Global Scientific JOURNALS

GSJ: Volume 10, Issue 9, September 2022, Online: ISSN 2320-9186

www.globalscientificjournal.com

Problématique de la conception d'un Entrepôt de données (Data Warehouse)

Par Jeannot FATAKI BAZONGA

Professeur

Résumé

D'entrée de jeux, cet article est destiné à ceux qui ont une certaine maitrise en ce qui concerne un environnement décisionnel avec comme concept central l'entrepôt de données ou le Data Warehouse. Nous nous sommes vraiment intéressé à présenter comment concevoir un entrepôt de données ou le Data Warehouse.

Sur ce, nous nous sommes posés les quelques questions suivantes : Quelle structure permet-elle d'avoir les fonctionnalités requises pour un entrepôt de données ? Quelles sont les techniques utilisées pour bien concevoir ? Quels sont les indicateurs d'une bonne conception ?

Dans cet article nous allons commencer par introduire les concepts fondamentaux de l'informatique décisionnelle (nécessaires pour la compréhension de cet article), nous continuerons par l'explication des méthodes de conception d'entrepôt de données via une étude de cas, et nous terminerons par une critique de ces techniques et une conclusion mentionnant les indicateurs d'une bonne conception d'entrepôt.

Abstract

From the outset, this article is intended for those who have a certain mastery regarding a decision-making environment with the central concept of the « entrepôt de données » or the Data Warehouse. We were really interested in how to design a « entrepôt de données » or the Data Warehouse.

With this in question, we ask the following questions: What structure does it allow to have the functionality required for a data warehouse? What techniques are used to design well? What are the indicators of good design?

In this article we will begin by introducing the fundamental concepts of decision-making computing (necessary for understanding this article), we will continue by explaining data

warehouse design methods via a case study, and we will end with a critique of these techniques and a conclusion mentioning indicators of good warehouse design.

Les mots-clés : Entrepôt de données (Data Warehouse), Magasin de donnés (Data Mart), Dimension, Fait, ETL ou ETC pour les francophiles, Etoile, Flocon, Constellation, Top-Down, Botton-Up, Middle-Out, OLAP.

Mots clés

1. Entrepôt de données (Data Warehouse)
2. Data Mart, ou magasin de données
3. Dimension
4. Fait
5. ETL, ou ETC
6. Étoile
7. Flocon
8. Modélisation en étoile
9. Modélisation en flocon
10. Conception d'entrepôts de données
11. Constellation

Table des matières

1. Introduction
2. Concepts fondamentaux
    2.1. Entrepôt de données (Data Warehouse)
    2.2. Data Mart, ou magasin de données
    2.3. Dimension
    2.4. Fait
    2.5. ETL, ou ETC pour les francophiles
    2.6. Étoile
    2.7. Flocon
3. Modélisation en étoile : étude d'un cas
    3.1. Le cas
    3.2. L'analyse
    3.3. La solution
4. Modélisation en flocon : étude d'un cas
5. Conception d'entrepôts de données
    5.1. Constellation
    5.2. Construire un entrepôt de données,
    5.3. Comparatif entre les bases de données de l'entreprise
6. Critique des méthodes de conception d'entrepôts
7. Conclusion
8. Bibliographie sommaire

© GSJ

1. Introduction

Comme nous l'avons dit précédemment, que cet article était destiné à ceux qui ont une certaine maitrise en ce qui concerne un environnement décisionnel avec comme concept central l'entrepôt de données ou le Data Warehouse. Intéressons-nous maintenant à comment concevoir un entrepôt de données en nous posant les quelques questions ci-après :

- Quelle structure permet-elle d'avoir les fonctionnalités requises pour un entrepôt de données ?
- Quelles en sont les techniques utilisées pour bien concevoir ?
- Quels sont les indicateurs d'une bonne conception ?

Comme hypothèses, nous proposons ce qui suit :

- Un entrepôt de données utilise une base de données ou un groupe de bases de données comme base.
- Outils OLAP, qui aident à construire un entrepôt de données multidimensionnel et permettent l'analyse des données d'entreprise à partir de nombreux points de vue.
- Des *techniques* d'analyse temporelles *sont* requises *pour* mieux tirer parti des avantages de cet outil. La *technique* de modélisation *utilisée* est celle du modèle relationnel décrit par des méthodes de développement agiles et itératives qui garantissent la fourniture d'informations de qualité aux utilisateurs finaux sur la base de données.
- Un des indicateurs clés d'une bonne conception d'entrepôt est la grosseur des constellations. En effet, plus la constellation est grosse, plus cela veut dire que vous avez réutilisé vos dimensions, et qui dit réutilisation de dimension, dit dimensions complètes, centralisées et avec une vue orientée entreprise.

Sur ce, nous avons utilisé deux des approches d'entrepôt de données les plus largement discutées et expliquées qui sont la méthode Inmon et la méthode Kimball. Pendant des années, les gens ont débattu de l'approche d'entrepôt de données la meilleure et la plus efficace pour les entreprises. Cependant, il n'y a toujours pas de réponse définitive car les deux méthodes ont leurs avantages et leurs inconvénients.

Fadila Bentayeb, Maître de conférences dans son mémoire intitulé « Entrepôts et analyse en ligne de données complexes centrés utilisateur : un nouveau défi » soutenu publiquement le 24 novembre 2011 devant un jury à *l'Université de Lyon - Lumière Lyon 2* a, dans ce mémoire a visé à proposer des solutions innovantes dans le domaine de la personnalisation dans les entrepôts de données complexes. L'originalité de nos travaux de recherche a consisté, a-t-il écrit, à montrer qu'il est pertinent d'intégrer la sémantique dans tout le processus d'entreposage, soit en invitant l'utilisateur à exprimer ses propres connaissances métier, soit en utilisant les méthodes de fouille de données pour extraire des connaissances cachées.

Marie-Chantal Denis, a présenté dans son mémoire intitulé « Conception et réalisation d'un entrepôt institutionnel de données dans une perspective de support à la prise de décision » comme exigence partielle de la maîtrise en mathématiques et informatique appliquées de

l'Université du Québec à Trois-Rivières. Son projet vise à concevoir et à appliquer une approche méthodologique afin de construire un prototype d'entrepôt de données de l'UQTR (Université du Québec à Trois-Rivières) et ainsi faciliter la prise de décision par l'élaboration de tableaux de bord présentés à l'intention des gestionnaires. Le projet, constitué d'une partie théorique et d'une seconde partie hautement pratique, est appliqué à un cas de gestion universitaire d'analyse des clientèles étudiantes. Préalablement à la phase d'analyse et d'historisation, nous avons procédé, dit-il, à la réalisation de l'étude des méthodologies de conception. La méthodologie développée dans ce projet est fondée sur celle utilisée par Ralph Kimball qui préconise une conception de bout en bout sur un seul processus d'affaires «orienté-sujet». La méthodologie ainsi conçue, a-t-il renchérit, permettra de poursuivre le projet et d'amalgamer les systèmes actuels de l'UQTR successivement afin d'offrir une homogénéité qui permettra aux demandeurs d'obtenir la compatibilité nécessaire à une consultation efficace et simple de l'ensemble des différents types de données. Dans un deuxième temps, l'analyse d'outils existants a été complétée. Cette étude a mis en relief les caractéristiques des différents outils menant à une première ébauche pour une recommandation d'achat d'une solution informatique pour l'UQTR. Deux types d'outils étaient nécessaires : un outil de chargement de l'entrepôt (ETL : extract-transform-load) et un outil de présentation en ligne des données. Finalement, une proposition d'historisation des définitions de tables a été développée permettant ainsi d'enrichir les métadonnées notamment en ce qui concerne les valeurs manquantes dues à la modification des structures de données dans le temps. Cette étape a offert alors l'opportunité d'en expliquer les causes à l'utilisateur.

Enfin comme dit aussi ci-haut, dans cet article nous allons commencer par introduire les concepts fondamentaux de l'informatique décisionnelle (nécessaires pour la compréhension de cet article), nous continuerons par l'explication des méthodes de conception d'entrepôt de données via une étude de cas, et nous terminerons par une critique de ces techniques et une conclusion mentionnant les indicateurs d'une bonne conception d'entrepôt.

2. Concepts fondamentaux

2.1. Entrepôt de données (Data Warehouse)

Un entrepôt de données, ou data Warehouse, est une vision centralisée et universelle de toutes les informations de l'entreprise. C'est une structure (comme une base de données) qui a pour but, contrairement aux bases de données, de regrouper les données de l'entreprise pour des fins analytiques et pour aider à la décision stratégique. La décision stratégique étant une action entreprise par les décideurs de l'entreprise et qui vise à améliorer, quantitativement ou qualitativement, la performance de l'entreprise. En gros, c'est un gigantesque tas d'informations épurées, organisées, historiées et provenant de plusieurs sources de données, servant aux analyses et à l'aide à la décision. L'entrepôt de données est l'élément central de l'informatique décisionnelle. En effet, l'entrepôt de données est le meilleur moyen que les professionnels ont trouvé pour modéliser de l'information pour des fins d'analyse, et il ne serait pas étonnant que d'ici quelques années un nouveau concept apparaisse pour révolutionner l'informatique décisionnelle.

2.2. Data Mart, ou magasin de données

Les Data Warehouses étant, en général, très volumineux et très complexes à concevoir, nous avons décidé de les diviser en bouchées plus faciles à créer et à entretenir. Ce sont les Data Marts[1]. On peut faire des divisions par fonction (un data mart pour les ventes, pour les commandes, pour les ressources humaines) ou par sous-ensemble organisationnel (un data mart par succursale). Nous verrons plus tard comment organiser les data marts pour créer un entrepôt proprement dit.

2.3. Dimension

Lorsqu'on réalise un schéma de Base de Données pour un système d'information classique, on parle en termes de tables et de relations, une table étant une représentation d'une entité et une relation une technique pour lier ces entités[2]. En BI (Business Intelligence), on parle en termes de Dimension et de Faits. C'est une autre approche des données, on entend par dimensions les axes avec lesquels on veut faire l'analyse. Il peut y avoir une dimension client, une dimension produit, une dimension géographie (pour faire des analyses par secteur géographique), etc.

Une dimension est tout ce que nous utiliserons pour réaliser nos analyses.

2.4. Fait

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. Nous aurons des tables de faits pour les ventes (chiffre d'affaire net, quantités et montants commandés, quantités facturées, quantités retournées, volumes des ventes, etc.) par exemple ou sur les stocks (nombre d'exemplaires d'un produit en stock, niveau de remplissage du stock, taux de roulement d'une zone, etc.), ou peut être sur les ressources humaines (performances des employés, nombre de demandes de congés, nombre de démissions, taux de roulement des employés, etc.).

Un fait est tout ce qu'on voudra analyser

2.5. ETL, ou ETC pour les francophiles

L'ETL (ou *Extract, Transform and Load* ou encore *datapumping*) pour extraction des données de production, transformations éventuelles et chargement de l'entrepôt, dont nous avons expliqué les fondements plus haut dans cet article, sert à transposer le modèle entité-relation des bases de données de production ainsi que les autres modèles utilisés dans les opérations de l'entreprise, en modèle à base de dimensions et de faits, nous verrons ces modèles dans les deux prochaines définitions.

---

[1] Alain Venot, Anita Burgun et Catherine Quantin, *Informatique Médicale, e-Santé – Fondements et applications*, Springer Science & Business, 18 janvier 2013.

[2] Isabelle Comyn-Wattiau, Jacky Akoka, *Les bases de données*, PUF, Que sais-je?, 978-2130533139, chap. ix *Les bases de données décisionnelles*, 2003.

2.6. Étoile

Une étoile est une façon de mettre en relation les dimensions et les faits dans un entrepôt de données. Nous le verrons plus tard, mais le principe est que les dimensions sont directement reliées à un fait (schématiquement, ça fait comme une étoile).

2.7. Flocon

Un autre modèle de mise en relation des dimensions et des faits dans un entrepôt de données. Le principe étant qu'il peut exister des hiérarchies de dimensions et qu'elles sont reliées aux faits.

> Les flocons et les étoiles peuvent être vus comme une manière de diviser les entrepôts de données et les magasins de données. On peut les voir comme l'atome de l'informatique décisionnelle : le plus petit élément avec lequel nous pouvons réaliser des analyses et avec lequel nous pouvons créer des magasins de données qui, mis ensemble, forment un entrepôt de données.

3. Modélisation en étoile : étude d'un cas[3]

Nous allons utiliser un exemple pour expliquer la modélisation en étoile. L'important en Business Intelligence (BI) est de toujours garder à l'esprit que ce que nous faisons est différent des bases de données traditionnelles. Le schéma créé sera accessible par les utilisateurs et doit donc être le plus simple et explicite possible !

3.1. Le cas

On vous demande de créer un data Mart (une étoile) pour l'analyse de l'activité des représentants d'une entreprise de vente d'imprimantes. Le chef d'entreprise veut savoir ce qui se passe pour ses vendeurs. Les employés font-ils leur travail, quelle est la zone de couverture des vendeurs, où sont les endroits où les vendeurs sont le moins efficaces, quelle est la moyenne de ventes des représentants, etc. L'entreprise possède un système de gestion de ressources humaines, un système de gestion des ventes et des feuilles de routes avec des informations concernant les vendeurs : kilomètres parcourus, litres d'essence utilisés, frais de voyage, ventes, promesses de ventes, etc.

3.2. L'analyse

Notre objectif est d'analyser l'activité des représentants. Il semble que nous ayons toutes les informations pour ce faire... Mais dans différents systèmes.

Commençons l'analyse :

[3] R. Kimball, L. Reeves, M. Ross, W. Tomthwaite, le data Warehouse : Guide conduite de projet, Eyrolles, .

Le but du jeu est de déceler les axes d'analyses (les dimensions) avec leurs attributs ainsi que les éléments à analyser (les faits). La meilleure façon de faire, selon nous, est l'étude approfondie de ce qui se passe dans l'entreprise : documents échangés, rapports périodiques, interviews des personnes clés, étude des besoins. Il faut vraiment faire un travail d'acteur, et rentrer dans la peau de chaque utilisateur, savoir comment les analystes organisent leurs raisonnements, savoir ce que voient les décideurs avant de décider, connaître les indicateurs de bonne santé de l'entreprise et de la concurrence.

Une manière très pratique de modéliser un cas en BI se fait comme suit :

	Date	**Vendeur**	**Produit**	**Zone géographique**	**Client**
	Années	Nom	Catégorie	Pays	Nom
	Mois	Prénom	Type	Province	Adresse
	Jours	Salaire	Groupe	Ville	Pays
	Heures				
Analyse : consommation d'essence, Qté commandée, Qté précommandée, kilométrage, nombre de visites, etc.					

Tableau n°1 : Modélisation d'un cas de Business Intelligent rempli pendant la phase d'analyse, en posant des questions aux décideurs du type :

- Que voulez-vous analyser (la dernière ligne du tableau) ?
- Quels sont vos critères d'analyse (la première ligne du tableau) ?
- Jusqu'à quel niveau de détail voulez-vous aller (les cellules à l'intérieur) ?

Remarque : L'axe du temps (dimension Temps) est toujours présent dans un entrepôt de données, c'est le type d'analyse le plus commun et le plus fréquent en entreprise.

La structure d'un entrepôt étant plus rigide que les systèmes conventionnels (se basent sur des ETL, des validations créées par l'homme, etc.), il est capital d'avoir une analyse des besoins exhaustive et conforme aux attentes des décideurs.

Il faut savoir :

- D'où provient chaque champ ?
- Comment transite l'information ?
- Où trouver l'information voulue?

Se poser des questions du type :

- Ai-je assez de données pour répondre aux besoins ?
- Si non, qu'est-ce que cela impliquerait de les créer ?
- Comment alimenter mes dimensions ?
- Comment alimenter mes faits ?

- Comment valider mes chargements ?
- Etc.

3.3. La solution

La modélisation en étoile découle naturellement du tableau ci-dessus, il en résulte le schéma suivant :

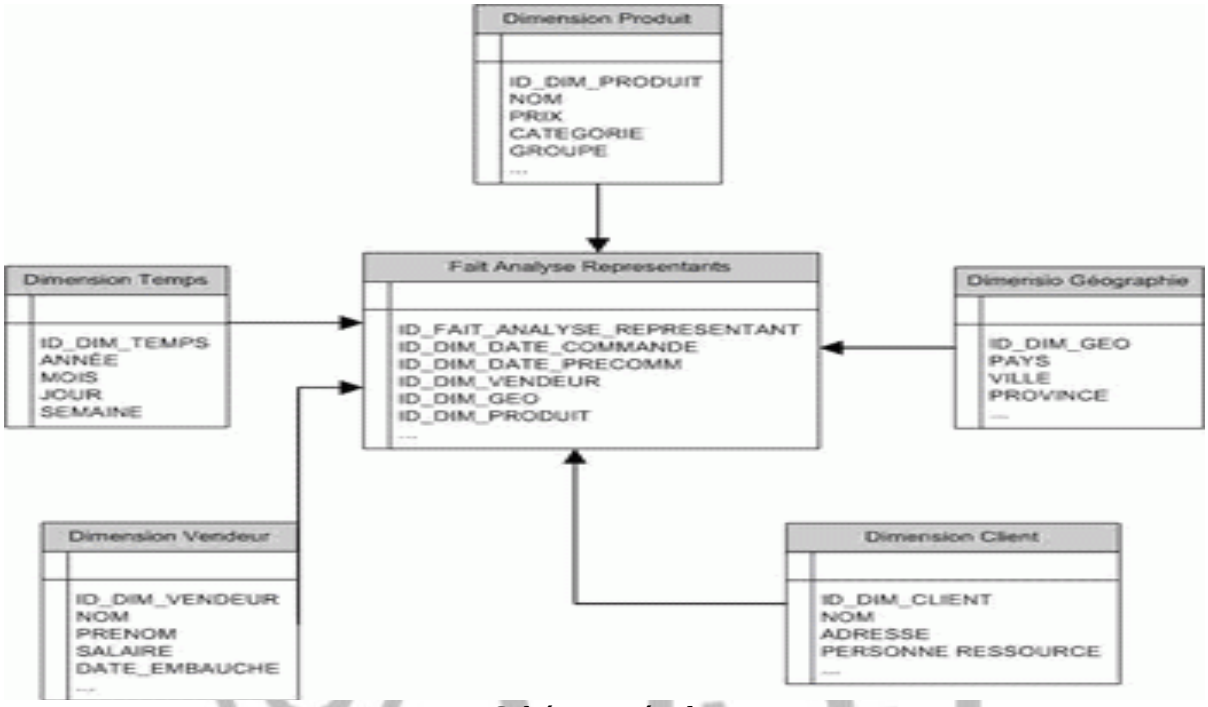


Schéma en étoile

Vous comprenez maintenant pourquoi on appelle ce schéma " modèle en étoile ". Toutes les dimensions sont directement reliées à la table de faits, qui contient les données à analyser. Plusieurs remarques sont à formuler pour ce schéma :

- La table de fait contient ce qu'on appelle des " mesures ", des champs (numériques pour la plupart) sur lesquels on va faire nos analyses, on peut y trouver le montant des ventes nettes, les quantités vendues, les kilomètres parcourus, les quantités en pré commande, etc. La table de faits est reliée aux dimensions par des relations (1, n). Pour analyser une ligne de fait par client par exemple, il faut qu'il y ait une relation entre cette ligne et la dimension client.
- Les tables de dimension contiennent les éléments qu'utiliseront les décideurs pour voir la table de faits. Les utilisateurs pourront ainsi apprécier les montant des ventes par vendeur, par client, ou le kilométrage pour un vendeur pour un client donné (pour voir si ce client est rentable), calculer le coût de revient d'un produit par rapport aux activités des vendeurs, etc.
- On n'utilise JAMAIS la clé d'un système de production comme clé de dimension : pour préserver l'historique des modifications dans l'entrepôt de données.
- La granularité des tables de dimensions et de faits doit être la même : imaginez que la table de faits regroupe les informations par heures et que la table de dimension du temps gère les minutes, il ne sera pas possible de lier la dimension temps et la table de faits (multi détermination).

- Chaque ligne de la table de faits doit avoir une relation avec chacune des tables de dimensions : dans le cas contraire, nous aurons la perte d'informations ou une analyse erronée.
- Il n'existe de relations qu'entre les dimensions et les tables de faits. Il sera beaucoup trop compliqué de gérer et d'utiliser des dimensions liées entre elles. N'oubliez pas que le schéma doit être assimilable par des non informaticiens pour pouvoir l'exploiter.

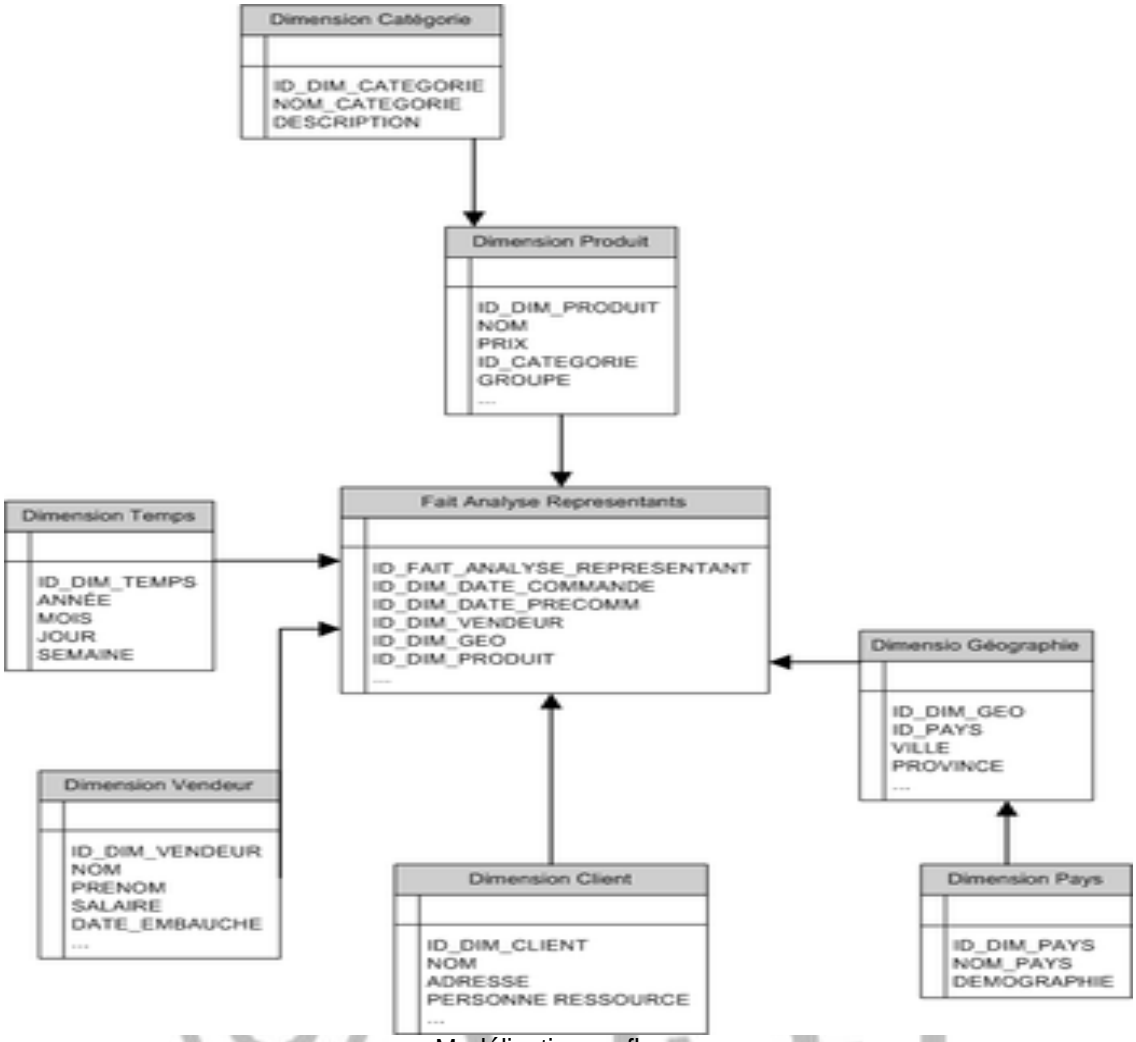
4. Modélisation en flocon : étude d'un cas

La modélisation en flocon étant une variante de la modélisation en étoile, nous prendrons le même cas avec la même analyse.

Il faut savoir que la modélisation en flocon existe pour des raisons de performances. En effet, des dimensions de plusieurs millions de lignes peuvent poser des problèmes de lenteur lors de l'exploitation des données.

Le principe de la modélisation en flocon est de créer des hiérarchies de dimensions, de telle manière à avoir moins de lignes par dimension. Cela va en contradiction avec la dernière remarque de la modélisation en étoile, à la seule chose près que la performance prime sur la structure. C'est la seule façon que les gens ont trouvé pour avoir des résultats clairs et rapides.
Le schéma d'une modélisation en flocon pourrait être comme suit :

© GSJ



Modélisation en flocon

Pour garder une structure simple, gérable et compréhensible, utilisez le plus possible la modélisation en étoile. La modélisation en flocon n'intervenant que lorsque des problèmes de performances apparaissent ou sont facilement prédictibles.

Une règle informelle en BI préconise de floconner que si l'on a la relation (1-1000)[4]. C'est-à-dire que si l'on réussit à créer une hiérarchie de deux dimensions avec une ligne de la dimension père (groupe produit par exemple) faisant référence à plus de 1000 lignes de la dimension fille (produit par exemple). Dans ce cas, il est peut-être temps de penser aux flocons.

> Note : cette règle fût émise en prenant en considération les technologies logicielles et matérielles actuelles. Il ne serait pas étonnant, à notre sens, de voir disparaître la modélisation en flocon avec les avancées technologiques (rapidité des disques durs, technologies OLAP, etc.)

5. Conception d'entrepôts de données

[4] Raph Kimball, Maguy Ross, John Wiky & Sons ltd, The data Warehouse toolkit : The definitive guide to dimensional, seconde edition,

Un entrepôt de données est une vue complète et centralisée des données de l'entreprise. La modélisation en étoile ou en flocon, elle, ne s'intéresse qu'à la conception d'un sous ensemble d'entrepôt, une seule table de fait. On ne peut même pas dire qu'une étoile ou un flocon représente un data Mart, car une fonction de l'entreprise peut comporter plusieurs tables de faits. La fonction commerciale d'une entreprise peut comporter une étoile pour les ventes, un flocon pour les commandes, une autre étoile pour les retours, etc. Ce qui est juste, c'est qu'un entrepôt de données est l'ensemble de ces étoiles et/ou flocons. Mais comment organiser tout ça ?

5.1. Constellation

Une constellation est une série d'étoiles ou de flocons reliées entre eux par des dimensions. Il s'agit donc d'étoiles avec des dimensions en commun. Un environnement décisionnel idéal serait une place où il serait possible de naviguer d'étoile en étoile, de constellation en constellation et de Data Mart en DataMart à la recherche de l'information si précieuse.

Un des indicateurs clés d'une bonne conception d'entrepôt est la grosseur des constellations. En effet, plus la constellation est grosse, plus cela veut dire que vous avez réutilisé vos dimensions, et qui dit réutilisation de dimension, dit dimensions complètes, centralisées et avec une vue orientée entreprise.

Remarque : dès qu'une dimension existante ne correspond pas parfaitement aux besoins d'une nouvelle étoile, on en crée une autre, même si elle est " presque " comme la dimension que nous allions utiliser[5]. C'est pour cela qu'il faut créer, autant que possible, des dimensions génériques et qui soient vraies tout le temps, pour toutes les fonctions de l'entreprise. Ces dimensions pourront être réutilisées et assurer une pérennité des données. Si de telles dimensions ne peuvent pas être créées, il ne faut pas avoir de remords à créer des dimensions similaires mais adaptées aux besoins de la nouvelle étoile. Mais si vous voyez que dans chaque étoile vous êtes obligés de créer une nouvelle dimension " client " par exemple, posez-vous des questions sur votre conception.

5.2. Construire un entrepôt de données.

Récapitulons, nous avons vu comment créer une étoile ou un flocon, nous avons vu que les data marts sont des étoiles regroupées par fonction ou par utilité dans l'entreprise et nous savons qu'un entrepôt est l'ensemble de tous les data marts de l'entreprise. Nous savons concevoir une étoile, mais comment les regrouper pour mettre en œuvre un entrepôt de données ? Trois méthodes existent pour cela :

- ***Top-Down*** : c'est la méthode la plus lourde, la plus contraignante et la plus complète en même temps. Elle consiste en la conception de tout l'entrepôt (exemple : toutes les étoiles), puis en la réalisation de ce dernier. Imaginez le travail qu'une telle méthode implique : savoir à l'avance toutes les dimensions et tous les faits de l'entreprise, puis réaliser tout ça… Le seul avantage que cette méthode comporte est

[5] Alain Fernandez, les nouveaux tableaux du bord des managers, le projet décisionnel en totalité, Eyrolles 6ème édition.

qu'elle offre une vision très claire et très conceptuelle des données de l'entreprise ainsi que du travail à faire.

- ***Bottom-Up*** : c'est l'approche inverse, elle consiste à créer les étoiles une par une, puis les regrouper par des niveaux intermédiaires jusqu'à l'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).
- ***Middle-Out*** : c'est l'approche hybride, et conseillée par les professionnels du BI. Elle consiste en la conception totale de l'entrepôt de données (exemple : concevoir toutes dimensions, tous les faits, toutes les relations), puis créer des divisions plus petites et plus gérables et les mettre en œuvre. Cela équivaut à découper notre conception par éléments en commun et réaliser les découpages un par un. Cette méthode tire le meilleur des deux précédentes sans avoir les contraintes. Il faut juste noter que cette méthode implique, parfois, des compromis de découpage (dupliquer des dimensions identiques pour des besoins pratiques).

La conception d'entrepôts de données est donc un processus en perpétuelle évolution. Sous cet angle, on peut finalement voir l'entrepôt de données comme une architecture décisionnelle capable à la fois de gérer l'hétérogénéité et le changement et dont l'enjeu est de transformer les données en informations directement exploitables par les utilisateurs du métier concerné.

5.3. Comparatif entre les bases de données de l'entreprise

Caractéristique	Base de données de production	Data Warehouses	Datamarts
Opération	gestion courante, production	référentiel, analyse ponctuelle	analyse récurrente, outil de pilotage, support à la décision
Modèle de données	entité/relation	3NF, étoile, flocon	étoile, flocon
Normalisation	fréquente	maximum	rare (redondance d'information)
Données	actuelles, brutes, détaillées	historisées, détaillées	historisées, agrégées
Mise à jour	immédiate, temps réel	souvent différée, périodique	souvent différée, périodique
Niveau de consolidation	faible	faible	élevé

Perception	verticale	transversale	horizontale
Opérations	lectures, insertions, mises à jour, suppressions	lectures, insertions, mises à jour	lectures, insertions, mises à jour, suppressions
Taille	en giga-octets	en téraoctets	en giga-octets

Tableau n° 2 : Comparaison entre différentes bases de données d'une entreprise

Ces différences tiennent au fait que les entrepôts permettent des requêtes qui peuvent être complexes et qui ne reposent pas nécessairement sur une table unique. On peut résumer les conséquences de la transformation d'un Data Waterhouse en Datamart comme suit : *un gain de temps de traitement et une perte de puissance d'utilisation*.

Exemples de requêtes OLAP :

- *Quel est le nombre de paires de chaussures vendues par le magasin « OnVendDesChaussuresIci » en mai 2022 ET Comparer les ventes avec le même mois de 2020 et 2021 ?*
- *Quelles sont les composantes des machines de production ayant eu le plus grand nombre d'incidents imprévisibles au cours de la période 2017-2021 ?*

6. Critique des méthodes de conception d'entrepôts

Les méthodes décrites plus haut sont une très bonne façon de faire du BI avec les moyens techniques d'aujourd'hui. Bien que nous appliquions des compromis entre conception logique et réelle (étoile et flocon) et bien que la réalisation ne ressemble pas toujours à la conception (création de tables d'agrégats, division de tables pour des questions de performance, recréation de dimensions identiques pour des questions de performance, etc.), la représentation des données à base de dimensions et de faits offre un regard très analytique sur les data de l'entreprise et permet de sublimer les limitations du modèle relationnel en troisième forme normale en matière de manipulation de gros volumes des données.

Il reste que, en utilisant ces méthodes régulièrement, nous nous sommes rendu compte qu'il y a beaucoup de bidouillages (bricolages) et beaucoup de gestions d'intégrité manuelle (grâce aux ETL), à un point tel que si l'on n'est pas extrêmement rigoureux dans sa gestion de projet, l'environnement décisionnel peut facilement se transformer en une vraie usine à gaz.

En résumé, étant la meilleure manière de faire du décisionnel pour l'instant, la modélisation en étoile reste une façon très efficace d'organiser les données pour des fins d'analyse. Mais le temps, et la veille technologique, nous diront s'il existera une meilleure manière de faire du décisionnel avec les nouvelles technologies logicielles et matérielles.

7. Conclusion

En conclusion, nous avons démontré qu'étant la meilleure manière de faire du décisionnel pour l'instant, la modélisation en étoile reste une façon très efficace d'organiser les données pour des fins d'analyse. Mais le temps, et la veille technologique, nous diront s'il existera une meilleure manière de faire du décisionnel avec les nouvelles technologies logicielles et matérielles.

Sur ce et en résumé, voici les éléments qui vous feront déduire que votre conception est bonne, à notre avis :

- Que votre entrepôt de données permettra de réaliser toutes les opérations analytiques et donnera aux décideurs des moyens chiffrés pour évaluer les faits voulus.
- Que vos dimensions seront orientées entreprise et pas fonction, avoir le plus possible des dimensions génériques et réutilisables.
- Pas trop de flocons dans votre entrepôt, si c'est le cas, pensez plutôt à changer de serveur ou de moteur de stockage. C'est plus une technique d'optimisation que de conception.
- Avoir des noms d'attributs et de tables compréhensibles par les utilisateurs.
- Bien documenter car un entrepôt non documenté est un entrepôt qu'on ne peut pas faire évoluer, comprendre ou modifier.
- N'oubliez pas, pendant votre phase d'analyse, de lister les outputs et les questionnements des analystes et décideurs de votre entreprise. Ceux-ci serviront de fil conducteur tout au long de votre projet.

© GSJ

8. Bibliographie sommaire

1. Alain Venot, Anita Burgun et Catherine Quantin, *Informatique Médicale, e-Santé – Fondements et applications*, Springer Science & Business, 18 janvier 2013.
2. Isabelle Comyn-Wattiau, Jacky Akoka, *Les bases de données*, PUF, Que sais-je?, 978-2130533139, chap. ix *Les bases de données décisionnelles*, 2003.
3. R. Kimball, L. Reeves, M. Ross, W. Tomthwaite, le data Warehouse : Guide conduite de projet, Eyrolles, 576pages.
4. Raph Kimball, Maguy Ross, John Wiky & Sons ltd, The data Warehouse toolkit : The definitive guide to dimensional, seconde edition, 600pages.
5. Alain Fernandez, les nouveaux tableaux du bord des managers, le projet décisionnel en totalité, Eyrolles 6ème édition, 495pages.

Sites internet

6. http://www.smiler.fr/ consulté le 18 mars 2020
7. http://support.pentaho.com http://www.pentaho.com/ consulté 24 avril 2021