ENTREPÔTS DE DONNÉES

Guide pratique de modélisation dimensionnelle

2e édition

Ralph Kimball et Margy Ross

Traduction de Claude Raimond

Chapitre 3

Stocks

Nous avons développé au chapitre 2 un modèle dimensionnel pour les transactions de vente d'une grande entreprise de distribution alimentaire. Le présent chapitre reste dans la même branche d'activité mais remonte d'un niveau dans la chaîne de valeur pour traiter le processus de gestion des stocks. Les concepts développés dans ce chapitre s'appliquent à une grande variété de cas de gestion des stocks, dans la distribution comme dans d'autres secteurs d'activité.

Plus important encore, ce chapitre contient une présentation détaillée de l'architecture de bus des entrepôts de données. Cette architecture de bus est indispensable à la création d'un entrepôt de données intégré recouvrant un ensemble de processus d'entreprise apparentés. Elle fournit un cadre pour la conception générale de l'entrepôt, même s'il est construit de façon incrémentale. Enfin, nous soulignerons à quel point il importe d'utiliser des dimensions et des faits communs et conformes, dans tous les modèles dimensionnels de l'entrepôt.

Le chapitre 3 traite des concepts suivants :

- *les implications de la chaîne de valeur ;*
- *modèle d'instantané périodique du stock, modèles d'instantané de transaction et d'instantané récapitulatif ;*
- *faits semi-additifs ;*
- *faits de stock améliorés ;*
- *architecture de bus et matrice d'entrepôt de données ;*
- *dimensions et faits conformes.*

3.1 Présentation de la chaîne de valeur

La plupart des organisations ont une chaîne de valeur sous-jacente constituée de leurs principaux processus d'entreprise. La chaîne de valeur correspond au flux logique des activités principales d'une organisation. Une société de distribution adresse par exemple une commande à un fabricant. Les produits sont livrés à un entrepôt du distributeur, qui les stocke. Une livraison est faite ensuite à un magasin particulier où les produits sont à nouveau stockés jusqu'à leur achat par les clients. Ce sous-ensemble de la chaîne de valeur d'un distributeur est présenté à la figure 3.1. À l'évidence, des produits provenant d'un fabricant qui livre directement au magasin de détail court-circuiteraient les étapes d'entreposage de la chaîne de valeur.

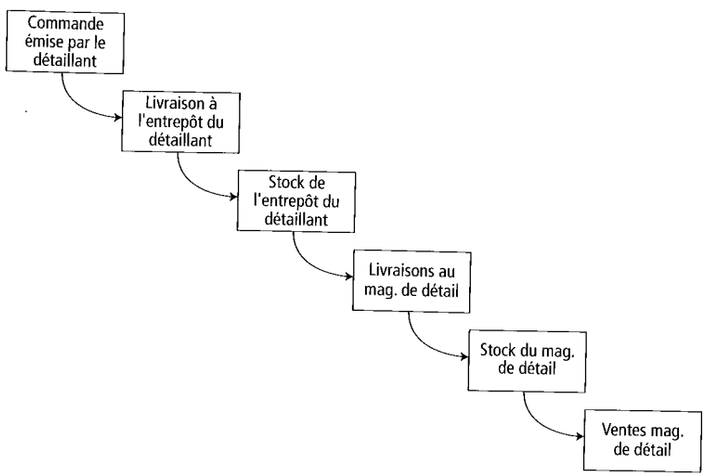

Figure 3.1 **Sous-ensemble d'une chaîne de valeur de distribution**

Les systèmes sources opérationnels fournissent généralement des transactions ou des instantanés à chaque étape de la chaîne de valeur, comportant des mesures de performance. L'objectif principal de la plupart des systèmes d'aide à la décision est de mesurer les performances de certains processus clés. Chaque processus d'entreprise fournissant ses propres mesures à des intervalles spécifiques à un niveau de granularité et avec des dimensions spécifiques, donne naissance généralement à une ou plusieurs tables de faits. C'est pourquoi le concept de chaîne de valeur est indispensable à la compréhension d'un entrepôt de données global d'entreprise. Nous l'évoquerons de nouveau plus loin dans ce chapitre.

3.2 Modèles de stock

Auparavant, examinons divers modèles de stock complémentaires. Le premier est l'instantané périodique de stock. Chaque jour (ou à un autre intervalle de temps), nous mesurons les niveaux de stock de chaque produit et les plaçons sur des lignes distinctes d'une table de faits. Ces lignes d'instantané périodique se présentent au bout d'un certain temps comme des séries de couches de données du modèle dimensionnel, un peu comme les strates géologiques représentent l'accumulation de sédiments sur de longues périodes de temps. Nous examinerons d'assez près ce modèle de stock très courant. Nous verrons également un deuxième modèle de stock où nous enregistrons toute transaction modifiant les niveaux de stock à l'occasion du mouvement des produits à travers les lieux de stockage. Nous verrons enfin dans un troisième modèle l'instantané récapitulatif de stock, où nous insérons une ligne de table de faits pour chaque livraison de produit et où nous tenons cette ligne à jour jusqu'à ce que le produit quitte le lieu de stockage. Chacun de ces trois modèles de stock raconte une histoire différente. Dans certaines applications de stock, deux ou même trois de ces modèles peuvent être utilisés en même temps.

Instantané périodique de stock

Revenons à notre étude de cas de la distribution. L'optimisation des niveaux de stock dans les magasins peut avoir une incidence importante sur la rentabilité de la chaîne de distribution. S'assurer que le bon produit est dans le bon magasin au bon moment minimise les ruptures de stock (cas où le produit est absent de l'étagère de vente) et réduit le coût global de maintien en stock. Le distributeur a besoin d'analyser les niveaux quotidiens des quantités disponibles par produit et par magasin.

Le moment est venu d'utiliser à nouveau le processus de conception d'un modèle dimensionnel en quatre étapes. Le processus d'entreprise que nous voulons analyser est le stock d'un magasin de vente. Pour ce qui est de la granularité, nous voulons voir le stock journalier par produit dans chaque magasin et nous supposons qu'il s'agit du niveau de détail atomique fourni par le système opérationnel de gestion des stocks. Les dimensions découlent immédiatement de cette déclaration du grain : date, produit et magasin. Nous sommes incapables d'envisager d'autres dimensions descriptives à cette granularité. Le stock n'est pas associé à une dimension de promotion. Une promotion peut avoir lieu tandis qu'un produit est en stock, mais la promotion n'est généralement associée au produit que lorsqu'il est vendu. Après une promotion, les produits peuvent toujours être en stock. Typiquement, les dimensions promotion sont associées aux mouvements des produits, comme la commande, la livraison ou la vente.

La vue la plus simple du stock n'utilise qu'un seul fait: la quantité disponible. Ceci conduit à une structure dimensionnelle sans fioriture, telle que celle de la figure 3.2.

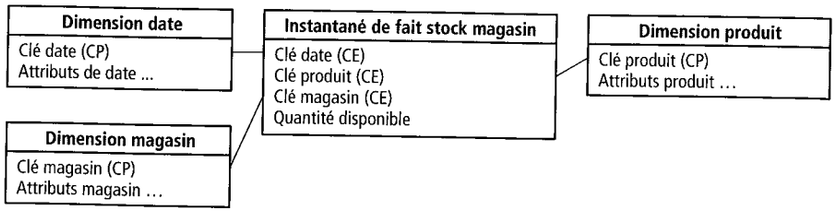

Figure 3.2 **Schéma de l'instantané périodique de stock**

La table de dimension date de cette étude de cas est identique à celle développée pour l'étude de cas précédente des ventes au détail. Les dimensions produit et magasin peuvent également être identiques. Mais nous pouvons vouloir enrichir ces dimensions de plusieurs attributs supplémentaires liés aux analyses du stock. Par exemple, la dimension produit peut inclure des colonnes telles que la quantité minimale pour une commande de réapprovisionnement, en supposant qu'il s'agisse de descripteurs constants et particuliers à chaque unité de stock. De même, dans la dimension magasin, en plus de l'attribut surface de vente rencontré au chapitre 2, nous pourrions inclure des attributs pour identifier les surfaces réfrigérées ou réservées aux produits congelés. Nous verrons plus loin dans ce chapitre ce qu'implique l'ajout de ces attributs.

Si nous analysons les niveaux de stock dans un entrepôt du distributeur et non plus dans un magasin de vente, le schéma obtenu reste très voisin de celui de la figure 3.2. Évidemment la dimension magasin doit être remplacée par une dimension entrepôt. Quand nous surveillons les niveaux de stock de l'entrepôt, nous ne conservons pas la dimension magasin comme quatrième dimension, à moins que le stock de l'entrepôt ne soit affecté à un magasin spécifique.

Même un schéma d'une telle simplicité peut rendre de grands services. S'il y a de nombreux produits et de nombreux entrepôts et si l'inventaire est effectué fréquemment (éventuellement chaque jour), on peut en retirer nombre d'indications très utiles. Si nous analysons les niveaux de stock des magasins d'une chaîne de distribution géante, la base de données correspondante peut servir à rééquilibrer les stocks des magasins chaque nuit après la fermeture.

Cependant, cette table de faits stock présente un problème nouveau par rapport à la table de faits des ventes au détail du chapitre 2. La table de faits vente au détail était raisonnablement éparse, parce que seulement 10 % environ des produits de chacun des magasins étaient vendus chaque jour. Si un produit n'était pas vendu un certain jour, il n'y avait pas d'enregistrement dans la table de faits pour la combinaison de clés correspondante. En revanche, les stocks, par nature, génèrent des

tables de faits denses. Le distributeur s'efforçant d'éviter les manquants, il y a une ligne dans la table de faits pour pratiquement chaque produit, chaque jour et chaque magasin. Nous pouvons inclure des mesures à la valeur zéro sous forme d'enregistrements explicites. Notre distributeur de produits alimentaires avec 60000 produits dans chacun de ses 100 magasins générerait environ 6 millions de lignes (60000 produits × 100 magasins) à chaque chargement de la table de faits. Avec une largeur de ligne de 14 octets, cette table de faits augmenterait de 84 Mo à chaque fois que nous lui rajouterions son contingent quotidien de lignes. En un an, les instantanés quotidiens auraient consommé plus de 30 Go. La densité des bases de données de stock nécessite l'acceptation de certains compromis.

Le compromis le plus évident consiste à espacer les intervalles de temps pour l'historique le plus ancien. Il peut être acceptable de conserver les 60 derniers jours de stock au niveau journalier et de passer au niveau hebdomadaire pour les données historiques. De cette façon, au lieu de 1 095 instantanés sur une durée de trois ans, nous n'aurons que 208 instantanés en tout (60 quotidiens + 148 hebdomadaires dans deux tables de faits distinctes compte tenu de leurs périodicités différentes). Nous réduisons ainsi le volume total des données par un facteur supérieur à 5.

Faits semi-additifs

Au chapitre 2, nous avons souligné l'importance de l'additivité des faits. En modélisant le passage des produits devant les caisses enregistreuses, nous n'avons mesuré que les produits effectivement vendus. Un produit vendu ne peut pas être compté de nouveau dans une vente ultérieure. La plupart des mesures du schéma de vente au détail sont donc parfaitement additives sur toutes les dimensions.

Dans le schéma de l'instantané du stock, on peut cumuler les quantités pour des produits et des magasins différents et obtenir des totaux valides. Mais ces niveaux de stock ne sont pas additifs sur les dates, parce qu'ils représentent des instantanés d'un niveau ou d'un solde à un point dans le temps. Il n'est pas possible de dire si le stock d'hier est le même que celui d'aujourd'hui ou s'il est différent en regardant seulement les niveaux de stock. Les niveaux de stock (et toutes les formes de solde de comptes financiers) sont additifs sur certaines dimensions mais pas sur toutes et c'est pourquoi nous disons que ce sont des *faits semi-additifs*.

La nature semi-additive des faits de solde de stock est encore plus compréhensible si nous pensons aux soldes de notre compte en banque. Supposons que votre solde créditeur soit de 50 euros le lundi. Le mardi, le solde est toujours le même. Le mercredi, vous déposez 50 euros de sorte que le solde passe à 100 euros. Il n'y a pas d'autre activité jusqu'à la fin de la semaine. Le vendredi, vous ne pouvez pas simplement additionner les soldes de la semaine et déclarer que votre solde est passé à 400 euros (50 + 50 + 100 + 100 + 100). La manière la plus utile de combiner les soldes et les niveaux de stock est d'en établir la moyenne (ce qui donne 80 euros dans l'exemple du compte bancaire).

Toutes les mesures enregistrant un niveau à un moment donné (niveaux de stock, soldes de compte et mesures d'intensité telles que des températures) sont intrinsèquement non additives sur les dimensions date et éventuellement sur d'autres dimensions. De telles mesures peuvent être agrégées utilement dans le temps, par exemple, en faisant la moyenne des différentes périodes.

Les derniers mots de ce principe de conception comportent un piège. Vous ne pouvez malheureusement pas vous servir de la fonction SQL AVG pour calculer la moyenne dans le temps. La fonction AVG du SQL fait la moyenne de toutes les lignes reçues par la requête et pas seulement le nombre de dates. Par exemple, si une requête demande le stock moyen pour un groupe de trois produits dans quatre magasins à sept dates différentes (c'est-à-dire, quel est le stock moyen journalier d'une marque dans une région géographique une certaine semaine), la fonction SQL AVG divisera le total des valeurs de stock par 84 (3 produits × 4 magasins × 7 dates). Il faudrait évidemment faire une division par 7, qui est le nombre de périodes journalières. L'absence d'une fonction du genre AVG_DATE_SUM en SQL complique les calculs de stock. Une application de stock bien conçue doit isoler la contrainte date et en obtenir la cardinalité seule (ici, les 7 jours que comporte la semaine de la requête). Ensuite, l'application doit diviser le cumul des valeurs de stock par la cardinalité de la contrainte de date. Ceci peut être réalisé par un appel intégré aux commandes générales du SQL ou en faisant une requête séparée sur la dimension date et en stockant le résultat obtenu dans une application qui le passe ensuite à l'instruction standard SQL.

Faits de stock améliorés

La vue simplifiée des stocks développée dans notre table de faits instantané périodiques est une série temporelle de niveaux successifs de stock. Pour la plupart des analyses de stock, on ne peut se contenter de la quantité disponible. Elle doit être utilisée en combinaison avec d'autres faits permettant de mesurer la rapidité des mouvements et d'autres mesures intéressantes comme la rotation du stock, le nombre de jours d'approvisionnement et la marge brute sur stock.

Si nous ajoutions à chaque ligne de fait de stock la quantité vendue (ou de façon parallèle, la quantité prélevée ou expédiée dans le cas d'un lieu de stockage), nous pourrions calculer la rotation et le nombre de jours d'approvisionnement. Dans le cas d'un instantané de stock journalier, la rotation mesurée chaque jour est la quantité vendue divisée par la quantité en stock. Le nombre de jours de disponibilité résulte d'un calcul similaire. Pour une période de temps plus importante, le nombre de jours d'approvisionnement est la quantité finale en stock divisée par la quantité moyenne vendue durant la période.

Si nous disposons de la quantité livrée, nous pouvons probablement connaître aussi la valeur totale à prix coûtant ainsi que la valeur au dernier prix de vente. La

différence entre ces deux valeurs est, bien sûr, le profit brut. La marge brute est égale au bénéfice brut divisé par le dernier prix de vente.

Enfin, nous pouvons multiplier la rotation par la marge brute pour obtenir le retour de marge brute sur stock, qui peut être exprimé par la formule:

RMBSS = quantité totale livrée × (valeur au dernier prix de vente — valeur à prix coûtant)
÷ (quantité moyenne journalière en stock × valeur au dernier prix de vente)

Bien que la formule soit compliquée, ce ratio correspond à une idée simple. En multipliant la marge brute par la rotation, nous obtenons une mesure du rendement de notre investissement dans le stock. Un ratio élevé signifie que nous faisons circuler le produit dans le magasin rapidement (nombreux tours) et que nous gagnons beaucoup d'argent sur la vente du produit (marge brute élevée). Un ratio faible veut dire que le produit circule lentement (peu de tours) et que nous ne gagnons pas beaucoup d'argent sur ce produit (marge brute faible). Le retour de marge brute sur stock est une mesure standard efficace utilisée par les analystes pour juger la qualité de l'investissement d'une entreprise dans ses stocks.

Pour aller au-delà du schéma initial de la figure 3.2, nous devons inclure dans notre table de faits instantané de stock la quantité livrée, la valeur au coût et la valeur au dernier prix de vente. C'est ce que nous appelons l'instantané de stock étendu (figure 3.3).

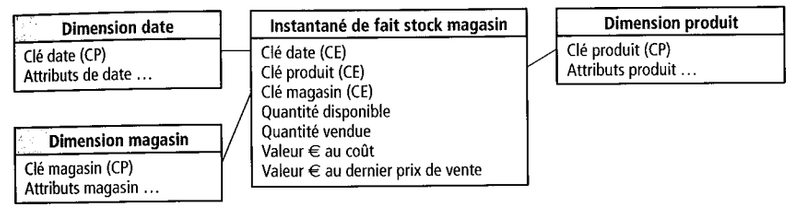

Figure 3.3 **Instantané de stock étendu pour supporter l'analyse de retour de marge brute sur stock**

Notez que si la quantité en stock est semi-additive, les autres mesures de l'instantané étendu sont pleinement additives sur les trois dimensions. La quantité vendue est au grain retenu pour la table de faits et que nous supposons dans ce cas être journalier. Les colonnes de montants contiennent des totaux eux-mêmes additifs. Nous ne plaçons pas le retour de marge brute sur stock dans la table de faits parce qu'il n'est pas additif. Nous pouvons le calculer à partir de ses constituants pour un nombre quelconque de lignes de faits en additionnant les colonnes voulues avant d'effectuer le calcul, mais il ne sert à rien de conserver ce ratio qui ne se prête lui-même à aucune combinaison sur plusieurs lignes de faits.

L'instantané périodique est le schéma de stock le plus courant. Nous évoquerons brièvement deux autres possibilités pour compléter l'instantané périodique. Nous

allons changer un peu de décor et au lieu de décrire ces autres modèles pour le stock des magasins, nous remontons d'un cran dans la chaîne de valeur pour traiter le stock situé dans les entrepôts.

Transactions de stock

Une deuxième façon de modéliser le processus d'entreprise du stock consiste à enregistrer chacune des transactions affectant le stock. Voici une liste de transactions ayant une incidence sur le stock d'un entrepôt :

- réception du produit ;
- mise du produit à l'inspection ;
- sortie d'inspection ;
- retour au fournisseur pour défaut à l'inspection ;
- placement du produit en casier de stock ;
- autorisation du produit à la vente ;
- prélèvement du produit dans un casier de stockage ;
- emballage du produit pour expédition ;
- envoi de produit à un client ;
- réception du produit retourné par un client ;
- remise d'un produit en stock suite à retour client ;
- suppression d'un produit en stock.

Chaque transaction de stock identifie la date, le produit, l'entrepôt, le fournisseur, le type de transaction et dans la plupart des cas, un montant unique représentant l'incidence de la transaction sur le stock. Si le grain de notre table de faits est une ligne par transaction de stock, elle s'inscrit dans un schéma tel que celui de la figure 3.4.

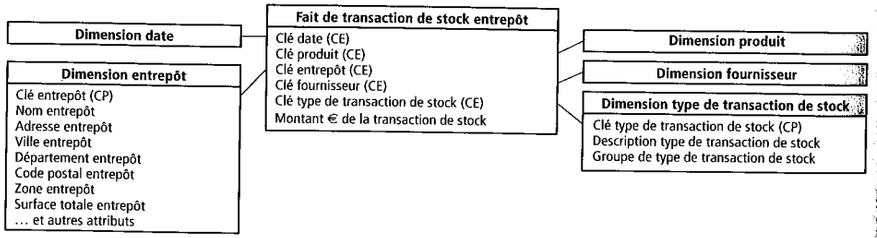

Figure 3.4 **Modèle de transaction de stock d'entrepôt**

Bien que la table de faits au niveau de la transaction soit elle aussi très simple, elle contient la plupart des informations détaillées relatives au stock, car elle reflète des manipulations de stock à une échelle très fine. Elle est utile pour mesurer la fréquence et le timing de types de transaction spécifiques. Par exemple, seule une

table de faits de stock au grain de la transaction permet de répondre aux questions suivantes :

- combien de fois avons-nous mis un produit dans un casier de stockage un jour donné et prélevé le produit du même casier le même jour ?
- combien de livraisons distinctes avons-nous reçues d'un fournisseur donné et quand les avons-nous reçues ?
- sur quels produits avons-nous eu des rejets à l'inspection répétés, provoquant le retour du produit au fournisseur ?

Malgré tout, il est malcommode d'utiliser cette table comme unique moyen d'analyse de la performance des stocks. Bien qu'il soit théoriquement possible de reconstruire la position de stock exacte à tout moment en décomptant toutes les transactions possibles à partir d'une position connue, une telle approche est trop pesante et compliquée pour des questions générales mettant en jeu des dates ou des produits.

> **Souvenez-vous que les transactions ne suffisent pas pour faire toutes les analyses. Une forme ou une autre de table d'instantané, donnant une vue plus cumulative d'un processus accompagne le plus souvent une table de faits transaction.**

Instantané récapitulatif de stock

Pour en terminer avec le modèle de stock, examinons brièvement l'instantané récapitulatif de stock. Dans ce modèle, la table de faits comporte une ligne par expédition de produit parvenue à l'entrepôt. Dans une même ligne de table de faits nous suivons l'évolution du contenu de cette expédition jusqu'à ce qu'il ait quitté l'entrepôt. Ce modèle d'instantané n'est possible que si nous pouvons distinguer à coup sûr les produits livrés lors d'une expédition de ceux livrés ultérieurement. Cette approche convient également si nous suivons les mouvements à des niveaux très détaillés tels que des numéros de série de produit ou des numéros de lot.

Supposons que le stock passe par une série d'événements ou d'étapes bien définis à l'intérieur de l'entrepôt, tels la réception, l'inspection, la mise en casier, l'autorisation à la vente, le prélèvement, l'emballage et l'expédition. La raison d'être de la table de faits de l'instantané récapitulatif est de fournir l'état mis à jour d'un lot de produits reçus, au fur et à mesure de son passage par ces différentes étapes. Chaque ligne de fait est mise à jour jusqu'à ce que le produit quitte l'entrepôt. Comme on le voit à la figue 3.5, la table de faits instantané récapitulatif de stock, avec sa multitude de dates et de faits, a un aspect très différent de celui des schémas d'instantanés périodiques ou de transactions.

Les instantanés récapitulatifs sont le troisième type important de table de faits. Leur originalité tient à la fois à leurs multiples clés étrangères basées sur des valeurs de date au début de la liste des clés et au fait que nous reprenons un grand nombre de fois les mêmes lignes de fait pour les modifier. Ce type d'instantanés

étant rarement utilisé pour des processus mettant en jeu des stocks permanents régulièrement renouvelés, nous reportons son examen approfondi au chapitre 5. Notez simplement au passage les quatre mesures non additives à la fin de la table de faits.

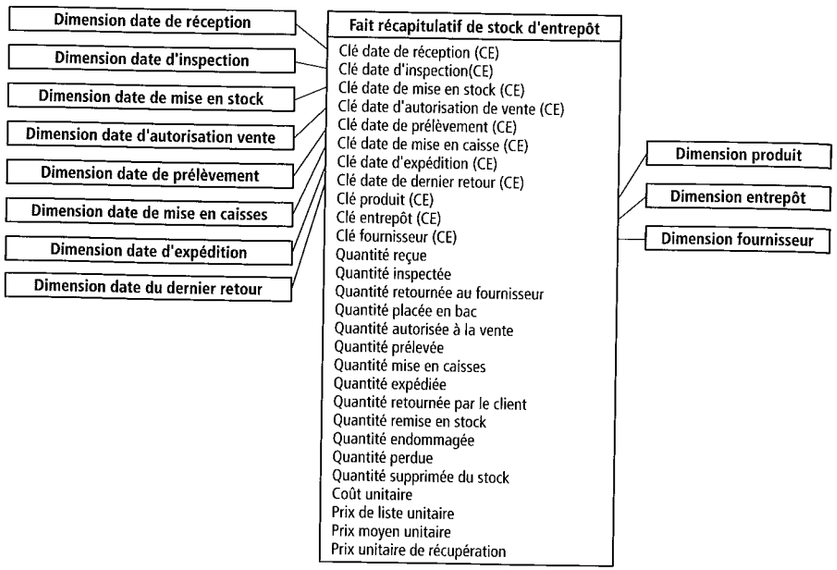

Figure 3.5 **Instantané récapitulatif de stock en entrepôt**

3.3 Intégration de la chaîne de valeur

Ayant terminé la conception des trois variantes de modèle de stock, revenons à la chaîne de valeur du distributeur. Aussi bien les gestionnaires que les informaticiens sont en général très intéressés par l'intégration de la chaîne de valeur. C'est plus particulièrement aux niveaux supérieurs du management que se manifeste le besoin d'examiner l'ensemble de l'activité pour mieux évaluer la performance. De nombreux projets récents d'entrepôt de données répondent à la volonté du management de mieux comprendre les relations avec les clients en les suivant d'un bout à l'autre de la chaîne de valeur. Il faut pour cela disposer d'informations cohérentes relatives aux clients dans les différents processus d'entreprise, qu'il s'agisse des propositions, des commandes, de la facturation, des règlements ou du service aux clients. Même si la vision de votre management n'est pas aussi élevée, prenez en compte l'irritation des gestionnaires qui reçoivent des informations contradictoires de la part de différentes applications ou équipes.

Les responsables informatiques savent parfaitement que l'intégration est indispensable pour que l'entrepôt de données puisse tenir ses promesses. Beaucoup se sentent responsables de gérer au mieux les richesses informationnelles de l'entreprise. Ils savent qu'ils n'assument pas leurs responsabilités s'ils permettent la prolifération de bases de données indépendantes, non intégrées. L'intégration permet non seulement de mieux répondre aux besoins de l'entreprise, elle permet aussi de mieux tirer parti de ressources limitées et de gagner en efficacité par la réutilisation de certains composants.

Heureusement, les personnes pour qui l'intégration présente généralement le plus d'intérêt ont aussi la volonté politique et le pouvoir économique nécessaire à sa réalisation. Si ces personnes n'attachent pas une grande importance à l'intégration, vous êtes confronté à un défi bien plus considérable que la construction d'un entrepôt de données. L'obtention d'un consensus au sein de l'organisation en faveur d'une architecture d'entrepôt de données intégrée couvrant toute la chaîne de valeur ne doit pas être la seule tâche du responsable de l'entrepôt de données. Il a besoin de l'appui politique des plus hauts responsables pour accomplir sa mission sans rencontrer des obstacles dont l'élimination n'est pas de son ressort.

Aux chapitres 1 et 2 nous avons modélisé des données relatives à différents processus d'entreprise de la chaîne de valeur. Bien que les données de chacun d'eux figurent dans des marchés d'infos qui lui sont propres, les modèles ont plusieurs dimensions en commun, les dimensions date, produit et magasin. Cette utilisation commune des mêmes dimensions est représentée à la figure 3.6. Elle est indispensable à la réalisation de marchés d'infos susceptibles d'être intégrés. Elle nous permet de combiner dans un même état des mesures de performance relatives aux différents processus d'entreprise. Nous utilisons des requêtes SQL parallèles portant sur les marchés d'infos distincts, puis nous faisons des jointures externes sur les résultats de ces requêtes en utilisant un attribut dimensionnel commun. Cette approche, souvent désignée par l'expression anglaise *drill across* (forage transversal), est simple à mettre en œuvre si les attributs des tables de dimension sont identiques.

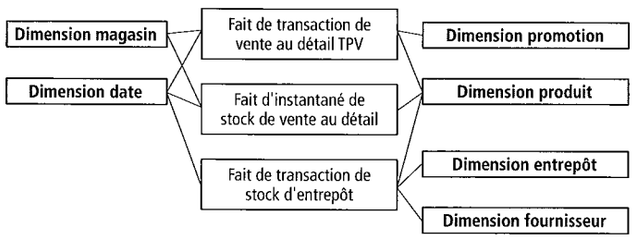

Figure 3.6 **Partage de dimensions entre différents processus d'entreprise**

3.4 Architecture de bus de l'entrepôt de données

De toute évidence, la construction d'un entrepôt de données d'entreprise en une seule étape est une tâche quasi insurmontable et par ailleurs la construction de

morceaux indépendants ne permet pas d'atteindre l'objectif primordial de cohérence. C'est pourquoi l'utilisation d'une approche incrémentale, dans le cadre d'une architecture commune, est la condition *sine qua non* de la réussite durable d'un entrepôt de données. Nous appelons cette approche l'architecture de bus de l'entrepôt de données.

Le mot *bus* est un terme utilisé depuis longtemps en anglais dans la distribution d'électricité et qui sert maintenant couramment en informatique. Un bus est une structure commune où on peut connecter n'importe quoi et où n'importe quoi peut tirer de la puissance. Le bus de votre ordinateur est conforme à une spécification d'interface standard vous permettant de connecter un lecteur de disques, un cédérom ou un nombre quelconque de cartes ou d'appareils différents. Grâce au standard du bus de l'ordinateur, les périphériques fonctionnent ensemble et peuvent coexister de manière productive, bien qu'ils aient été produits à des moments différents par des fournisseurs différents.

> **Grâce à la définition d'une interface de bus standard pour l'environnement de l'entrepôt de données, des marchés d'infos distincts peuvent être réalisés par des groupes différents à des moments différents. Ces marchés d'infos peuvent être interconnectés et utilisés conjointement s'ils adhèrent au standard.**

En nous remémorant le diagramme de la chaîne de valeur de la figure 3.1, nous pouvons imaginer de nombreux processus d'entreprise raccordés au bus de l'entrepôt de données de la manière suggérée à la figure 3.7. À terme, tous les processus de la chaîne de valeur d'une organisation seront représentés par des modèles dimensionnels partageant un ensemble complet de dimensions communes et conformes. Nous verrons de plus près ce que sont des dimensions conformes dans la suite de ce chapitre, mais vous pouvez dès maintenant supposer que ce terme veut dire *similaires*.

L'architecture de bus permet de décomposer le planning de réalisation d'un entrepôt de données d'entreprise selon une approche rationnelle. Au cours d'une phase de durée limitée, la phase de définition de l'architecture, l'équipe conçoit une suite principale de dimensions et de faits standardisés qui doivent être interprétés de la même manière dans toute l'entreprise. C'est ce qui fonde le cadre de l'architecture. On peut ensuite aborder la réalisation de marchés d'infos distincts, chacun d'eux respectant les normes architecturales. Une fois opérationnels, ces marchés d'infos s'assemblent à la manière d'un puzzle. À un certain stade, il existe suffisamment de marchés d'infos pour que le projet d'un entrepôt de données intégré couvrant toute l'entreprise soit devenue réalité.

L'architecture de bus permet aux responsables d'entrepôt de données de gagner sur deux tableaux. Ils disposent d'un cadre architectural servant de guide à la conception d'ensemble, mais le problème a été divisé en marchés d'infos susceptibles

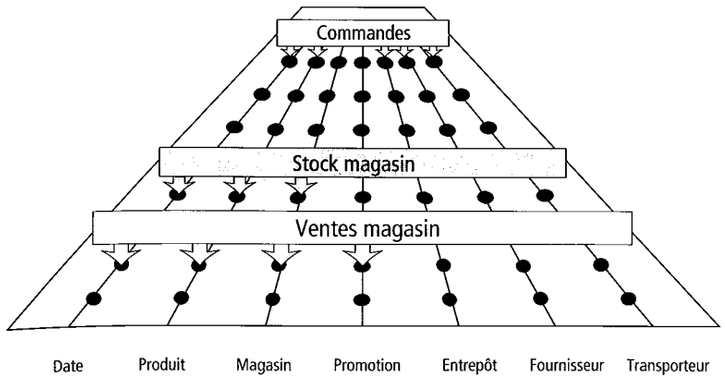

Figure 3.7 **Des dimensions communes sur toute la chaîne de valeur**

d'être réalisés dans des délais réalistes. Les équipes de développement des différents marchés d'infos peuvent suivre les lignes directrices de l'architecture tout en travaillant de manière relativement indépendante et asynchrone.

L'architecture de bus est indépendante de la technologie et de la plate-forme matérielle utilisées. Toutes les sortes de marchés d'infos basées sur l'approche relationnelle ou l'approche OLAP (*Online Analytical Processing*) peuvent participer sans réserve à l'architecture de bus si elles ont été conçues autour de dimensions et de faits conformes. Les entrepôts de données comporteront inévitablement à terme de nombreuses machines distinctes avec des systèmes d'exploitation et des systèmes de gestion de bases de données différents. S'ils sont développés de manière cohérente, selon une architecture commune de dimensions et de faits conformes, ils pourront être fusionnés en un tout intégré.

Matrice de bus de l'entrepôt de données

L'outil que nous utilisons pour créer, documenter et diffuser l'architecture de bus est la matrice de bus d'entrepôt de données, représentée à la figure 3.8. Nous représentons les processus d'entreprise de l'organisation sous forme de lignes d'un tableau. Souvenez-vous que nous distinguons les processus d'entreprise en les identifiant étroitement aux sources de données et non aux différents services ou fonctions. Les lignes de la matrice correspondent aux marchés d'infos relatifs aux activités de base de l'organisation. Nous commençons par lister les marchés d'infos découlant d'un unique système source de base et que nous appelons *marchés d'infos de premier niveau*. Ces marchés d'infos sont le prolongement d'une source opérationnelle à laquelle ils sont visiblement liés.

PROCESSUS D'ENTREPRISE	Date	Produit	Magasin	Promotion	Entrepôt	Fournisseur	Contrat	Transporteur
	DIMENSIONS COMMUNES							
Ventes au détail	X	X	X	X				
Stock vente détail	X	X	X					
Livraisons pour vente détail	X	X	X					
Stock entrepôt	X	X			X	X		
Livraisons entrepôt	X	X			X	X		
Commandes	X	X			X	X	X	X

Figure 3.8 **Exemple de matrice de bus d'entrepôt de données.**

Les lignes de la matrice de bus correspondent à des marchés d'infos. Vous devez créer des lignes de matrice distinctes si les sources sont différentes, si les processus sont différents ou si le travail de réalisation d'un même processus lié à une même source demande un effort trop important pour être mené à bien en une seule fois.

Le moment venu, nous vous recommandons de commencer la réalisation par les marchés d'infos de premier niveau, imposant des efforts de réalisation plus modestes. La plupart des échecs résultent de l'ampleur des efforts de conception et de développement associés aux étapes ETC (extraction, transformation, chargement) de la préparation des données. Dans de nombreux cas, les marchés d'infos de premier niveau fournissent des données suffisamment intéressantes pour faire le bonheur des utilisateurs, lesquels vous laisseront résoudre dans le calme les problèmes plus ardus.

Ayant terminé l'énumération des marchés d'infos de premier niveau, nous pouvons ensuite identifier dans une deuxième phase des marchés d'infos plus complexes puisant à plusieurs sources. Nous appelons ces marchés d'infos des *marchés d'infos consolidés* parce qu'ils recouvrent généralement des processus d'entreprise distincts. Les marchés d'infos consolidés présentent un intérêt immense pour l'organisation, mais ils sont plus difficiles à réaliser parce que l'effort de développement au niveau ETC (extraction, transformation, chargement) croît dangereusement à chaque source de données supplémentaire intégrée à un même modèle dimensionnel. Il est prudent de se concentrer d'abord sur les composants que représentent les marchés d'infos de premier niveau avant d'aborder la consolidation. Dans certains cas, le marché d'infos consolidé est en fait beaucoup plus que la simple réunion d'ensembles de données des marchés d'infos de premier niveau.

La rentabilité est un exemple classique de marché d'infos consolidé. Des éléments de chiffre d'affaires et de coût provenant de différents marchés d'infos sont combinés pour fournir une vue complète de la rentabilité. Il est très motivant d'envisager une approche de la rentabilité à un grain très fin, permettant d'apercevoir les bénéfices au niveau des produits et des clients, mais ce n'est certainement

pas le premier marché d'infos que vous devez chercher à réaliser. Vous pourriez être submergé en tentant de préparer tous les composants de coût et de chiffre d'affaires. Si vous êtes absolument obligé de vous concentrer d'emblée sur la rentabilité pour votre premier marché d'infos, vous devez alors commencer par allouer les coûts de façon approximative et non essayer de tenir compte de tous les éléments sous-jacents du coût. Cependant, la recherche d'un consensus au sein de l'entreprise sur l'allocation des coûts peut s'avérer difficile et même faire obstacle à la réalisation du projet, compte tenu du caractère sensible de ces allocations, qui peuvent aller jusqu'à mettre en jeu des intérêts financiers personnels. L'une des conditions préalables d'un tel projet, nettement extérieure aux responsabilités de l'équipe concernée, est l'accord de l'entreprise sur les règles d'allocation des coûts. Il est en tout cas préférable d'aborder les difficultés de la rentabilité en ayant derrière soi les premières réussites de l'entrepôt de données.

Les colonnes de la matrice représentent les dimensions utilisées dans toute l'entreprise. Il est souvent utile de dresser une liste complète des dimensions avant de remplir la matrice. Cet exercice sollicite votre créativité et vous conduit à vous demander de quelle manière une dimension donnée peut être associée à un marché d'infos.

Les cases comportant une croix spécifient que la colonne de dimension est liée à la ligne du processus d'entreprise. La matrice obtenue est étonnamment dense. Une ligne permet de voir d'un coup d'œil les dimensions d'un marché d'infos. Mais les colonnes font l'intérêt principal de cette matrice, car elles décrivent l'interaction entre les marchés d'infos et les dimensions communes.

La matrice est aussi très utile comme outil de planning et de communication. Bien qu'il soit relativement facile de remplir les lignes et les colonnes, ce simple exercice aboutit à définir l'architecture d'ensemble de l'entrepôt de données. Nous voyons immédiatement les dimensions qui méritent une attention particulière compte tenu de leur participation à de multiples marchés d'infos. Ces dimensions doivent être traitées en priorité et leur conformité définie avec le plus grand soin.

La matrice sert à la communication au sein des équipes de développement des marchés d'infos, ainsi qu'aux différents échelons de l'organisation. Ce document succinct présente la totalité du plan. Il est bien adapté aux communications avec les décideurs en matière d'informatique et de gestion de l'entreprise.

La matrice de bus de l'entrepôt de données est le document le plus important parmi tous ceux dont la préparation doit précéder toute réalisation d'un entrepôt de données. C'est un document hybride, pour partie outil de conception, pour partie outil de gestion de projet, pour partie outil de communication.

Il va sans dire que la construction d'un marché d'infos ignorant ce cadre commun est inacceptable. Des marchés d'infos indépendants et isolés font plus que priver l'entreprise de possibilités d'analyse. Ils produisent des vues de l'organisation

qui sont inconciliables, contribuent à la propagation d'états non comparables les uns avec les autres et deviennent par leur existence même un obstacle à la réalisation d'un environnement d'entrepôt de données cohérent.

Que faire si votre projet d'entrepôt de données ne commence pas par une page blanche ? Peut-être a-t-on déjà construit plusieurs marchés d'infos sans se tenir à des dimensions conformes. Pouvez-vous récupérer ces réalisations isolées et les convertir à l'architecture de bus ? Pour répondre à cette question, vous devez commencer par une évaluation honnête des marchés d'infos non intégrés. Ceci implique des discussions avec les différentes équipes (y compris les équipes clandestines au sein de l'organisation) pour déterminer les différences entre l'environnement actuel et l'objectif architectural de l'organisation. Une fois ces différences établies, vous devez définir un plan de conversion progressive à l'architecture de l'entreprise. Ce plan doit être solide. Les responsables de l'informatique et de la gestion de l'entreprise doivent bien comprendre le désordre actuel des données, les risques encourus si l'on ne fait rien et l'intérêt que présente la mise en œuvre de votre plan. Ils doivent aussi se rendre compte que la conversion nécessite leur engagement en termes de soutien, de ressources et de financement.

Si un marché d'infos existant repose sur une conception dimensionnelle saine, peut-être suffit-il de faire correspondre une dimension existante à la version standard. La table de dimension d'origine sera reconstruite au moyen d'un système de références croisées. De la même façon, la table de faits devra être retraitée pour remplacer les clés de dimension d'origine par des clés de dimension conformes. Naturellement, si les tables de dimension d'origine et les tables de dimension conformes contiennent des attributs différents, la révision des requêtes existantes devient inévitable. Il est plus courant de rencontrer des marchés d'infos existants affectés de défauts de modélisation qui vont plus loin que l'absence d'adhésion à des dimensions standardisées. Dans certains cas, le marché d'infos indépendant et isolé perdure au-delà de sa durée de vie utile. Un marché d'infos isolé est souvent construit pour un domaine fonctionnel particulier. Lorsque d'autres essaient de s'en servir, ils découvrent généralement que le marché d'infos a été réalisé avec une granularité qui ne convient pas et qu'il lui manque des dimensions nécessaires. L'effort que représente l'adaptation de ces marchés d'infos à l'architecture peut excéder celui d'une construction à partir de zéro. Bien que cela soit difficile à admettre, les marchés d'infos isolés doivent souvent être abandonnés et reconstruits selon l'architecture de bus adoptée.

Dimensions conformes

Ayant vu tout l'intérêt de l'architecture de bus, nous pouvons examiner les dimensions conformes standardisées, qui sont la pierre angulaire du bus de l'entrepôt. Les dimensions conformes sont soit identiques à la dimension la plus granulaire et la plus détaillée, soit des sous-ensembles stricts de celle-ci, au sens mathématique.

Les dimensions conformes utilisent les mêmes clés de dimension, les mêmes noms de colonnes d'attributs, les mêmes définitions d'attributs et les mêmes valeurs d'attributs (ce qui garantit des intitulés d'états et des regroupements d'information cohérents). Les tables de dimension ne sont pas conformes si leurs attributs sont nommés différemment ou contiennent des valeurs différentes. Si une dimension client ou produit n'est pas conforme, les marchés d'infos ne peuvent pas être utilisés conjointement ou pis encore, les tentatives d'utilisation conjointe produiront des résultats invalides.

Il existe plusieurs types de dimensions conformes. Au niveau le plus élémentaire, les dimensions conformes ont exactement la même signification pour toutes les tables de faits auxquelles elles peuvent être jointes. La table de dimension date reliée aux faits de vente est identique à la table de dimension date reliée aux faits de stock. En fait, la dimension conforme peut être la même table physique au sein de la base de données. Toutefois, compte tenu de la complexité de l'environnement technique de l'entrepôt de données, il est plus probable que les dimensions soient dupliquées en synchronie sur de multiples plates-formes de bases de données. En tout cas, les dimensions date des deux marchés d'infos doivent avoir le même nombre de lignes, les mêmes valeurs de clés, les mêmes appellations d'attributs, les mêmes définitions d'attributs et les mêmes valeurs d'attributs. Les informations, leur interprétation et la présentation aux utilisateurs sont les mêmes.

La plupart des dimensions conformes sont définies au niveau de granularité le plus fin possible. Le grain de la dimension client sera naturellement le client individuel. Celui de la dimension produit sera le niveau le plus bas auquel les produits sont suivis par les applications sources. Le grain de la dimension date sera la journée.

On a parfois besoin de dimensions à un niveau de granularité supérieur. Peut-être a-t-on besoin d'une dimension à ce niveau parce que la table de faits représente des faits agrégés à associer à des dimensions agrégées. Ce serait le cas si nous avions un instantané de stock hebdomadaire en plus de notre instantané journalier. Des faits peuvent aussi être générés simplement par un autre processus d'entreprise à un degré de granularité plus élevé. Un processus d'entreprise tel que les ventes capte les données au niveau atomique du produit, tandis que les prévisions génèrent des données au niveau des marques. Vous ne pourriez pas utiliser une table de dimension unique pour les schémas de ces deux processus d'entreprise, parce que leurs granularités diffèrent. Les dimensions produit et marque seront néanmoins conformes si la table des marques est un strict sous-ensemble de la table produit au niveau atomique. Les attributs communs aux deux tables de dimension, celle qui est détaillée et celle qui est condensée, comme les descriptions des marques et des catégories, doivent être nommés et définis de la même manière pour les deux tables et comporter des valeurs identiques, comme on peut le voir à la figure 3.9.

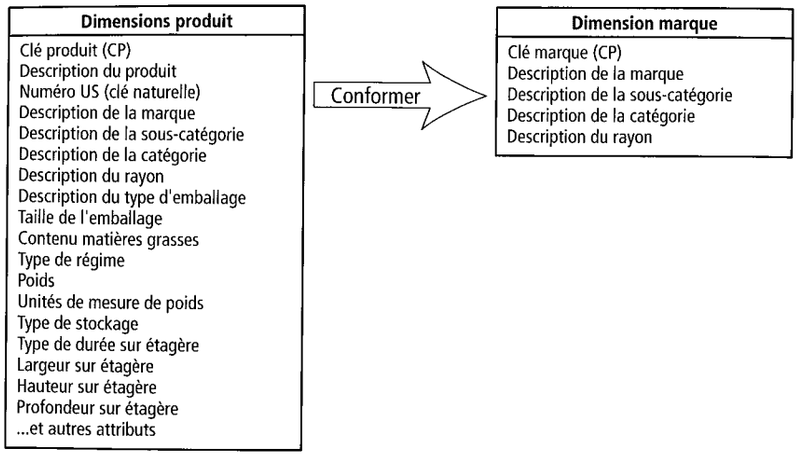

Figure 3.9 **Dimension conforme sous-ensemble d'une autre dimension**

Des dimensions synthétiques sont conformes à une dimension au niveau atomique si elles sont un sous-ensemble strict de cette dimension atomique.

Nous pouvons aussi rencontrer d'autres dimensions conformes dans le cas de tables qui sont des sous-ensembles de tables de dimension au même degré de granularité. Par exemple, dans le schéma de l'instantané de stock nous avons ajouté des attributs supplémentaires aux dimensions produit et magasin qui peuvent être inutiles dans le cas du schéma des transactions de vente. Les tables de dimension utilisées dans ces deux marchés d'infos n'en sont pas moins conformes si les clés et les colonnes communes sont identiques. Évidemment, comme les attributs supplémentaires ne concernent que le marché d'infos du stock, nous ne pourrions pas nous en servir pour des analyses portant sur les deux marchés d'infos.

Un autre cas de sous-ensemble constituant une dimension conforme est celui dans lequel deux dimensions sont au même niveau de détail, mais où l'une d'elle ne comporte qu'un sous-ensemble des lignes de l'autre. Par exemple, nous pouvons avoir une dimension produit au niveau global d'une grande entreprise, contenant les données de tous les produits de divers secteurs d'activité, comme dans le cas de la figure 3.10. Les analystes de ces secteurs individuels peuvent souhaiter ne voir que leur sous-ensemble de la dimension globale, limité aux lignes de leur secteur. L'utilisation d'un sous-ensemble de lignes leur évite d'être encombrés par la totalité des produits de l'entreprise. Naturellement, la table de faits jointe à cette dimension réduite doit être limitée au même sous-ensemble de produits. Si un utilisateur essaie d'utiliser une dimension réduite tout en accédant à une table de faits contenant l'ensemble des produits, il risque d'obtenir des résultats de requête inattendus. Techniquement, l'intégrité référentielle ne serait plus assurée. Nous devons avoir conscience du risque éventuel d'erreur ou de méprise des utilisateurs lorsque

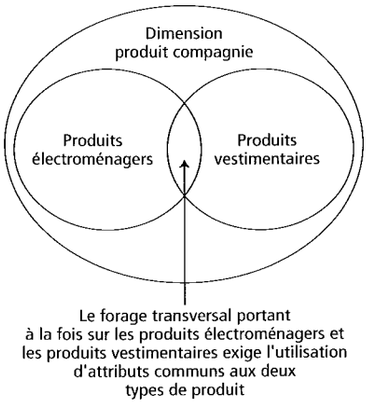

Figure 3.10 **Sous-ensembles de dimensions conformes à une même granularité**

nous créons des dimensions dont les lignes sont un sous-ensemble des lignes de la dimension complète. Nous reviendrons sur les sous-ensembles de dimensions dans le cadre des produits hétérogènes, traités au chapitre 9.

La dimension date conforme utilisée dans le scénario des ventes journalières et des prévisions mensuelles est un exemple unique de sous-ensemble de dimension à la fois pour les colonnes et pour les lignes. Évidemment nous ne pouvons pas utiliser simplement la même table de dimension à cause de la différence de granularité. Toutefois, la dimension mois peut n'être que le sous-ensemble des lignes relatives aux dates de fin de mois, en excluant toutes les colonnes qui ne s'appliquent pas à la granularité mensuelle. Les colonnes exclues comprendraient celles concernant la date du jour comme la description de la date, le numéro de jour dans l'époque, l'indication de jour de semaine ou de week-end, la date de fin de semaine, l'indicateur de jour férié, le numéro de jour dans l'année et d'autres encore. Vous pourriez envisager d'inclure un indicateur de fin de mois à la dimension des dates journalières pour faciliter la création de cette table mensuelle.

Les dimensions conformes doivent être dupliquées logiquement ou physiquement dans toute l'entreprise ; toutefois, elles ne doivent être construites qu'une seule fois dans la zone de préparation. La responsabilité de chaque dimension conforme est assumée par un groupe que nous appelons *autorité de dimension.* L'autorité de dimension est chargée de définir et de tenir à jour une dimension particulière avec ses sous-ensembles éventuels et de la diffuser à tous les marchés d'infos qui en sont les clients. Il lui appartient de préparer les données étalons de la dimension. Elle peut même être amenée à puiser des éléments à de multiples sources opérationnelles en vue de publier une table de dimension complète et vérifiée.

La responsabilité majeure de la fonction centrale appelée autorité de dimension est de préparer, de tenir à jour les dimensions conformes et de les diffuser à tous les marchés d'infos client.

Une fois les dimensions maîtresses définies pour toute l'entreprise, il est de la plus haute importance que toutes les équipes des marchés d'infos les utilisent effectivement. L'engagement à l'égard des dimensions conformes est bien plus qu'une décision d'ordre technique. C'est une décision de politique d'entreprise conditionnant le fonctionnement de l'entrepôt de données. L'accord sur les dimensions conformes se heurte davantage à des obstacles d'ordre politique qu'à des difficultés techniques. C'est pourquoi elles doivent bénéficier d'emblée du soutien des plus hauts niveaux de l'organisation. Les responsables doivent en souligner l'importance auprès de leurs équipes respectives, même si les dimensions conformes imposent certaines contraintes. Le responsable informatique doit obtenir de chaque équipe de marché d'infos l'engagement formel de toujours les utiliser.

Il est évident que les dimensions conformes exigent des efforts de réalisation et de coordination. Les modifications d'attributs existants ou l'addition de nouveaux attributs doivent être passées en revue avec toutes les équipes de marché d'infos utilisant la dimension conforme à modifier. Vous devez aussi prévoir des règles pour la diffusion d'une dimension conforme. Les modifications portant sur des dimensions identiques doivent être appliquées de manière synchrone à tous les marchés d'infos concernés. Seule une publication du type *push* peut garantir l'indispensable cohérence au sein de toute l'organisation.

Ayant suffisamment démontré l'importance des dimensions conformes, voyons des cas où il peut ne pas être réaliste ni nécessaire d'établir des dimensions conformes pour une organisation. Si vous êtes un conglomérat avec des filiales recouvrant des industries diverses, il se peut qu'il n'y ait guère d'intérêt à essayer d'intégrer. Si vous ne voulez pas vendre les produits des différents secteurs d'activité aux mêmes clients, ni vendre des produits qui sont communs à différents secteurs d'activité, ni confier la vente de produits de secteurs d'activité différents à un même commercial, se lancer dans une architecture globale d'entrepôt de données peut ne pas avoir de sens. La perception du gain apporté par les dimensions conformes sera très atténuée. L'obtention d'un consensus pour rechercher une définition commune des produits ou des clients fait figure de test pour une organisation théoriquement désireuse de construire un entrepôt de données global. Si une organisation n'a pas la volonté de s'accorder sur des définitions communes à tous les marchés d'infos, elle ne doit pas essayer de construire un entrepôt de données recouvrant tous ses marchés d'infos. Il vaut mieux construire des entrepôts de données distincts, reflétant les besoins de chaque filiale.

D'après notre expérience, alors que beaucoup d'organisations pensent que combiner les données de leurs secteurs d'activité disparates est une mission impossible, elles envisagent généralement un certain degré d'intégration comme un objectif

lointain. Au lieu de lever les bras au ciel et de déclarer que c'est impossible, nous vous suggérons de vous engager sur la voie de la conformité. Peut-être y a-t-il quelque attributs qui peuvent être rendus conformes à travers différents domaines d'activité. Même s'il ne s'agit que d'un attribut de description de produit, de catégorie ou de secteur d'activité, cette recherche du plus petit commun dénominateur est déjà un pas dans la bonne direction. Rien ne vous oblige à attendre l'accord improbable de toutes les activités sur tout ce qui concerne une dimension pour commencer.

Faits conformes

Nous avons évoqué jusqu'ici la tâche centrale consistant à établir des dimensions conformes pour relier toutes les données de nos marchés d'infos. Ceci représente 90 % de l'effort de mise en place de l'architecture. Les 10 % restant concernent l'établissement de définitions de faits conformes.

Le chiffre d'affaires, le bénéfice, les prix standard, les coûts standard, les mesures de qualité, les mesures de satisfaction des clients et d'autres indicateurs de performance clé sont des faits qui doivent être conformes. En général, les données des tables de faits ne sont pas dupliquées explicitement dans de multiples marchés d'infos. Toutefois, si des faits existent effectivement à plusieurs endroits, tels que des marchés d'infos de premier niveau et des marchés d'infos consolidés, les définitions et les modes de calcul sous-jacents doivent être les mêmes si on les appelle de la même manière. S'ils portent les mêmes noms, ils doivent être définis dans le même contexte dimensionnel et utiliser les mêmes unités de mesure d'un marché d'infos à l'autre.

Nous devons observer une discipline dans l'attribution d'appellations. S'il est impossible de conformer un fait exactement, vous devez alors donner des noms différents aux différentes interprétations. Il y aura alors moins de chance que des faits incompatibles soient utilisés dans un même calcul.

Parfois, un fait a une unité de mesure naturelle dans une table de faits et une autre unité de mesure naturelle dans une autre table de faits. Par exemple, le flux des produits descendant la chaîne de valeur de distribution est mesuré plutôt en nombre de cartons au niveau de l'entrepôt, mais il est mesuré dans le magasin en nombre d'unités scannées. Même si l'on a correctement pris en compte toutes les considérations dimensionnelles, il restera difficile d'utiliser conjointement ces deux unités de mesure incompatibles dans un même état recouvrant des frontières de marché d'infos. Une solution aussi fréquente que discutable à ce genre de problème consiste à renvoyer l'utilisateur à un facteur de conversion caché dans la table de dimension produit, en espérant qu'il trouvera ce facteur de conversion et l'utilisera correctement. Ce procédé est inacceptable tant par le surcroît d'effort qu'il impose à l'utilisateur qu'à cause des risques d'erreur. La solution correcte

consiste à enregistrer le fait dans les deux unités de mesure pour qu'un état puisse glisser le long de la chaîne de valeur en sélectionnant des faits compatibles. Nous reviendrons sur les unités de mesure multiples au chapitre 5.

Résumé

Le stock est un processus qu'il est important de mesurer et de suivre dans de nombreuses industries. Nous avons développé dans ce chapitre des modèles dimensionnels pour les trois vues complémentaires du stock. Les modèles d'instantané périodique ou d'instantané récapitulatif fournissent une bonne description indépendante du stock. L'instantané périodique sera choisi pour des scénarios de stocks permanents, constamment renouvelés. L'instantané récapitulatif convient pour des produits stockés une seule fois et dont le séjour en stock a un début et une fin bien définis. Les applications de stock plus élaborées doivent compléter l'un de ces modèles ou les deux par le modèle des transactions.

Nous avons présenté des concepts clé relatifs à l'architecture et à la matrice de bus de l'entrepôt de données. Tout processus d'entreprise de la chaîne de valeur supporté par un système principal de données sources se traduit par un marché d'infos, ainsi que par une ligne dans la matrice de bus. Les marchés d'infos partagent un nombre important de dimensions standardisées et conformes. Le développement et le respect d'une architecture de bus est la condition *sine qua non* du succès d'un entrepôt de données comportant un ensemble intégré de marchés d'infos.