Année Academique 2023- 2024

Ministère de l'Economie Numérique, des Télécommunication et de l'Innovation

ESATIC
ECOLE SUPÉRIEURE AFRICAINE DES TIC

Cours d'informatique décisionnel

Système d'Information Décisionnel

Cycle Master (1)

Spécialité : SIGL

Chapitre 1: Introduction aux Entrepôts de Données

1. Introduction

Formalisé au début des années 1990, le concept d'Entrepôt de Données (ED) en anglais Datawarehouse (DW) est devenu la clé de voûte de ce qui est appelé aujourd'hui «l'information décisionnelle ». Son fort développement a été poussé par un effet de mode renforcé par l'occasion saisie par les fournisseurs du marché informatique (constructeurs, intégrateurs de systèmes, éditeurs de logiciels), qui y ont trouvé le moyen de proposer des nouveaux produits remplissant plus ou moins bien les fonctions demandées par les utilisateurs.

Les systèmes d'information (SI) ont connu une longue suite d'innovations.

- Sur les infrastructures (en particulier les structures client/serveur, intranet/internet)
- Sur les outils (en particulier les bases de données relationnelles et le développement des langages objets).

Mais l'entrepôt de données ne s'inscrit pas dans cette séquence : L'entrepôt n'est pas une nouvelle plateforme technologique. Il ne s'agit pas uniquement d'outils et de techniques. Avant tout, l'entrepôt de données doit considérer les besoins de l'entreprise. Son développement demande une approche métier.

Il faut cependant se méfier d'une démarche simpliste de présentation :

- L'expertise associée à la notion de démarche métier n'est pas une simple connaissance des produits du marché,
- Bien qu'il s'agisse de traiter des grands volumes de données, les performances du système ne s'appuient pas nécessairement sur l'utilisation de machines ou d'architectures massivement parallèles,
- Un entrepôt de données n'est pas simplement un SGB haut de gamme,
- Un entrepôt de données ne se résume pas à un logiciel de présentation de données, même si on peut donner à cette partie du système une importance stratégique (partie visible, donc partie vedette).

En résumé, la définition d'un entrepôt de données ne se réduit pas à une interface de présentation associée à une base de données. Ce n'est pas faux, mais la clé est dans la pertinence des contenus : l'adéquation des données au mécanisme de décision.

2. Présentation

Un système d'information décisionnel (SID) viable implique avant tout la mise en place de deux éléments essentiels :

- Un modèle de données spécifique et
- Une infrastructure d'alimentation

La construction de ces modèles sera étudiée plus en détail dans les chapitres suivants. Leur élaboration est l'œuvre de génie logiciel et d'intégration.

Un système d'Information Décisionnel est bien un projet stratégique, ce qui ne signifie pas que c'est nécessairement un projet de taille démesurée. D'une façon réaliste les volumes de données sont plus de l'ordre de la centaine de giga-octets que du tera-octet.
Un SID ne se définit pas de façon académique, on peut cependant donner quelques caractéristiques fondamentales, en examinant les objectifs qu'il doit servir. Ces objectifs sont

définis par l'expression des besoins des gestionnaires d'entreprises pour réaliser des tâches de prise de décision. La réalisation de ces tâches nécessite :

- d'accéder aux (montagnes de) données, qui existent dans leur société,
- de faire des tris, des regroupements, des coupes en tranche, en dés et en toutes sortes de façon dans les données,
- d'accéder facilement et directement aux données qui leurs sont nécessaires,
- de montrer seulement ce qui est important,
- d'obtenir des valeurs cohérentes et comparables même lorsque les sources sont différentes,
- d'obtenir des valeurs fiables et objectives qui permettent de prendre des décisions.

On peut alors transformer l'expression de ces besoins, sous forme d'une sorte de cahier des charges d'un entrepôt de données, d'une part relativement aux fonctions qu'il doit remplir et d'autre part, relativement à sa structure.

L'idée de constituer une base de données orientée sujet, intégrée, contenant des informations datées, non volatiles et exclusivement destinées aux processus d'aide à la décision fut dans un premier temps accueillie avec une certaine perplexité.

Mais l'économie actuelle en a décidé autrement. Les entreprises sont confrontées à une concurrence de plus en plus forte, des clients de plus en plus exigeants, dans un contexte organisationnel de plus en plus complexe et mouvant.

Pour faire face aux nouveaux enjeux économiques, l'entreprise doit anticiper. L'anticipation ne peut être efficace qu'en s'appuyant sur de l'information pertinente. Cette information est à la portée de toute entreprise qui dispose d'un capital de données gérées par ses systèmes opérationnels et qui peut en acquérir d'autres auprès de fournisseurs externes.

Mais actuellement, les données sont surabondantes, non organisées, dans une perspective décisionnelle et éparpillées dans de multiples systèmes hétérogènes.

Pourtant, les données représentent une mine d'informations. Il devient fondamental de rassembler et d'homogénéiser les données afin de permettre d'analyser les indicateurs pertinents pour faciliter les prises de décisions.

Pour répondre à ces besoins, le nouveau rôle de l'informatique est de définir et d'intégrer une architecture qui serve de fondation aux applications décisionnelles : l'entrepôt de données ou Datawarehouse.

3. Pourquoi un Entrepôt de Données (Data Warehouse) ?

3.1. La problématique des Entreprises

L'entreprise construit un système décisionnel pour améliorer sa performance. Elle doit décider et anticiper en fonction de l'information disponible et capitaliser sur ses expériences.

Depuis plusieurs dizaines d'années, une importante masse d'informations est stockée sous forme informatique dans les entreprises. Les systèmes d'information sont destinés à garder la trace d'événements de manière fiable et intègre. Ils automatisent de plus en plus les processus opérationnels.

Parallèlement, les entreprises réalisent la valeur du capital d'information dont elles disposent. Au-delà de ce que l'informatique leur apporte en terme fonctionnel, elles prennent conscience de ce qu'elle pourrait apporter en terme de contenu informationnel.

Considérer le système d'information sous cet angle en tant que levier pour accroître leur compétitivité et leur réactivité n'est pas nouveau. Par contre, étant donné l'environnement concurrentiel actuel, cela devient une question de survie.

L'informatique a un rôle à jouer, en permettant à l'entreprise de devenir plus entreprenante et d'avoir une meilleure connaissance de ses clients, de sa compétitivité ou de son environnement.

Il est intéressant de calculer les retours sur investissement rendus publics. Ils se calculent rarement en termes de baisse de coûts, mais en terme de gains. Par exemple, ils permettent un meilleur suivi des ventes, une meilleure compréhension des habitudes d'achats des clients, d'une adaptation des produits à une clientèle mieux ciblée.

A ce titre, le Datawarehouse doit être rapproché de tous les concepts visant à établir une synergie (association) entre le système d'information et sa stratégie.

3.2. La réalité des systèmes d'information

A première vue, les systèmes opérationnels seraient des mines d'or informationnelles. En fait, il n'en est rien.

Les données contenues dans ces systèmes sont :

- Eparpillées : il existe souvent de multiples systèmes, conçus pour être efficace pour les fonctions sur lesquelles ils sont spécialisés.
- Peu structurées pour l'analyse : la plupart des systèmes informatiques actuels ont pour objet de conserver en mémoire l'information, et sont structurés dans ce but.
- Focalisées pour améliorer le quotidien : toutes les améliorations technologiques se sont focalisées pour améliorer cette capacité en termes de volume, qualité, rapidité d'accès. Il manque très souvent la capacité à nous donner les moyens de tirer parti de cette mémoire pour prendre des décisions.
- Utilisées pour des fonctions critiques : la majorité des systèmes existants est conçue dans le but unique de nous servir avec des temps de réponse corrects.

Le Tableau *Tab1* Présente les différences entre les données opérationnelles et les données décisionnelles.

Données opérationnelles	Données décisionnelles
Orientées application, détaillées, précise au moment de l'accès	Orientées activité (thème, sujet), condensées, représente des données historiques
Mise à jour interactive possible de la part des utilisateurs	Pas de mise à jour de la part des utilisateurs
Accédées de façon unitaire par une personne à la fois	Utilisées par l'ensemble des analystes, gérées par sous-ensembles
Cohérence atomique	Cohérence globale
Haute disponibilité en continu	Exigence différente, haute disponibilité ponctuelle
Unique (pas de redondance en théorie)	Peuvent être redondantes
Structure statique, contenu variable	Structure flexible
Petite quantité de données utilisées par un traitement	Grande quantité de données utilisées 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

Tab 1. Différences entre données du système de production et données décisionnelles

S'il existe effectivement des informations importantes, il n'en est pas moins nécessaire de construire une structure pour les héberger, les organiser et les restituer à des fins d'analyse.

Cette structure est « l'entrepôt de données » ou le Data Warehouse . Ce n'est pas une usine à produire l'information, mais plutôt un moyen de la mettre à la disposition des utilisateurs de manière efficace et organisée.

La mise en œuvre du Datawarehouse est un processus complexe. L'objectif à atteindre est de recomposer les données disponibles pour en donner :

- Une vision intégrée et transversale aux différentes fonctions de l'entreprise,
- Une vision métier au travers de différents axes d'analyse,
- Une vision agrégée ou détaillée suivant le besoin des utilisateurs.

Le Data Warehouse permet la mise en place d'un outil décisionnel s'appuyant sur les informations pertinentes pour l'entreprise, centrées sur le métier utilisateur.

3.3 Pourquoi pas un SGBD ?

Fonctions d'un SGBD (*Fig.1*)

- Systèmes transactionnels (OLTP),
- Permettre d'insérer, modifier, interroger rapidement, efficacement et en sécurité les données de la base,
- Sélectionner, ajouter, mettre à jour, supprimer des tuples,
- Répondre à de nombreux utilisateurs simultanément.

Fonctions d'un DW (*Fig.1*)

- Aider à la prise de décision (OLAP)
- Regrouper, organiser des informations provenant de sources diverses

- Intégrer et stocker les données pour une vue orientée métier
- Retrouver et analyser l'information rapidement et facilement

	OLTP	DW
Utilisateurs	Nombreux Employés	Peu Analystes
Données	Alphanumériques Détaillées / atomiques Orientées application Dynamiques	Numériques Résumées / agrégées Orientées sujet Statiques
Requêtes	Prédéfinies	« one-use »
Accès	Peu de données (courantes)	Beaucoup d'informations (historisées)
But	Dépend de l'application	Prise de décision
Temps d'exécution	Court	Long
Mises à jour	Très souvent	Périodiquement

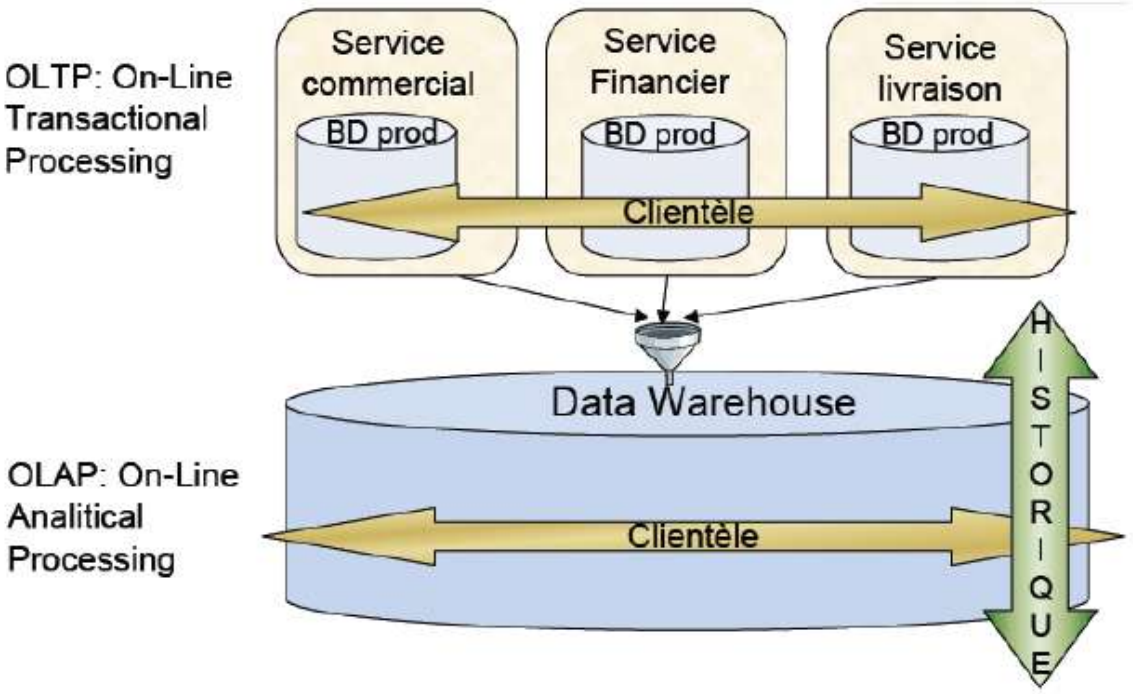


Fig. 1. OLTP vs DW

3.4. Les fonctions attendues d'un entrepôt de données

- L'entrepôt de données doit rendre les données de l'organisation facilement accessibles. Le contenu de l'entrepôt doit être facile à comprendre. Les données doivent être parlantes et leur signification évidente pour l'utilisateur. Pour être lisible, le contenu doit être étiqueté de manière significative.
- Les outils d'accès doivent être simples et faciles à utiliser, avec des temps de réponses minimes. Ils doivent permettre de séparer et combiner les données de toutes sortes de façons.
- L'entrepôt de données doit présenter l'information de manière cohérente. Les données doivent être crédibles, même si elles sont assemblées à partir de plusieurs sources d'information. Si deux mesures portent le même nom, elles doivent vouloir dire la même chose. Inversement, si deux mesures ne veulent pas dire la même chose, elles doivent avoir des noms différents. La cohérence implique une qualité des données élevée. Elle suppose que l'on a tenu compte de toutes les données, qu'elles sont complètes.
- L'entrepôt de données doit être adaptable et résistant aux changements. Les besoins des utilisateurs, les conditions d'activité, les données et la technologie sont en perpétuelle évolution. Ces modifications doivent être prises en compte par l'entrepôt de données, sans remettre en cause les données existantes. Elles ne doivent pas invalider les données existantes et les applications ne doivent pas être modifiées ou bouleversées lorsque les utilisateurs posent de nouvelles questions ou que de nouvelles données sont adjointes à l'entrepôt. Si les données descriptives de l'entrepôt doivent être modifiées, il faut pouvoir en rendre compte convenablement.
- L'entrepôt de données doit être protégé. Il contient de précieuses informations sur l'entreprise, ce qu'elle vend, à qui, à quel prix, quelles sont ses interrogations, etc. Il doit donc posséder un contrôle d'accès rigoureux aux informations confidentielles de l'organisation.
- L'entrepôt de données sert de socle à la prise de décision. Il doit contenir des données servant à étayer ces décisions.
- L'entrepôt de données doit être accepté par la communauté des utilisateurs.

4. Qu'est-ce qu'un entrepôt de données ?

4.1. Un entrepôt de données est une base de données

- Consolidant les données de bases de données opérationnelles,
- Utilisée en consultation et mise à jour périodiquement,
- Organisée pour permettre le traitement de requêtes "analytiques" plutôt que "transactionnelles" (OLAP par rapport à OLTP).
  - OLTP: On-line transaction processing. Petites transactions consistant en une recherche d'informations et, souvent, une mise à jour.
  - OLAP: On-line analytical processing. Grosses transactions impliquant une fraction importante des données réalisant, par exemple, un calcul statistique.

4.2. Intérêt de l’entrepôt de données

- Vision transversale de l’entreprise
- Intégration de différentes bases
- Données non volatiles (pas de suppression)
- Historisation
- Organisation vers prise de décision

4.3. Un projet complexe

- Rassembler des données hétérogènes
- Les homogénéiser et les restructurer
- Vérifier leur fiabilité
- Les éditer (publier)

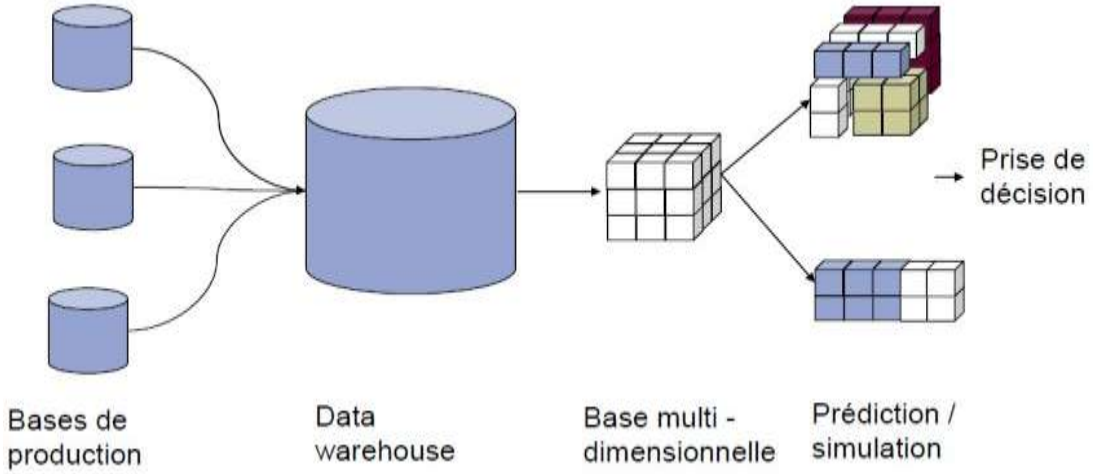


Fig. 2. Processus de prise de Décision

4.4. Exemples

a) Utilisation d'un entrepôt par une entreprise de distribution.

- Les données de vente sont enregistrées dans les différents magasins (OLTP).
- Chaque nuit, les données des différents magasins sont transférées dans un entrepôt de données au siège de la firme.
- Les données de l'entrepôt sont utilisées pour mettre au point des stratégies commerciales, des campagnes de promotion . . .

b) Marketing et gestion de la relation client (CRM)

- Qui sont mes clients?
- Pourquoi sont-ils mes clients?
- Comment les conserver ou les faire revenir?
- Ces clients sont-ils intéressants pour moi ?

c) Analyse des ventes dans les grandes surfaces utilisateurs de la partie décisionnelle du SI

d) Marketing, Actions Commerciales

- Où placer ce produit dans les rayons?
- Comment cibler plus précisément le mailing (communication publicitaire) concernant ce produit ?

4.5. Exemples de requêtes OLAP

a) Quel est le nombre de paires de chaussures vendues par le magasin X en mai 2003

b) Comparer les ventes avec le même mois de 2001 et 2002.

c) Quelles sont les composantes des machines de production ayant eu le plus grand nombre d'incidents imprévisibles au cours de la période 1992-97 ?

4.6. Objectifs

Toutes les données qu'elles proviennent du système de production de l'entreprise ou qu'elles soient achetées vont devoir être organisées, coordonnées, intégrées et stockées, pour donner à l'utilisateur une vue intégrée et orientée métier.

L'objectif d'un Data Warehouse est une sorte de point focal stockant en un endroit unique toute l'information utile provenant des systèmes de production et des sources externes.

Avant d'être chargée dans le Data Warehouse, l'information doit être extraite, nettoyée et préparée. Puis, elle est intégrée et mise en forme de manière compréhensible pour être comprise par l'utilisateur.

5. Types de Données dans un Data Warehouse

5.1. Définition

De nombreuses définitions ont été proposées, soit académiques, soit par des éditeurs d'outils, de bases de données ou par des constructeurs, cherchant à orienter ces définitions dans un sens mettant en valeur leur produit. La définition la plus consensuelle est celle de Bill Inmon, père fondateur du Data Warehouse, en 1990:

"*Un entrepôt de données (data Warehouse) est une collection de données thématiques, intégrées, non volatiles et historisées pour la prise de décisions*".

5.2. Les données sont thématiques

La vocation du Data Warehouse est de prendre des décisions autour des activités majeures de l'entreprise. Les données sont ainsi structurées autour de thèmes, ce qui facilite l'analyse transversale. Pour éviter le doublonnage des données, on regroupe les différents sujets dans une structure commune. Ainsi, si le sujet client contient des informations dans les sujets marketing, ventes, analyse financière, on regroupera ces trois sujets au sein du thème client. Dès lors, chaque donnée n'est présente qu'à un endroit et le Data Warehouse joue bien un rôle de point focal.

5.3. Les données sont intégrées

Elles proviennent de sources hétérogènes. Avant d'être intégrées dans le Data Warehouse, les données doivent êtres mises en forme et unifiées afin d'avoir un état cohérent. Par exemple, la consolidation de l'ensemble des informations concernant un client donné est nécessaire pour donner une vue homogène de ce client. Une donnée doit avoir une description et un codage unique (cohérence, normalisation, maîtrise de la sémantique, prise en compte des contraintes référentielles et des règles de gestion).

5.4. Les données sont historisées et non volatiles

- **Données historisées**
  Suivre dans le temps l'évolution des différentes valeurs des indicateurs → couches de données

- **Données non volatiles (**Traçabilité → non suppression)

  La non volatilité des données est en quelque sorte une conséquence de l'historisation. Une même requête effectuée à quelques mois d'intervalle en précisant la date de référence de l'information recherchée donnera le même résultat.

  Le Tableau *Tab2* présente les principales différences entre le système de production et le data Warehouse.

Critère	Système de production	Datawarehouse
Niveau de détail des informations utilisateurs	Très détaillé	Synthétique, parfois détaillé
Utilisateurs	Une ou quelques fonctions de l'entreprise	plusieurs fonctions de l'entreprise
Données figées	Non-évolution en temps réel	Oui-archivage
Historique	Non	Oui
Opérations sur les données	Mise à jour/consultation	Consultation uniquement

Tab. 2. Différences entre Système de Production et Data Warehouse

6. Les concepts de base

6.1 La structure

Un Data Warehouse se structure en quatre classes de données, organisées selon un axe historique et un axe synthétique (*Fig3*).

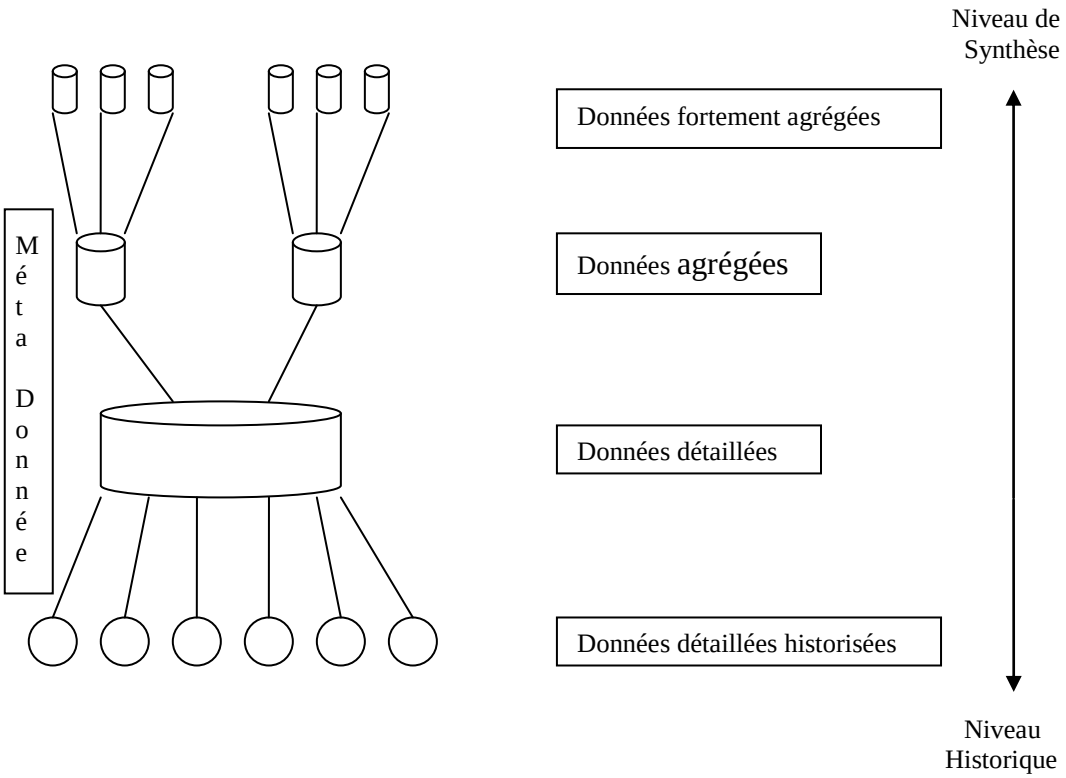


Fig. 3. Structure d'un Data Warehouse

- **Les Données Détaillées**

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

  Les volumes à traiter sont plus importants que ceux gérés en transactionnel. Attention: le niveau de détails géré dans le Data Warehouse n'est pas forcément identique au niveau de détails géré dans les systèmes opérationnels. La donnée insérée dans le Data Warehouse peut être déjà une agrégation ou une simplification d'informations tirées du système de production.

  Exemple: l'étude du panier de la ménagère nécessite de stocker le niveau de finesse du ticket de caisse.

- **Les Données Agrégées**

  Elles correspondent à des éléments d'analyse représentatifs des besoins 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. La facilité d'accès est apportée par des structures multidimensionnelles qui permettent aux utilisateurs de naviguer dans les données suivant une logique intuitive, avec des performances optimales. (Certains SGBD du marché sont conçus pour faciliter la mise en place des agrégations et la navigation au sein de celles-ci).

La définition complète de l'information doit être mise à la disposition de l'utilisateur pour une bonne compréhension. Dans le cas d'un agrégat, l'information est composée du contenu présenté (moyenne des ventes, …) et de l'unité (par mois, par produit,…).

- **Les Métadonnées**

  Elles regroupent l'ensemble des informations concernant le Datawarehouse et les processus associés. Elles constituent une véritable aide en ligne permettant de connaître l'information contenue dans le Datawarehouse. Elles sont idéalement intégrées dans un référentiel.

  Les principales informations sont destinées :

  - A l'utilisateur (sémantique, localisation).
  - Aux équipes responsables des processus de transformation des données du système de production vers le Datawarehouse (localisation dans les systèmes de production, description des règles, processus de transformation).
  - Aux équipes responsables des processus de création des données agrégées à partie des données détaillées.
  - Aux équipes d'administration de la base de données (structure de la base implémentant le Datawarehouse).
  - Aux équipes de production (procédures de changement, historique de mise à jour,…)

- **Les Données Historisées**

  Un des objectifs du Datawarehouse est de conserver en ligne les données historisées. Chaque nouvelle insertion de données provenant du système de production ne détruit pas les anciennes valeurs, mais crée un nouvelle occurrence de la donnée. Le support de stockage dépend du volume des données, de la fréquence d'accès, du type d'accès. Les supports les plus couramment utilisés sont les disques, les disques optiques numérique, les cassettes.

  La logique d'accès aux données la plus utilisée est la suivante : les utilisateurs commencent à attaquer les données par le niveau le plus agrégé, puis approfondissent leur recherche vers les données les plus détaillées (Drill Drown).

  L'accès des données se fait également directement par les données détaillées et historisées, ce qui conduit à des brassages de données lourds, demandant des machines très puissantes.

  Le Datawarehouse est une réussite dans une entreprise lorsque le nombre d'utilisateur accédant aux données de détail augmente.

6.2 Les Architectures

Pour implémenter un Data Warehouse, trois types d'architectures sont possibles :

- L'architecture réelle,
- L'architecture virtuelle,
- L'architecture remote.

a) L'architecture réelle

Elle est généralement retenue pour les systèmes décisionnels. Le stockage des données est réalisé dans un SGBD séparé du système de production. Le SGBD est alimenté par des extractions périodiques. Avant le chargement, les données subissent d'importants processus d'intégration, de nettoyage, de transformation (*Fig4*).

L'avantage est le fait de disposer de données **préparées** pour les besoins de la décision et répondant aux objectifs du Datawarehouse.

L'inconvénient est d'une part le coût de stockage supplémentaire et d'autre part le temps d'accès en réel.

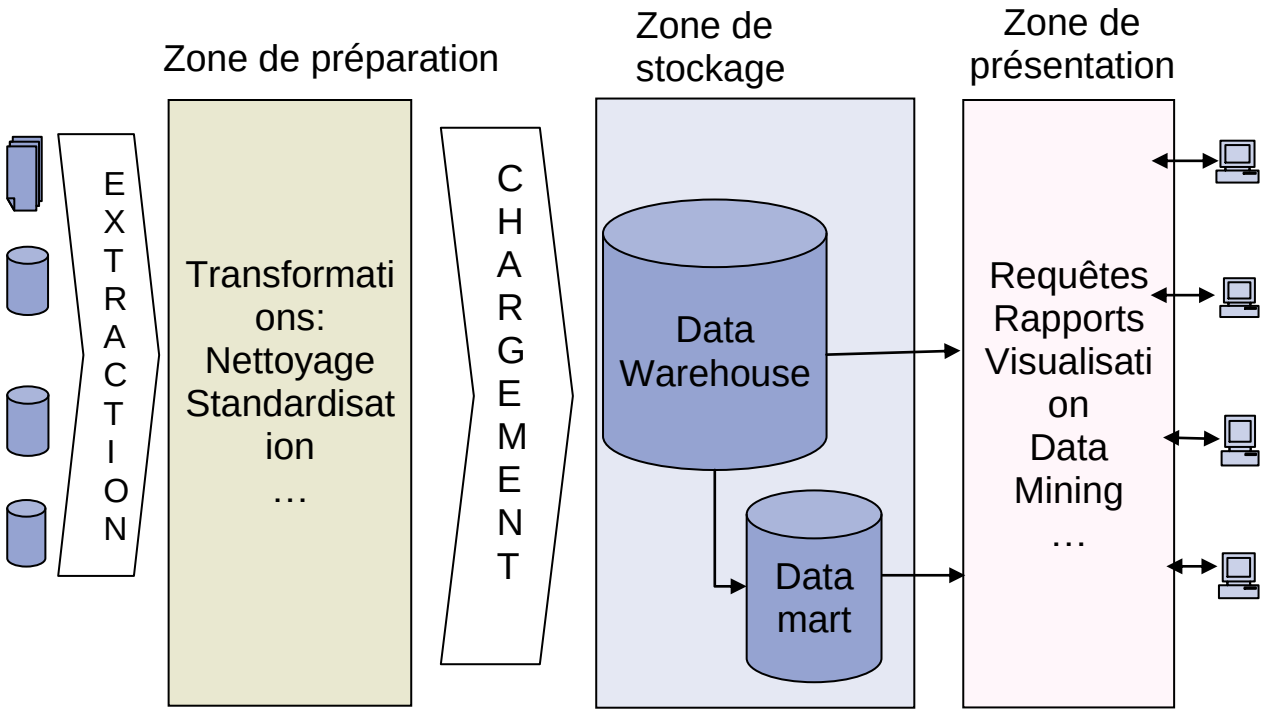


Fig. 4. Architecture d'un Data Warehouse

- **Les Flux de Données**
  - Flux entrant
  - Extraction: multi-source, hétérogène
  - Transformation: filtrer, trier, homogénéiser, nettoyer
  - Chargement: insertion des données dans l'entrepôt
  - Flux sortant
  - Mise à disposition des données pour les utilisateurs finaux
- **Les Différentes Zones de l'Architecture**
  - Zone de préparation (Staging area)
  - Zone temporaire de stockage des données extraites
    - Réalisation des transformations avant l'insertion dans le DW
    - Nettoyage
    - Normalisation…
    - Données souvent détruites après chargement dans le DW
  - Zone de stockage (DW, DM)
    - On y transfère les données nettoyées

- Contient les données de l'entreprise

– Zone de présentation

- Donne accès aux données contenues dans le DW
- Peut contenir des outils d'analyse programmés: Rapports, Requêtes…

b) L'architecture virtuelle

Cette architecture n'est pratiquement pas utilisée pour le Datawarehouse. Les données résident dans le système de production. Elles sont rendues visibles par des produits middleware ou par des passerelles.

Il en résulte deux avantages : pas de coût de stockage supplémentaire et l'accès se fait en temps réel.

L'inconvénient est que les données ne sont pas préparées.

c) L'architecture REMOTE

C'est une combinaison de l'architecture réelle et de l'architecture virtuelle. Elle est rarement utilisée.

L'objectif est d'implémenter physiquement les niveaux agrégés afin d'en faciliter l'accès et de garder le niveau de détail dans le système de production en y donnant l'accès par le biais de middleware ou de passerelle.

d) Synthèse

Les différents éléments d'appréciation sont repris dans le tableau récapitulatif *Tab. 3*.

	Architecture réelle	Architecture virtuelle	Architecture remote
Utilisation	Retenue pour les systèmes décisionnels	Rarement utilisée	Rarement utilisée
Stockage	SGBD séparé du système de production, alimenté par es extractions périodiques	Données résidant dans le système de production	Combinaison des architectures réelles et virtuelles
Avantages	Données préparées pour les besoins de la décision	Pas de coût de stockage supplémentaire, accès en temps réel	
Inconvénients	Coût de stockage supplémentaire, manque d'accès temps réel	Données non préparées	

Tab. 3. Tableau de Synthèse des Architectures de Data Warehouse

7. Les Magasins de Données (Datamarts)

Le Datamart minimise la complexité informatique. Il est donc plus facile de se concentrer sur les besoins utilisateurs.

Définition

Le Datamart est une base de données moins coûteuse que le Datawarehouse, et plus légère puisque destinée à quelques utilisateurs d'un département. Il séduit plus que le data Warehouse les candidats au décisionnel.

C'est une petite structure très ciblée et pilotée par les besoins utilisateurs. Il a la même vocation que le Data Warehouse (fournir une architecture décisionnelle), mais vise une problématique précise avec un nombre d'utilisateurs plus restreint.

En général, c'est une petite base de données (SQL ou multidimensionnelle) avec quelques outils, et alimentée par un nombre assez restreint de sources de données.

Mais pour réussir, il y a quelques précautions à prendre, gage de son évolutivité vers le Data Warehouse.

Donc le Datamart peut préparer au Datawarehouse. Mais il faut penser grand, avenir, et adopter des technologies capables d'évoluer.

En résumé un Datamart est :

- Un sous-ensemble d'un entrepôt de données
- Destiné à répondre aux besoins d'un secteur ou d'une fonction particulière de l'entreprise
- Point de vue spécifique selon des critères métiers

8. Histoire

Les principales dates à retenir construisant l'histoire de l'Entrepôt de données sont les suivantes :

- Années 1960 - General Mills et l'Université Dartmouth, dans un projet conjoint, créent les termes "faits" et "dimensions".
- 1983 - Teradata introduit dans sa base de données managériale un système exclusivement destiné à la prise de décision.
- 1988 - Barry Devlin et Paul Murphy publient l'article "Une architecture pour les systèmes d'information financiers" ("An architecture for a business and information systems") où ils utilisent pour la première fois le terme "Datawarehouse".
- 1990 - Red Brick Systems crée Red Brick Warehouse, un système spécifiquement dédié à la construction de l'Entrepôt de données.
- 1991 - Bill Inmon publie Building the Datawarehouse (Construire l'Entrepôt de Données).
- 1995 - Le Datawarehousing Institute, une organisation à but lucratif destinée à promouvoir le datawarehousing, est fondé.
- 1996 - Ralph Kimball publie The Data Warehouse Toolkit (La boîte à outils de l'Entrepôt de données).

CHAPITRE 2

Modélisation et implémentation d'un entrepôt de données

1. Modélisation dimensionnelle

Les données d'un Data Warehouse sont de deux types:

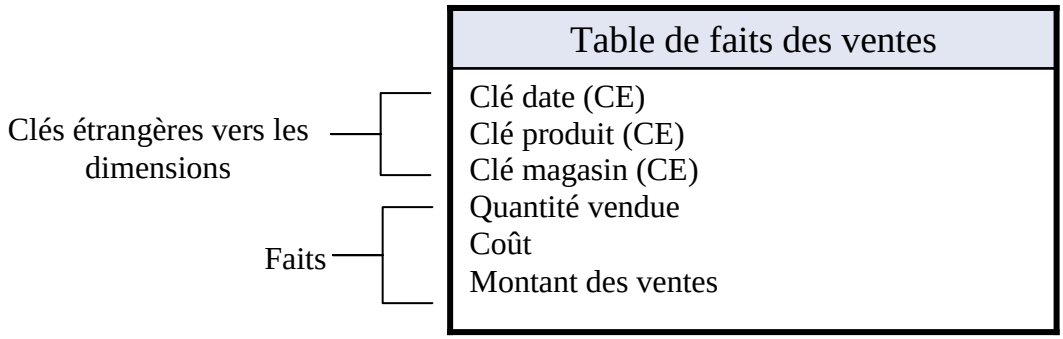
- Les faits: grosse accumulation de données reprenant des faits simples. Exemple: chiffres de ventes,
- Les données "dimensionnelles": données en quantité réduite, souvent statiques qui précisent des informations sur les éléments apparaissant dans les faits.

La modélisation dimensionnelle sert à modéliser l'activité que l'on souhaite analyser.

1.1. Table des faits (clé multiple)

- Table principale du modèle dimensionnel
- Contient les données observables (les faits ou un agrégat de faits) sur le sujet étudié selon divers axes d'analyse (les dimensions), autrement dit elle Contient un ou plusieurs faits numériques qui se produisent pour la combinaison de clés définissant chaque enregistrement

Exemple



Fait: Ce que l'on souhaite mesurer: Quantités vendues, …
Table des faits : Contient les clés des axes d'analyse (dimension) et les faits.

1.2. Typologie des faits

- **Additif**: additionnable suivant toutes les dimensions
  - Quantités vendues, chiffre d'affaire
  - Peut être le résultat d'un calcul:
  - Bénéfice = montant vente - coût
- **Semi additif**: additionnable suivant certaines dimensions
  - Solde d'un compte bancaire:
  - Pas de sens d'additionner sur les dates car cela représente des instantanés d'un niveau
  - Σ sur les comptes: on connaît ce que nous possédons en banque
- **Non additif**: fait non additionnable quelque soit la dimension
  - Prix unitaire: l'addition sur n'importe quelle dimension donne un nombre dépourvu de sens

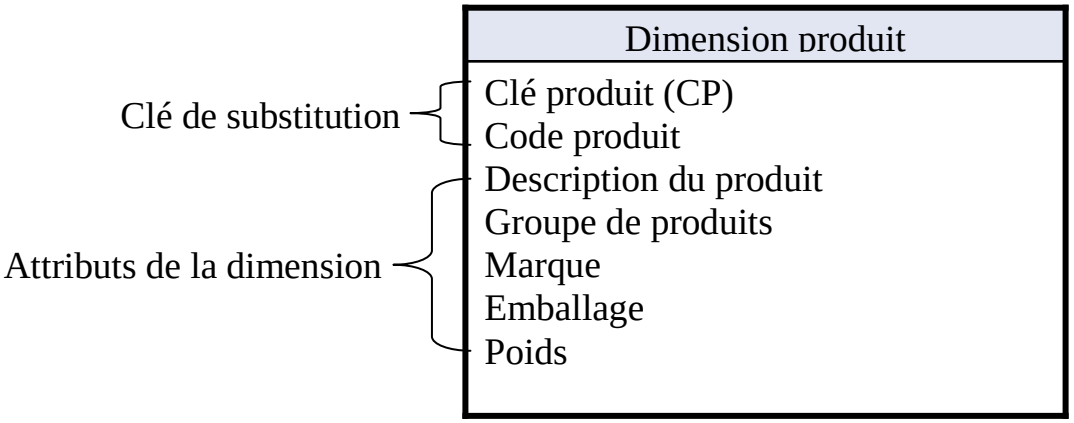
1.3. Table "Dimension" (Dimension = Axe d'analyse)

- Axe d'analyse selon lequel vont être étudiées les données observables (faits)
- Contient les détails sur les faits

- Sa clé est primaire et correspond à l'un des composants de la clé multiple de la table de faits.

Exemple: Client, Produit, Période de temps, …

Une table Dimension contient souvent un grand nombre de colonnes, un ensemble d'informations descriptives des faits et beaucoup moins d'enregistrements qu'une table de faits.



a) Dimensions et Indicateurs

- **Dimensions**
  - Produit
  - Client
  - Vendeur
  - Date
- **Indicateur**
  - Chiffre d'affaires

Une dimension prend une liste de valeurs, un indicateur est un nombre.

b) Hiérarchie de dimensions

Mois ⟶ ***Semaine*** ⟶ ***Jour***

2. Les types de modèles

2.1. Modèle en flocons (snowflake schema)

La représentation directe d'un contexte dimensionnel dans une base de données relationnelle est un réseau de tables jointes selon un schéma en flocon. Dans ce mode de représentation l'association conceptuelle qui contient les faits devient la table de faits, et chacune des entités dimensionnelles devient une table distincte (Fig. 6).

La table de faits contient en plus des indicateurs significatifs qu'elle comporte par définition, un ensemble de clés étrangères, dont chacune assure la liaison avec la table du niveau le plus fin de chaque dimension.

La table des faits est généralement une très grande table, puisqu'elle comporte autant d'enregistrements qu'il existe de combinaisons pertinentes entre les tables "Dimension". Dans le cas de la figure 1, le nombre d'enregistrements de la table de faits "Activité" peut théoriquement être égal au produit du nombre d'Etablissements par le nombre de Produits et par le nombre de Jours de l'historique mémorisé.

C'est une borne maximum, car il n'y a pas nécessairement eu d'activité pour chaque combinaison possible. Même si le nombre d'activité réelle est une faible proportion de ce maximum, la table de faits a pratiquement toujours une taille supérieure d'un ordre de

grandeur à la taille de la plus grande table dimensionnelle. Elle occupe en général 95 à 99 % du volume total de la base de données.

La génération de clés techniques est impérative. Pour être logiquement connectée, une table de faits doit posséder une clé pour chaque dimension. Dans chaque enregistrement d'une table de faits, les clés prennent une place importante. Si la table de faits possède des centaines de milliers, voire de millions, d'enregistrements, l'espace occupé par les clés dans la base de données est loin d'être négligeable. Il faut donc chercher à minimiser cet espace.

Il ne faut donc pas chercher à utiliser les clés signifiantes, qui ne sont pas faites pour économiser de la place, mais pour signifier quelque chose. Il faut utiliser les clés techniques numériques, générées éventuellement lors du chargement dans la base de diffusion (ou dans l'entrepôt de données).

Le format des clés doit être homogène, le plus petit possible compte-tenu de la cardinalité de chaque table. Il faut penser également aux possibilités d'extension de la base en volume. (Fig. 5).

En résumé, le modèle en flocons contient:

- Une table de fait et des dimensions décomposées en sous hiérarchies
- Plusieurs niveaux de Tables de Dimension
- 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.

Avantages

- Normalisation des dimensions
- Économie d'espace disque

Inconvénients

- Modèle plus complexe (jointures)
- Requêtes moins performantes

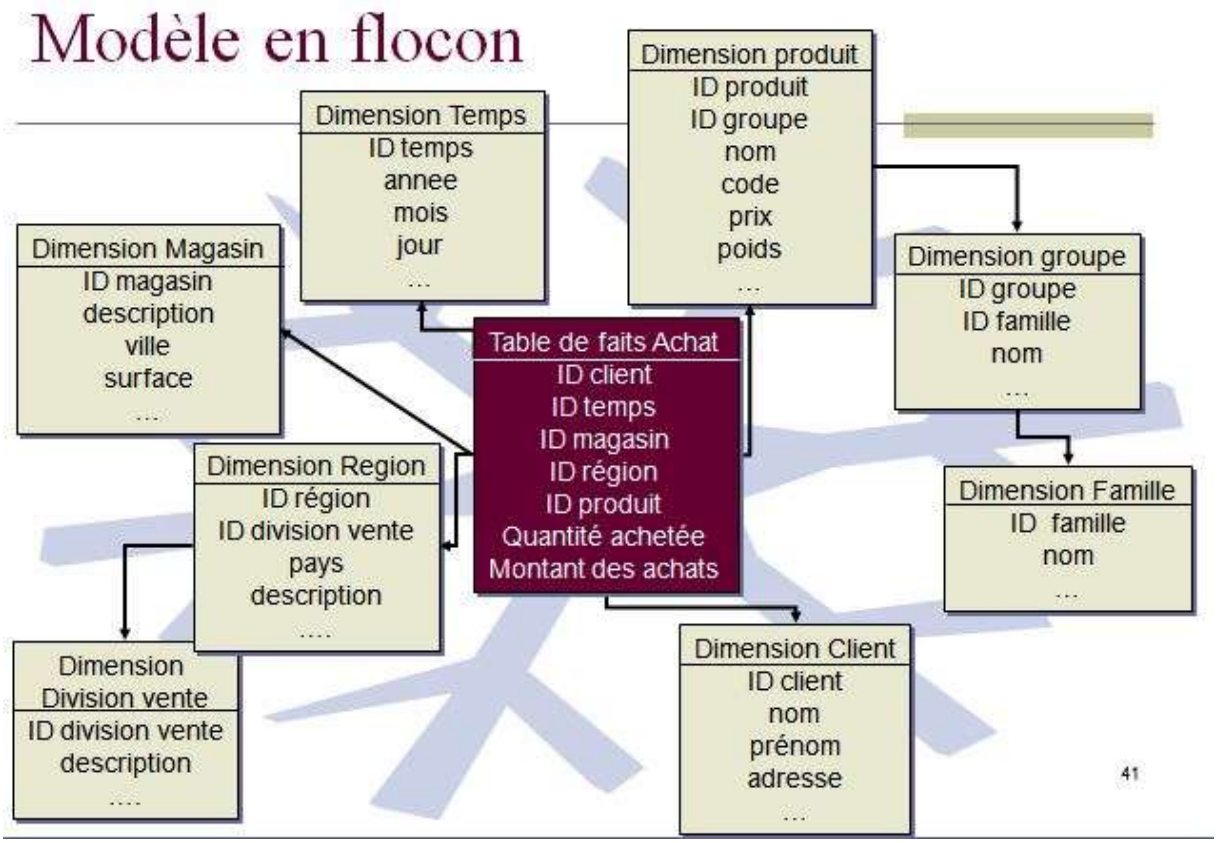


Fig. 5. Exemple de Schéma en flocons

2.2. Modèle en étoile (star schema)

Dans la pratique on préfère une forme dénormalisée du schéma en flocon : le schéma en étoile.

La figure 6 représente le schéma en étoile, dérivé du même modèle que le schéma en flocons de la figure 5.

Le schéma en étoile ne comporte, en plus de la table de faits qu'une table par dimension. Cette simplification est obtenue au prix d'une forte dénormalisation.

Dans la dimension « Client », par exemple, toutes les propriétés descriptives de l'Entreprise et du Groupe sont regroupées dans la même table que les propriétés de l'Etablissement. Dans le cas d'un groupe contrôlant 100 établissements, la description du Groupe sera répétée dans 100 enregistrements.

Ce modèle est donc générateur d'une forte redondance mais:

- La redondance des données ne compromet pas la cohérence d'une base ne subissant pas de mise à jour transactionnelle,
- L'espace occupé par les tables dimensionnelles est insignifiant par rapport au volume de la table de faits,
- Toutes les tables dimensionnelles ont une liaison directe avec la table de faits. Quelle que soit la complexité des dimensions, le nombre de tables pouvant être impliquées dans une requête, en plus de la table de faits, est inférieur ou égal au nombre de dimensions du contexte. Le temps d'exécution d'une requête est indépendant du niveau hiérarchique des propriétés conditionnelles invoquées.

En résumé, le modèle en étoile contient:

- Une table de fait centrale.
- Un ensemble de tables de Dimension (1 niveau).
- Les dimensions n'ont pas de liaison entre elles.

Avantages

- Facilité de navigation
- Nombre de jointures limité

Inconvénients:

- Redondance dans les dimensions
- Toutes les dimensions ne concernent pas les mesures

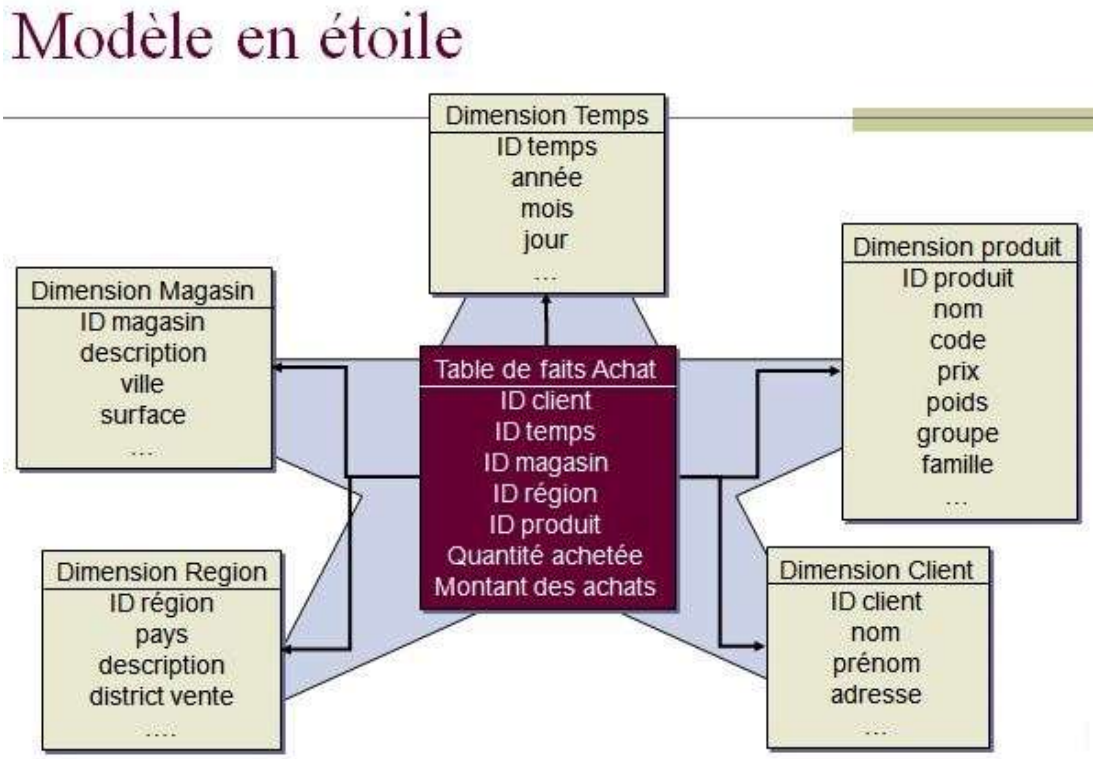
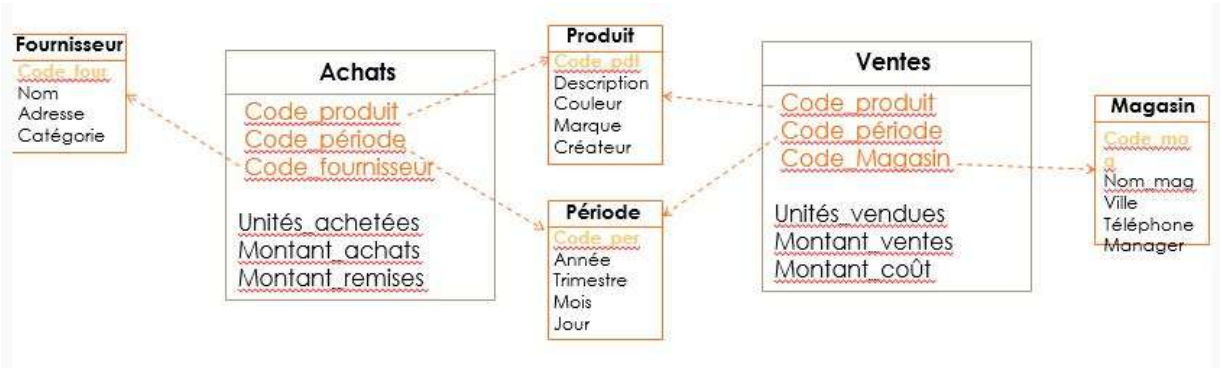


Fig. 6. Exemple de Modèle en Etoile

2.3 Le Modèle en Constellation de faits

Ce modèle est un ensemble de schémas en étoiles et/ou en flocon dans lesquels les tables de faits se partagent certaines tables de dimensions. C'est de cette accumulation que découle un modèle en constellation.



3. BD Multidimensionnelles et OLAP

3.1. Introduction

La vocation d'un entrepôt de données est l'analyse de données pour l'aide à la décision dans les entreprises. La modélisation multidimensionnelle est la base des entrepôts de données et de l'analyse multidimensionnelle (OLAP). Donc, la modélisation multidimensionnelle est une réponse à un besoin analytique.

Les bases de données relationnelles, modélisées selon les principes classiques de normalisation, s'adaptent très mal à un contexte analytique (OLAP). En analyse, l'utilisateur doit disposer d'un modèle relativement intuitif et capable de stocker le résultat de nombreux calculs d'agrégation.

L'intérêt pour l'analyse de données s'est développé énormément ces dernières années. Les entreprises se sont rendues compte de l'efficacité de la technologie OLAP (On-line Analytical Processing) dans l'analyse et l'exploration des données. Cette technologie est utilisée dans les systèmes d'aide à la décision. Le plus souvent, ces systèmes sont basés sur des techniques d'entreposage de données pour exploiter la grande masse d'informations disponibles dans les entreprises à des fins d'analyse et d'aide à la décision.

La modélisation multidimensionnelle propose donc d'analyser des *indicateurs* numériques (par exemple chiffre d'affaires, nombre d'individus, ratios, etc.) dans un contexte précisé par le croisement de plusieurs *dimensions* d'analyse (par exemple temps, géographie, organisation, produits, ...).

Par exemple, considérons les trois dimensions Temps, Pays et Produits, utilisées pour analyser les ventes. L'indicateur Chiffre d'Affaires sera calculable sur l'ensemble des combinaisons possibles entre ces trois axes. L'ensemble des combinaisons possibles peut être représenté par un cube.

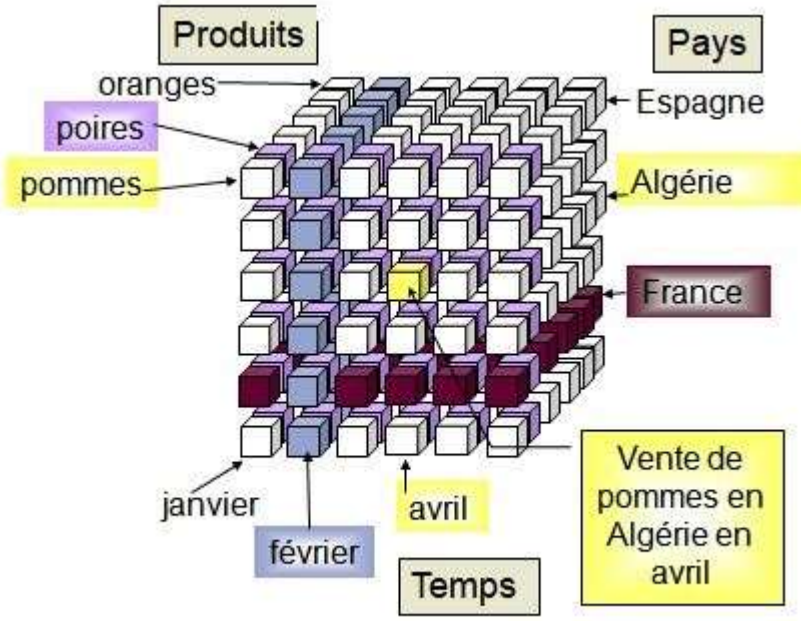


Fig. 7. Exemple de Cube de Données

Au-delà de trois dimensions, cela devient mathématiquement un hypercube (qu'il est beaucoup plus difficile de représenter graphiquement). Une base de données multidimensionnelle typique peut donc s'envisager comme un hypercube d'une dizaine de dimensions comprenant plusieurs millions de cellules (on parle plus couramment de *Cube OLAP*).

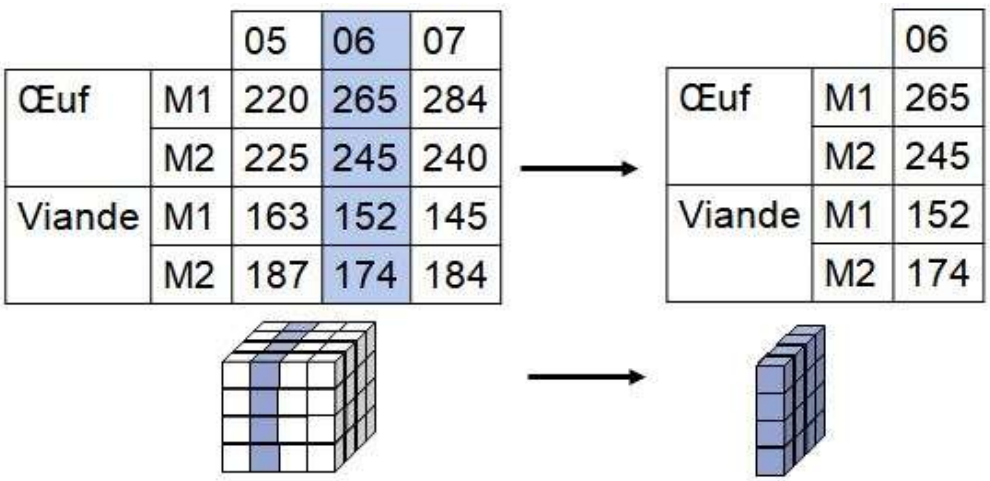
3.2. OLAP (On-Line Analytical Processing)

L'OLAP ou Online Analytical Processing est une technique informatique d'analyse multidimensionnelle, qui permet aux décideurs, d'avoir accès rapidement et de manière interactive à une information pertinente présentée sous des angles divers et multiples, selon leurs besoins particuliers. L'OLAP est donc une technique dont les fonctionnalités servent à faciliter l'analyse multidimensionnelle: opérations réalisables sur l'hypercube pour extraire les données.

a) Opérations OLAP

- **Opérations Agissant sur la Structure**

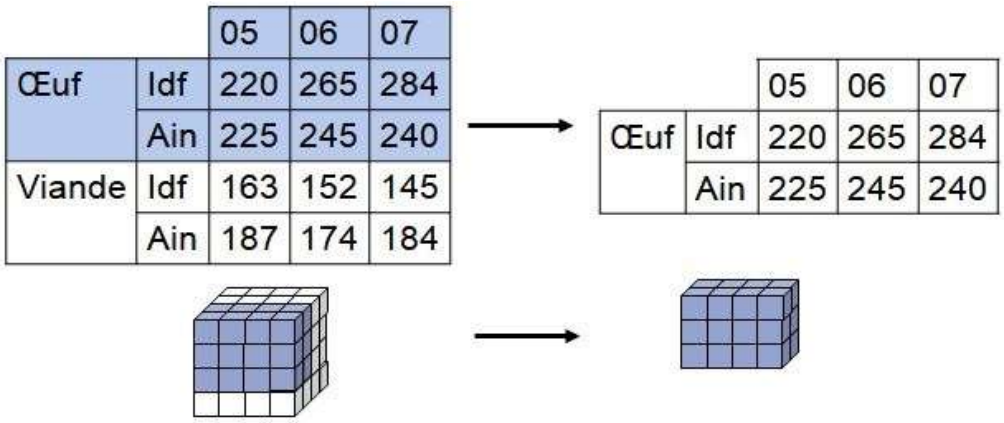
  Tranchage (slicing): Consiste à ne travailler que sur une tranche du cube. Une des dimensions est alors réduite à une seule valeur.



		05	06	07
Œuf	M1	220	265	284
	M2	225	245	240
Viande	M1	163	152	145
	M2	187	174	184

		06
Œuf	M1	265
	M2	245
Viande	M1	152
	M2	174

Extraction d'un bloc de données (dicing): Consiste à travailler uniquement sur un sous-cube.



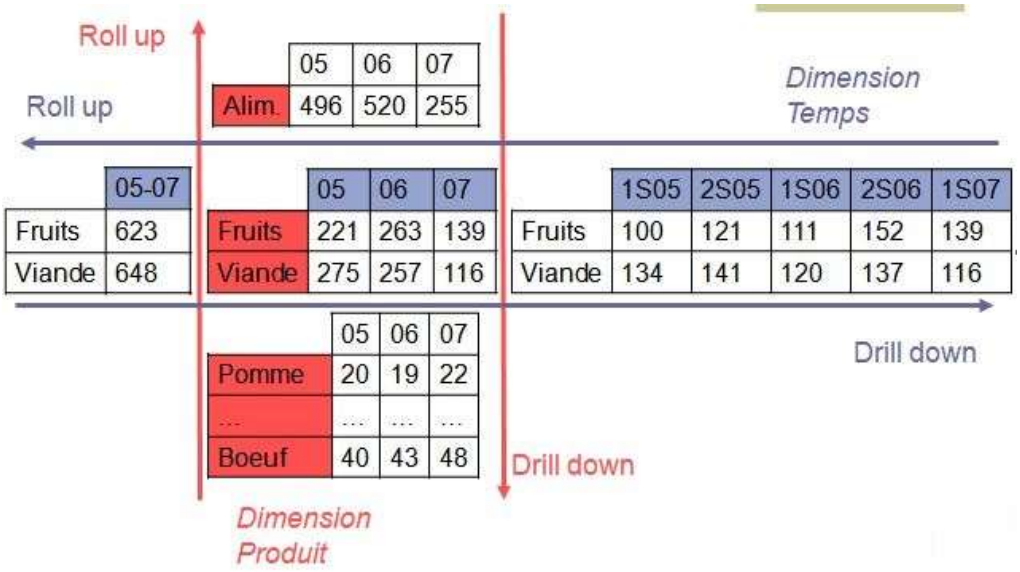
- **Opérations agissant sur la Granularité**

  Forage vers le haut (roll-up): « dézoomer »

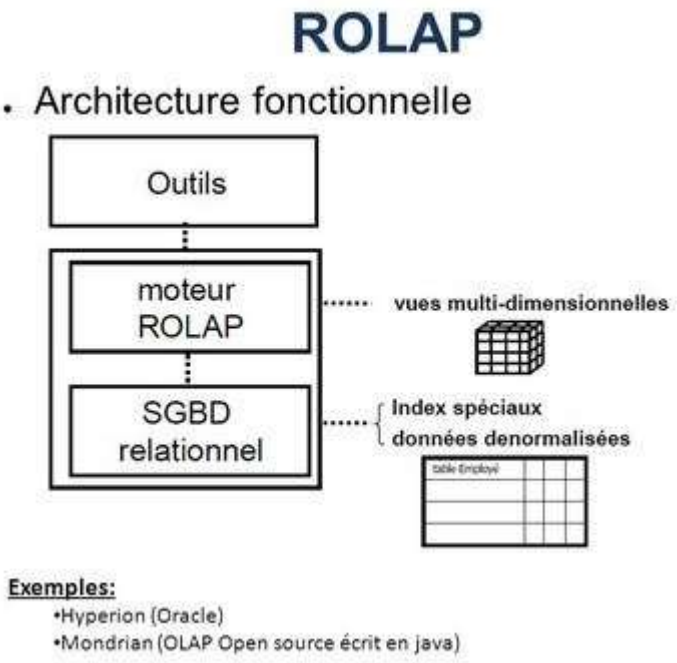
  - Obtenir un niveau de granularité supérieur
  - Utilisation de fonctions d'agrégation

  Forage vers le bas (drill-down): « zoomer »

  - Obtenir un niveau de granularité inférieur
  - Données plus détaillées



- **ROLAP: Relational OLAP**



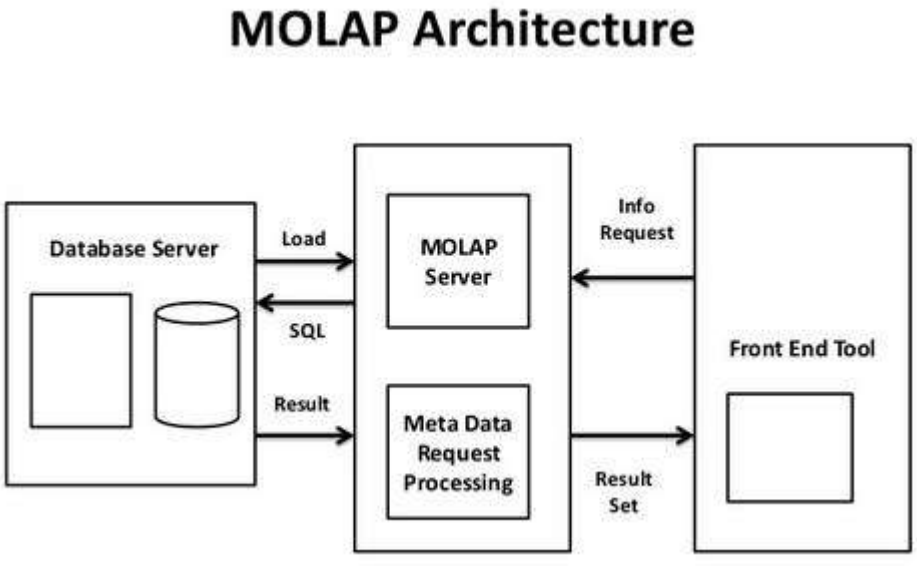
L'obtention des données se fait via des tables relationnelles et des jointures qui vont avec celles-ci. Donc, la requête créée sera relativement complexe, selon la granularité, et, sera d'une longueur plus ou moins importante. Comme le résultat n'est pas stocké, à chaque consultation, la requête devra être relancée à chaque consultation la requête devra être relancée.

Les différents inconvénients de la méthode ROLAP: Le temps de réponse est d'une longueur assez conséquente étant donné que les requêtes fonctionnent via des tables. Les bases sont donc utilisées à chaque relance du rapport.

Les avantages de la méthode ROLAP: Le coût est relativement faible, en effet, cette méthode utilise des ressources déjà existantes comme des ressources matérielles, des licences etc.

Exemples: Microsoft Analysis Services, Oracle 10g, MetaCube d'Informix, Mondrian de Pentaho, DSS Agent de MicroStrategy.

- **MOLAP: Multidimentional OLAP**



On stock les données dans un CUBE qui est en fait une base de données multidimensionnelles. De cette façon, le concept de relationnel n'est plus présent. Pré calculer tous les croisements envisageables est l'objectif de cette base de données multidimensionnelle, de cette manière la restitution des données se fait de façon instantanée. Les données étant stockées, le temps gagné pendant la restitution des données sera considérable.

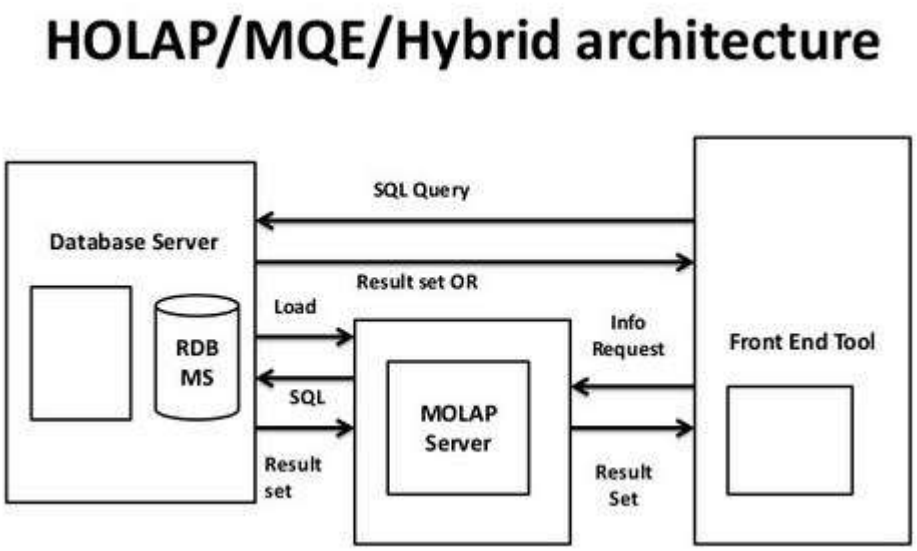
Inconvénients des cubes MOLAP : Le coût est important, en effet, elle nécessite souvent des licences pour les bases multidimensionnelles et des coûts pour le développement des CUBES.

Avantage des cubes MOLAP: Le temps de réponse est extrêmement court car la totalité des données sont stockées au sein d'un CUBE.

Remarque: Les bases de données multidimensionnelles possèdent leur propre langage permettant de faire des requêtes, appelé le MDX, qui est l'équivalent du SQL utilisé pour les bases de données relationnelles.

Exemples: Board M.I.T., Essbase, IBM TM1, Jedox Palo, icCube server, Infor Alea, Microsoft Analysis Services, Oracle OLAP.

- **HOLAP: Hybrid OLAP**



L'HOLAP est un mélange du ROLAP et du MOLAP. Les cubes HOLAP sont donc Hybrides. On se sert du MOLAP lorsque l'on veut accéder aux données agrégées. Si l'on souhaite arriver à un niveau de détail plus important, nous utilisons le ROLAP.

Par exemple, les données sont stockées et accessible via un Cube multidimensionnel, mais on fait également de la restitution via un outil de reporting comme SSRS par exemple. L'utilisateur pourra donc avoir accès à un rapport contenant les données issues du CUBE ainsi qu'à un autre rapport détaillé contenant les données en provenance de tables, cette fois relationnelles.

Inconvénients de la méthode HOLAP: Elle est inutilisable en cas de complexité trop élevée des rapports ou qu'ils fassent appel à trop de croisements de données. **Avantages de la méthode HOLAP**: Un investissement financier moindre que la méthode MOLAP, en effet la partie développement sera beaucoup moins importante. De plus le temps de réponse est relativement court.

Exemple: Oracle OLAP, Microsoft Analysis Services.

CHAPITRE 3

Alimentation du Data Warehouse

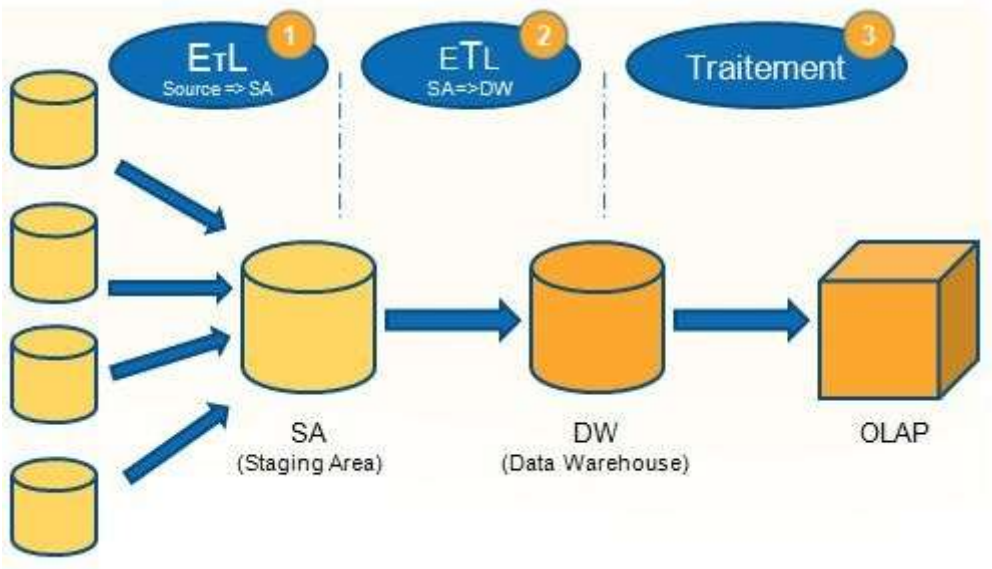
1. Introduction

Après la conception de l'ED, on a besoin d'Acquérir des données pour l'alimentation ou la mise à jour régulière de l'entrepôt.

Besoin d'un outil pour automatiser les chargements de l'entrepôt : ETL

ETL, acronyme de *Extraction, Transformation, Loading* (ou : *data pumping*), en français (Extraction, Transformation et Chargement) est un système de chargement de données depuis les différentes sources d'information de l'entreprise (hétérogènes) jusqu'à l'entrepôt de données.

Il est important de savoir que la réalisation de l'ETL constitue 70% d'un projet décisionnel en moyenne. Et ce n'est pas pour rien, ce système est complexe et ne doit rien laisser s'échapper, sous peine d'avoir une mauvaise information dans l'entrepôt, donc des données fausses, donc inutilisables.



Les sources de données peuvent être de plusieurs types:

- **Enterprise ressource planning (ERP)**

  Gèrent les processus opérationnels d'une entreprise (ex: ressources humaines, finances, distribution, approvisionnement, etc.).

- **Customer Relationship Management (CRM)**

  Gèrent les interactions d'une entreprise avec ses clients (ex: marketing, ventes, après-vente, assistance technique, etc.).

- **Systèmes Legacy (systèmes légataires)**

  Matériels et logiciels dépassés mais difficilement remplaçables.

- **Point of sale (POS) (point de vente)**

  Matériels et logiciels utilisés dans les caisses de sorties d'un magasin.
- **Externes**

  Données concurrentielles achetées, données démographiques.

  Les sources de données peuvent présenter des problèmes tels que:

- Sources diverses et hétérogènes;
- Sources sur différentes plateformes et OS;
- Applications *legacy* utilisant des BD et autres technologies obsolètes;
- Historique de changement non-préservé dans les sources;
- Qualité de données douteuse et changeante dans le temps;
- Structure des systèmes sources changeante dans le temps;
- Incohérence entre les différentes sources;
- Données dans un format difficilement interprétable ou ambigu.

2. Définitions

Les processus ETL sont les composants les plus critiques et les plus importants d'une infrastructure décisionnelle. Bien que cachés de l'utilisateur de la plate-forme décisionnelle, les processus ETL rassemblent les données à partir des systèmes opérationnels et les prétraitent pour les outils d'analyse et de reporting. La précision et la vitesse de la plate-forme décisionnelle toute entière dépendent des processus ETL

Ils regroupent plusieurs étapes, qui ont pour objet de transférer des données depuis les applications de production vers les systèmes décisionnels :

L'alimentation se déroule en 3 phases. Ces trois étapes décrivent une mécanique cyclique qui a pour but de garantir l'alimentation du Datawarehouse en données homogènes, propres et fiables.

2.1. Extraction des données

But: Extraction de données des applications et des bases de données de production (ERP, CRM, SGBDR, fichiers, ...)

L'extraction est la première étape du processus d'apport de données à l'entrepôt de données. Extraire, cela veut dire lire et interpréter les données sources et les copier dans la zone de préparation en vue de manipulations ultérieures.

Elle consiste en :

- Cibler les données,
- Appliquer les filtres nécessaires,
- Définir la fréquence de chargement,

Lors du chargement des données, il faut extraire les nouvelles données ainsi que les changements intervenus sur ces données. Pour cela, il existe trois stratégies de capture de changement :

- **Colonnes d'audit**: la colonne d'audit, est une colonne qui enregistre la date d'insertion ou du dernier changement d'un enregistrement. Cette colonne est mise à jour soit par des triggers ou par les applications opérationnelles, d'où la nécessité de vérifier leur fiabilité.
- **Capture des logs**: certains outils ETL utilisent les fichiers logs des systèmes sources afin de détecter les changements (généralement logs du SGBD). En plus de l'absence de cette fonctionnalité sur certains outils ETL du marché, l'effacement des fichiers logs engendre la perte de toute information relative aux transactions.
- **Comparaison avec le dernier chargement**: le processus d'extraction sauvegarde des copies des chargements antérieurs, de manière à procéder à une comparaison lors de

chaque nouvelle extraction. Il est impossible de rater un nouvel enregistrement avec cette méthode. L'extraction des données des sources hétérogènes nécessite d'identifier les sources utiles et de comprendre les schémas.

2.2. Transformation des données

But: Rendre cohérentes les données issues de différentes sources C'est une suite d'opérations qui a pour but de rendre les données cibles homogènes pour être traitées de façon cohérente.

Cette opération se solde par la production d'informations dignes d'intérêt pour l'entreprise et les données sont donc prêtes à être entreposées.
La transformation de ces données vise à les réconcilier entre les différentes sources, pour effectuer des calculs ou du découpage de texte, pour les enrichir avec des données externes et aussi pour respecter le format requis par les système cibles (Troisième Forme Normale, Schéma en Etoile, etc.)

Avant de commencer, il faut visualiser le schéma d'un entrepôt et sa façon de fonctionner (gérer l'historique, dimensions, faits, etc.). Le but du jeu est de faire rentrer toutes les données de l'entreprise dans ce modèle, les données doivent donc être :

- **Dé-normalisées**: Dans un DW (datawarehouse), avoir des doublons n'est pas important; avoir un schéma en troisième forme normale est même déconseillé. Il faut que les données apparaissent là où elles doivent apparaître.
- **Nettoyées**: Dans un système de production, les utilisateurs introduisent les données. Les risques d'erreurs sont donc élevés. Ces erreurs ont des répercutions directes sur les analyses. Il faut pouvoir détecter et corriger ces erreurs.
- **Contextualisées**: Imaginez un système de production, où les informations sur l'activité du personnel sont enregistrées, et un système de RH où les informations personnelles des employés sont stockées. Un entrepôt de données possède une vision universelle, les informations relatives à un employé ne seront stockées qu'une seule dans une seule dimension.

La transformation consiste donc en:

- Une détection et correction d'erreurs,
- Une discrétisation, réduction, normalisation,
- Une détection de redondance, fusion, intégration,
- Une transformation dans le modèle cible.

Exemple

Donnés sources	**données cibles (après intégration)**
Source 1 : male, femelle	**m, f**
Source 2 : 1, 0	**m, f**
Source 3 : Masculin, féminin	**m, f**

Donnés sources	**données cibles (après intégration)**
Source 1 : $150,000	**1 050 000 DA**
Source 2 : 16 000 €	**1 600 000 DA**
Source 3 : 200.000 RS	**4 000 000 DA**

2.3 Chargement

But: Introduire les données dans l'entrepôt.

C'est la dernière phase de l'alimentation d'un entrepôt de données, le chargement est une étape indispensable. Elle reste toutefois très délicate et exige une certaine connaissance des structures du système de gestion de la base de données (tables et index) afin d'optimiser au mieux le processus.

Le chargement est l'étape la plus complexe, il s'agit ici d'ajouter les nouvelles lignes, voir si des lignes ont été modifiées et faire une gestion d'historique, voir si des lignes ont été supprimées et le mentionner dans l'entrepôt, tout en faisant attention de ne pas charger des données en double.

La latence des processus d'ETL varie, du mode batch (traitement effectué par lot) (parfois mensuel ou hebdomadaire, le plus souvent quotidien) jusqu'au quasi-temps réel avec des rafraîchissements plus fréquents (toutes les heures, toutes les minutes, etc.).

3. Conception d'un ETL

Il n'existe pas de méthodes de conception d'ETL. Chaque entreprise possède ses propres systèmes, sa propre logique de fonctionnement et sa propre culture. Un ETL va prendre toutes les données de l'entreprise et les charger dans un DW.

3.1. Comment procéder?

Deux cas sont à prendre en compte, le chargement initial et les chargements incrémentiels.

Le chargement initial est effectué au tout premier chargement de l'entrepôt et dans des cas spéciaux comme après la perte des données de l'entrepôt. Dans ce cas, on charge toutes les données de l'entreprise dans l'entrepôt.

Le chargement incrémentiel est le fait d'ajouter des données à un entrepôt existant, c'est l'opération qui va se répéter dans le temps (chaque jour par exemple). Il faudra faire attention dans ce cas à ne charger que les informations nouvelles, et ne pas charger deux fois la même information.

3.2. Comment sont mes sources ?

Avant de faire un ETL, il faut bien étudier les sources de données. En effet, c'est d'après les sources que les stratégies de chargement vont se faire.

3.2.1 Politiques de l'alimentation

Le processus de l'alimentation (rapatriement des données) peut se faire de différentes manières. Le choix de la politique de chargement dépend des disponibilités et des accessibilités des sources de données. Ces politiques sont:

- **Push:** dans cette méthode, la logique de chargement est dans le système de production. Il "pousse" les données vers le Staging quand il en a l'occasion. L'inconvénient est que si le système est occupé, il ne poussera jamais les données.
- **Pull:** au contraire de la méthode précédente, le Pull " tire " les données de la source vers le Staging. L'inconvénient de cette méthode est qu'elle peut surcharger le système s'il est en cours d'utilisation.
- **Push-pull:** vous le devinez, c'est le mélange des deux méthodes. La source prépare les données à envoyer et prévient le Staging qu'elle est prête. Le Staging va récupérer les données. Si la source est occupée, le Staging fera une autre demande plus tard.

Staging (ou zone de préparation) est le terme désignant l'endroit où se fait l'ETL. C'est une machine dédiée à ce travail dans la plupart des cas. Considérez le *Staging* comme une zone tampon entre les sources de données et l'entrepôt.

Une fois la bonne stratégie choisie (en fonction des spécificités de l'entreprise), il est temps de se poser les questions fondamentales qui dessineront les caractéristiques de votre système:

- Quelle est la disponibilité de mes sources de données ?
- Comment y accéder ?
- Comment faire des chargements incrémentiels ?
- Quel est le temps d'un chargement incrémentiel moyen, ai-je la possibilité de recharger des données dans le cas où mon processus de chargement échoue ?
- Quelle politique vais-je utiliser dans le cas d'échec de chargement ?

Ces questionnements aideront à établir une stratégie de chargement des données sources dans le *Staging*.

3.3. Comment traiter les données ?

Maintenant que les données sont dans le Staging, il va falloir les nettoyer. C'est l'opération la plus importante du processus. En effet, une erreur dans un champ affecte forcément les analyses (exemple de Canada et Canonda). Voici les questions à se poser à cette étape :

- Quels sont les champs les plus sujets à erreurs ?
- Ai-je les moyens de corriger les erreurs automatiquement ?
- Comment permettre à un utilisateur de corriger les erreurs ?
- Quelle politique vais-je utiliser pour le traitement des erreurs.
- Comment montrer à l'utilisateur final que des données n'ont pas été totalement chargées à cause d'erreurs ?

3.4. Comment charger les données dans l'entrepôt ?

La dernière mission de l'ETL, charger les données dans le DW. Le point critique dans cette étape est qu'il faut avoir, à tout moment, un contrôle total sur les processus. Exemple : pendant un projet de construction d'entrepôt, vous commencez à automatiser les chargements incrémentiels. Mais un jour, la machine tombe en panne au beau milieu du chargement, c'est-à-dire qu'une partie des données a été chargée et une autre non. Que faire ??

Et bien si vous n'aviez pas prévu cela, vous n'avez qu'à vider la base et la recharger, avec toutes les pertes d'historique que cela implique, ou sinon prendre le temps et chercher une à une, les informations qui ont été chargées. Voici les questions qu'il faut se poser pour cette étape :

- Que faire si un chargement échoue ?
- Ai-je les moyens de revenir à l'état avant le chargement ?
- Puis-je revenir dans le temps d'un chargement donné ?
- Comment valider mon chargement, comment détecter les erreurs ?

3.5. Les métadonnées

C'est une des clés du succès de tout projet décisionnel. Les méta-données, en informatique décisionnelle, sont des informations décrivant notre environnement décisionnel. Il ne s'agit pas seulement des informations concernant le schéma des entrepôts ou la politique d'attribution de noms aux champs de l'entrepôt, mais de tout ce qui, de près ou de loin, peut ajouter de la compréhension aux chiffres présentés.

En effet, il est peut être pertinent pour notre l'analyste de savoir que la colonne prix qu'il est en train d'analyser provient des archives et non des données courantes. Il est peut être utile aussi de savoir que les chiffres devant nos yeux sont issus d'un chargement qui a échoué mais qu'on a réussi à recharger correctement. Il est important pour le grand patron d'une entreprise d'avoir une petite info bulle qui lui indique que les données de son tableau de bord sont ceux de l'avant-veille car le chargement ne s'est pas bien déroulé. Imaginez la catastrophe si le décideur prenait des décisions sur des données erronées !!

Il est très important, dans un environnement décisionnel, de non seulement documenter tout le projet, mais de rendre aussi disponible toutes ses méta-données aux utilisateurs de l'environnement pour générer encore plus de connaissance. Car n'oubliez pas que le but finalement est de créer de la connaissance.

3.6. Comment contrôler cet ETL ?

Dans tout ce que vous ferez dans la vie, le contrôle est la clé du succès. Le non contrôle amenant inévitablement le risque et le risque entraînant l'erreur. Si vous ne savez pas d'où provient une erreur, il est fort probable qu'elle soit du côté d'un élément dont vous n'avez pas le contrôle. Le meilleur système étant celui qui laisse le moins de place au risque. Intéressons nous à comment contrôler un ETL, quels sont les points clés à surveiller et, surtout, que faire lorsqu'un élément ne fait pas son travail correctement.

Les ETL sont, malheureusement, la plus grande faiblesse des environnements décisionnels. Tout est critique et sujet à contrôle dans un ETL. Sans oublier que le contrôle sans action n'est pas utile !

Comment retourner en arrière, c'est la principale question qu'on se pose et à laquelle on essaie de répondre quand on conçoit un ETL. Comment retourner en arrière si un chargement s'arrête brusquement, comment revenir à l'état initial si les données semblent incohérentes, comment valider mes transformations…

Le but de tous ces questionnements est de préserver l'intégrité et la " vérité " de l'entrepôt de données. Car c'est le seul point de défaillance de l'environnement. (à part la phase de chargement, tout le système est en lecture seule).

4. Outils ETL

4.1. Éléments à prendre en compte lors du choix de l'ETL

Les solutions d'ETL existent, sont nombreuses et répondent à toutes les demandes de performance et de portefeuille.

Cependant, devant un choix si diversifié, on se retrouve un peu perdu : Open Source ou payant, solution intégrée ou indépendante, sous-traitance ou développement. Les éléments à prendre en compte dans le choix de votre ETL sont les suivants :

- **Taille de l'entreprise**: j'entends par là taille des structures. S'il s'agit d'une multinationale avec des milliers de succursales à travers le monde, on ira plus pour une solution complète et, en général, très coûteuse. Si on est une PME, on optera plutôt pour des solutions payantes (comme Microsoft Integration Services) assurant un certain niveau de confort sans impliquer des mois de développement.
- **Taille de la structure informatique**: une entreprise avec une grosse structure informatique pourra se permettre d'opter pour une solution Open Source et la personnaliser selon les besoins de l'entreprise. Une PME ne pourra sûrement pas faire cela.
- **Culture d'entreprise**: évidemment, si une entreprise a une culture de l'Open Source très prononcée, l'application d'une solution payante risquera fortement de subir un phénomène de rejet.

- **Maturité des solutions**: il existe des solutions bien rodées, qui fonctionnent bien et qui bénéficient d'un bon retour d'expérience, c'est en général les plus chères (Business Objects, Oracle, SAP). Il existe d'autres solutions, moins matures, bénéficiant d'un " effet de mode " et qui semble offrir de très bonnes performances (Microsoft). Enfin, il existe des solutions Open Source qui, de part leur jeunesse, n'offrent pas autant de flexibilité et de facilité de mise en œuvre que les solutions précédemment citées. Il faudra compter avec le temps pour que ces solutions émergent et arrivent à un niveau de maturité acceptable…

5. Les Challenges de l'ETL

L'implémentation de processus d'ETL efficaces et fiables comprend de nombreux challenges.

- Les volumes de données sont en croissance exponentielle, et les processus d'ETL doivent traiter des quantités importantes de données granulaires (produits vendus, appels téléphoniques, transactions bancaires, etc.). Certains systèmes décisionnels sont mis à jour de façon incrémentale, alors que d'autres sont rechargés dans leur totalité à chaque itération.
- Alors que les systèmes d'information se complexifient, la variété des sources de données s'accroît également. Les processus d'ETL doivent disposer d'une large palette de connecteurs à des progiciels (ERP, CRM, etc.), bases de données, mainframes, fichiers, Services Web etc.
- Les structures et applications décisionnelles incluent des Data Warehouses, des Datamarts, des applications OLAP - pour l'analyse, le reporting, les tableaux de bord, le scorecarding, etc. Toutes ces structures cibles présentent des besoins différents en termes de transformation de données, ainsi que des latences différentes.

Les transformations des processus d'ETL peuvent être **très complexes**. Les données doivent être agrégées, parsées, calculées, traitées statistiquement, ...

6. Solutions d'Intégration Open Source pour l'ETL

Les solutions d'intégration de données *Talend* sont optimisées pour les besoins ETL de l'entreprise.

Exemple: Talend Open Studio (TOS)

Talend est la solution d'intégration de données Open Source permettant de répondre avec efficacité à un très large éventail de besoins: alimentation de Data Warehouse, synchronisation de bases de données, transformation de fichiers de divers formats (XML, VSAM, délimités, positionnels...), ...

Talend Open Studio sait déjà se connecter à un nombre plutôt impressionnant de bases de données, sait manipuler un grand nombre de formats de données. Cependant, il rester encore à étendre et ce par exemple à vos applications qui doivent gagner en ouverture vers tout autre logiciel.

Exemples de d'Outils ETL

- Oracle Warehouse Builder;
- IBM Infosphere Information Server;
- Microsoft SQL Server Integration Services (SSIS);
- SAS Data Integration Studio.

En résumé, Le secret d'un bon ETL réside dans sa complétude et dans son exhaustivité dans la prise en charge des données depuis les sources de données jusqu'à l'entrepôt.

7. La phase de restitution

C'est la dernière étape d'un projet Data Warehouse, soit son exploitation, soit restitution des résultats. Elle se fait par le biais d'un ensemble d'outils analytiques développés autour du Data Warehouse. Les outils de restitution sont la partie visible offerte aux utilisateurs. Par leur biais,

les analystes sont à même de manipuler les données contenues dans les entrepôts et les marchés de données. Les intérêts de ces outils sont l'édition de rapports et la facilité de manipulation. En effet, la structure entière du système décisionnel est pensée pour fournir les résultats aux requêtes des utilisateurs, dans un temps acceptable et sans connaissance particulière dans le domaine de l'informatique.

On distingue à ce niveau plusieurs types d'outils différents :

- Les outils de reporting et de requêtes
- Les outils d'analyse OLAP
- La phase de Datamining

Les outils de reporting et de requêtes permettent la mise à disposition de rapports périodiques, pré-formatés et paramétrables par les opérationnels. Ils offrent une couche d'abstraction orientée métier pour faciliter la création de rapports par les utilisateurs eux-mêmes en interrogeant le Data Warehouse grâce à des analyses croisées. Ils permettent également la production de tableaux de bord avec des indicateurs de haut niveau pour les managers, synthétisant différents critères de performance.

Le tableau de bord est un ensemble d'indicateurs peu nombreux conçus pour permettre aux gestionnaires de prendre connaissance de l'état et de l'évolution des systèmes qu'ils pilotent et d'identifier les tendances qui les influenceront sur un horizon cohérent avec la nature de leurs fonctions.

Les tableaux de bord sont prédéfinis et consultables à l'écran, ils génèrent des états tels que les histogrammes.

Les outils d'analyse OLAP permettent de traiter des données et de les afficher sous forme de cubes multidimensionnels et de naviguer dans les différentes dimensions. Cet agencement des données permet d'obtenir immédiatement plusieurs représentations d'un même résultat, en une seule requête sous une approche descendante des niveaux agrégés vers les niveaux détaillés (Drill-down, Roll-up).

Les outils de Datamining offrent une analyse plus poussée des données historisées permettant de découvrir des connaissances cachées dans les données comme la détection de corrélations et de tendances, l'établissement de typologies et de segmentations ou encore des prévisions. Le Datamining est basé sur des algorithmes statistiques et mathématiques, et sur des hypothèses métier.

8. Maintenance et expansion

La mise en service du Data Warehouse ne signifie pas la fin du projet, car un projet Data Warehouse nécessite un suivi constant compte tenu des besoins d'optimisation de performance et ou d'expansion. Il est donc nécessaire d'investir dans les domaines suivants:

- Support: Assurer un support aux utilisateurs pour leur faire apprécier l'utilisation de l'entrepôt de données. En outre, la relation directe avec les utilisateurs permet de détecter les correctifs nécessaires à apporter.
- Formation: il est indispensable d'offrir un programme de formation permanant aux utilisateurs de l'entrepôt de données.
- Support technique: un entrepôt de données est considéré comme un environnement de production. Naturellement le support technique doit surveiller avec la plus grande vigilance les performances et les tendances en ce qui concerne la charge du système.
- Management de l'évolution: il faut toujours s'assurer que l'implémentation répond aux besoins de l'entreprise. Les revues systématiques à certain point de contrôle sont un outil clé pour détecter et définir les possibilités d'amélioration. En plus du suivi et de la maintenance du Data Warehouse, des demandes d'expansion sont envisageables pour de nouveaux besoins, de nouvelles données ou pour des améliorations.

Ces travaux d'expansion sont à prévoir de façon à faciliter l'évolution du schéma du Data Warehouse.

9. Exercice

Une entreprise de fabrication de vaisselle jetable souhaite mettre en place un système d'information décisionnel sous la forme d'un Datamart pour observer son activité de ventes au niveau des différents lieux de distributions de ses articles et cela dans plusieurs villes. Ces lieux de distributions sont identifiés par leur enseigne, leur type (en fonction de leur surface), leur adresse (code postal et ville), leur département et leur région. Les informations relatives aux ventes sont: une période en mois, en trimestre et année. Les ventes sont observées par le nombre d'articles selon le type, et le chiffre d'affaire.

1. Quel est le fait à observer?

2. Quels sont les axes d'analyse et les mesures (indicateurs)?

3. Construire le modèle en étoile de ce Datamart.

4. Modifier ce modèle en un modèle en flocon. Modéliser explicitement les hiérarchies des dimensions.

5. Donner une représentation en sommant les éléments selon les hiérarchies représentées en 4.