le business intelligence :
<http://business-intelligence.developpez.com/cours/>

Pour concevoir un data warehouse (à lire absolument pour les membre du groupe Alimentation)
<http://taslimanka.developpez.com/tutoriels/bi/>