

# Conception d'un entrepôt de données (Data Warehouse)

par Yazid Grim ([Business Intelligen\(ce\)](#))

Date de publication :

Dernière mise à jour :

Nous avons vu dans mes articles précédents ce qu'était le BI, ce que comprenait un environnement décisionnel et qu'il avait comme concept central l'entrepôt de données ou le Data Warehouse.

Intéressons nous maintenant à comment concevoir un entrepôt de données. 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 ? Ce mini cours commencera par introduire (ou réintroduire) les concepts fondamentaux de l'informatique décisionnelle (nécessaires pour la compréhension de cet article), continuera par l'explication des méthodes de conception d'entrepôt de données via une étude de cas, et terminera par une critique de ces techniques et une conclusion mentionnant les indicateurs d'une bonne conception d'entrepôt.

- I - Introduction
- II - Concepts fondamentaux
  - II-A - Entrepôt de données (Data Warehouse)
  - II-B - Data Mart, ou magasin de données
  - II-C - Dimension
  - II-D - Fait
  - II-E - ETL, ou ETC pour les francophiles
  - II-F - Étoile
  - II-G - Flocon
- III - Modélisation en étoile, un cas
  - III-A - Le cas
  - III-B - L'analyse
  - III-C - La solution
- IV - Modélisation en flocon, un cas
- V - Conception d'entrepôts de données
  - V-A - Constellation
  - V-B - Construire un entrepôt de données, un vrai !
- VI - Critique des méthodes de conception d'entrepôts
- VI - Conclusion
- VIII - Remerciements

## I - Introduction

Nous avons vu dans mes articles précédents ce qu'était le BI, ce que comprenait un environnement décisionnel et qu'il avait comme concept central l'entrepôt de données ou le Data Warehouse.

Intéressons nous maintenant à comment concevoir un entrepôt de données.

- 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 ?

Ce mini cours commencera par introduire (ou réintroduire) les concepts fondamentaux de l'informatique décisionnelle (nécessaires pour la compréhension de cet article), continuera par l'explication des méthodes de conception d'entrepôt de données via une étude de cas, et terminera par une critique de ces techniques et une conclusion mentionnant les indicateurs d'une bonne conception d'entrepôt.

## II - Concepts fondamentaux

### II-A - Entrepôt de données (Data Warehouse)

J'estime en avoir assez parlé **ici** et **ici** : mais un peu de répétition ne fait pas de mal !!!

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 à 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, historisées 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 à l'heure où j'écris ce tutorial. 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# Mais intéressons nous à ce qui existe pour l'instant#

### II-B - Data Mart, ou magasin de données

Les Data Warehouses étant, en général, très volumineux et très complexes à concevoir, on a décidé de les diviser en bouchées plus faciles à créer et entretenir. Ce sont les Data Marts. 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.

### II-C - Dimension

Lorsqu'on fait un schéma de BD 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. Et bien en BI, 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 qu'on utilisera pour faire nos analyses.**

### II-D - 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. On aura 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.**

### II-E - ETL, ou ETC pour les francophiles


L'ETL, dont j'ai expliqué les fondements 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).

## II-F - 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).

## II-G - 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 au faits, ça fait comme un flocon :)

 *Note : 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 on peut faire des analyses et avec lequel on peut faire des magasins de données qui, mis ensemble, forment un entrepôt de données.*


### III - Modélisation en étoile, un cas

Nous allons utiliser un exemple pour expliquer la modélisation en étoile. L'important en 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 !

#### III-A - 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, ou sont les endroits où les vendeurs sont le moins efficaces, quelle est la moyenne de ventes des représentants, etc., 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 utilisée, frais de voyage, ventes, promesses de ventes, etc.

#### III-B - L'analyse

 *Note : cette méthode m'a été apprise à l'université Sherbrooke par Monsieur R. Laurin.*

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 :

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 meilleur façon de ce faire, selon moi, 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. Un vrai travail de fourmi et des heures de plaisir :)

Les techniques d'acquisition d'information et d'analyse des besoins étant un sujet à eux seuls, je passerais la main pour ce point # Nous supposons que tout a été fait selon les règles de l'art et nous nous contenterons de compiler :)


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				
<b>Analyse :</b> consommation d'essence, Qte commandée, Qte précommandée, kilométrage,					

nombre de visites, etc.					
----------------------------	--	--	--	--	--

Explications : le tableau suivant a été 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., etc., etc.

Vous pouvez penser que c'est de la paranoïa (comme certains clients) et croire que tous ces problèmes n'apparaîtront pas forcément. Mais rappelez vous qu'un entrepôt ça coûte très cher, et qu'un entrepôt avec des données incomplètes, invalides ou non-conformes à la demande est tout simplement à mettre à la poubelle#

### III-C - La solution

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

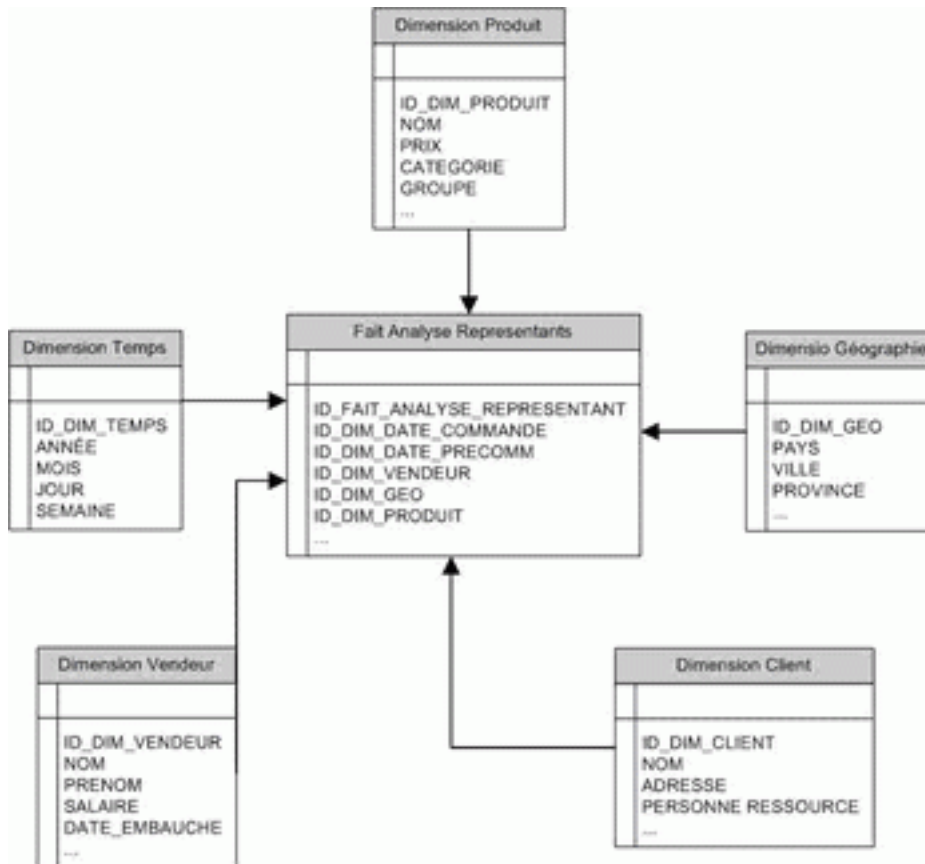


Schéma en étoile

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 faire 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 relation (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 donnée (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 (**voir l'article sur la gestion de l'historique dans un 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, on aurait perte d'information ou 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. N'ayons pas peur de créer des doublons !



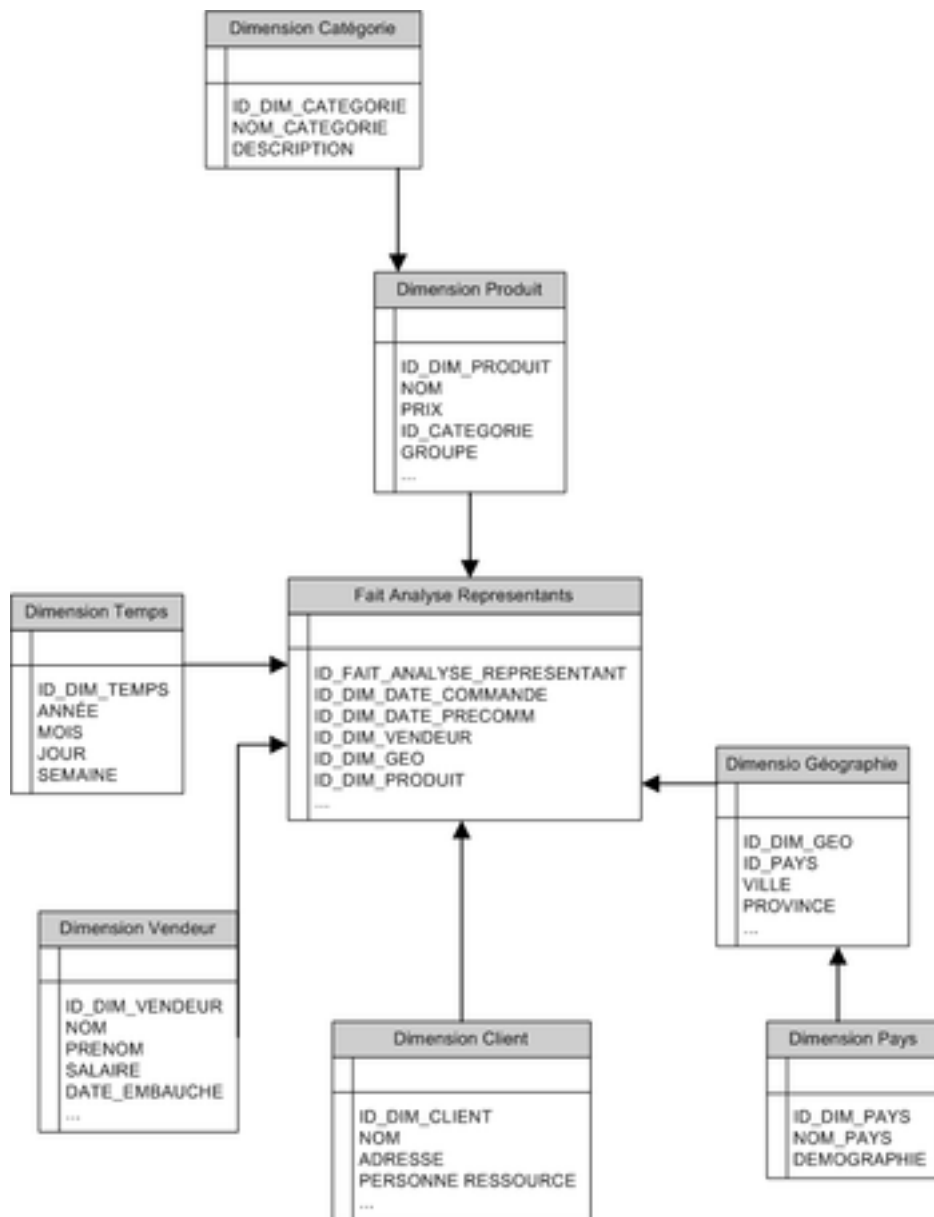
## IV - Modélisation en flocon, un cas

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 dimensions. Vous me direz que cela va en contradiction avec la dernière remarque de la modélisation en étoile, et je vous dirai que vous avez raison, à la seule chose près que la performance prime sur la structure. C'est la seule façon que les gens ont trouvée pour avoir des résultats clairs et rapides.


Le schéma d'une modélisation en flocon pourrait être comme suit :



*Modélisation en flocon*

Conseil : ne "floconisez" pas à tort et à travers. En effet, 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). 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 mon sens, de voir disparaître la modélisation en flocon avec les avancées technologiques (rapidité des disques durs, technologies OLAP, etc.)

## V - Conception d'entrepôts de données

Je sais ce que vous vous dites : mais c'est pas ce qu'on vient de faire la !! Relisez les titres et voyez si je parle d'entrepôts :)

Plus sérieusement, un entrepôt de données, un vrai, selon la définition officielle et pas celle des commerciaux, 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 ?

### V-A - Constellation

Vous remarquez que tous ces termes sont empruntés à l'astronomie et à la météo : étoile, flocon, constellation. Hubert Reeves n'a qu'à bien se tenir :)

Une constellation est une série d'étoiles (tu m'étonnes !) 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.

Je m'explique :

En conception d'entrepôt, il ne faut pas se casser la tête, 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 allons utiliser. 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. Et 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.

### V-B - Construire un entrepôt de données, un vrai !

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 faire une étoile, mais comment les regrouper pour mettre en œuvre un entrepôt de données ? Et bien trois méthodes s'offrent à nous :

- **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 (ie : 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 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 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 (ie : concevoir toutes dimensions, tous les faits, toutes les relations), puis créer des divisions plus petites et plus gérables et les mettre en #uvre. 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).

## VI - Critique des méthodes de conception d'entrepôts

C'est très humblement que j'ajoute cette section car je ne suis pas un chef de file dans le domaine. Le BI me passionne, je lis énormément sur le sujet, mais je n'ai pas encore proposé de méthode de conception :)

Mon avis est que 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 le 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, l'on se rend compte qu'il y a beaucoup de bidouillage et beaucoup de gestion 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.

## VI - Conclusion

Je citerais, en conclusion, les éléments qui vous feront déduire que votre conception est bonne :

- Que votre entrepôt de données permettra de faire 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.
- Documenter, documenter, documenter. N'oubliez pas qu'un entrepôt non documenté est un entrepôt qu'on ne peut pas faire évoluer, comprendre ou modifier. Gare à la rétention d'information !!
- 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.

## VIII - Remerciements

Encore un grand merci à toute l'équipe de développez.com pour leurs soutien et leurs conseils avisés qui ont contribué à l'aboutissement de ce document. Mention spéciale pour l'équipe de correcteurs, Olsimare et à Adrien Artero ;)

