**Résumé DATA** **WAREHOUSE**

**Introduction**

Un *data* *Warehouse*, ou entrepôt de données, est un ensemble de données organisées et traitées de telle sorte que des décideurs puissent prendre des décisions à partir des informations produites.

Un entrepôt de données est à différencier d’une base de données. Ils n’ont pas les objectifs : les uns servent à la prise de décision historisée quand les autres font du stockage de données qui doivent être accessibles rapidement. Ils ne font pas appel aux données de la même manière : les entrepôts croisent les données de différentes sources, les bases de données fonctionnent en silos. Le volume de données, le public utilisateur, la forme des requêtes sont également des points de divergence.

**Implémentation** **d’un** **entrepôt** **de** **données**

Il existe trois approches d’implémentation :

- ROLAP : à partir d’une base de données, on va simuler un fonctionnement multidimensionnel

- MOLAP : les données sont stockées de manière multidimensionnelle, donc le fonctionnement est nativement multidimensionnel

- HOLAP : les données sont stockées de manière relationnelle, et on va “multidimensionnaliser” ces données, pour un fonctionnement intermédiaire

**Cube**

Un cube est une représentation des données selon plusieurs dimensions. Ces dernières sont des points vues depuis lesquels les données peuvent être observées. Les données observées selon les dimensions sont ainsi comparables entre elles : c’est le coeur de l’analyse.

On peut effectuer des opérations sur un cube : vue par tranche, extraction de sous-cube, observation des données avec une granularité différente sur une dimension (inférieure ou supérieure), la granularité d’une dimension étant le niveau de découpage de son unité.

**Modèles** **en** **étoile** **et** **en** **flocon**

Les mesures relevées constituent des faits, stockés dans des tables de faits. Ces mesures ont été faites selon un certain point de vue, en fonction des dimensions.

Ainsi, représenter une ou plusieurs tables de faits reliée à ses dimensions associées dénormalisées (c’est-à-dire que tous les descripteurs d’une dimension sont contenus dans sa table) constitue un modèle en étoile.

Si les dimensions sont normalisées (c’est-à-dire qu’elles sont décomposées selon leur granularité), alors cela constitue un modèle en flocon.

Fusionner des modèles en étoile qui utilisent des dimensions communes donne un modèle en constellation.

**Conception** **d’un** **entrepôt** **de** **données**

Pour concevoir un entrepôt de données, il faut en premier lieu décider à quoi il va servir, définir le modèle de données, puis obtenir les données que l’on souhaite exploiter. Étant donné que les sources peuvent être nombreuses, et par conséquent que des redites peuvent avoir eu lieu, on fait un “nettoyage des données”. Une fois cela réalisé, on peut décider du mode de restitution pour établir l’architecture de l’entrepôt. On peut également créer des “data marts”, sous-ensemble de notre entrepôt.

**De l’entrepôt à la décision**

Le besoin des entreprises et des organisations est généralement de disposer des informations stratégiques. C'est-à-dire des informations qui aideront l’entreprise ou l’organisation à prendre des décisions stratégiques dans le temps en fonction de ces objectifs.

Pour accéder à ces informations stratégiques, il faut d’abord centraliser toutes les données sources dans des entrepôts de données. Les données de ces entrepôts dérivent des sources de données structurées (bases de production, fichiers, ...) et de données semi-structurées et multimédias (XML, HTML, ...). La fiabilité et le nettoyage de toutes ces données sont indispensables pour assurer de bonnes prises de décision et prédictions d’où l’importance, en amont, de la phase d’entreposage proprement dite.

**Définition et concept d’un entrepôt de données**

L’entrepôt de données n’a pas une définition unique. Plusieurs définitions ont été proposées pour essayer de le définir parmi lesquelles on en cite deux :

-Définition 1 : l’entrepôt de données est un ensemble de données historisées variant dans le temps, organisé par sujets, consolidé dans une base de données unique, géré dans un environnement de stockage particulier, aidant à la prise de décision dans l’entreprise.

-Définition 2 (de Bill Inmon en 1996 ) qui présente un entrepôt comme étant ‘’Une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision’’.

Le principe de l’entrepôt de données est de mettre en place une base de données multidimensionnelle utilisée à des fins d’analyse. L'organisation des données est conçue pour que les personnes intéressées aient accès rapidement et sous forme synthétique à l'information stratégique dont elles ont besoin pour leurs prises de décision.

Les trois fonctions essentielles dans l’entrepôt de données sont : collecte de données de bases existantes et leur chargement, la gestion des données dans l’entrepôt et enfin l’analyse de données pour la prise de décision, en passant souvent par les Datamarts.

L’entrepôt de données est séparé en trois zones : la zone de préparation permet un un stockage temporaire des données ainsi que leur transformation (nettoyage, normalisation …), la zone de stockage permanent des données nettoyées, puis la zone de présentation qui permet  d’accéder aux données et de procéder à des analyses.

Le processus d’alimentation consiste à rassembler de multiples données sources souvent hétérogènes et les homogénéiser. Ainsi pour mettre les entrepôts à jour régulièrement on utilise un outil en mesure d’automatiser les chargements dans l’entrepôt tel que ETL (Extract, Transform, Load). **ETL** offre un environnement de développement, des outils de gestion des opérations et de maintenance permettant de découvrir, extraire et analyser les données à partir de sources hétérogènes mais aussi de nettoyer et standardiser les données afin de charger les données dans un entrepôt.

Un *DataMart* est une partie de l’entrepôt destiné à répondre aux besoins d’un secteur ou d’une fonction particulière de l’entreprise (point de vue spécifique, critères métiers...). Un DataMart comporte moins de données que l’entrepôt ce qui le rend plus facile à comprendre, à manipuler mais aussi améliore les temps de réponse aux requêtes.

**Modélisation**

La modélisation multidimensionnelle souvent appelée modélisation OLAP (On-Line Analitical Processing) se présente comme une alternative au modèle relationnel classique des bases de données. Elle correspond mieux aux besoins du décideur tout en intégrant la modélisation par sujet. Cette méthode de conception logique vise à présenter les données sous une forme standardisée et intuitive qui permet des accès plus performants. Elle aboutit à présenter les données non plus sous forme de tables mais de cube ce qui est mieux adapté à l’intelligence d’affaires. La modélisation multidimensionnelle a donné naissance aux concepts de *fait* et de *dimension* (Kimball 1996).

Le terme de fait est utilisé pour représenter une mesure économique. Par exemple, lors de la vente de produits sur un marché. La table des faits contient les données observables (les fait) sur le sujet étudié selon les divers axes d’analyse (les dimensions).

Les tables de dimensions sont les entités complémentaires à la conception de la table de faits. Les tables de dimension sont des axes d’analyse selon lesquels vont être étudiées les données observables (par exemple il y a la dimension temps).

**Les types de modèles**

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 clefs étrangères. Les dimensions n’ont pas de liaison entre elles.

Les avantages du schéma en étoile sont : Facilité de navigation, Nombre de jointures limité.

Et les principaux inconvénients sont : Redondance dans les dimensions et toutes les dimensions ne concernent pas les mesures

**Le schéma en flocon** : Il a la même composition que le schéma en étoile sauf que les tables de fait et de dimensions sont décomposées en sous hiérarchies. Mais une table de dimension contient un seul niveau hiérarchique. La table de dimension de niveau hiérarchique le plus bas est reliée à la table de fait. On dit qu’elle a la granularité la plus fine.

Les avantages du modèle en flocon sont : la normalisation des dimensions et l’Économie d’espace disque.

Mais les inconvénients sont : Modèle plus complexe (jointure) et requêtes moins performantes.

**Les schémas en constellation de faits :** Dans un schéma en constellation, plusieurs modèles dimensionnels se partagent les mêmes dimensions,c'est-à-dire que les tables de faits ont des tables de dimensions en commun.

**Pour une bonne modélisation il est conseillé de suivre les 9 étapes de Kimball :**

|  |  |
| --- | --- |
| 1-Choisir le sujet  2-Choisir la granularité des faits  3-Identifier et adapter les dimensions  4-Choisir les faits  5-Stocker les pré-calculs | 6-Établir les tables de dimensions  7-Choisir la durée de la base  8-Suivre les dimensions lentements évolutives  9-Décider des requêtes prioritaires |

**Le cube**

La modélisation multidimensionnelle des données facilite l’analyse d’une quantité selon différentes dimensions : Temps, Localisation géographique etc. Les calculs sont réalisés lors du chargement ou de la mise à jour du cube.

**Manipulation des données multidimensionnelles**

Le langage permettant de définir, d'utiliser et de récupérer des données à partir d'objets multidimensionnels est le MDX (Multidimensional Expressions), permettant d’effectuer les opérations. MDX est l’équivalent de SQL pour le monde OLAP. Ainsi les opérations ensemblistes possibles sur un cube sont : Rotate, Slicing, Dicing, Scoping et Drill-up / down.

Les données peuvent être fouillées grâce au Data Mining qui est une recherche de connaissance, sous forme de modèle de comportement, cachés dans les données. Les techniques de fouille en évolution les plus répandues sont : régression linéaire, induction d’arbres de décision, algorithmes génériques, réseaux de neurones.