# ENTREPÔTS DE DONNÉES

## Guide pratique de modélisation dimensionnelle

2e édition

Ralph Kimball et Margy Ross

Traduction de Claude Raimond

Chapitre 1

# Initiation à la modélisation dimensionnelle

Dans ce premier chapitre, nous posons les bases des études de cas qui suivront. Nous commencerons par prendre du recul pour examiner les entrepôts de données d'un point de vue macroscopique. Certains lecteurs seront peut-être déçus d'apprendre qu'il ne s'agit pas uniquement d'outils et de techniques — avant tout et par-dessus tout, l'entrepôt de données doit considérer les besoins de l'entreprise. Nous poserons des jalons pour marquer les objectifs de l'entrepôt de données tout en observant les étranges similitudes entre les responsabilités d'un gestionnaire d'entrepôt de données et celles d'un éditeur. Dans cette perspective globale, nous examinerons les principales composantes du paysage de l'entrepôt de données, y compris le rôle des modèles normalisés. Pour finir, nous définirons les termes fondamentaux de la modélisation dimensionnelle. Notre souhait est qu'à la fin du chapitre vous ayez compris que pour assumer vos responsabilités en matière d'entrepôt de données, vous devez être pour moitié administrateur de bases de données et pour moitié analyste doté de compétences en administration des entreprises.

*Le chapitre 1 traite des concepts suivants :*

- *objectifs de l'entreprise pilotant un entrepôt de données;*
- *la publication dans un entrepôt de données;*
- *principales composantes d'un entrepôt de données global;*
- *importance de la modélisation dimensionnelle pour la présentation des données;*
- *terminologie pour les tables de faits et les tables de dimensions;*
- *mythes autour de la modélisaton dimensionnelle;*
- *quelques pièges à éviter en matière d'entrepôt de données.*

## 1.1 Des univers d'information différents

L'une des plus grandes richesses d'une organisation est son information. Cette richesse est presque toujours conservée sous deux formes : le système opérationnel de stockage des données et l'entrepôt de données. Disons pour simplifier que les données entrent par les applications opérationnelles et que nous les retirons par l'entrepôt de données.

Les utilisateurs d'une application opérationnelle font tourner les rouages de l'organisation. Ils prennent des commandes, signent des contrats avec de nouveaux clients et enregistrent des réclamations. Ils ont presque toujours affaire à un enregistrement à la fois. Ils répètent les mêmes tâches opérationnelles un grand nombre de fois.

Les utilisateurs d'un entrepôt de données, quant à eux, regardent tourner les rouages de l'organisation. Ils comptent les nouvelles commandes et les comparent avec les commandes de la semaine dernière, demandent pourquoi les nouveaux clients ont signé et sur quoi portent les réclamations des clients. Les utilisateurs d'un entrepôt de données n'ont presque jamais affaire à une seule ligne d'information à la fois. Au contraire, leurs questions exigent souvent de chercher des réponses parmi des centaines ou des milliers de lignes et de les condenser dans un unique jeu de réponses. Pour compliquer davantage les choses, les utilisateurs d'un entrepôt de données modifient tout le temps le type de questions qu'ils posent.

Dans la première édition de *Entrepôt de données* (1996), Ralph Kimball avait consacré un chapitre entier à la description des différences entre les deux mondes du traitement opérationnel et des entrepôts de données. On reconnaît aujourd'hui sans discussion que l'entrepôt de données a des besoins, des clients, des structures et des rythmes profondément différents de ceux des applications opérationnelles. Malheureusement, nous rencontrons encore de prétendus entrepôts de données qui ne sont que de simples copies du système d'enregistrement des données opérationnelles, sur une plate-forme matérielle distincte. Cette solution répond à la nécessité de séparer les deux environnements pour des raisons de performance, mais elle n'apporte rien en ce qui concerne les autres différences entre ces deux types de système. Les utilisateurs ne sont pas satisfaits de la commodité d'emploi et des performances de ces pseudo-entrepôts de données. Ces impostures font du tort aux entrepôts de données, en méconnaissant le fait que les utilisateurs des entrepôts de données ont des besoins radicalement différents de ceux des applications opérationnelles.

## 1.2 Objectifs d'un entrepôt de données

Avant de plonger dans les détails de la modélisation et de son implémentation, il convient de s'interroger sur les objectifs fondamentaux de l'entrepôt de données. On peut les comprendre en se promenant dans les couloirs de toute organisation et



en écoutant les gestionnaires d'entreprise. On retrouve inévitablement des thèmes récurrents :

- « Nous avons des montagnes de données dans cette société mais nous ne pouvons pas y accéder. »
- « Nous avons besoin de faire des coupes en tranches et en dés dans les données de toutes sortes de façons. »
- « Il faut faire en sorte que les gestionnaires puissent accéder aux données facilement et directement. »
- « Montrez-moi seulement ce qui est important. »
- « Je suis fou de rage quand deux personnes me présentent les mêmes mesures de performance dans une réunion, mais avec des chiffres différents. »
- « Nous voulons que les gens puissent utiliser les informations pour baser leurs prises de décision sur des faits. »

Nous avons constaté que ces préoccupations sont tellement universelles qu'elles représentent le fondement du cahier des charges d'un entrepôt de données. Transformons ces commentaires de gestionnaires d'entreprise en articles du cahier des charges d'un entrepôt de données.

**L'entrepôt de données doit rendre les données de l'organisation facilement accessibles.** Le contenu de l'entrepôt de données doit être facile à comprendre. Les données doivent être parlantes et leur signification évidente pour l'utilisateur gestionnaire et pas seulement pour le développeur. Le fait qu'elles soient compréhensibles implique qu'elles soient aussi lisibles; le contenu de l'entrepôt de données doit être étiqueté de manière significative. Les utilisateurs gestionnaires veulent séparer et combiner les données de toutes sortes de façons, un processus que l'on nomme découpage *en tranches et en dés*. Les outils d'accès à l'entrepôt de données doivent être simples et faciles à utiliser. Ils doivent aussi renvoyer à l'utilisateur les résultats de ses requêtes avec des temps d'attente minimes.

**L'entrepôt de données doit présenter l'information de l'organisation de manière cohérente.** Les données de l'entrepôt doivent être crédibles. Elles doivent être assemblées à partir de différentes sources de l'organisation et nettoyées. Il faut contrôler leur qualité et ne les publier que lorsqu'elles sont propres à la consommation par les utilisateurs. Les informations d'un processus d'entreprise doivent correspondre à celles d'un autre processus d'entreprise. Si deux mesures de performance portent le même nom, elles doivent vouloir dire la même chose. Inversement, si deux mesures ne veulent pas dire la même chose, elles doivent être appelées différemment. La cohérence implique une qualité élevée. Elle suppose aussi que l'on a tenu compte de toutes les données, qu'elles sont complètes. La cohérence exige, en outre, que les définitions communes du contenu de l'entrepôt de données soient disponibles pour les utilisateurs.

**L'entrepôt de données doit être adaptable et résistant aux changements.** Nous ne pouvons tout simplement pas éviter les changements. Les besoins des utilisateurs, les conditions de l'activité, les données et la technologie sont les uns comme les autres exposés au temps qui passe. Les données de l'entrepôt doivent être conçues pour traiter ces changements inévitables. Les modifications de l'entrepôt de données doivent se faire en douceur, ce qui veut dire qu'elles ne doivent pas invalider les données existantes ou les applications. Les données existantes et les applications ne doivent pas être modifiées ou bouleversées lorsque la communauté des utilisateurs pose de nouvelles questions ou que de nouvelles données sont adjointes à l'entrepôt de données. Si les données descriptives de l'entrepôt de données sont modifiées, nous devons rendre compte convenablement de ces modifications.

**L'entrepôt de données doit être un bastion sûr protégeant notre richesse informationnelle.** Les plus précieuses parmi les informations d'une organisation sont conservées dans l'entrepôt de données. Au minimum, l'entrepôt de données contient le plus souvent des informations sur ce que nous vendons, à qui, à quel prix — des précisions dangereuses si elles tombent en de mauvaises mains. L'entrepôt de données doit efficacement contrôler l'accès aux informations confidentielles de l'organisation.

**L'entrepôt de données doit être le socle sur lequel repose l'amélioration des prises de décision.** Il doit contenir les données servant à étayer les décisions. L'entrepôt de données ne produit qu'une seule sortie : les décisions prises sur la base des réalités qu'il révèle. Ces décisions sont la valeur ajoutée de l'entrepôt de données. L'appellation antérieure à celle d'entrepôt de données est toujours celle qui décrit le mieux ce que nous mettons en place : un système d'aide à la décision.

**L'acceptation de l'entrepôt de données par la communauté des utilisateurs est l'une des conditions de sa réussite.** Cela ne sert à rien de concevoir une solution élégante utilisant les produits et les plates-formes les plus prisés. Si la communauté des utilisateurs n'a pas accueilli l'entrepôt de données à bras ouverts et continué de s'en servir régulièrement six mois après la formation, alors nous avons échoué au test de réception. Contrairement à la nouvelle version d'une application opérationnelle, dont les utilisateurs sont toujours obligés de se servir, l'entrepôt de données peut n'être qu'une option pour ses utilisateurs potentiels. L'acceptation des utilisateurs est avant tout liée à la simplicité d'utilisation.

Comme le montre cette liste, le succès d'un entrepôt de données demande bien plus qu'une excellente maîtrise de la technique et des bases de données. Lors d'un projet d'entrepôt de données, nous avons un pied solidement posé dans les technologies de l'information et l'autre sur le terrain moins familier des gestionnaires. Nous devons être à cheval sur les deux domaines et faire évoluer certains de nos



talents confirmés pour les adapter aux exigences particulières des entrepôts de données. Il est clair que nous avons besoin de compétences variées requises par un comportement hybride administrateur de données/gestionnaire.

## Métaphore de la publication

Ayant à l'esprit les objectifs de l'entrepôt de données, comparons notre rôle de responsable d'un entrepôt de données à celui de rédacteur en chef. Comme le rédacteur en chef d'un magazine de grande qualité, vous disposez d'une grande latitude pour gérer le contenu du magazine, son style et sa distribution. Toute personne avec ce titre et cette fonction aurait presque certainement les activités suivantes :

- Identifier les lecteurs au point de vue démographique.
- Découvrir ce que les lecteurs veulent voir dans cette sorte de magazine.
- Identifier les « meilleurs » lecteurs qui vont renouveler leur abonnement et acheter les produits des annonceurs du magazine.
- Trouver de nouveaux lecteurs potentiels et leur faire prendre conscience de l'existence du magazine.
- Choisir le contenu du magazine pour qu'il convienne le plus possible aux lecteurs visés.
- Prendre des décisions sur la mise en page et l'aspect pour maximiser le plaisir des lecteurs.
- Faire respecter des standards exigeants quant à la qualité de l'écriture et de l'édition, tout en conservant un style de présentation cohérent.
- Vérifier en permanence l'exactitude des articles et celle des affirmations des annonceurs.
- Développer un bon réseau de rédacteurs ou autres sources de copie.
- Attirer des annonceurs et gérer le magazine de façon rentable.
- Publier le magazine régulièrement.
- Entretenir la confiance des lecteurs.
- Faire en sorte que les propriétaires du magazine demeurent pleinement satisfaits.

Nous pouvons aussi identifier des objectifs qui n'ont pas leur place parmi ceux du rédacteur en chef. Ce pourrait être de bâtir le magazine autour des particularités techniques d'un procédé d'impression, de focaliser le management uniquement sur l'efficacité opérationnelle, d'imposer un style de rédaction technique que les lecteurs auront du mal à comprendre ou de créer une mise en page complexe difficile à parcourir et à lire.

En basant votre activité éditoriale sur un objectif de service efficace aux lecteurs, vous augmenterez les chances de réussite du magazine. Vous pouvez à l'inverse examiner la liste et vous demander ce qui arrivera si l'un de ses éléments n'est pas respecté; tôt ou tard, votre magazine aura de sérieuses difficultés.

Cette métaphore établit un parallèle entre l'édition et la direction d'un entrepôt de données. Nous sommes convaincus que la description correcte du travail d'un responsable d'entrepôt de données est *éditeur des bonnes données*. Guidés par les besoins de l'activité, les dirigeants d'entrepôts de données doivent publier des données qui ont été rassemblées à partir de sources multiples et corrigées pour en assurer la qualité et la cohérence. Votre responsabilité principale en tant que dirigeant d'un entrepôt de données est de servir vos lecteurs, autrement dit vos utilisateurs gestionnaires. La métaphore de l'édition souligne la nécessité d'une orientation externe vers les clients de préférence à une orientation interne vers les produits et les processus. Vous utilisez la technologie pour établir l'entrepôt de données, mais la technologie n'est au mieux qu'un moyen au service d'une fin. C'est pourquoi la technologie et les techniques que vous utilisez pour construire votre entrepôt de données ne doivent pas apparaître directement parmi vos principales responsabilités.

Traduisons les responsabilités de l'éditeur d'un magazine en celles du responsable d'un entrepôt de données :

- Comprendre vos utilisateurs en tenant compte de leur domaine d'activité, des responsabilités associées à leur fonction et de leur tolérance à l'informatique.
- Déterminer les décisions que les utilisateurs veulent prendre à l'aide de l'entrepôt de données.
- Identifier les « meilleurs » utilisateurs qui prennent des décisions efficaces, à forte répercussion, grâce à l'entrepôt de données.
- Trouver de nouveaux utilisateurs éventuels et les informer sur l'entrepôt de données.
- Parmi toutes les données possibles de l'organisation, choisir le sous-ensemble le plus pertinent et le plus exploitable pour les présenter dans l'entrepôt de données.
- Mettre en place des interfaces et des applications simples, pilotées par des modèles adaptés aux processus cognitifs des utilisateurs.
- Veiller à ce que les données soient exactes, fiables et modélisées de manière cohérente dans toute l'entreprise.
- Surveiller en permanence l'exactitude des données et du contenu des états livrés aux utilisateurs.
- Rechercher de nouvelles sources de données et adapter constamment l'entrepôt à l'évolution des profils de données, des besoins en information et des priorités.
- Revendiquer le mérite qui vous revient pour les bonnes décisions prises grâce à l'entrepôt de données et utiliser ces succès pour justifier les dépenses de personnel, de logiciel et de matériel.
- Publier les données régulièrement.
- Garder la confiance des utilisateurs.



- Veiller à la satisfaction des utilisateurs, des responsables qui ont décidé l'installation de l'entrepôt de données et de votre chef.

Si vous assumez toutes ces responsabilités, vous serez un excellent dirigeant d'entrepôt de données. Inversement, parcourez cette liste et imaginez ce que peut provoquer l'omission d'un seul de ses éléments. Tôt ou tard, votre entrepôt de données aurait de sérieuses difficultés. Nous vous invitons à comparer la liste avec votre propre définition de fonctions. Notre liste met probablement davantage l'accent sur ce qui concerne les utilisateurs et le fonctionnement de l'organisation et peut même sembler ne pas décrire un rôle d'informaticien. À notre avis, c'est ce qui fait l'intérêt des entrepôts de données.

## 1.3 Composants d'un entrepôt de données

Connaissant désormais les objectifs de l'entrepôt de données, voyons maintenant tout ce qu'il y a dans un entrepôt de données et dans son environnement. Il est souhaitable de bien comprendre tous les éléments séparément avant de commencer à les combiner pour créer un entrepôt de données. Toute confusion entre les fonctions des différents éléments peut compromettre la réussite de l'ensemble.

Comme on peut le voir à la figure 1.1, l'environnement de l'entrepôt de données comporte quatre parties différentes — les applications opérationnelles sources, la préparation des données, la présentation des données et les outils d'accès aux données.

### Les applications opérationnelles sources

Ce sont les applications opérationnelles qui « capturent » les transactions de l'entreprise. On doit se les représenter comme extérieures à l'entrepôt de données, car nous n'avons vraisemblablement guère ou pas du tout de moyens d'influencer le contenu ou le format des données qu'ils traitent. Les principales priorités de ces applications sources sont la performance des traitements et la disponibilité. Les requêtes adressées à ces applications sont étroites, concernent un enregistrement à la fois, font partie d'un flux de transactions volumineux et la charge de travail imposée au système par chaque transaction est sévèrement limitée. Les requêtes des applications sources contrastent avec celles des entrepôts de données, généralement très étendues et imprévisibles. Les applications sources ne conservent que très peu de données historiques et si vous avez un bon entrepôt de données, il peut libérer les applications sources d'une bonne partie de leurs responsabilités concernant la représentation du passé. Chaque application sources est souvent une application verticale naturelle, pour laquelle peu d'investissements ont été consacrés au partage, avec d'autres applications, de données communes de l'organisation telles que les produits, les clients, la géographie ou les calendriers. Dans le meilleur des cas, les applications opérationnelles peuvent être refondues pour offrir une vue

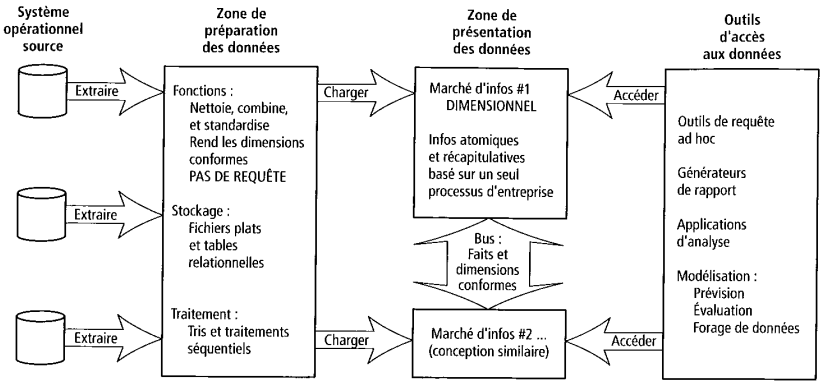


Figure 1.1 **Composants de base de l'entrepôt de données**

cohérente des données. L'intégration des applications au niveau de l'entreprise facilite considérablement le travail de conception d'un entrepôt de données.

## Préparation des données

La partie préparation de l'entrepôt de données est à la fois une zone de stockage et un ensemble de processus couramment appelés ETC, *extraction/transformation/chargement*. La préparation inclut tout ce qu'il y a entre les applications opérationnelles sources et la présentation des données. Elle est comparable à la cuisine d'un restaurant, où des produits nutritifs crus sont transformés en un agréable repas. Dans l'entrepôt de données, des données opérationnelles crues sont transformées en éléments accessibles aux requêtes des utilisateurs qui les consommeront. Comme la cuisine du restaurant, la zone de préparation des données est à l'arrière-plan et n'est accessible qu'à des professionnels qualifiés. Le personnel de cuisine de l'entrepôt de données est occupé par la préparation des repas et ne peut pas en même temps répondre à des demandes d'information de clients. Les clients ne sont pas invités à manger dans la cuisine, où ils ne seraient pas en sécurité. Ne voulant pas que les clients de l'entrepôt de données soient exposés aux équipements dangereux, aux surfaces brûlantes et aux couteaux aiguisés de la cuisine, nous leur interdisons l'entrée dans la zone de préparation des données. D'ailleurs, il se passe dans la cuisine des choses qui ne regardent pas les clients.

> **Un point très important, dans l'aménagement de l'entrepôt de données, est d'interdire aux utilisateurs l'accès à la zone de préparation des données, qui ne fournit *aucun* service de requête ou de présentation.**

L'extraction est la première étape du processus d'apport de données à l'entrepôt de données. Extraire, cela veut dire lire et interpréter les données sources et les copier dans la zone de préparation en vue de manipulations ultérieures.



Une fois dans la zone de préparation, les données peuvent être transformées de nombreuses manières, par exemple nettoyées (correction orthographique, résolution de conflits de domaine, traitement du problème des éléments manquants, conversion à des formats standard), combinées à partir de sources multiples, dédoublées s'il y a lieu et pourvues de clés propres à l'entrepôt de données. Ces transformations sont le prélude au chargement dans la zone de présentation de l'entrepôt de données.

Malheureusement, il y a encore beaucoup de discussions, dans notre industrie, sur la question de savoir si les données utilisées dans ce processus ou celles qui en résultent doivent être mises sous forme de structures physiques normalisées avant leur chargement dans la zone de présentation où elles serviront aux requêtes et à la préparation d'états. Ces structures normalisées sont parfois appelées par la profession *entrepôt de données d'entreprise*; cependant, nous croyons que cette appellation est incorrecte, parce que l'entrepôt de données est beaucoup plus vaste que cet ensemble de tables normalisées. L'entrepôt de données d'entreprise devrait désigner la réunion des zones de préparation et de présentation des données. Par la suite, quand nous mentionnons dans ce livre l'entrepôt de données d'entreprise, nous nous référons à l'union des divers composants de l'entrepôt et pas seulement à la zone de préparation des données.

La zone de préparation des données est surtout le siège de simples opérations de tri et de traitement séquentiel. Dans de nombreux cas, elle n'est pas basée sur la technologie relationnelle et se contente d'un système de fichiers plats. Une fois rendues conformes aux règles définissant les relations de type un-à-un et de type un-à-plusieurs, il peut ne plus y avoir aucun intérêt à passer à l'étape finale de construction d'une base de données physique de troisième forme normale.

Il y a cependant des cas où les données arrivent sur le seuil de la zone de préparation des données dans un format relationnel de troisième forme normale. En pareil cas, il peut être plus commode pour les responsables de la préparation de réaliser les opérations de nettoyage et de transformation au moyen d'un ensemble de structures normalisées. Cette approche est acceptable, mais elle nous suggère cependant certaines réserves. Créer d'une part, des structures normalisées pour la préparation et d'autre part, des structures dimensionnelles pour la présentation, revient à extraire, transformer et charger les données deux fois — une fois dans la base de données normalisée et une nouvelle fois lors du chargement du modèle dimensionnel. Ce processus en deux étapes exige plus de temps et de ressources au niveau du développement, plus de temps lors du chargement ou de la mise à jour des données et une plus grande capacité de stockage pour les copies multiples des données. En fin de compte, le développement, la maintenance et les budgets pour les plates-formes matérielles sont plus importants. Malheureusement, certaines équipes de projet d'entrepôt de données échouent en concentrant toute leur énergie et leurs ressources sur la construction de structures normalisées plutôt que de consacrer du temps au développement d'une zone de présentation contribuant à la

prise de meilleures décisions par l'entreprise. Nous pensons que la cohérence des données au niveau de toute l'entreprise est un objectif fondamental pour l'entrepôt de données, mais il y a des moyens tout aussi efficaces et moins coûteux pour y parvenir que la création physique d'un ensemble normalisé de tables dans la zone de préparation, quand ces structures n'existent pas déjà.

**Il est acceptable de créer une base de données normalisée pour supporter les processus de préparation; toutefois, ce n'est pas l'objectif principal. Les structures normalisées doivent être écartées au niveau des requêtes des utilisateurs car elles sont difficiles à comprendre et entraînent des performances médiocres. Dès lors qu'une base de données supporte des services de requête et de présentation, elle doit être considérée comme faisant partie de la zone de présentation de l'entrepôt de données. Par défaut, les bases de données normalisées sont exclues de la zone de présentation qui doit rigoureusement être structurée selon le modèle dimensionnel.**

Que l'on ait affaire à des fichiers plats ou des structures de données normalisées dans la zone de préparation, l'étape finale du processus ETC est le chargement des données. Cette étape consiste à transmettre aux programmes assurant le chargement en masse des données de chacun des marchés d'infos récepteurs, des tables dimensionnelles dont la qualité a été vérifiée une première fois au cours de la phase d'extraction. Le marché d'infos ciblé doit indexer les données pour assurer la performance des requêtes. Lorsque chaque marché d'infos a été réapprovisionné en données fraîches, qu'il les a indexées, complétées des agrégats appropriés et qu'il les a soumises à d'autres opérations d'assurance qualité, la communauté des utilisateurs est informée de la publication des nouvelles données. La publication doit inclure l'indication de tout changement affectant les dimensions sous-jacentes et de toute nouvelle condition relative aux faits mesurés ou calculés.

## Présentation des données

La zone de présentation des données est le lieu où les données sont organisées, stockées et offertes aux requêtes directes des utilisateurs, aux programmes de reporting et autres applications d'analyse. Alors que l'accès à la zone de préparation est interdit, la zone de présentation est l'entrepôt de données, tel qu'il est perçu par la communauté des utilisateurs. Elle est tout ce que les utilisateurs voient et touchent par le biais des outils d'accès. Le titre provisoire de l'édition précédente de l'ouvrage était *Getting the Data Out*, sortir les données. C'est en fait le rôle de la zone de présentation avec ses modèles dimensionnels.

Nous décrivons habituellement la zone de présentation des données comme une série de marchés d'infos intégrés. Un marché d'infos est une part du gâteau qui recouvre la zone de présentation. Dans le cas le plus simple, un marché d'infos représente les données d'un unique processus d'entreprise. Ces processus d'entreprise peuvent être à cheval sur les frontières de différents services.



Nous avons des idées très arrêtées au sujet de la zone de présentation. En premier lieu, les données doivent être présentées, stockées et consultées sous forme de schémas dimensionnels. Heureusement, la profession a mûri au point que nous n'avons plus à débattre de cette exigence. Elle est parvenue à la conclusion que la modélisation dimensionnelle est la seule technique viable pour fournir des données à des utilisateurs d'entrepôt de données.

La modélisation dimensionnelle est un nouveau nom pour une ancienne technique permettant de rendre les bases de données simples et compréhensibles. À partir des années soixante-dix, les services informatiques, les consultants, les utilisateurs finals et les fournisseurs se sont ralliés à une structure dimensionnelle adaptée au besoin de simplicité fondamental des humains. Imaginez un responsable d'entreprise décrivant son activité en disant «Nous vendons des produits dans différents marchés et nous mesurons nos performances dans le temps». En tant que concepteurs dimensionnels, nous relevons l'insistance du locuteur sur les notions de produit, de marché et de temps. La plupart des gens trouvent naturel de penser à cette activité sous la forme d'un cube de données dont les arêtes s'appellent produit, marché et temps. Nous pouvons imaginer effectuer des coupes en tranches et en dés le long de chacune de ces dimensions. Les points à l'intérieur du cube sont les endroits où sont stockées les mesures pour chaque combinaison de produit, marché et temps. La possibilité de visualiser quelque chose d'aussi abstrait qu'un ensemble de données d'une manière concrète et tangible est la clé de l'intelligibilité. Si cela vous paraît trop simple, tant mieux ! Un modèle de données qui est simple au départ a une chance de rester simple jusqu'à la fin de la conception. Un modèle qui commence par être compliqué sera nécessairement trop compliqué en fin de parcours. Les modèles trop compliqués fonctionnent lentement et sont rejetés par les utilisateurs.

La modélisation dimensionnelle est très différente de la modélisation en troisième forme normale (3NF). La modélisation en 3NF est une technique qui vise à supprimer les redondances dans les données. Les données sont divisées en entités discrètes, chacune devenant une table de la base de données relationnelle. Une base de données de commandes client peut commencer par un enregistrement pour chaque ligne de commande mais se transforme en un diagramme en forme de toile d'araignée étonnamment complexe une fois mise en 3NF, avec éventuellement des centaines ou des milliers de tables normalisées.

La profession appelle quelquefois les modèles 3NF des *modèles ER*. ER est l'acronyme de *entité-relation*. Les diagrammes entité-relation sont des dessins de rectangles et de lignes indiquant les relations entre des tables. Les modèles dimensionnels comme les modèles 3NF peuvent être représentés en diagrammes ER parce les uns comme les autres sont des tables relationnelles jointes; la différence clé entre les deux types est le degré de normalisation. Comme le modèle ER s'applique aussi bien aux deux types de modèles, nous éviterons de désigner les

modèles 3NF par l'appellation modèle ER et nous les appellerons *modèles normalisés*, pour limiter les risques de confusion.

La modélisation normalisée apporte d'immenses gains de performance dans les traitements opérationnels parce qu'elle permet à une transaction de mise à jour ou à une insertion de ne toucher la base de données qu'à un seul endroit. Mais les modèles normalisés sont trop compliqués pour les requêtes d'entrepôt de données. Les utilisateurs ne peuvent pas comprendre, parcourir, ni se rappeler des modèles normalisés dont la complexité s'apparente à l'enchevêtrement des autoroutes d'un centre urbain congestionné. De même, les systèmes de gestion de bases de données relationnelles (SGBDR) ne peuvent pas faire de requêtes efficaces sur un modèle normalisé; sa complexité dépasse les possibilités de ses fonctions d'optimisation, ce qui conduit à des performances désastreuses. L'utilisation de la modélisation normalisée dans un entrepôt de données va à l'encontre du but recherché, qui est de permettre de trouver rapidement des données par une recherche intuitive.

Il existe un syndrome commun à certaines grandes organisations informatiques. C'est une sorte de maladie provoquée par des schémas d'entrepôt de données excessivement complexes. Les symptômes peuvent être :

- un investissement en matériel et en logiciel d'une dizaine de millions d'euros ne permettant de traiter qu'une poignée de requêtes par jour;
- un service informatique dont les grands prêtres doivent écrire toutes les requêtes de l'entrepôt de données;
- des requêtes apparemment simples qui requièrent plusieurs pages de code SQL (*Structured Query Language*);
- un département marketing qui est malheureux parce qu'il ne peut pas accéder directement au système (et ne sait toujours pas si la société est rentable à Amiens);
- un responsable informatique inquiet, fermement décidé à faire quelques changements si les choses ne s'améliorent pas radicalement.

Heureusement, la modélisation dimensionnelle est une solution au problème des schémas excessivement complexes dans la zone de présentation. Un modèle dimensionnel contient les mêmes informations qu'un modèle normalisé, mais il organise les données dans un format dont les objectifs de conception sont l'intelligibilité, la performance des requêtes et l'aptitude au changement.

Le deuxième principe auquel nous tenons concernant les marchés d'infos de la zone de présentation est qu'ils doivent contenir des données atomiques détaillées. L'atomicité est requise pour résister aux assauts des requêtes imprévisibles des utilisateurs, les requêtes *ad hoc*. Les marchés d'infos peuvent contenir des données cumulées dites « agrégats » pour améliorer les performances, mais il ne suffit pas de fournir ces cumuls sans fournir aussi les grains eux-mêmes (les atomes) sous une forme dimensionnelle. En d'autres termes, il n'est pas acceptable de ne stocker que les données cumulées dans des modèles dimensionnels et d'enfermer les données



atomiques dans des modèles normalisés. On ne peut envisager qu'un utilisateur fasse un forage à travers des données dimensionnelles presque jusqu'au niveau des grains et perde les avantages de la présentation dimensionnelle lors de l'étape finale. Nous verrons au chapitre 16 que toute application utilisateur peut descendre sans effort jusqu'aux données granulaires sous-jacentes en navigant parmi les agrégats, mais que cela n'est possible que si toutes les données sont disponibles sous une même forme dimensionnelle cohérente. Bien que les utilisateurs d'un entrepôt de données ne regardent que très rarement une ligne ou une commande isolée, ils peuvent être très intéressés par les commandes de la semaine dernière pour les produits d'une certaine taille (ou d'une certain goût, type d'emballage, fabricant) pour des clients qui ont acheté pour la première fois au cours des six derniers mois (ou qui résident dans une région particulière ou bénéficient de certaines conditions de crédit). Nous avons besoin de données du grain le plus fin dans la zone de présentation pour que les utilisateurs puissent poser les questions les plus précises possibles. Comme les exigences des utilisateurs sont imprévisibles et changent constamment, nous devons offrir l'accès aux détails les plus infimes pour qu'ils puissent être regroupés selon les questions du moment.

Tous les marchés d'infos doivent être construits à partir de dimensions et de faits communs, auxquels nous appliquons l'adjectif *conforme*. C'est la base de l'architecture de bus de l'entrepôt de données, que nous verrons de plus près au chapitre 3. L'adhésion à l'architecture de bus est notre troisième exigence concernant la zone de présentation des données. Privé de dimensions et de faits partagés et conformes, un marché d'infos n'est qu'une application verticale indépendante. Ces marchés d'infos isolés sont un frein au développement des entrepôts de données. Ils ne font que perpétuer des visions contradictoires de l'entreprise. Pour garder l'espoir de construire un entrepôt de données robuste et intégré, vous devez souscrire au principe de l'architecture de bus. Nous montrons dans ce livre que si les marchés d'infos ont été conçus avec des dimensions et des faits conformes, ils peuvent être combinés et utilisés ensemble. La zone de présentation des données d'un entrepôt de grande entreprise pourra comporter finalement une bonne vingtaine de marchés d'infos. Les modèles dimensionnels de ces marchés pourront contenir plusieurs tables de faits, avec chacun de cinq à quinze tables de dimension. Si la conception a été faite correctement, un grand nombre de ces tables de dimension seront partagés par des tables de faits distinctes.

L'architecture de bus est le secret de la construction de systèmes d'entrepôts de données distribués. Soyons réalistes — nous n'avons pour la plupart pas le budget, le temps et le pouvoir pour construire un entrepôt de données entièrement centralisé. Quand l'architecture de bus est utilisée comme cadre, l'entrepôt de données d'entreprise peut être développé de manière décentralisée (et beaucoup plus réaliste).

**Les données servant aux requêtes dans la zone de présentation de l'entrepôt de données doivent être dimensionnelles, atomiques et doivent adhérer à l'architecture de bus d'entrepôt de données.**

Si la zone de présentation utilise une base de données relationnelle, on appelle alors les tables dimensionnelles des *schémas en étoile*. Si elle est sur une base de données multidimensionnelle utilisant une technologie OLAP (*Online Analytic Processing*), les données sont alors stockées dans des *cubes*. Cette technologie ne s'appelait pas OLAP à l'origine, mais beaucoup de fournisseurs des premiers systèmes d'aide à la décision les ont construits autour du concept de cube, de sorte qu'actuellement les fournisseurs de systèmes OLAP se sont naturellement alignés sur l'approche dimensionnelle des entrepôts de données. La modélisation dimensionnelle s'applique aussi bien aux bases de données relationnelles qu'aux bases de données multidimensionnelles. L'une et l'autre ont en commun le concept logique des dimensions; cependant, la réalisation physique est différente. Heureusement, la plupart des recommandations de ce livre s'appliquent indépendamment de la plate-forme utilisée comme base de données. Bien que les possibilités de la technologie OLAP s'améliorent constamment, au moment où nous écrivons, la plupart des grands marchés d'infos sont encore réalisés sur des bases de données relationnelles. En outre, la plupart des cubes OLAP sont dérivés de forages dans des schémas de relations dimensionnelles en étoile ou font des forages dans des schémas de relations dimensionnelles en étoile utilisant une variante de la navigation à base d'agrégats. C'est pourquoi nous utilisons les termes applicables aux plates-formes relationnelles pour la plupart des discussions relatives à la zone de présentation.

Contrairement aux croyances de départ sur les entrepôts de données, les marchés d'infos modernes peuvent très bien être mis à jour, dans certains cas très souvent. À l'évidence, les données incorrectes doivent être corrigées. Les modifications d'appellation, de hiérarchie, de statut et de propriété des entreprises entraînent de nécessaires modifications aux données des marchés d'infos qui composent un entrepôt, mais le plus souvent il s'agit de mises à jour par le biais de la gestion des chargements et non par des traitements transactionnels.

## Outils d'accès aux données

Le dernier composant majeur d'un environnement d'entrepôt de données est l'ensemble des outils d'accès aux données. Nous appliquons ce terme général à un ensemble de moyens fournis aux utilisateurs pour exploiter la zone de présentation en vue de prendre des décisions basées sur des analyses. Par définition, tous les outils d'accès aux données font des requêtes sur les données de la zone de présentation. Les requêtes sont évidemment la raison d'être de l'entrepôt de données.

Un outil d'accès aux données peut être une chose aussi simple qu'un outil de requête *ad hoc* ou aussi complexe qu'une application de forage de données ou de modélisation. Les outils de requête *ad hoc*, quelle qu'en soit la puissance, ne peuvent être compris et utilisés efficacement que par une petite fraction de la population des utilisateurs potentiels d'entrepôts de données. La plupart des utilisateurs accèdent aux données par l'intermédiaire d'applications d'analyse préfabriquées, pilotées par des paramètres. Environ 80 à 90 % des utilisateurs potentiels sont desservis par ces applications qui sont essentiellement des modèles préétablis leur évitant d'avoir à construire eux-mêmes des requêtes relationnelles. Certains outils d'accès plus sophistiqués, comme les outils de modélisation ou de prévision, sont en mesure de renvoyer leurs résultats vers les applications opérationnelles ou vers les zones de préparation ou de présentation de l'entrepôt de données.



## Observations complémentaires

Pour clore notre discussion des composants de l'entrepôt de données, voici quelques autres concepts importants.

### Métadonnées

Les métadonnées sont toutes les informations de l'environnement de l'entrepôt de données qui ne sont pas les données elles-mêmes. Les métadonnées s'apparentent à une encyclopédie de l'entrepôt de données. Les équipes d'un entrepôt de données passent un temps considérable avec les métadonnées : ils en parlent, s'en inquiètent et ont mauvaise conscience à leur égard. Comme tous les développeurs ont une aversion naturelle pour la préparation et le classement ordonné de la documentation, les métadonnées sont souvent omises dans le plan du projet, bien que tout le monde reconnaisse leur importance.

Les métadonnées ont des formes très diverses et servent aux besoins des différentes parties prenantes de l'entrepôt de données : techniciens, administrateurs et utilisateurs. Il y a les métadonnées des applications opérationnelles sources, comprenant les schémas et les dossiers qui facilitent le processus d'extraction. Une fois les données dans la zone de préparation, nous avons affaire à des métadonnées pour piloter les transformations et les chargements, y compris la structure des fichiers de la zone de préparation et celle des tables cibles de la zone de présentation, ainsi que les règles à observer pour le nettoyage, les dimensions conformes et les définitions de faits, les définitions des agrégats, les plannings des opérations ETC (extraction/transformation/chargement), ainsi que le journal de ces opérations. Le code des programmes spéciaux que nous écrivons pour la zone de préparation fait partie des métadonnées.

Les métadonnées relatives au SGBD de l'entrepôt de données comprennent les tables système, l'organisation des partitions, les index, la définition des vues, ainsi que les privilèges et les autorisations accordées au titre de la sécurité du SGBD. Enfin, les métadonnées des outils d'accès incluent les noms et les définitions des tables et des colonnes de la zone de présentation ainsi que les filtres de contrainte, les spécifications des modèles des applications, les statistiques d'accès et d'utilisa-

tion et d'autres documents destinés aux utilisateurs. Et bien sûr, n'oubliez pas tous les paramétrages de sécurité, depuis ceux relatifs aux données sources des applications transactionnelles jusqu'à ceux applicables au bureau de l'utilisateur.

Le but ultime est de rassembler, de cataloguer, d'intégrer ces formes disparates de métadonnées et de s'en servir, un peu comme les ressources d'une bibliothèque. L'effort demandé par la construction de modèles dimensionnels paraît minime par comparaison. Toutefois, ce n'est pas parce que la tâche paraît considérable que nous pouvons nous dispenser de créer un cadre général pour les métadonnées de l'entrepôt de données. Nous devons établir un plan des métadonnées et donner un haute priorité à des opérations telles que l'achat ou la construction d'une application pour la conservation et le suivi de toutes les métadonnées.

### Magasin de données opérationnelles

Certains d'entre vous se demandent peut-être où se place un éventuel magasin de données opérationnelles (en anglais *operational data store*) dans le diagramme des composants de notre entrepôt de données. Comme il n'y a pas de définition universelle d'un magasin de données opérationnelles, cela dépend de votre situation particulière. Il s'agit souvent de copies fréquemment mises à jour et modérément intégrées des données opérationnelles proprement dites. La fréquence de mise à jour et le degré d'intégration varient en fonction de besoins spécifiques.

Le plus souvent, un magasin de données opérationnelles est prévu au niveau de l'entrepôt pour fournir des états sur les opérations, notamment quand ni les anciennes applications en place, ni les applications de traitement transactionnel en ligne ne fournissent d'états opérationnels adéquats. Ces états se caractérisent par un ensemble limité de requêtes de format fixe qui peuvent être codées en dur dans une application de reporting. Les états répondent le plus souvent à des besoins de décision tactique d'une organisation. Les agrégats visant à l'amélioration des performances, les historiques correspondant à des découpages du temps significatifs et les attributs descriptifs multiples n'ont pas leur place à ce niveau. Les magasins de données opérationnelles peuvent être une étape intermédiaire pour alimenter les données opérationnelles dans l'entrepôt de données.

Dans d'autres cas, les magasins de données opérationnelles sont construits pour supporter des interactions en temps réel, notamment dans des applications de gestion des relations client, comme par exemple l'accès à votre itinéraire de consultation sur un site web ou votre historique des demandes au service d'assistance aux clients. L'entrepôt de données traditionnel n'est généralement pas en mesure de produire des réponses presque en temps réel ou immédiates. Tout comme pour les états opérationnels, les requêtes de données qui correspondent à ces interactions en temps réel ont une structure fixe. Il est intéressant de noter que ce type de magasin de données opérationnelles exploite souvent des éléments de l'entrepôt de données, comme dans le cas d'un centre d'appel client qui utilise des informations sur le comportement des clients fournis par l'entrepôt de données pour précalculer des valeurs moyennes et les stocker dans le magasin de données opérationnelles.



Dans chacun de ces deux scénarios, le magasin de données opérationnelles peut être soit un troisième système physique placé entre les applications opérationnelles et l'entrepôt de données, soit une partition spécialement gérée de l'entrepôt de données lui-même. Toute organisation a évidemment besoin d'applications opérationnelles. De la même façon, toute organisation pourrait tirer avantage d'un entrepôt de données. On ne peut pas en dire autant d'un magasin de données opérationnelles distinct, sauf si aucun des deux autres systèmes n'est en mesure de traiter les questions opérationnelles immédiates. C'est seulement dans ce cas qu'il convient d'allouer des ressources à la construction d'un troisième système physique. Nous pensons que la tendance en matière d'entrepôt de données est de prévoir un magasin de données opérationnelles en tant que portion gérée de manière spécifique d'un entrepôt de données classique. Nous reviendrons plus en détail au chapitre 5 sur le sujet d'une partition «chaude» réservée au magasin de données opérationnelles.

Notons pour finir que l'équivalent anglais de magasin de données opérationnelles, ODS *(Operational Data Store)*, est parfois défini comme l'endroit de l'entrepôt de données où sont stockées les données atomiques. Nous considérons que ces données détaillées font naturellement partie de la zone de présentation de l'entrepôt de données et ne sont pas une entité distincte. À partir du chapitre 2, nous montrerons que les transactions du plus bas niveau dans une entreprise sont le fondement de la zone de présentation de l'entrepôt de données.

## 1.4 Vocabulaire de la modélisation dimensionnelle

Tout au long du livre nous nous référerons à des tables de faits et des tables de dimension. Contrairement au folklore des entrepôts de données, Ralph Kimball n'a pas inventé cette terminologie. Autant que nous le sachions, les termes *dimensions* et *faits* sont apparus dans un projet de recherche commun de General Mills et de l'université de Dartmouth au cours des années soixante. Dans les années soixante-dix, aussi bien AC Nielsen que IRI ont utilisé ces termes systématiquement pour décrire leurs offres d'information, que nous décririons exactement aujourd'hui comme étant des marchés d'infos dimensionnels destinés à des données de vente. Bien avant que la simplicité ne devienne une tendance générale, ces précurseurs dans la fourniture de bases de données se sont appuyés sur ces concepts pour simplifier la présentation des informations d'analyse. Ils comprenaient qu'une base de données ne serait utilisée qu'à condition d'être présentée de manière simple.

**L'approche dimensionnelle n'est vraisemblablement pas l'invention d'une seule personne. Il s'agit d'une force irrésistible dans la conception de bases de données, qui s'impose dès lors que le concepteur fait de l'intelligibilité et de la performance ses objectifs prioritaires.**

## Table de faits

Une table de faits est la table principale d'un modèle dimensionnel où les mesures de performance sont stockées, comme dans l'exemple de la figure 1.2. Nous nous efforçons de stocker les informations de mesure d'un processus d'entreprise dans un unique marché d'infos. Les données de mesure étant de très loin la portion la plus volumineuse d'un marché d'infos, nous nous abstenons de les dupliquer à différents endroits de l'entreprise.

| Table de faits des ventes journalières |
|---|
| Clé date (CE) |
| Clé produit (CE) |
| Clé magasin (CE) |
| Quantité vendue |
| Montant des ventes (€) |

Figure 1.2 **Exemple de table de faits**

Nous utilisons le terme *fait* pour représenter une mesure économique. Nous pouvons nous imaginer installés sur le marché, observant des produits en train d'être vendus et notant la quantité vendue et le montant de la vente chaque jour pour chaque produit dans chaque magasin. Une mesure est prise à l'intersection de toutes les dimensions (jour, produit et magasin). La liste des dimensions définit le *grain* de la table et nous dit quelle est la portée de la mesure.

**Une ligne dans une table de faits correspond à une mesure. Une mesure est une ligne dans une table de faits. Toutes les mesures figurant dans une table de faits doivent être au même grain.**

Les faits les plus utiles sont des faits numériques, additifs, tels que des montants en euros. Dans tout le livre nous utiliserons l'euro comme monnaie standard pour rendre les études de cas plus concrètes.

L'additivité est cruciale parce que les applications d'entrepôt de données ne récupèrent presque jamais une seule ligne de table de faits. En fait, elles en récupèrent des centaines, des milliers ou même des millions à la fois et la chose la plus utile que l'on puisse faire avec un tel nombre de lignes est de les additionner. À la figure 1.2, quelle que soit la tranche de données choisie par l'utilisateur, nous pouvons ajouter les quantités et les euros pour former un total valide. Nous verrons, plus loin dans le livre, qu'il existe des faits semi-additifs et d'autres qui sont non-



additifs. Les faits semi-additifs ne peuvent être additionnés que pour certaines dimensions et les faits non-additifs ne peuvent pas du tout être additionnés. Dans le cas des faits non-additifs, nous sommes obligés d'utiliser des comptages ou des moyennes si nous souhaitons cumuler les lignes, sinon nous en sommes réduits à afficher les lignes de faits une par une. Ce serait monotone dans le cas d'une table de faits d'un milliard de lignes.

**Les faits les plus utiles d'une table de faits sont numériques et additifs.**

Nous disons souvent que les faits sont valorisés de façon continue pour aider les concepteurs à faire la distinction entre un fait par opposition à un attribut de dimension. Le fait montant vendu en euros est valorisé de façon continue parce qu'il peut prendre pratiquement n'importe quelle valeur à l'intérieur d'une plage très étendue. En tant qu'observateurs, nous devons nous tenir sur le marché et attendre que la mesure soit faite pour avoir une idée de sa valeur.

Il est théoriquement possible qu'un fait mesuré soit du texte, mais c'est rare. Dans la plupart des cas, une mesure textuelle est la description de quelque chose et elle est extraite d'une liste de valeurs. Le concepteur doit faire son possible pour faire de ces mesures textuelles des dimensions parce qu'elles peuvent être corrélées plus efficacement avec les autres attributs de dimension textuels et consommeront beaucoup moins d'espace. Nous ne stockons pas des informations textuelles redondantes dans des tables de faits. À moins que le texte ne soit différent pour chaque ligne de la table de faits, il doit aller dans une table de dimension. Un véritable fait texte apparaît rarement dans un entrepôt de données, car le contenu d'un fait texte étant imprévisible, comme un commentaire en texte libre, il est pratiquement impossible de l'analyser.

Dans notre exemple de table de faits (voir figure 1.2), s'il n'y a pas d'activité de vente un certain jour dans un magasin donné pour un produit donné, nous ne plaçons pas de ligne dans la table. Il est très important de ne pas chercher à remplir la table de zéros manifestant que rien ne s'est produit, car alors la plupart de nos tables de faits succomberaient à une invasion de zéros. À condition de n'y inclure que la véritable activité, les tables de faits tendent à être creuses. Malgré cela, les tables de faits occupent 90 % ou davantage de l'espace consommé par une base de données dimensionnelle. Les tables de faits tendent à être profondes en ce qui concerne le nombre de lignes, mais étroites au niveau des colonnes. Compte tenu de leur taille, nous sommes attentifs à l'espace utilisé par ces tables.

Au fur et à mesure que nous développerons les exemples du livre, nous constaterons que le grain des tables de faits se range dans l'une des trois catégories : transaction, instantané périodique et instantané récapitulatif. Les tables de faits au grain de la transaction sont les plus courantes. Nous présentons les tables de faits de transaction au chapitre 2, les instantanés périodiques au chapitre 3 et les instantanés récapitulatifs au chapitre 5.

Toutes les tables de faits ont deux clés étrangères ou plus, désignées par la notation CE sur la figure 1.2. Ces clés étrangères se connectent aux clés primaires des tables de dimension. Par exemple, la clé produit de la table de faits correspond toujours à une clé de produit spécifique de la table de dimension produit. Lorsque toutes les clés de la table de faits renvoient effectivement aux clés primaires appropriées dans les tables de dimension correspondantes, nous disons que les tables satisfont l'*intégrité référentielle*. Nous accédons à la table de faits par le biais des tables de dimension qui lui sont jointes.

La table de faits elle-même a généralement sa propre clé primaire faite d'un sous-ensemble des clés étrangères. Cette clé est parfois appelée *clé composite* ou *clé concaténée*. Chaque table de faits d'un modèle dimensionnel a une clé composite et inversement, toute table qui a une clé composite est une table de faits. On peut dire aussi que dans un modèle dimensionnel, toute table qui exprime une relation de plusieurs-à-plusieurs doit être une table de faits. Toutes les autres tables sont des tables de dimension.

**Dans les modèles dimensionnels, les tables de faits expriment des relations de plusieurs-à-plusieurs entre les dimensions.**

Généralement, un sous-ensemble des éléments de la clé composite d'une table de faits suffit à distinguer chacune des lignes. On trouve environ une demi-douzaine de dimensions ayant entre elles de solides relations de plusieurs-à-plusieurs et qui identifient chaque ligne de manière unique. Parfois, il n'y a que deux dimensions, comme le numéro de facture et la clé produit. Une fois ce sous-ensemble identifié, les autres dimensions ne peuvent prendre qu'une seule valeur pour la clé primaire d'une ligne de la table de faits. En d'autres termes, elles suivent le mouvement. Dans la plupart des cas, il n'est pas judicieux de prévoir une clé numéro de ligne pour servir de clé primaire à la table de faits. Cette pratique élargit la table de faits et tout index sur cette clé primaire serait sans objet. Cependant, une telle clé peut être nécessaire pour satisfaire une exigence du SGBD, notamment s'il peut être légitime, du point de vue de l'activité concernée, de charger de multiples lignes identiques dans la table de faits.

## Tables de dimension

Les tables de dimension sont les compagnes obligatoires d'une table de faits. Les tables de dimension contiennent les descriptions textuelles de l'activité, comme dans l'exemple de la figure 1.3. Dans un modèle dimensionnel bien conçu, les tables de dimension ont de nombreuses colonnes ou attributs. Ces attributs décrivent les lignes de la table de dimension. Nous essayons d'introduire autant de descriptions textuelles compréhensibles que possible. Il n'est pas rare, pour une table de dimension, d'avoir de cinquante à cent attributs. Les tables de dimension tendent à être relativement peu profondes c'est-à-dire à n'avoir que peu de lignes (en



général beaucoup moins qu'un million de lignes), mais elles sont larges c'est-à-dire qu'elles ont beaucoup de colonnes. Chaque dimension est définie par son unique clé primaire dénotée à la figure 1.3 par l'abréviation CP . Cette clé primaire sert de base à l'intégrité référentielle de toute table de faits à laquelle elle est jointe.

| Table de dimension produit |
|---|
| Clé produit (CP) |
| Description du produit |
| Numéro US (clé naturelle) |
| Description de la marque |
| Description de la catégorie |
| Description du rayon |
| Description du type d'emballage |
| Taille de l'emballage |
| Description matières grasses |
| Description type de régime |
| Poids |
| Unités de mesure de poids |
| Type de stockage |
| Type de durée étagère |
| Largeur sur étagère |
| Hauteur sur étagère |
| Profondeur sur étagère |
| ... et bien d'autres attributs |

Figure 1.3 **Exemple de table de dimension**

Les attributs de dimension sont la principale source des contraintes de requête, de groupement et d'intitulés de colonne des états. Dans une requête servant à la préparation d'un état ou à une recherche, les attributs sont identifiés par des mots *by*. Par exemple, quand un utilisateur indique qu'il veut voir le montant des ventes en euros *par* marque et semaine, la semaine et la marque doivent être disponibles en tant qu'attributs de dimension.

Les attributs de dimension jouent un rôle vital dans un entrepôt de données. Comme ils sont la source de pratiquement toutes les contraintes et de tous les intitulés intéressants dans les états, ils sont le moyen par lequel l'entrepôt de données est rendu utilisable et compréhensible. À bien des égards, l'entrepôt de données vaut ce que valent les attributs des dimensions. La puissance de l'entrepôt de données est directement proportionnelle à la qualité et à la profondeur des attributs de dimension. Plus on passera de temps à fournir des attributs utilisant la terminologie de l'entreprise, meilleur sera l'entrepôt de données. Plus on passera de temps à assurer la qualité des valeurs dans une colonne d'attribut, meilleur sera l'entrepôt de données.

**Les tables de dimension sont les points d'entrée dans la table de faits. Des attributs de dimension nombreux permettent de varier les possibilités d'analyse en tranches et en dés. Les dimensions établissent l'interface homme/entrepôt de données.**

Les meilleurs attributs sont des valeurs distinctes, textuelles. Ils doivent être des mots véritables plutôt que des abréviations absconses. Les attributs typiques d'une dimension produit contiendraient une brève description (10 à 15 caractères), une longue description (30 à 50 caractères), un nom de marque, un nom de catégorie, un type d'emballage, une taille et de nombreuses autres caractéristiques de produit. Bien que la taille soit en général numérique, ce n'en est pas moins un attribut de dimension parce qu'elle se comporte davantage comme une description textuelle que comme une mesure numérique. La taille est une constante discrète décrivant un produit spécifique.

Il peut arriver, en concevant une base de données, de ne pas savoir immédiatement si un champ de données numérique extrait d'une source de données de production est un fait ou un attribut de dimension. Nous pouvons souvent en décider en nous demandant si le champ est une mesure qui peut prendre de nombreuses valeurs et participer à des calculs (c'est alors un fait) ou si c'est une valeur discrète plus ou moins constante et qui participe à des contraintes (c'est alors un attribut dimensionnel). Par exemple, le coût standard d'un produit ressemble à un attribut constant du produit mais il peut être changé si souvent qu'en fin de compte nous décidons qu'il s'agit davantage d'un fait mesuré. Il peut arriver que nous ne puissions décider de la classification à retenir. C'est alors au concepteur de choisir à son gré l'une ou l'autre solution.

Nous nous efforçons de réduire l'utilisation de codes dans nos tables de dimension en les remplaçant par un plus grand nombre d'attributs textuels. Nous comprenons que vous puissiez déjà avoir formé les utilisateurs à l'emploi des codes opérationnels, mais à l'avenir, nous souhaitons leur éviter d'avoir à coller sur leurs écrans d'ordinateur des petites notes leur rappelant la signification des codes. Le décodage des codes opérationnels doit être disponible sous forme d'attributs de dimension, de telle sorte que les appellations utilisées pour les requêtes et les états soient cohérentes. Il ne faut pas encourager les fonctions de décodage cachées dans des applications de reporting, ce qui conduit immanquablement à des incohérences. Parfois, les codes ou les identificateurs opérationnels ont une véritable signification pour les utilisateurs ou leur sont nécessaires pour pouvoir communiquer en retour avec le monde opérationnel. Ces codes doivent alors apparaître en tant qu'attribut dimensionnel explicite, en plus des descriptions textuelles conviviales. Dans les figures du livre représentant des dimensions nous avons identifié ces codes opérationnels éventuels en leur adjoignant la mention (clé naturelle).

Les codes opérationnels sont souvent pourvus de significations implicites. Par exemple, les deux premières positions peuvent identifier la branche d'activité et les deux suivantes la région. Plutôt que de forcer les utilisateurs à faire des interrogations ou des filtrages sur le code opérationnel, nous en extrayons les significations implicites et les présentons aux utilisateurs sous forme d'attributs de dimension



distincts pouvant facilement servir à des filtrages et des groupements ou figurer dans des états.

Les tables de dimension indiquent souvent des relations hiérarchiques dans une activité. Dans notre table de dimension prise comme exemple, les produits peuvent être regroupés par marques, puis par catégories. Dans chaque ligne de la dimension produit, nous stockons les descriptions de la marque et de la catégorie associées à chaque produit. Nous sommes conscients que l'information hiérarchique descriptive est stockée de façon redondante, mais nous le faisons pour faciliter l'utilisation et améliorer les performances. Nous résistons à notre envie de ne stocker que le code de la marque dans la dimension produit et de créer une table spéciale de consultation des marques. On appellerait cela un *flocon de neige*. Les tables de dimension typiques sont fortement dénormalisées. Elles sont généralement assez petites (moins de 10% de l'espace de stockage). Les tables de dimension ayant des tailles considérablement inférieures à celles des tables de faits, les gains de place obtenus pour elles par la normalisation ou la création de flocons de neige sont sans incidence sur la taille globale de la base de données. Nous donnons presque toujours la priorité à la simplicité et l'accessibilité sur l'encombrement des tables de dimension.

## Relier les faits et les dimensions

Ayant compris ce que sont les tables de faits et de dimension, rassemblons ces deux pièces de Meccano dans un modèle dimensionnel. Comme le montre la figure 1.4, la table de faits contenant des mesures numériques est jointe à un ensemble de tables remplies d'attributs descriptifs. Cette structure caractéristique en étoile est appelée *schéma de jointures en étoile*. Ce terme remonte aux premiers temps de l'utilisation des bases de données relationnelles.

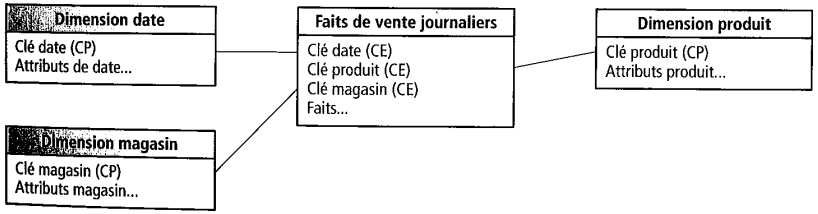


Figure 1.4 **Tables de faits et tables de dimension dans un modèle dimensionnel**

La première chose que nous remarquons à propos de ce schéma est sa simplicité et sa symétrie. Les utilisateurs bénéficient évidemment de cette simplicité puisqu'ils peuvent ainsi plus facilement comprendre et parcourir des données. Le charme de la conception de la figure 1.4 est qu'il est immédiatement reconnaissable par les

utilisateurs. Nous avons observé littéralement des centaines de cas où les utilisateurs sont immédiatement d'accord sur le fait que le modèle dimensionnel *est* leur activité. De plus, le nombre réduit de tables ainsi que l'utilisation de descriptions significatives diminuent les possibilités d'erreur.

La simplicité du modèle dimensionnel entraîne aussi des gains de performance. Les fonctions d'optimisation des SGBD traitent ces simples schémas plus efficacement et avec moins de jointures. Un moteur de base de données peut envisager avec assurance de commencer par des contraintes sur les tables de dimension fortement indexées, puis s'attaquer à la table de faits d'un seul coup avec le produit cartésien des clés des tables de dimension correspondant aux contraintes de l'utilisateur. Cela peut paraître surprenant, mais par cette approche il est possible d'évaluer n'importe quelle jointure multiple à une table de faits en un seul passage dans les index de la table de faits.

Enfin, les modèles dimensionnels sont facilement extensibles pour accueillir les changements. La structure prévisible d'un modèle dimensionnel résiste aux changements imprévisibles dans les comportements des utilisateurs. Toutes les dimensions sont équivalentes; toutes sont autant de points d'entrée symétriques dans la table de faits. Le modèle logique ne privilégie pas *a priori* des structures de requête particulières. Il n'y a pas de préférence pour les questions que nous poserons ce mois par comparaison aux questions que nous poserons le mois suivant. Nous ne voulons certainement pas avoir à ajuster nos schémas si les utilisateurs inventent de nouvelles façons d'analyser les activités.

Nous verrons à maintes reprises dans ce livre que les données ayant la plus fine granularité, c'est-à-dire atomiques, ont le plus de dimensions. Les données atomiques qui n'ont pas été agrégées sont les données les plus expressives; ces données doivent être le fondement de la conception de chaque table de faits. L'atomicité des tables de faits leur permet de résister aux assauts des requêtes inattendues (*ad hoc*) des utilisateurs. Dans le cadre du modèle dimensionnel, nous pouvons ajouter au schéma n'importe quelle dimension complètement nouvelle dès lors qu'une valeur unique de cette dimension est définie pour chaque ligne de la table de faits. De même, nous pouvons ajouter à la table de faits de nouveaux faits qui n'avaient pas été prévus, dans la mesure où leur niveau de détail est cohérent avec la table de faits existante. Nous pouvons complémenter des tables de dimension préexistantes avec de nouveaux attributs qui n'avaient pas été prévus. Nous pouvons aussi fragmenter des lignes de dimension existantes jusqu'à un niveau de granularité plus poussé à partir d'un certain point dans le temps. Dans chacun de ces cas, les tables existantes peuvent être modifiées sur place soit en leur ajoutant simplement de nouvelles lignes de données, soit en exécutant une instruction SQL `ALTER TABLE`. Les données n'ont pas à être rechargées. Toutes les applications d'accès existantes continueront de marcher sans donner des résultats différents. Nous examinons cette extensibilité en douceur au chapitre 2.



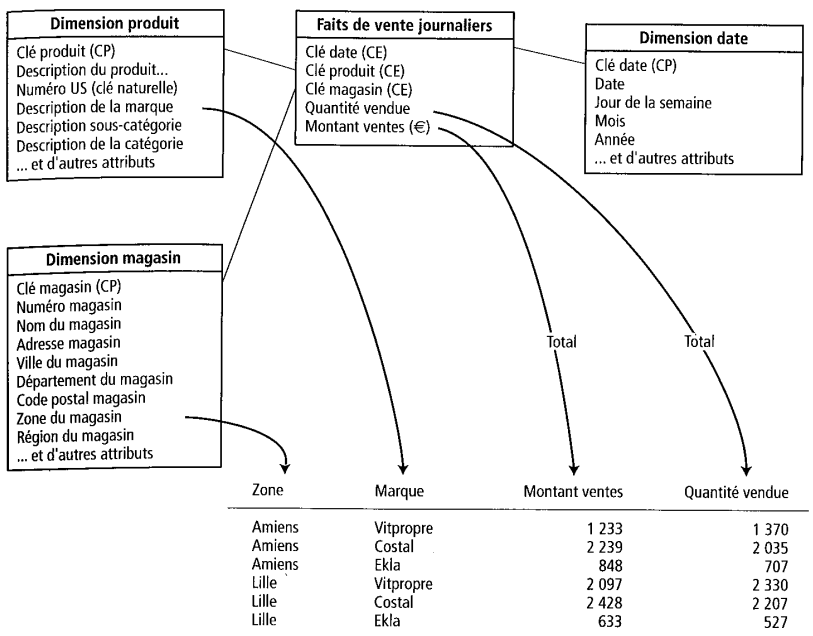


Figure 1.5 **Glisser/déposer des attributs et des faits dans un état**

Une bonne façon d'illustrer la nature complémentaire des tables de faits et de dimension est de voir comment cette complémentarité se manifeste dans un état. Comme le montre la figure 1.5, les attributs de dimension fournissent les titres de l'état, tandis que les tables de faits en fournissent les valeurs numériques.

Enfin, comme nous l'avons déjà souligné, nous tenons à ce que les données de la zone de présentation soient structurées selon le mode dimensionnel. Cependant, il y a une relation naturelle entre les modèles dimensionnels et les modèles normalisés. La clé pour comprendre cette relation est qu'un unique modèle normalisé se fragmente souvent en de multiples schémas dimensionnels. Un grand modèle normalisé pour une organisation peut présenter des visites client, des commandes, des factures client, des règlements client et des retours de produits sur un même diagramme. D'une certaine façon, le diagramme normalisé se fait du tort en représentant sur le même dessin des processus d'entreprise multiples qui ne peuvent coexister dans un même ensemble de données à aucun point dans le temps. Il n'est pas surprenant que le modèle normalisé semble complexe.

Si vous disposez déjà d'un diagramme normalisé, la première étape de sa conversion en modèles dimensionnels consiste à le segmenter en processus d'entreprise distincts, puis à modéliser chacun d'eux séparément. La seconde étape consiste à sélectionner les relations de plusieurs-à-plusieurs du diagramme normalisé conte-

nant des faits additifs numériques qui ne sont pas des clés et de les désigner comme étant des tables de faits. L'étape finale consiste à dénormaliser toutes les tables restantes en tables plates comportant des clés d'un seul élément, qui les joignent directement aux tables de faits. Ces tables deviennent des tables de dimension.

## 1.5 Mythes de la modélisation dimensionnelle

Bien que la modélisation dimensionnelle soit généralement acceptée, certains malentendus continuent de se propager dans la profession. Ce sont les *mythes de la modélisation dimensionnelle.*

**Mythe n° 1.** *Les modèles dimensionnels et les marchés d'infos ne servent que pour les données cumulées.* Ce premier mythe est à l'origine de la mauvaise conception de nombreux entrepôts de données. Comme nous ne pouvons pas prévoir toutes les questions que poseront les utilisateurs, nous devons leur fournir un accès jusqu'au niveau le plus détaillé des données, pour qu'ils puissent les cumuler de la manière exigée par toute question éventuelle. Les données au plus fin niveau de détail sont pratiquement insensibles aux surprises et aux changements. Nos marchés d'infos incluront aussi dans les schémas dimensionnels des données cumulées couramment demandées. Ces données cumulées doivent compléter les détails au grain le plus fin dans le seul but d'améliorer les performances des requêtes courantes, mais ne pas chercher à remplacer les données détaillées.

Un corollaire de ce premier mythe est qu'il ne faudrait stocker dans les structures dimensionnelles qu'une quantité limitée de données historiques. Il n'y a rien dans un modèle dimensionnel qui interdise le stockage d'informations historiques substantielles. Le volume de données historiques disponible dans un marché d'infos doit dépendre des besoins de l'activité concernée.

**Mythe n° 2.** *Les modèles dimensionnels et les marchés d'infos sont des solutions par service et non au niveau de l'entreprise.* Plutôt que de tracer des limites basées sur les services d'une organisation, nous maintenons que les marchés d'infos doivent être organisés autour de processus d'entreprise, tels que les commandes, les factures et les visites de maintenance. Des services différents veulent souvent analyser les mêmes valeurs de mesure propres à un processus d'entreprise donné. Nous nous efforçons d'éviter la duplication des mesures fondamentales dans de multiples bases de données dans toute l'organisation.

Les partisans de l'approche normalisée pour l'entrepôt de données tracent parfois des diagrammes en toile d'araignée comportant de multiples extraits d'une même source pour alimenter de multiples marchés d'infos. Ce genre d'illustration prétend décrire les périls encourus en ne partant pas d'un entrepôt de données normalisé pour alimenter les marchés d'infos. Les partisans de la normalisation mettent en garde contre les coûts supplémentaires et les risques d'incohérence en cas de modification dans les systèmes sources du fait que les modifications devraient être prises en compte dans le processus ETC (extraction/transformation/chargement) de chaque marché d'infos.



Cet argument ne tient pas debout parce que personne ne recommande de faire de multiples extraits à partir de la même source. Les diagrammes en toile d'araignée ne tiennent pas compte de ce que les marchés d'infos sont centrés sur les processus de l'organisation et non ses services, ni de ce que les données sont extraites une seule fois de la source opérationnelle et présentées à un seul endroit. Il est clair que les responsables du support d'une application opérationnelle verraient d'un mauvais œil une approche impliquant de multiples extractions. Nous aussi.

**Mythe n° 3.** *Les modèles dimensionnels et les marchés d'infos ne sont pas extensibles.* Les tables de faits modernes abritent des milliards de lignes. Les modèles relationnels de nos marchés d'infos sont extrêmement extensibles. Les fournisseurs de SGBD ont accueilli les entrepôts de données à bras ouverts et ont incorporé de nombreuses fonctions dans leurs produits pour augmenter l'extensibilité et optimiser la performance des modèles dimensionnels.

Un corollaire du mythe n° 3 est que les modèles dimensionnels ne conviendraient qu'aux données de la distribution. Cette croyance prend racine dans l'origine historique de la modélisation dimensionnelle, mais est sans rapport avec les réalités actuelles. La modélisation dimensionnelle a été appliquée à pratiquement toutes les activités, notamment la banque, les assurances, le courtage, le téléphone, les journaux, les produits pétroliers, les administrations, la fabrication, les voyages, les jeux, la santé, l'éducation et bien d'autres. Si nous utilisons la distribution dans cet ouvrage pour illustrer plusieurs des concepts initiaux, c'est surtout parce nous avons tous eu des contacts avec cette activité; mais ces concepts sont hautement transférables à d'autres activités.

**Mythe n° 4.** *Les modèles dimensionnels et les marchés d'infos ne conviennent que s'il existe des schémas d'utilisation prévisibles.* Un corollaire est que les modèles dimensionnels ne prendraient pas en compte les modifications dans les besoins des activités. Bien au contraire, grâce à leur symétrie, les structures dimensionnelles de nos marchés d'infos sont extrêmement souples et adaptables aux changements. Le secret de la souplesse en matière de requête est la construction de tables de faits au grain le plus fin. À notre avis, la source du mythe n° 4 est le problème du concepteur en butte à des tables de faits qui ont été prématurément agrégées sur la base d'une regrettable croyance dans le mythe n° 1 sur les données cumulatives. Les modèles dimensionnels qui ne fournissent que des données cumulatives sont forcément pro-

blématiques. Les utilisateurs se heurtent à un mur quand ils veulent faire des forages jusqu'à des détails qui sont absents des tables de cumul. Les développeurs eux aussi se heurtent à un mur parce qu'ils ne peuvent pas introduire facilement de nouvelles dimensions, de nouveaux attributs ou de nouveaux faits à cause de ces tables cumulatives. Le point de départ correct de vos modèles dimensionnels est d'exprimer les données au plus petit niveau de détail possible en vue d'une souplesse et d'une extensibilité maximales.

**Mythe n° 5.** *Les modèles dimensionnels et les marchés d'infos ne peuvent pas être intégrés et conduisent par suite à des solutions verticales.* Les modèles dimensionnels et les marchés d'infos peuvent certainement être intégrés s'ils sont conformes à l'architecture de bus des entrepôts de données. Les bases de données de zone de présentation qui n'adhèrent pas à cette architecture de bus conduisent à des solutions indépendantes les unes des autres. On ne peut rendre la modélisation dimensionnelle responsable de l'incapacité de certaines organisations à respecter l'un de ses principes fondamentaux.

## Dix erreurs fréquentes

Nous pouvons fournir des recommandations positives sur les entrepôts de données dimensionnels, mais certains lecteurs sont friands de listes de pièges courants dans lesquels d'autres sont déjà tombés. Voici notre liste des dix erreurs les plus courantes à ne pas commettre lors de la construction d'un entrepôt de données. Ce sont toutes des erreurs fatales — une seule suffit à provoquer l'échec d'un projet d'entrepôt de données. Nous les évoquerons plus à fond au chapitre 16, mais nous vous mettons ici en garde à titre préliminaire.

**Erreur n° 10.** Porter une dévotion excessive à la technologie et aux données au lieu de se concentrer sur les exigences et les objectifs de l'activité.

**Erreur n° 9.** Ne pas s'assurer l'appui d'un membre du management à la fois influent, disponible et raisonnable comme sponsor visionnaire de l'entrepôt de données.

**Erreur n° 8.** Se lancer dans un projet de proportions galactiques sur plusieurs années au lieu de prévoir des efforts de développement plus gérables, quoique ardus et itératifs.

**Erreur n° 7.** Consacrer de l'énergie à la construction d'une structure normalisée et néanmoins dépasser les budgets avant d'avoir construit une zone de présentation basée sur les modèles dimensionnels.

**Erreur n° 6.** S'intéresser à la performance opérationnelle de la zone de préparation et à la facilité de développement plutôt qu'à la performance des requêtes et à la commodité d'utilisation de la zone de présentation.

**Erreur n° 5.** Rendre les données destinées aux requêtes de la zone de présentation excessivement complexes. Les concepteurs de bases de données qui préfèrent une présentation plus complexe devraient passer un an à supporter les utilisateurs; ils pourraient ainsi beaucoup mieux apprécier la nécessité de rechercher des solutions plus simples.



**Erreur n° 4.** Renseigner les modèles dimensionnels indépendamment les uns des autres sans respecter une architecture commune représentée par des dimensions communes et conformes.

**Erreur n° 3.** Ne charger que des données cumulatives dans les structures dimensionnelles de la zone de présentation.

**Erreur n° 2.** Supposer que l'activité de l'organisation, ses besoins et ses modes d'analyse, ainsi que les données elles-mêmes et les technologies qui les traitent sont statiques.

**Erreur n° 1.** Ne pas voir que le succès de l'entrepôt de données est lié directement à son acceptation par les utilisateurs. Si les utilisateurs ne reconnaissent pas l'entrepôt de données comme un moyen fondamental d'amélioration des prises de décision, vos efforts auront été vains.

## Résumé

Dans ce chapitre nous avons évoqué les objectifs essentiels de l'entrepôt de données ainsi que les différences entre les entrepôts de données et les applications opérationnelles sources. Nous avons exploré les principaux composants de l'entrepôt de données et indiqué le rôle éventuel de modèles normalisés dans la zone de préparation, bien qu'ils ne soient pas une fin en soi. Nous avons ensuite porté notre attention sur la modélisation dimensionnelle au sein de la zone de présentation et défini les premiers éléments du vocabulaire applicable aux faits et aux dimensions. Tous ces concepts vont servir pour la première étude de cas, traitée au chapitre suivant.