# Mettre en place un entrepôt de données multidimensionnel

par Carlos Da Costa

Date de publication : 19 décembre 2011

L'objectif de ce livre blanc est de démystifier le concept d'entrepôt de données multidimensionnel, à travers une approche simple, courte et pragmatique. J'espère que ce livre blanc vous permettra de comprendre les tenants et aboutissants d'un entrepôt de données multidimensionnel et qu'il vous assistera dans sa mise en oeuvre et sa maintenance.
Ce court mémoire regroupe l'ensemble des points clés d'une approche multidimensionnelle.

# 1 - Auteur

Consultant indépendant et ingénieur en informatique. Carlos Da Costa propose d'assister les organisations dans la mise en place d'un système décisionnel et de gestion de la performance (notamment à partir des technologies proposées par Information Builders : WebFOCUS / iWay software / Performance Management Framework).

Après un master en audit et conception des systèmes d'informations, Carlos évolue au sein de la division en charge du système décisionnel de la 1re banque de détail luxembourgeoise. Ses expériences réussies lui ont permis d'approfondir ses connaissances dans l'architecture d'un système décisionnel et de mettre en place des applications décisionnelles au niveau stratégique, analytique et opérationnel.

Lors de sa dernière mission, Carlos collabore pendant près de deux ans dans un projet de large envergure portant sur la mise en place d'un entrepôt de données multidimensionnel à partir d'une approche « balanced scorecard ». Ce livre blanc est en partie le fruit de cette expérience.

# 2 - L'entrepôt de données multidimensionnel en théorie

## 2-A - Pourquoi mettre en place un entrepôt de données ?

L'entrepôt de données a pour objectif de centraliser et faire converger l'ensemble des données d'une organisation dans le but de faciliter l'accès à l'information, l'analyse et la prise de décision. Ce point de convergence de l'information devrait permettre *in fine* l'automatisation et la standardisation (définition unique) d'indicateurs.

Seul un entrepôt de données, global, transverse et "historisé" permet une mise en perspective équilibrée de l'organisation. En effet, dans la mesure où chaque indicateur pertinent est directement ou indirectement corrélé à un autre, il influe directement ou indirectement sur un objectif, la tactique et la stratégie de l'organisation.

La démarche de centralisation et de mise en perspective entreprise à travers la mise en place d'un entrepôt de données doit pouvoir vous permettre de produire une vue « balanced scorecard (1) » de votre organisation afin de planifier, vérifier, réagir et être efficace.

## 2-B - L'analyse multidimensionnelle une approche naturelle

Même si le mot « multidimensionnel » peut paraître savant et complexe, cette approche se veut avant tout simple et naturelle.

Lorsque vous produisez un rapport ou que vous vous interrogez sur la valeur d'un indicateur, il y a de fortes chances que cela corresponde à l'expression d'un **contexte** (vos **dimensions**) et de **mesure**(s) (les **faits**), exemple :

« Je souhaite connaître mon **chiffre d'affaires**, **par** produit, **par** agence **pour** le mois d'avril.»

Les clauses « POUR » et « PAR » (respectivement les clauses WHERE et GROUP BY de vos requêtes) correspondent à vos dimensions. Le chiffre d'affaires lui correspond à la mesure (le fait). Vous utilisez donc déjà, de façon consciente ou non, une approche "multidimensionnel".

Si le mot dimension vous met encore mal à l'aise, remplacez celui-ci par axe d'analyse (c'est "blanc bonnet, bonnet blanc"). Nous aborderons en détail la notion des faits et des dimensions tout au long de ce papier.

## 2-C - Entrepôt multidimensionnel contre Entrepôt relationnel ?

Personnellement, je pense que l'on peut associer la production d'un modèle multidimensionnel avec celle d'un système d'entrepôt de données de type relationnel. Constituer un marché de l'information multidimensionnel n'est en aucun cas incompatible avec une approche relationnelle, bien au contraire, elles sont complémentaires.

En effet, la convergence des systèmes de transactions opérationnels sera difficile à modéliser à travers un schéma multidimensionnel. De la même manière, la mise en place d'un système de mesure et d'exploitation de l'information sera difficile à réaliser à travers un simple entrepôt relationnel. En tant que concepteur du système d'information décisionnel, vous n'avez pas à faire le choix entre une approche plutôt qu'une autre. Vous devez connaître les tenants et aboutissants de votre projet décisionnel afin de proposer la combinaison « relationnel/multidimensionnel » gagnante.

## 2-D - L'entrepôt de données multidimensionnel

Le concept d'entrepôt de données multidimensionnel consiste (cf. partie : « Pourquoi mettre en place un entrepôt de données ») en la mise en place d'un marché de l'information unique et transversal à partir des différents systèmes transactionnels d'une organisation. Ce marché de l'information est construit à partir de l'analyse des processus clés de l'organisation dans le but de faciliter l'accès à l'information, l'analyse et la prise de décisions.

### 2-D-1 - En pratique...

La méthode de modélisation dimensionnelle est basée sur une architecture de type « bus (2) ». Une table de fait est produite à partir des données transactionnelles et de l'analyse d'un processus afin de répondre aux questions des analystes. Une table de faits est constituée de mesures ainsi que de clés de dimension. Ces clés de dimension permettent la jointure entre la table de faits et les tables dimensions. Les dimensions sont communes à l'ensemble des services de l'organisation et représentent des axes d'analyses stratégiques.

| | Dimension Organisation | Dimension Produit | Dimension Localité | ... | Dimension Temps |
|---|---|---|---|---|---|
| Délai moyen de traitement | | X | X | | X |
| Volume des ventes | X | X | | | X |
| ... | | | | | |
| Processus « xyz » | X | | X | | X |

L'interrogation des tables de faits à travers les tables de dimensions produit des rapports agrégés qui sont théoriquement capables de répondre à l'ensemble des besoins en information des utilisateurs.

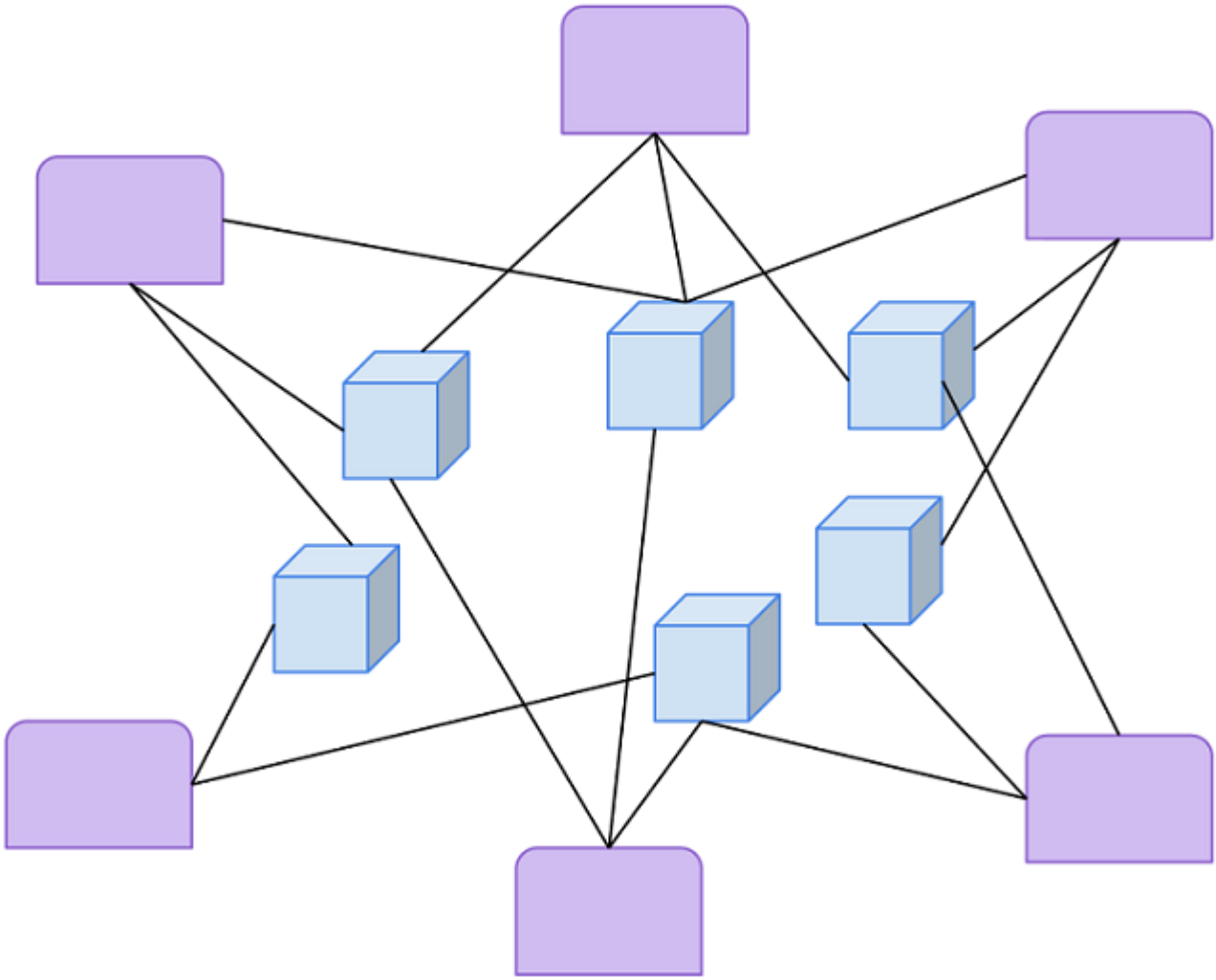


Figure 1 : modélisation entrepôt de données multidimensionnel

DIMENSION (peut être représentée à travers plusieurs tables)

FAIT (plusieurs mesures peuvent être calculées à partir d'une seule table. La représentation d'un processus peut exiger plusieurs tables de fait)

## 2-E - Une construction itérative en quatre étapes

La construction d'un entrepôt de données multidimensionnel est itérative et incrémentale. Au fur et à mesure de sa construction l'entrepôt de données doit devenir global et transversal via l'intégration de mesures de processus clés. L'intégration d'un nouveau processus passe par quatre étapes (3) :

### 2-E-1 - Étape 1 : sélection d'un processus clé

Le choix du processus clé doit se faire en fonction de son importance au sein de l'organisation. C'est aux stratèges et aux responsables opérationnels de prendre cette décision.

Voici quelques pistes de réflexions pour vous permettre de prioriser vos processus clés :

- quel est le but premier de mon organisation et quels sont les processus qu'il serait nécessaire d'analyser ? Ici on parle bien de processus et non de départements, de services, ou de divisions... Une organisation est un tout, votre schéma de pensées doit être transversal et d'abord lié aux processus fondamentaux ;
- quels sont les indicateurs de performance, d'éclairage ou de risque pertinents aux niveaux stratégique, analytique et opérationnel ? Même si les interrogations d'un stratège sont différentes des questions d'une personne opérationnelle, la source de données reste la plupart du temps a même. Pour mettre en place une stratégie de modélisation à long terme vous devez avoir conscience des besoins aux différents niveaux des prises de décisions ;
- quels exemples de décisions concrètes pourraient être pris à partir de ce système de mesure, et ce pour chaque personne de chaque niveau mentionné. Si aucune décision ne peut découler de votre système décisionnel c'est qu'il est sans intérêt. N'oubliez jamais qu'à chaque mesure doit correspondre un ou plusieurs leviers d'actions (plus le niveau s'approche de l'opérationnel, plus les leviers doivent être simples) ;
- quels gains réels pour l'organisation apporteraient ces décisions : hausse de la productivité ? Optimisation des coûts ? Meilleure réactivité sur les marchés ? Accroissement de la prise de part de marché ? Améliorer la fidélisation clientèle ?

Gardez à l'esprit que la mise en place et l'évolution d'un système a un coût. L'étape du choix du processus à modéliser est fondamentale, si celui-ci n'est pas pertinent, les décisions qui en découlent n'auront pas d'impact sur votre activité.

### 2-E-2 - Étape 2 : choix de la granularité stockée

Dans cette étape vous devez faire le choix du niveau de détail de l'information que vous souhaitez conserver.

Exemple, sur un ticket de caisse que conserveriez-vous ? Chaque ligne de produit d'achat ? Chaque produit différent ? Regrouperiez-vous le montant par famille de produits ? Ou seul le montant total vous intéresse ?

Les responsables opérationnels ont tendance à vouloir tout stocker, et ils ont raison. En effet, plus le degré de granularité est fin plus la capacité d'analyse est élevée (en terme de croisement entre vos différents axes d'analyses). Mais attention, malgré les progrès réalisés dans le traitement et le stockage de grands volumes de données, plus l'information est détaillée et moins votre possibilité d'historique dans le temps est importante. Le choix cornélien de la granularité stockée est donc rempli de compromis. Dans tous les cas, il doit se faire avec l'aval de l'utilisateur final, du DBA (4) et du responsable de l'entrepôt de données.

### 2-E-3 - Étape 3 : choix des axes d'analyses (dimensions)

Si le mot dimension est utilisé dans la majorité des outils de modélisation, je préfère l'expression « axe d'analyse ». Chaque processus peut-être analysé sous différents axes : par période de temps, par typologie de client, par typologie de produits, par localité, par site de production, etc. Le but de cette étape est de choisir quels sont les axes d'analyse adéquats pour le processus en question. Rappelez-vous que l'architecture d'un entrepôt de données dimensionnel repose sur le principe de « bus », c'est-à -dire que ce sont les données de votre processus qui vont venir se brancher à vos dimensions et non l'inverse. En d'autres termes, les dimensions définies ici sont communes à l'ensemble de

l'organisation et à l'ensemble des processus. Avant de créer un nouvel axe d'analyse, veillez à ce que celui-ci puisse être suffisamment générique pour convenir à l'ensemble de votre organisation.

Pour trouver les dimensions adéquates au processus mesuré, posez-vous les questions suivantes :

- à qui ces données pourraient-elles être utiles ?
- comment les analystes regrouperaient-ils les données ?
- comment les analystes filtreraient-ils les données ?
- quels sont les titres de colonne des rapports actuellement produits ?

Dans tous les cas, faites preuve de bon sens, restez simple et pragmatique et n'essayez pas d'être original. C'est à partir du croisement de dimensions simples et compréhensibles que vous allez pouvoir prendre des décisions (pour plus de détail, je vous renvoie vers le chapitre dédié aux tables de dimensions). Cette étape du choix des axes d'analyses doit-être faite par une équipe hétérogène (opérationnels, analystes et représentant du système décisionnel en place). Elle doit être simple et naturelle pour tous les intervenants.

### 2-E-4 - Étape 4 : quelles mesures (faits)

Quelles mesures de performance, d'éclairage ou de risque serait-il pertinent de rattacher à ce processus clé ? Cette réponse doit être fournie à la fois par des représentants opérationnels expérimentés et des collaborateurs stratégiques (manager, stratège). Ici votre mantra doit être : "Not everything that can be counted counts, and not everything that counts can be counted." [Albert Einstein]. Ce qui compte ne peut pas toujours être compté, et ce qui peut être compté ne compte pas forcément. En gardant cette phrase à l'esprit vos avez plus chance d'assurer la pérennité de votre système.

**Ce qui compte ne peut pas toujours être compté** : en effet, votre système décisionnel se limitera toujours à des faits tangibles. Vos collaborateurs les plus imaginatifs vous proposeront sans doute des techniques pour mesurer certain faits intangibles (exemple : la satisfaction client), si c'est le cas attention... Même s'il est possible de sonder ses clients via un questionnaire et calculer un ratio de satisfaction, ces données de type « exceptionnelles » n'ont rien à faire dans votre système décisionnel. Celui-ci doit d'abord reposer sur des faits tangibles issus d'activités tracées par votre système d'information (c'est-à-dire : pas de données disponibles de façon récurrente = pas de mesure). Dans ce cas précis, on préférera représenter la satisfaction client à travers un objectif qui encapsule des mesures tangibles en rapport avec votre activité, exemple : le nombre de réclamations traités, le nombre de retours de marchandises ou le nombre de connexions hebdomadaires,etc.

**Ce qui peut être compté ne compte pas forcément** : mesurez des faits qui ont un enjeu et un sens pour les décideurs ainsi que pour les opérationnels. Mesurez d'abord des indicateurs actés et connus. Rappelez-vous que rien n'est gratuit ! Le coût de votre système décisionnel sera proportionnel au nombre de mesures stockées. Plus il y a de mesures, plus il y a de volume de données, plus il y a de source de données différentes, plus il y a de maintenance lors de l'évolution des systèmes opérationnels, plus vous aurez besoin de personnel qualifié affecté à ces travaux de maintenances, etc. Avant d'intégrer une mesure, vous devez intégrer la notion « pertinence / coût ». Chaque mesure doit représenter un coût raisonnable proportionnel à la valeur ajoutée qu'elle apporte. Seuls les responsables techniques pourront réellement vous indiquer le coût de l'intégration : volume de données, coût de stockage, coût CPU, coût de maintenance, impact sur le système actuel, etc. Rappelez-vous que vous ne mesurez pas par plaisir mais par nécessité. Les personnes issues des activités opérationnelles avec une forte expérience du métier sont souvent pragmatiques et de très bon conseil.

## 3 - L'entrepôt de données multidimensionnel en pratique

### 3-A - Les tables de fait

Tout d'abord démystifions la chose... Une table de fait n'est rien d'autre qu'un ensemble de données structurées, composé de champs de type dimension (le contexte) et champs de type mesure (les faits).

Un processus d'entreprise peut être représenté à l'aide d'une ou plusieurs tables de fait. À ce jour on dénombre trois types de tables de fait (5) : les tables de fait de transaction, les tables de fait périodique et les tables de fait récapitulative.

## 3-A-1 - Les règles d'or des tables de fait

Voici une série de conseils qui pourront vous être utiles lors de la mise en place de n'importe quelle table de fait (n'hésitez pas à les relire avant d'en concevoir une nouvelle, vous économiserez sans doute un temps précieux) :

- l'unicité d'une ligne d'une table de fait doit toujours pouvoir être garantie par la concaténation de ses champs dimensions. En effet, même si les concepteurs rajoutent souvent une clé physique de type « auto-incrémenté » la clé logique de table reste l'unicité du contexte. En d'autres termes, si un fait (une occurrence) a exactement le même contexte (même valeurs de dimensions) qu'un autre fait, cela doit être la même ligne ;
- une table de fait contient toujours la dimension temps. Bien que la durée de rétention varie en fonction de la granularité de la mesure, TOUTES vos mesures doivent comprendre un historique pour vous permettre de produire des tendances. Un entrepôt de données quel qu'il soit doit comprendre la dimension temps (c'est à l'équipe métier de juger de la durée minimale de rétention requise pour leur permettre de prendre des décisions et non pas aux administrateurs de base de données) ;
- les mesures stockées dans une table de fait sont (presque) toujours de types numériques et additifs. Cela implique les règles suivantes : les ratios sont toujours stockés à travers deux champs distincts (numérateur et dénominateur). Les mesures stockées ne correspondent jamais à une moyenne (la somme des moyennes ne correspond pas à la moyenne des sommes !). Prenez garde au calcul des délais, vous ne stockez pas les dates mais la durée effectif du délai calculé quel qu'en soit l'unité (jours, minutes ou secondes) ;
- les données d'une table de fait sont figées. Une table de fait stocke une situation passée et révolue (sauf table de fait récapitulative). Il ne doit pas y avoir d'opération de mise à jour sur la table une fois le chargement effectué et que ses données sont à disposition des utilisateurs (sauf correction). Si certains de vos utilisateurs génèrent le rapport « situation financière du 31 mars 2011 » le 5 avril et que d'autres utilisateurs génèrent ce même rapport le 10 mai, ils doivent absolument avoir accès aux mêmes chiffres ;
- une table de fait est toujours interrogée à partir d'un contexte donné. Sa volumétrie ainsi que sa nature transverse (multitude de dimensions) vous obligent à interroger une table de fait à partir d'un contexte bien particulier (filtre de dimensions) ;
- des vues d'une même table de fait peuvent être produites avec des filtres entièrement différents parce que les besoins d'un service à un autre sont entièrement différents. Ce qui importe, c'est que les deux services aient la même définition de la mesure, car les rapports sont produits à partir de la même source de données ;
- une table de fait ne doit pas contenir de ligne artificielle valorisée à zéro. Il faut donc éviter les alimentations de type « produit cartésien de dimension ». Exemple, si votre système opérationnelle ne contient pas d'information sur la vente du produit « REF-0001 » vendu dans l'agence « ABC » qu'il en soit ainsi ! Ne créez pas la ligne « REF-001 | ABC | ... | 0 » sous peine d'explosion du volume de données ;
- une table de fait ne comprend que les clés des dimensions, sous forme de clé étrangère (numérique de préférence et dénuée de sens pour faciliter la maintenance, cf. chapitre sur les tables de dimensions). Si les tables de fait peuvent être très longues en termes de nombre d'occurrences, elles doivent être étroites en largeur pour pouvoir les compresser en terme d'espace et être performante ;
- le volume d'une table de fait dépend (en partie) du nombre de dimensions AINSI QUE de la structure de celle-ci (profondeur, nombre d'occurrences). En d'autres termes plus le contexte est précis et plus votre table de fait sera volumineuse et difficile à maintenir.

## 3-A-2 - Quelques patrons liés aux tables de fait

Comme nous l'avons déjà précisé le concepteur d'entrepôt de données distingue différents types de tables de fait : de transaction, périodique et récapitulative. Même si nous classons les tables de fait par type, le principe reste identique : c'est une structure contenant des champs « clé étrangère » de table de dimensions (contexte) et des champs de type mesure.

Connaître et maîtriser les différents types de tables de fait vous aidera tout au long de la conception de votre marché de l'information. Cela vous permettra de mettre en place plus rapidement un SID complet. En effet, malgré la

transversalité du système, les besoins de reporting sont redondants. En connaissant les patterns à appliquer vous accélérerez la mise en place ainsi que la qualité et la pertinence de l'information fournie.

### 3-A-2-A - Table de fait de transaction

C'est le type le plus commun et le plus fondamental, l'ensemble de votre entrepôt repose sur les tables de fait de transaction.

Principe : comme son nom l'indique elle repose sur la transaction du système opérationnel. Vous devez définir les clés des dimensions de chacune des transactions opérationnelles et extraire les mesures qui vous intéressent. Ici, la difficulté provient de la gestion des dimensions (notamment le traitement des nouvelles occurrences) ainsi que du compromis volumétrie/précision du contexte fourni (cf. étape choix de la granularité de la table de fait).

Attention, il ne s'agit pas d'avoir une ligne par transaction, bien au contraire. Ici l'unicité de l'occurrence est marquée par l'unicité de son contexte (ensemble des clés de dimensions, cf. règle 1 des propriétés des tables de fait). Notez que plus vous agrégez les faits, moins il sera possible de proposer des dimensions dégénérées (cf. chapitre « Les Dimensions »).

Vous n'avez pas à faire d'opération de mise à jour sur une table de fait de transaction (sauf erreur de chargement). La table de fait de transaction représente le niveau le plus détaillé que peut proposer votre entrepôt sur le processus en question, c'est pourquoi le choix de la granularité de celle-ci est si importante.

### 3-A-2-B - Table de fait périodique

Une table de fait périodique est généralement construite à partir d'une table de fait de transaction. Elle représente soit l'image d'une table de fait de transaction à un moment T ; soit la synthèse d'une table de fait de transaction à travers l'agrégation des mesures sur ses dimensions.

Les tables de faits périodiques permettent d'analyser un volume de transaction beaucoup plus important et ainsi de dégager des tendances. En contrepartie, le nombre et les profondeurs des dimensions sont réduits. En d'autres termes, plus vous agrégez les mesures de la table de transaction plus votre contexte d'analyse sera limité mais plus vous pourrez traiter de transactions et dégager des tendances. Comme toujours tout est histoire de compromis...

Une bonne approche peut être de proposer une table de fait de transaction avec un contexte d'analyse très riche et très détaillé sur un axe de temps limité (trois mois) couplé à une table de fait périodique avec un contexte moins riche (profondeurs de dimension moins importante ou suppression de dimensions) mais un axe de temps plus important (36 mois).

Tout comme sur les tables de faits de transaction, hormis une erreur de chargement, les tables de fait périodiques ne doivent pas être soumises à des opérations de mise à jour, la production de ce type de tables consiste à agréger et à charger les données.

Si votre processus possède un nombre de dimensions restreint avec très peu d'occurrences, le volume traité par votre table de fait de transaction peut éventuellement vous permettre de proposer un axe temps suffisant à vos analystes, auquel cas la mise en place d'une table de fait périodique n'est pas nécessaire, voir déconseillée.

### 3-A-2-C - Table de fait récapitulatif

Très utile et complémentaire aux autres types de tables de fait, les tables de fait récapitulatives sont aussi plus complexes et plus difficiles à maintenir.

La table de fait récapitulative n'est pas fondée sur la transaction ou sur l'axe temps mais sur l'analyse détaillée d'une dimension. Exemple : l'activité d'un site de production. Dans les autres types de table de fait la référence d'un site de production était un axe d'analyse possible mais le but premier était l'analyse du processus à travers ses différentes

mesures. Dans les tables de fait récapitulatives les rôles s'inversent, on souhaite analyser une dimension à travers différents processus. La table de fait récapitulative va donc contenir différentes mesures liées à différents processus pour une même occurrence de dimension.

Les tables de fait récapitulatives permettent une analyse transversale d'une dimension clé stratégique. Contrairement aux autres types de tables de fait, l'essentiel des opérations effectuées sur ce type de table sont des mises à jour. Les tables de fait récapitulatives sont complémentaires aux autres types de table fait et leur contenu repose entièrement sur les tables de fait de type transactionnelle.

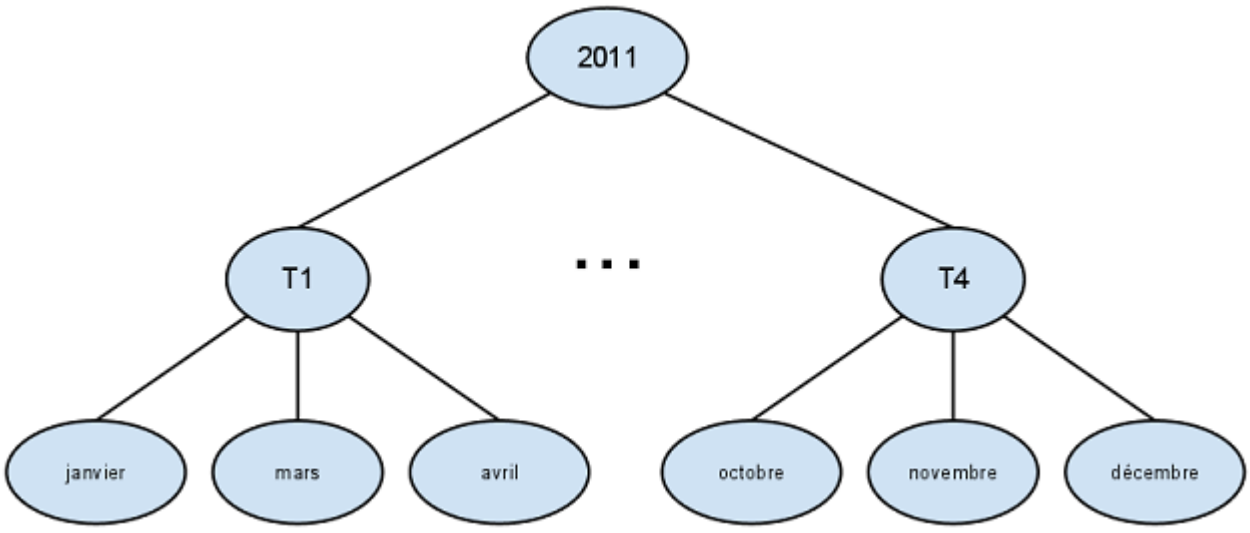
La notion de table de fait récapitulative peut encore vous paraître abstraite, prenons un exemple basique pour illustrer ces propos. Imaginons que vous disposiez de deux tables de faits de type transaction. La première est liée à un processus de vente (dimensions utilisés : Temps, Magasin, Employé, **Produit**, Age client, Localité client). La seconde est liée à un processus de réclamation (dimensions utilisés : Temps, **Produit**). Vous pourriez donc créer une table de fait récapitulative « Mesure Produit » qui intègre pour chaque produit la notion de ventes et de réclamation.

## 3-B - Les tables de dimension

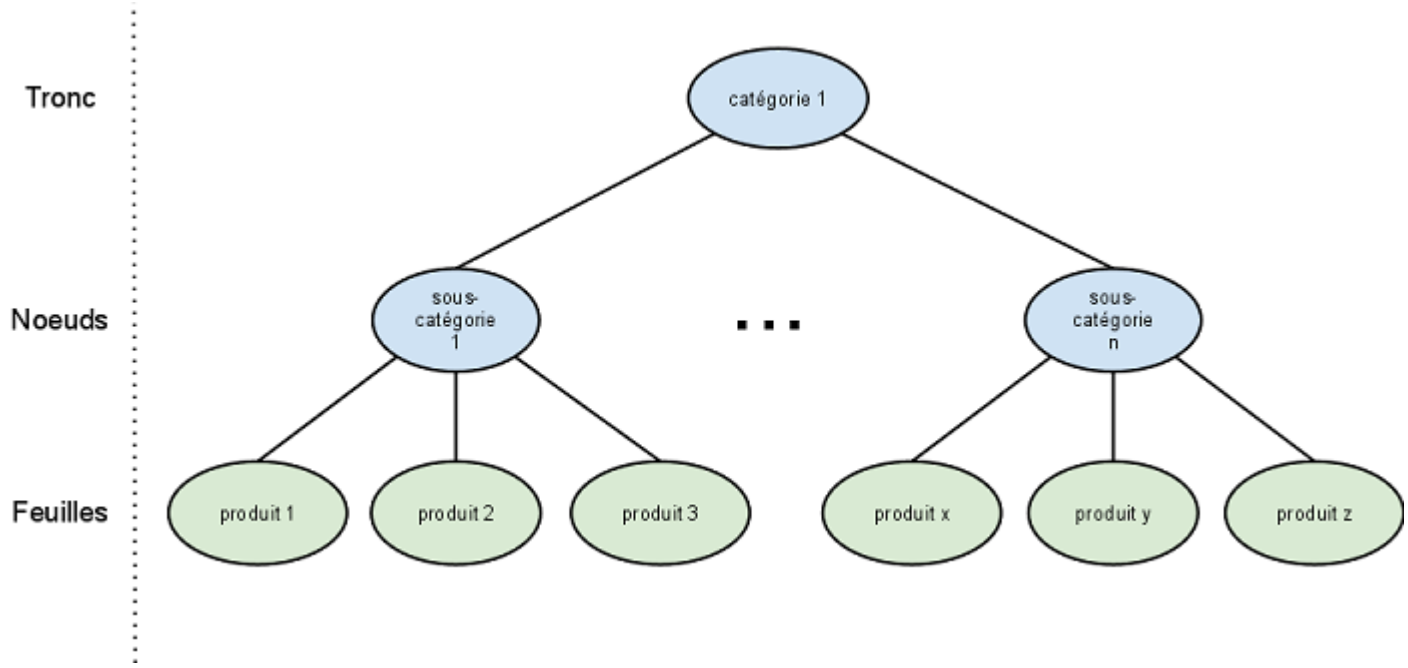
Les dimensions correspondent donc à vos axes d'analyses, exemple : Produits, Agences, Temps, Rating Client, Localité, Ancienneté, Service, etc. On retrouve souvent les mêmes dimensions dans un même secteur d'activité et pour cause, les utilisateurs métier expérimentés savent très bien distinguer l'utile de l'inutile.

Une dimension correspond à un arbre, le but étant d'être capable de faire joindre les faits à un noeud ou à une feuille de cet arbre. Grâce à cette jointure entre mesures et dimensions, vous permettez aux analystes de calculer un agrégat par noeud, par branche ou pour l'ensemble du tronc, le tout sur plusieurs dimensions à la fois (croisement de dimensions).

Exemple de dimensions temps :




Exemple de dimension produit :




Le décideur peut alors analyser l'agrégat des mesures associées à ces dimensions par trimestre et sous-catégories de produits.

Considérez les dimensions comme l'interface entre l'homme et les données de votre système de mesure. Tout rapport sera toujours construit à travers les dimensions que vous mettez à disposition des utilisateurs. C'est pourquoi, elles se doivent d'être simples et d'avoir du sens pour l'ensemble de vos utilisateurs.

En croisant les dimensions, les analystes construisent « le contexte » du rapport. Si le « contexte » est incompréhensible vos rapports le seront tout autant.

## 3-B-1 - Les règles d'or des tables de dimensions

Voici une série de conseils qui pourront vous être utiles lors de la mise en place de vos tables de dimensions :

- même si une table de dimension est souvent très dé-normalisée, veuillez à toujours respecter la première forme normale (unicité de la clé et valeur atomique des champs) ;
- afin de faciliter la maintenance due à l'évolution des dimensions et dans le but de faciliter les jointures avec la table de fait et de réduire la volumétrie de celle-ci, préférez l'utilisation d'une clé physique générique et dépourvue de sens et qui soit de type « auto-incrément » plutôt qu'une clé logique composée de plusieurs champs pour vos tables de dimensions ;
- même s'il existe différents types de processus pour mettre à jour une table de dimension, un conseil : ne supprimez jamais une ligne d'une table de dimension. Si vous ne souhaitez pas afficher des dimensions obsolètes, utilisez un champ ETAT ou FLAGAFFICHAGE. Il est très difficile de mesurer l'impact de la suppression d'une valeur de dimension sur l'ensemble des rapports de vos utilisateurs. Si un rapport est produit un jour « j » ses valeurs doivent être identique s'il est produit jour « j+30 » ou « j+365 ». Si vous supprimez physiquement une ligne d'une dimension cela ne sera plus le cas ;
- pour faciliter la production de rapports et le calcul d'agrégats, faites en sorte de créer des dimensions respectant une hiérarchie simple de type 1 vers n. Imaginez les relations au sein d'une même dimension comme celles des poupées russes qui s'emboîtent ou comme un arbre vis-à-vis de ses branches et de ses feuilles. Cette approche est très naturelle pour l'utilisateur qui facilite la manipulation des données. Exemple :
  - un domaine de produit contient une à plusieurs catégories de produits ; une catégorie de produit contient une ou plusieurs références de produit,
  - une région contient un ou plusieurs départements, un département contient un ou plusieurs sites de production, un site de production contient 1 ou plusieurs services ;

- En d'autres termes, éviter les relations n vers n dans une dimension. Délimitez strictement la hiérarchie de vos dimensions : un site de production appartient à un et un seul département, un département n'appartient qu'à une et une seule région. Si « un département peut appartenir à plusieurs régions » dans votre dimension, vous risquez de complexifier l'accès à l'information et la production de rapports. Pensez aux poupées Russe !
- Évitez les modèles de dimensions en flocon, ils sont plus difficiles à maintenir et moins efficaces. Normalement, vous pouvez toujours les éviter en créant deux dimensions complètement distinctes. Exemple, un produit peut appartenir à une typologie de produit et un produit appartient également à un regroupement de produits. Dans ce cas il est préférable de créer deux axes d'analyses : Typologie Produit et Produits.

Vous l'avez compris, l'approche multidimensionnelle est simple et naturelle. Une organisation compte, selon la complexité de son activité, de trois à douze dimensions (6) . Si vous en avez plus, vous devez simplifier votre modèle. Le croisement de dimensions est souvent interprété comme une dimension en soit. Exemple, vous êtes peut être tenté par une dimension « Segmentation clientèle », mais celle-ci ne revient-elle pas au croisement de votre dimension Métier et de votre dimension ? Âge ?

N'oubliez pas que la valeur ajoutée d'un système multidimensionnel provient essentiellement de ses dimensions et de la possibilité de les croiser, ce qui permet de mettre en perspective des mesures dans un contexte riche de sens.

## 3-B-2 - Quelques patrons liés aux tables de dimensions

Voici quelques modèles connus et reconnus à appliquer lors de la modélisation de vos dimensions.

### 3-B-2-A - Les dimensions à « jeux de rôle »

Ce modèle représente le fait d'implémenter une seule dimension physique pouvant représenter plusieurs axes d'analyses pour les utilisateurs ; typiquement la dimension « Temps ». En tant que concepteur vous pouvez produire différentes vues de cette de table de dimension « Temps » afin de permettre aux utilisateurs d'utiliser différents axes d'analyses (exemple : date d'achat, date de livraison...). Vous ne gérez donc qu'une seule dimension physique tout en mettant à disposition de vos utilisateurs plusieurs axes d'analyses.

### 3-B-2-B - Les dimensions à changement rapide

Lorsqu'une dimension comporte un très grand nombre de champs évoluant en continu, il est préférable de regrouper ces champs dans une table satellite (relation 1->N). Les champs à évolution constante se retrouvent alors isolés et disponibles à travers une simple jointure. Le fait d'isoler ces champs à évolution rapide facilite la maintenance et le traitement des erreurs de chargement. Cela permet également de réduire l'impact sur votre dimension en cas d'erreurs. Exemple, les produits financiers comme les actions avec leurs statuts et valorisations à un moment T.

### 3-B-2-C - Les flags dimensionnels (à utiliser avec modération)

Il est parfois intéressant de connaître si une occurrence de table de fait appartient à un « groupe », sans pour autant que ce « groupe » soit suffisamment important pour en faire une dimension. Dans ce cas précis l'utilisation de flags dimensionnels peuvent s'avérer utile. Le flag dimensionnel est un champ de nature dimensionnelle qui n'a pas une table de dimension. Le flag dimensionnel n'est pas une clé étrangère mais un flag avec ses valeurs propres (celles-ci doivent être connues des utilisateurs). Exemple, pour chaque occurrence d'une table de fait liée à un processus de vente, nous pourrions créer le flag dimensionnel « Vente Direct » avec les valeurs possible ['O' | 'N']. Il permet aux analyses de faire des statistiques comparatives entre les ventes dites « directes » et les autres à partir d'une seule table de fait.

### 3-B-2-D - Les mini-dimensions ou dimensions déportées

Lorsque la table de fait comporte beaucoup d'attributs spécifiques liés aux mesures, il est d'usage de déporter l'ensemble de ces attributs dans une table de type « mini-dimension » uniquement utilisée pour interroger cette table de fait. Une « mini-dimension » est donc généralement liée à une seule table de fait. Exemple, dans une table de fait qui comporte des mesures liées aux crédits, les attributs du crédit peuvent être déportés dans une mini-dimension.

### 3-B-2-E - Les dimensions dégénérées

Un nom bien savant pour un modèle relativement simple. Dans certains rapports, il s'avère parfois utile de proposer un accès vers le détail ou d'ajouter des informations sur la transaction traitée. La dimension dégénérée consiste en un champ dans votre table de fait qui fait référence à l'identifiant de la transaction dans la base opérationnelle ou vers votre ODS (7) . Ce modèle très utile permet de vérifier la véracité des chiffres fournis ainsi que la valeur des dimensions indiquées. Exemple, un champ « référence du contrat » dans une table de fait de transaction vous permettra de créer un lien vers votre système opérationnel afin d'accéder à ses propriétés.

### 3-B-2-F - Les dimensions horodatées

Ce modèle consiste à ajouter des dates de début et de fin de validité à une table de dimension. Cette pratique peut s'avérer particulièrement utile lorsqu'on sait que les données de la dimension sont « périssables », exemple : une date de transfert ou de licenciement d'un employé (l'employé ne fait plus partie du service il ne doit pas apparaître dans la dimension) ; une date de fin de disponibilité d'un produit (le produit n'est plus référencé, il ne doit pas apparaître dans la dimension)...

### 3-B-2-G - La dimension audit

Véritable « cerise sur le gâteau », la dimension audit consiste à loger l'ensemble des informations sur le chargement de vos tables de fait et de dimensions. Cette dimension contient, entre autres, l'ensemble des métadonnées renseignées dans vos job ETL. Chaque feuille de cette dimension correspond à un job ETL et l'ensemble des mesures (nombre d'opérations INSERT, temps d'exécution, temps CPU, temps CPU réservé à la qualité de données, etc.) sont bien entendu regroupées dans une table de fait.

Vous pouvez également rattacher la dimension audit à toutes les occurrences de vos tables de fait à travers un champ « référence job ETL ». Cela vous permettra de connaître « qui à chargé quoi, quand et comment ? ».

La dimension audit permet une traçabilité et démontre une certaine rigueur ainsi qu'un certain professionnalisme dans la mise en place de votre système décisionnel.

## 3-C - Maintenance d'un entrepôt de données multidimensionnel

La maintenance et l'évolution des dimensions à long terme est un processus délicat. Dans une approche multidimensionnelle, vos dimensions étant communes à l'ensemble des tables de fait, une simple erreur de maintenance sur une dimension aura un impact sur l'ensemble de votre système décisionnel. Vous marchez donc sur des oeufs...

La bonne nouvelle, c'est qu'il n'existe pas « trente-six » façons de maintenir une dimension. Si vous respectez les quelques conseils prodigués dans la partie « les tables de dimensions » lors de la conception de vos dimensions, vous pourrez mettre en place ces conseils de maintenance sans grande difficulté.

## 3-C-1 - Opérations de maintenance liées aux dimensions

### 3-C-1-A - Ajout

Avant d'ajouter une nouvelle occurrence de dimension, posez-vous les questions suivantes : cette occurrence a-t-elle sa place dans la structure hiérarchique de ma dimension ? Les catégories parentes de celle-ci sont-elles réellement adéquates ? Un chargement rétroactif de certaines tables de fait est-il nécessaire ?

L'ajout d'une feuille dans un arbre de dimension est rarement problématique, il faut juste veiller à conserver une cohérence hiérarchique et logique dans toute la hauteur de l'arbre.

### 3-C-1-B - Suppression

Votre arbre de dimension évoluant, les utilisateurs ont maintenant accès à des valeurs d'occurrence obsolètes. Vous souhaitez supprimer l'accès à ses occurrences, pour cela deux méthodes possibles :

- première méthode (conseillée) : mettez en place un champ « visibilité de l'utilisateur » dans vos tables de dimensions et changez la valeur de ce champs lorsque l'occurrence est obsolète ;
  - avantage : méthode simple à mettre en oeuvre, aucune opération de mise à jour des tables de fait n'est nécessaire,
  - inconvénient : des valeurs de dimension étant masquées à l'utilisateur et les tables de fait n'ayant subit aucune modification, il se peut que l'agrégat d'un rapport puisse ne pas représenter la somme des valeurs après un « drill-down » sur la dimension ;
- seconde méthode : vous décidez de supprimer l'occurrence de votre table de dimension. Dès lors, l'ensemble des occurrences dans vos tables de fait qui font référence à cette valeur de dimension seront obsolètes. Vous devez alors mettre à jour l'ensemble des tables de fait, en remplaçant la référence de la dimension supprimée par une référence de dimension « valeur non applicable » que vous aurez créée au préalable dans votre dimension.

Quelle méthode utiliser ?

Si l'occurrence de dimension à supprimer n'est plus référencée par vos tables de fait depuis longtemps (exemple, une référence de produits qui n'existe plus depuis plusieurs mois), la méthode via un champ « flag » semble être la plus pertinente.

### 3-C-1-C - Mise à jour

La modification d'une dimension est l'opération la plus délicate. Comme pour la suppression nous pouvons distinguer deux méthodes :

- première méthode : vous remplacez simplement l'ancienne valeur par la nouvelle ;
  - avantage : méthode est simple à mettre en oeuvre et il n'y a aucun processus de traçage d'évolution à mettre en oeuvre,
  - inconvénient : il y a un impact sur l'ensemble des rapports programmés avant la modification ;
- seconde méthode (conseillée) : en cas d'évolution, vous ajoutez la nouvelle valeur sans supprimer l'ancienne que vous marquez comme « non applicable » à partir d'un champ « flag » prévu à cet effet. Ainsi lors du processus d'alimentation de la table de fait, c'est bien la nouvelle valeur qui est référencée et non l'ancienne. Ici le nombre d'occurrences de la table de dimension augmente à chaque modification.
  - avantage : aucun rapport n'est impacté,
  - inconvénient : le processus de maintenance est un peu plus complexe.

Les opérations d'ajout, de suppression et de mise à jour de vos dimensions sont moins triviales qu'il ni paraît. J'espère que cette courte présentation vous permettra de cerner rapidement la problématique de maintenance des dimensions dans un entrepôt de données multidimensionnel.

### 3-C-2 - Opérations de maintenance liées aux mesures

Hormis les tables de fait de type récapitulatif, les seules opérations effectuées sur vos tables de fait seront des insertions (sauf erreur de chargement...). Toutefois, vous pourrez rencontrer des difficultés de maintenance liée à vos tables de fait, celles-ci peuvent provenir :

#### 3-C-2-A - Du volume de données traitées :

Une granularité très fine dans une table de faits peut amener à avoir à traiter des millions d'occurrences. Les gros volumes sont toujours délicats à traiter, si l'alimentation de cette table est en dépendance avec d'autres tables (des bases métiers par exemple) c'est l'ensemble de vos traitements de nuit qui seront retardés... (Une nuit de traitement à une durée maximum de 12 heures). Vous devez absolument garder un oeil sur le temps de traitement de chacune de vos tables et sur les différentes dépendances de votre nuit batch.

#### 3-C-2-B - De la complexité du calcul des mesures :

Des calculs de mesure complexes peuvent devenir problématiques, exemple : les calculs de délais ou les calculs qui doivent tenir compte de l'état d'autres occurrences de tables opérationnelles. Hélas, ici, les compromis avec les analystes ne seront probablement pas possibles. Une définition de mesure ne doit pas être changée pour cause de problème technique de mise en oeuvre. Essayez de mettre en place des vues des tables opérationnelle réduites et divisez la complexité du calcul en différentes étapes plus simples : « break it and make it easier ».

#### 3-C-2-C - Des jointures avec les tables de dimensions :

Pour déterminer le contexte des occurrences de vos tables de fait, vous allez devoir d'une manière ou d'une autre, déterminer quels sont vos clés de dimension pour chaque occurrence traitée. Or, il arrive que pour diverses raisons (évolution des tables opérationnelles ; erreur de chargement de vos dimensions ; problème de qualité des données) que vos lignes de mesure ne trouvent pas de correspondance dans vos dimensions. Ici attention, il ne faut surtout pas perdre ces enregistrements ! Ces occurrences existent, elles doivent donc être prises en compte dans le calcul de votre mesure. Pour cela, je vous conseille de mettre en place une catégorie « Non Applicable » dans chacune de vos dimensions. Si votre enregistrement ne trouve pas de correspondance, il doit tomber par défaut dans cette catégorie « Non Applicable ». Vous serez ensuite apte à isoler les enregistrements problématiques pour les corriger.

D'après ma propre expérience, ces trois points sont des problématiques récurrentes liées aux tables de fait. Cette courte présentation vous permettra sans doute de mettre en place un système plus robuste.

## 4 - Pour conclure

### 4-A - Rappel et derniers conseils

Ce petit traité autour du concept d'entrepôt de données multidimensionnel a donc mis en avant les différents concepts et points clés sur la mise en place ainsi que sur la maintenance d'un tel entrepôt.

Pour conclure ce témoignage, j'aimerai partager avec le lecteur encore quelques conseils techniques concernant l'implémentation ainsi que les facteurs clés de succès d'un tel projet.

#### 4-A-1 - Derniers conseils techniques

Si vous avez rigoureusement lu l'ensemble de ce livre blanc, vous allez vous rentre compte que je radote un peu.

- Pour vos dimensions, utilisez plutôt des clés physiques qui n'ont pas de sens logique. Cela facilite les opérations de maintenance et force vos utilisateurs à ne pas interroger les tables de fait en direct (sans passer par une dimension).
- Sauvegardez des images de vos dimensions à moment « T » par exemple à travers la mise en place de tables mensuelles (image de la table en fin de mois).
- Utilisez et abusez des « flags » pour connaître l'état et la visibilité des occurrences, ils vous seront souvent très utiles.
- Soyez conscient de l'impact de la modification de feuilles ou de noe?uds de votre arbre de dimension. Une mauvaise gestion des dimensions peut avoir de graves incidences sur l'ensemble de votre marché de l'information.
- Faites vous conseiller par un administrateur de base de données. Il saura vous guider sur les types d'index et le schéma de partitionnement à mettre en place. L'optimisation de votre base de données n'est pas une option.
- « Break it and make it easier » (diviser pour mieux régner)
- « Not everything that can be counted counts, and not everything that counts can be counted. » (Ce qui compte ne peut pas toujours être compté, et ce qui peut être compté ne compte pas forcément).

## 4-B - Les cinq Facteurs clés de succès

### 4-B-1 - Compréhension du métier

Lorsque vous choisissez de mesurer un processus, assurez-vous que vous êtes entouré de personnes qui maîtrisent totalement et en profondeur les « workflow (8) » en jeux.

Avant d'essayer de définir « Quoi mesurer ? » l'équipe en charge doit comprendre et définir « Pourquoi le mesurer ? ». Plus l'équipe en charge maîtrise le métier, plus vos mesures seront justes et pertinentes.

### 4-B-2 - Atomicité de la table des faits

Le choix de la granularité des données de votre table des faits est primordial. C'est ce choix de granularité qui détermine votre capacité à répondre aux questions de vos utilisateurs. Le coût de stockage de la donnée diminuant, les organisations ont tendance à descendre au niveau le plus fin disponible, cela semble censé : qui peut le plus peu le moins. Toutefois attention, un volume trop important vous obligera soit à créer des tables de fait périodiques soit à imposer un temps de traitement plus important à vos utilisateurs (sachant que vos collaborateurs perdront vite patience d'avoir à attendre 10 minutes entre chaque clic). Comme toujours, tout est histoire de compromis.

### 4-B-3 - Acceptation par les utilisateurs

Les données que vous mettez à disposition de vos utilisateurs ne doivent pas être remises en cause. Il y a deux principales variables jouant sur l'acceptation par les utilisateurs :

- la véracité des données : « est-ce que les données qu'on me présente sont exactes ? Comment pourrais-je les vérifier ? »
  Pour ne pas que vos utilisateurs puissent remettre en question la véracité des données, vous devez leur fournir un accès à un niveau de détail opérationnel qu'ils ne peuvent pas remettre en cause. Si les rapports sont remis en question pour cause d'erreur de données, c'est tant mieux ! Chaque erreur détectée et corrigée en profondeur (correction de la source initiale de la donnée) doit être considérée comme un pas de plus dans votre processus de la qualité de données.
- la pertinence des indicateurs : « mais qu'elle est l'utilité de cette mesure ? Qu'elle est mon moyen d'action ? ». Construisez votre système de mesure avec des personnes issues du métier. Pour impliquer vos utilisateurs, assurez-vous qu'il existe une variable d'action attaché aux mesures que vous mettez en oeuvre.

Si vos utilisateurs pensent que les données présentées sont mal calculées ou ne sont pas pertinentes par rapport au processus mesuré, c'est l'échec assuré. Ils n'accepteront pas le système de mesure que vous vous efforcez de mettre en place.

### 4-B-4 - Le nombre et la pertinence des dimensions

Rappelez-vous que les dimensions sont communes à l'ensemble des processus mesurés. Elles doivent donc être simples et génériques.

Limitez-vous dans le nombre de dimensions. Selon la complexité du secteur d'activité de votre organisation vous devriez avoir de 3 à 15 dimensions.

Si vous avez plus de 15 dimensions : supprimez-en ! La complexité de votre système vous mènera sans doute à des problèmes de qualité de données récurrents, voir à l'abandon du projet.

Si vous avez moins de trois dimensions : vous ne maîtrisez sans doute pas le concept de dimension et vous ne comprenez sans doute pas la puissance de l'analyse multidimensionnelle.

### 4-B-5 - Définition commune et métadonnées

La définition de vos mesures doit être unique et commune à l'ensemble de l'organisation. L'équipe en charge du système d'information décisionnel doit être la garante de cette unicité. Le processus de gestion des métadonnées doit faire partie intégrante du processus de maintenance et de mise en place des mesures.

### 4-B-6 - Le SPONSOR

Unifier et standardiser à travers un système de mesures transverses a un coût et exige une certaine autorité. Vous devez avoir un SPONSOR de poids. Si vous n'avez pas le soutien direct de la direction du système d'information ou de la direction générale, toute tentative de mise en place d'un système de reporting global sera un échec.

## 5 - Remerciements

Je tiens à remercier l'ensemble de la rédaction de developpez.com pour m'avoir donné la chance de partager ce livre blanc et plus particulièrement **KalyParker**, **jacques_jean** Pierre-André G. ainsi qu'Aurélie M. pour leurs disponibilités et leurs conseils avisés.

1 : balanced scorecard : tableau de bord équilibré permettant d'analyser de façon globale l'activité d'une organisation. Généralement composé de quatre perspectives (finance, client, processus clés, innovation et développement) regroupant des objectifs (eux-mêmes composés d'indicateurs clés) reliés les uns aux autres, formant ainsi une carte stratégique. Le but étant d'exprimer les corrélations existantes entre chaque objectif.
2 : Bus : ensemble de plusieurs fils conducteurs ou de circuits, disposés en lignes parallèles, reliant différents blocs fonctionnels et composants.
3 : Étapes déterminées par Ralph Kimball et Margy Ross dans le livre « Entrepôts de données Guide pratique de modélisation dimensionnelle »
4 : DBA : Database Aministrator (Administrateur de base de données)
5 : Typologie déterminée par Ralph Kimball et Margy Ross dans le livre « Entrepôts de données Guide pratique de modélisation dimensionnelle »
6 : Selon Ralph Kimball (gourou de la modélisation dimensionnelle).
7 : ODS : « Operational Data Store » est une base de données conçue pour centraliser les données issues de sources hétérogènes afin de faciliter les opérations d'analyse et de reporting (source : wikipedia).
8 : Workflow : un workflow (anglicisme) est la représentation d'une suite de tâches ou opérations effectuées par une personne, un groupe de personnes, un organisme, etc. Le terme flow renvoie au passage du produit, du document, de l'information, etc. d'une étape à  l'autre. (source : wikipedia)