# *GIRAF*

## *Dossier DATAWAREHOUSE*

*Ingénierie des Systèmes d'Information 2 (ISI 2)*

*Groupe 2*

*Université Paris-Dauphine*


Michael GOLETTO
Négué DIALLO
Benjamin FAIBIS
Salim GUENNOUNI (salim.guennouni@gmail.com)

26/04/11

# GIRAF

Ce document est destiné à présenter un Datawarehouse avec ses finalités, les techniques utilisées pour le concevoir, sa place dans le système d'information d'une entreprise, ainsi que les outils et différents outils associés au DTW.

## Sommaire

Dans la première partie nous analyserons pourquoi et comment créer un DTW dans une deuxième partie nous verrons comment utiliser un DTW au sein d'une l'entreprise puis dans un dernier temps nous illustrerons un exemple de requête. Dans chacune de ces parties, l'application d'un DTW dans GIRAF est intégrée.

# Présentation du Datawarehouse

## Définition

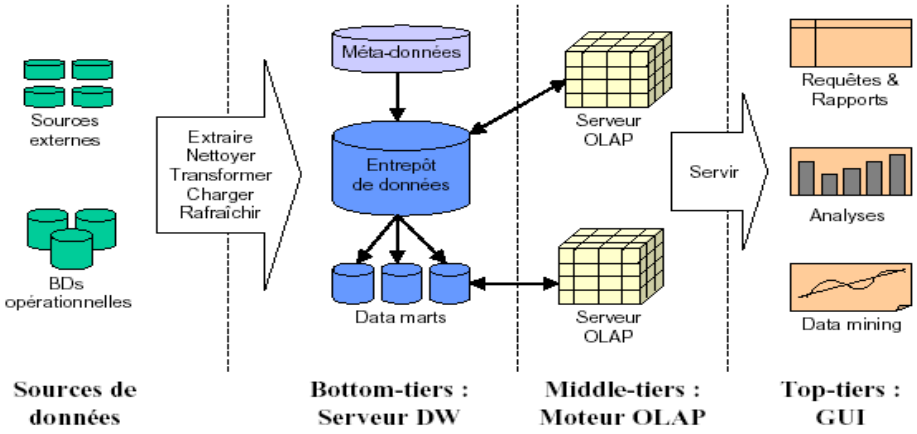
Le Datawarehouse (ou entrepôt de données) est une collection regroupant toutes les informations de l'entreprise. Il s'agit d'une structure utilisée pour rassembler toutes les données concernant l'entreprise et son activité. C'est un ensemble de données qui est consolidé dans une base de données unique. On utilise cette base de données pour collecter, ordonner, historiser et stocker des informations.

Le DTW (ou entrepôt de données) entre dans le cadre de l'informatique décisionnelle. Il en est même l'élément central. En effet, il fournit une aide à la décision stratégique de l'entreprise et permet d'en améliorer ses performances.

Les trois fonctions importantes du DTW sont donc :

- La collecte de données de base existantes et chargement
- Gestion de données dans l'entrepôt
- Analyse de données pour la prise de décision

# Architecture type d'un Datawarehouse




1. Schéma de l'architecture type d'un Datawarehouse

# Pourquoi et comment créer un Datawarehouse

## Pourquoi créer un Datawarehouse

Les entreprises ont un besoin d'avoir accès à toutes données concernant leurs actions. Il faut ensuite regrouper les informations disséminées. En fonction de tout cela, les dirigeants doivent analyser et prendre des décisions rapidement.

Le datawarehouse est une base de données utilisée pour l'analyse de données. Le datawarehouse est ainsi le lieu unique de consolidation de l'ensemble des données de l'entreprise. Toutes les données historisées et informations que possède l'entreprise sont stockées dans le DTW. Le DTW est orienté sujet, les données sont organisées en « métier ». De cette façon les données utilisées par un métier sont reliées entre elles dans un sous ensemble du DTW que l'on appelle des Datamarts.

Le DTW peut être utilisée pour obtenir un historique des données ainsi que pour analyser ces données dans des fins stratégiques. En effet un datawarehouse est essentiellement utilisé comme outils analytique et pour aider aux décisions stratégiques de l'entreprise. Ces décisions ont pour but d'améliorer les performances de l'entreprise.

La société Jolyfringues envisage un datawarehouse pour des activités de pilotage, donc dans des fins stratégiques. En effet, il existe un grand déficit dans l'exploitation des statistiques de ventes des boutiques. Le service Marketing veut s'appuyer sur cela et faire des études sur les ventes de leurs différentes boutiques ou franchises afin de mieux préparer les collections suivantes.

Exemple d'applications concernées :

- Grande distribution : Marketing, maintenance
  - Produits à succès, modes, habitudes d'achat
  - Préférences par secteur géographique
- Bancaire : suivi des clients, gestion de portefeuilles
  - Envoi de mails ciblés pour le marketing
- Télécommunications : pannes, fraudes, mobiles, …
  - Classification des clients, détection des fraudes, fuites de clients

## Comment créer un Datawarehouse

Définir un DTW est une tache compliquée. Chaque entreprise a un mode de fonctionnement propre et des nécessités uniques. Il faut donc savoir manipuler ces données avant de pouvoir définir le schéma du DTW.

Obtenir un schéma logique entre les différentes relations de données de l'entreprise peut se révéler compliqué. Rappelons-nous que le DTW est une base de données.

Les différentes tables, relations et clés qui permettent d'obtenir un schéma structuré peuvent se révéler complexe, surtout quand le nombre de données à traiter sont nombreuses. Sans oublier que la création d'un tel schéma se doit d'être évolutive. Le maintient d'un schéma de base de données étant un « on going process ».

## Faits et dimensions

Cette base de données est organisée avec des relations et des tables de faits. On entend par dimensions les axes avec lesquels on veut faire l'analyse. Par exemple pour GIRAF on a les dimensions boutiques, contacts, produits,… Les faits, en complément aux dimensions, sont ce sur quoi va porter l'analyse. Ce sont des tables qui contiennent des informations opérationnelles et qui relatent la vie de l'entreprise. Par exemple les faits GIRAF seraient les ventes d'une boutique, le rattachement de telle boutique à une certaine collection, … En réunissant ces faits en métier on crée des datamarts.

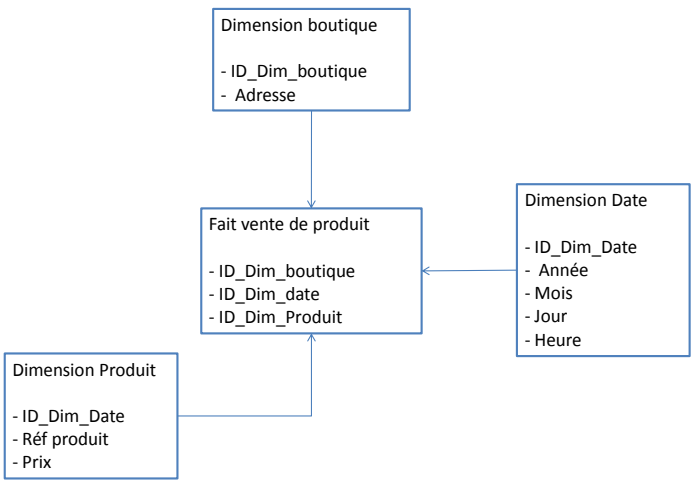
## Datamarts

Les datamarts sont un sous ensemble du DTW global. Les données d'un datamart sont les données concernées par un certain secteur de l'entreprise (direction, finance,…) La qualité de ce type de stockage de données est la suivante. Le nombre de données étant réduites il est donc bien plus rapide de pouvoir accéder, analyser ces données que dans le DTW au complet. De plus les données comprises dans un datamart sont moins volumineuses que dans le DTW.

Un datamart c'est aussi la réunion de modélisation de faits en étoile. Par exemple dans GIRAF si on souhaite faire un datamart pour le métier marketing, il nous faudrait réunir chaque fait inclus dans le cadre du service marketing. Un des faits inclus au datamart marketing serait par exemple les ventes de produit des boutiques.

## Modélisation en étoile

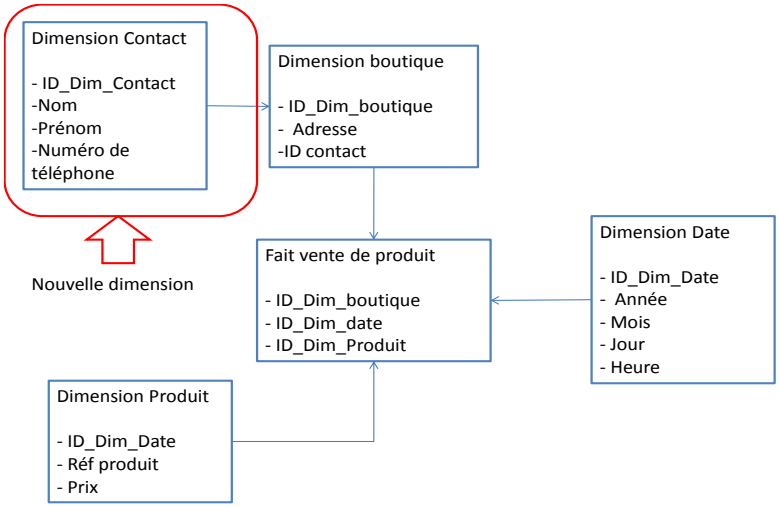
On illustre ci-dessous la modélisation en étoile d'un fait. La table de fait étant la vente d'un produit, les dimensions sont les axes selon lesquels on désire faire l'analyse. Ici les dimensions sont les produits (pour analyser les produits les plus vendus), les boutiques, et les dates. Un diagramme en étoile pour la table de fait vente produit ressemble à :




2. Modélisation d'une table en étoile

On peut si l'on souhaite être plus précis faire une modélisation en flocon d'une table de fait. Le principe de la modélisation en flocon est de créer des hiérarchies de dimensions. On dit alors que l'on a une hiérarchie en n-dimensions (ci-dessous on est en 2 dimensions).

Exemple pour la table de fait vente de produit :




3. Modélisation d'une table en flocon

C'est en cumulant les tables de faits (ainsi que ses dimensions reliées) qui concernent un métier que l'on peut créer un datamart. Exemple en réunissant les tables de fait vente de produit, rattachement boutiques, détachement boutiques, … ainsi que leurs dimensions reliées on peut créer le datamart marketing.

Enfin, en réunissant les datamarts marketing, approvisionnement, achat et informatique on peut créer un DTW complet.

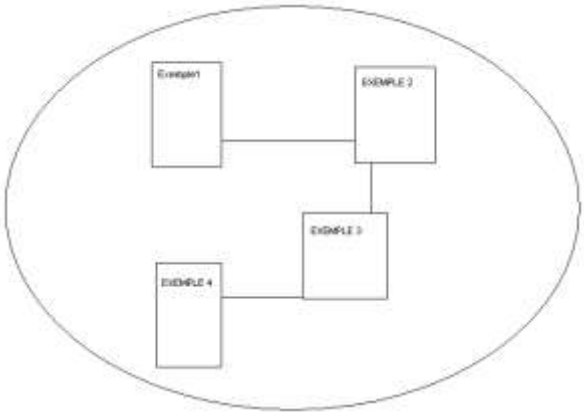
Sami Rébiensefut dit « on envisage un datawarehouse qui pompera dans les différents systèmes», par cette phrase il envisage qu'on ne crée pas un DTW complet mais seulement une partie des datamarts et par conséquent une partie du DTW. La solution parait bonne, mais dans la quasi-totalité de nos sources, les experts déconseillent la création d'un datawarehouse partiel et encourage de relier tous les datamarts (groupe d'étoiles ou flocons) en DTW global (créer une constellation avec nos étoiles et flocons)

Ainsi dans la création d'un DTW, deux approches principales sont utilisées : l'approche Top-down et l'approche Bottom-up.

## L'approche Top-Down

La méthode top-down est la plus complète. Cependant, c'est la méthode la plus longue, la plus coûteuse et la plus contraignante. Elle consiste en la conception de tout l'entrepôt (toutes les étoiles), puis en la réalisation de ce dernier.

Le seul avantage que cette méthode comporte est qu'elle offre une vision très claire et très conceptuelle des données de l'entreprise ainsi que du travail à faire.

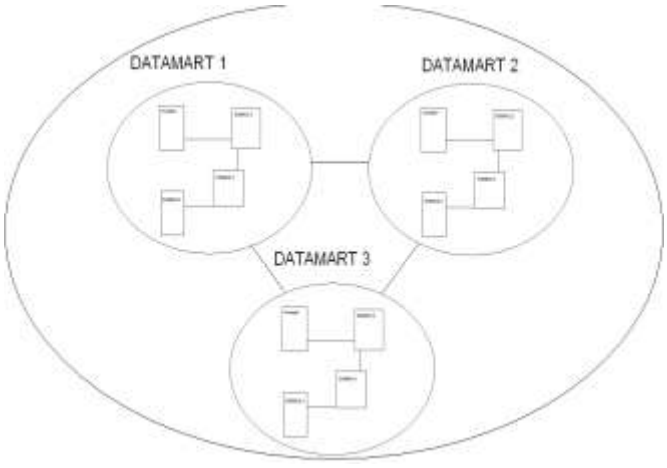


4. Schéma aproche Top-Down

On établie le schéma de l'entreprise avant de créer notre DTW. Puis on crée toutes nos tables de faits qui correspondent à tous les métiers de l'entreprise sans diviser les activités.

## Approche Bottum-up

L'autre approche en bottom-up est différente. C'est l'approche inverse, elle consiste à créer les étoiles une par une, puis les regrouper par datamarts jusqu'à obtention d'un véritable entrepôt pyramidal avec une vision d'entreprise. L'avantage de cette méthode est qu'elle est simple à réaliser (une étoile à la fois). L'inconvénient est le volume de travail d'intégration pour obtenir un entrepôt de données ainsi que la possibilité de redondances entre les étoiles (car elles sont faites indépendamment les unes des autres). Ce processus est plus rapide et moins cher.




5. Schéma approche Bottum-up

On établie le schéma de nos datamarts avant de créer notre DTW. On crée nos datamarts en réunissant les tables de faits de nos différents métiers dans des sous entrepôt de données (on réuni les étoiles des différents faits organisées par activité métier). Puis on relie ses différentes constellations de table de faits en une grande constellation qui représente le DTW.

Pour le projet GIRAF, certains services sont clairs sur les tables de faits qu'ils veulent étudier. Par exemple le service marketing veut étudier les ventes des produits. On commencerait d'abord par créer une table de fait vente de produit puis d'autre table de fait du service marketing. Les autres services sont moins précis sur leur utilisation d'un datawarehouse. Il nous semblerait alors qu'une création en bottom-up, service par service, semble bien adapté. Cependant il ne faut pas oublier que Sami Rébiensefut veut un datawarehouse qui couvre l'ensemble de l'activité de l'entreprise. Il ne faudra pas donc juste créer le datamart lié au métier marketing mais bien tous les datamarts.

## Contraintes d'un Datawarehouse

### Contraintes

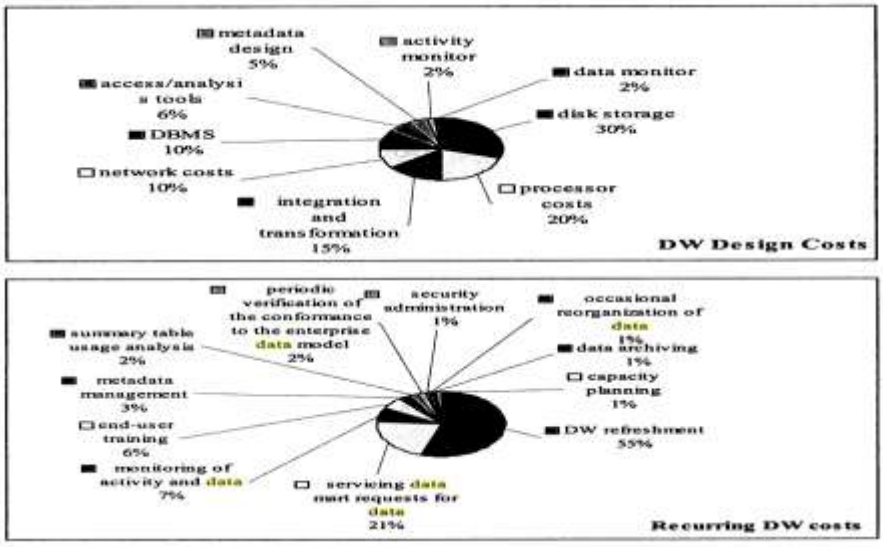
Dans l'élaboration d'un DTW une des grandes difficultés est du à la multitude des provenances des informations. Bases de données traditionnelles, document Excel, document Word, document du web, … Un des principaux défis lors de l'élaboration d'un datawarehouse est donc d'importer ces informations vers le DTW. Par exemple lorsque Jolyfringues reçoit des informations en .xls il faut les importer dans la base de données.

De plus, un seul datawarehouse ne suffit pas à rassembler des données et engendre des contraintes techniques lourdes :

- la capacité de stockage qui peut dépasser les 100 Tera octets pour les plus grosses bases.
- la capacité à traiter des requêtes de plus en plus complexes et de plus en plus nombreuses.
- l'alimentation en continue des données (pour pouvoir analyser des données les plus fraîches possibles).

### Coût de création d'un Datawarehouse

Créer une datawarehouse reste un projet couteux. En moyenne il faut entre 1 et 2 ans pour construire un DTW et débourser entre 1 M€ et 1.5 M€ pour la réaliser. Evidemment dans le cas de GIRAF, l'élaboration du DTW serait bien moins couteuse et moins longue vu l'étendue des données à traiter.




6. Illustration des coûts d'un Datawarehouse

Dans le premier tableau on peut voir la répartition des coûts lors de la création d'un datawarehouse. Plus de 60 % des coûts sont compris dans le hardware (mémoire et réseaux essentiellement) et le reste par des services. Le deuxième tableau, quant à lui présente, les coûts récurent, donc après la création du DTW. Le principal coût provient de la mise à jour du DTW. Il est difficile de savoir quels seront les coûts de création d'un datawarehouse pour Jolyfringues. Les données ne sont pas si nombreuses. Cependant il faudra au moins un ingénieur pour la mise à jour du datawarehouse et pour effectuer les analyses demandées par le marketing. Du coté du hardware au moins 250 Go de mémoire sont nécessaires.

Tant de données apportent évidemment des problèmes dans les requêtes. Toute requête SQL basique prend du temps à effectuer vu la taille des données. L'usage d'index pour chacune des tables est donc fortement recommandé. Cependant, aujourd'hui grâce à l'efficacité du hardware ce temps est vraiment revu à la baisse.

## Utilisation et mise à jour du Datawarehouse

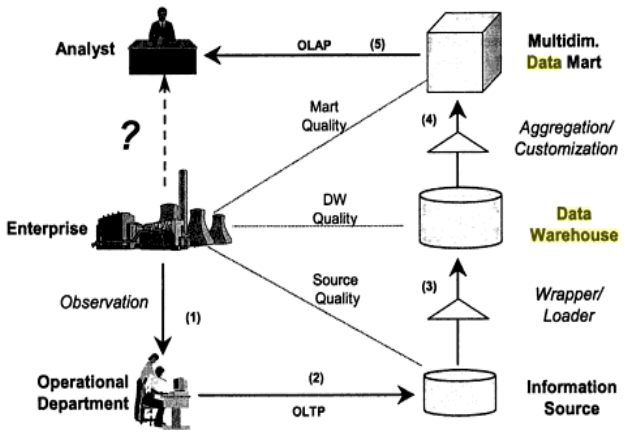


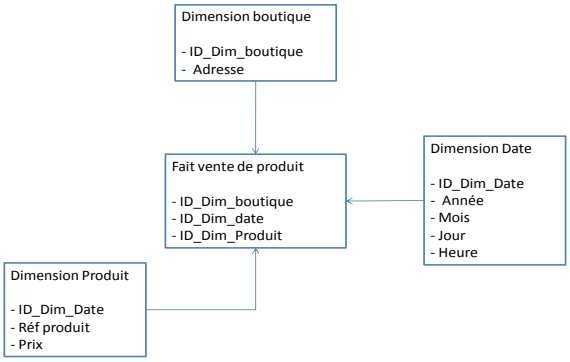

**Fig. 2.3.** Data warehousing in the context of an enterprise

7. Utilisation d'un Datawarehouse

La figure ci-dessus permet de bien comprendre le principe d'utilisation d'un datawarehouse. Les informations de l'entreprise sont observées (1) puis saisies par le département opérationnel (traitement transactionnel en ligne). Ces informations sont alors transformées pour être adaptées au format du DTW. Par exemple si ces données sont fournies en document Excel, il faut importer toutes ces données dans le DTW (en se servant d'un programme d'importation par exemple). Le DTW est alors mis à jour, mais les informations sont bruts et n'apportent pas toujours beaucoup d'intérêt pour la compagnie, excepté celui de stocké des données en grand volume. Pour apporter une plus value à ses données, on les regroupe en datamart sorte de mini DTW (environ 50Go de données) focalisé sur un sujet spécialisé souvent lié à un métier de l'entreprise (finance, marketing, commercial…). Ces données ainsi organisée, subissent l'intervention d'une application informatique de **traitement analytique en ligne(OLAP)** (5) avant d'être utilisées par un analyste. Grace à ce traitement analytique en ligne l'utilisateur à désormais accès à des manipulations des données utiles pour un analyste.

## Exemple d'utilisation d'un Datawarehouse

Mettons nous à la place du service marketing. Prenons l'hypothèse que toutes les ventes sont, depuis l'année 2008, historisées dans le datawarehouse. On a modélisé la table de fait vente de produit en étoile, comme ci-dessous :




8. Exemple : modélisation de la table de faits en étoile

Le service marketing souhaite connaitre le prix moyen des produits vendus par boutique en 2009. Grâce à notre datawarehouse cette information est facilement accessible avec la requête SQL suivante :

```
SELECT ID_Dim_boutique, AVG(prix)
FROM fait_vente_de_produit fvdp, dimension_date dd, Dimension_produit dp
WHERE fvdp.id_dim_date=dp.id_dim_date
        AND fvdp.id_dim_date=dd.id_dim_date
        AND dd.annee=2009
GROUP BY ID_Dim_boutique
```

On voit tout de suite l'intérêt de ce genre de requête. Une stratégie à mettre en place devant l'étude des résultats de cette requête pourrait être :

- Si le prix moyen des produits achetés dans la boutique A est bas, alors on privilégie l'exposition de produit bon marché en vitrine.

- Si le prix moyen des produits achetés dans la boutique B est haut, alors on privilégie l'exposition de produit cher en vitrine.

Evidemment il pourrait exister d'autres stratégies, mais ce travail est réservé aux spécialistes marketing. Cependant sans datawarehouse il serait impossible d'effectuer ce genre d'analyse !

On peut également faire des analyses bien plus poussées comme par exemple avec du datamining ou grâce à des requêtes OLAP plus adaptées dans le cadre d'une base de données rangée en dimension.

## Conclusion

La conception de datawarehouse ne date pas d'aujourd'hui. Tous les grands organismes ont déjà des DTW.C'est donc un processus en perpétuelle évolution. Au cours des prochaines années, la croissance des datawarehouse sera énorme avec l'évolution des nouvelles technologies. Afin de profiter de cette évolution, il sera important que dirigeants et développeurs aient une idée claire de ce qu'ils recherchent et d'alors établir une stratégie garantissant les meilleures performances.