Entrepôt de données [1]
(data warehouse)

Introduction

1 Présentation

Le concept d'entrepôt de données a été formalisé pour la première fois en 1990 par Bill Inmon. Il s'agissait de constituer une base de données **orientée sujet, intégrée et contenant des informations historisées, non volatiles et exclusivement destinées aux processus d'aide à la décision**.

En effet, la simple logique de production (produire pour répondre à une demande) ne suffit plus pour pérenniser l'activité d'une entreprise. Elle est un système ouvert sur son environnement au coeur des systèmes d'informations confrontée à des phénomènes économiques et sociaux lourd de conséquences.

Pour faire face aux nouveaux enjeux, l'entreprise doit collecter, traiter, analyser les informations de son environnement pour anticiper. Mais cette information produite par l'entreprise est **surabondante, non organisée et éparpillée** dans de multiples systèmes opérationnels hétérogènes et peut provenir de toutes les places de marchés (mondialisation des échanges).

Il devient fondamental de rassembler et d'homogénéiser les données afin de permettre l'analyse des **indicateurs** pertinents pour faciliter la **prise de décisions**. L'objet de l'entrepôt de données est de définir et d'intégrer une architecture qui serve de fondation aux applications décisionnelles.

Définition : **Un entrepôt de données est une collection de données thématiques, intégrées, non volatiles et historisées pour la prise de décisions (**Bill Inmon)

L'infrastructure technique mise en œuvre est capable d'intégrer, d'organiser, de stocker et de coordonner de manière intelligible des données produites au sein du Système d'Information (issues des applications de production) ou importées depuis l'extérieur du SI (louées ou achetées) dans lesquelles les utilisateurs finaux puisent des informations pertinentes à l'aide d'outils de restitution et d'analyse (OLAP[2], Datamining[3]).

Les points clefs garantissant le succès d'un entrepôt de données sont les suivants :

- Les informations d'un entrepôt de données doivent être accessibles et fiables (de qualité).
- La conception d'un entrepôt de données doit répondre à un besoin de ROI[4] élevé.
- La réponse aux demandes très diverses des utilisateurs.
- L'entrepôt de données doit évoluer avec les besoins des utilisateurs et du système d'information.

---

1 On utilise souvent le nom anglais : data warehouse.
Remerciements à Fabien Angot « Intégration de données autour d'un entrepôt de données » Projet CNAM 2004
2 OLAP : On-Line Analytical Processing. Désigne une catégorie d'applications et de technologies permettant de collecter, stocker, traiter et restituer des données multidimensionnelles, à des fins d'analyse.
3 Datamining : Désigne une catégorie d'outils d'exploitation d'un entrepôt de données permettant d'effectuer des fouilles " mining " ou d'extraire des connaissances permettant de faire apparaître des corrélations jusqu'alors cachées entre les données.
4 ROI : Return On Investment, retour sur investissement

2 Les données du système d'information

Les données permettant la prise de décisions diffèrent des données opérationnelles :

Données opérationnelles	Données décisionnelles
Orientées application, détaillées, précises au moment de l'accès	Orientées activité (thème, sujet), condensées, représentent des données historiques
Mise à jour interactive possible de la part des utilisateurs	Pas de mise à jour interactive de la part des utilisateurs
Accédées de façon unitaire par une personne à la fois	Utilisées par l'ensemble des analystes, gérées par sous-ensemble
Haute disponibilité en continu	Exigence différente, haute disponibilité ponctuelle
Uniques (pas de redondance en théorie)	Peuvent être redondantes
Petite quantité de données utilisées par un traitement	Grande quantité de données utilisée par les traitements
Réalisation des opérations au jour le jour	Cycle de vie différent
Forte probabilité d'accès	Faible probabilité d'accès
Utilisées de façon répétitive	Utilisée de façon aléatoire

Tableau 1 : Différences entres les données opérationnelles et les données décisionnelles

Données orientées sujet

L'entrepôt de données est organisé autour des sujets majeurs et des métiers de l'entreprise. Les données sont organisées par thème, contrairement aux données des systèmes de production, organisées par processus fonctionnels.

L'avantage de cette représentation demeure dans le fait qu'il devient possible de réaliser des analyses sur des sujets transversaux aux structures fonctionnelles et organisationnelles de l'entreprise. Et ainsi, de pouvoir analyser un processus dans le temps à différentes étapes de sa conception au sein du SI. Cette orientation permet également de faire des analyses par itération, sujet après sujet. L'intégration dans une structure unique est indispensable pour éviter aux données concernées par plusieurs sujets d'être dupliquées. Dans la pratique il existe également des Datamart[5] pouvant supporter l'orientation sujet.

Données intégrées

Un Entrepôt de données est un projet d'entreprise et concerne les différents services et métiers de l'entreprise. L'intégration de données, au sein d'un entrepôt de données, est donc un processus déterminant sur la qualité et la quantité d'informations disponibles aux utilisateurs pour le processus de décision.
Cette phase, que nous verrons plus en détail avec les outils ETL[6], implique que les données doivent êtres mises en forme et unifiées afin d'avoir un état cohérent. Pour parfaire cette cohérence, l'intégration

5 Datamart ou Magasin de données : petit entrepôt de données, en général spécialisé dans un domaine « métier »
6 ETL : acronyme de Extract Transform and Load

nécessite une forte normalisation de données. Mais aussi la maîtrise de la sémantique, la prise en compte des contraintes référentielles et des règles de gestion. Ces notions sont énoncées, détaillées et administrées au sein des métadonnées de l'entrepôt de données.
C'est ainsi que l'on pourra donner une bonne vision de l'entreprise via l'utilisation d'indicateurs.

Données historisées

L'historisation est nécessaire pour suivre dans le temps l'évolution des différentes valeurs des indicateurs à analyser. Ainsi, un référentiel temps doit être associé aux données afin de permettre l'identification dans la durée de valeurs précises.

Données non volatiles

Afin de conserver la traçabilité des informations et des décisions prises, les informations stockées au sein de l'entrepôt de données ne peuvent être supprimées.

3 Les classes de données

Un entrepôt de données peut se structurer en quatre classes de données organisées selon un axe historique et un axe de synthèse.

Les données agrégées

Les données agrégées correspondent à des éléments d'analyse représentant les besoins des utilisateurs. Elles constituent déjà un résultat d'analyse et une synthèse de l'information contenue dans le système décisionnel, et doivent être facilement accessibles et compréhensibles.

Les données détaillées

Les données détaillées reflètent les événements les plus récents. Les intégrations régulières des données issues des systèmes de production vont habituellement être réalisées à ce niveau.

Les métadonnées

Les métadonnées constituent l'ensemble des données qui décrivent des règles ou processus attachés à d'autres données. Ces dernières constituent la finalité du système d'information.

Les données historisées

Chaque nouvelle insertion de données provenant du système de production ne détruit pas les anciennes valeurs, mais créée une nouvelle occurrence de la donnée.

4 Modélisation de données

La modélisation par sujet

Un entrepôt de données est généralement basé sur un SGBD relationnel.
La modélisation par sujet est une technique de conception logique qui vise à organiser et classifier les informations des bases légataires en données classées par sujet fonctionnel. Elle est basée sur la modélisation " Entité/Relation " et est préliminaire à la modélisation dimensionnelle. Chaque sujet correspond à une table gérée au sein de l'entrepôt. Il faut isoler les données stratégiques, déterminer les informations de détails nécessaires (profondeur, granularité) et conserver les métadonnées.

La modélisation dimensionnelle

La modélisation dimensionnelle (modèle multidimensionnel) souvent appelée modélisation OLAP (Codd 1993) se présente comme une alternative au modèle relationnel. Elle correspond mieux aux besoins du décideur tout en intégrant la modélisation par sujet.
C'est une méthode de conception logique qui vise à présenter les données sous une forme standardisée intuitive et qui permet des accès hautement performants. Elle aboutit à présenter les données non plus sous forme de tables mais de **cube**[7] centré sur une activité. Un cube de dimension n (n > 3) est aussi dit **hyper cube**.

Faits, indicateurs et dimensions

La table de faits est la clef de voûte du modèle dimensionnel où sont stockés les indicateurs de performances. Le concepteur s'efforce de considérer comme indicateurs les informations d'un processus d'entreprise dans un système d'information. Les indicateurs étant les données les plus volumineuses d'un système d'information, on ne peut se permettre de les dupliquer dans d'autres tables mais de les rationaliser au sein de la table de faits.

La table de faits

Table de faits des ventes journalières
Clé date (CE)
Clé produit (CE)
Clé magasin (CE)
Quantité vendue
Montant des ventes (€)

Tableau 2 : Modèle conceptuel d'une table de faits

Le terme de fait est utilisé pour représenter une mesure économique. Pour exemple, lors de la vente de produits sur un marché, on comptabilise les types de produits vendus, leur quantité et le montant de chaque vente au jour le jour et ce, pour chaque produit et pour chaque magasin.
La mesure des quantités et des prix est réalisée à l'intersection de toutes les dimensions (produit, magasin, temps). Le nombre des dimensions détermine la finesse, la granularité de la table et indique la portée de l'indicateur.

Additivités des indicateurs

7 Cube : Une construction multidimensionnelle formée de la conjonction de plusieurs dimensions. Chaque cellule est définie par une seule valeur de chaque dimension.

Les indicateurs les plus utiles d'une table de faits sont numériques et additifs. L'additivité des attributs d'une table de faits est cruciale pour les outils décisionnels. Les utilisateurs demandent rarement l'analyse d'une seule ligne. Dans notre exemple, constater les ventes de produits sur une année pour les magasins d'une région demande l'analyse de plusieurs milliers de lignes à la fois.
Pour autant, tous les attributs utiles ne sont pas additifs. Certains sont semi additifs et ne peuvent être additionnés que pour certaines dimensions.
D'autres sont non additifs et ne peuvent pas être additionnés par dimensions. Pour cette dernière catégorie, on utilise des fonctions d'agrégations tel que, le calcul de moyenne, le ratio ou le comptage de lignes.

Les dimensions

Les tables de dimensions sont les entités complémentaires à la conception de la table de faits. Elles contiennent, autant que possible, des attributs sous forme de descriptions textuelles permettant de qualifier ou d'expliquer l'activité.
Des attributs de dimensions, nombreux, permettent de varier les possibilités d'analyse (par tranches ou en dés). Ces attributs rendent utilisables et intelligible les données de l'entrepôt de données. Ils établissent, en quelque sorte une interface homme/entrepôt de données.
En général, les tables de dimensions tendent à être peu profondes mais elles sont larges (l'inverse de la table de faits), en d'autres termes elles ont peu de lignes mais beaucoup de colonnes.

Tables de dimension "Produit"
Clé produit (CP)
Description du produit
Numéro US (clé naturelle)
Description de la marque
Description de la catégorie
Description du rayon
Description du type d'emballage
… et bien d'autre attributs

Tableau 1-3 : Modèle conceptuel d'une table de dimension

Structure de la base de données

Au sein de l'entrepôt de données les données sont redondantes et dénormalisées, nous sommes loin de la modélisation en troisième forme normale (3NF) et pour cause, cela permet de faciliter l'utilisation et d'améliorer les performances lors de l'analyse des données.
Trois types de schémas sont fréquemment rencontrés, le schéma en étoile, le schéma en flocon et le schéma en constellation de faits.

Le schéma en étoile

Dans un schéma en étoile, une table centrale de faits contenant les faits à analyser, référence les tables de dimensions par des clefs étrangères. Chaque dimension est décrite par une seule table (feuille de l'arbre de tables) dont les attributs représentent les diverses granularités possibles.

Le schéma en flocon

Dans un schéma en flocon, cette même table de faits, référence les tables de dimensions de premier niveau, au même titre que le schéma en étoile. La différence réside dans le fait que les dimensions sont décrites par une succession de tables (à l'aide de clefs étrangères) représentant la granularité de l'information. Ce schéma évite les redondances d'information mais nécessite des jointures lors des agrégats de ces dimensions.

Les schémas en constellation de faits

Dans un schéma en constellation, plusieurs modèles dimensionnels se partagent les mêmes dimensions, c'est-à-dire, les tables de faits ont des tables de dimensions en commun.
Pour conclure, les différences entre ces trois modèles sont faibles et ne peuvent donner lieu à des comparaisons de performance. Ce sont des schémas issus de la modélisation dimensionnelle utilisés par les outils décisionnels.