N^{0}d'ordre :

UNIVERSITE*MOHAMED BOUDIAF* DE M'SILA
FACULTE DES SCIENCES ET SCIENCES DE L'INGENIEUR
DEPARTEMENT D'INFORMATIQUE

MEMOIRE

Présenté pour l'obtention du diplôme de :

Magister

Spécialité : Informatique
Option : Informatique industrielle

Par :

DJEDDAOUI Mohamed

Thème :

Elaboration d'un système multidimensionnel

Soutenu publiquement le : 23 / 09 / 2006 devant le jury composé de :

AHMED Nacer Mohamed	Pr. USTHB	Président
BOUDERAH Brahim	M.C. Université de M'sila	Rapporteur
MAAMRA Mohamed Said	C. R. USTHB	Co-encadreur
ABBAS Moncef	Pr. USTHB	Examinateur
BELOUADAH Hocine	M.C. Université de M'sila	Examinateur

Remerciement

*Je remercie en premier lieu **Dieu** le tout puissant qui m'a donné la force et la patience pour bien achever ce modeste travail.*

Je tiens à remercier les membres du jury pour l'honneur qu'ils me font en acceptant de juger mon travail.

Puis je tiens à remercier :

Monsieur Brahim BOUDERAH, Maître de conférence à l'Université Med BOUDIAF de M'sila pour avoir accepté de m'encadrer et de me fournir de précieux conseils. Je lui exprime ma profonde gratitude pour m'avoir encouragé et soutenu.

Monsieur Mohamed Said MAAMRA, Chargé de recherche à l'USTHB, co-encadreur de ce mémoire. Ses orientations ses encouragements m'ont donné la volonté d'achever mon travail. Je le remercie encore pour sa proposition de ce sujet.

En fin Je tiens à remercier tous ceux qui ont contribué de prés ou de loin à l'élaboration de ce travail.

A mes parents,
A mes frères,
A mes amis.

Résumé

Le processus d'analyse en ligne (OLAP) décrit l'approche multidimensionnelle de l'aide à la décision. Ce processus permet de structurer et de stoker les données provenant des bases de production dans un format multidimensionnel connu sous le nom de cube de données. Ce cube organise les données selon les perspectives d'analyse et selon les niveaux de détail.

Notre travail présente une démarche dédie à l'élaboration d'un système multidimensionnel basé sur le concept OLAP. Notre travail repose essentiellement sur une modélisation conceptuelle du cube de données, une mécanisme pour l'extraction de données en utilisant un graphe connexe et finalement, sur une démarche pour manipuler les données du cube en utilisant une représentation tabulaire.

Mots clés:

OLAP, cube de données, entrepôt de données, magasin de données, systèmes multidimensionnels, système d'aide à la décision.

Sommaire

Table des matières

Introduction générale……...……..... **1**

CHAPITRE I : Les systèmes décisionnels

1. Introduction ……. 4
2. Système OLTP versus système OLAP……..……. 4
3. Systèmes décisionnels 5
   3.1 Les sources de donnée……..…..……. 6
   3.2 L'entrepôt de données (data warehouse) ….…..…….. 7
   3.3 Les magasins de données (Data mart)……..... 7
   3.4 Les outils d'analyse…….……..……. 7
4. Le concept OLAP……...……. 7
   4.1 Les 12 règles OLAP……..…… 8
5. Les tendances de modèle…….…….……... 10
   5.1 stocker un maximum des données…….……..... 10
   5.2 définir de nombreuses dimensions……..…….. 11
   5.3 rechercher le meilleur temps de réponse……....…… 11
6. Synthèse…… ……. ... ……...…..... 11

CHAPITRE II: L'entrepôt de données

1. Introduction …….……..…… 12
2. Les besoins décisionnels …….…..……. 12
3. l'infocentre…….…….……. 13
4. L'entrepôt de données…….……. 13
   4.1 L'intégrité de données…….…….…… 15

4.2 L'orientation sujet ……..……..…….. 15
4.2.1 Le magasin de données……..……... 16
4.3 L'historisation de Données……...… 16
5. Modélisation de l'entrepôt de données ……..... 16
6. L'extraction des données ……...……. 17
6.1 Les vues matérialisées…….……. 17
6.1.1 Problème lies aux vues matérialisées……..... 18
6.1.1.1 Problème de maintenance……. 18
6.1.1.2 Problème de sélection…….…….……. 18
7 L'historisation……..……..… ……. 19
7.1 Manipulation de temps……..…… 20
7.2 L'évolution de données……..…… 20
7.3 Le rafraîchissement des données……..…… 21
8 L'agrégation…… …… 22
9 Synthèse…….……. 23

CHAPITRE III : la modélisation multidimensionnelle

1. Introduction……..… 24
2. Modélisation multidimensionnelle…….…..… 25
2.1 Modélisation conceptuelle…….……..……. 25
2.1.1 Le fait …….……..... 26
2.1.2 La dimension…….……..... 26
2.1.2.1 Notion de hiérarchie…….…….. 27
2.1.2.1.1 Hiérarchie simple ou multiple…….. 27
2.2 Modèles en étoile……..... 28
2.3 Modèles en flocon…….……. 29
2.4 Modèles en constellation …….……..… 29
3. Les opérations multidimensionnelles ……....…… 30
3.1 Opérations classiques …….……..……. 30
3.1.2 Opérations de sélection et projection …… 31

3.2 Opérations associées à la granularité ……..…. 31
3.3 Opérations agissant sur la structure……..... 33
3.3.1 La rotation…….……. 33
3.3.2 La permutation …….…….. 33
3.3.3 La division……..... 34
3.3.4 L'emboîtement……...…….…. 34
3.3.5 L'enfoncement ……. 34
3.3.6 La factualisation……... 35
3.3.7Cube….……..… …… 35
4. Les implémentations desmodèlesmultidimensionnels……... 36
4.1 Les systèmes MOLAP…….. 36
4.2 Les systèmes ROLAP…….…….. 37
4.2.1 Schéma en étoile……..... 37
4.2.2 Schéma en flocon …….. 38
4.2.3 La normalisation …….…… 39
4.3 Le modèle HOLAP…….……..... 39
5. L'exemple du système oracle…….. 40
5.1 Introduction …….…….……..……... 40
5.2 Les dimensions…….……..…… 40
5.3 Les variables 41
5.4 Les relations……..……. 41
5.5 Les formules ……..…….…….…….. 42
5.6 Limitation de la vue sur les données……..……..... 43
5.7 Calcule des données agrégées……..……... 43
6.Synthèse……... 44

CHAPITRE IV : L'élaboration du système multidimensionnel

1. Introduction ……...……. 45
2. Description du système……..……. 45

3. Architecture du système…….. 46
4. Modélisation du magasin de données……..…..... 47
4.1 Fait…….……..... 47
4.2 Dimensions et hiérarchies…….……..... 48
4.3 Schéma multidimensionnel…….…….. 50
5. L'extraction de données……..... 52
5.1 Modélisation par un graphe ……..... 52
5.1.1 Le graphe connexe……. 55
5.2 Détermination du fait et des dimensions …….. 56
5.2.1 Détermination des dimensions…….. 56
5.2.2 Détermination du fait ……..... 57
5.2.3 Détermination des fonctions de groupement……. 58
6. Manipulation des données du magasin……..... 59
6.1 Représentation tabulaire du schéma multidimensionnel…..... 59
6.2 Opérations associées à la granularité…….……..... 61
6.2.1 Opérations de forage…….…….. 61
6.2.2 Opération d'agrégation……..…….. 63
6.2.3 Opération de calcul des totaux…….…….. 64
6.3 Opérations agissant sur la structure…….…… 65
6.3.1 Opérations de rotation…….…….. 65
6.3.2 Opération de permutation…….…… 66
6.3.3 Opération d'emboîtement…….…… 67
6.4 Opérations classiques ……..……..... 68
7. Conception d'interfaces ……..……..... 69
7.1 L'interface d'administration ……..……..... 69
7.2 L'interface décideur…….. 70
8. Exemple complet d'élaboration et manipulation d'un magasin de données…… 71
8.1 Définition des liens……..... 72
8.2 Détermination des dimensions ……. 73
8.3 Manipulation des données du magasin multidimensionnel …….. 75

8.3.1 Représentation graphique des données ……. 76
9. Limitation……. 76
10. Synthèse ……..…….. 77
Conclusion ……..... 78
Bibliographie …… **80**
Annexe A……. **83**

Introduction

Introduction générale

La prise de décision est devenue cruciale et indispensable pour diriger les entreprises dans un monde concurrent. Ces entreprises ont des importants volumes d'informations provenant de système opérationnel de l'entreprise ou d'environnement extérieur. Les dirigeants de l'entreprise sont à la recherche d'outils et de systèmes exploitant efficacement ces informations pour supporter la prise de décision.

Les systèmes traditionnels sont conçus pour gérer les activités quotidiennes de l'entreprise. Ils reflètent les exigences de l'informatique transactionnelle mais s'avèrent inadaptés au décisionnel [1] [2]. Face à ce problème plusieurs systèmes sont apparus, on parle de systèmes décisionnels. Ces systèmes sont basés sur l'approche multidimensionnelle de l'aide à la décision OLAP.

Objectifs

Notre objectif est d'appliquer les techniques d'aide à la décision tout en se basant sur les traitements OLAP pour élaborer un système multidimensionnel. Notre système peut être considéré comme un outil ROLAP(Relational OLAP) qui permet d'organiser les données dans un magasin de données à partir d'un entrepôt de données de manière multidimensionnelle(on utilise un schéma en étoile). La réorganisation multidimensionnelle des données constitue une tâche essentielle dans l'élaboration d'un système décisionnel [4]. Nous choisissons le modèle relationnel parce qu'il s'intègre facilement dans un contexte de base de données relationnel, de plus il ne nécessite pas de redéfinir des nouvelles opérations pour manipuler les structures de données.

Pour atteindre notre objectif nous présentons:

- Une modélisation conceptuelle du magasin.
- Un mécanisme pour l'extraction et la transformation des données dans le magasin à partir d'un entrepôt.
- Une algèbre pour la manipulation du magasin.

Nous complétons notre travail par un prototype de logiciel comprenant deux interfaces homme-machine permettant la création et la manipulation du magasin (cube de donnés) :

- Une interface administrateur dédiée à la construction d'une base de données multidimensionnelle (cube de données) en utilisant un modèle en étoile [1].

- Une interface utilisateur (décideur) permettant la manipulation de données selon les dimensions du cube de manière simple et interactive (Roll up, Drill down…). Cette interface offre des représentations tabulaires et graphiques pour visualiser les données.

Organisation du mémoire

Pour présenter notre travail, nous avons retenu pour ce mémoire une organisation en quatre chapitres.

Dans le premier chapitre, nous présentons les concepts de base de systèmes décisionnels, notamment le concept OLAP, les critères et les tendances de ce dernier.

Dans le deuxième chapitre, nous étudions le concept d'entrepôt de données, nous présentons aussi quelque Techniques (approches) liées au développement d'entrepôt de données tels que l'extraction et l'historisation.

Dans le troisième chapitre, nous étudions la modélisation multidimensionnelle et les différents modèles logiques pour implémenter les schémas multidimensionnels. Nous concluons ce chapitre par la représentation d'un exemple relatif au système MOLAP.

Dans le quatrième chapitre, nous présentons notre travail en se basent sur une modélisation conceptuelle pour le magasin, nous proposons un mécanisme d'extraction de données en utilisant une modélisation par un graphe connexe, une algèbre pour la manipulation de données du magasin (opération multidimensionnelle). Nous présentons un exemple complet d'élaboration et manipulation d'un magasin de données à travers les fenêtres de notre prototype de logiciel.

CHAPITRE I :

Les systèmes décisionnels

1. Introduction :

Les bases de données et les systèmes de gestion de bases de données jouent un rôle vital dans le développement et le succès d'entreprises. Le modèle le plus utilisé au sein d'entreprise est le modèle relationnel, il permet de stocker et de manipuler des importants volumes de données. L'exploitation de ces données pour l'analyse et la prise de décision est réalisée souvent par des outils de reporting (requêtes), graphiques, tableaux...

Le décideur dans le monde transactionnel, pose des questions relativement simples par exemple << quel est le chiffre d'affaires du moins derniers >>. Les besoins du décideur sont totalement différents. IL ne pose pas la question précédente, mais << quel est le chiffre d'affaires par produit et par zone géographique pour les trois derniers moins >>. Cette requête est complexe, lente et pénalise le système de production. Face à ce problème les entreprises sont recours à des systèmes d'aide à la décision spécifiques, basés sur l'approche multidimensionnel (OLAP).

2. Système OLTP versus système OLAP

Dans l'entreprise les données sont stockées et gérées sous différent systèmes et différents formats selon des processus transactionnels en ligne (OLTP : "On-Line Transactional Processing"[2]), ces processus mettent en avant la mise à jours en linge, La capacité de stockage, la rapidité d'accès, la garantie d'intégrité et la sécurité des transactions effectuées (notamment les bases réparties en plusieurs lieux physiques). Ces processus reflètent l'exigences de l'informatique transactionnelle, mais s'avère inadapté au décisionnel [1] [2].

La base de production contient des informations concernant l'activité de l'entreprise (les ventes, les produit, les stockes, les client, les fournisseurs …).

Les décideurs dans l'entreprise sont à la recherche de systèmes exploitant efficacement ces informations pour des fins d'analyse et de prise de décision.
Ces systèmes utilisent des processus d'analyse en ligne de données (OLAP : "On-Line analytical Processing"), tout comme OLTP, OLAP a été proposé par E.F.Codd. E.F.Codd définit douze règles représentant des critères pour les systèmes décisionnels. Ces processus mettent en avant la rapidité et l'interactivité d'analyses pour des données généralement volumineuses, varies et historisées (gardent l'évolution de données). Les utilisateurs de ces systèmes sont peu nombreux que les utilisateurs des systèmes OLTP, on parle de décideurs.
Le Tableau 1 compare les caractéristiques des processus OLTP et OLAP [10].

	Processus OLTP	**Processus OLAP**
Données	Exhaustives Courantes Dynamiques Orientées applications	Résumées Historiques Statiques Orientées sujets (d'analyse)
Utilisateurs	Nombreux Variés (employés, directeurs,…) Concurrents Mises à jours et interrogations Requêtes prédéfinies Réponses immédiates Accès à peu d'information	Peu nombreux Uniquement les décideurs Non concurrents Interrogations Requêtes imprévisibles et complexes Réponses moins rapides Accès à de nombreuses informations

Tableau 1 : Comparaison des processus OLTP et OLAP.

3. Systèmes décisionnels

Cette exploitation de l'information contenue dans les bases de donnée de l'entreprise met en place de nouveau systèmes informatiques dédiés à l'analyse et à la prise de décision, on parle de systèmes décisionnels.

Ces systèmes collectent les d'information pertinentes aux besoins de décideurs à partie d'informations collectées, on peut extraire des données et les organiser de manière multidimensionnelle pour permettre une analyse interactive et rapide. La réorganisation multidimensionnelle des données constitue une tâche essentielle dans l'élaboration d'un système décisionnel [4].

De manière synthétique, le système décisionnel peut être schématisé ainsi:

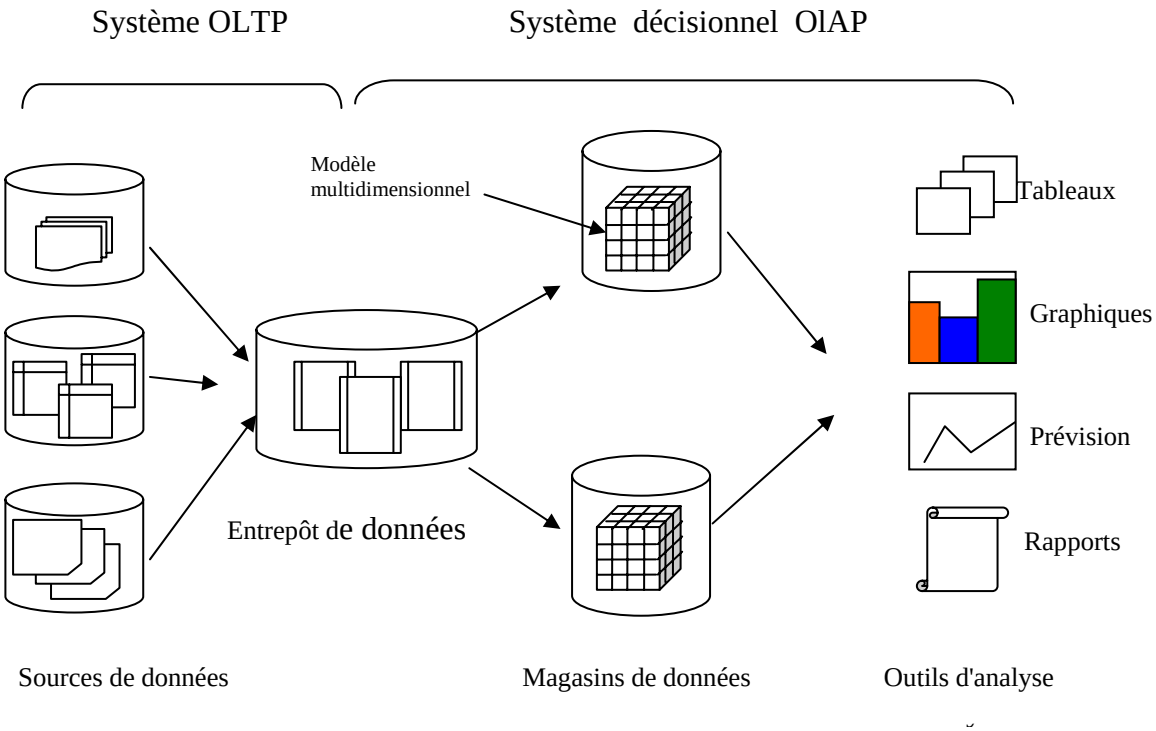


Figure 1 : architecture des systèmes décisionnels **.**

3.1 Les sources de données

Les sources de donnée sont nombreuses et variées :

-les bases de production (différents format, reparties…). Ces bases peuvent enregistrer l'ensemble des informations formelles.

-les informations externes à l'entreprise, mais nécessaires aux décideurs telles celles obtenues au cours de forums, salons, séminaires, auprès d'un expert, internet. Ces informations entrant plutôt dans la catégorie informelle.

3.2 L'entrepôt de données (data warehouse)

L'entrepôt de données est le lieu centralisé pour stocker les informations décisionnelles. C'est une base de données conçue spécifiquement pour l'analyse et l'interrogation (requêtes complexes). Il collecte les données provenues des différentes sources et garde leur historique (conserve l'évolution de données).

3.3 Les magasins de données (Data mart)

Les magasins de données sont des extraits de l'entrepôt orientés sujet. Les données sont organisées de manière adéquate pour permettre des analyses rapides à des fins de prise de décision.

3.4 Les outils d'analyse

Les outils d'analyse sont des outils d'accès aux données des magasins. Ils permettent de manipuler les données suivant des axes d'analyses (Modèle multidimensionnel). L'information est visualisée au travers d'interfaces interactives et fonctionnelles dédiées à des décideurs souvent non informaticiens (directeurs, chefs de services,…) pour leur aider de prendre les décisions. Ces outils représentent la partie visible dans le système décisionnel pour les utilisateurs (décideurs). Ils offrent de différentes représentations de l'information: tableaux, graphiques, courbe de prévision (statistique).

4. Le concept OLAP

Le processus **OLAP** repose sur le principe: restructurer et stocker les données dans un format **multidimensionnel** issus de bases de production ou de l'extérieur de l'entreprise. Ce format est connu sous le nom de hypercube, organise les données suivant des dimensions. Ainsi les utilisateurs analysent

les données suivant les dimension propres à leur activité. On peut dire que le terme OLAP décrit l'approche multidimensionnelle de l'aide à la décision.

4. 1 Les 12 règles OLAP

La philosophie OLAP comprend douze règles [2]. Dans cette section nous citions ces règles avec explication de chaque règle [9].

-Vue multidimensionnelle

L'utilisateur a l'habitude de raisonner en vue multidimensionnelle comme par exemple lorsqu'il souhaite analyser les ventes par produit mais aussi par région ou par période. Ces modèles permettent des manipulations simples : rotation, pivot ou vues par tranche, analyse de type permutations d'axes ou en cascade (drill anywhere).

-Transparence du serveur OLAP à différents types de logiciels

Cette transparence se traduit pour l'utilisateur par un complément à ses outils habituels garantissant ainsi sa productivité et sa compétence. Elle s'appuie sur une architecture ouverte permettant à l'utilisateur d'implanter le système OLAP sans affecter les fonctionnalités du système central.
Par ailleurs, l'utilisateur ne doit pas être concerné par l'intégration des données dans OLAP provenant d'un environnement homogène ou hétérogène.

-Accessibilité à de nombreuses sources de données

Le système OLAP doit donner accès aux données nécessaires aux analyses demandées. Les outils OLAP doivent avoir leur propre schéma logique de stockage des données physiques hétérogènes, doivent accéder aux données et réaliser n'importe quelle conversion afin de présenter à l'utilisateur une vue simple et cohérente. Ils doivent aussi savoir de quel type de systèmes proviennent les données.

-Performance du système de Reporting

L'augmentation du nombre de dimensions ou du volume de la base de données ne doit pas entraîner de dégradation visible par l'utilisateur.

-Architecture Client/Serveur

La plupart des données pour OLAP sont stockées sur des gros systèmes et sont accessibles via des PC. Il est donc nécessaire que les produits OLAP soient capables de travailler dans un environnement Client/Serveur.

-Dimensions Génériques

Toutes les dimensions doivent être équivalentes en structure et en calcul. Il ne doit exister qu'une seule structure logique pour toutes les dimensions. Toute fonction qui s'applique à une dimension doit être aussi capable de s'appliquer à une autre dimension.

-Gestion dynamique des matrices creuses

Le schéma physique des outils OLAP doit s'adapter entièrement au modèle d'analyse spécifique créé pour optimiser la gestion des matrices creuses. En effet, dans une analyse à la fois sur les produits et les régions, tous les produits ne sont pas vendus dans toutes les régions.

-Support multi-utilisateurs

Les outils OLAP doivent supporter les accès concurrents, garantir l'intégrité et la sécurité afin que plusieurs utilisateurs accèdent au même modèle d'analyse.

-Calculs à travers les dimensions

Les opérations doivent pouvoir s'effectuer sur toutes les dimensions et ne doivent pas faire intervenir l'utilisateur pour définir un calcul hiérarchique.

-Manipulation intuitive des données

Toute manipulation doit être accomplie via une action directe sur les cellules du modèle sans utiliser de menus ou des chemins multiples à travers l'interface utilisateur.

-Souplesse et facilité de constitution des rapports

La création des rapports dans les outils OLAP doit permettre aux utilisateurs de présenter comme ils le désirent des données synthétiques ou des résultats en fonction de l'orientation du modèle.

-Nombre illimité de niveaux d'agrégation et de dimensions

Tout outil OLAP doit gérer au moins 15 à 20 dimensions.

5. Les tendances de modèle

Actuellement, l'expansion non maîtrisée constitue le principal problème des systèmes OLAP. En effet, la tendance naturelle conduit à trois éléments suivant[6] :

-stocker un maximum des données.
-définir de nombreuses dimensions.
-rechercher le meilleur temps de réponse.

5.1 stocker un maximum des données

Les utilisateurs auront besoin de rapprocher de nombreuses données et de conserver un maximum d'historiques. Actuellement, avec un rythme de renouvellement accéléré des gammes de produits des entreprises, les données évoluent très rapidement et demandent des bases de plus en plus volumineuses.

5.2 définir de nombreuses dimensions

Les dimensions doivent être prévus à l'origine du chargement de la base OLAP. Pour éviter des impasses et faciliter les analyses, les utilisateurs ont tendance à définir de nombreuses dimensions pour un même ensemble de données. Souvent ces dimensions reste inexploitées et occupent une place inutile. A l'analyse des bases existantes, on constate qu'elles contiennent de nombreuses cases vides et les données sont très clairsemées.

5.3 rechercher le meilleur temps de réponse

Pour être acceptable, un maximum de requêtes doit être traité en moins de 5 secondes. Quelque unes uniquement seront traitées en un temps maximal de 30 secondes. Au-delà, c'est le domaine de l'inacceptable. Pour un maximum de performance, de nombreux agrégats sont pré-caclués et stockés.

6. Synthèse

Nous avons présenté dans ce chapitre le concept de **système décisionnel** comme un système d'information suivant un processus de type OLAP.

Le processus **OLAP** repose sur le principe: restructurer et stocker les données issus de bases de production ou de l'extérieur de l'entreprise dans un format **multidimensionnel**. Cette restructuration permet de faire des analyses rapides et interactives.

Le système décisionnel comprend quatre éléments : les sources des données, l'entrepôt de données, les magasin des données et les outil d'analyse.

Le Système OLAP à des tendances tels que : le stockage maximum de données avec nombreuses dimension, l'interrogation de ces données doit être en temps optimal.

La base de données appropriée pour le processus OLAP est l'entrepôt de données. Ce concept sera exposé dans le chapitre suivant.

CHAPITRE II:

L'entrepôt de données

1. Introduction

L'informatique transactionnelle met en avant la rapidité d'exécution et la sécurité, notamment en cas d'accès concurrent, l'utilisateur dans un monde de production pose des questions relativement simples ou tout moins prédéfinies et pouvant ainsi être optimisées a priori.

Dans le cadre de l'aide à la décision. L'utilisateur sera peut-être plus exigeant en matière de complexité de requête. Il cherchera à créer des informations en agrégeant des données qui, a priori, ne correspondent pas à un modèle de production.

La construction de ces agrégats, recherche spécifique au calcul coûte relativement cher en temps machine et de plus risque de perturber la production.

En règle générale, les requêtes complexes porteront sur des données historisées autrement dit, un décideur souhaitant connaître l'état des ventes d'une région pour les trois derniers mois, peut se contenter de la mise à jour de veille.

2. Les besoins décisionnels

Pour un besoin décisionnel; les données devront être stable et cohérentes au moment de leur utilisation [9].

-Stables, il ne faut pas qu'elles évoluent entre deux questions successives (deux requêtes) qui seront évaluées l'une par rapport à l'autre**.**

-cohérentes, elles devront correspondre à un cliché de l'entreprise à une instante donnée. On ne va pas comparer le chiffre d'affaires du magasin B daté du 29, avec le chiffre d'affaires du magasin A daté du 30, s'il y a eu une mise à jour entre-temps.

Il y a de cela quelques années, un nouveau concept est né **: l'infocentre**.

3. l'infocentre

L'infocentre est un SGBDR présentant une copie de travail d'une partie de la base de production, mise à jour périodiquement. L'infocentre soulage le système de production des requêtes complexes du décideur. Il permet de transférer les données de base de type non relationnel dans un univers plus propice à l'interrogation impromptue [6].

L'infocentre est une collection de données orientées sujet, intégrées, volatiles, actuelles, organisées pour le support d'un processus de décision ponctuel [9].

Dans un infocentre, chaque nouvelle valeur remplace l'ancienne valeur. Il est donc impossible de retrouver une valeur calculée dans une session préalable aux dernières alimentations(volatilité des données). L'infocentre sert à prendre des décisions opérationnelles basées sur des valeurs courantes (pas d'historique de données). Dans un infocentre, l'intégration des données est plus ou moins poussée et le processus d'alimentation est simple.

4. L'entrepôt de données

Les entrepôts de données apportent des solutions au problème d'exploitation efficace d'importants volumes de données pour l'analyse et l'aide à la décision. W.H.Immon définit l'entrepôt de données dans son ouvrage "Building the data warehouse " [3] comme "une collection de données intégrées, orientées sujet, non volatiles, historisées, résumées et disponibles pour l'interrogation et l'analyse".

-Intégrées : les données de l'entrepôt proviennent de différentes sources éventuellement hétérogènes. L'intégration consiste à résoudre les problèmes d'hétérogénéité des modèles, des schémas, de la sémantique.

-Orientées sujet : Les données de l'entrepôt peuvent être réorganisée autour de thèmes tels que patient, diagnostic, médicaments…

-Non volatiles : les données de l'entrepôts sont essentiellement utilisées en mode de consultation ; elles sont très rarement modifiées.

-Historisées : la prise en compte de l'évolution des données est primordiale pour la prise de décision et notamment les prédictions.

-Résumées : les informations issues des sources de données doivent être agrégées et réorganisées afin de faciliter le processus de prise de décision.

-Disponibles pour l'interrogation et l'analyse : les utilisateurs doivent pouvoir consulter les données réorganisées de l'entrepôt en fonction de leurs droits d'accès au travers d'outils interactifs d'aide à la manipulation et l'analyse.

La figure suivante illustre l'emplacement d'un entrepôt de données dans l'architecture de système décisionnel.

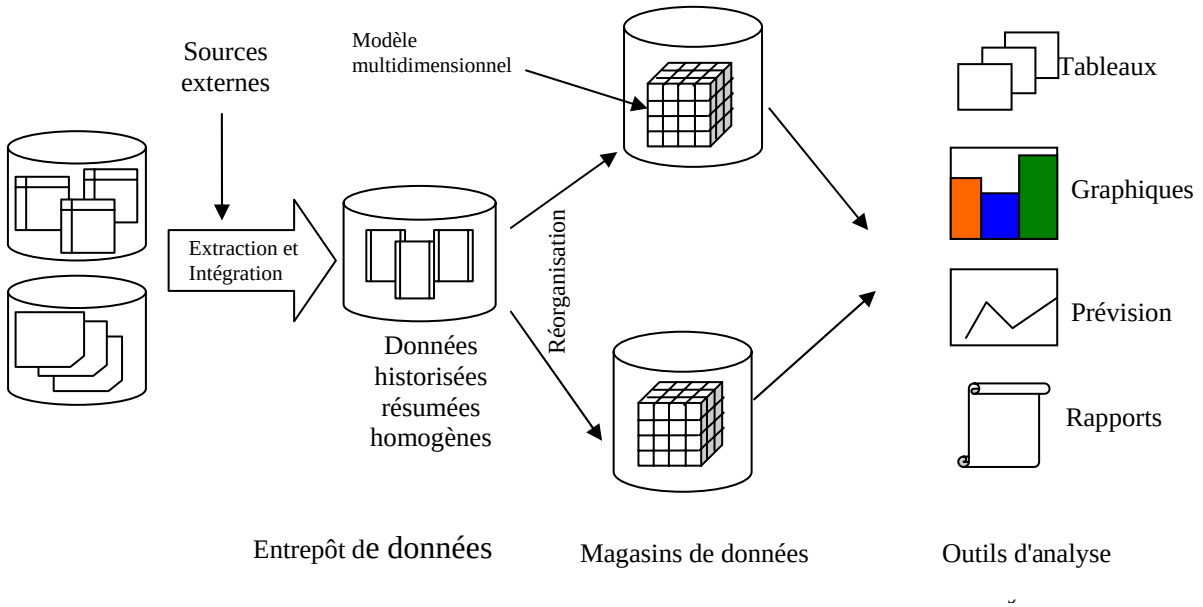


Figure 2 :L'entrepôt de données dans le système décisionnel.

De plus, l'entrepôt de données offre à l'entreprise les avantages suivants[13] :

- Il constitue une collection de données centralisée disponible pour l'aide à la décision (OLAP, datamining,….).
- Les évolutions des données de l'entrepôt sont conservées (historisation des données).

- Il contient un ensemble de données consolidées (données homogènes et fiables).
- Il contient des données agrégées permettant une analyse à différent niveaux de détails.
- Il permet de développer différents thèmes d'analyse (réorganisation en fonction des sujets à analyser).

4.1 L'intégrité de données

Un entrepôt de données est un projet d'entreprise. Par exemple dans la distribution, le même indicateur de chiffre d'affaires intéressera autant les forces de vente que le département financier ou les acheteurs. Pour y parvenir, les données doivent êtres intégrées.

Avant d'être intégrées dans l'entrepôt de données, les données doivent êtres mises en forme et unifiées afin d'avoir un état cohérent. Par exemple, la consolidation de l'ensemble des informations concernant un client donné est nécessaire pour donner une vue homogène de ce client. Une donnée doit avoir une description et un codage unique. Cette phase d'intégration est très complexe.

4.2 L'orientation sujet

L'entrepôt de données est organisé autour des sujets majeurs de l'entreprise, contrairement aux données des systèmes de production. Ceux-ci sont généralement organisés par processus fonctionnels. Les données sont structurées par thème.

L'intégration dans une structure unique est indispensable car les informations communes à plusieurs sujets ne doivent pas être dupliquées. Dans la pratique, une structure supplémentaire appelée Data Mart (magasin de données) peut être créée pour supporter l'orientation sujet.

On peut définir le magasin de données comme suite**.**

4.2.1 Le magasin de données

Le magasin de données est un extrait de l'entrepôt. Les données extraites sont adaptées à une classe de décideurs ou à un usage particulier (recherche de corrélation, logiciel de statistiques,...). L'organisation des données suit un modèle spécifique qui facilite les traitements décisionnels [10].

4.3 L'historisation de Données

Dans un système de production ; la donnée est mise à jour à chaque nouvelle transaction. Dans un entrepôt de données, la donnée ne doit jamais être mise à jour. Un référentiel temps doit être associé à la donnée afin d'être capable d'identifier une valeur particulière dans le temps.

La prise en compte de l'évolution des données est essentielle pour la prise De décision qui par exemple utilise des techniques de prédiction en s'appuyant sur les évolutions passées pour prévoir les évolutions futures. L'aspect d'historisation sera détaillé dans la section 7.

5. Modélisation de l'entrepôt de données

Les modèles utilisés pour la conception, des entrepôts de données reposant sur une conception multidimensionnelle des données. C'est à dire que les objets à modéliser sont décrits par plusieurs attributs différents; indépendants à priori; qui permettent chacun en particulier de pouvoir classifier et/ou ordonner ces objets. Prenons par exemple un entrepôt de données d'une entreprise commerciale multi-sites. Les information pertinentes concernant les informations sur les sites. Sur les produit, sur le temps calendaire. Le vocabulaire utilisé est alors la dimension: dimension produit ; dimension temporelle, dimension site chaque dimension possède sa propre table. Les tables de dimension sont regroupées autours d'une table contenant ces informations (mesures). On parle de schéma en étoile.

Il est aussi possible d'utiliser un schéma en flocon ou en constellation. La modélisation multidimensionnelle sera exposée dans le chapitre 3.

6. L'extraction des données

L'extraction des données désigne la collection des données utiles de la base de production. Le mécanisme d'extraction repose essentiellement sur la technique des vues matérialisées. Une vue matérialisée consiste à calculer une vue exprimée sur une source de données et à stocker physiquement les données obtenues dans l'entrepôt [10].

6.1 Les vues matérialisées

Une vue est une requête nommée. Une vue matérialisée est une table contenant les résultats d'une requête. Les vues améliorent l'exécution des requêtes en précalculant les opérations les plus coûteuses comme la jointure et l'agrégation, et en stockant leurs résultats dans la base. En conséquence, certaines requêtes nécessitent seulement l'accès aux vues matérialises et sont ainsi exécutées plus rapidement. Les vues dans le contexte OLTP ont été largement utilisées pour répondre a plusieurs rôles : la sécurité, la confidentialité, l'intégrité référentielle, etc**.** les vue sont utilisés pour construire les entrepôts de données et pour défini les données qui serons copiées dans l'entrepôt [18][19].

Le modèle le plus approprié pour représenter et manipuler des données dans le contexte d'analyse en ligne (OLAP) est le modèle multidimensionnel. Dans la plupart des projets de recherche sur les entrepôts de données, le modèle de représentation logique des données est le modèle relationnel. Le modèle conceptuel peut être un modèle en étoile ou en flocons qui reflète mieux la représentation multidimensionnelle des données. Au niveau logique, le schéma de l'entrepôt est défini par un ensemble de vues relationnelles matérialisées. Ces vues sont définies par des requêtes de type SPJ (Sélection

Projection Jointure) avec éventuellement des agrégat. Elles sont par la suite matérialisées. Une vue matérialisée est une vue dont les données sont calculées et stockées d'une manière redondante par rapport aux données sources [13].

6.1.1 Problème lies aux vues matérialisées

Noms avons vu que les vues matérialisées sont utilisées pour extraire les données dans l'entrepôt. Ces vues sont dérivées à partir de table de la base source. Les données de table sources sont changées et évoluer dans le temps à cause des mise à jour (base transactionnelle), alors que les vue dérivées de ces tables restent inchangé, c'est à dire les données de l'entrepôt ne sont pas cohérentes avec celles des sources. Ce problème est appelé le problème de maintenance des vues matérialisées.

6.1.1.1 Problème de maintenance

Trois stratégies fondamentales ont été proposées. Dans la première stratégie, les vues sont mises à jour périodiquement. Dans la deuxième stratégie. Les vues sont mises à jour immédiatement à la fin de chaque transaction. Dans la dernière stratégie, les modifications sont propagées d'une manière différée. Dans ce cas, une vue est mise à jour uniquement au moment où elle est utilisée par une requête d'un utilisateur.

La maintenance donc consiste à recalculer le contenu des vues matérialisées. Cette approche est inefficace (très coûteuse).

Une bonne solution est propager d'une manière incrémentale les changements sur les vues matérialisées. Seules les vues effectuées par les mises à jour sont modifiées, sans recalculer complètement leurs contenus.

6.1.1.2 Problème de sélection

Les vues matérialisées améliorent l'exécution des requêtes, mais il faut remarquer leur coût de maintenance, donc la matérialisation de toutes les vues est inefficace, on parle de problème de sélection des vues matérialisées.

Plusieurs travaux traitent ce problème, l'objectif est de sélectionner un ensemble des vues matérialisées en optimisant le coût de maintenance et le coût d'exécution de requêtes [13].

7 L'historisation

Les entrepôts de données améliorent le processus de prise de décision dans les entreprises. Ce processus utilise systématiquement les données actuelles de l'entreprise ainsi que "l'histoire" de ces données. Par exemple, pour prévoir le chiffre d'affaire d'une entreprise, les décideurs s'appuient sur les résultats passés pour se projeter dans l'avenir via des techniques (régression linéaire…). Conserver les données passées afin de mieux appréhender le présent et anticiper le futur est une caractéristique majeure des entrepôts de données[13], le terme d'historisation des données est employé.

Les données sources sont issues essentiellement des systèmes opérationnels (OLTP) de l'entreprise, et éventuellement des partenaires de l'entreprise, des serveurs du réseau Internet… Ces données sont peu ou pas historisées :

- certaines évolutions ne sont pas conservées (par exemple, l'ancienne adresse d'un fournisseur n'est pas conservées dans les bases de production).
- la période de conservation des données temporelles (données associées à une date) est courte, de l'ordre de quelques mois alors que la prise de décision réclame une conservation des évolutions sur des périodes de temps plus importantes, de l'ordre de plusieurs années (par exemple, les ventes effectuées dans un magasin sont conservées de 6 mois à 2 ans maximum dans les bases de production).
- la représentation du temps est souvent minimale, les dates sont stockées sous forme d'un simple attribut.

L'entrepôt de données doit palier ces insuffisances en terme de conservation des évolutions pour servir de support efficace aux techniques d'analyse de

type On-Line Analytical Processing (OLAP) utilisées lors de la prise de décision[13].

7.1 Manipulation de temps

Dans le contexte des entrepôts de données, de nombreux problèmes liés à la représentation du temps se posent, et il est nécessaire de concevoir des moyens permettant de manipuler simplement le temps.

Plusieurs temps orthogonaux coexistent : le temps de validité, le temps d'entreposage, la série temporelle des temps d'extraction.

- Le temps de validité correspond au temps pendant lequel l'information est considérée valide dans la réalité.
- Le temps d'entreposage est le temps pendant lequel l'information est active dans la réalité.
- La série temporelle des temps d'extraction correspond à la série des temps de demande d'information de l'entrepôt (mise à jour).

Ce dernier temps est utile plus particulièrement dans la gestion de la sécurité de l'entrepôt. En effet, l'entrepôt de données constitue un élément très sensible pour les entreprises puisqu'il contient des informations confidentielles vis de l'extérieur. Ainsi, il est important de prévoir des mécanismes de sécurité permettant par exemple de savoir quelles données ont été interrogées et par quelles personnes. L'estampille temporelle relative au temps d'extraction doit comporter d'autres informations permettant de savoir par exemple qui a extrait l'information à cette date.

7.2 L'évolution de données

La matérialisation des données dans l'entrepôt permet de conserver dans l'entrepôt les évolutions des données. Lorsque les données sources évoluent, il est alors nécessaire de répercuter ces modifications dans l'entrepôt de données tout en conservant les anciennes valeurs[13].

Un premier problème posé concerne la pertinence des évolutions: toutes les évolutions ne sont pas nécessairement utiles pour la prise de décision (par exemple, si une modification du nom d'un client survient pour corriger une erreur de saisie, l'ancienne valeur fausse ne doit pas être stockée dans l'entrepôt). De plus, les évolutions des sources ne doivent pas (ou ne peuvent pas dans le cas d'un accès limité) forcément être répercutées avec la même fréquence. Il est donc nécessaire d'offrir des moyens permettant à l'administrateur d'adapter le processus de rafraîchissement suivant les parties de l'entrepôt.

7.3 Le rafraîchissement des données

Le rafraîchissement des données de l'entrepôt est abordé suivant deux méthodes spécifiques : le rafraîchissement périodique et le rafraîchissement incrémental[13].

- L'approche de rafraîchissement périodique consiste à extraire les modifications survenues à la sources suivant une période de temps définie par l'administrateur. Le processus de rafraîchissement propage les modifications de la source dans l'entrepôt périodiquement. Cette approche fait appel à des techniques de comparaison de fichier ou de comparaison de dates de mise à jour des données.
- L'approche de rafraîchissement incrémental consiste à répercuter dans l'entrepôt chaque modification de manière immédiate ou quasi immédiate. Cette approche utilise soit des déclencheurs qui repèrent les modifications à la source, soit le journal de la source où les changements intéressants sont recherchés.

Ces deux approches se heurtent au coût de rafraîchissement qui doit être minimisé pour garantir les meilleures performances possibles pour l'entrepôt en terme de temps de réponse pour l'interrogation et l'analyse. Le processus de rafraîchissement a un coût qui est en conflit avec les besoins de temps de réponse rapide.

Enfin, un problème crucial est celui du volume des données stockées dans l'entrepôt lié en particulier à l'historisation des données. Ce problème est d'autant plus accru que l'entrepôt conserve l'évolution des données sur des périodes de temps importantes (plusieurs années).

8 L'agrégation

L'agrégation est le problème majeur lié à la construction des entrepôts de données car elle permet de résumer les données opérationnelles élémentaires à un niveau de détail plus élevé ; ce niveau plus élevé devient le niveau élémentaire dans l'entrepôt. L'agrégation est utilisée afin de définir la granularité des données stockées dans l'entrepôt.

Le choix de la granularité influence de manière significative la pertinence des données contenues dans l'entrepôt, et dans le même temps, affecte le type des requêtes possibles. Une grande granularité limite les requêtes possibles, en particulier les niveaux de détails fins ne sont pas disponibles. Inversement, un niveau de détail fin dans l'entrepôt a pour conséquence une faible granularité.

Plusieurs niveaux de granularité peuvent être présents dans l'entrepôt. Les données récentes sont stockées généralement avec un niveau de détail très fins (par exemple, le détail des ventes d'un magasin sont conservées avec un niveau de détail journalier, durant plusieurs mois), tandis que les données plus anciennes sont agrégées à un niveau de granularité plus grand (par exemple, les ventes du magasin sont résumées au bout d'un an avec un niveau de détail mensuel). Enfin, les données très anciennes peuvent être archivées avec un taux d'agrégation très fort (par exemple, les ventes du magasin peuvent être réduite à l'année). Il faut remarquer que cette opération d'agrégation provoque une perte d'information. Le chois de la granularité consiste à déterminer la perte acceptable suivant des contraintes de volume de données (c'est-à-dire de coût de stockage pour l'entreprise).

Une autre utilisation importante des agrégations est celle de l'amélioration des performances de l'entrepôt lors de l'interrogation. La méthode utilisée consiste à déterminer des pré-agrégations [13]; des algorithmes permettent de déterminer une politique optimale permettant de déterminer quels agrégats doivent être calculés. Lorsque les requêtes sont posées, les agrégats pré-calculés sont utilisés pour améliorer les temps de réponses. Là encore, le besoin de haute performance pour l'interrogation est en conflit avec la contrainte d'un faible coût de maintenance de l'entrepôt, ainsi que la contrainte du volume des données stockées.

Pour finir, il est essentiel de comprendre que dans les entrepôts de données l'agrégation n'induit pas une réduction du volume des données, au contraire on parle du paradoxe de la volumétrie. Plus on monte le niveau de granularité plus le volume des informations à stocker est grand[13]. En effet, l'information de synthèse est plus volumineuse que celle de détail.

9 Synthèse

Les **entrepôts** de données sont nés pour répondre aux besoins de décideurs. Ils regroupent dans un format homogène les données provenant généralement de plusieurs sources de production pouvant être réparties et avoir des format varies.

La conception et la mise en œuvre d'entrepôt de données ont plusieurs difficultés tels que : L'**extraction**, l'i**ntégration** et l'**historisation** de données.

L'entrepôt de données (magasins de données) repose sur un modèle spécifique plus souple et plus adapté que le modèle relationnel pour supporter l'analyse et l'aide à la décision, on parle de **modèle multidimensionnel**. Ce modèle sera présenté en détail dans le chapitre suivant.

CHAPITRE III :

La modélisation multidimensionnelle

1. Introduction

Il existe deux façons d'analyser, soit par reporting, soit par une analyse on-line (c'est ce qui nous intéresse ici). Une société désire construire une base de données pour suivre l'évolution de ses ventes par mois et par catégories. Elle dispose d'une table qui contient les ventes.

Mois	Catégorie	Quantité	Montant
Janvier 2000	Alimentation	15	150.00
Janvier 2000	Nettoyage	10	200.00
…	…	…	…

Pour analyser ces données, on peut par exemple placer les moins en colonne et les catégories en ligne.

Mois / Catégorie	Nettoyage	Alimentation	Vêtement	Papeterie
Janvier	200.00	150.00	180.00	50.00
Février	180.00	140.00	250.00	55.00
Mars	220.00	155.00	100.00	48.00

Tableau 2 : Tableau bidimensionnel.

Il s'agit ici d'un tableau à deux dimension et non d'une table !

Si l'on veut compare les ventes par catégories, par région et pour plusieurs années, nous considérons plusieurs tables, relatives aux ventes de chaque année par exemple entre 2004 et 2006, On peut alors observer les données dans un espace à trois dimensions : la dimension géographie, la dimension produits et la dimension temps. Chaque intersection de ces dimensions représente une cellule comportant le total des ventes. On parle alors de cube de données (un cube n'a pas forcément 3 dimensions).

La figure 2 montre la représentation multidimensionnelle (cube de données) permettant l'analyse des ventes pour les derniers trois années, par région et par catégorie.

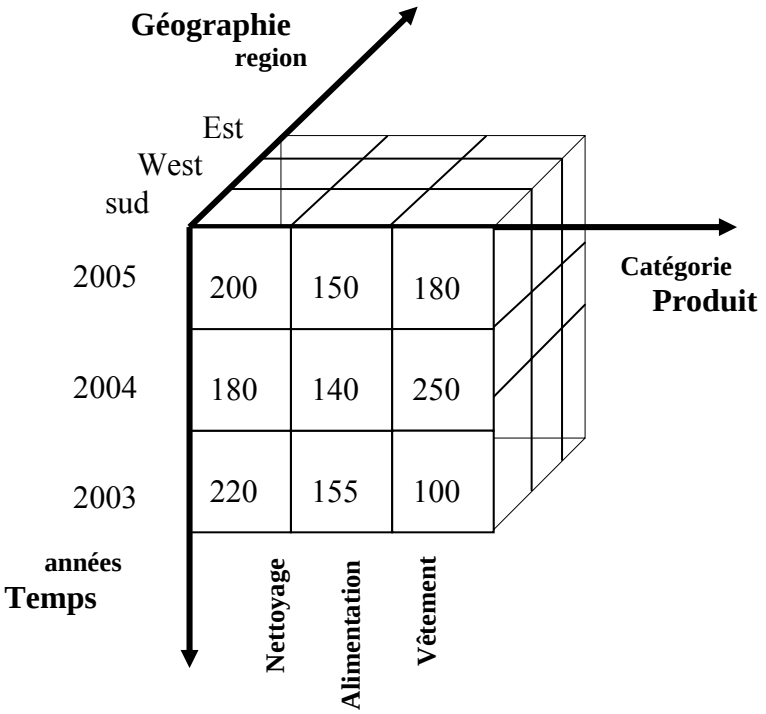


Figure 2 : Représentation multidimensionnelle

2. Modélisation multidimensionnelle

Les analyses décisionnelles sont basées sur des traitements OLAP directement liés à la modélisation de l'information sous une forme conceptuelle proche de la perception qu'en a l'analyste. Cette perception de l'information est basée sur une vision multidimensionnelle des données. Cette modélisation multidimensionnelle consiste à considérer un sujet analysé comme un point dans un espace à plusieurs dimensions (axes d'analyses). Les données sont organisées de manière à mettre en évidence le sujet analysé et les différentes perspectives de l'analyse [10].

Dans notre exemple le sujet analysé est le vente qui prend comme mesure le montant des ventes et les dimensions (axes d'analyses) sont : la dimension Temps, la dimension Géographie et la dimension Produit.

2.1 Modélisation conceptuelle

La modélisation multidimensionnelle repose sur le concept de fait et de dimension [1].

2.1.1 Le fait

Le fait représente le sujet analysé qui est formé de mesures d'activité analysée, les mesures sont généralement numériques pour permettre de résumer un grand nombre d'enregistrements (les additionner, calculer le minimum, le maximum,…). Ils sont aussi additives ou semi-additives pour pouvoir les combiner au moyen d'opérateurs arithmétiques.

Exemple : le fait de Vente de notre exemple peut être constitué des mesures d'activités suivantes : quantité de produits vendus et montant des ventes. On représente le fait par un rectangle qui contient le nom du fait et les mesures.

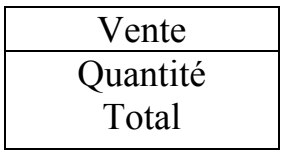


Figure 3 : Exemple de fait.

2.1.2 La dimension

Le fait est analysé suivant différentes perspectives. Ces perspectives correspondent à une catégorie utilisée pour caractériser les mesures d'activité analysées [5] ; c'est-à-dire les dimensions. Une dimension se compose de paramètres (attributs). Chaque paramètre peut prendre différentes valeurs faisant varier les mesures de l'activité.

Exemple: Poursuivons l'exemple précèdent, les dimension sont : Temps, Produit et Géographie. La dimension temps peut prendre plusieurs paramètres par exemple jour, mois, trimestre et année. La dimension géographie peut être composée en région, ville et magasin. Dans le modèle multidimensionnel les paramètres sont essentiellement discrets et leurs valeurs sont bien déterminées.

Comme le fait, la dimension est représentée par un rectangle contient le nom de la dimension et ses paramètres.

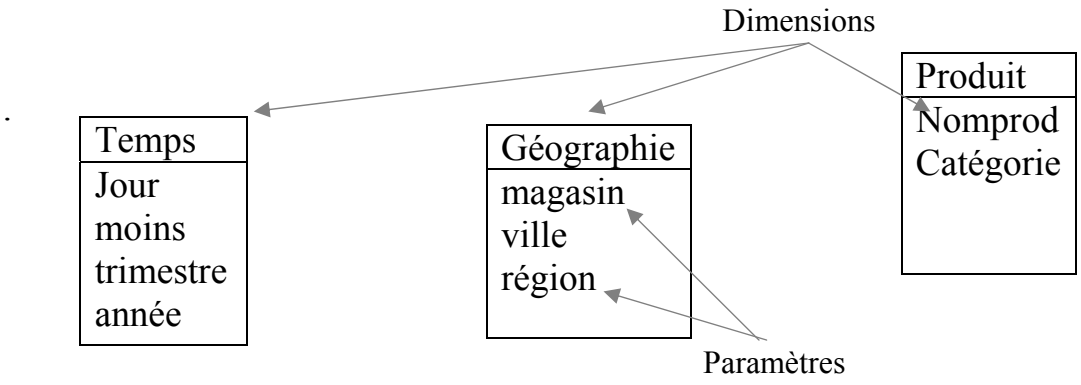


Figure 4 : Exemple de dimensions.

2.1.2.1 Notion de hiérarchie

Lors du processus OLAP, les paramètres d'un dimension sont organisés en une ou plusieurs **hiérarchie(s)** pour permettre d'analyser les données en partant d'un faible niveau de détail vers des données plus détaillées "forage vers le bas", ou réciproquement d'un niveau détaillé vers un niveau supérieur " forage vers le haut ". La hiérarchie spécifie une relation d'ordre entre les paramètres.

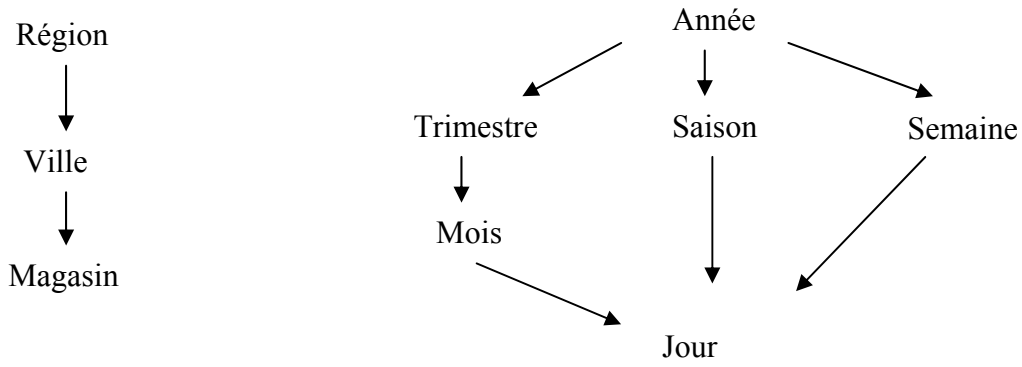
Exemple: pour la dimension Géographie Les paramètres des dimensions sont organisés suivant une hiérarchie :

Magasin → ville → Région

Ainsi chaque magasin appartient à une ville qui est situé dans une région.

2.1.2.1.1 Hiérarchie simple ou multiple

Dans le cas d'une hiérarchie simple; chaque paramètre n'a qu'un seul paramètre supérieur. Dans le cas d'hiérarchie multiple, on a différents chemins pour créer une hiérarchie.



Hiérarchie simple Hiérarchie multiple

Figure 5 : type d'hiérarchies.

2.2 Modèles en étoile

Il est possible d'établir un model multidimensionnel simple à partir du fait et des dimensions. Ce modèle est constitué du fait central et des dimensions représentant visuellement une étoile, on parle de modèle en étoile (star schema [1]).

Exemple: On peut modéliser les analyses des quantités et des montants des ventes selon trois dimensions : le Temps, la Catégorie et la situation Géographique par le schéma en étoile décrit dans La Figure 6. On utilise la lettre **D** pour les dimensions et la lettre **F** pour le fait.

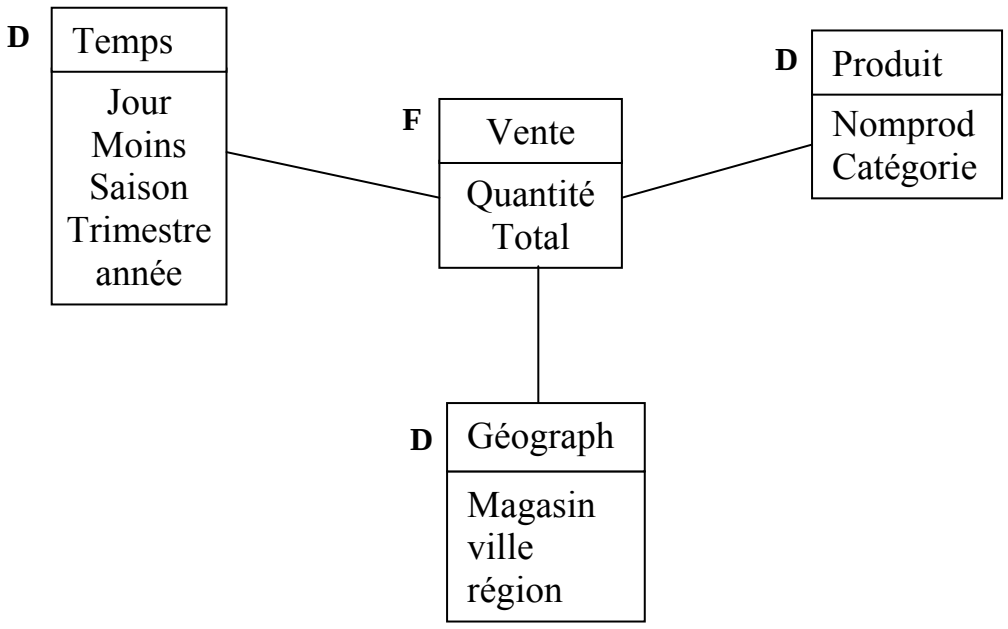


Figure 6 : Modélisation en étoile.

2.2 Modèles en flocon

Il existe d'autres types de modélisation multidimensionnelle, la modélisation en **flocon** (snowflake) consiste à décomposer les dimensions du modèle en étoile en sous hiérarchies. La modélisation en flocon est donc une extension de la modélisation en étoile ; le fait est conservé et les dimensions sont éclatées selon leurs hiérarchies des paramètres.

Cette modélisation permet de formaliser une hiérarchie au sein d'une dimension.

Exemple: La Figure 7 illustre la modélisation en flocon en normalisant les dimensions du modèle en étoile de la Figure 6.

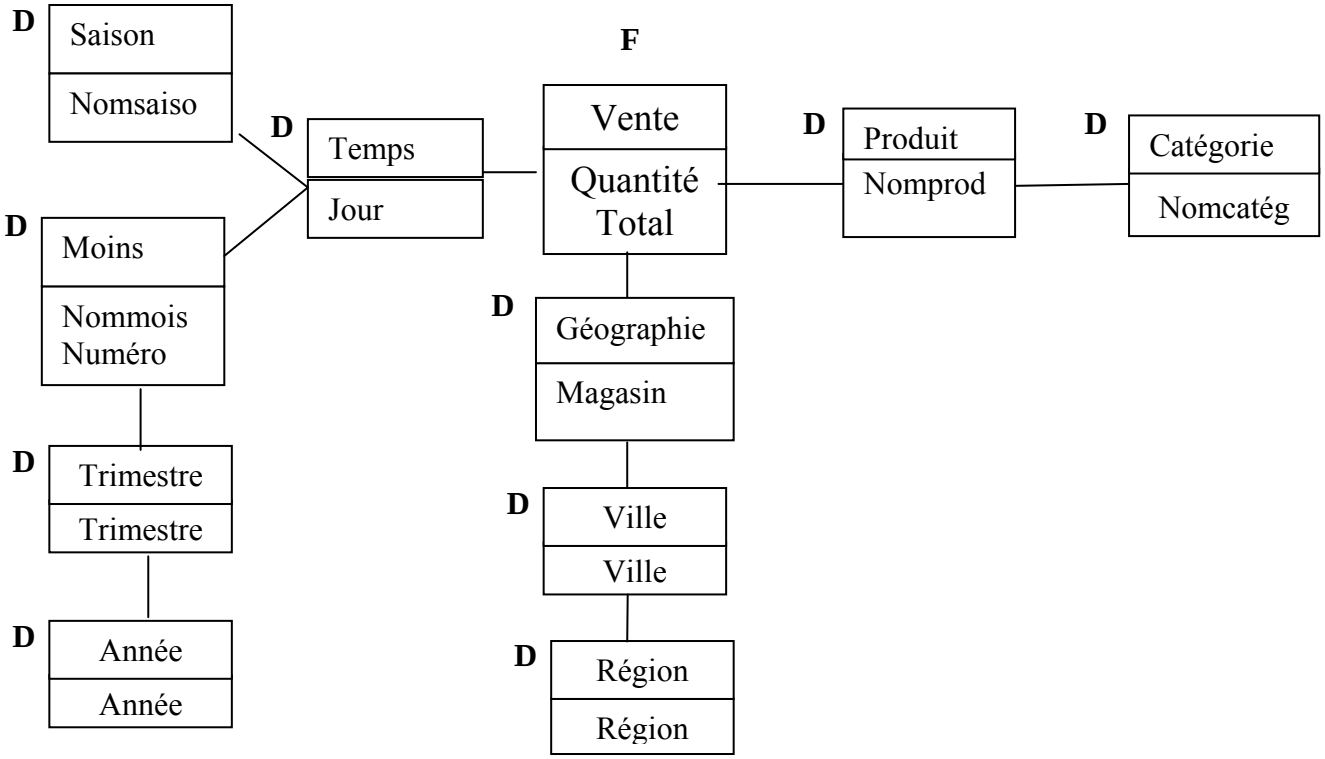


Figure 7 : Modélisation en flocon.

2.3 Modèles en constellation

Une autre technique de modélisation, issue du modèle en étoile, est la modélisation en constellation. Il s'agit de fusionner plusieurs modèles en

étoile qui utilisent des dimensions communes. Un modèle en constellation comprend donc plusieurs faits et des dimensions communes ou non.

Exemple : La Figure 8 illustre la modélisation en constellation de deux schémas en étoile : l'un analyse les ventes et l'autre correspond aux fournisseurs.

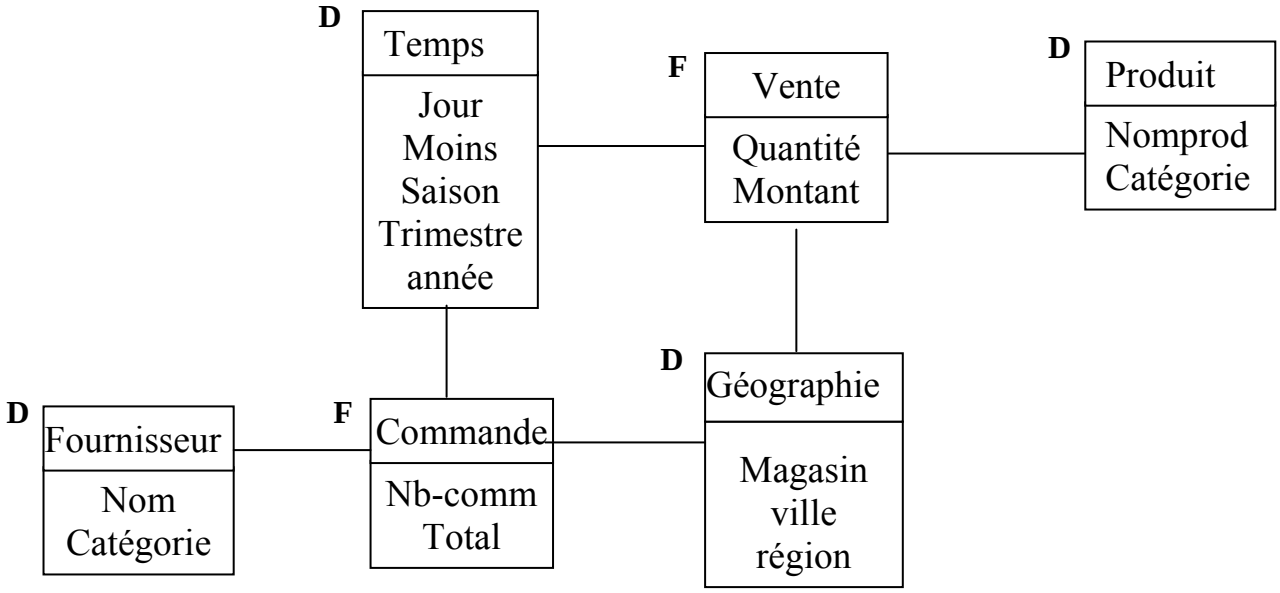


Figure 8 : Modélisation en constellation.

3. Les opérations multidimensionnelles

Les opérations multidimensionnelles permettent de manipuler et visualiser les données multidimensionnelles. Elles offrent la possibilité de la navigation dans les données selon les axes d'analyse, elles consistent aussi à faire pivoter le cube, le couper en tranches, interchanger ou combiner les paramètres et les mesures.

Dans les sous-sections suivantes, nous décrivons les opérations attachées à la manipulation des données multidimensionnelles.

3.1 Opérations classiques

On peut citer parmi les opérations classiques (issues de l'algèbre relationnelle) les opérations suivantes :

- le **produit cartésien** et la **jointure**.
- les opérations ensemblistes d'**union**, d'**intersection** et de **différence**.
- la **suppression**, l'**ajout** d'une mesure.
-le **renommage** de mesures. De paramètres,...

3.1.2 Opérations de sélection et projection

L'opération slice and dice (couper des tranches) permet de sélectionner un sous-cube. Lors l'opération Slice on spécifie une valeur pour un paramètre c'est-à-dire une position dans une dimension d'un cube, tandis que L'opération dice on spécifie des positions sur une ou plusieurs dimensions, dans la figure ci-dessous, on représente les deux opérations slice et dice séparément. L'opération slice permet d'obtenir les montant des vents dans le premier trimestre.

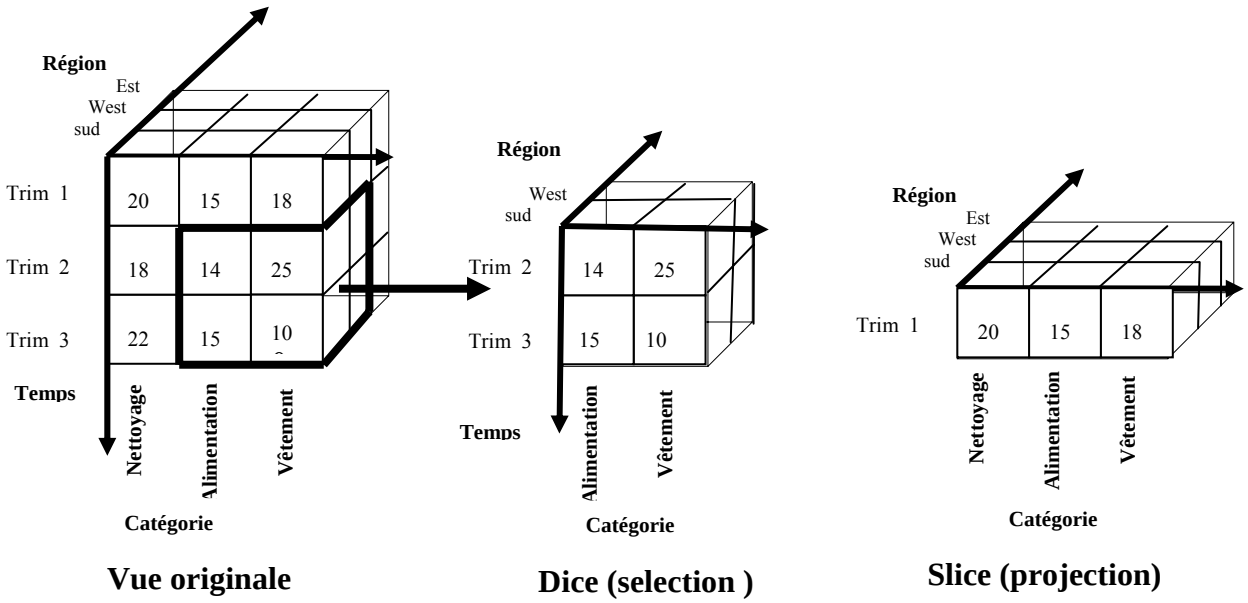


Figure 9 : Opérations de sélection et projection

3.2 Opérations associées à la granularité

Les opérations agissant sur la granularité autorisent l'observation des données selon les hiérarchies associées aux dimensions et permettent

la navigation dans l'hypercube du niveau global vers le niveau détaillé, et inversement.

Le forage vers le haut (drill-up ou roll-up ou scale-up) consiste à représenter les données de l'hypercube à un niveau de granularité supérieur conformément à la hiérarchie définie sur la dimension. Une fonction d'agrégation (somme, moyenne…) en paramètre de l'opération indique comment sont calculées les valeurs du niveau supérieur à partir de celles du niveau inférieur.

- **Le forage vers le bas** (drill-down ou roll-down ou scale-down) consiste à représenter les données de l'hypercube à un niveau de granularité inférieur, donc sous une forme plus détaillée.

Exemple : la figure 10 illustre le mécanisme de roll-up sur la dimension "Temps" qui permet d'aller du niveau "trimestre" vers le niveau "année" avec la fonction d'agrégation"Somme", et le mécanisme de drill-down sur la dimension " Géographique " en partant du niveau "région" vers un niveau plus détaillé "ville".

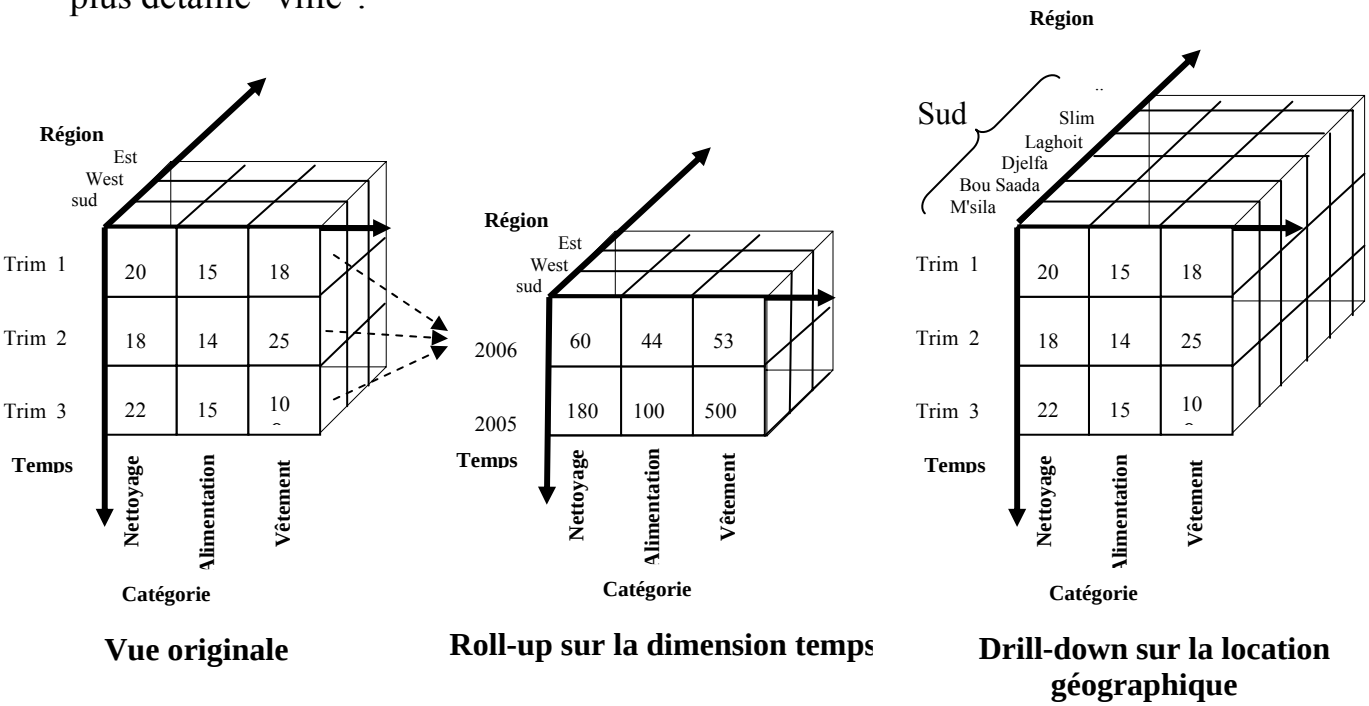


Figure 10: Operations Roll-up et Drill-down.

3.3 Opérations agissant sur la structure

Les opérations agissant sur la structure multidimensionnelle visent à changer le point de vue des données observées.

3.3.1 La rotation

La rotation (rotate ou pivot) consiste à effectuer à un hypercube une rotation pour présenter une face différente. Par exemple on peut analyser les ventes par catégorie de produits pour une période de temps donnée, une simple permutation d'axes permet de l'analyse par région et par période, comme montre la figure 10.

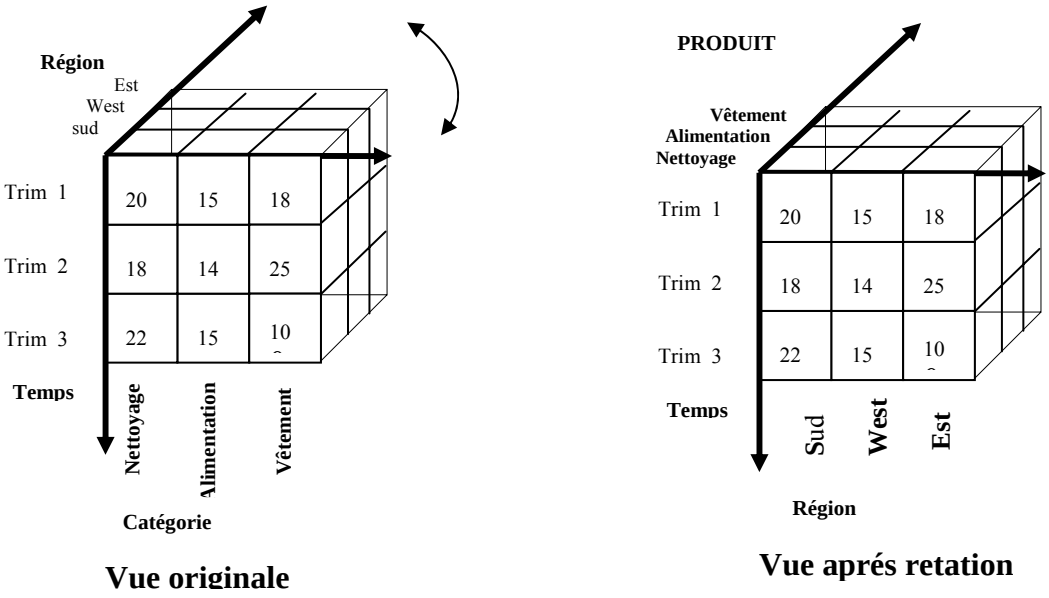


Figure 11 : la rotation d'axes d'analyses.

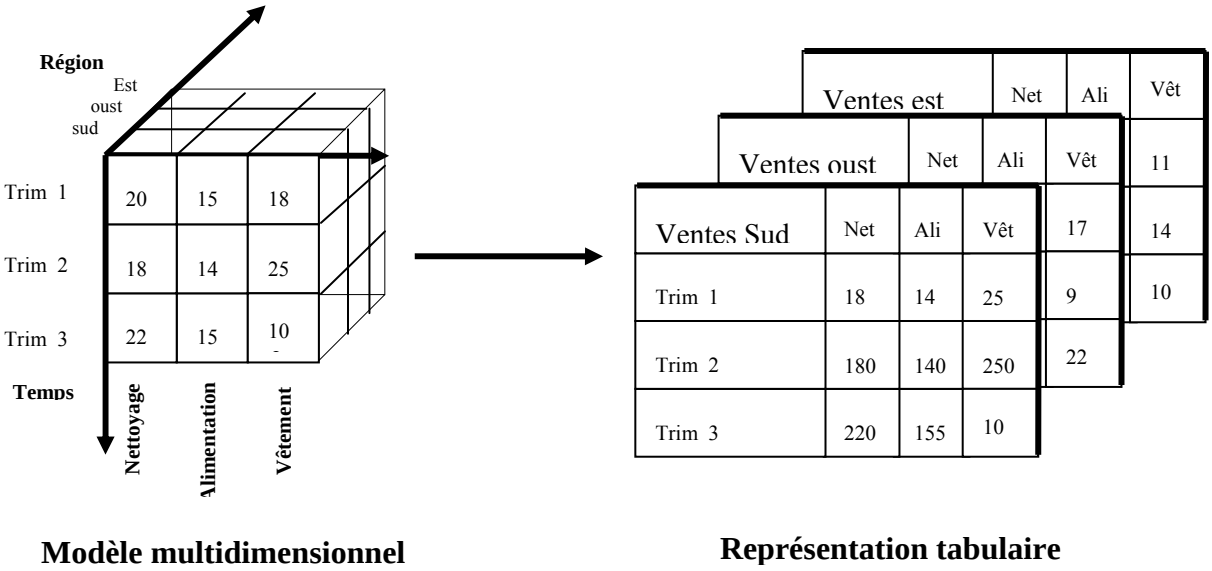
3.3.2 La permutation

La permutation (switch) consiste à inverser deux positions d'un paramètre d'une dimension données. La figure 12 illustre la permutation de positions du paramètre région de la dimension Géographie.

Figure12: Opération de permutation réorganise les régions.

3.3.3 La division

La division (split) consiste à représenter chaque tranche de l'hypercube de passer d'une représentation multidimensionnelle (n-dimensions) à une représentation tabulaire (2-dimensions). Sa généralisation permet de découper un hypercube en cubes (réduire le nombre de dimensions).



Modèle multidimensionnel **Représentation tabulaire**

Figure13: Représentation tabulaire avec l'Opération Split.

3.3.4 L'emboîtement

L'opération d'emboîtement (nest) permet d'imbriquer les positions (valeurs) d'un paramètre d'une dimension avec un autre paramètre. L'intérêt de cette opération est qu'elle permet de représenter de manière bidimensionnelle toutes les informations de l'hypercube quelque soit le nombre de dimensions.

3.3.5 L'enfoncement

L'enfoncement (push) consiste à combiner les positions (valeurs) d'un paramètre d'une dimension aux mesures du fait et donc de transformer un paramètre en mesure. L'opération inverse de **retrait** (pull) permet de transformer une mesure en paramètre en changeant le statut de certaines mesures de l'hypercube pour constituer une nouvelle dimension.

3.3.6 La factualisation

La factualisation (fold) consiste à transformer une dimension en mesure(s) ; cette opération permet de transformer en mesure l'ensemble des paramètres d'une dimension. La **paramétrisation** (unfold) permet de transformer une mesure en paramètre dans une nouvelle dimension.

3.3.7 Cube

L'opération **Cube** permet de calculer des sous-totaux et un total final dans l'hypercube.

Le tableau suivant présente quelques travaux et les opérations associées **[**10**].**

		Li, Wang [22]	**Cabbibo, Tyrone [21]**	**Gyssen, Lakshmanan [23]**	**Agrawal, et al [24]**	**Marcel [5]**	**Lehner, et al [25] Lehner [26]**	**Pedersen,Jensen [27]**	**Mendelzon, Vaisman[28]**
opérations	Renommage	x	x	x		x		x	
	Ajout de mesures		x			x			
	Suppression					x			
	Sélection		x	x	x	x	x	x	x
	Jointure	x	x	x	x			x	x
	Union,Différence Intersection	x	x	x	x			x	
	Agrégation	x	x	x	x		x	x	x
	Roll-up	x	x	x	x	x	x	x	x
	Drill-down						x	x	x
	Cube			x		x			
	Switch			x					
	Push				x	x			
	Pull				x	x			
	Fold		x	x					
	Unfold		x	x					
	Rotate					x			
	Nest					x			
	Split			x		x			x

Le tableau 3: Quelques travaux et les opérations associées

4. Les implémentations des modèles multidimensionnels

Selon la façon dont le cube de données (nous utilisons le terme cube pour désigner des structures à deux, trois ou à plus de trois dimensions) est stocké et géré (dans un magasin), il existe deux approches fondamentales. L'approche MOLAP (Multidimensional OLAP) utilise des tableaux multidimensionnels (multi-dimensional array). Par contre, l'approche ROLAP (Relational OLAP) utilise un SGBD relationnel pour implémenter les cubes de données.

4.1 Les systèmes MOLAP

Les systèmes MOLAP stockent les données dans un SGBD multidimensionnel pur sous la forme d'un tableau multidimensionnel (multi-dimensional array). Les dimensions de ce tableau sont associées aux dimensions du cube (voir Figure 14). Seules les valeurs de données de chaque cellule de cube sont stockées. Ces systèmes demandent un pré-calcul de toutes les agrégations possibles [12].

L'intérêt de ces systèmes est que les temps d'accès aux données sont meilleurs, mais la capacité de stockage est limité (quelques gigaoctets). De plus le modèle MOLAP nécessite de redéfinir des opérations qui manipulent ces structures de données. Mais, lorsque les données sont éparses, ces systèmes sont consommateurs d'espace [14].

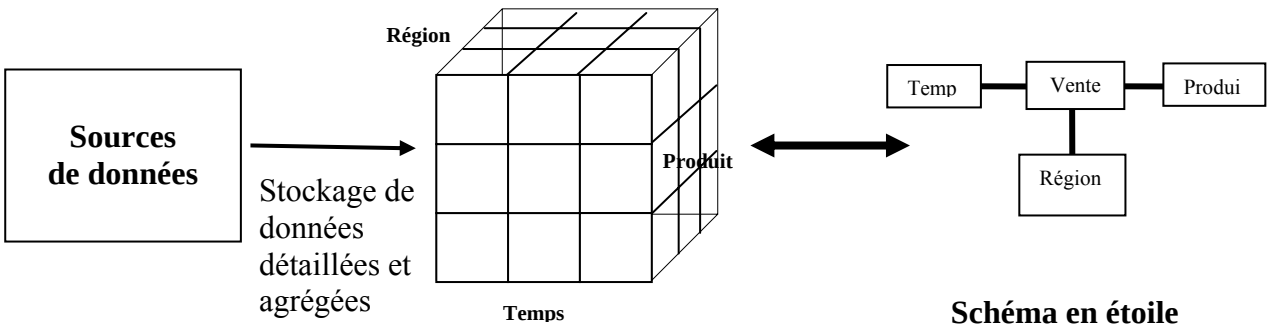


Figure 14 : système MOLAP.

4.2 Les systèmes ROLAP

L'approche ROLAP est le plus utilisé dans les système multidimensionnels, il stocke les données dans une base relationnelle mais la gère selon le concept multidimensionnel. Ce modèle s'intègre facilement dans un contexte de données de production relationnelles et permet de gérer de grande taille de données de l'ordre du téraoctet. Les analyses sont par contre plus lentes [6].
Le tableau suivant illustre une comparaison entre l'approche ROLAP et l'approche MOLAP [15].

	Avantages	Inconvénients
ROLAP	Technologie familière	Lent
	Scalable	
	Ouvert	
MOLAP	Modèle multidimensionnel	Technologie non prouvée
	Traitement de requête spécialisé	Non scalable
	Techniques d'indexation spécialisées	

Tableau 4 : ROLAP vs MOLAP**.**

4.2.1 Schéma en étoile

Dans ce type de schéma, chaque fait correspond à une table de fait (constituée d'attributs représentant les mesures), et chaque dimension correspond à une table de dimensions. La table des faits référence les tables de dimensions par une clé étrangère pour chacune d'elles et stocke les valeurs des mesures pour chaque combinaison de clés. Les tables de dimension contiennent les paramètres et une clé primaire permettant de réaliser des jointures avec la table de fait. Les tables de dimension sont généralement dénormalisées pour minimiser le nombre de jointures (meilleur temps de réponse).
Exemple : Nous considérons l'exemple du schéma en étoile de la Figure15. La table Vente correspond au fait et les tables Temps, Géographie et Catégorie représentent les dimensions:

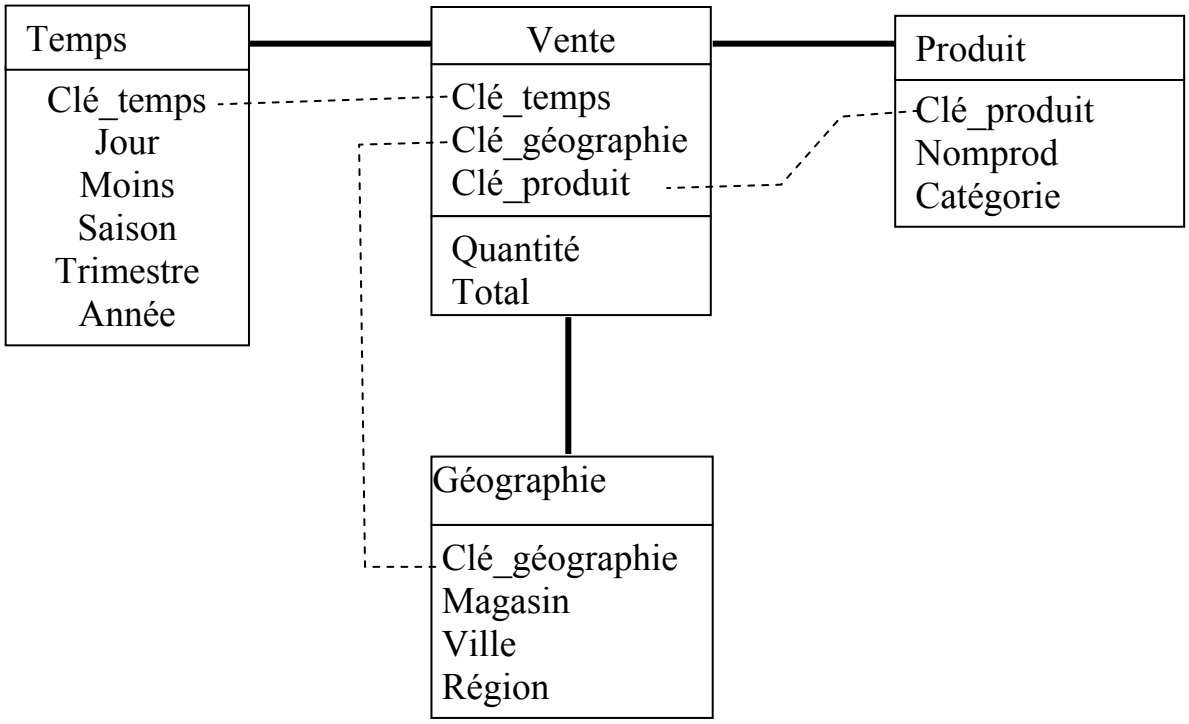


Figure 15: Schéma en étoile

Cependant, un schéma relationnel utilisé par les applications OLTP est généralement orienté vers les mises à jour, par opposition au schéma en étoile dédie aux interrogations (requête). Les requêtes typiques de ce schéma sont appelées les requêtes de jointure en étoile (star-join queries) qui ont les caractéristiques suivantes [12] :

1. Il y a des jointures multiples entre la table des faits et les tables de dimension.
2. Il n'y a pas de jointure entre les tables de dimensions.
3. Chaque table de dimension impliquée dans une opération de jointure a plusieurs prédicats de sélection sur ses attributs descriptifs.

4.2.2 Schéma en flocon

Le schéma en étoile ne reflète pas les hiérarchies associées à une dimension. Le schéma en flocon (voir la figure 7) a été propose pour résoudre ce problème. Il est une émanation du schéma en étoile. Il consiste à conserver la table des faits et à éclater les tables de dimensions pour permettre une

représentation plus explicite de la hiérarchie [7]. L'éclatement des tables de dimensions peut être considéré comme un normalisation de ces derniers.

4.2.3 La normalisation

La table de fait atteint naturellement un haut degré de normalisation, car elle est essentiellement constituée de clés étrangères. Tout effort de normalisation des tables de données dimensionnelles afin d'économiser de l'espace disque est une perte de temps. Les tables de dimensions ne doivent pas être normalisées, mais rester des table plates (éviter le flocon mais plutôt tenter d'obtenir un schéma en étoile). Des tables de dimension normalisées ne sont plus navigables. Les gains d'espace disque sont de l'ordre de 1% par rapport à l'ensemble du schéma.

4.3 Le modèle HOLAP

Le modèle HOLAP comme Hybride OLAP Est une croisement des deux concepts pour profiter des avantages de chacun : la performance en matière d'analyse de la base MOLAP et la capacité de stockage de ROLAP.
Pour plus grand efficacité, il suffit de stocker les données les fréquemment utilisées dans la base MOLAP et les 80% restant qui ne représentent que 20% des transactions décisionnelles, dans la base ROLAP[6].

Le tableau suivant montre quelques produits des systèmes OLAP.

Modèle	Caractéristiques	produits
ROLAP	Analyse plus lente Taille de stockage grande	Discover de Oracle
MOLAP	Efficacité d'analyse	Express de oracle Essbase de Hyperson
HOLAP	Croisement de deux Modèles	Holos de seagate SQL7 server de microsoft

Tableau 5 : Quelques Produits OLAP

La suite de ce chapitre présente un exemple d'un système MOLAP décrit par le groupe d'Evolution [13].

5. L'exemple du système oracle

5.1 Introduction

Les systèmes OLAP permettent d'analyser des données (les mesures) suivant différents axes d'analyse (les dimensions). Cette analyse se fait à différents niveaux de détail en étudiant des mesures plus ou moins agrégées. On peut par exemple analyser une mesure des ventes suivant les dimensions du produit, du temps et du magasin concernés par l'acte de vente.

Un des éléments principaux d'un système OLAP est le système de gestion de base de données multidimensionnelles (SGBDM). Oracle express est un SGBDM utilisé comme serveur de données par des applications client d'analyse multidimensionnelle. Les principaux éléments manipulés par oracle express sont : les dimensions, les variables, les relations, les formules.

Ces différents éléments peuvent être manipulés à l'aide d'un langage de commande (L4G) comportant des opérateurs multidimensionnels.

5.2 Les dimensions

Les dimensions sont les axes d'analyse qui structurent les données de la base express. Les occurrences de ces axes d'analyse sont appelées les positions. La figure suivante présente un exemple d'une dimension "Produit".

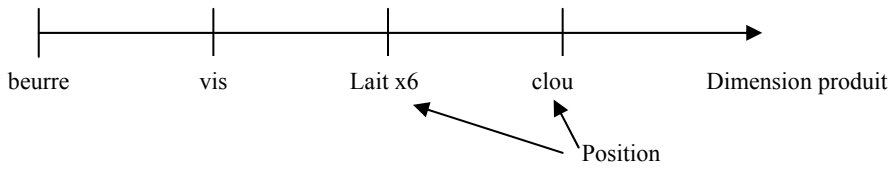


Figure 16: Exemple de dimension.

5.3 Les variables

Les variables sont des cellules de tableaux à n-dimensions. Elles correspondent aux mesures de l'analyse. Elle sont correspondent aux mesures de l'analyse. Elles sont structurées par les dimensions. Les valeurs des variables sont accessibles par les positions des dimensions. La figure suivante présente un exemple d'une variable "Chiffre Affaire" structurée suivant deux dimensions "Produit" et " Mois".

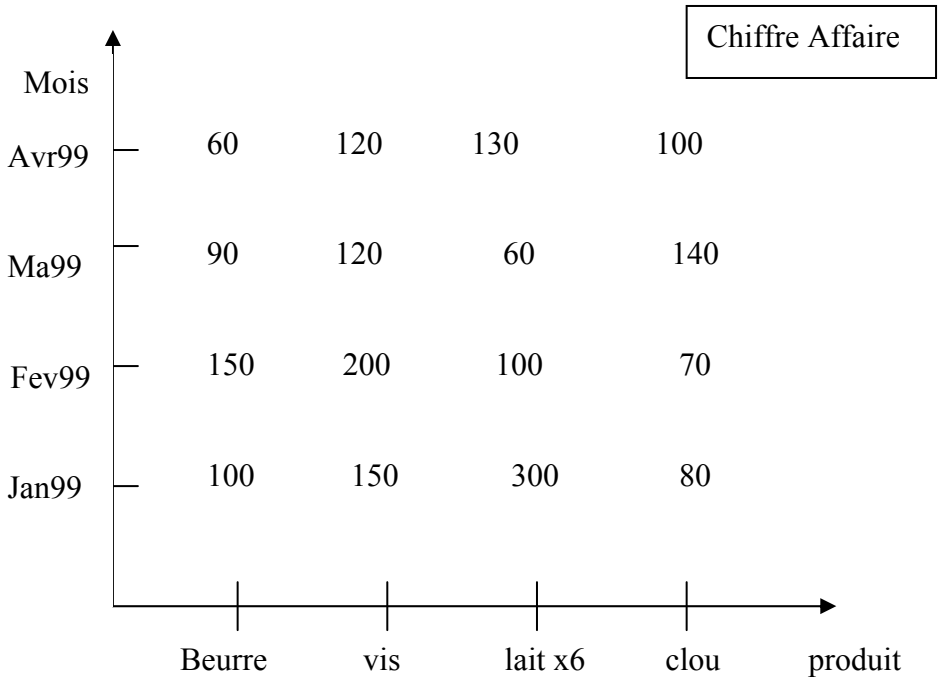
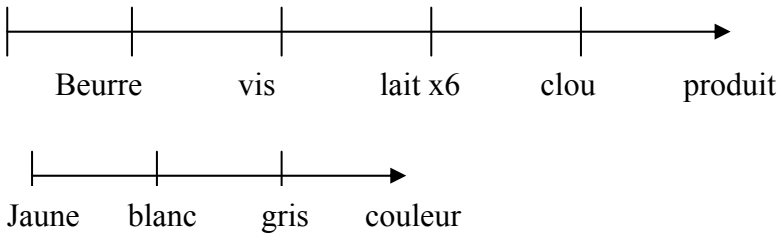


Figure 17: Exemple d'une variable "Chiffre Affaire "**.**

5.4 Les relations

Les relations permettent d'établir des correspondances entre les valeurs de dimensions distinctes ou d'une même dimension. La figure suivante présente un exemple d'une relation "ProdCoul" qui associe des valeurs des dimensions "Produit"et "Couleurs", voir la figure suivante.



produit	couleur
Beurre	jaune
Vis	Blanc
Lait x6	Blanc
clou	Gris

Figure 18: Exemple d'une relation "ProdCoul"**.**

Les relations sont utilisées pour structurer hiérarchiquement les valeurs d'une dimension. Chaque niveau de la hiérarchie présente un niveau de détail de l'analyse. Par example, une relation "ProduitClassseproduit" permet de hiérarchiser les produits et les classes de produits de la manière suivante :

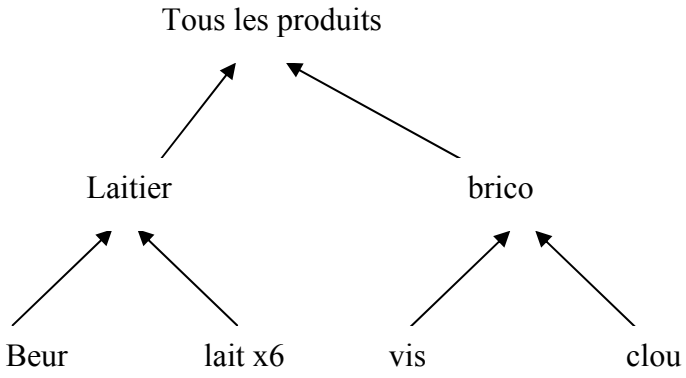


Figure 19: Exemple d'une hiérarchie.

5.5 Les formules

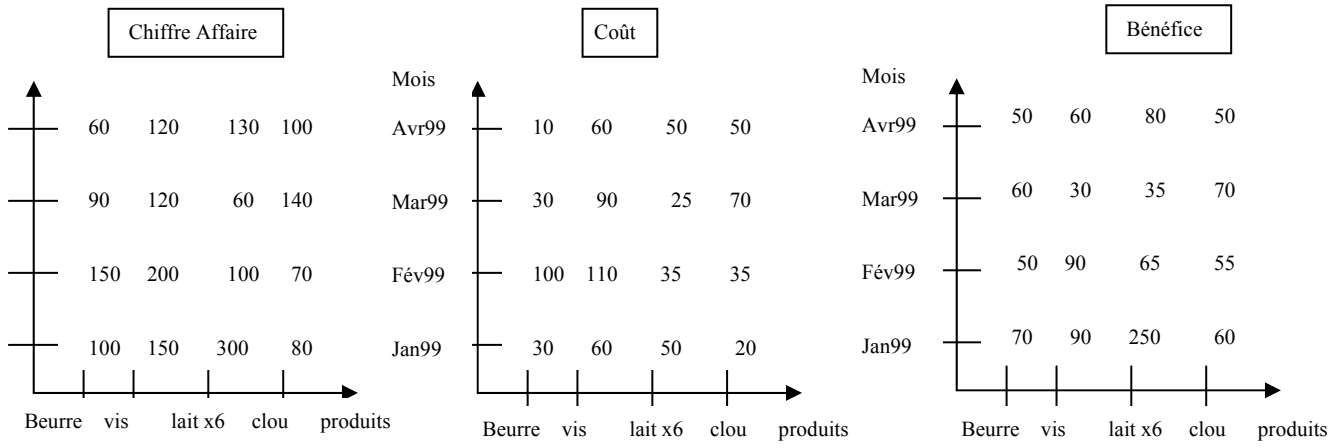


Figure 20: Exemple de formules.

Les formules sont des expressions de calcul constituées à partir des variables déclarées dans la base de données express. Le résultat d'une formule est évalué dynamiquement à chaque appel de la formule.

Le dimensionnement du résultat d'une formule est automatiquement calculé en fonction des dimensions des variables intervenant dans l'expression.

5.6 Limitation de la vue sur les données

L'opération LIMIT permet de selectioner des positions dans une dimensions afin d'accéder à certaines valeurs de variables. L'utilisateur peut, par example, sélectioner les chiffres d'affaire pour les vis et les packs de lait pour les mois de février et mars 1999.

5.7 Calcul des données agrégées

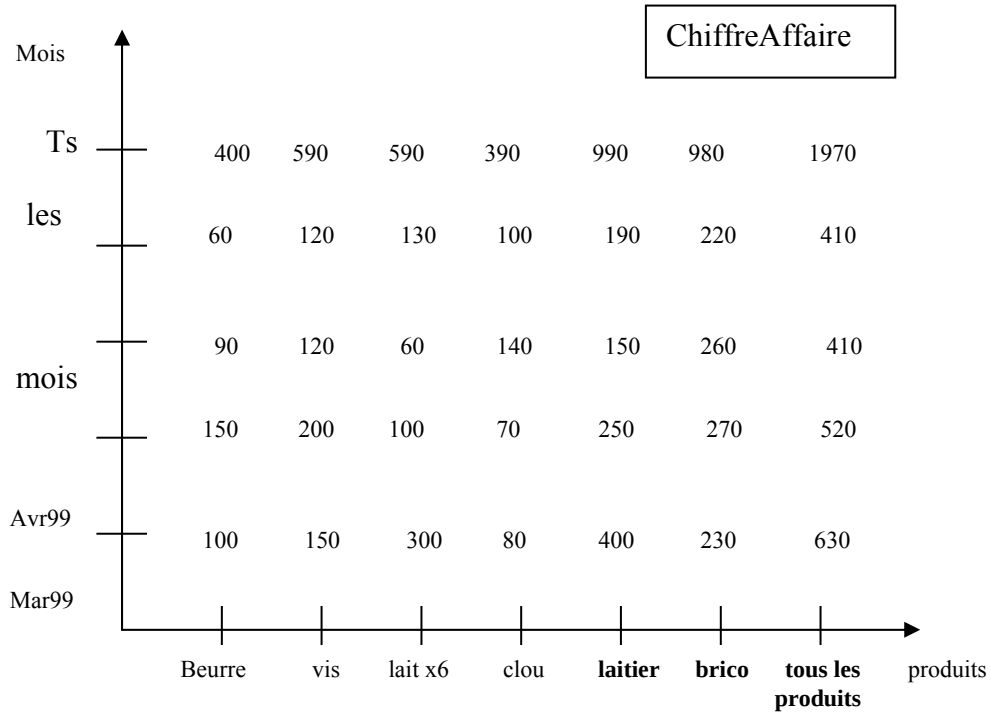


Figure 21 : résultat d'une opération ROLLUP

L'opération ROLLUP permet de créer des valeurs agrégées de variables à partir de valeurs détaillées. Le calcule des valeurs agrégées se fait en suivant les hiérarchies associées à chaque dimension. La figure précédente illustre ce principe. La fonction d'agrégation appliquée lors de l'opération ROLLUP est une somme.

Synthèse

Ce chapitre est consacré à la **modélisation multidimensionnelle**. L'approches multidimensionnelle correspond à une structuration adaptée des données (selon plusieurs axes d'analyses) reflétant la vision des analystes. Ce modèle comprend un **fait** contenant les mesures à analyser et des **dimensions** contenant les paramètres de l'analyse. Dans chaque dimension, les **paramètres** sont **hiérarchisés** selon des niveaux de détail.
L'hypercube de données peut avoir une représentation relationnelle ou un représentation multidimensionnelle (MOLAP, ROLAP). Le modèle le plus utilisé pour représenter l'hypercube est le **modèle en étoile**. Ce modèle tire son nom de la position centrale de la **table de fait** à partir de laquelle partent les références vers les **tables des dimensions**.
La manipulation de données multidimensionnelle est faite par des opérations permettant la navigation dans les données librement selon les axes d'analyse et selon leurs hiérarchies associées.

Dans la suite de ce mémoire, nous présentons une conception dédie à l'élaboration d'un système multidimensionnel repose sur les concepts précédents.

CHAPITRE IV :

L'élaboration
du système multidimensionnel

1. Introduction

Nous avons présenté dans les premiers chapitres des notions concernant l'informatique décisionnelle notamment le concept OLAP, la modélisation et la manipulation des structures multidimensionnelles. Notre objectif est d'utiliser ces aspects pour élaborer un système multidimensionnel basé sur les traitements OLAP effectuées sur un magasin de données. Dans ce chapitre nous présentons une conception tout en se basant sur des définitions algébriques concernant :

- le magasin de données (la base multidimensionnelle de notre système).
- Une démarche pour l'extraction et la transformation de données dans le magasin.
- La manipulation de données du magasin.

Nous présentons aussi un exemple complet pour l'élaboration et la manipulation d'un magasin en utilisant les fenêtres de notre prototype de logiciel.

2. Description du système

Notre système peut être considéré comme un outil ROLAP(relationel OLAP) permet d'organiser les données dans un magasin de données à partir d'un entrepôt de données de manière multidimensionnelle(on utilise un schéma en étoile). La réorganisation multidimensionnelle des données constitue une tâche essentielle dans l'élaboration d'un système décisionnel [4]. Nous choisissons le modèle relationnel parce qu'il intègre facilement dans un contexte de base de données relationnel, de plus il ne nécessite pas de redéfinir des nouvelles opérations pour manipuler les structures de données. Il caractérise aussi par sa grande capacité de stockage.

La base de notre système est accessible par plusieurs outils d'analyses et de visualisation (tableaux, graphiques, rapport …). Ces outils repose essentiellement sur une représentation tabulaire qui permet d'analyser,

visualiser et naviguer dans les données de manière simple selon les dimensions (axes d'analyse) et les hiérarchies associées à ces dimensions (cet aspect est détaillé dans la section 6 de ce chapitre). Nous ne traitons pas la visualisation graphique parce qu'il est simple de transférer les données d'un tableau en graphiques (courbes, gantt …).

On peut situer notre outil dans le système décisionnel décrit dans le chapitre 1 comme suit :

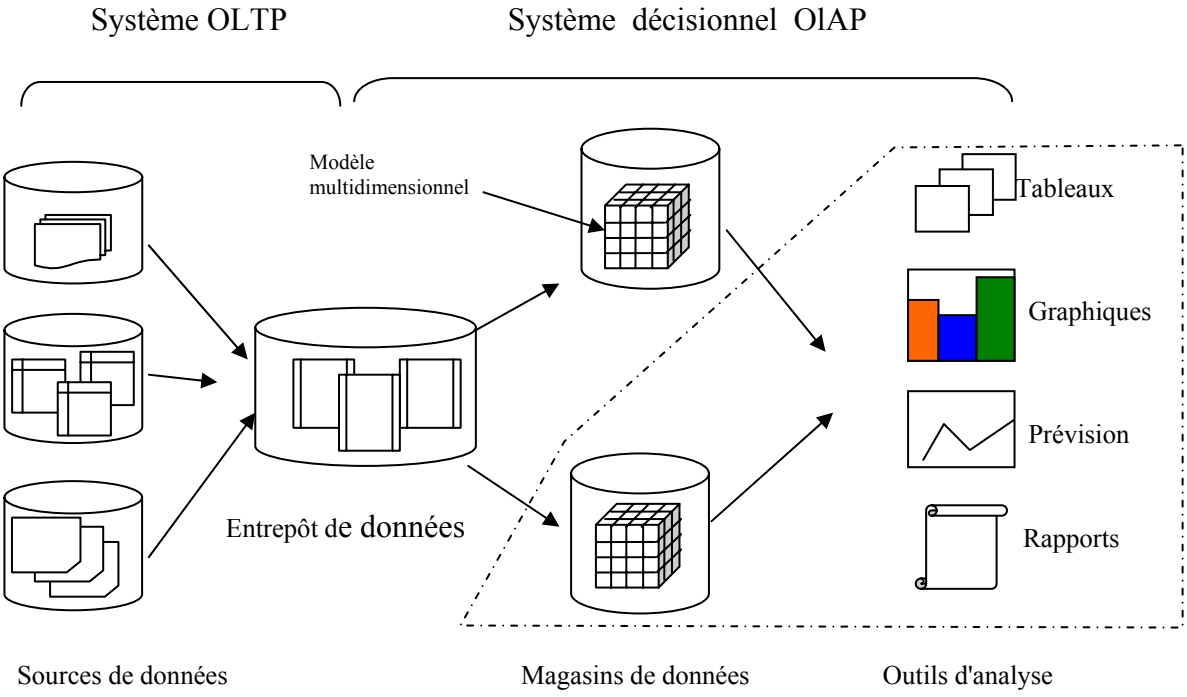


Figure 22 : L'emplacement de notre outil dans le système décisionnel.

3. Architecture du système

L'objectif de notre travail est de proposer un outil décisionnel pour améliorer la prise de décision. Notre système comprend deux modules :

- Module pour les administrateurs permettant de construire et modifier la base multidimensionnelle (hypercube de données) extraite à partir de l'entrepôt de données. Le modèle utiliser est le modèle en étoile.

-Module pour les utilisateurs permettant manipuler les données multidimensionnelles. Il offre une interface interactive pour l'interrogation et l'analyse.

La figure suivante illustre l'architecture de notre système.

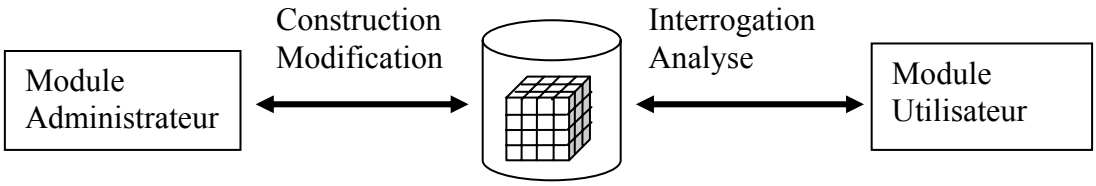


Figure 23: l'architecture de notre système.

Dans la suite de ce chapitre nous présentons une conception pour élaborer le magasin de données. Nous proposons une démarche pour l'extraction et la transformation de données dans le magasin, nous présentons aussi une algèbre pour la manipulation de données du magasin. Pour mieux comprendre notre démarche nous présentons un exemple d'élaboration d'un magasin (son schéma est décrit dans l'annexe) à travers les fenêtre de notre prototype de logiciel.

4. Modélisation du magasin de données

Le magasin de données permet un accès de l'utilisateur à toutes les données pertinentes de l'entrepôt. Les données doivent être organisées d'une manière qui soit à la fois conforme au point de vue de l'utilisateur et compatible avec les techniques de mise en œuvre. Dans notre système, le magasin de données repose sur un modèle multidimensionnel représentant les données sous la forme d'un schéma regroupant un fait et des dimensions.

4.1 Fait

Dans les systèmes décisionnels, un magasin de données est conçu pour rependre aux besoins des décideurs (requêtes). Il concerne un type d'activité, un type d'analyse [10]. Il contient les informations pertinentes (choisis par l'administrateur selon les objectifs des décideurs) issues de l'entrepôt.

Notre magasin de données comporte un seule **fait**. Il s'agit du sujet de l'analyse, c'est à dire des différentes **mesures** de l'activité à analyser.

Définition:

Fait F se définit par NomF (MF) où

- NomF désigne le fait,
- MF= {AF1, AF 2,…, AF f} est un ensemble des mesures de l'activité contenues dans le magasin.

Exemple: nous souhaitons analyser des ventes, nous choisissons comme mesures : la quantité vendue et le prix de vante

L'administrateur définit le fait comme suite :

Vente (quantité, prix).

4.2 Dimensions et hiérarchies

On peut observer les mesures de l'activité selon plusieurs perspectives. Chaque perspective est modélisée par une **dimension** qui regroupe des **paramètres** déterminant les niveaux de détail de l'analyse.

Les **hiérarchies** de paramètres sont utilisées pour analyser les données à différents niveaux de détail (notamment les opérations de forage vers le haut ou vers le bas). Les paramètres d'une dimension sont rattache entre eux par des liens d'appartenance ou de groupement hiérarchique. Un lien d'appartenance entre deux paramètres P_i et P_j indique qu'à une position (valeur) de Pi est associée exactement une position de P_j et à une position de P_j est associée une (ou plusieurs) position(s) de P_i (lien de groupement).

Définition:

Une relation d'appartenance entre deux paramètres P_i et P_j, notée $P_i \rightarrow P_j$.

Exemple: Soit une dimension "localisation géographique" contient deux paramètres "région" et "ville". Nous considérons les positions (valeurs) pour chaque paramètre :

- Positions (région) = {Est, ouest, sud}.
- Positions (Ville)={M'sila, bba, Oran, El 'oued, Annaba}.

On observe que ville → région (à une position de "ville" correspond une seule position de "région" et à une position de "région" correspond plusieurs positions de "ville").

A partir des relations d'appartenance, on peut définir une hiérarchie.

Définition:

On peut définir une hiérarchie sur un ensemble de paramètres d'une même dimension Dim par:

$H^{dim} = \{P^{dim}_1, P^{dim}_2, \ldots, P^{dim}_n\} / P^{dim}_1 \rightarrow P^{dim}_2 \rightarrow \ldots \rightarrow P^{dim}_n$

Exemple: Nous considérons l'ensemble des paramètres {ville, région, état}. On peut définir relations d'appartenance suivantes :

ville→ région →état

Il existe donc une hiérarchie sur cet ensemble de paramètres :

H^{dim}= {ville, région , état}.

Il est possible de définir plusieurs hiérarchies pour un même ensemble de paramètres, on parle de **hiérarchies multiples**.

Le magasin est constitué de dimensions dont les paramètres sont organisés selon une hiérarchie. Il s'agit des paramètres de l'analyse.

Définition:

Une dimension Dim se définit par le triplet (NomDim, ParamDim, HDim)où :

- NomDim désigne la dimension,

-ParamDim=$\{P^{dim}_1, P^{dim}_2, \ldots, P^{dim}_n\}$ est un ensemble de paramètres contenues dans la dimension,

- HDim =H^{dim} est la hiérarchie définie sur la dimension, formant une hiérarchie simple (on peut définir une hiérarchie multiple). Cette hiérarchie est utilisée pour visualiser les données de la dimension Dim par l'utilisateur.

On pose la fonction Param(H^{dim}) retourne l'ensemble des paramètres d'une hiérarchie.

Exemple: La vente des produits peut être analysé selon la dimension localisation géographique D1 définie de la manière suivante :

NomD1="Géographie",

ParamD1={magasin , ville, région},

HD1={ H^{D1}}\ H^{D1}={magasin, ville, région}

4.3 Schéma multidimensionnel

Notre magasin de données est caractérisé par un **schéma multidimensionnel**. Nous définissons un schéma multidimensionnel comme un schéma formé d'un fait et de dimensions, parmi lesquelles deux dimensions sont considérées visible. La visualisation de l'information contenue dans le schéma multidimensionnel est réalisée sous forme de tableaux, nous présentons la visualisation des données d'un schéma multidimensionnel dans la section 6.

Définition:

Un schéma multidimensionnel est défini par le triple (NomS, Fait, Dim,) où

- NomS est le nom du schéma multidimensionnel,
- Fait est le fait représente les mesures,
- DIM={D_1, D_2,…, D_p} est l'ensemble des dimensions,
- on pose Fdim une fonction renvoie la liste des dimensions.

Fdim={D_1, D_2,…, D_p} .

D_1, D_2 sont les dimensions visibles utilisées pour afficher les données du fait dans un plan.

Exemple: Nous poursuivons l'exemple relatif aux ventes dans entreprise. Il est possible de définir le schéma multidimensionnel SV comme suivant :

-Noms=" Ventesanalysés",

-Vente (quantité, montant),

- DIM={D_1, D_2, D_3} où

D_1=(temps,(Jour, Moins ,Saison ,Trimestre, année), H^{D1}),

D_2=(géographie(magasin, ville ,région), H^{D2}),

D_2=(produit(Nomprod, Catégorie), H^{D3})

On peut définir sur chaque dimension du schéma multidimensionnel une hiérarchie comme suit :

-H^{D1} ={ Jour, Moins ,Trimestre, année }

-H^{D2} ={magasin, ville, région}

-H^{D3} ={ Nomprod, Catégorie }

Nous complétons l'exemple précédent en donnant les graphes associés au schéma multidimensionnel intitulé "Ventesanalyses".

Cette représentation facilite la compréhension du schéma multidimensionnel

- La Figure 24 décrit le schéma en étoile issu du schéma multidimensionnel,
- La Figure 25 décrit les trois graphes de hiérarchie représentant les relations d'appartenance des paramètres contenus dans les trois dimensions D1, D2 et D3.

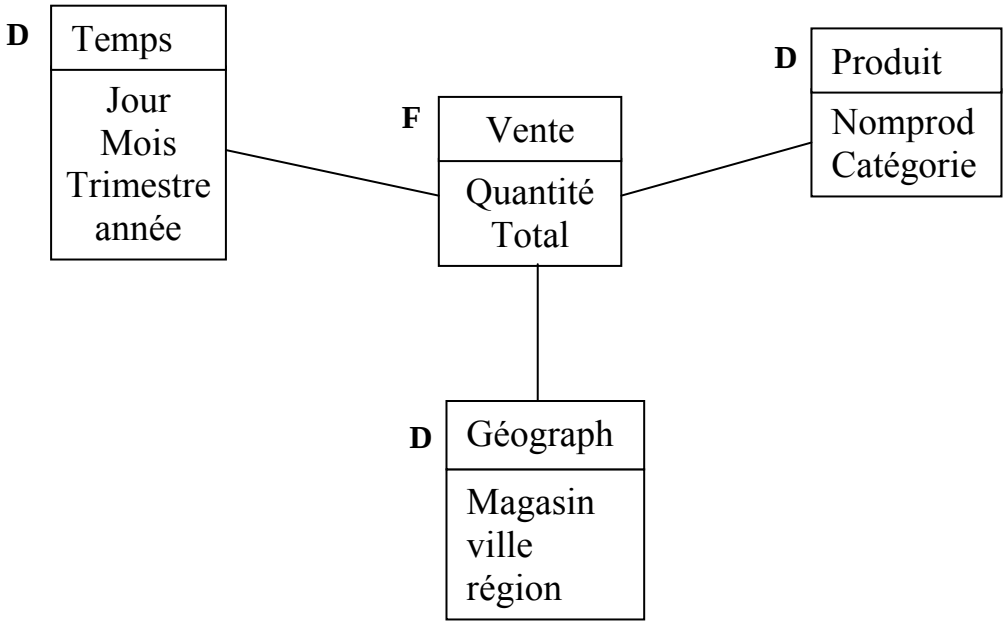


Figure 24 : Exemple de schéma en étoile (avec trois dimensions).

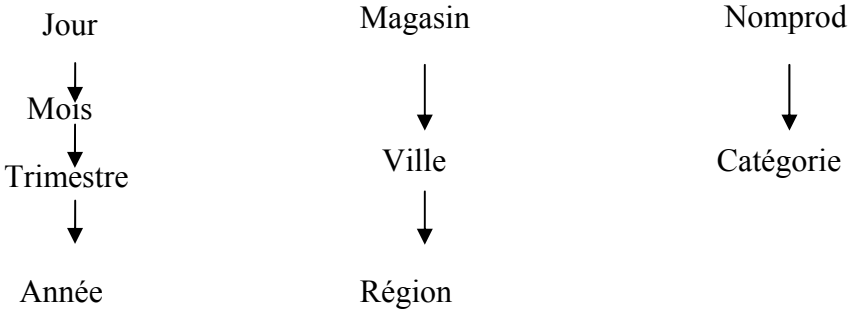


Figure 25 : Exemple de hiérarchies.

5. L'extraction de données

Selon la description de notre système le magasin de données est élaboré à partir d'un entrepôt de données, donc le schéma multidimensionnel est dérivé du schéma de source (l'entrepôt).

La création du schéma multidimensionnel nécessite une connaissance de la méta-donnée de l'entrepôt (description de données tels que, les tables, les types d'attribut…). Cette phase est faite par l'administrateur qui définit la correspondance entre le schéma multidimensionnel est tel de l'entrepôt. Il définit les éléments du schéma multidimensionnel (les mesures les dimensions,…) et les éléments sources correspondants.

Définition:

L'extraction de données est faite par l'opération :

Extract(SR) = S telle que : SR : schéma relationnel de l'entrepôt.

S : schéma multidimensionnel.

5.1 Modélisation par un graphe

Nous modélisons le schéma de source de données (l'entrepôt) par un graphe non orienté composé de nœuds et des arcs, un graphe peut représente de manière simple les tables (relations) de l'entrepôt et les liens entre ces table.

Définition:

Un graphe G est défini par le couple G(X,A) :

-X est l'ensemble des nœuds.

-A est l'ensemble des arcs.

Chaque nœud représente une table relationnelle.

Chaque arc représente un lien (référence par des clés étrangères) entre deux tables.

Exemple: dans cet exemple nous présentons un modèle conceptuel d'un entrepôt de données décrit selon le modèle entité/ relation, comme montre la figure suivante :

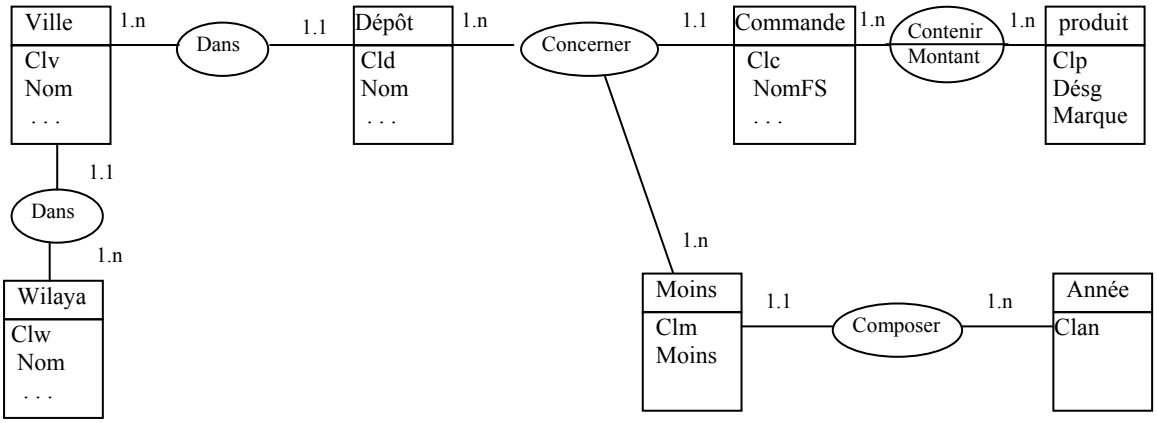


Figure 26 : Présentation de l'entrepôt au travers d'un modèle entité/relation.

On peut obtenir les schémas relationnels suivants :

1- Produit (Clp , désg , marque).
2- Contenir (Clc*, Clp*, montant).
3- Commande (Clc,NomFs,Cld*,Clm*).
4- Dépôt (Cld, nom,Clv*).
5- Ville(Clv , nom , clw*).
6- Wilaya (Clw , nom).
7- Moins (Clm,moins , Clan*).
8- Année (Clan,année).

Selon la définition précédente, on peut modéliser les données de l'entrepôt par le graphe suivant :

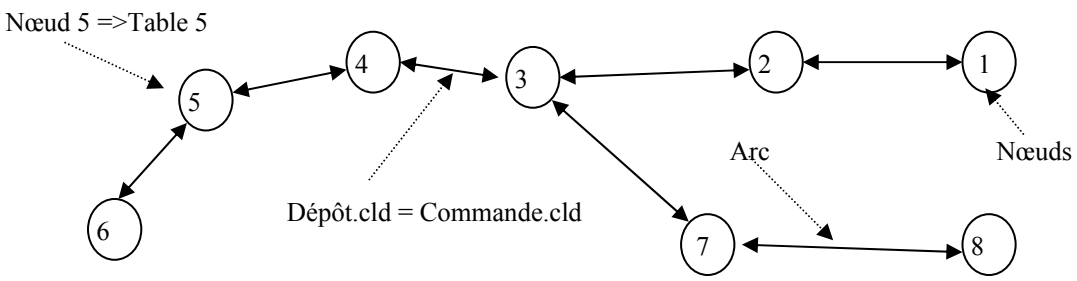


Figure 27 : modélisation par un graphe.

Pour agir sur la structure du graphe, on utilise les fonctions suivantes :

Définition:

La fonction Nœud(G) = E renvoie l'ensemble des nœuds du graphe.

En entrée

G est le graphe associé à la base de données (l'entrepôt).

En sortie

$E = (C_1 , C_2 ,\ldots, C_n)$ avec

$C_n = (C_{n1} , C_{n2}) / C_{n1} , C_{n2} \in X$ deux nœuds liés par un arc.

Exemple: nous reprenons l'exemple du graphe précédent, la résultat obtenue est :

Nœud (G)= ((1,2), (2,3), (3,4), (4,5), (5,6),(3,7),(7,8)).

Définition:

La fonction Arc (SR,G) = L renvoi l'ensemble des arcs du graphe.

En entrée

SR est le schéma de l'entrepôt.

G est le graphe associé à ce schéma.

En sortie

$L = (A_1 , A_2 ,\ldots, A_n)$ avec

$A_n = (Cl_{n1} , Cl_{n2}) / Cl_{n1} , Cl_{n2}$ deux clés reliant les deux tables représentés par les deux nœuds C_{n1} , C_{n2} obtenus par la fonction nœud(G) .

Exemple: nous poursuivons l'exemple du graphe précédent, la résultat obtenue est :

Arc(S,G)= ((Clp,Clp), (Clc,Clc), (Cld,Cld), (Clv,Clv), (Clw,Clw) , (Clm,Clm) , (Clan,Clan)).

Remarque : les deus fonctions Nœud et Arc permettent de parcourir le graphe et de référencer les tables pour faire les jointures nécessaire dans la phase d'extraction de données.

5.1.1 Le graphe connexe

Nous vérifions la connexité du graphe (existence d'un nœud ou d'un parti du graphe isolé), en utilisant une matrice d'adjacence. La matrice d'adjacence est une matrice carrée associée à un graphe. Les lignes et les colonnes représentent les nœuds. Pour chaque combinaison, la valeur "1" indique l'existence d'un arc, c'est-à-dire la table correspondant à cette ligne est référencée par une clé étrangère de la table représentée par la colonne , la valeur "0" dans le cas contraire.

Exemple: à partir de l'exemple précédent, on peut construire la matrice suivante:

Tables	Produit	Contenir	Command	Dépôt	Ville	Wilaya	Moins	année
Produit	0	1	0	0	0	0	0	0
Contenir	1	0	1	0	0	0	0	0
Commande	0	1	0	1	0	0	1	0
Dépôt	0	0	1	0	1	0	0	0
Ville	0	0	0	1	0	1	0	0
Wilaya	0	0	0	0	1	0	0	0
Moins	0	0	1	0	0	0	0	1
Année	0	0	0	0	0	0	1	0

Figure28 : Matrice d'adjacence.

Pour construire cette matrice, on fait l'appel à la fonction:

Construire(Nœud (G),Arc(SR,G))=M (M:matrice d'adjacence).

On pose la fonction connexe (M) pour parcourir la matrice M et vérifier que la matrice connexe ou non. La connexité du graphie est indispensable pour assurer que les tables sont référencé par de clés étrangère (opération de jointure pour extraire les données).

5.2 Détermination du fait et des dimensions

Pour élaborer notre magasin, nous proposons une démarche comportant trois étapes :

- Détermination des dimensions en spécifiant les paramètres et les hiérarchies associées.
- Détermination de fait et des mesures.
- Détermination des fonctions de groupement (granularité de données de l'analyse).

Dans notre architecture, nous supposons que l'ensemble d'attributs représentant les mesure et les paramètres des dimensions sont inclus directement dans l'ensemble d'attributs des tables de l'entrepôt. Si ce n'est pas le cas, l'administrateur peut définir des requêtes SQL pour créer des tables représentant les données multidimensionnelles (mesures, paramètres,…).

5.2.1 Détermination des dimensions

La première étape consiste à déterminer l'ensemble Dim des dimensions. Pour chaque dimension, on spécifie l'ensemble Attributdim des paramètres. La définition d'hiérarchie sur la dimension est faite de manière implicite selon l'ordre de choix des paramètres.

Définition:

Le nom d'une dimension D_i est obtenu par la fonction Nommage(D_i) , par exemple : Nommage(D_i) ="Géographie".

Définition:

La détermination d'un paramètre A_i de la dimension D_i est réaliser par l'opération Param(Nom,D_i,T,attrib) où

- Nom est le nom du paramètre A_i.
- D_i est la dimension.
- T est la table de la base de données contenant l'attribut représentatif du paramètre A_i.
-attrib est l'attribut représentatif du paramètre A_i.

Exemple: nous complétons l'exemple de cette section, l'administrateur peut définir une dimension suivante :

D_1 =(géographie,(dépôt,ville,wilaya),H^{D1}) où

- "géographie" = Nommage(D_i).
- "dépôt" = Param(dépôt, géographie,dépôt,nom).
- "ville" = Param(ville, géographie,vile,nom).
- "wilaya" = Param(wilaya, géographie,wilaya,nom).
- H^{D1} ={ dépôt ,ville,wilaya}.

On peut schématiser ainsi :

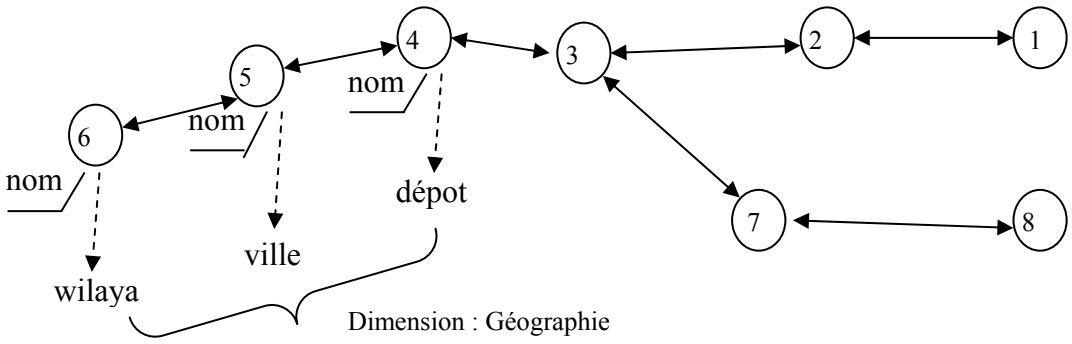


Figure 29: Détermination d'une dimension.

5.2.2 Détermination du fait

La deuxième étape consiste à déterminer le fait et ces mesures. Les mesures sont des attributs issus des tables de l'entrepôt de données.

Définition:

Le nom du fait F est obtenu par la fonction Nommage(F) , par exemple : Nommage(F) =" vente ".

Définition:

La détermination d'une mesure AF_i du fait F est réaliser par l'opération Mesure(Nom,T,attrib) où

- Nom est le nom de la mesure AF_i.

- T est la table de la base de données contenant l'attribut représentatif de la mesure AF_i.

-attrib est l'attribut représentatif de la mesure AF_i.

Exemple: avec le même exemple on peut définir un fait comme suit :

Vente(montant) avec :

- "Vente"=nommage(F).

-"montant" =mesure(montant,contenir,montant) où

-contenir est la table contenant l'attribut représentatif de la mesure "montant".

- montant est l'attribut représentatif de la mesure "montant".

La figure suivante illustre la détermination du fait "Vente".

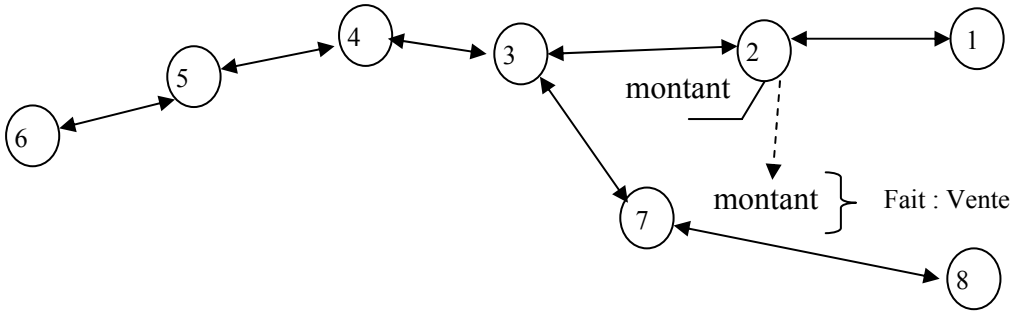


Figure 30: Détermination du fait.

5.2.3 Détermination des fonctions de groupement

On peut spécifier le type de regroupement devant être appliqué sur l'ensemble de valeur d'une mesure, c'est-à-dire on spécifie l'agrégation à appliquer sur un niveau de granularité (selon les hiérarchies des dimensions).

Définition:

L'agrégation des mesures est réalisée par l'opération :

Agrégation (AF_i, Fonction) où

-AF_i est une mesure

- Fonction est fonction d'agrégation telle que sum,avg,min,max,…

Exemple: on peut définir une fonction d'agrégation "la somme" sur la mesure "montant" de l'exemple passé comme suivant :

Agrégation (montant , sum).

6. Manipulation des données du magasin

Dans l'architecture des systèmes décisionnels, les magasins de données sont dédiés à un groupe d'analystes. Pour cela, nous avons défini un modèle multidimensionnel qui représente les données analysées en modèle en étoile, ce modèle est constitué d'un fait et de dimensions munies de hiérarchies.

6.1 Représentation tabulaire du schéma multidimensionnel

La représentation sous forme de tableau est permet d'un visualisation simple et intuitive pour les utilisateurs d'un magasin de données (décideurs).

Un tableau donne une vision d'un schéma multi dimensionne en deux dimensions.

Soit le schéma multidimensionnel S =(NomS, Fait, Dim) où

- Noms est le nom du schéma multidimensionnel,
- Fait est le fait qui représente les mesures,
- Dim=$\{D_1, D_2,\ldots, D_p\}$ est l'ensemble des dimensions.

Fait est le fait dont les mesures sont visualisées par les utilisateurs. D_1, D_2 sont les deux dimensions visibles dont les paramètres sont utilisés pour visualiser les mesures du fait suivant leur hiérarchie.

La Figure 31 illustre l'affichage schéma multidimensionnel sous forme tabulaire organisé en colonnes et en lignes.

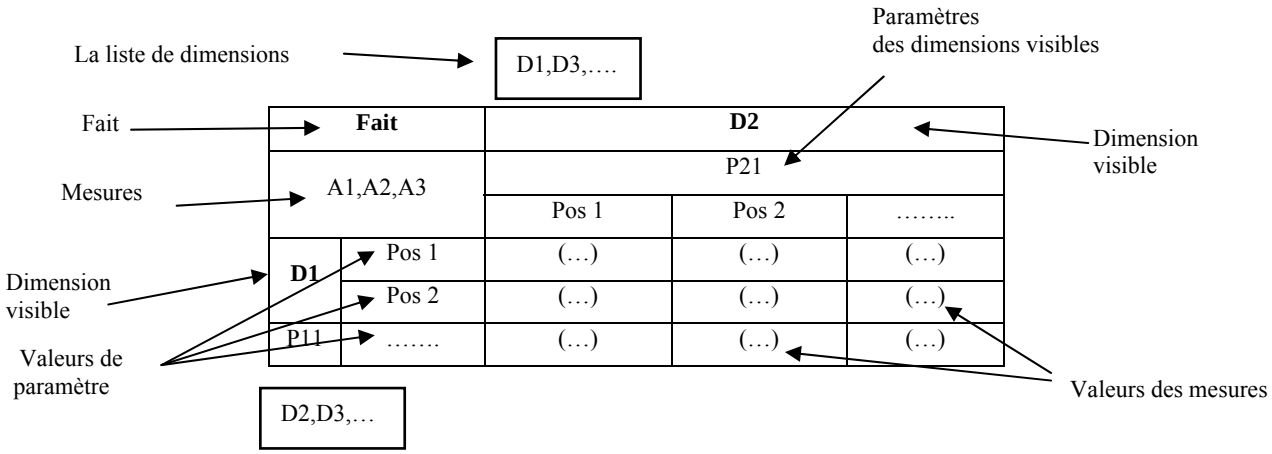


Figure 31 : Représentation tabulaire d'un schéma multidimensionnel.

- Le fait est affiché.

- Les deux dimensions visibles sont affichées en ligne et en colonne. La liste des autres dimensions est disponible en bas et en haut de la figure. On peut ensuite permuter les dimensions pour changer le plan de visualisation.

- Pour les deux dimensions visibles, les paramètres sont affichés suivant leurs positions dans la hiérarchie définie sur chaque dimension.

- Les valeurs des mesures du fait sont placées à l'intersection d'une ligne et d'une colonne.

Exemple: Nous reprenons le magasin de données caractérisé par le schéma multidimensionnel présenté à la section 4.3 relatif aux ventes. Le tableau suivant présente les données du fait vente contenu dans le schéma.

Vente		**Géographie**		
Montant		Région		
		Est	ouest	Sud
Temps	Trim1	250	200	130
trimestre	Trim2	140	150	300
	Trim3	130	210	310

Tableau 6: Représentation des ventes par trimestre et par région pour la catégorie des produits "alimentation".

Dans la suite de cette section, nous montrons les principales opérations multidimensionnelles applicables sur les schémas multidimensionnels :

- Des opérations permettant de modifier la granularité des données affichées pour observer les données suivants différents niveaux de détail,
- Des opérations permettant de modifier les structures affichées, c'est à dire de transformer les dimensions et les hiérarchies pour visualiser les données suivants différentes perspectives.
- Des opérations classiques de l'algèbre relationnelle.

6.2 Opérations associées à la granularité

Ces opérations permettent de visualiser les mesures d'un schéma multidimensionnel en modifiant leur niveau de détail (granularité). Elles consistent à augmenter ou à diminuer les paramètres de l'analyse conformément aux hiérarchies des dimensions.

6.2.1 Opérations de forage

Les opérations de forage sont parmi les plus importantes du processus OLAP. Elles permettent d'augmenter ou de diminuer ("forer") le détail avec lequel l'information est représentée, ces opérations de forage manipulent les différents niveaux de granularité des données multidimensionnelles.

L'opération de **forage vers le bas** consiste à introduire un paramètre de granularité inférieure dans la hiérarchie d'une dimension. L'utilisateur commence par analyser les données à un faible niveau de détail, puis l'utilisateur analyse à un niveau de détail inférieur.

Définition:

Le forage vers le bas introduit un niveau de détail inférieur dans la hiérarchie d'une dimension. La syntaxe de l'opération est la suivante :

DrillDown(S, D, P) = S'

En entrée

S =(NomS, Fait, Dim,) est un schéma multidimensionnel,

D est une dimension de ce schéma \ D ∈ Dim,

P est un paramètre de la dimension D.

En sortie

S' =(NomS, Fait, Dim') avec Dim' est issue de Dim dont la dimension D en affichant les données suivant les valeurs du paramètre P'/ P' → P .

Exemple: Nous considérons le schéma multidimensionnel décrit au Tableau 6. L'utilisateur souhaite visualiser les données en descendant d'un niveau de détail, il souhaite afficher les données suivant le paramètre ville en plus de celui de région sur la dimension géographie.

L'utilisateur définit l'opération suivante :

DrillDown(S, géographie, ville).

Le résultat de cette opération est décrit dans le tableau suivant.

Vente		Géographie							
		Région							
		ville							
Montant		est			ouest			sud	
		Oran	Rel	Chel	Set	BBa	Cons	M'sila	Slim
temps	Trim1	50	40	30	50	10	60	50	50
Trimestre	Trim2	40	30	10	20	70	15	19	20
	Trim3	14	20	15	41	20	30	24	10

Tableau 7 : Résultat du forage vers le bas au paramètre ville sur la dimension Géographie.

Le **forage vers le haut** permet de représenter les mesures à une granularité moins fine. L'opération permet d'afficher un paramètre hiérarchiquement supérieur dans une dimension.

Définition:

Le forage vers le haut introduit un niveau de détail supérieur dans la hiérarchie d'une dimension. La syntaxe de l'opération est la suivante :

RollUp(S, D, P) = S'

En entrée

S =(NomS, Fait, Dim,) est un schéma multidimensionnel,

D est une dimension de ce schéma \ D ∈ Dim,

P est un paramètre de la dimension D.

En sortie

S' =(NomS, Fait, Dim') avec Dim' est issue de Dim dont la dimension en affichant les données suivant les valeurs du paramètre P'/ P → P' .

Exemple: Nous poursuivons l'exemple précédent. Le Tableau 7 décrit le schéma S obtenu après un forage vers le bas. On réalise maintenant un forage vers le haut pour revenir à un niveau de granularité moins fin par:

RollUp(S, géographie, région).

Le Tableau 6 décrit le schéma obtenu (opération inverse).

6.2.2 Opération d'agrégation

L'opération d'agrégation permet de regrouper les mesures d'un fait suivant des valeurs de paramètres. Une fonction d'agrégation telle que sum, avg, moyen, min, max ou count peut ensuite s'appliquer sur les mesures regroupées.

Les données de granularité minimale peuvent être pré-agrégées suivant les différentes combinaisons d'attributs appartenant aux dimensions du schéma multidimensionnel. Les préagrégats sont utilisés pour accélérer le traitement des opérations d'agrégation car il est inutile de revenir aux données les plus détaillées en utilisant des données d'un niveau intermédiaire. Cependant, ces pré-agrégats présentes des inconvénients : ils occupent un espace de stockage important (ce qui limite leur nombre) et il est difficile de prévoir les pré-agrégats à stocker (avec la fonction d'agrégation adaptée) [8].

6.2.3 Opération de calcul des totaux

Cette opération permet de calculer des sous-totaux et un total final pour les mesures d'un fait selon une combinaison de deux dimensions.

Définition:

Le calcul des totaux d'un fait permet d'afficher les sous-totaux et le total final des mesures du fait , pour les dimensions visibles. La syntaxe de l'opération est la suivante :

Cube(S) = S'

En entrée

S=(NomS, Fait, Dim) est un schéma multidimensionnel.

En sortie

S' est le schéma multidimensionnel S résultat auquel les sous-totaux et le total final des mesures du fait sont ajoutés pour les dimensions visibles.

Exemple: Nous considérons le schéma multidimensionnel décrit au Tableau 6. On souhaite visualiser les ventes en affichant les totaux de chaque ligne et chaque colonne. On réalise donc l'opération suivante :

Cube(S) = S'

Le tableau suivant illustre le résultat obtenu. Une ligne et une colonne supplémentaires sont ajoutées pour afficher les sous-totaux , elles représentent l'ensemble des positions pour les trimestres (colonnes) et les régions (lignes).

Vente		**Géographie**			
Montant		Région			
		Est	ouest	Sud	
Temps	Trim1	250	200	130	580
trimestre	Trim2	140	150	300	590
	Trim3	130	210	310	650
		520	560	740	1820

Tableau 8 : Résultat des calculs des totaux.

6.3 Opérations agissant sur la structure

Ces opérations agissant sur la structure permettent de changer la représentation du schéma multidimensionnel en modifiant sa structure. L'objectif de ces opérations est de manipuler le schéma afin de mieux appréhender les informations.

6.3.1 Opération de rotation

Parmi les opérations classiques en analyse multidimensionnelle est l'opération de rotation qui permet d'analyser les données selon différentes perspectives. Cette opération s'applique aux dimensions d'un schéma multidimensionnel.

Définition:

La rotation de dimensions inverse deux dimensions dans la liste dim. La syntaxe de l'opération est la suivante :

Rotate(S, D_i, D_j) = S'

En entrées

S=(NomS, Fait, Dim) est un schéma multidimensionnel.

D_i , $D_j \in$ Dim \ D_i est une dimension visible, $D_j \in$ Dim/{D_j}.

En sortie

S'=(NomS, Dim) est un schéma multidimensionnel résultat tel que

D_j sera une dimension visible à la place de D_i.

Exemple: Nous considérons l'exemple de schéma multidimensionnel décrit au Tableau 6. On souhaite visualiser les ventes des produits suivant les deux dimensions du Produit et du Temps. On définit l'opération de rotation des dimensions Géographie et Produit :

Rotate("Ventesanalysés",vente, géographie, produit)=S'

Dans le schéma S' résultant de cette rotation, les deux dimensions Géographie et Produit sont permutées de telle sorte que Produit devienne une dimension visible, autrement dite, les valeurs des paramètres de Produit et Temps sont affichés. Le tableau suivant illustre le résultat obtenu.

Vente		**Produit**		
Montant		Catégorie		
		Alimentation	Vêtement	Netoyage
Temps	Trim1	250	200	130
trimestre	Trim2	140	150	300
	Trim3	130	210	310

Tableau 9: Résultat de la rotation des dimensions Géographie et Temps.

6.3.2 Opération de permutation

L'opération de permutation agissant sur la structure des paramètres des dimensions. Elle permet de changer l'ordre des positions(valeurs) affichées. Cette opération est s'applique pour réorganiser les données contenues dans les dimensions afin de mettre en évidence certaines valeurs de paramètres.
On utilise une fonction position (P) = $\{pos_1, pos_2,\ldots, pos_n\}$ qui retourne
L'ensemble ordonné des positions d'un paramètre P.

Définition:

La permutation inverse deux positions d'un paramètre. La syntaxe de l'opération est la suivante :

Switch(S, D, P, pos_i, pos_J) = S'

En entrée

S=(NomS, Fait, Dim) est un schéma multidimensionnel,

D ∈ Dim est une dimension du schéma multidimensionnel,

P est un paramètre de la dimension D,

pos_i, pos_J sont deux positions du paramètre P \ psition(p)={.., pos_i ,...,pos_j ,.. },

En sortie

S'=(NomS, Fait, Dim) est le schéma multidimensionnel en inversant les positions pos_i et pos_J ,donc psition(p)={.., pos_j ,...,pos_i ,.. }.

Exemple: Nous poursuivons l'exemple du schéma multidimensionnel décrit par le Tableau 6. On souhaite visualiser les ventes des produits en changeant l'ordre des régions. On réalise donc les permutations suivantes :

Switch(S, Géographie, région, sud,ouest) = S'

Switch(S', Géographie, région, sud,est) = S''

Le tableau suivant illustre le résultat obtenu.

Vente		**Géographie**		
		Région		
Montant		Sud	Est	Ouest
Temps	Trim1	130	250	200
trimestre	Trim2	300	140	150
	Trim3	310	130	210

Tableau 10: Résultat des permutations des positions du paramètre région.

6.3.3 Opération d'emboîtement

L'emboîtement permet de modifier l'ordre des paramètres d'une dimension. Il s'agit de permuter deux paramètres d'une dimension. L'intérêt de cette opération réside dans la possibilité pour l'utilisateur de modifier l'ordre d'importance entre les paramètres en modifiant l'ordre hiérarchique des paramètres.

Définition:

L'emboîtement permute deux paramètres dans la hiérarchie H^D associée à une dimension D.

La syntaxe de l'opération est la suivante :

Nest(S, D, P_i, P_j) = S'

En entrée

S=(NomS, Fait, Dim) est un schéma multidimensionnel,

D ∈ Dim est une dimension du schéma multidimensionnel,

P_i, P_j sont deux paramètres de la dimension D tels que H^D= {…, P_i, ..,P_j,…,}.

En sortie

S'=(NomS, Fait, Dim') est le schéma multidimensionnel résultat dans lequel

La dimension D' issue de $D \setminus H^D = \{\ldots, P_j, .., P_i, \ldots,\}$.

Exemple : Nous considérons le schéma multidimensionnel décrit par le Tableau 6. On souhaite visualiser les ventes des produits par trimestre, pour les trois années 2004,2005et 2006. On réalise donc l'emboîtement suivant :

Nest(S, Tps, Année, Trimestre) = S'

Le tableau suivant illustre le résultat obtenu.

Vente		**Géographie**								
		Région								
		ville								
Montant		Trim1			Trim2			Trim3		
		2004	2005	2006	2004	2005	2006	2004	2005	2006
Temps	Trim1	50	40	30	50	10	60	50	40	20
Trimestre	Trim2	40	30	10	20	70	15	19	30	26
	Trim3	14	20	15	41	20	30	24	40	23

Tableau 11 : Résultat de l'emboîtement du paramètre trimestre dans le paramètre année.

6.4 Opérations classiques

On peut **renommer** les noms d'attributs, c'est à dire les noms des mesures et des paramètres, et même les nom de dimensions et de fait. D'autre part, les **suppressions** d'attributs (c'est à dire de mesures, de paramètres, de fait ou de dimensions).

L'ajout d'une mesure dans un fait est réalisable par l'intermédiaire d'une expression de calcul qui combine d'autres mesures, par exemple, si l'on dispose d'une mesure représentant le prix d'une vente et d'une autre mesure représentant la quantité vendue, il est alors possible d'ajouter une mesure calculée dynamiquement (prix * quantité) et représentant le montant total de la vente.

Il est possible d'utiliser les opérations des langages classiques en base de données issues de l'algèbre relationnelle, notamment l'opération de **sélection** et **projection** (Slice and Dice qui signifie littéralement "couper des tranches"), qui permet de sélectionner un sous-schéma. Elle correspond à

l'opération de sélection de l'algèbre relationnelle mais s'applique ici à un schéma multidimensionnel**.**

7. Conception d'interfaces

Au niveau d'interfaces, nous proposons deux interfaces selon l'architecture de notre système, une pour les administrateurs et l'autre pour les décideurs.

7.1 L'interface d'administration

Cette interface permet de construire et de modifier le schéma multidimensionnel. Elle présente la source (entrepôt de données) travers son schéma relationnel. Cette interface comporte trois zones comme montre la figure suivante.

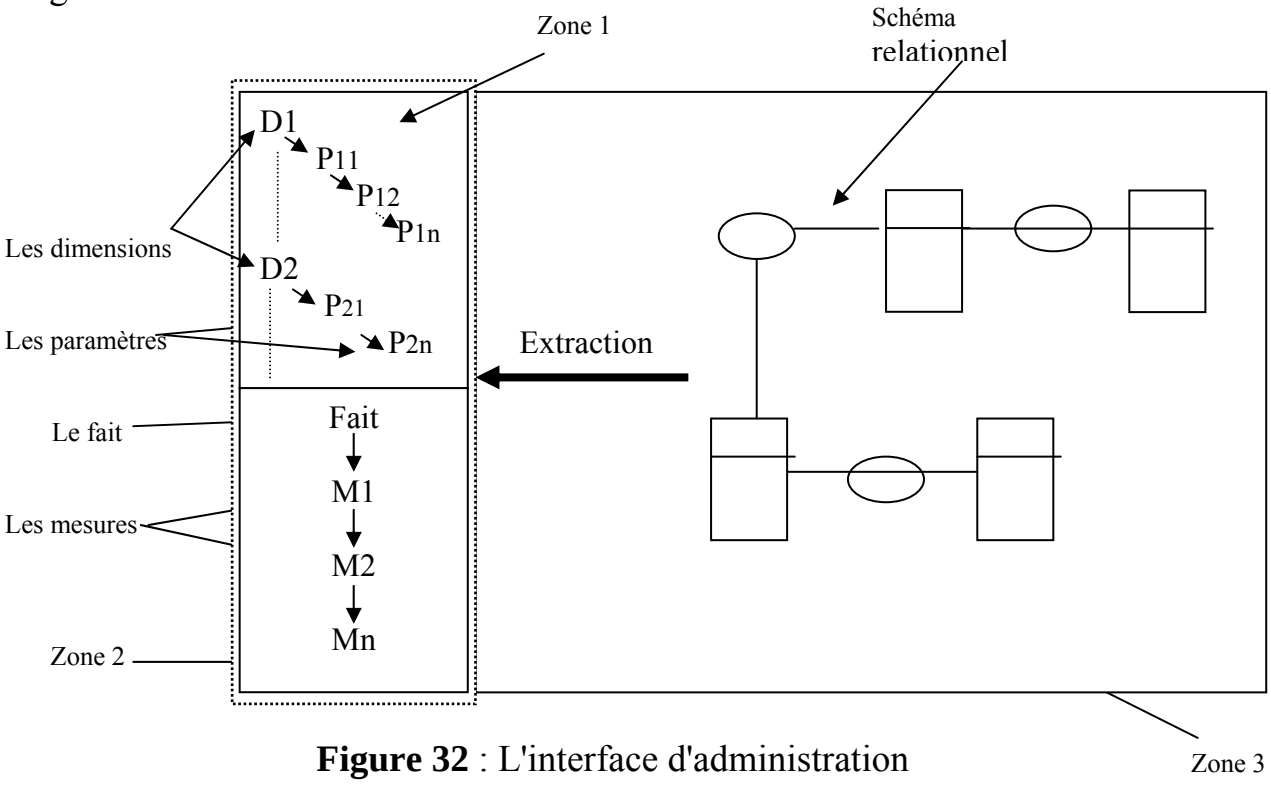


Figure 32 : L'interface d'administration

- La zone 1 permet de définir les dimensions, les paramètres et les hiérarchies.
- La zone 2 permet de définir le fait et les mesures.
- La zone 3 offre la possibilité d'obtenir le schéma relationnel de l'entrepôt (source de Données).

Au niveau d'extraction, cette interface permet à l'administrateur d'agir de manière interactive sur les deux schémas pour faire la correspondance entre les éléments multidimensionnels et les éléments de l'entrepôt.

7.2 L'interface décideur

Cette interface est orientée vers les décideurs. Elle permet de naviguer dans les données de manière simple et interactive et de visualiser le cube de données essentiellement sous forme de tableaux. Il est possible d'utiliser des formats graphiques (courbes, histogrammes,...). Comme l'interface précédente cette interface est composée de trois zones.

La figure suivante illustre une représentation tabulaire d'un schéma multidimensionnel.

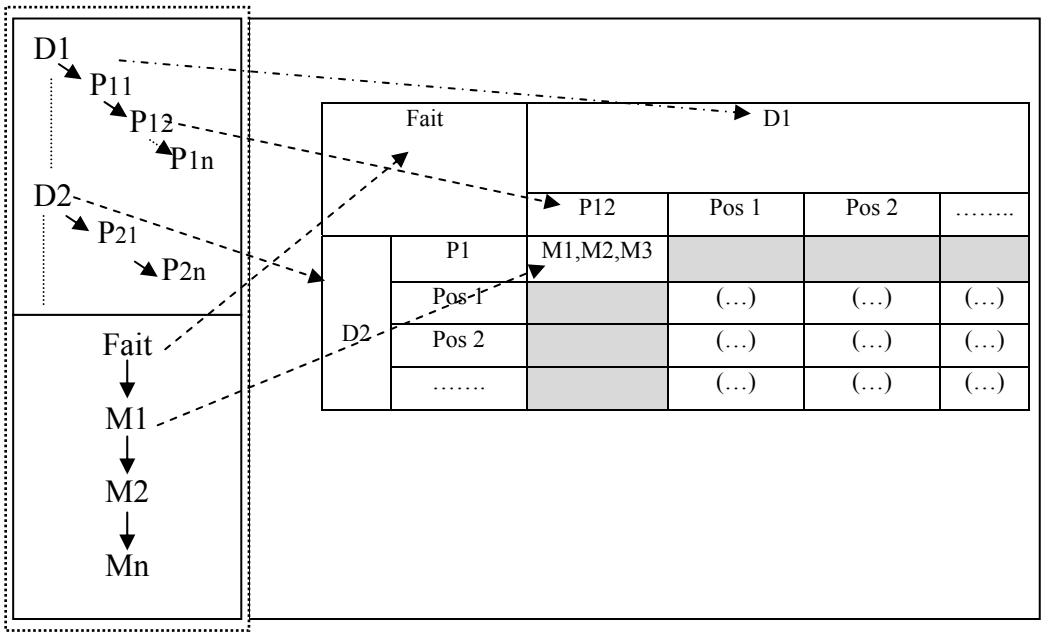


Figure 33 : L'interface décideur.

- La zone 1 permet de l'interaction avec l'hypercube en s'appuyant sur les dimensions, les paramètres et les hiérarchies.
- La zone 2 permet de choisir les mesures à analyser.
- La zone 3 affiche une tranche de l'hypercube en plant de deux dimensions. Il permet de visualiser les données au travers des tableau ou des graphiques.

8. Exemple complet d'élaboration et manipulation d'un magasin de données

Dans cette section, nous décrivons un exemple complet de construction d'un magasin de données (son schéma est décrit dans l'annexe A) à partir d'un entrepôt (décrit dans la section 5.1) selon les fenêtres de notre prototype de logiciel.

L'administrateur peut connecter facilement la base de l'entrepôt pour obtenir l'ensemble des tables et l'ensemble d'attributs correspondants.

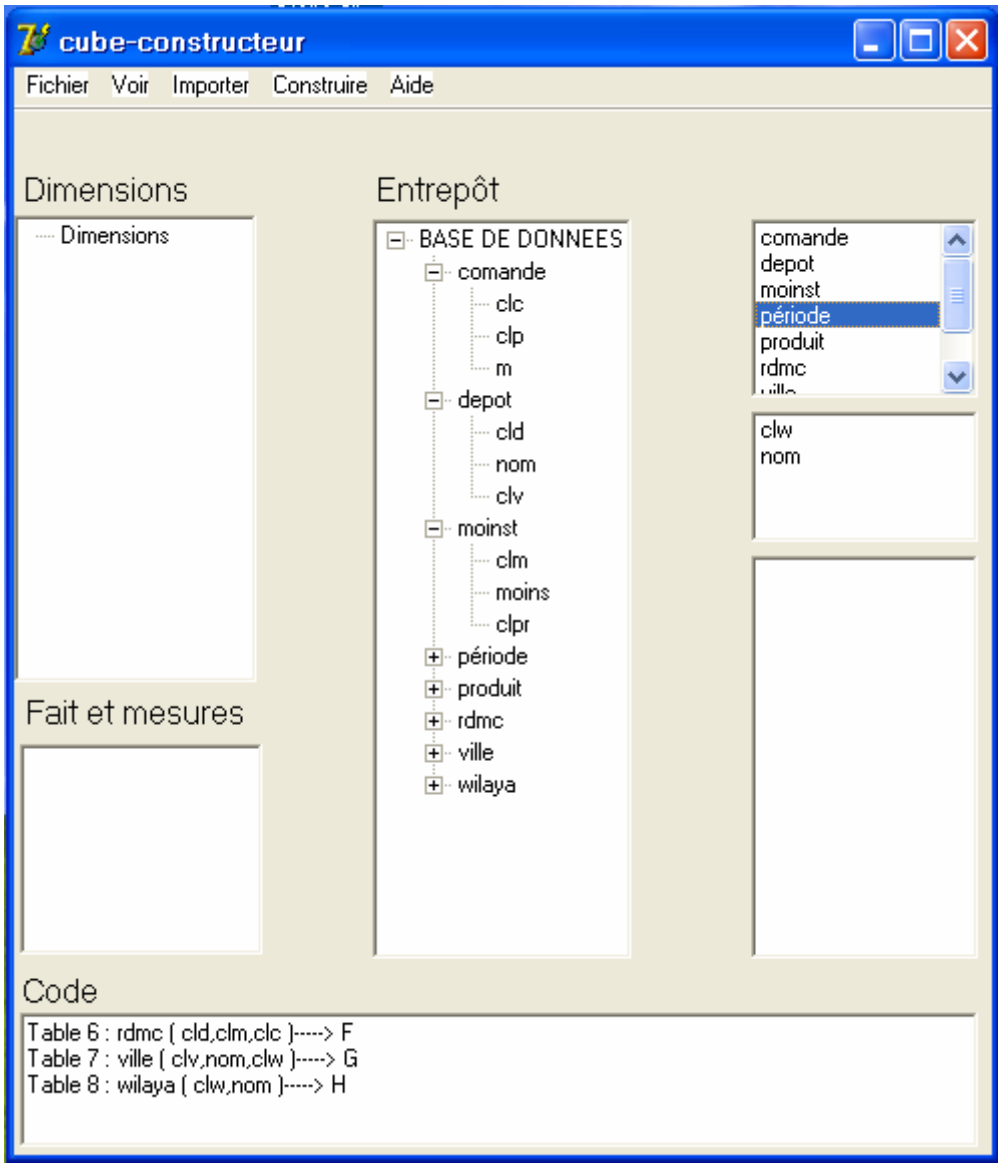


Figure 34: Représentation de la base de données de l'entrepôt.

L'administrateur élabore le magasin (cube de données) en définissant :

- Les liens entre les tables.
- Les dimensions (noms, les paramètres, les hiérarchies).
-Le fait (nom, les mesures).

8.1 Définition des liens

L'administrateur détermine les champs pertinents pour référencer les tables.

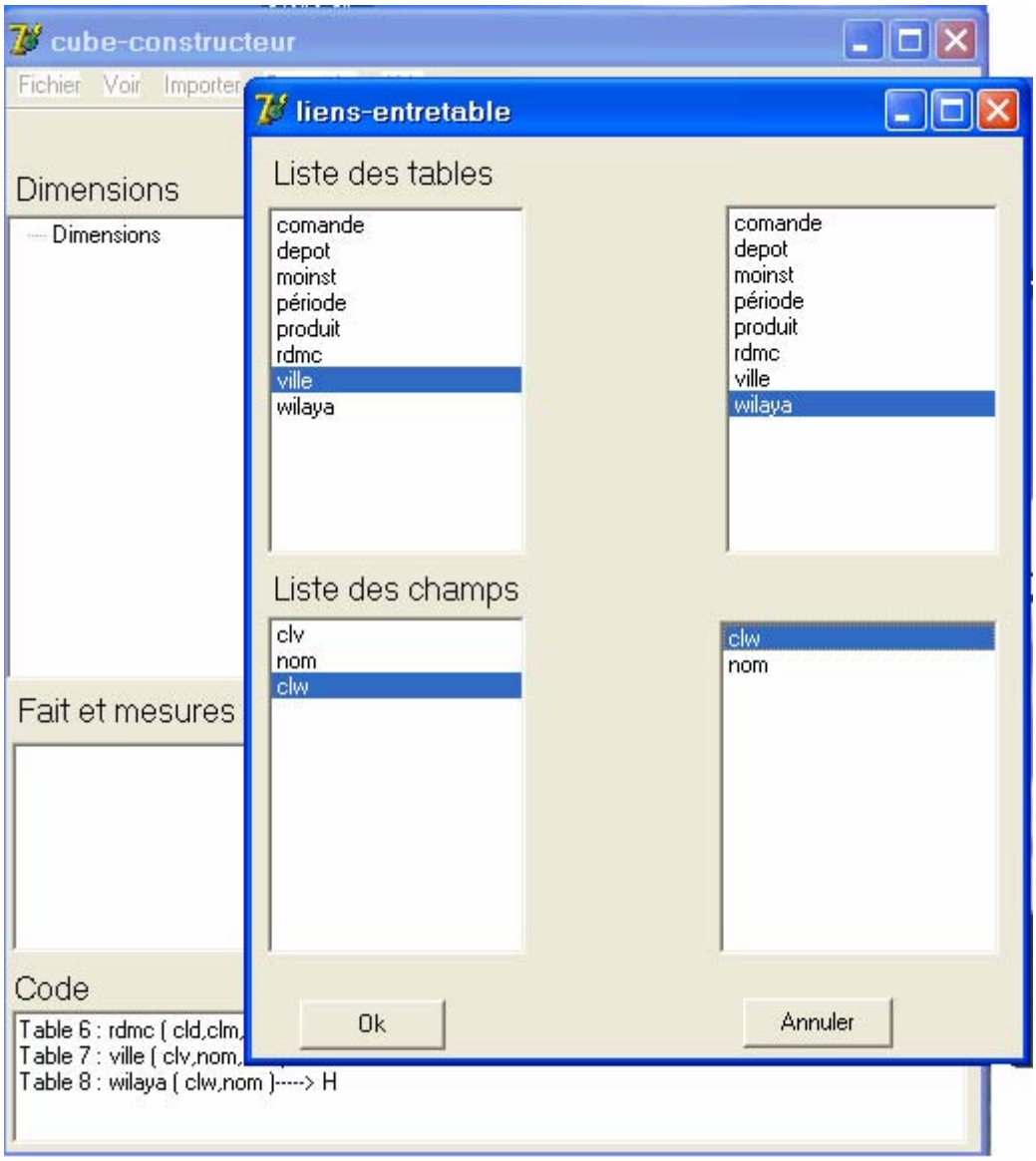


Figure 35 : Définition des liens entre les tables de la base source.

8.2 Détermination des dimensions

La seconde étape de notre démarche consiste à déterminer les dimensions en définissant les tables et les champs représentatifs.

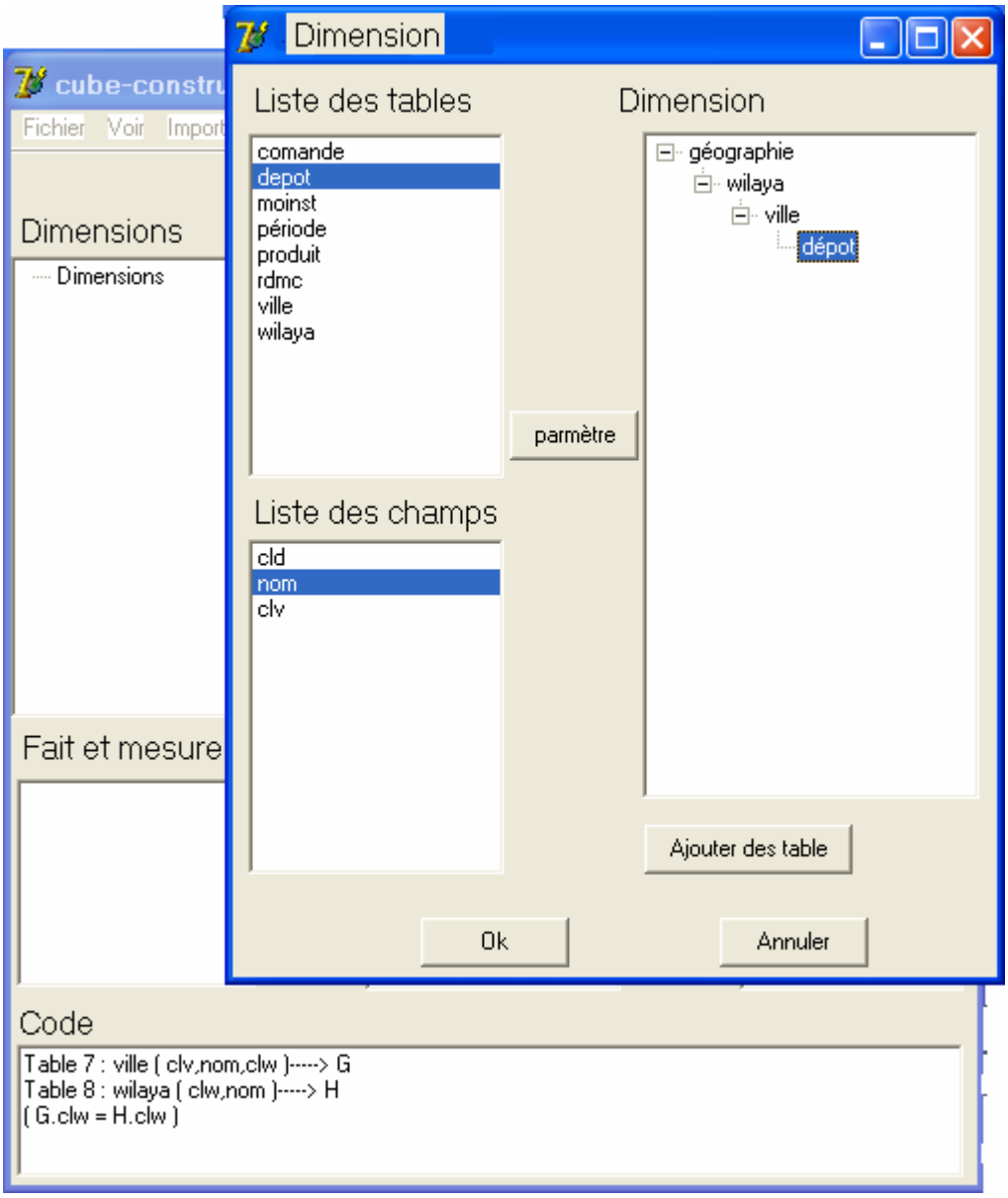


Figure 36 : Détermination d'une dimension.

De manière similaire on peut déterminer le fait et les mesures, c'est-à-dire les tables et les champs représentatifs.

Le résultat obtenu après la détermination des liens, des dimensions et le fait est représenté dans la figure suivant :

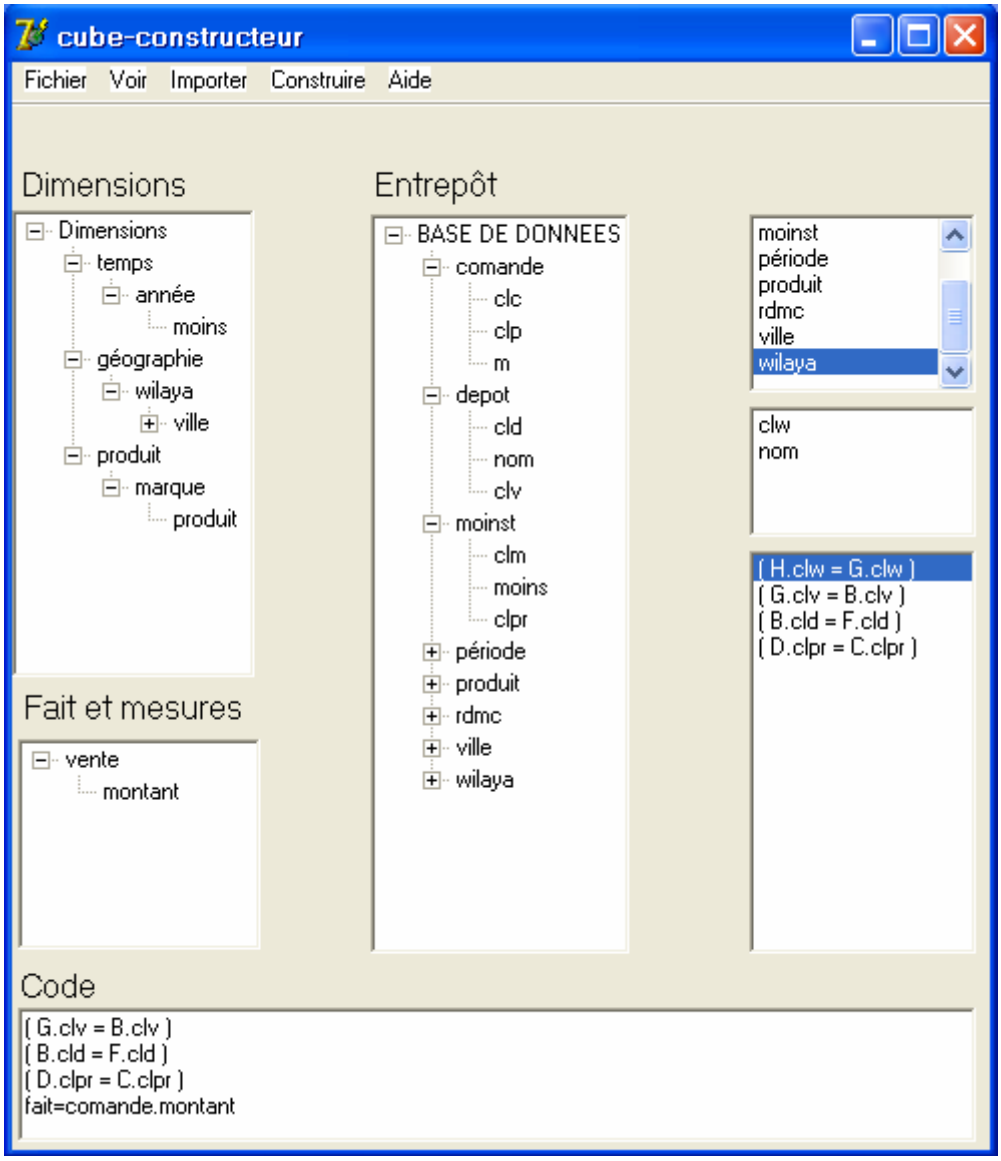


Figure 37 : Représentation de magasin obtenu.

Après la construction du magasin par l'administrateur, l'utilisateur (décideur) peut manipuler et visualiser les données du magasin de manière simple et interactive. Dans la suite nous présentons l'interface dédie à la manipulation du magasin.

8.3 Manipulation des données du magasin multidimensionnel

Notre module d'utilisateur permet de manipuler la base multidimensionnelle du magasin. Les données sont représentées sous forme de hypercube où les arêtes sont constituées par les dimensions et les cases contiennent les mesures. Notre outil permet de visualiser le plan d'affichage de deux dimensions. Les autres plans sont visualisés en changeant les dimensions de manière interactive avec le schéma multidimensionnel.

Par exemple, on peut visualiser l'hypercube correspondant à la mesure "montant" suivant les dimensions "Temps" et "Géographie". La figure suivante affiche le plan des dimensions "années" (en lignes) et "wilaya" (en colonne).

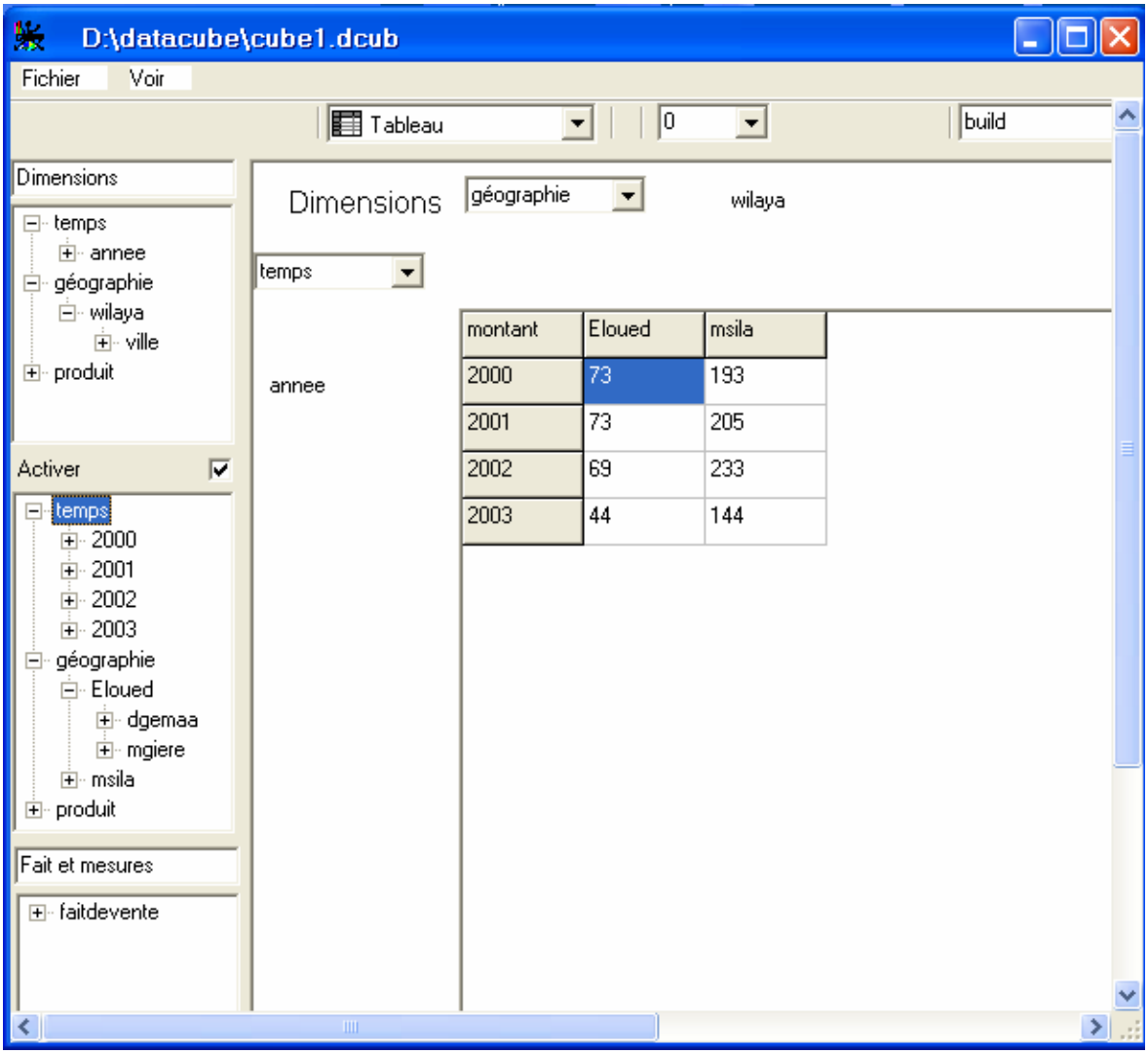


Figure 38 : Représentation tabulaire des données.

8.3.1 Représentation graphique des données

Il est possible de visualiser les tranches d'hypercube sous forme graphique (histogrammes, sectoriels,..). Sur la figure suivante on affiche le plan des dimension "Temps" (année) et "géographie" (wilaya) sous forme d'histogramme.

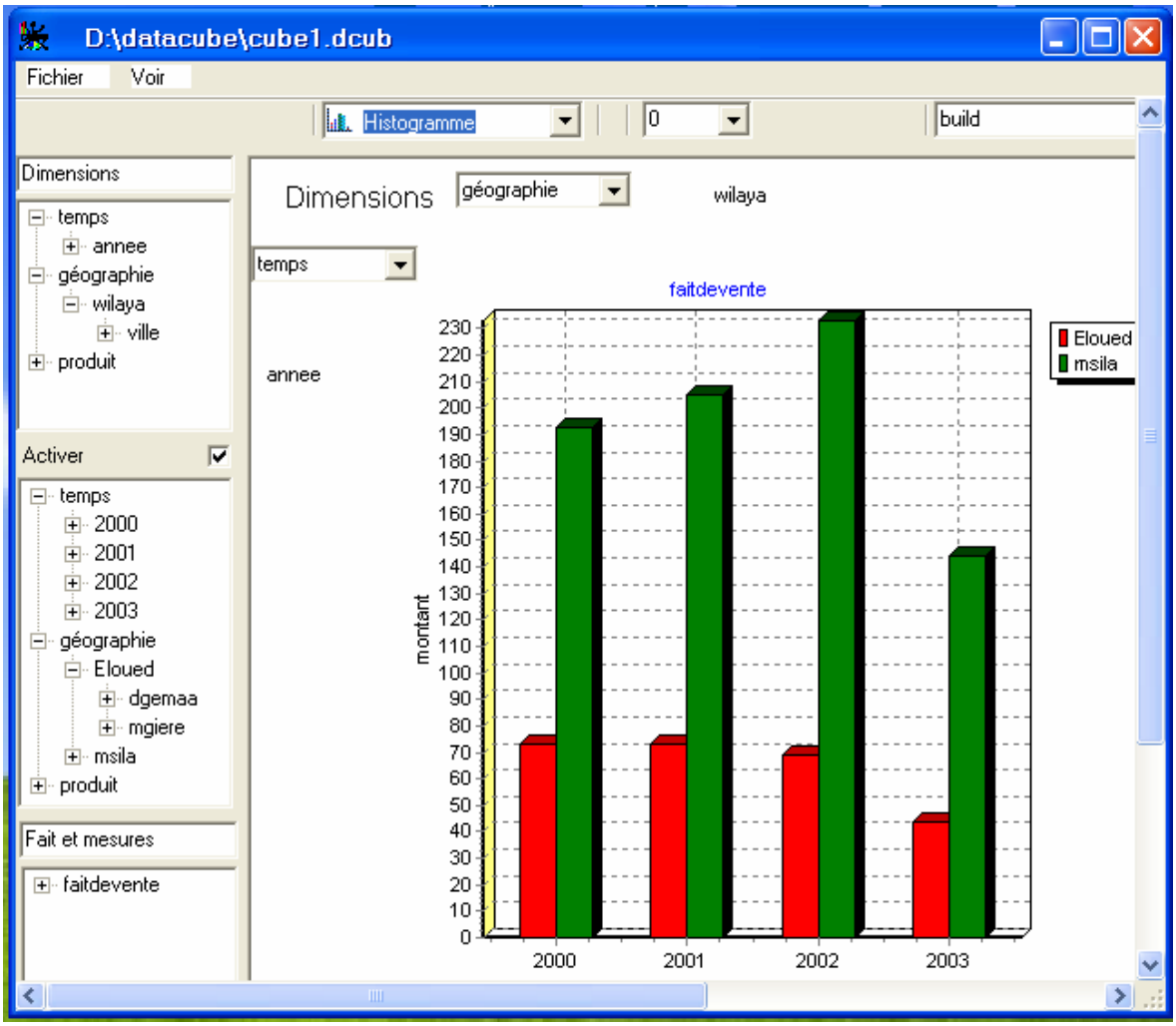


Figure 39 : Représentation graphique.

9. Limitation

Nous avons présenté un système multidimensionnel. Il permet de construire un magasin à partir de données issues d'un entrepôt de données et il donne la possibilité de manipuler cette base multidimensionnelle en visualisant les information sous forme de tableau (ou graphique) correspondant à un plan dans l'hypercube.

Mais notre système a des limites:

- La définition du schéma est fastidieuse et demande une expertise complète de la part de l'administrateur. Ce dernier doit connaître parfaitement l'information contenue dans l'entrepôt de données pour effectuer manuellement toutes les correspondances avec les données du magasin.
- La maintenance du magasin n'est pas réalisée et il est nécessaire d'effectuer une extraction complète à chaque fois que le magasin doit être rafraîchi.
- L'aspect temporelle n'est pas traité (le temps est représenté par un simple attribut)

10. Synthèse

Ce chapitre présente notre démarche pour élaborer le système multidimensionnel. Ce système repose sur une modélisation adaptée (schéma en étoile) dont un fait contient des mesures à analyser, autour ce fait il existe plusieurs dimensions reflétant les perspectives de l'analyse.

La construction du magasin de données nécessite une phase importante pour importer les données de l'entrepôt, on parle d'opération d'extraction.

La manipulation de données du magasin représente le point visible pour les décideurs dans notre système. Elle est faite grâce aux opérations multidimensionnelles.

Nous proposons deux interfaces, l'une dédiée aux administrateurs pour construire le magasin et l'autre pour les décideurs permettant d'accès rapide et interactif au données de magasin (hypercube de données).

Conclusion

Conclusion

Le sujet de ce mémoire s'insert dans la problématique de prise de décision dans les entreprises en exploitant les données de production ou les données provenu de l'extérieur.
La solution est de collecter ces information dans un lieu centralisé de données (entrepôt de données), structurer ces données de manière reflète la vision de décideurs (modèle multidimensionnel), mettre cette structure (magasin de données) disponible pour l'analyse et l'interrogation.
Cette solution présente une démarche connue comme OLAP (on-line analytical processing). Plusieurs travaux sont proposés autour ce processus concernant par exemple : la modélisation, l'organisation multidimensionnelle de données et les opérations multidimensionnelles.

Notre travail rentre dans le domaine d'organisation de données de l'entrepôt dans un magasin de données tout en se basant sur un modèle multidimensionnel (schéma en étoile). L'objectif est d'offrir à l'utilisateur (décideur) un système qui peut être traduire à une interface interactive pour l'analyse, l'interrogation et la visualisation de données de manière multidimensionnelle. L'élaboration de ce système repose sur :

- Une modélisation conceptuelle du magasin.
- Un mécanisme pour l'extraction et la transformation des données dans le magasin à partir d'un entrepôt.
- Une algèbre pour la manipulation du magasin.

Perspectives

Parmi les perspectives que nous envisageons on peut citer:

- **Extension de la modélisation du magasin**. Dans ce mémoire nous avons utilisé un modèle à un seule fait (modèle en étoile), mais dans l'analyse multidimensionnel, il est nécessaire d'utiliser plusieurs faits (modèle en constellation).
- **Opération d'extraction** reste difficile à réaliser et nécessite un expert pour élaborer et alimenter le magasin de données. Ils est utile de proposer des modèles et des méthode pour automatiser cette opération.
- **L'aspect temporel** jeu un rôle vital dans les systèmes décisionnel, donc la représentation du temps par un simple attribut ne permette pas de mieux exploiter les données de l'entrepôt. L'utilisation des modèles s'appuyant sur cet aspect augmente l'utilité d'analyse multidimensionnel.
- **Les décideur** ne sont pas des informaticiens, il est nécessaire d'offrir des interfaces simples et interactives pour agir sur les données du magasin
- Nous avons vu que les entreprises ont recours à des systèmes d'aide à la décision spécifiques, basés sur l'approche multidimensionnelle. Ces systèmes Comprennent des outils d'analyse multidimensionnelle, repondant à des requêtes généralement complexes, par exemple de type <<quel est le chiffre d'affaires par produit et par localisation géographique pour les trois derniers moins >>.

  Ces outils sont les outils meilleurs pour l'analyse, mais ne permettent pas de remarquer que << dans les magasin X, les acheteurs du produit A sont souvent les acheteurs du produit B >>.

  Un autre concept, s'appuyant sur les techniques d'intelligence artificiel peut rependre à cette question; on parle de **Datamining**. Le datamining assiste l'utilisateur dans la recherche de corrélations informelles entre les données.

  Pour ces besoin spécifies d'analyse, il est comme tendance d'intégrer la technique de datamining à notre système pour mieux supporter l'aide à la décision.

Bibliographie

[1] R.Kimball. "The data warehouse toolkit", John Wiley and Sons, 1996.

[2] E.F.Codd. "Providing OLAP (on-line analytical processing) to user-analysts :an IT mandate", Technical Report, E.F. Codd and Associates, 1993.

[3] W.H.Inmon. "Building the Data Warehouse", John Wiley&Sons, ISBN 0471-14161-5, 1994.

[4] M.Golfarelli, D.Maio, S.Rizzi. "Conceptual Design of Data Warehouses from E/R Schemes", In Proceedings of the 31st Hawaii International Conference on System Sciences, Kona (Hawaii, USA), 1998.

[5] P.Marcel . "Manipulation de Données Multidimensionnelles et Langages de Règles", Thèse de Doctorat de l'Institut des Sciences Appliquées de Lyon, 1998.

[6] ALain Fernandez ." Les nouveux tableaux de bord des décideurs" Edition d'organisation, 2e édition 2000.

[7] H. Jagadish, L. V. S. Lakshmanan, and D. Srivastava. "What can hierarchies do for your data warehouses". Proceedings of the International Conference on Very Large Databases, pages 530{541, September 1999.

[8] V.Harinarayan, A.Rajaraman, J.Ullman. "Implementing Data Cubes Efficiently", In Proceedings of ACM SIGMOD Conference, Montreal, Canada, June 1996.

[9] Conservatoire National des Arts et Métiers de Lille."Data warehouse et data mining", Version 1.1, Le 15 Juin 1998.

[10] Olivier Teste."Modélisation et manipulation d'entrepôts de données complexes et historisées", Thèse de Doctorat de DE L'Université Paul Sabatier 2000.

[11] Todd Eavis . "Parallel relational OLAP" , Thèse de Doctorat ,Université Dalhousie , Nova Scotia 2003.

[12]Ladjel Bellatreche ." Techniques d'optimisation des requêtes dans les data warehouses", LISI/ENSMA.
[13]Groupe d'Evolution ." Entrepôt de données pour l'aide à la décision", (http://www.prism.uvsq.fr/dataware/coop/évolution.html).

[14] S. Chaudhuri, U. Dayal. "An overview of data warehousing and olap technology". Sigmod Record, 26(1) :65-74,March 1997.
[15] A. Mendelzon. "Olap : Concepts and products". Talk at University of Toronto, 1997.
[16] M.Djeddaoui ,B.Bouderah,M.S.Maamra ."L'entrepot de données pour l'aide à la decision'", congrès international en informatique appliquée,C.U de BBA ,Algérie 2005.
[17] HEG ."l'informatique décisionnelle", support de cours d'Haute école de gestion de Neuchâtel 2003.

[18] A.Gupta, I.S.Mumick. "Maintenance of Materialized Views:Problems, Techniques, and Applications", IEEE Data Engineering Bulletin, 1995.

[19] J.Widom . "Research problems in data warehousing", Proceedings of the 4th International Conference on Information and Knowledge Management - ACM CIKM'95, November 29-December 2 1995, Baltimore (Maryland, USA).

[20] J.M.Franco, S. de Lignerolles ."Péloter l'entreprise grâce au data warehouse ". Eyrolles 2001 .

[21] L.Cabibbo, R.Torlone. "Querying Multidimensional Databases", Proceedings of the 6th DBPL Workshop, pp 253-269, 1997.

[22] C.Li, X.Wang. "A Data Model for Supporting On-Line AnalyticalProcessing", In Proceedings of the 5th International Conference on Information andKnowledge Management - ACM CIKM'96, November 1996.

[23] M.Gyssen , L.V.S.Lakshmanan. "A Foundation for Multi-Dimensional Databases", In Proceedings of 23rd International Conference on Very Large Data Bases - VLDB'97, Athens (Greece), August 25-29 1997.

[24] R.Agrawal, A.Gupta, A.Sarawagi. "Modeling Multidimensional Databases", ICDE'97.

[25] W.Lehner . "Modeling Large Scale OLAP Scenarios", 6th International Conference on Extending Database Technology - EDBT'98, Valencia (Spain), March, 1998.

[26] W.Lehner, J.Albrecht, H.Wedekind. "Multidimensional Normal Forms", In Proceedings of the 10th International Conference on Scientific and Statistical Data Management - SSDBM'98, Capri (Italy), July 1-3, 1998.

[27] T.B.Pedersen , C.S.Jensen. "Multidimensional Data Modeling for Complex Data", In Proceedings of the International Conference on Data Engineering - ICDE'99, March 1999.

[28] A.O.Mendelzonm, A.A.Vaisman. "Temporal Queries in OLAP",Proceedings of 26th International Conference on Very Large Data Bases - VLDB 2000, Cairo (Egypt), September 10-14, 2000.

Annexe A

Exemple d'un schéma multidimensionnel d'un magasin de données

Le magasin de données contient des données organisées de manière multidimensionnelle issues d'un entrepôt de données pour supporter l'analyse. Dans la suite nous présentons un exemple d'un schéma multidimensionnel décrit algébriquement.

Noms=" Ventesanalysés",

- Vente (montant),
- DIM={D_1, D_2, D_3} où

D_1=(temps,(Moins , année), H^{D1});

D_2=(géographie(dépôt, ville ,wilaya), H^{D2});

D_3=(produit(Nomprod, marque), H^{D3});

Chaque dimension du schéma multidimensionnel est munie d'une hiérarchie.

-H^{D1} ={ , Moins , année }

-H^{D2} ={ dépôt, ville ,wilaya }

-H^{D3} ={ Nomprod, marque }

Au niveau logique on peut obtenir **les schémas relationnels** suivant :

-fait-de-Vente (Cltemps,Clgéographie,Clproduit ,montant).

-temps (Cltemps ,Moins, année) .

-géographie (Clgéographie ,dépôt, ville, wilaya).

-produit (Clproduit ,Nomprod, marque)

Le schéma en étoile correspondant est illustré dans la figure suivante :

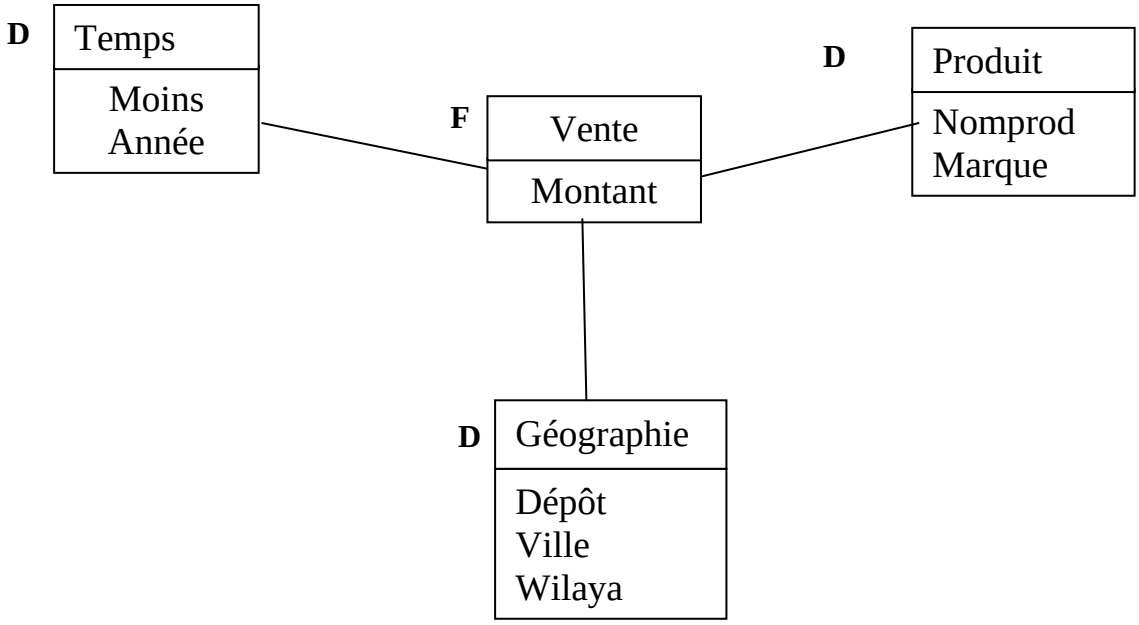


Figure 33 : Schéma en étoile de l'exemple

Remarque : Dans notre étude Le magasin est présenté par un schéma contenant un seul fait, mais le magasin peut présenté par un schéma contenant plusieurs faits représentant les sujets d'analyse d'une activité.

ملخص:

المعالجات التحليلية السريعة (OLAP) أصبحت أحد أهم مكونات أنظمة اتخاذ القرارات الحديثة، حيث تسمح بتحليل كم كبير من المعطيات. هذه الأنظمة تعتمد أساسا على نموذج متعدد الأبعاد يعرف باسم مكعب المعطيات. تصورا، مكعب المعطيات يعطي لمتخذي القرارات إمكانية متابعة المعطيات من وجهات مختلفة (محاور) إجمالا وتفصيلا.

هدفنا هو استغلال هذا النسق لتطوير نظام متعدد الأبعاد يسمح بتنظيم المعطيات الواردة من مخازن المعلومات ضمن نمط متعدد الأبعاد، حيث يتيح هذا الأخير إمكانية التحليل والملاحظة بصورة فعالة.

في دراستنا قدمنا نموذج تصوري لمكعب المعطيات معتمدين على النموذج النجمي، وبالنسبة لعملية جلب المعطيات من مصادرها اقترحنا استعمال الشبكة المترابطة لنمذجة مصدر المعلومات (قاعدة بيانات علائقية).

وفي الأخير عرضنا أسلوبا للتعامل مع مكعب المعلومات يعتمد على التمثيل بالجداول.

الكلمات المفتاحية :
تعدد الأبعاد ، مكعب المعطيات ، مخازن المعلومات ، التحليل ، اتخاذ القرارات.

Abstract:

On-line Analytical Processing (OLAP) has become a fundamental component of contemporary decision support systems and represents a means by which knowledge workers can efficiently analyze vast amounts of organizational data. OLAP relies heavily upon a multidimensional data model known as the data cube. Conceptually, the data cube allows users to view organizational data from different perspectives (axes) and at a variety of summarization levels.

Our subject is to use this approach to elaborate a multidimensional system allows organizing the data extracted from the data warehouse (the data source) in a multidimensional format. This organization permits the data to be observed and analyzed efficiently.

In our study, we present a conceptual modeling for the data cube (we use the star model); we also propose an extraction mechanism based on a connected graph to represent the data source (Relational data base). Finally we present a demarche to manipulate the data cube relies upon a tabular representation.

Key words

OLAP, data cube, data warehouse, data mart, multidimensional systems, decision support systems.

Résumé

Le processus d'analyse en ligne (OLAP) décrit l'approche multidimensionnelle de l'aide à la décision. Ce processus permet de structurer et de stoker les données provenant des bases de production dans un format multidimensionnel connu sous le nom de cube de données. Ce cube organise les données selon les perspectives d'analyse et selon les niveaux de détail.

Notre travail présente une démarche dédie à l'élaboration d'un système multidimensionnel basé sur le concept OLAP. Notre travail repose essentiellement sur une modélisation conceptuelle du cube de données, une mécanisme pour l'extraction de données en utilisant un graphe connexe et finalement sur une démarche pour manipuler les données du cube en utilisant une représentation tabulaire.

Mots clés:

OLAP, cube de données, entrepôt de données, magasin de données, systèmes multidimensionnels, système d'aide à la décision.