N° 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 : **GHELLAB Abdelkrim**

Thème :

CONCEPTION D'UNE BASE DE DONNEES DECISIONNELLE

Soutenu publiquement le : ----/----/-------- devant le jury composé de :

Mr. **BOUDERAH BRAHIM**,	Maître de Conférence, Université de M'sila	**Président**
Mr. **BELOUADAH HOCINE**,	Maître de conférence, Université M'sila	**Rapporteur**
Mr. **MAAMRA Mohamed Said**,	Chargé de recherche, USTHB	**Co-encadreur**
Mr. **ABASSI Moncef**,	Professeur, USTHB	**Examinateur**
Mr. **MIHOUBI Douadi**,	Maître d e conférence - Université de M'sila	**Examinateur**
Mr. **BRAHIMI Mahmoud**,	Chargé de cours, Université de M'sila	**Invité**

Résumé

Le travail s'inscrit dans la problématique générale de la modélisation conceptuelle des Entrepôt de données. Le Domaine que nous avons choisi est relatif à la dérivation d'un schéma conceptuel pour les Entrepôt de données à partir des schémas opérationnels et en particulier les schémas E/R.
Dans ce mémoire, nous avons présenté une approche de conception de schéma conceptuel multidimensionnel à partir du schéma E-R opérationnel, inspiré essentiellement de l'approche de Daniel L. Moddy et Mark A.R. Kortink comme approche de base, ainsi que les travaux de B. Husemman pour son formalise, sans oublier la démarche de R. Kimball, qui s'inscrit dans le cycle de vie dimensionnel. L'objectif principal de cette approche est d'exploiter le schéma E-R initial opérationnel et déduire les faits et les dimensions en utilisant la méthode de classification des entités du schéma E-R initial en trois classes (Transactionnelle, composante, classification), et puis déterminer les différentes hiérarchies existantes, et en premier définir les spécifications des besoins sous forme d'une série des questions ou requêtes pour les futures analyses OLAP, et un tableau de spécification obtenu à partir de l'analyse du schéma E-R initial et les requêtes des décideurs pour classifier les attributs (Mesure, Dimensionnel, Optionnel), et enfin produire le modèle multidimensionnel où nous avons un large choix d'options pour la réalisation de ce modèle. Chacune de ces différentes options (Etoile, Flocon de neige, Galaxie, Plat, etc.) représente le compromis entre la complexité et la redondance, et obéit à des règles de passage du modèle E-R d'entreprise vers le modèle multidimensionnel, et nous avons défini des niveaux de restriction pour toutes les mesures le long des différents chemins d'agrégation dans chaque schéma multidimensionnel de fait.

Les Mots Clés: Entrepôt de données, Fait, Dimension, Magasins, Mesure, OLAP, Hierarchie

---

Abstract

The present work is interested in the general problem of the conceptual modeling of data warehouse. The Field that we have chosen relates to the derivation of a conceptual scheme for data warehouse starting from the operational scheme and in particular E/R ones.

We have presented an approach that permits the conception (design) of a multidimensional conceptual scheme starting from operational E-R scheme, inspired essentially from the approach of Daniel L Moddy and Mark A.R. Kortink as a basic approach and the work of B.Husemman for its formalism, without forgetting the work of R. Kimball which falls under the dimensional cycle of life.

The principal objective of this approach is to exploit the initial operational E-R scheme to deduce the facts and dimensions by using the entities classification method of the initial E-R scheme into three classes (Transactional, component, classification) then determine the various hierarchies that exist. Firstly, the needs specifications are defined in a form of series of questions or requests for the future OLAP analyses. Secondly, a specification table is obtained from the analysis of the initial E-R scheme and the requests of the deciders to classify the attributes (Measurement, Dimensional, Optional). Finally, to produce the multidimensional model where we have a broad choice of options (Star, Snowflake, Galaxy, Flat, etc.) each one represents the compromise between complexity and redundancy and obeys to many rules of passage from the company's E-R model towards the multidimensional model. We have defined levels of restriction for all measurements along the various aggregation paths in each fact multidimensional scheme.

Keys Word: Data Warehouse, Data Mart, Fact, dimension, hierarchie, OLAP, Measurement.

ملخص

هذا العمل يندرج في إطار النمذجة التصورية لمخازن المعطيات. المجال الذي اخترناه يتمثل في إيجاد تصميم تصوري لمخازن المعطيات انطلاقا من تصاميم عملية و بالأخص التصميم (E/R).

في هذه المذكرة، اقترحنا مقاربة لإيجاد تصميم تصوري متعدد الأبعاد انطلاقا من التصميم (E/R) العملي، مستوحاة أساسا من مقاربة دانيال ل.مودي و مارك أ.ر. كورتينك، و كذا باستعمال تشكيل ب.أوزمان، دون أن ننسى سيرورة ر.كمبال، التي تندرج في إطار دورة الحياة البعدوية.

الهدف الأساسي للمقاربة المقترحة هو استعمال التصميم العملي الأولي (E/R) و استخراج الأحداث و الأبعاد باستعمال طريقة تصنيف كائنات التصميم (E/R) الأولي إلى ثلاثة أصناف (معاملاتي، مركب، نصنيف) ، و بعد ذالك تعيين مختلف التدرجات الموجودة، لكن في البداية تحديد المتطلبات على شكل سلسلة من الأسئلة أو طلبات لأجل محللي OLAP فيما بعد، و إنشاء جدول مواصفة انطلاقا من تحليل التصميم (E/R) الأولي و طلبات أصحاب القرار من أجل تصنيف الخصائص (قياس، بعدوي، اختياري) ، وفي النهاية التوصل إلى نموذج متعدد الأبعاد أين يكون لدينا عدة خيارات من أجل إنجاز هذا النموذج. كل من هذه الاختيارات(نموذج نجمي، كومة الثلج، مجرة، طبق، الخ) يشكل الوصلة بين التعقيد و التكرار، و يخضع إلى قواعد المرور من نموذج (E/R) المؤسسة إلى النموذج المتعدد الأبعاد، و كذالك قمنا بتعريف مستويات اقتصار من أجل كل القياسات عبر مختلف سبل التجميع في كل تصميم الأحداث متعدد الأبعاد.

الكلمات المفتاحية

مخزن المعطيات ، الحدث ، بعد، مخزن ، قياس ، OLAP ، تدرج.

Remerciements

En tout premier lieu, je remercie mon dieu, tout puissant, qui m'a éclairé le bon chemin et qui m'aide à réaliser dans les meilleurs conditions (الحمد لله) ,ainsi que mes encadreurs respectivement Mr **BELOUADAH Hocine**, et Mr **Maamra Med Said** qui m'ont guidé avec patience et gentillesse et m'ont fait profiter de leurs grandes expériences ainsi que leurs précieuses remarques qui ont grandement contribué à améliorer la qualité de ce mémoire. Qu'il soit ici assuré de ma profonde gratitude et de mon très grand respect.

Je tiens à remercier sincèrement l'ensemble des membres du Jury qui me font grand honneur d'avoir accepté de juger mon travail.

Je remercie Mr. **BOUDERAH BRAHIM**, Maître de Conférence, Université de M'sila pour l'honneur qu'il me fait en acceptant la présidence de ce jury. Qu'il trouve donc ici l'assurance de ma profonde gratitude.

Je tiens à exprimer tout ma reconnaissance à Messieurs **ABASSI Moncef**, professeur à USTHB, et **MIHOUBI Douadi** Maître d e conférence à l'université de M'sila, et **BRAHIMI Mahmoud** Chargé de cours à l'université de M'sila, pour avoir accepté d'être examinateurs de ce travail et pour le temps qu'il ont investi à l'évaluer malgré leurs nombreuses obligations.

Je voudrais remercier tous mes collègues de la promotion Poste graduation, sans oublier Mr **MAHDJOUBI Rosafi** Chef Département Informatique, et MR **GASMi Abdelkader** Président du Conseil Scientifique, et à toutes les personnes qui ont contribué de prés ou de loin à l'élaboration de notre travail et tout particulièrement à Mr **HIMEUR Fares**, et **BACHIRI Hamza**.

Merci infiniment à toutes mes sœurs et frères, mes collègues du travail.

Dédicaces

A ma chère mère et mon chèr père

A ma chère femme

A mon frère Khaled

A mes Sœurs et mes frères

A mon fils Abdelillah

A ma chère poupée adorée Aichouche Sara

A tous mes amies

Table des matières

INTRODUCTION :

Devant le phénomène de la mondialisation qui a engendré un environnement de concurrence grandissant, la prise de décision est devenue cruciale et vitale pour les dirigeants et les chefs d'entreprises. L'efficacité de cette prise de décision repose sur la mise à disposition des informations pertinentes et d'outils adaptés aux décideurs et gestionnaires. Le problème des entreprises est d'exploiter efficacement d'importants volumes d'informations, provenant soit de leurs systèmes opérationnels, soit de leur environnement extérieur (et souvent mis en quarantaine), pour supporter la prise de décision. Les systèmes traditionnels s'avèrent inadaptés à une telle activité. Afin de pallier cet inconvénient, des systèmes décisionnels ont été développés; fondés sur les entrepôts de données et des outils d'investigation de ces données pour guider, voire suggérer des décisions cruciales. Cependant la mise en oeuvre d'un entrepôt ne peut actuellement s'effectuer à coûts et délais acceptables.

Selon le "Meta Group", 95% des 500 entreprises les plus importantes aux Etats-Unis ont déjà ou sont en train de finaliser la mise en place d'un tel système [**TEST00**], et d'après une enquête *d'information Week* publie en Juin 1998 fait du Data Warehouse la priorité numéro 1 en matière d'évolution des système d'information d'entreprise [**KIM97**]. La plupart de ces systèmes reposent sur un espace de stockage centralisé, appelé **entrepôt de données** (***data warehouse*** *en anglais*) ; son rôle est d'intégrer et de stocker l'information utile aux décideurs et de conserver l'historique des données pour supporter les analyses effectuées lors des prises de décision.

Le schéma multidimensionnel est considéré comme le cœur de l'entreposage de données et comme l'élément de base dans le cycle globale de développement et de

maintenance du data warehouse. Afin d'exprimer les caractéristiques du paradigme multidimensionnel, les approches de modélisation peuvent être basées sur des nouveaux modèles, ce qui nécessite des efforts additionnels pour leurs constructions tel qu'il est le cas dans les travaux de Agrawal [**AGR97**], et Cabibbo [**CAB98**]. C'est pour cette raison que la majorité des approches de modélisation multidimensionnelle sont basées sur l'extension des modèles existants afin d'exprimer les besoins décisionnels. Parmi les travaux basés sur cette approche nous citons les suivants : [**TRU02**], [**AKO01**], [**GOLF98**], l'objectif majeur de ces approches ou travaux est de proposer des modèles multidimensionnels standards permettant de représenter les données opérationnelles dans un niveau plus élevé **[ASK03]**.

Contexte général de la thèse de mémoire

Alors que peu de travaux de recherche à l'instant ont été conduits dans le domaine de présenter une méthode ou démarche complète pour dériver un schéma conceptuel pour les entrepôts de données à partir des schémas opérationnels et en particulier les schémas E/R qui représentent le modèle de données de la quasi majorité des schémas des bases de données opérationnelles en Algérie , pour cela nous essayons de faire un survol sur les méthodes de conceptions des schémas conceptuels multidimensionnels à partir du modèle relationnel et de faire une comparaison entre ces méthodes afin de présenter une démarche "Complète " qui peut aider un informaticien d'une organisation à exploiter leurs schémas opérationnelles de données et de tirer le schéma conceptuel du data warehouse. Cette approche sera inspirée essentiellement des travaux de *Daniel L. Moody et Mark A.R. Kortink* et al, et *Bobo Huseman* et al, et *R. Kimball*

Pour cela nous allons suivre dans ce mémoire les lignes suivantes:

- **P**résentations de tous les concepts relatifs aux systèmes décisionnels, Les entrepôts de données (Data Warehouse), les magasins (data marts), [1ere Partie, 2eme Partie, 3eme Partie]
- **U**ne présentation des principales méthodes de conception des entrepôts de données, [4eme Partie].
- **E**tude de cas pour illustrer la démarche proposée, [5eme Partie].

.

PARTIE 1 SYSTEME D'INFORMATION DECISIONNEL

I.1. INFORMATIQUE DECISIONNELLE VS OPERATIONNELLE

I.1.1. Informatique Décisionnelle Vs Opérationnelle:

Loin d'être qu'un simple phénomène de mode, l'informatique décisionnelle est devenue incontournable pour toute entreprise moderne.

Les enjeux économiques et financiers des entreprises sont devenus tels que l'intuition et la réflexion ne suffisaient plus à prendre une décision. Dans un monde où la concurrence fait rage et où chaque choix stratégique peut être vital, les décideurs devaient se munir d'outils informatiques puissants et fiables visant à faciliter la tâche de pilotage de l'entreprise. Et ceci en offrant les moyens d'exploiter les données stockées dans les bases de données de l'entreprise souvent restées au stade d'archives, et d'en extraire le contenu informationnel qui servira de miroir de l'état de l'entreprise et qui permettra la compréhension du passé et du coup à une meilleure anticipation du futur.

Définition 1 :

Une définition simple consiste à dire que l'informatique décisionnelle en anglais *« business intelligence »* est la branche de l'informatique qui permet l'exploitation des données de l'entreprise dans le but de faciliter la prise de décision. C'est-à-dire, la compréhension du fonctionnement actuel et l'anticipation des actions pour un pilotage éclairé de l'entreprise.

Définition 2 :

Une deuxième définition plus technique veut que nous présentions en premier lieu les différents niveaux d'un système d'information et que nous définissions ce qu'est un système d'information décisionnel.

I.1.1.1 Les différents niveaux d'un système d'information d'une organisation:

Résumé dans le schéma ci-dessous, un système d'information se décompose en trois niveaux :

A- Le niveau opérationnel : Concerne les données relatives aux différentes fonctions de l'entreprise, il s'agit des bases de données résultantes des sources d'information internes.

B- Le niveau décisionnel : Constitue une synthèse des données opérationnelles, choisies pour leur pertinence. Ce niveau concerne des données qui, agrégées, intégrées et organisées sur base de structures particulières de stockage volumineux, résultent en informations pertinentes à la décision.

C- Le niveau stratégique : le niveau le plus élevé dans la hiérarchie, concerne l'orientation des informations résidant au niveau décisionnel en vue de fournir des indicateurs pertinents. Ce niveaux fourni au décideur d'une part, des systèmes de pilotage qui lui fournissent une série de tableaux de bord et de synthèse, très souvent enrichis de fonctionnalités statistiques et de simulations, et d'autre part sur des outils d'extraction, de gestion de connaissances qui permettent de mettre en évidence des corrélations entre des événements apparemment non liés. La finalité du niveau stratégique est le pilotage de l'entreprise dans une vision stratégique à long terme [**Van01**].

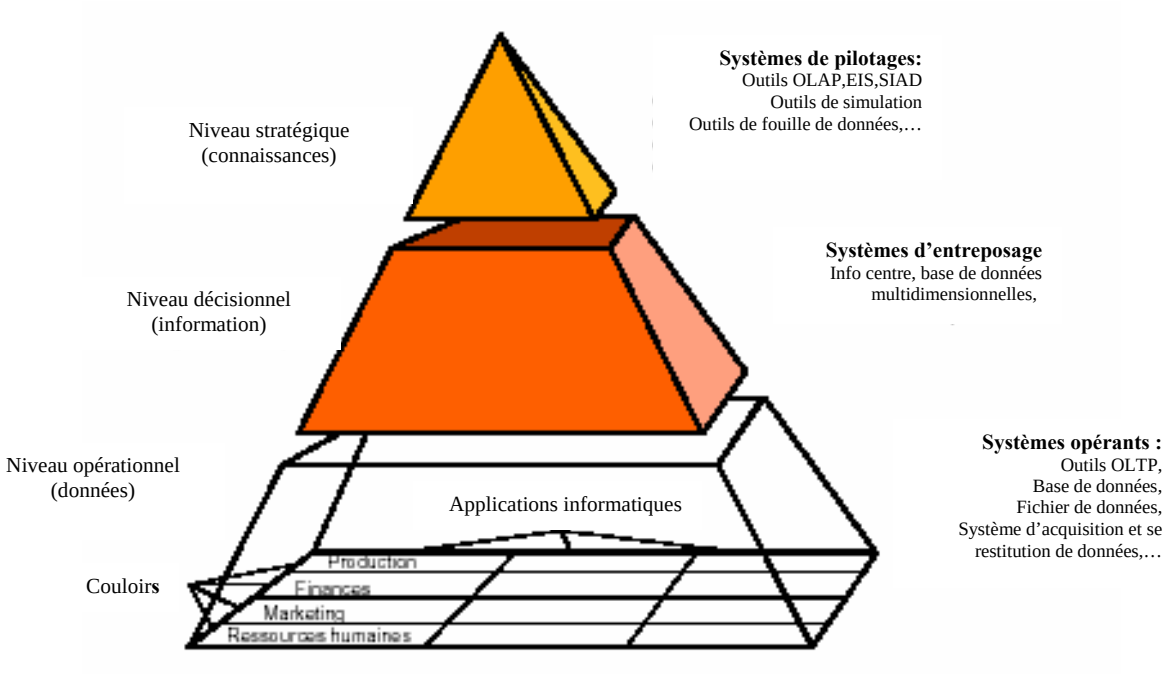


Figure N°1 :Les différents niveaux d'un système d'information

I.1.2 - Définition d'un Data Warehouse :

Toute entreprise possède actuellement d'importants volumes de données, stockés le plus souvent dans différents médias (bases de données, documents papiers,...) et a besoin d'outil permettant une exploitation efficace et performante de ces données pour l'aider dans ses prises de décision.

Les entrepôts de données apportent des solutions à cette problématique. Un entrepôt se définit par Bill Immon référence, considère comme le père du concept, dans son livre "Using the Data Warehouse", il en donne la définition suivante, qui fait référence : "*Le Data Warehouse est une collection de données intégrées, orientées sujet, non volatiles, historisées, résumées et disponibles pour l'interrogation et l'analyse*" **[FRA00], [EVO00]**.

Il permet de stocker des données nécessaires à la prise de décision ; il est alimenté via des extractions de données portant sur les bases de production (sources de données) et la saisie de données quotidiennes **[EVO99]**.

Les données d'un Entrepôt de données respectent donc les caractéristiques suivantes :

– **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 sont organisées autour des thèmes qui ont un intérêt majeur pour l'entreprise, le but de cette organisation est de disposer de l'ensemble des informations utiles sur un thème le plus souvent transversal aux structures fonctionnelles et organisationnelles de l'entreprise tels que : Le client, Le Produit, Les Ventes…

Cette orientation par thème va permettre à l'entreprise de développer son système décisionnel progressivement, c'est une approche par itération.

- **Non volatiles** :

Les données de l'entrepôt sont essentiellement utilisées en mode de consultation; elles sont très rarement modifiées, la non volatilité des données est en quelque sorte une conséquence de l'historisation.

– **Historisées** :

La prise en compte de l'évolution des données est primordiale pour la prise de décision et notamment les prédictions. Dans un système de production ; la donnée est mise à jour à chaque nouvelle transaction. Dans un Data Warehouse, 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.

– **Résumées** :

Les informations issues des sources de données doivent être agrégées (ou résumé) 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.

De plus, l'entrepôt de données offre à l'entreprise les avantages suivants [**EVO00**] :

– Il constitue une collection de données centralisées disponibles 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 différents niveaux de détails.

– Il permet de développer différents thèmes d'analyse (réorganisation en fonction des sujets à analyser).

I.1.3 Composantes Architecturales d'un Data Warehouse:

L'architecture des systèmes décisionnels met en jeu quatre éléments essentiels : les sources de données, l'entrepôt de données, les magasins de données et les outils d'analyse et d'interrogation.

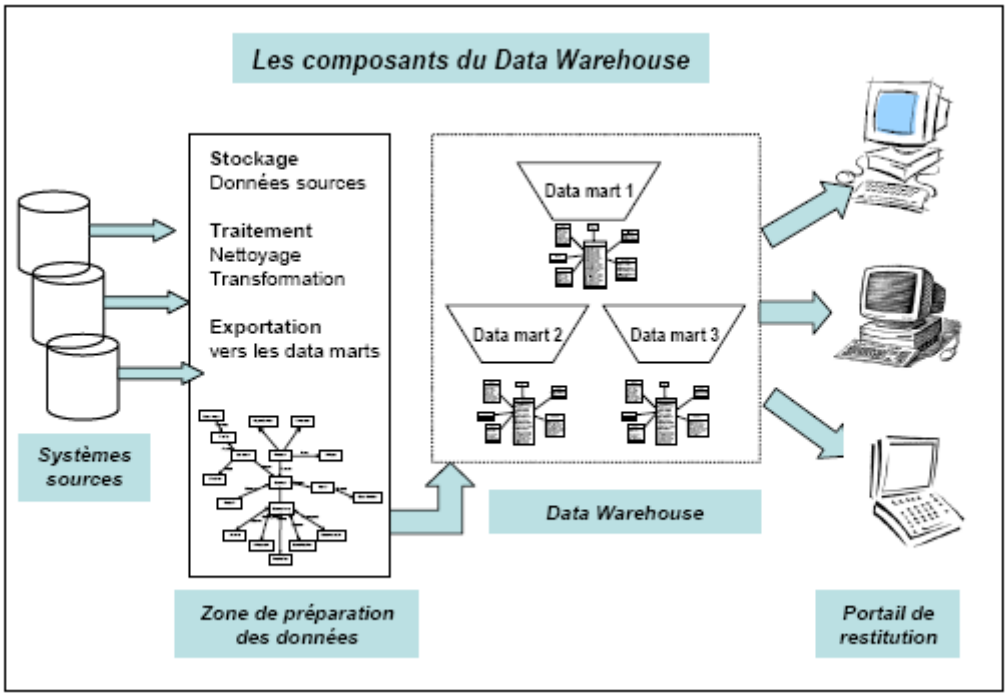


Figure N° 2 : Les Composants du Data Warehouse [BOL02]

√ Les sources de données: sont nombreuses, variées, distribuées et autonomes. Elles peuvent être internes (bases de production de l'entreprise) ou externes (Données fournies par les partenaires tels que: Les fournisseurs, clients, Administration Publiques, Documentation Juridiques) à l'entreprise.

√ L'entrepôt de données: est le lieu de stockage centralisé des informations utiles pour les décideurs. Il met en commun les données provenant des différentes sources et conserve leurs évolutions.

√ 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, et principalement dédie à une classe de décideurs.

√ Les outils d'analyse : permettent de manipuler les données suivant des axes d'analyses. L'information est visualisée au travers d'interfaces interactives et fonctionnelles dédiées à des décideurs souvent non informaticiens (directeurs, chefs de services,…).

I.1.5. Entrepôts des données et Magasins des données

Il est important de distingue les entrepôts et les magasins [**TEST00**]:

☒ *L'entrepôt de données* est le lieu de stockage centralisé d'un extrait des bases de production. Cet extrait concerne les données pertinentes pour le support à la décision. Elles sont intégrées et historisées. L'organisation des données est faite selon un modèle qui facilite la gestion efficace des données et leur historisation.

☒ *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.

I.1.5. LA MISE EN OEUVRE D'UN DATA WAREHOUSE

Quatre caractéristiques ont des effets déterminants sur la démarche de conception d'un Data Warehouse **[FRA00]**:

a- Les évolutions technologiques : un système d'information peut se construire par intégration d'un certain nombre de composants, chacun pouvant être choisi par rapport à son contexte d'utilisation. L'entreprise défini son architecture en fonction de ses besoins.

b- *La stratégie de l'entreprise:* le Data Warehouse est très proche de la stratégie de l'entreprise. L'objectif du Data Warehouse se définit en terme de métier. Il faut donc impliquer les utilisateurs ayant le plus de connaissances dans leur entreprise ou dans leur métier.

c- *L'amélioration continue:* un Data Warehouse doit évoluer en fonction des demandes utilisateurs ou des nouveaux objectifs de l'entreprise.

d- *La maturité de l'entreprise:* certaines entreprises ont déjà un système décisionnel. D'autres n'ont aucun acquis. Dans tous les cas, il n'existe pas de cadre figé pour la conception d'un Data Warehouse. Chaque entreprise doit adapter le projet à son contexte, en ne perdant pas les objectifs de vue. Cet objectif est de mettre en place un système d'information cohérent et intégré, le système devant être décomposé en applications, chacune s'intégrant dans le Data Warehouse.

I.1.5.1 - DECOUVRIR ET DEFINIR LES INITIATIVES

Cette phase consiste en l'étude stratégique du Data Warehouse et la définition du plan d'action.

I.1.4.1.1 - L'étude Stratégique

Pendant l'étude stratégique, il faut :

- Informer et motiver les personnes concernées dans l'entreprise.
- Impliquer les managers, les équipes opérationnelles, les équipes informatiques: phase d'identification et de compréhension des enjeux métier/entreprise.
- Identifier les projets Data Warehouse.

L'étude stratégique permet d'identifier la stratégie de l'entreprise, son organisation, les processus qu'elle met en œuvre, la culture de l'entreprise.
Le but est de déterminer les domaines pour lesquels la mise en place d'un Data Warehouse peut être le plus bénéfique.

I.1.5.1.2 - Le Plan d'action

Pour mettre en place le plan d'action, il faut :

- Vérifier la faisabilité de chaque projet (s'assurer de l'existence et de la qualité des données, des possibilités techniques, des possibilités organisationnelles).
- Estimer les ressources pour chaque projet.
- Séquencer et planifier les projets.

I.1.5.2 *- L'INFRASTRUCTURE*

Il s'agit de déterminer l'infrastructure technologique et organisationnel nécessaire à la mise en place du Data Warehouse.

I.1.5.2.1- L'infrastructure Technique

Des choix technologiques en phase avec la politique de l'entreprise doivent être faits à plusieurs niveaux :

- Les fournisseurs : faut-il prendre un seul fournisseur (ce qui facilite la politique d'intégration et en réduit les coûts de mise en œuvre) ou assembler les meilleurs offres du marché (ce qui apporte une flexibilité, une adaptation à chaque projet, mais coûte beaucoup en intégration).
- Les outils : faut-il construire, acheter ou faire avec l'existant.
- Comment sera utilisé le Data Warehouse, par qui, comment sera structuré l'organisation qui l'exploitera.
- Faut-il une architecture centralisé (Data Warehouse), distribuée (plusieurs Data Mart), ou une architecture répliquée (un Data Warehouse et plusieurs Data Mart).
- La structure de stockage, sera-t-elle relationnelle, multidimensionnelle, hybride (Data Warehouse en relationnel, Data Mart en multidimensionnel).
- Choisir le matériel : selon les volumes envisagés, les utilisateurs concernés, l'architecture visée, la flexibilité attendue.

- Organiser l'administration des systèmes et la gestion de la sécurité.

I.1.5.2.2 - L'infrastructure Organisationnelle

Parallèlement aux choix technologiques, il faut :

- Déterminer la logistique et l'organisation nécessaires à la concrétisation des initiatives.
- Répartir les tâches entre les équipes de développement et les équipes d'exploitation : déterminer l'alimentation du Data Warehouse, l'administration.
- Déterminer les flux d'informations entre le Data Warehouse et les utilisateurs.

I.1.5.3 - LA MISE EN ŒUVRE DES APPLICATIONS

La démarche proposée est une démarche en cinq étapes :

- La Spécification,
- La Conception,
- La Mise En Œuvre et l'intégration,
- Le Déploiement Et La Mise En Place Des Accompagnements,
- Les Mesures.

Ces étapes correspondent à celles de mise en place d'un projet informatique.

Pendant l'étape de spécification, les différentes étapes des initiatives sont définies et planifiées de manières plus détaillées.

Il est recommandé de faire attention aux coûts cachés que peuvent entraîner les technologies informatiques.

L'étape de mesure permet de faire le bilan de la réalisation et de capitaliser les réussites et échecs rencontrés pendant le développement de l'application.

Deux visions du Data Warehouse cohabitent dans l'approche précédente :

- Une vision entreprise : chaque projet défini dans la première phase (initiative) est construit de manière indépendante et répond à un objectif métier délimité, tout en s'intégrant dans le Data Warehouse.

- Une vision projet : les projets identifiés deviennent des applications. Donc le processus est itératif.

Il n'existe pas de démarche complète et universelle pour la mise en œuvre d'un data Warehouse. Toute approche doit être adaptée à l'entreprise.

I.1.5.4 - L'ADMINISTRATION DES DONNEES

Comme tout autre système informatique, un Data Warehouse s'administre.

Dès la phase de conception de l'architecture, il faut penser à l'administration des données, c'est une des fonctions les plus importantes du Data Warehouse.

Cette fonction est d'autant plus importante que le Data Warehouse évolue au fur et à données, permettant de décrire, stocker et diffuser les méta-données associées.

Cette mise en place passe par l'organisation d'une fonction d'administration des données à plusieurs niveaux, par la définition de normes et de procédure d'administration des référentiels.

I.1.6. SYSTEMES OLAP VERSUS SYSTEMES OLTP

I.1.61. OLAP Versus OLTP

Les bases de données sont utilisées dans les entreprises pour organiser les importants volumes d'informations contenus dans leurs systèmes opérationnels. Ces données sont gérées selon des processus transactionnels en ligne (OLTP : "*On-Line Transactional Processing*".

L'exploitation de l'information contenue dans ces systèmes opérationnels est devenue une préoccupation essentielle pour les dirigeants qui sont appelles à prendre des décisions par une meilleure connaissance de leur activité, et par conséquent les entreprises sont donc à la recherche de systèmes supportant efficacement les applications d'aide à la décision. Ces applications décisionnelles

utilisent des processus d'analyse en ligne de données (OLAP : "*On-Line Analytical Processing*". Ces processus répondent aux besoins spécifiques des analyses d'information **[Codd93] [TEST00]**.

*I.1.6.2. - Les 12 Règles OLAP de **E.F Codd***

OLAP est le terme pour décrire l'approche dimensionnelle de l'aide à la décision. Tout comme OLTP, OLAP a été proposé par E. F. Codd. Cette philosophie comprend douze critères qui représentent l'étalon de mesure servant à comparer les systèmes d'aide à la décision. A ces douze critères, 6 ont été ajoutés en 1995. Il faut noter que ces six critères supplémentaires sont rarement cités et utilisés **[HES03]**.

1 - Vue multidimensionnelle

Permet d'avoir une vision multidimensionnelle des données.

2 – Transparence du serveur OLAP à différents types de logiciels

L'utilisateur ne doit pas se rendre compte de la provenance des données si celles-ci proviennent de sources hétérogènes; ces sources peuvent être un fichier Excel, une base de données de production ou même un fichier texte.

3 – Accessibilité à de nombreuses sources de données

OLAP est décrit comme un middleware qui se place entre les sources de données hétérogènes et un front-end (sous la forme d'un data warehouse). Il doit donner accès aux données nécessaires aux analyses demandées afin de présenter à l'utilisateur une vue simple et cohérente. Ils doivent aussi savoir de quel type de systèmes proviennent les données.

4 - Performance du système de Reporting

Les performances ne doivent pas être diminuées lors de l'augmentation du nombre de dimension ou de la taille de la base de données, mais proportionnelles à la taille des réponses retournées.

5 - Architecture Client/Serveur

Il est essentiel que le produit soit en Client-Serveur mais aussi que les composants serveurs d'un produit OLAP intègrent facilement ses différents clients.

6 - Dimensions Génériques

Chaque dimension doit être équivalente par rapport à sa structure et à ses capacités opérationnelles.

7 - Gestion dynamique des matrices creuses

Le système OLAP ajuste automatiquement son schéma physique pour s'adapter au type du modèle et au volume des données (plus on dispose de place plus on peut agréger).

8 - Support Multi-Utilisateurs

Les outils OLAP doivent fournir des accès concurrents, l'intégrité et la sécurité.

9 - Calculs à travers les dimensions

Les calculs doivent être possibles à travers toutes les dimensions (les agrégats doivent être faits dans toutes les dimensions).

10 - Manipulation intuitive des données

La manipulation des données se fait directement à travers les cellules d'une feuille de calcul, sans recourir aux menus ou aux actions multiples. Il doit permettre l'analyse intuitive dans plusieurs dimensions au final.

11 - Souplesse et facilité de constitution des rapports

Lors de la création de rapport, les dimensions peuvent être présentées de n'importe quelle manière.

12 - Nombre illimité de niveaux d'agrégation et de dimensions

Dimensions et niveaux d'agrégations illimités.

I.1.6.3.Comparaison des caractéristiques des processus OLAP et OLTP

	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
Administration Des Données	Forte disponibilité Sauvegardes fréquentes Peu de maintenance off-line	Disponibilité faible Sauvegardes peu fréquentes mais très volumineuses Beaucoup de maintenance mais en off-line

Tableau 1: **Comparaison des processus OLTP et OLAP.**

I.2. CYCLE DE VIE DECISIONNEL:

La réussite de l'implémentation d'un entrepôt de données dépend de l'intégration adéquate de nombreux composants et taches, Il ne suffit pas de posséder le modèle de données parfait ou la meilleur technologie; il s'agit de

coordonner les multiples facettes du projet de data warehouse. Donc une exposition d'une méthodologie globale pour l'implémentation d'entrepôt de données par le cycle de vie dimensionnel est nécessaire, le schéma ci-dessous représente une succession de taches de haut niveau nécessaire à la conception au développement et au déploiement d'un entrepôt de données efficace **[KIM00]**.

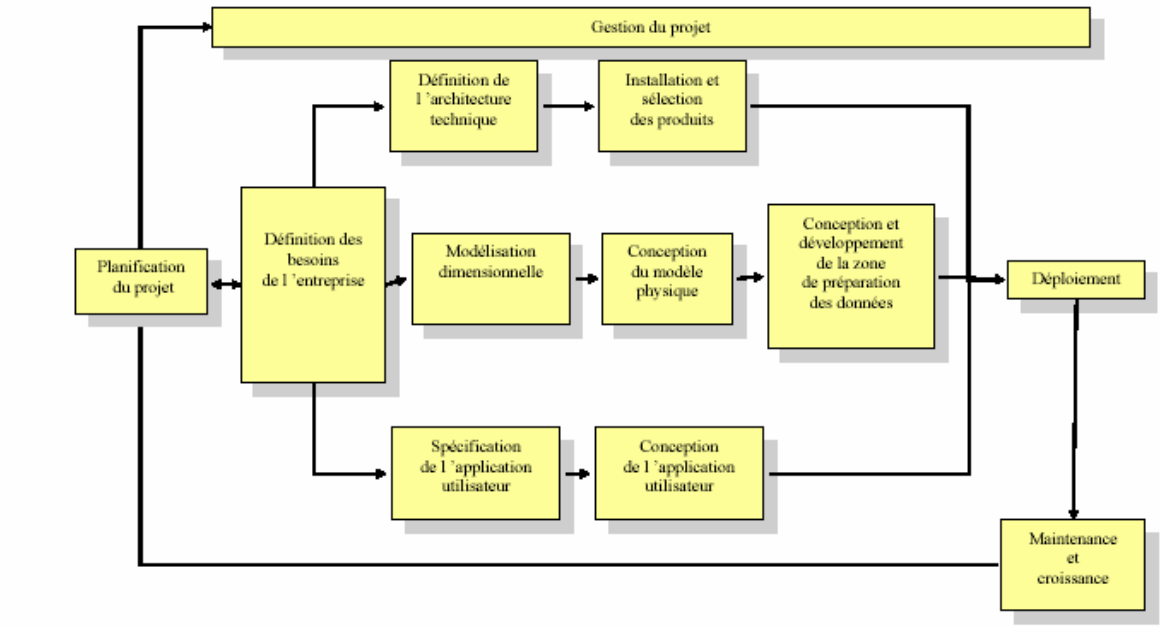


Figure N° 3 : Schéma d'un cycle de vie multidimensionnel **[KIM00]**.

I.2.1. Planification du projet:

La planification du projet aborde la définition et l'étendue du projet de data warehouse, elle se concentre sur les besoins en termes de ressources et de niveau de qualification, couples aux affectations des taches, a leurs durées et a leurs séquencement. La planification dépend des besoins comme l'indique la flèche double dans le schéma.

I.2.2. Définition des besoins:

Il est essentiel de bien comprendre les utilisateurs et leurs besoins, sinon l'entrepôt deviendra rapidement un exercice vain de la part de l'équipe des concepteurs.

L'approche utilisée pour identifier les besoins analytiques diffère de manière significative de la traditionnelle analyse des besoins basés sur les données. Les besoins une fois définis constituent le point de départ de trois trajectoires parallèles que sont la technologie, les données et les interfaces utilisateurs.

I.2.3 Modélisation dimensionnelle

C'est la définition des besoins qui détermine quelles sont les données requises pour répondre aux besoins d'analyse des utilisateurs. La conception du modèle logique de données commence par la construction d'une matrice représentant les processus métier clé et leurs dimensionnalités. A partir de cette matrice, il faut effectuer une analyse plus détaillée des données du (des) système(s) source(s) opérationnels. Le résultat de cette analyse est le modèle dimensionnel.

Ce modèle identifie la granularité de la table de fait, les dimensions associées avec leurs attributs et leurs hiérarchisations. Cet ensemble d'activités s'achèvera sur le développement d'une mise en correspondance des données sources et cibles dans des méta données.

I.2.4. Conception du modèle physique

La conception physique d'une base de données définit les structures nécessaires pour l'implémentation du modèle dimensionnel. Les éléments fondamentaux sont la détermination des règles de nommage des objets, la mise en place de l'environnement de la base de données. L'indexation primaire, les stratégies de partitionnement et les agrégations primaires sont également définies. La

conception du modèle physique est fortement dépendante de la machine utilisée pour l'entrepôt.

I.2.5. Conception et développement de la zone de préparation des données

La conception de la zone de préparation des données (staging area) constitue généralement la tache la plus sous-estimée du projet entrepôt de données. Le processus de préparation se déroule en trois phases majeures :

- Extraction
- Transformation
- Chargement (Loading)

Le processus d'extraction des données révèle généralement le problème de la qualité des données, qui influence de manière significative la crédibilité de l'entrepôt. L'extraction est compliquée par le fait qu'il faut la construire avec deux processus distincts, le peuplement initial et le chargement régulier et incrémentiel.

I.2.6. Définition de l'architecture technique

Cette étape définit la vision globale de l'architecture technique à mettre en Œuvre. Elle nécessite la prise en compte de trois facteurs :

- Les besoins
- L'environnement existant
- Les orientations techniques stratégiques planifiées

En plus de l'architecture supportant l'entrepôt, il est nécessaire de mener des réflexions sur les outils de conception de la zone de préparation des données et des outils de restitutions

I.2.7. Choix technologique et mise en œuvre

A partir de l'étude de l'architecture technique il faut sélectionner les composants spécifiques, telle plate-forme(s)matérielle(s)et logicielle(s), SGBD

outils d'extraction et restitution à mettre en Œuvre. Une fois les produits évalués et sélectionnés, ceux-ci doivent être installés et testés méticuleusement afin de garantir une intégration adéquate d'un bout à l'autre de l'environnement de l'entrepôt.

I.2.8. Développement de l'application utilisateur

Il est recommandé de définir une série d'applications standard destinées aux utilisateurs finaux, car tous n'ont pas besoin d'un accès ad hoc à l'entrepôt. Les spécifications de l'application décrivent les maquettes d'états, les critères de sélection laissés à l'utilisateur et les calculs nécessaires.

I.2.9. Déploiement:

Le déploiement est le point de convergence de la technologie, des données et des applications utilisateurs. Une planification est indispensable pour gérer le déploiement qui comprend également la formation des utilisateurs, les processus de communication, le support utilisateur, la prise en compte des demandes d'évolution et de correction.

I.2.10. Maintenance et croissance:

Après le déploiement initial de l'entrepôt, c'est sa vie qui commence. Il faut s'assurer de fournir un service de support et de formation continue. Il faut également s'assurer que les processus mis en place pour la gestion de la zone de construction vont faire fonctionner l'entrepôt en continu et efficacement. Il est également important de mesurer périodiquement les performances de l'entrepôt et de son acceptation dans l'entreprise. L'entrepôt va donc évoluer et croître et le changement doit être perçu comme un facteur de succès et non d'échec. Des processus de hiérarchisation des priorités doivent bien sur être mis en place afin de gérer les demandes des utilisateurs en termes d'évolution et de croissance.

I.2.11. Gestion du Projet:

La gestion du projet garantit que les activités du cycle de vie restent sur la bonne voie et sont bien synchronisées. Cela consiste à contrôler l'état d'avancement du projet, la détection et la résolution des problèmes et le contrôle des changements afin de garantir l'accès aux objectifs du projet.

PARTIE II: ARCHITECTURE D'UN SYSTEMES DECISIONNELS:

II.1 - ARCHITECTURE D'UN SYSTEMES DECISIONNELS:

Un **système décisionnel** est un système d'information dédié aux applications décisionnelles **[Test00]**, dans cette partie nous allons présenter l'ensemble des concepts de base de l'architecture du data Warehouse et qui regroupe l'ensemble d'informations et d'outils mis à la disposition des décideurs pour supporter de manière efficace la prise de décision.

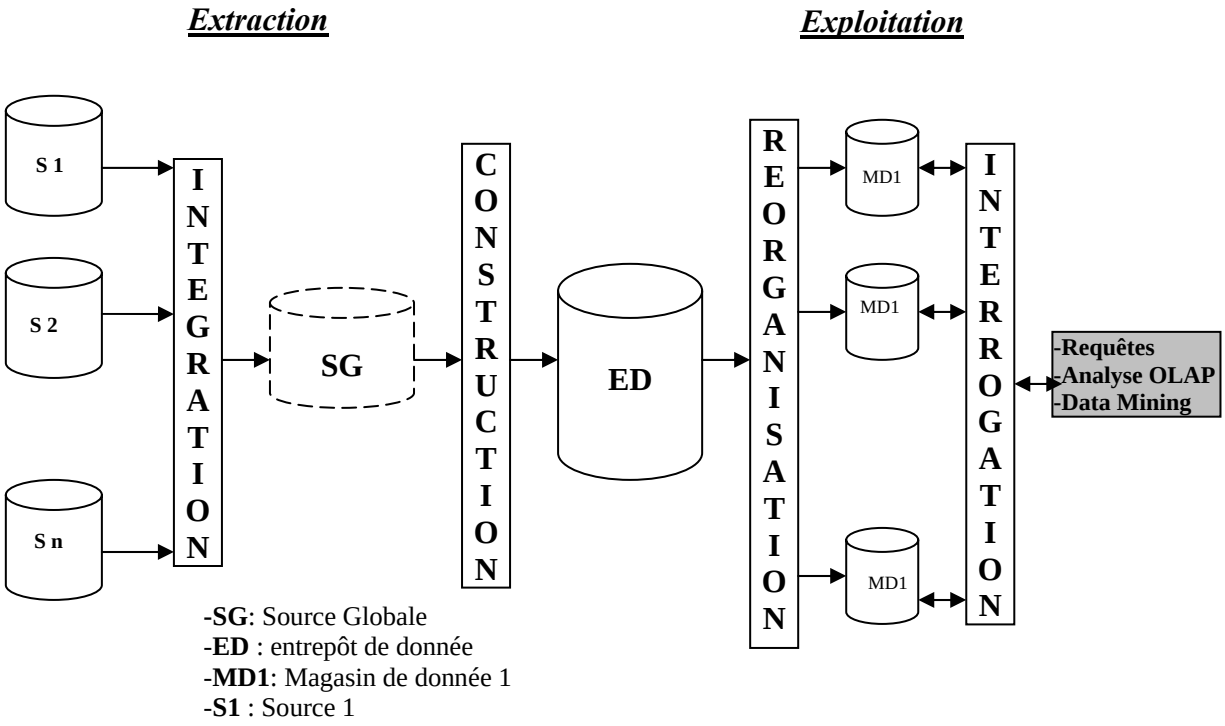


Figure N° 4 : Architecture détaillée d'un système décisionnel

II.1.1. Intégration/Construction/Réorganisation/Interrogation :

II.1.1.1 L'intégration

Se propose de résoudre les problèmes d'hétérogénéité (systèmes, modèles, formats et sémantiques des données,…) des différentes sources de données en intégrant celles-ci dans une source globale. Cette source globale est virtuelle,

c'est à dire que les données utilisées pour la décision restent stockées dans les sources de données et sont extraites uniquement au moment des mises à jour de l'entrepôt **[TEST00]**.

Diverses architectures réalisent la production de données intégrées. Elles ont en commun les composantes suivantes **[EVO99]** :

(1) - Des méta données,

(2) - Intégration des schémas des sources

(3) - Des outils d'extraction de données,

(4) - Des outils de nettoyage de données,

(5) - Des outils d'intégration de données.

II.1.1.1.1- Les meta-données

Dans le contexte d'un entrepôt de données, "décrire une donnée" consiste principalement à indiquer comment l'obtenir à partir des sources. Les meta-données jouent un rôle important d'une part dans les algorithmes d'extraction, de rafraîchissement et d'intégration, et d'autre part dans la présentation d'une vision globale des données aux administrateurs et aux utilisateurs.

Les meta-donnees regroupent l'ensemble des informations concernant le Date Warehouse et les processus associés. Elles sont intégrées dans un référentiel. Les principales informations sont destinées [**FRA00**] :

- A l'utilisateur final (Information sur la sémantique des données utilisées).
- Aux équipes responsables de l'acquisition des données du Date warehouse (Information sur la localisation de la donnée dans le système de production, etc.).
- Aux équipes d'administration de la base de données (Information sur la structure de la base de données qui implémente le Data Warehouse).
- Aux équipes de production (Information sur les procédures de chargement, historique des Mise a jour, etc.).

Une partie des meta-données est propre à une application (ex : description des attributs des sources), et des données de connaissances générales (ex : tables de conversion de monnaies). D'autres sont obtenues par personnalisation de connaissances générales (ex : dictionnaire de synonymes).
Généralement, elles représentent toutes les informations nécessaires à l'accès, à la compréhension et à l'exploitation des données du Data Warehouse. [**CNAM98**], [**FRA00**].

Type d'information et Signification

Sémantique : Que signifie la donnée ? A quoi sert cette donnée?

Origine : D'où vient-elle, où, par qui est-elle créée ou mise à jour

Règle de calcul : Règle de calcul.

Règle d'agrégation : Périmètre de consolidation

Stockage, format : Où, comment est-elle stockée, sous quel format

II.1.1.1.2- Intégration des schémas des sources

L'intégration des schémas des sources fait apparaître des conflits, depuis longtemps bien répertoriés dans la littérature. Les principaux conflits pouvant survenir entre deux schémas sont les suivants :

- - Problèmes de terminologie.
- - Incompatibilités de contraintes.
- - Conflits de structures.
- - Conflits de représentation.

II.1.1.1.3- L'extraction des données

L'hétérogénéité physique des sources est traitée par l'ajout au-dessus de chaque source d'un extracteur -en anglais "wrapper"- qui extrait les données désirées et les formate dans un modèle commun. Ce modèle commun est

généralement le modèle relationnel, principalement du fait que les données extraites représentent un gigantesque volume d'informations, et donc sont dans un premier temps stockées à l'aide d'un SGBD relationnel adapté, tel ORACLE 7 ou SYBASE.

L'idéal est de ne recharger l'entrepôt qu'avec les données modifiées ou ajoutées depuis la dernière extraction. Certains outils fournissent ce type de fonctionnalité, et s'appuie par exemple sur un mécanisme de marquage des données (date de la dernière modification associée à la donnée) [**EVO99**].

II.1.1.1.4- Le nettoyage des données

Le "nettoyage" (appelé aussi "épuration", ou "analyse de la qualité des données") des données ont pour but de résoudre le problème de la consistance des données. Ces inconsistances peuvent être locales à un enregistrement (ex : une erreur de frappe), locales à une source (ex : une même personne a deux adresses différentes), ou peuvent survenir lors de la mise en commun de deux sources (ex: une personne a une adresse différente dans chaque source). Une centaine de type d'inconsistances ont été répertoriées. Elles peuvent être dues :

1- à la présence de données fausses dès leur saisie,

2- à la persistance de données obsolètes,

3- à la confrontation de données exactes, sémantiquement identiques, mais syntaxiquement différentes.

De nombreux outils comme *ACTAWorks, EDD Datacleanser* et *GENIO,* sont disponibles sur le marché pour nettoyer les données. La plupart des outils de nettoyage traitent en profondeur le problème des adresses et des noms de clients. C'est en effet un des problèmes pratiques les plus cruciaux des entrepôts de données, cette donnée étant d'une part de la plus haute importance, et d'autre part à la fois subjective, sans format fixe et volatile.

II.1.1.1.5- L'intégration des données

Deux principales approches permettent un accès unifié à des sources de données hétérogènes : une approche virtuelle (souvent appelée approche par médiateur) et une approche matérialisée (approche par entrepôt).

- ***Les approches virtuelles:*** Sont basées sur une hiérarchie de médiateurs, correspondant à des vues virtuelles, au-dessus des extracteurs. Les données ne sont stockées que dans leur source d'origine.
- ***L'approche matérialisée*** : les données sont effectivement extraites, nettoyées, intégrées et stockées dans un entrepôt. Les requêtes sont posées directement sur les données de l'entrepôt, un des problèmes majeurs à résoudre dans cette approche est celui de la répercussion dans l'entrepôt des mises à jour effectuées sur les sources.

L'approche virtuelle a été traditionnellement utilisée pour des systèmes répartis et hétérogènes.

II.1.1.2 La *construction* Consiste à extraire les données pertinentes pour la prise de décision, puis à recopier dans l'entrepôt de donnée [**TEST00**].

II.1.1.3 *La réorganisation* Permet de restructurer les données dans des magasins de données la réorganisation des données vise à supporter efficacement les processus d'interrogation et d'analyse tels que les applications OLAP [**TEST00**].

II.1.1.4 *L'interrogation* Consiste à manipuler les données multidimensionnelles afin d'analyser les tendances passées pour prendre des décisions. Les données sont représentées sous une forme visant à faciliter leur compréhension et leur manipulation pour les décideurs non informaticiens (tableaux à *n* dimensions, graphiques,…).

Le Tableau 2 résume le rôle de chaque module constituant le système décisionnel. A chacun de ces modules correspond des problématiques distinctes et spécifiques.

		Entrées	Sorties	Problématiques
INTEGRATION		$S_1,\ldots, S_n$	SG	Hétérogénéité (systèmes, modèles, formats et sémantiques des données), Distribution
CONSTRUCTION		SG	ED	Extraction des données, Modélisation de l'entrepôt, Maintenance de l'entrepôt, Historisation des données.
REORGANISATION		ED	$MD_1,\ldots,MD_p$	Extraction des données, Modélisation multidimensionnelle.
INTERROGATION	*expert*	ED	résultats	Manipulations temporelles, Manipulations multidimensionnelles, Interaction et présentation des résultats (graphes, statistiques,…)
	non expert	MD_i		

Tableau 2: les modules constituent le système décisionnel et leur rôle

II.2. Les différentes architectures d'un data warehouse:

Pour implémenter un Data Warehouse, trois types d'architectures sont possibles [**CNAM98**]:

- L'architecture réelle,
- L'architecture virtuelle,
- L'architecture remote.

II.2.1. L'architecture réelle

Elle est généralement retenue pour les systèmes décisionnels.

Le stockage des données est réalisé dans un SGBD séparé du système de production. Le SGBD est alimenté par des extractions périodiques.

Avant le chargement, les données subissent d'importants processus d'intégration, de nettoyage, de transformation.

L'avantage est de disposer de données préparées pour les besoins de la décision et répondant aux objectifs du Data Warehouse.

Les inconvénients sont le coût de stockage supplémentaire et le manque d'accès en temps réel.

II.2.2. L'architecture Virtuelle

Cette architecture n'est pratiquement pas utilisée pour le Data Warehouse.
Les données résident dans le système de production. Elles sont rendues visibles par des produits middleware ou par des passerelles.
Il en résulte deux avantages : pas de coût de stockage supplémentaire et l'accès se fait en temps réel. L'inconvénient est que les données ne sont pas préparées.

II.2.3. L'architecture Remote

C'est une combinaison de l'architecture réelle et de l'architecture virtuelle. Elle est rarement utilisée. L'objectif est d'implémenter physiquement les niveaux agrégés afin d'en faciliter l'accès et de garder le niveau de détail dans le système de production en y donnant l'accès par le biais de middleware ou de passerelle.

PARTIE III: LES CONCEPTS DE BASE DE LA MODELISATION MULTIDIMENSIONNELLE

III.1 LA MODELISATION MULTIDIMENSIONNELLE

III.1.1. Concepts de bases Multidimensionnels

III.1.1.1. Concept de Cube:

Le Cube de données offre une abstraction très proche de la façon dont l'analyse voit et interroge les données. Il organise les données en une ou plusieurs *Dimensions* qui déterminent une *mesure* d'intérêt. Une dimension spécifie la manière dont on regarde les données pour les analyser, alors qu'une mesure est un objet d'analyse. Chaque dimension est formée par un ensemble d'attributs et chaque attribut peut prendre différentes valeurs. Les dimensions possèdent en général des hiérarchies associées qui organisent les attributs à différents niveaux pour observer les données à différentes granularités. Une dimension peut avoir plusieurs hiérarchies associées, chacune spécifiant différentes relations d'ordre entre ses attributs **[BEL00]**.

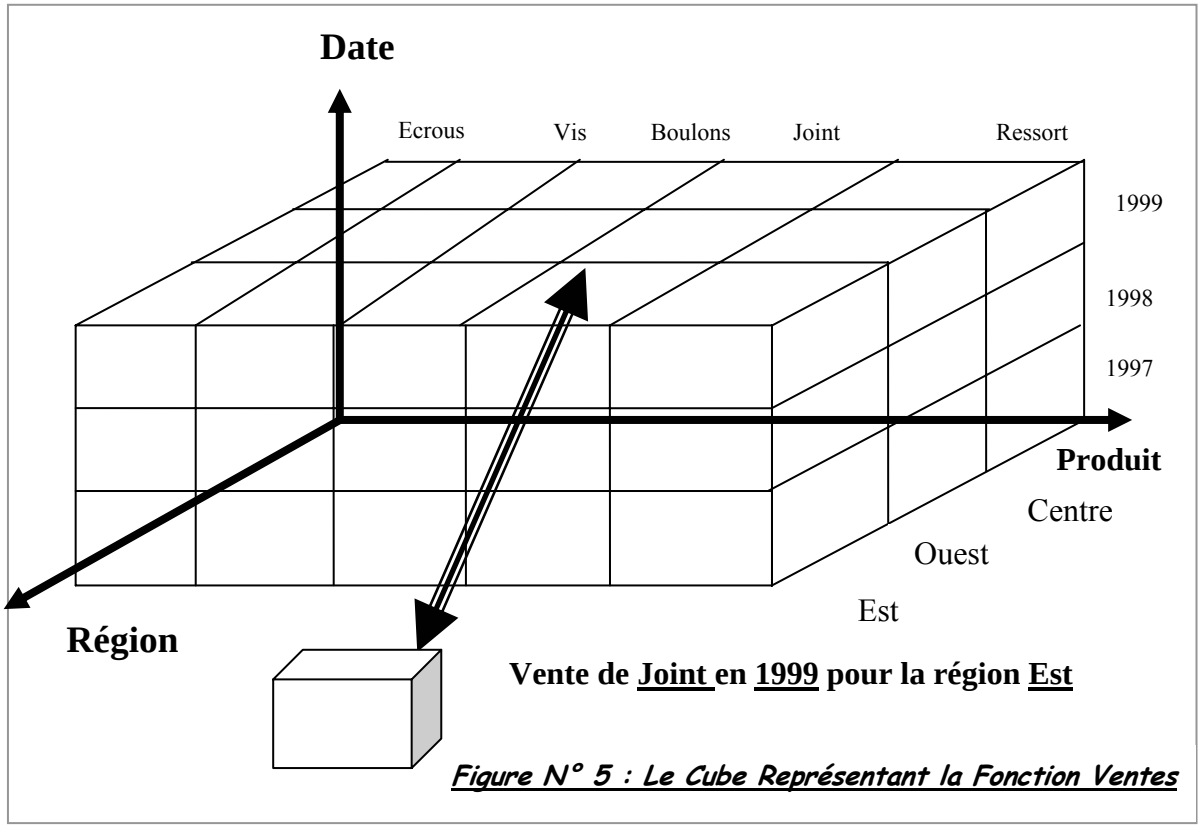


Figure N° 5 : Le Cube Représentant la Fonction Ventes

Plus formellement, la structure d'un cube de données est la suivante [**AGR96**]:

- Chacune des ***d*** dimensions porte un nom D_i. A chaque dimension D_i est associe un domaine de valeur Dom_{di}.
- Une référence de cellule est un n-uple $< v_1,v_2,...,v_d>$ appartenant à $Dom_{D1}X$ $Dom_{D2}X....Dom_{Dd}$. Le contenu d'une cellule d'un cube est soit la constante 0, soit la constante 1, soit un n-uple $< v'_1,v'_2,...,v'_k>$ appartenant a $Dom_{D'1}X$ $Dom_{D'2}X....Dom_{D'k}$

On dit encore que $v_1,v_2,...,v_d$ sont les *dimensions membres* et que $v'_1,v'_2,...,v'_k$ sont les *dimension mesures.*

Un cube ***C*** de **d** dimensions est une fonction F_C associant à chaque cellule de coordonnées $<v_1,v_2,...,v_d>$ l'un des éléments suivants:

- La constante 0 si la cellule de référence $< v_1,v_2,...,v_d>$ n'existe pas pour ***C***.

- La constante 1 si la cellule de référence $< v_1,v_2,...,v_d>$ existe mais ne contient pas de mesure;

- Un n-uple $< v'_1,v'_2,...,v'_k>$ si la cellule de coordonnées $< v_1,v_2,...,v_d>$ contient $< v'_1,v'_2,...,v'_k>$.

Pour illustrer les opérations liées à la structure multidimensionnelle (Pivot, swith, Slice, Dice), et Opérations liées à la granularités (Roll-up, et Drill-down) nous utilisons l'exemple Suivant :

Produit	Région	Vente	Période: 97	Produit	Région	Vente	Période: 98	Produit	Région	Vente	Période: 99
P1	Centre	**15**		P1	Centre	**34**		P1	Centre	**33**	
P2	Centre	**24**		P2	Centre	**25**		P2	Centre	**15**	
P3	Centre	**43**		P3	Centre	**37**		P3	Centre	**26**	
P2	Est	**54**		P1	Est	**41**		P2	Est	**21**	
P3	Est	**59**		P3	Est	**26**		P3	Est	**35**	
P1	Sud	**23**		P1	Sud	**43**		P1	Sud	**12**	
P3	Sud	**34**		P3	Sud	**44**		P3	Sud	**19**	

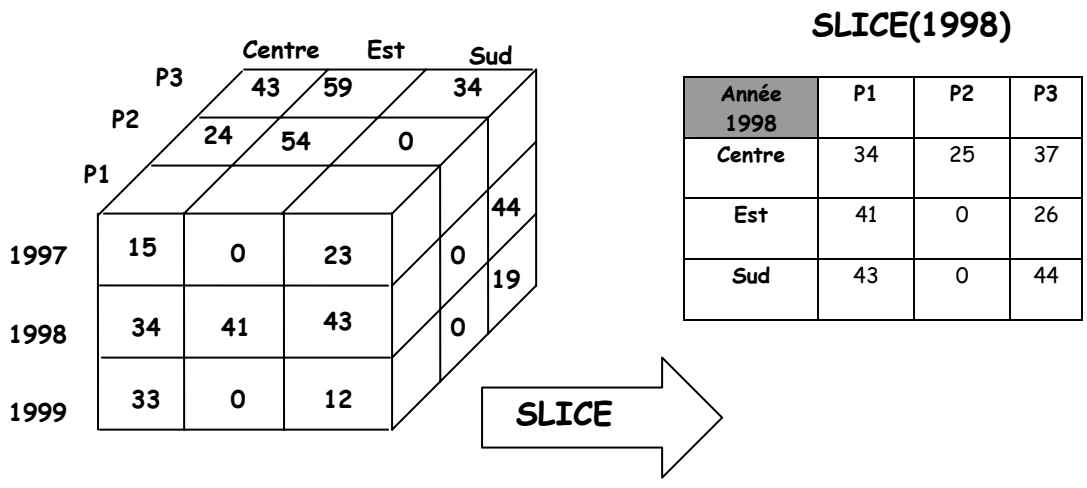
Tableau 3 les Données d'une activité de Vente

III.1.1.1.1. Opérations liées à la structure :

Les opérations agissant sur cette structure multidimensionnelle de l'information sont motivées par l'aspect interactif de l'analyse en ligne de données, et le souci d'offrir des possibilités d'animation de la représentation. De plus, elles illustrent l'importante des liens entre la manipulation des données et la représentation du cube à l'écran.

Ces opérations sont regroupées sous le nom de restructuration. Tout cube obtenu par une opération de restructuration d'un cube initial contient tout ce qu'il faut pour régénérer le cube initial par restructuration réciproque. Ces opérations sont **[BEL00]** :

- ***Slicing :*** Cette opération consiste en général à filtrer une dimension selon une valeur ou une plage de valeurs afin de se concentrer sur une partie de données, exemple : **Slice** (1998) : on ne retient que la partie du cube qui correspond à cette période.



Année 1998	P1	P2	P3
Centre	34	25	37
Est	41	0	26
Sud	43	0	44

Figure N° 6 : L'opération SLICE sur un Cube

- ***Dicing :*** Cette opération consiste à l'extraction d'un sous cube.

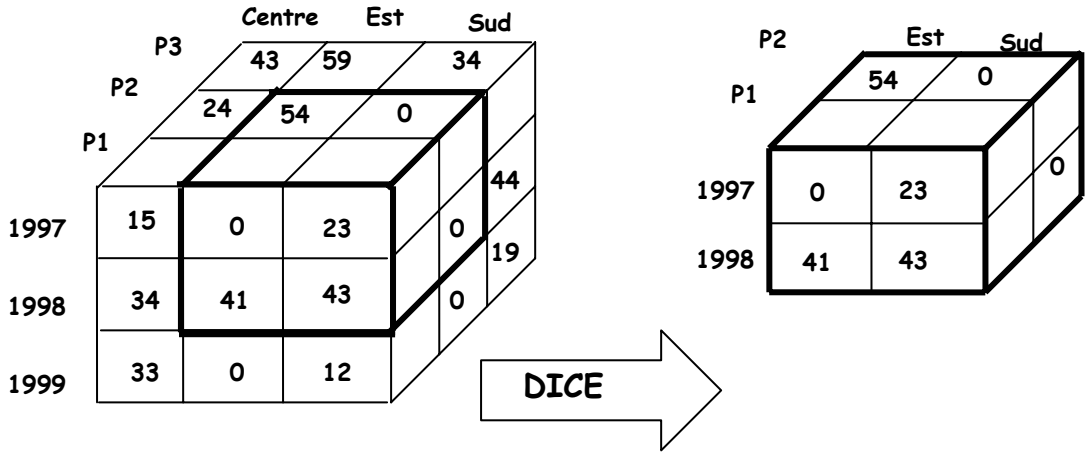


Figure N° 7 : L'opération DICE sur un Cube

- ***Pivot :*** Cette opération consiste à faire effectuer à un cube une rotation autour d'un des trois axes passant par le centre de deux faces opposées, de manière à présenter un ensemble de faces différent.

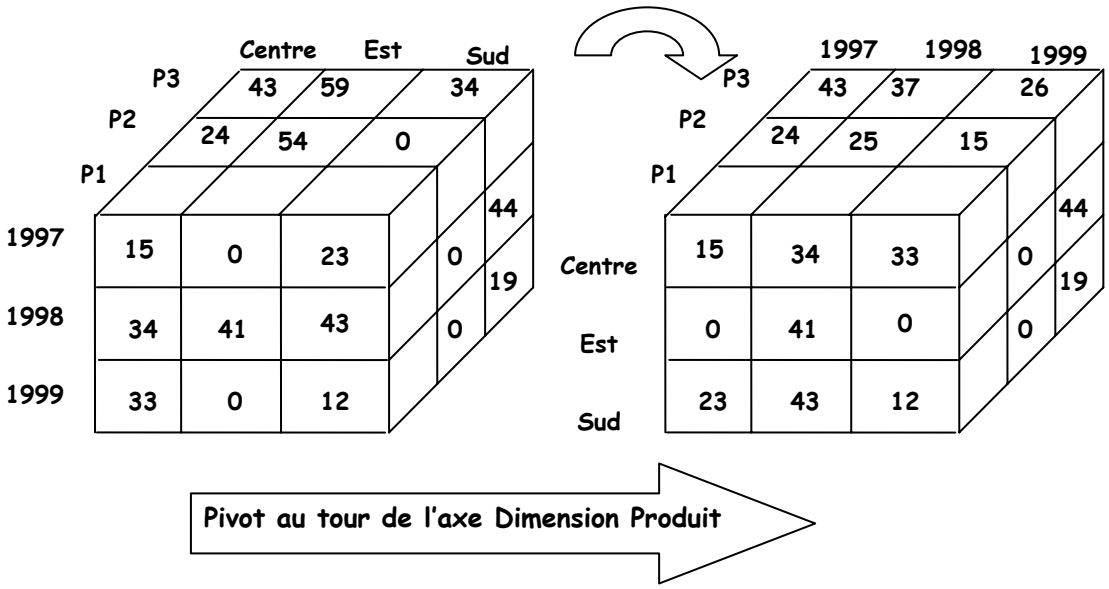


Figure N° 8 : L'opération PIVOT sur un Cube

- ***Switch :*** Cette opération consister à interchanger la position des membres d'une dimension.

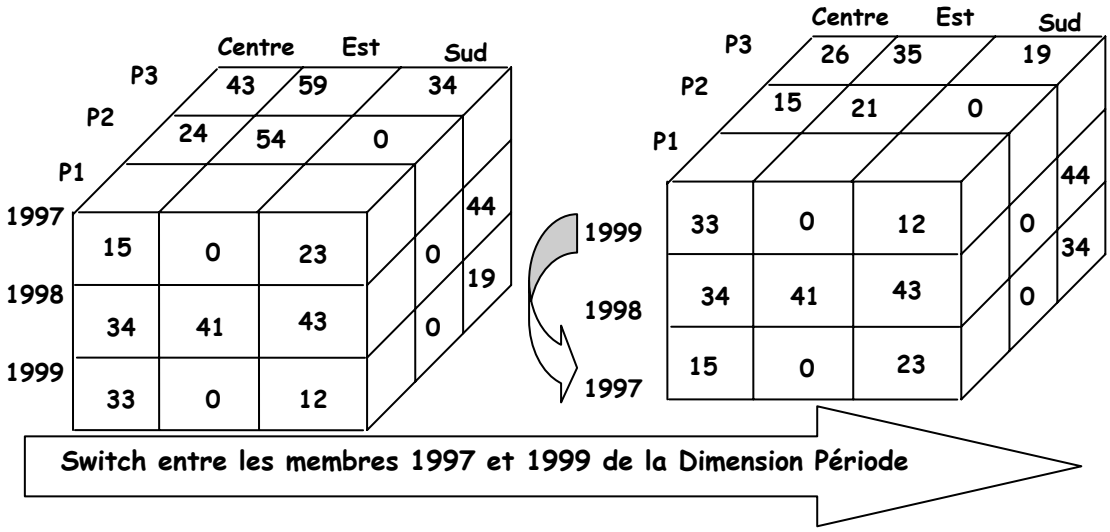


Figure N° 9 : L'opération SWITCH sur un Cube

III.1.1.1. Opérations liées à la granularités:

Le deuxième aspect de la vision de l'analyse est de hiérarchiser l'information en différents niveaux de détail appelés niveaux de *granularité.* Les opérations permettant la hiérarchisation sont [**BEL00**] :

Roll-up et drill-down. Ces deux opérations autorisent l'analyse de données à différents niveaux d'agrégations en utilisant des hiérarchies associées à chaque dimension.

- ***Roll-up:*** Cette opération effectue l'agrégation des mesures en allant d'un niveau particulier de la hiérarchie vers un niveau général.

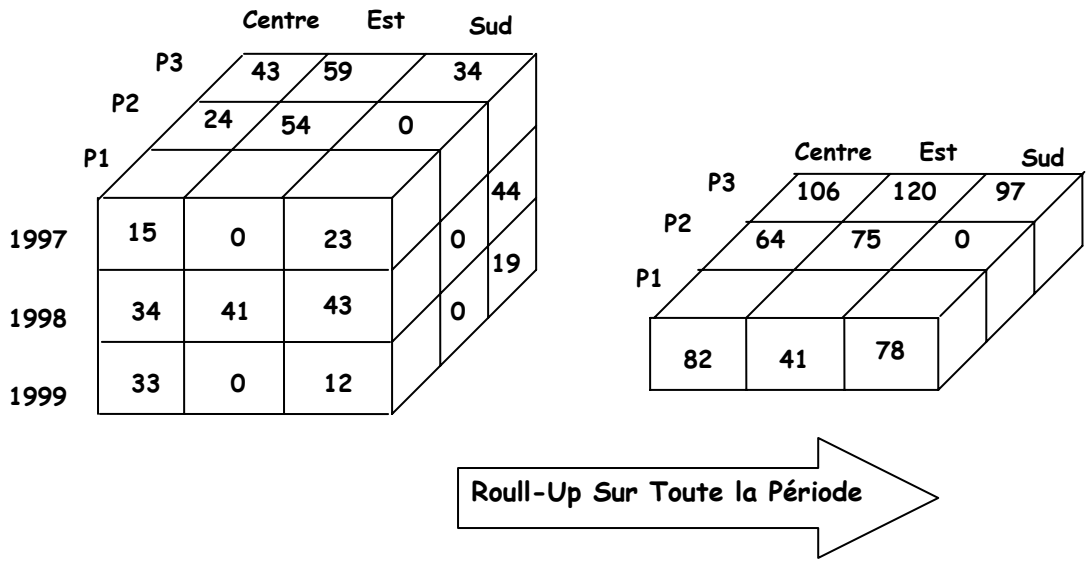


Figure N° 10 : L'opération ROULL-UP sur un Cube

- ***Drill-down:*** Elle consiste à représenter les données d'un cube à un niveau inférieur, et donc sous une forme détaillée. Elle peut être vue comme l'opération réciproque du *Roll-up.*

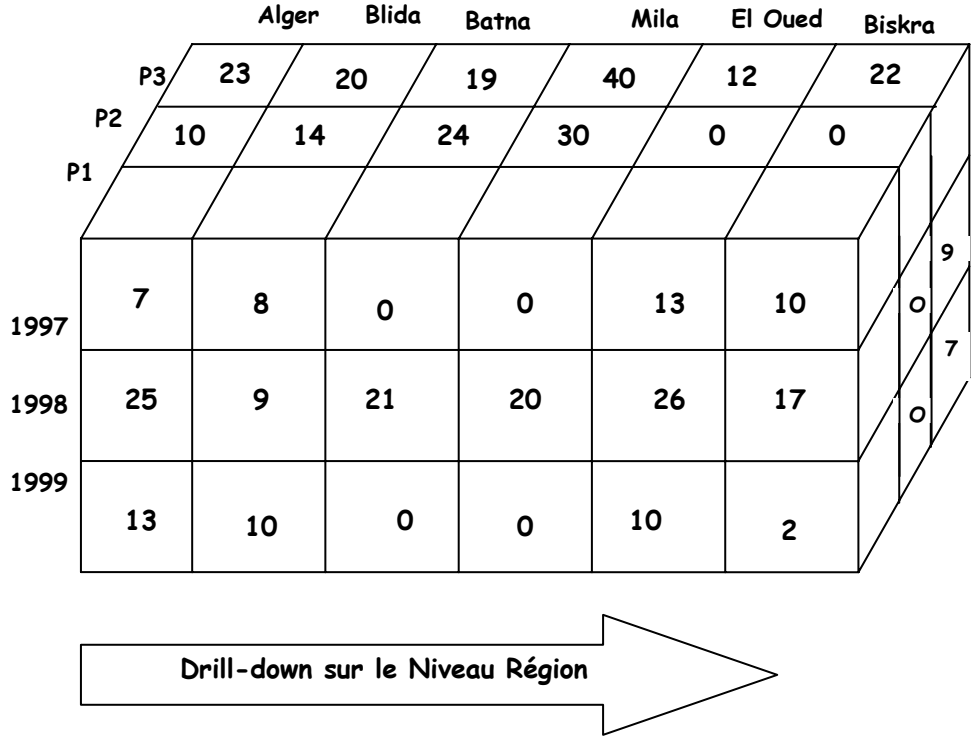


Figure N° 11 : L'opération DRILL-DOWN sur un Cube

III.1.1.2. Concept de Faits:

Le **fait** modélise le sujet de l'analyse. Un fait est formé de **mesures** correspondant aux informations de l'activité analysée [**TEST00**].

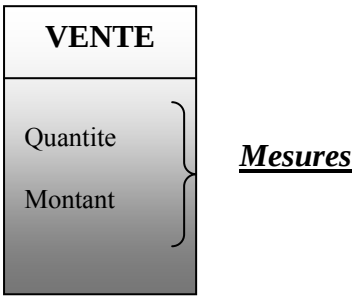


Figure N° 12 : Schéma d'un Fait

Les mesures d'un fait sont numériques et généralement valorisées de manière continue [**KIM96**]. Les mesures sont numériques pour permettre de résumer un grand nombre d'enregistrements en quelques enregistrements (on peut les additionner, les dénombrer ou bien calculer le minimum, le maximum ou la moyenne). Les mesures sont valorisées de façon continue car il est important de ne pas valoriser le fait avec des valeurs nulles. Elles sont aussi souvent :

- Additives : C'est-à-dire que l'opérateur d'agrégation (représenté par **SUM**) peut être appliqué pour regrouper les valeurs des mesures au cours de tous les dimensions (exemple : Quantités vendus, Chiffre d'affaire ….).
- Semi-additives : Additives selon quelques dimensions (exemple : Nombre de transactions d'un client).
- Non-additives : C'est-à-dire que l'opérateur d'agrégation (représenté par **SUM**) n'est pas applicable sur aucune dimension (exemple : un ratio de gestion).

III.1.1.2. Concept de Dimension:

Le sujet analysé, c'est à dire le fait, est analysé suivant différentes perspectives. Ces perspectives correspondent à une catégorie utilisée pour caractériser les mesures d'activité analysées [**MAR98**], on parle de dimensions.

Une **dimension** modélise une perspective de l'analyse. Une dimension se compose de **paramètres** correspondant aux informations faisant varier les mesures de l'activité.

Les dimensions servent à enregistrer les valeurs pour lesquelles sont analysées les mesures de l'activité. Une dimension est généralement formée de paramètres (ou attributs) textuels et discrets. Les paramètres textuels sont utilisés pour restreindre la portée des requêtes afin de limiter la taille des réponses. Les paramètres sont discrets, c'est à dire que les valeurs possibles sont bien déterminées et sont des descripteurs constants.

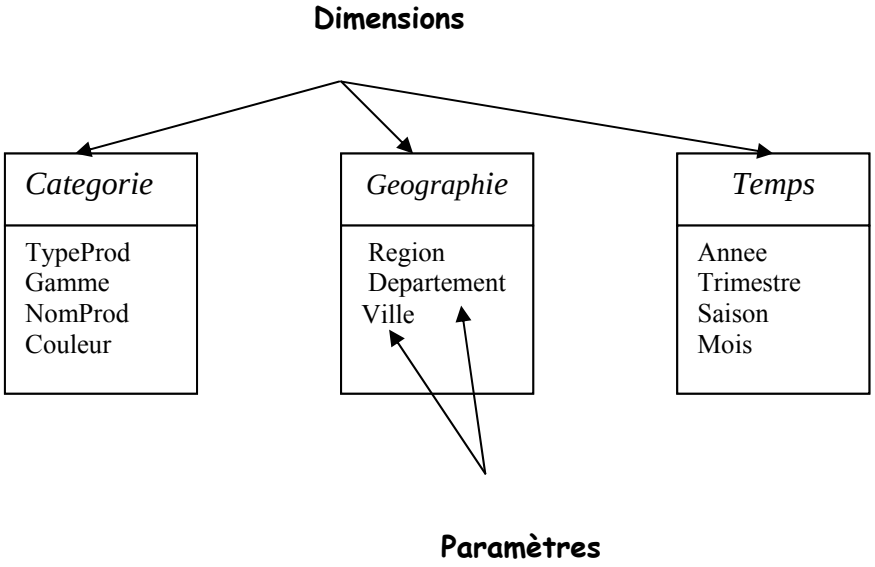


Figure N°13 : Exemple des dimensions d'un Fait

III.1.1.4. Concept d'hiérarchie

Une hiérarchie organise les paramètres d'une dimension selon une relation *"est_plus_fin"* conformément à leur niveau de détail [**TEST00**]. On distingue deux types d'hiérarchies Simple et Multiple:

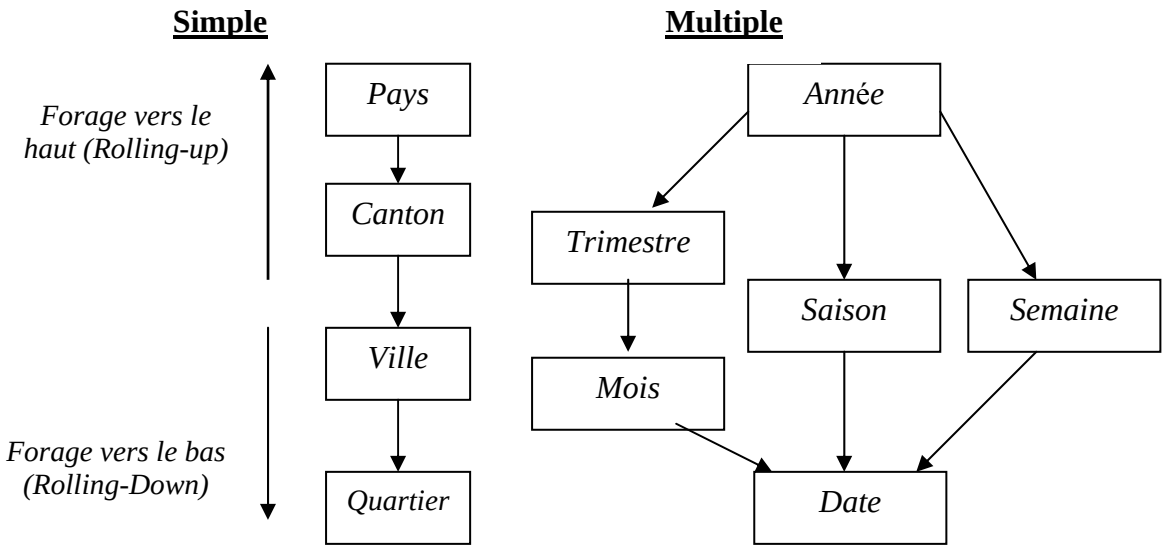


Figure N°14 : Différents Types d'hiérarchies

***III.1.2. Schémas Multidimensionnels*:**

D'une manière très générale, le modèle dimensionnel possède une table de faits centrale et des tables de dimensions situées autour. Chaque enregistrement de la table de faits stocke les clés des tables de dimensions et les mesures faites à un instant précis. La taille de la table de faits est de plusieurs millions d'enregistrements, et peut nécessiter plusieurs giga, voire tera octets d'occupation sur disque [**PROB01**].

III.1.2.1. Schéma en Etoile :

A partir des faits et des dimensions, il est possible d'établir une structure de données simple qui correspond au besoin de la modélisation multidimensionnelle. Cette structure est constituée du fait central et des dimensions. Ce modèle représente visuellement une étoile, on parle de *modèle en étoile* [**KIM96**], [**TEST00**].

- La table des faits contient des mesures (par exemple le prix des produits vendus, la quantité de produits vendus) qui peut être agrégée de diverses façons.
- Les tables de dimension fournissent la base pour l'agrégation des mesures de la table de fait.
- La table de fait est liée avec toutes les tables de dimension par des relations "Un-à-Plusieurs".
- La clef primaire de la table de fait est la concaténation des clefs primaires de toutes les tables de dimension.

L'avantage d'employer les schémas en étoile pour représenter les données est la réduction du nombre de tables, et le nombre de relations entre ces tables et donc le nombre de jointures exigé dans les requêtes des utilisateurs [**Kim96**].

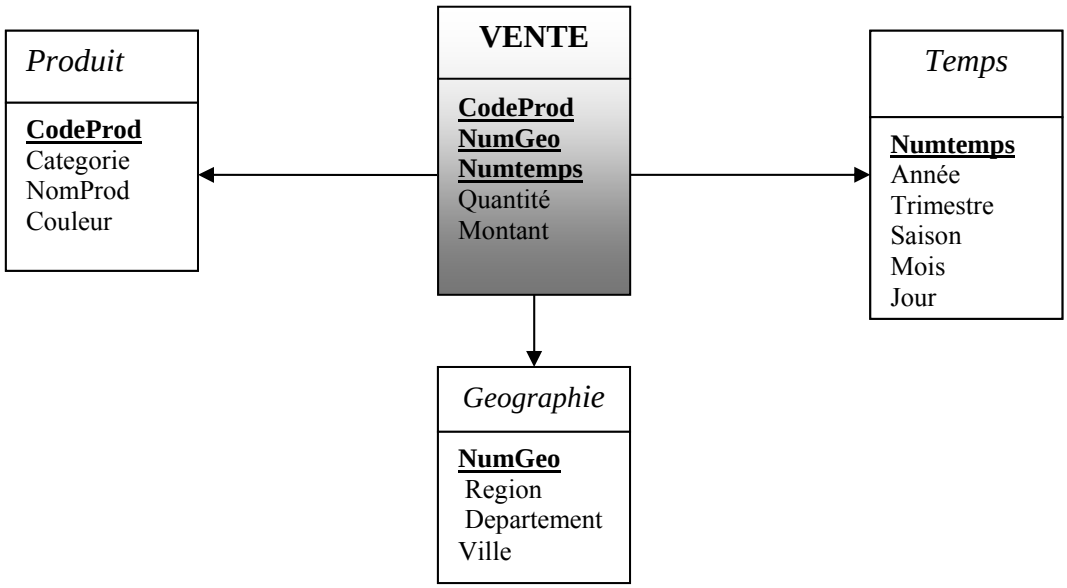


Figure N°15 : Schéma en Etoile

III.*1.2.2. Schéma en Flocon*

Dans une étoile, les dimensions sont *dénormalisées* : les hiérarchies d'agrégation sont implicites. Un *flocon* de neige est une étoile où les hiérarchies sont respectées et explicites.

Un flocon peut être formé depuis une étoile en normalisant les dimensions. Le flocon diffère de l'étoile de par sa constitution : un flocon reste relativement normalisé alors qu'une étoile est dénormalisée [**PROB01**].

Une modélisation en flocon consiste à décomposer les dimensions du modèle en étoile en sous hiérarchies. La modélisation en flocon est donc une émanation de la modélisation en étoile, le fait est conservé et les dimensions sont éclatées conformément à sa hiérarchie des paramètres.

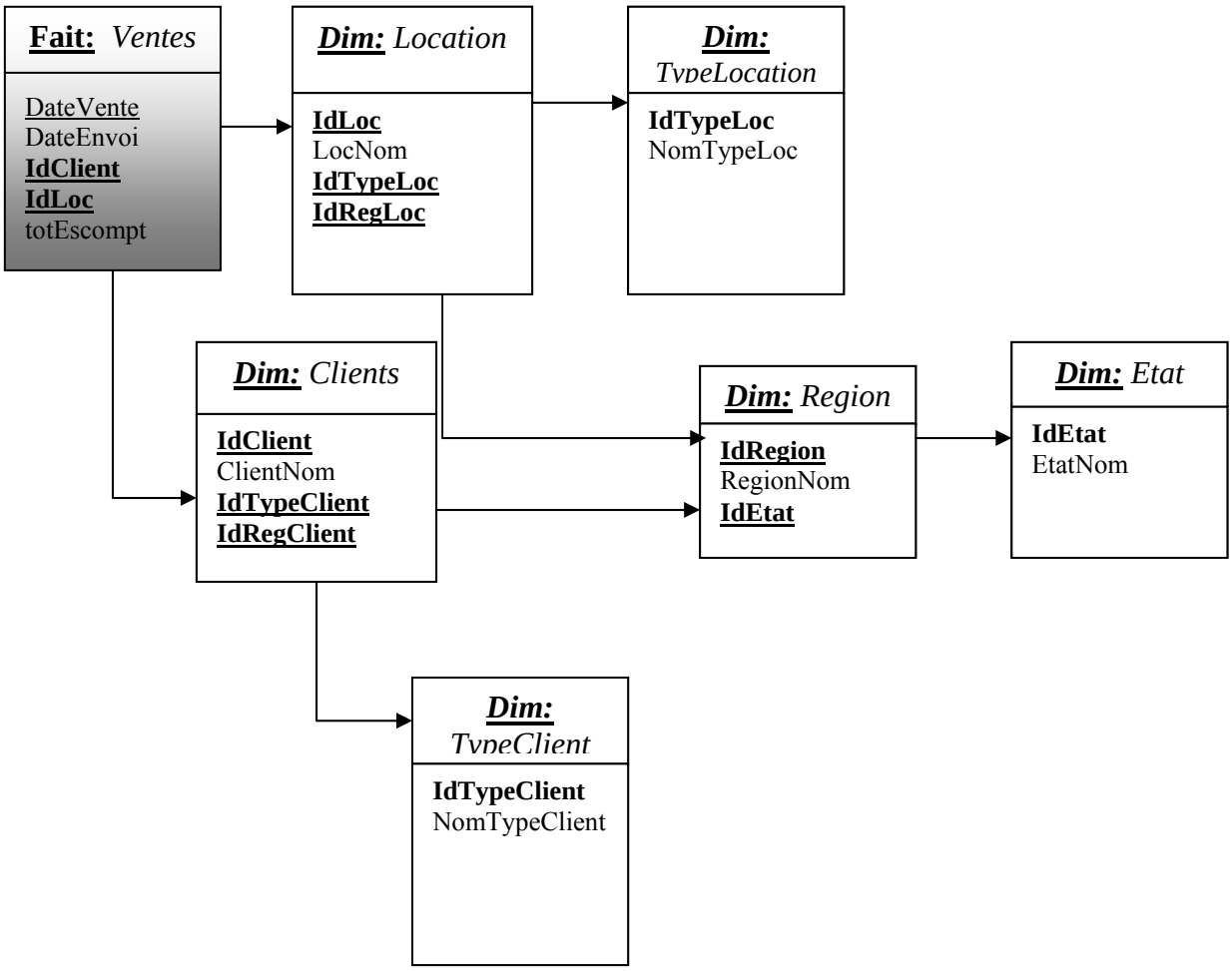


Figure N°16 : Schéma en Flocon de Neige

Le modèle en Flocon est plus propre (3eme Forme Normale), la hiérarchie multiple est simplifiée et il offre un gain d'espace de stockage (même s'il est négligeable par rapport à la table de Fait).

III.*1.2.3. Schéma en Constellation :*

Le schéma en constellation consiste à placer plusieurs schémas en étoile avec des tables de faits reliées hiérarchiquement. Les liens entre les différentes tables de faits permettent de visionner les différents niveaux de détail

[**PROB01**]. Donc ce schéma consiste à fusionner plusieurs modèles en étoile qui utilisent des dimensions Communes [**TEST00**].

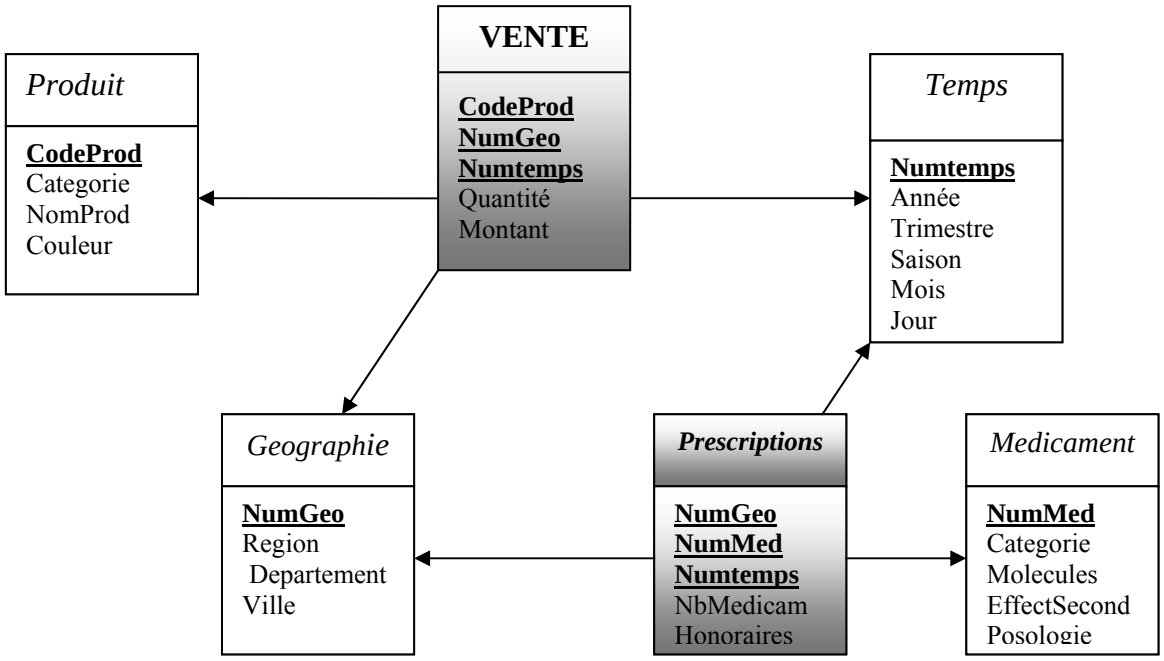


Figure N°17 : Schéma en Constellation

III.1.2.4. Schéma en Grappe:

Ce schéma est apparu car il n'existe pas de schéma en étoile ou de schéma en flocon parfait. Le schéma en grappe est une dérivation de ces deux schémas pour en former un troisième. *Kimball* déclare qu'un schéma en flocon n'est pas optimal, car il est trop complexe. Toujours pour les mêmes raisons de simplifications des tables, afin de pouvoir trouver facilement les informations dans l'entrepôt, le schéma en grappe apparaît alors comme un compromis entre le schéma en étoile et le schéma en flocon.

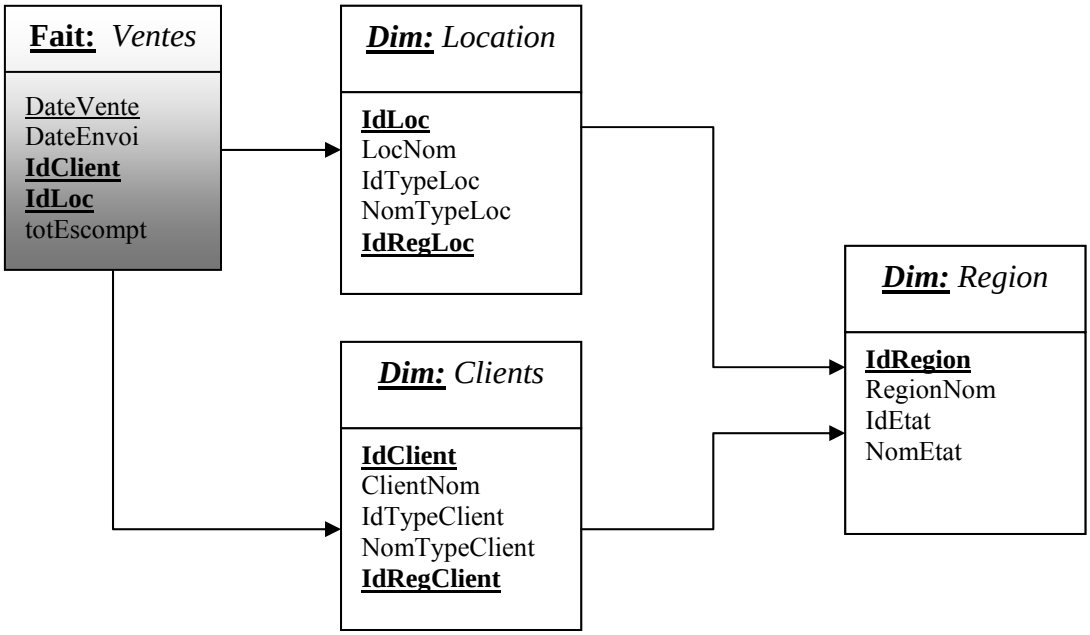


Figure N°18 : Schéma en Grappe

Si l'on prend la figure N°18, on remarque que les tables *Location* et *Clients* dépendent toutes deux des tables *Région* et *Etat*. De plus, *Client* dépend de *TypeClient*, et Location dépend de *TypeLocation*. La table des faits restant la table *Ventes*. Ainsi, par une agrégation des entités, le schéma en grappe permet de regrouper les tables *TypeLocation* dans la table *Location*, la table *Etat* dans la table *Region*, et la table *TypeClient* dans la table *Client* pour ne former que quatre tables. La figure n°18 donne un exemple d'un schéma en grappe [**PROB01**].

III.1.2.5. Complexité Vs Redondance

Tous ces modèles sont des modèles dimensionnels, que l'on peut ou non implémenter lors de la création d'un entrepôt de données. Cependant, il faut savoir que ces modèles augmentent soit la complexité des tables, soit la redondance des données. Chaque cas est unique, et on doit bien réfléchir au schéma que l'on va adopter. La figure N°19 illustre le classement des modèles présentés en fonction de la complexité ou de la redondance des données.

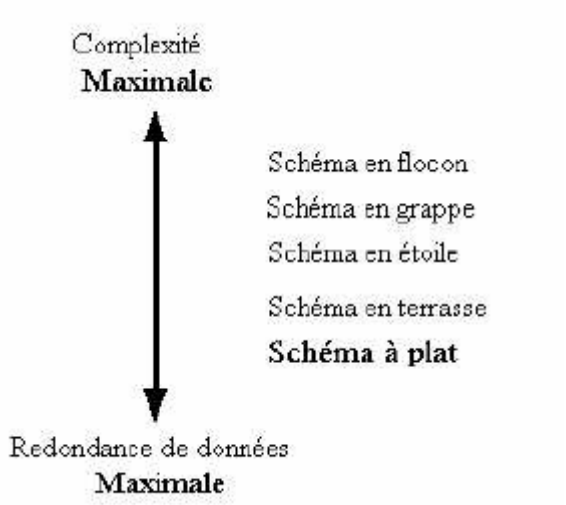


Figure N°19 : Complexité Vs Redondance

***III**.1.2.6 Agrégations :*

L'utilisation de récapitulation préenregistrées (agrégats) est l'outils le plus efficace dont dispose le concepteur d'entrepôt de données pour améliorer les performances.

Un agrégat est un enregistrement de table de faits représentant la récapitulation d'enregistrements élémentaires de cette table. Un enregistrement d'une table de faits agrégés est associé à un ou plusieurs enregistrements de tables de dimensions d'agrégats **[KIM97]**.

***III*.1.3 Les implémentations des modèles multidimensionnels**

Selon la façon dont le cube de données est stocké, il existe deux approches fondamentales pour construire des systèmes basés sur un modèle multidimensionnel et une troisième Hybride. L'approche MOLAP (Multidimensionnel OLAP) implémente le cube de données dans un tableau multidimensionnel (multidimensionnelle array). Par contre, l'approche ROLAP (Relationnel OLAP) utilise un SGBD relationnel pour gérer et stocker le cube de données [**BEL03**].

III.1.3.1 Les systèmes MOLAP

Les systèmes de type MOLAP stockent les données dans un SGBD multidimensionnel sous la forme d'un tableau multidimensionnel (multidimensionnelle array). Chaque dimension de ce tableau est associée à une dimension du cube. Seules les valeurs de données correspondant aux données de chaque cellule sont stockées. Ces systèmes demandent un pré calcul de toutes les agrégations possibles. En conséquence, ils sont plus performants que les systèmes traditionnels, mais difficiles à mettre à jour et à gérer [**BEL03**].
Les systèmes MOLAP apparaissent comme une solution acceptable pour le stockage et l'analyse d'un entrepôt lorsque la quantité estimée des données d'un entrepôt ne dépasse pas quelques gigaoctets et lorsque Le modèle multidimensionnel évolue peu. Mais, lorsque les données sont éparses, ces systèmes sont consommateurs d'espace [***CHA97***] et des techniques de compression doivent être utilisées.

III.1.3.2 Les systèmes ROLAP

Les systèmes de type ROLAP utilisent un SGBD relationnel pour stocker les données de l'entrepôt. Ils représentent une interface multidimensionnelle pour le SGBD relationnel. Le moteur OLAP est un élément supplémentaire qui fournit une vision multidimensionnelle de l'entrepôt, des calculs de données dérivées et des agrégations à différents niveaux. Il est aussi responsable de la génération des requêtes SQL mieux adaptées au schéma relationnel. Les mesures (par exemple les quantités vendues) sont stockées dans une table qu'on appelle la table des faits. Pour chaque dimension du modèle multidimensionnel, il existe une table qu'on appelle la table de dimension (comme Produit, Temps, Client) avec tous les niveaux d'agrégation et les propriétés de chaque niveau [**BEL03**].

Ces systèmes peuvent stocker de grands volumes de données, mais ils peuvent présenter un temps de réponse élevé. Les principaux avantages de ces systèmes sont :

1- Une facilité d'intégration dans les SGBDs relationnels existants.

2- Une bonne efficacité pour stocker les données multidimensionnelles.

Les exemples de produits de cette famille sont DSS Agent de MicroStrategy et MetaCube d'Informix.

III.1.3.3 Les systèmes HOLAP

Les données des cubes sont stockées dans une structure relationnelle et les agrégats dans une structure multidimensionnelle.

L'utilisateur formule une requête sur la base de données relationnelle et le résultat s'affiche sous une forme multidimensionnelle. La convivialité et ainsi respecté tout en permettant le stockage de grand volume de données.

PARTIE IV: LES APPROCHES DE CONCEPTION MULTIDIMENSIONNELLE

IV.1- Conception d'un schéma conceptuel selon l'approche KIMBALL [KIM97]

Considérée comme étant la référence dans le monde de la conception des entrepôts de données la méthode de conception d'une base de données multidimensionnelles suivant la démarche KIMBALL est abordée d'une manière systématique, en envisageant quatre étapes dans un ordre bien défini.

- *Première étape*
  - *Choisir le processus d'activité à modéliser.* Un processus d'activité est un processus opérationnel important pour l'organisation, étaye par une application existante à partir de laquelle des données peuvent être collectées au profit de l'entrepôt de données. Exemples de processus d'activité : Facturation, Ventes….
- *Deuxième étape*
  - *Choisir le Grain du processus d'activité.* Le Grain est le niveau de détail fondamental, atomique, des données figurant dans la table des faits pour ce processus. Des grains typiques sont des transactions individuelles, des récapitulations individuelles quotidiennes.
- *Troisième étape*
  - *Choisir les dimensions applicables a chaque enregistrement de la table de faits.* Des dimensions typiques sont le temps, le produit, le magasin, le client. Le choix d'une dimension s'accompagne de la définition de tous les attributs textuels qui garniront la table de dimension
- *Quatrième étape*
  - *Choisir les faits mesures que contiendra chaque enregistrement de la table de faits.* Des faits mesures typiques sont des quantités numériques additives.

Sachant qu'aucune méthode automatique ou semi-automatique n'est envisagée, mais KIMBALL se base sur une approche intuitive.

La construction d'un entrepôt de donnée multidimensionnel selon KIMBALL revient à faire correspondre les besoins de la communauté utilisateurs avec la réalité des informations disponibles. Les neuf décisions majeures qui jalonnent la conception de la base de données sont toutes subordonnées aux besoins des utilisateurs et aux réalités que représentent les données.

IV.2- Les neuf décisions

Les neuf décisions à prendre pour la conception complète d'un entrepôt de données dimensionnel portent sur les points suivants:

1. Les processus, et partant, l'identité des tables de faits.
2. Le grain de chaque table de fait.
3. Les dimensions de chaque table de faits.
4. Les faits, y compris les faits précalculés.
5. Les attributs des dimensions, avec des descriptions complètes et la terminologie adéquate.
6. Comment suivre les dimensions à évolution lente.
7. Les agrégats, les dimensions hétérogènes, les minidimensions, les modes de requêtes et autres décisions sur le stockage physique.
8. L'étendue historique de la base de données.
9. L'urgence avec laquelle les données doivent être extraites et chargées dans l'entrepôt de données.

Cette méthodologie est une méthodologie descendante, qui commence par identifier les processus majeurs de l'entreprise dans la quelle les informations sont collectées. Car les concepteurs d'entrepôt de donnée d'après KIMBALL doivent commencer avec des sources des données existantes et réellement

utilisées, pour éviter de perdre du temps en rêvant à des sources d'information qui n'existent pas.

Une fois les processus identifies, une ou plusieurs tables de faits sont construites à partir de chacun des processus choisis. Avant de pouvoir concevoir en détail une table de faits, une décision doit être prise sur la signification précise d'un enregistrement de plus bas niveau dans la table de faits. Cette signification est le grain de la table de faits. Lorsque le grain de la table de faits est connu, les dimensions et leurs grains respectifs peuvent être identifiés. Il y aura des dimensions supplémentaires qui ne seront pas strictement requises pour décider du grain de la table de faits.

Le choix des dimensions est le point clé de la conception, une fois les dimensions choisies, il faut définir toutes les mesures de la table de faits, on peut ensuite définir complètement le contenu des enregistrements de dimension.

A ce stade la conception de la structure logique principale est terminée, et nous pouvons porter notre attention sur les aspects les plus généraux de la structure physique. Ces aspects incluent la manière à suivre les dimensions à évolution lente, l'inclusion d'agrégats, les dimensions hétérogènes, les minidimensions et le mode de requêtes.

IV.2. Conception d'un schéma conceptuel selon l'approche M. Kortink et D.L.Moody

IV.2.1. Introduction:

D'après [**Kor00**], on peut arriver à un modèle dimensionnel en partant du modèle d'entreprise s'il existe. Ils donnent une méthode de conception basée sur un modèle existant. Ils proposent de définir le modèle de l'entrepôt de données à partir du modèle d'entreprise, dont la mise en place est en général très onéreuse.

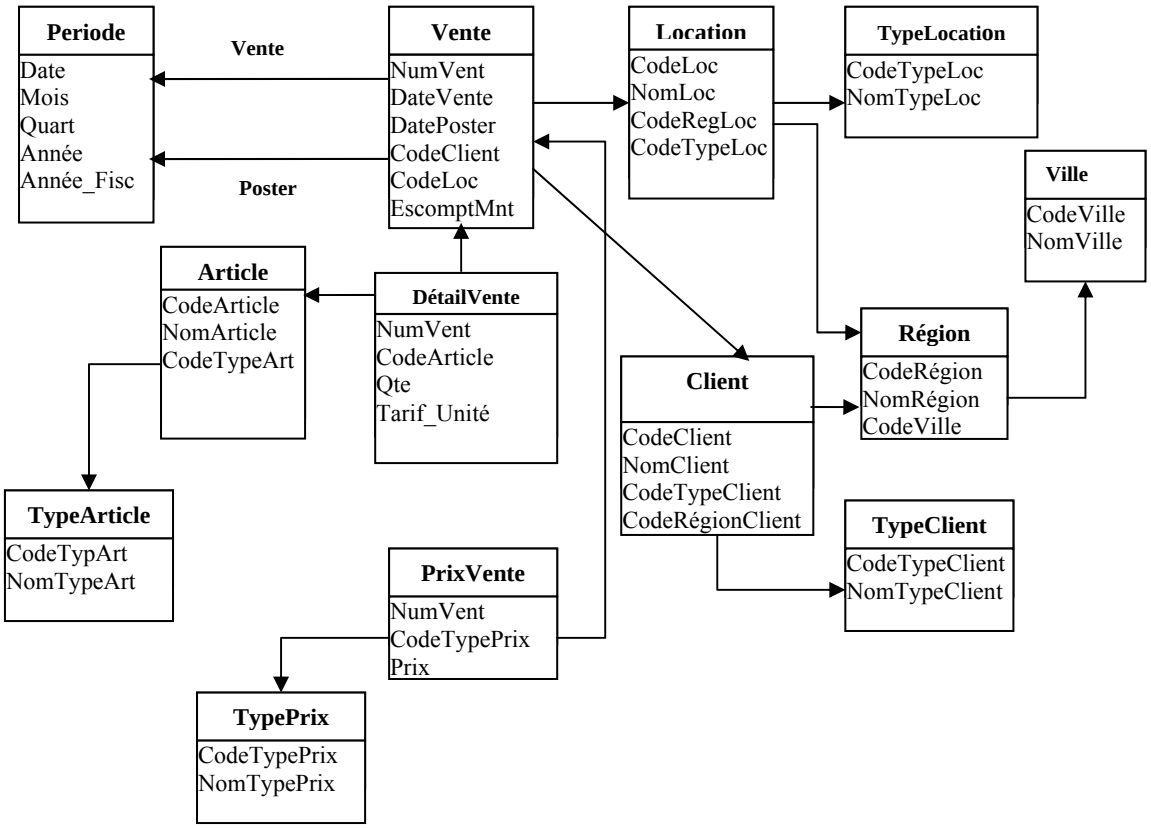


Figure N°20 : Exemple de Modèle de donnée

IV.2.2. Phase 01: Classification des entités

La première étape pour créer un modèle multidimensionnel à partir du modèle l'entreprise est de classer ses entités. Elles se décomposent en trois catégories

IV.2.2.1 Entité de transaction :

Elles enregistrent les détails d'événements particuliers comme les salaires, les réservations d'hôtels. Ce sont ces événements, entre autres, que les décideurs veulent analyser. Les caractéristiques des entités de transaction sont :

1°) Elles décrivent un événement qui se produit à un instant donné.

2°) Elles contiennent les mesures, comme le montant, le poids, les volumes.

Ces mesures forment la base des résultats que l'entrepôt permet d'étudier.

Les entités de transaction forment le noyau à partir duquel la table des faits d'un schéma en étoile peut être créée. Toutes les entités de transaction ne sont pas bonnes pour l'aide à la décision, et l'on doit choisir et identifier les plus intéressantes.

IV.2.2.2 Entité Composante :

Elle est directement liée à l'entité de transaction par une relation (un à plusieurs). Ces entités définissent la finesse des détails ou composants de chaque transaction. Les composants répondent aux questions **Q**ui, **Q**uoi, **O**ù, **Q**uand, **C**ombien, **P**ourquoi, d'un événement business (commercial):

- *Client* : **Q**ui a fait l'achat.
- *Produit* : ce qui a été vendu, le **Q**uoi.
- *Emplacement* : **O**ù il a été vendu.
- *Période* : **Q**uand il a été vendu.

Les entités composantes forment la base des tables de dimension dans les schémas en étoile. Chaque table de dimension correspond à une ou plusieurs entités composantes.

IV.2.2.3 Entité de classification :

Ce sont des entités qui sont apparentées à des entités composantes par une chaîne de relations (un à plusieurs). Ces entités représentent la hiérarchie entre les entités dans le schéma en étoile.

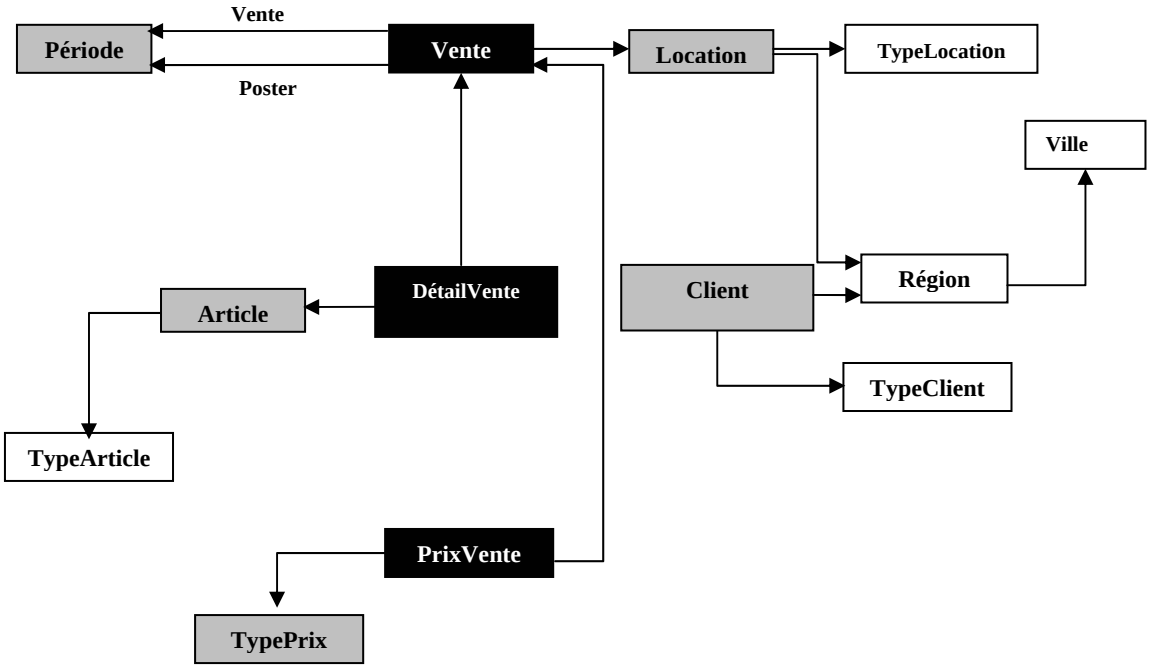


Figure N°21 : Classification des entités

Les entités en noir représentent les entités de transaction.

Les entités en gris représentent les entités composantes.

Les entités en blanc représentent les entités de classification.

IV.2.2.4 Ambiguïtés :

Parfois, des entités sont présentes dans plusieurs catégories. Les auteurs définissent la priorité des hiérarchies comme suit :

1°) Transaction : la plus haute.

2°) Classification : La Moyenne.

3°) Composante : la plus basse.

IV.2.3. Phase 02 : Identification des hiérarchies

La hiérarchie est un concept extrêmement important dans le modèle multidimensionnel, car c'est par la hiérarchie que l'on va pouvoir passer du modèle relationnel au modèle multidimensionnel.
Une hiérarchie est un modèle d'entité-association qui est identifié par des séquences d'entités reliées par des relations (un à plusieurs), toutes alignées dans le même sens. Nous parlons alors d'un schéma normalisé.

Hiérarchie maximale

Une hiérarchie est maximale si l'on ne peut l'étendre vers le haut ou vers le bas en ajoutant de nouvelles entités. Une entité est dite minimale si elle est placée au bas d'une hiérarchie maximale. Une entité est dite maximale si elle est située au plus haut de la hiérarchie. Dans l'exemple de la figure N°21, il y a deux entités minimales, DétailVente et PrixVente, alors qu'il y a six entités maximales : Periode, TypeClient, Ville, Type Location, TypeArticle et TypePrix.
Les entités minimales sont facilement identifiables car ce sont des entités qui n'ont pas de relations (un à plusieurs) ou Entités "feuille" dans la terminologie hiérarchique. A l'inverse, les entités maximales se décrivent par des entités qui n'ont pas de relations monovaluées (plusieurs à un) ou Entités "Racine" dans terminologie hiérarchique.

IV.2.4 Phase 03: Production du Modèle multidimensionnel

IV.2.4.1. Opérateur 01: Réduction de la hiérarchie

Des entités de rang supérieur peuvent être concaténées à des entités inférieures en respectant la hiérarchie. Ainsi, l'attribut *NomVille* peut être mis dans la table Région. Cela introduit une redondance et forme une dépendance transitive, ce qui viole la troisième forme normale de [**COD70**]. On introduit alors le concept de base dénormalisée. On peut aller plus loin, en concaténant les

attributs *NomRégion*, *CodeVille* et *NomVille* à l'entité *Location*. Il faut alors continuer dans le même sens pour se retrouver avec une seule table, la table *DétailVente* qui contient la concaténation de tous les attributs des autres tables.

IV.2.4.2. Opérateur 02: L'agrégation une fonction des entités de transaction

Une agrégation peut être appliquée à une entité de transaction pour former une nouvelle entité de transaction, constituée d'une synthèse de données. Cette fonction n'est possible qu'avec des attributs numériques, afin de faire ressortir des champs calculés, par exemple. La clé de cette nouvelle entité est une combinaison des attributs utilisés par l'agrégation. Attention, l'agrégation perd des informations, et l'on ne peut reconstruire les détails de la table *DétailVente* depuis la table Récap_Article, comme la montre la figure n°22 ci-dessous.

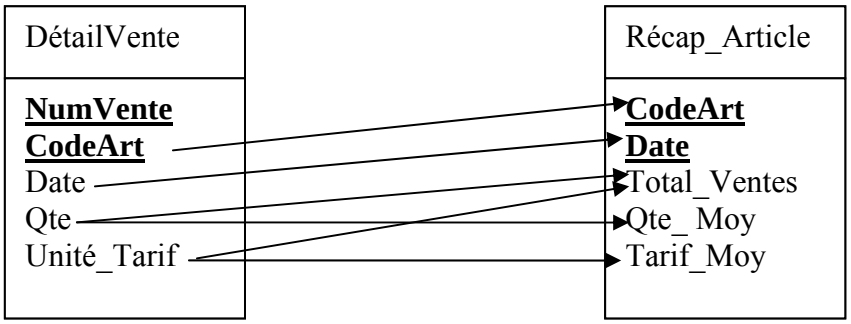


Figure N°22 : Agrégation d'entité

Il y a un large choix d'options pour la production des Modèles multidimensionnels à partir d'un modèle Entité Relation, et qui incluent :

- √ Schéma Plat (Flat)
- √ Schéma En terrasses (Terrace)
- √ Schéma d'étoile (Star)
- √ Schéma de Flocon de neige (Snowflake)
- √ Schéma en grappe (Star Cluster)

Chacune de ces différentes options représente le compromis entre la complexité et la redondance. Ici nous récapitulons comment les opérateurs précédemment définis peuvent être employés pour Produire des différents modèles multidimensionnels (.Schéma d'étoile, Schéma de Flocon de neige (Snowflake), Schéma en grappe (Star Cluster))

Option : 01 *Le schéma à plat :*

Ce schéma est le plus simple à réaliser sans perte de données. Il est formé par agrégation des entités dans le modèle de données vers une entité minimale. Cela minimalise le nombre de tables dans la base de données, mais sans perdre aucune information du modèle original, comme la montre la figure N°23.

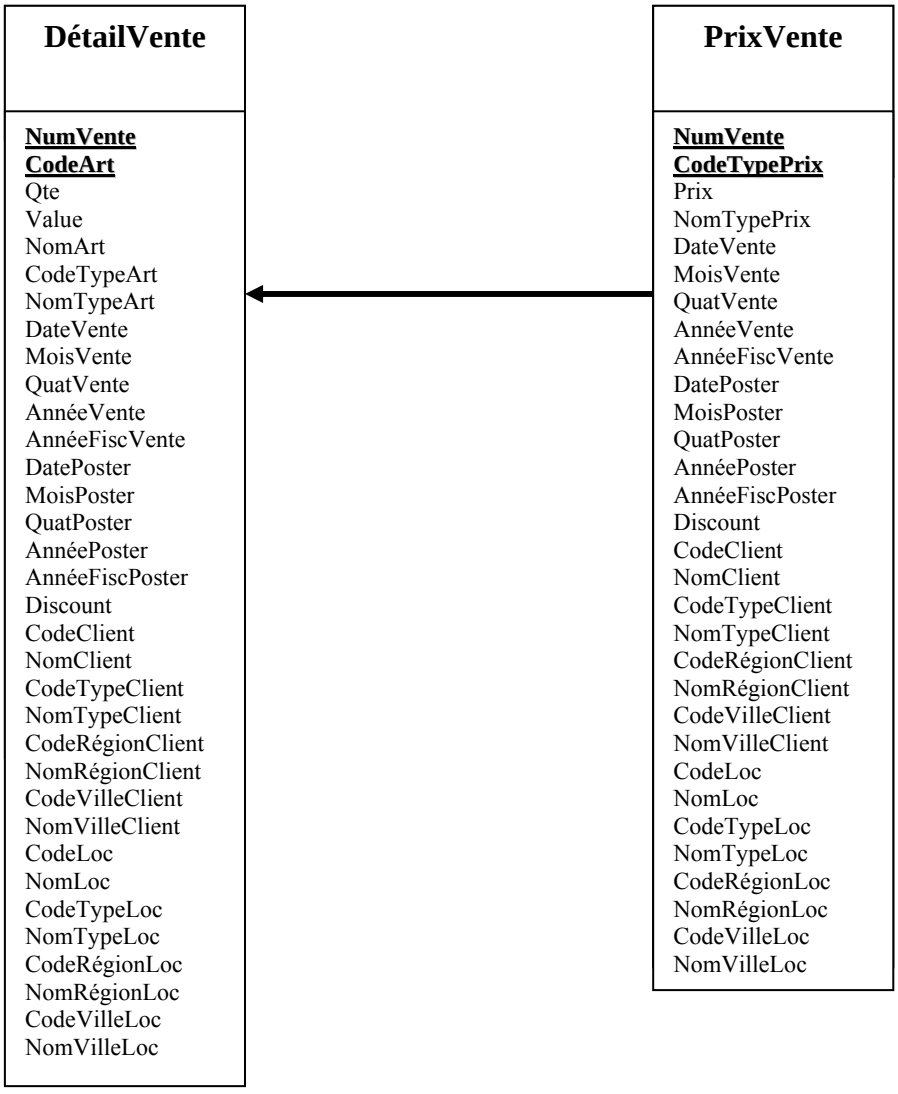


Figure N°23 : Schéma à Plat

Le problème de ce type de schéma est qu'il peut engendrer des erreurs lorsqu'il existe une relation entre les entités de transaction. Lorsque l'on agrège les montants numériques d'une entité de transaction vers une autre, cet agrégat est alors répété. Dans l'exemple de la figure n°23, si une vente contient trois articles de Ventes, le montant des ventes sera répété dans trois

enregistrements différents dans la table *DétailVente*. Ainsi, ces ajouts engendrent deux, voire trois, comptages. Un autre problème de ce schéma est qu'il propose un grand nombre d'attributs dans les tables, ce qui ne facilite pas la lecture et la compréhension. Lorsque le nombre de tables est minimal, la complexité de chaque table s'accroît.

***Option 02** : Le schéma en terrasse :*

Ce schéma est créé à partir d'une agrégation d'entités en dessous d'entités maximales, ainsi l'on s'arrête lorsque l'on rencontre une entité de transaction. Il en résulte une table pour chaque entité de transaction dans le modèle de données. Ce schéma n'est pas très apprécié car il peut poser des problèmes de compréhension aux utilisateurs non avertis, dus à la séparation entre les niveaux de transaction.

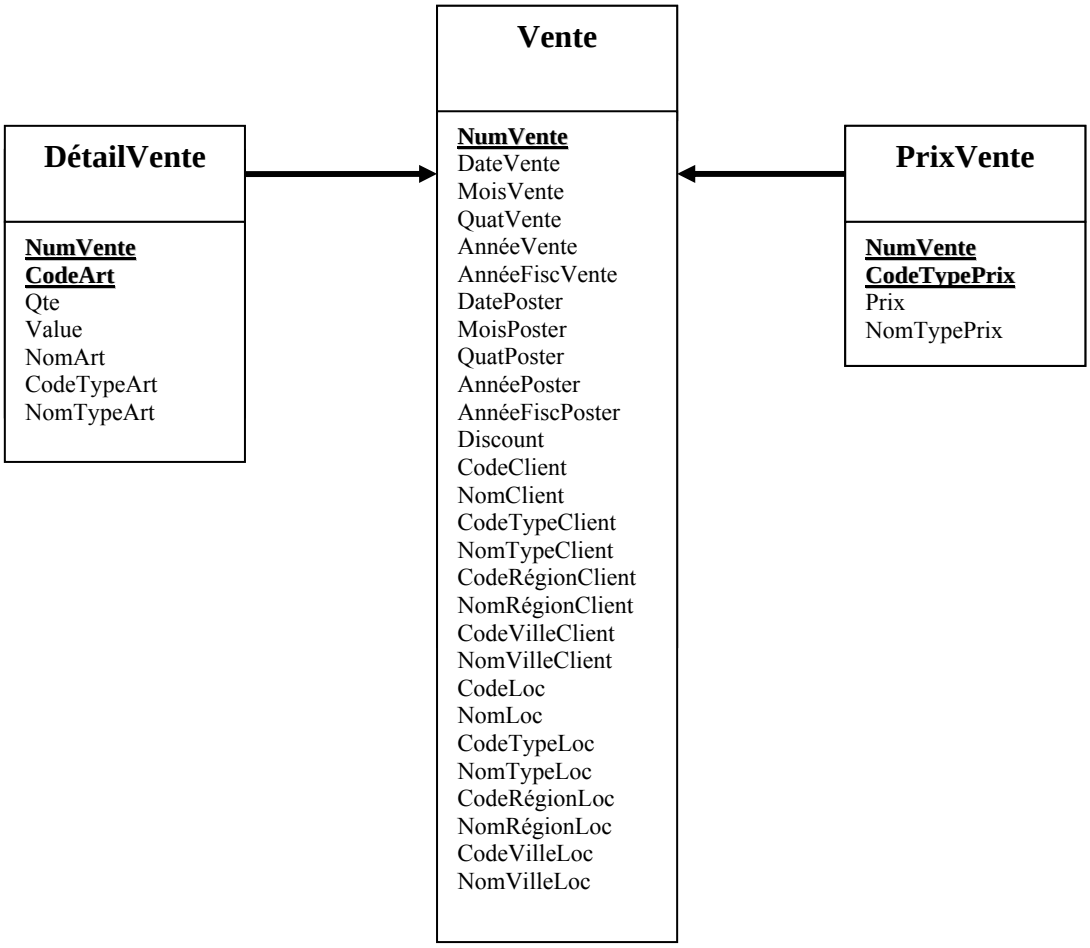


Figure N°24: Le schéma en terrasse

Option 03 *: Le schéma en étoile:*

Un schéma en étoile peut être facilement tiré d'un modèle entité -relation.

Chaque schéma en étoile est construit comme suit:

- Une table de fait est formée pour chaque entité de transaction. La clef de la table est la combinaison des clefs de ses entités associées composantes.
- Une table de dimension est formée pour chaque entité de composante, par réduction de la hiérarchie liée aux entités de classification dans cette entité composante.

- Quand des relations hiérarchiques existent entre des entités de transaction, l'entité d'enfant hérite toutes les dimensions (et clefs d'attributs) de l'entité parentale.

Cela fournit la capacité de " Drill Down " entre les niveaux de transaction

- Attributs numériques dans entités de transaction doit être agrégé par des attributs clefs (des dimensions).

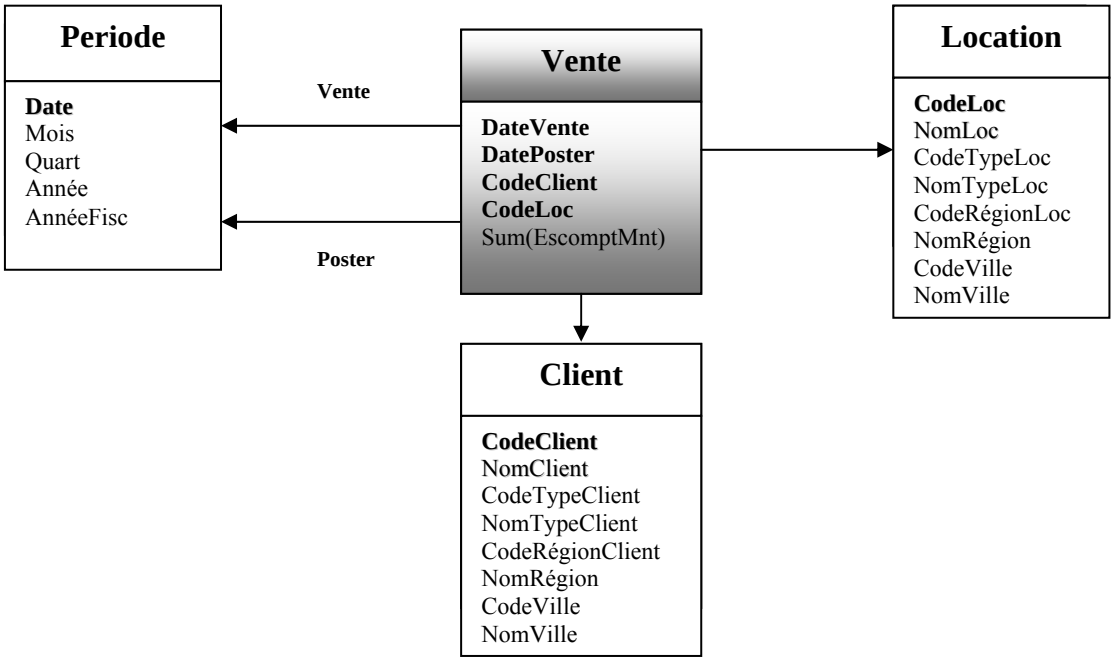


Figure N°25 : Le schéma en étoile pour les Ventes

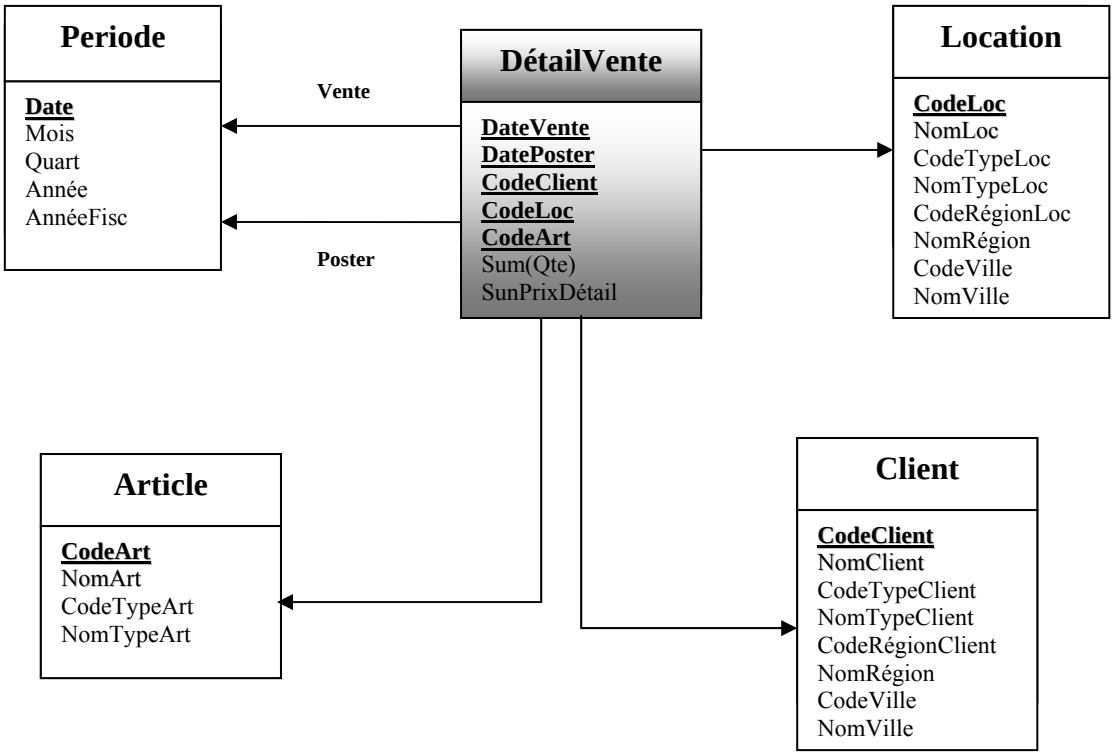


Figure N°26 : Le schéma en étoile pour Le Détail des Ventes

Schéma en Constellation: Au lieu de quelques schémas d'étoile discrets, le modèle de données peut être transformé dans un schéma en constellation. Un schéma de constellation consiste en un ensemble de schémas d'étoile avec des tables de fait hiérarchiquement liées. Les liaisons entre les diverses tables de fait fournissent la Capacité de " Drill Down " (forer en bas) entre les niveaux de détail (exemple. De *Vente* vers *DétailVente*).

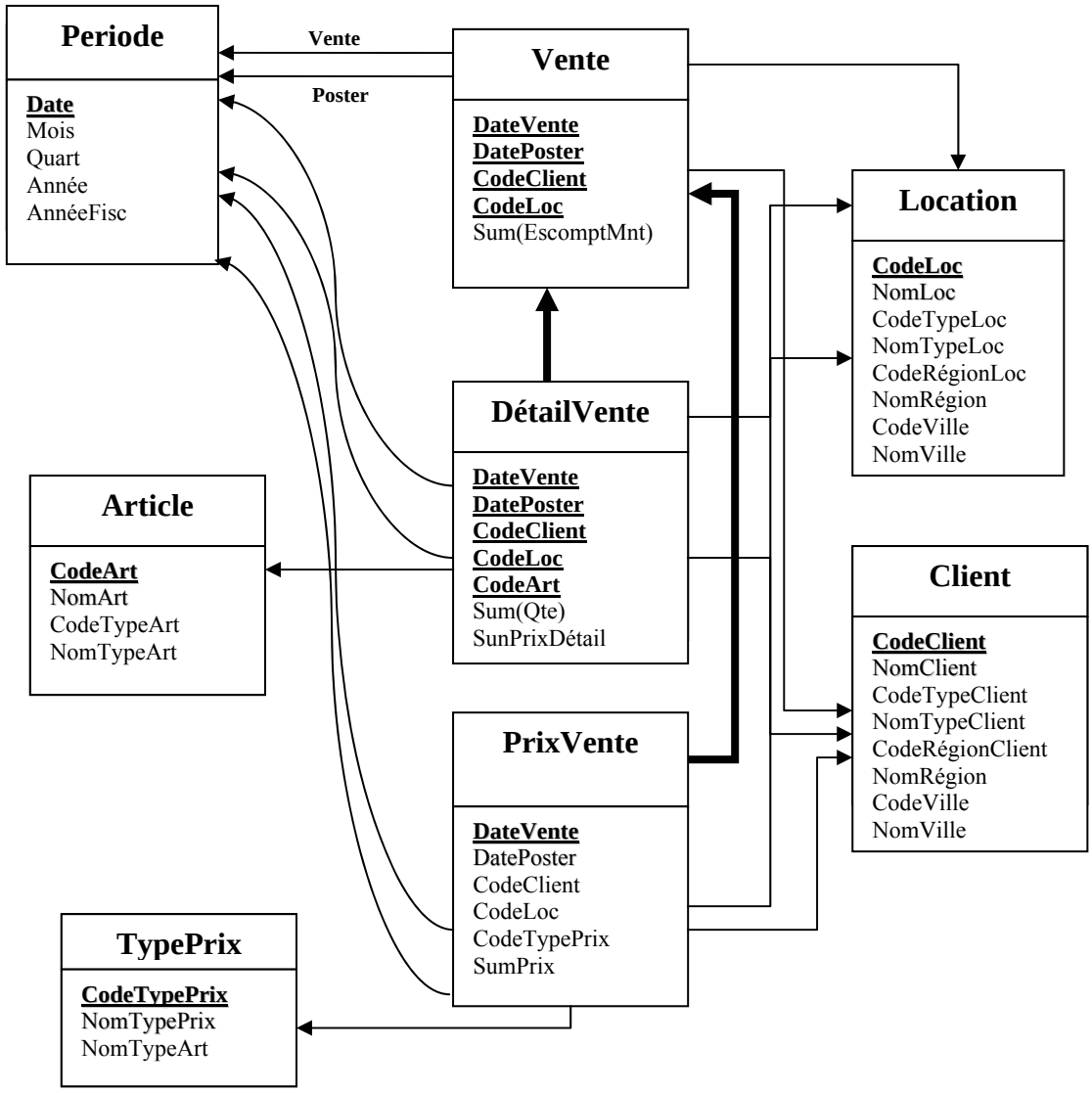


Figure N°27: Le schéma en Constellation pour les Ventes

Schéma de Galaxie

Plus généralement, un ensemble de schémas d'étoile ou constellations Peuvent être combinés ensemble pour former une galaxie. A La galaxie a d'une collection de schémas d'étoile avec des dimensions partagées, a la différence d'un schéma en constellation, les tables de fait dans une galaxie ne doivent pas être directement liées.

***Option 04** : Le schéma en Flocon de neige:*

Le schéma en flocon de neige peut être formé d'un schéma d'étoile en étendant des hiérarchies dans chaque dimension (la normalisation). Alternativement, un schéma de flocon de neige peut être Produit directement d'un modèle de Entité Relation Selon la procédure suivante :

- Une table de fait est formée pour chaque entité de transaction. La clef de la table est la combinaison des clefs de ses entités associées composantes.
- Chaque entité composante devient une table de dimension.
- Quand des relations hiérarchiques existent entre des entités de transaction, l'entité d'enfant hérite toutes les dimensions (et clefs d'attributs) de l'entité parentale.

Cela fournit la capacité de " Drill Down " entre les niveaux de transaction

- Attributs numériques dans entités de transaction doit être agrégé par des attributs clefs (des dimensions).

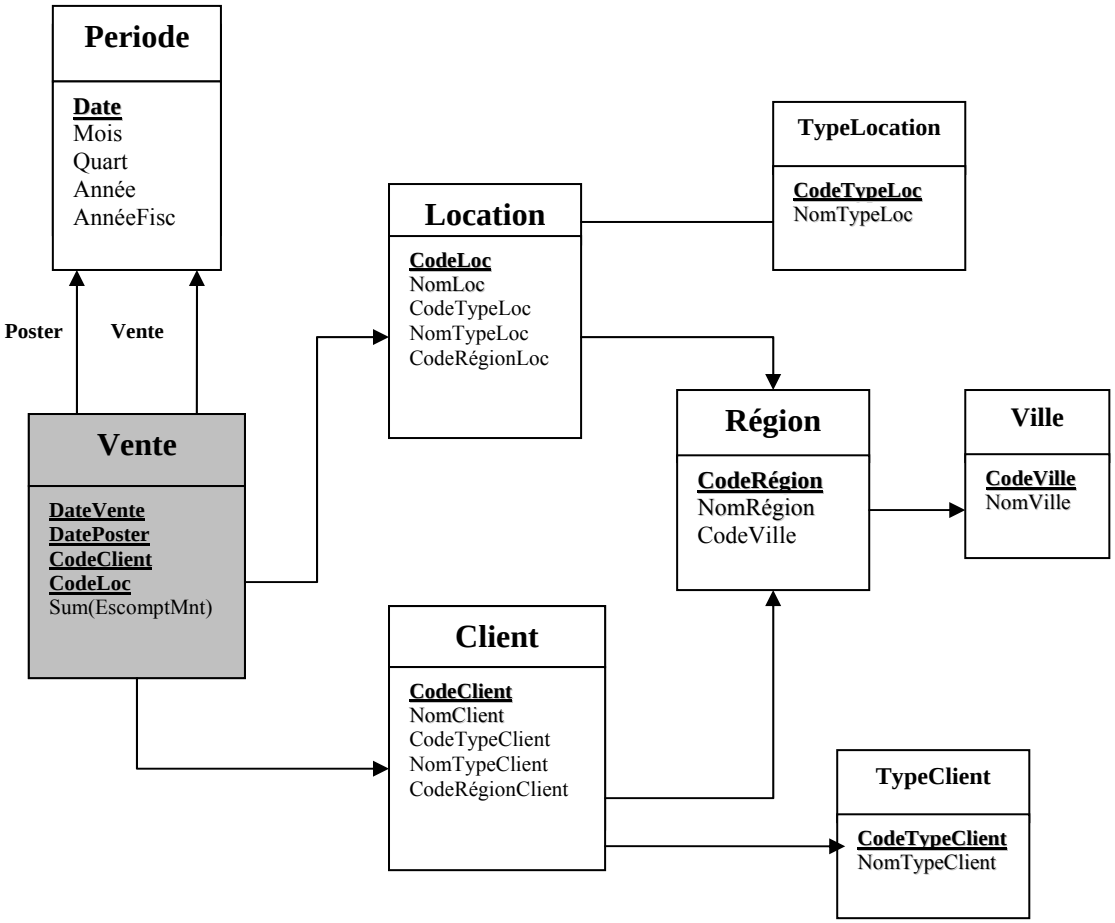


Figure N°28 : Le schéma en Flocon de Neige pour les Ventes

Option 05 *: Le schéma en groupe d'étoiles:*

Le problème de chevauchement sur des dimensions peut être identifié via "Fourchettes" dans les hiérarchies. Une fourchette arrive quand une entité se présente comme "un parent" dans deux hiérarchies dimensionnelles différentes Cela aboutit à l'effondrement de l'entité et tous ses ancêtres dans deux tables de dimension séparées. Les entités de fourchette peuvent être identifiées comme des entités de classification avec des relations multiples (un à plusieurs). Dans l'exemple du modèle de données, une fourchette arrive à l'entité Région. La région est un parent de : l'emplacement et le Client, Qui sont tous des composants de la transaction *Vente*.

Nous définissons un schéma *de groupe d'étoile* comme celui qui a Le nombre minimal de tables en évitant le chevauchement entre les dimensions. C'est un schéma d'étoile qui est sélectivement "Snowflaked" pour se séparer des segments hiérarchiques ou les sous-dimensions qui sont partagés entre les différentes dimensions. Les
sous-dimensions représentent effectivement "Le facteur le plus haut commun" entre dimensions.

Un schéma de groupe d'étoile peut être produit d'un Modèle Entité-Relation en utilisant la procédure suivante, Chaque groupe d'étoile est formé par :

- Une table de dimension est formée pour chaque entité de composante, par réduction de la hiérarchie liée aux entités de classification dans cette entité composante.
- Des entités de classification doivent être réduites en bas de leurs hiérarchies avant qu'ils n'atteignent chacun une entité de fourchette ou une entité composante. Si une fourchette est atteinte, une table de sous-dimension doit être formée, qui consistera une entité de fourchette plus tout ses ancêtres. La réduction des hiérarchies doit commencer de nouveau après l'entité de fourchette. Quand une entité composante est atteinte, une table de dimension doit être formée.
- Quand des relations hiérarchiques existent entre des entités de transaction, l'entité enfant hérite toutes les dimensions (et clefs d'attributs) de l'entité parentale.
- Attributs numériques dans entités de transaction doit être agrégé par des attributs clefs (des dimensions).

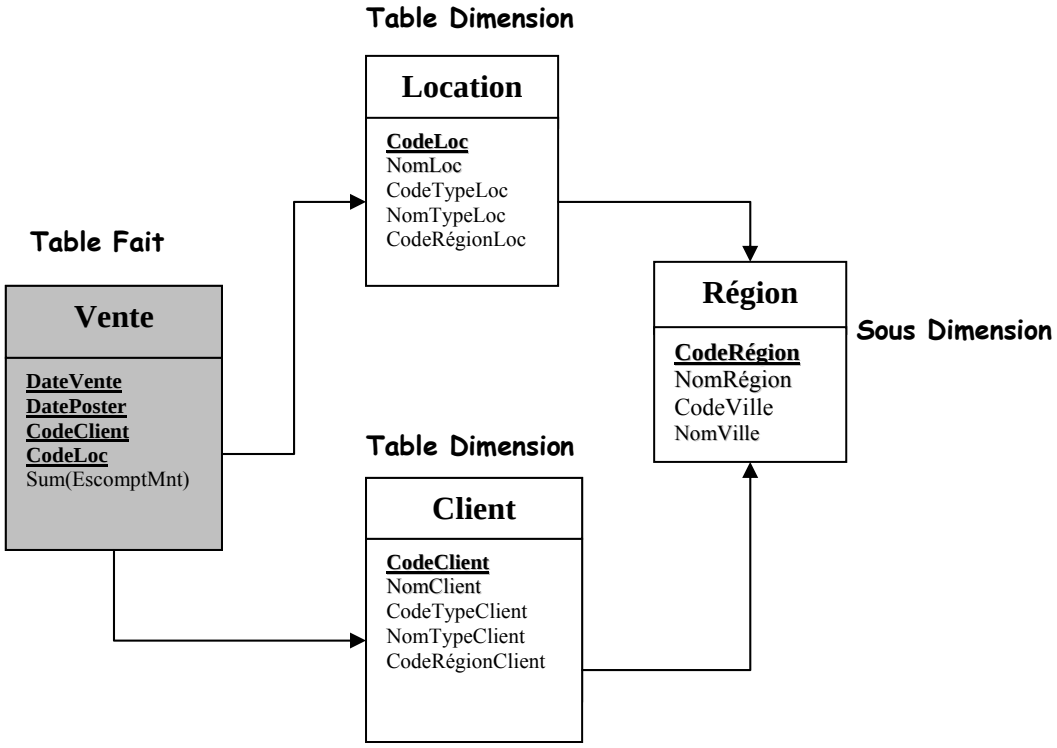


Figure N°29: Le schéma en groupe d'étoiles pour les Ventes

IV.2.5.Phase 04: Evaluation et Raffinement:

En pratique, la modélisation multidimensionnelle est un processus itératif.

Combinaison des Tables de Fait

Tables de fait avec les mêmes clefs primaires (c'est-à-dire les mêmes dimensions) doivent être combinées. Cela réduit le nombre de schémas d'étoile et facilite la comparaison entre les faits liés.

Combinaison des Tables de Dimension

Création de tables de dimension pour chaque entité composante aboutit souvent à un grand nombre de tables de dimension. A simplifier la structure de Data Marts, des dimensions liées doit se consolider ensemble dans une simple table de dimension.

Relations Plusieurs à Plusieurs:

La plupart des complexités qui surgissent dans la conversion du modèle Entité Relation traditionnel à un modèle dimensionnel résulte des relations " Plusieurs à Plusieurs ", ou intersection des entités. Il y a quelques options pour procéder avec ce type de relation:

(**a**) Ignorer l'entité d'intersection (éliminez-le du Data Mart correspondant)

(**b**) Convertir la relation " Plusieurs à Plusieurs " à une relation "un à Plusieurs ", en définissant une relation "primaire"

(**c**) Insérer cette relation "primaire" comme un relation " Plusieurs à Plusieurs" dans le Data Mart comme des entités peuvent être utiles pour Les analystes expert mais ne seront pas susceptibles d'être analyser en utilisant les outils OLAP.

Sous-types de Traitement

Les relations de supertype / soustype peuvent être convertis à une structure hiérarchique en enlevant les sous-types et on créant une entité de classification pour distinguer entre Sous-types.

IV.2.6. Conclusion:

Les auteurs **[KOR00]** proposent une méthodologie pour construire un modèle multidimensionnel. Les schémas cibles sont décrits en première partie, à savoir : le schéma en étoile, le flocon de neige, la constellation, la galaxie, le schéma en grappe, le schéma à plat et le schéma en terrasse.

Quelque soit le modèle choisit, la procédure de création de la table des faits et la création de ses clés est :

1°) Une table des faits est formée pour les entités de transaction les plus pertinentes.

La clé de cette table est une combinaison des clés des tables de dimension.

2°) Si une relation hiérarchique existe entre deux entités de transaction, l'entité fille hérite de tous les attributs de son père.

3°) Les attributs numériques appartenant aux entités de transaction doivent êtres agrégés par les clés.

4°) Les tables des faits avec les mêmes clés primaires doivent fusionner.

Les auteurs [**KOR00**] nous proposent de partir du modèle d'entreprise existant, et de le transformer, soit en schéma à plat, soit en terrasse, soit vers les autres modèles multidimensionnels. De plus, pour chaque schéma, il faut rajouter certaines spécificités propres à chacun. Par exemple :

1°) Dans le cas d'un modèle en étoile : une table de dimension est formée pour chaque entité composante par agrégation de la hiérarchie constituée de ses entités de classification.

2°) Pour le modèle en flocon de neige : chaque entité composante devient une table de dimension. A partir de chaque entité composante et ses entités de classification on dérive la hiérarchie d'une dimension.

IV.3 Conception d'un schéma conceptuel selon L'approche B.Husemann et J. Lechtenborger et G. Vossen:

IV.3.1 Introduction:

La contribution de la démarche de [**HLG00**] est triple :

1- L'établissement des directives pour répondre à la question si un attribut est un niveau de dimension ou un attribut de propriété.

2- Proposition d'un formalisme graphique pour la conception conceptuelle d'entrepôt

3- Généralisation de l a forme normale multidimensionnelle.

IV.3.2 - Terminologie et notation

IV.3.2.1 Structures multidimensionnelles de base

- ***Les faits*** représentent des éléments atomiques de l'information dans une base de données multidimensionnelle. Un fait consiste à évaluer quantitativement des

valeurs stockées dans des mesures et un contexte de qualification qui est déterminé par des niveaux de dimension (terminaux).

-***Niveau de dimension*** contient un ensemble d'instances ou des éléments.

- ***Chemin d'agrégation*** est une sous séquence de niveaux de dimension, qui commencent dans une niveau de dimension terminale et aboutit à un niveau de dimension implicite ***All*** contenant un élément simple ‹***All***›, soit un chemin d'agrégation (d_1,……..,d_n), où d_i est un niveau de dimension,$1 \leq i \leq n$, chaque élément de d_i appartient à (ou est associé à) au maximum un élément de d_{i+1}, $1 \leq i \leq n-1$; de plus, nous disons que d_{i+1} est un niveau (d'agrégation) plus haut que d_i.

- ***Une dimension*** : est structurée en termes de chemin d'agrégation ou plus qui partagent le même niveau de dimension terminal. Donc, chaque dimension comprend au moins un niveau de dimension terminal et le niveau implicite ***All***.

- ***La hiérarchie de dimension*** est le graphique consistant de tous les chemins d'agrégation pour une dimension donnée (bien qu'un niveau de dimension puisse avoir plus qu'un niveau parental). Les ensembles des niveaux de dimension de dimensions distinctes sont disjoints.

- ***Un attribut de propriété*** décrit l'information complémentaire liée à un niveau de dimension (par exemple, la propriété attribue *NomClient* pour le niveau de dimension *CodeClient* dans la Figure 30). Un attribut de propriété peut être employé pour délimiter l'ensemble résultant d'une question multidimensionnelle, mais ne détermine pas son niveau d'agrégation.

- ***Un attribut de propriété facultatif*** (par exemple, *AgeClient*) n'est pas toujours spécifié pour chaque élément du niveau de dimension correspondant et peut donc avoir des valeurs ‹nulles›.

IV.3.2.2 Hiérarchies de dimension

Les hiérarchies de dimension sont classifiées dans deux types de base.

- Une hiérarchie simple consiste en exactement une chemin linéaire d'agrégation dans une dimension (exemple: chemin *day* →*month* →*year* dans la dimension *temps* Figure 30).

- Une hiérarchie de dimension multiple contient au moins deux chemins d'agrégation différents dans une dimension (exemple: la dimension *Compte* de Figure 30).

1- Un groupe de chemins d'agrégation enracinés dans un niveau de dimension commun ***d*** est appelé *l'alternative*, si chaque élément de ***d*** appartient à exactement un élément de chaque niveau plus haut (par exemple, le chemin de Figure 3 départs dans le niveau de dimension *CodeOrg*).

2 - Un groupes de chemins d'agrégation facultatifs. Ou quelque niveau de dimension ***dl*** a deux ou plus de niveaux plus hauts tel que pour chaque élément de ***dl*** il y a exactement un élément dans un niveau plus haut auquel cet élément appartient.

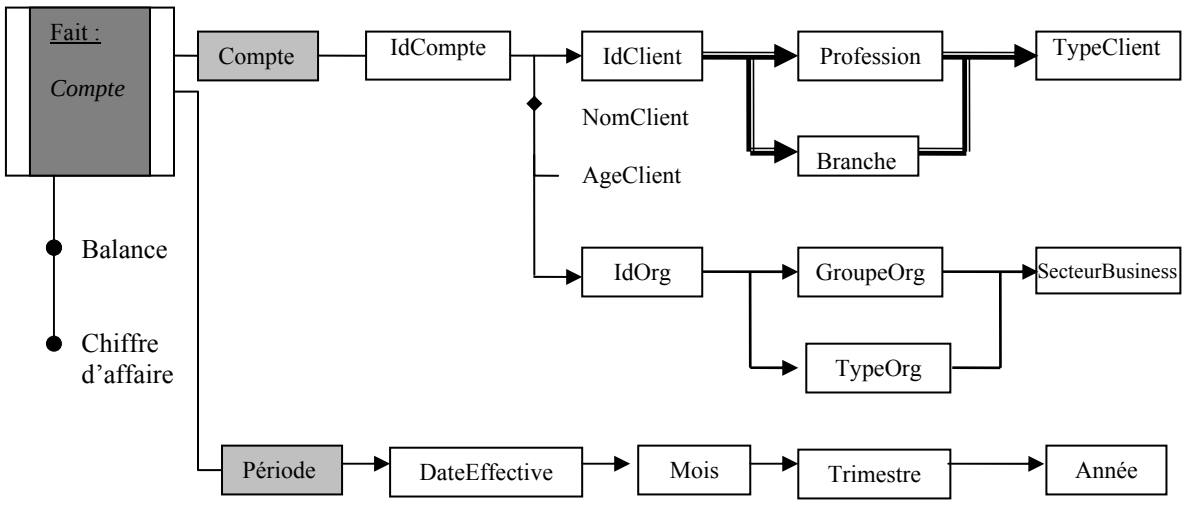


Figure N°30 : Schéma d'un Fait Compte et Dimension Compte et Période

Par exemple, dans la Figure 30 le groupe de chemins d'agrégation commençant dans le niveau de dimension *CodeClient* est facultatif et chaque élément de niveau de dimension *CodeClient* est rapproché ou bien de *profession* ou bien une *branche*); Dans un groupe facultatif de chemins d'agrégation un niveau de dimension comme *CodeClient* est appelé Le niveau division **(**ou partage) tandis que *TypeClient* est appelé le niveau joint.

Dans notre notation graphique,

- des hiérarchies simples et les groupes alternatifs de chemins d'agrégation sont indiqués par des flèches simples.
- Les groupes facultatifs de chemins d'agrégation sont spécifiés par des flèches doubles alignées.
- Un attribut de propriété obligatoire est connecté via un diamant à son niveau de dimension et un attribut de propriété facultatif est sans connexion.

IV.3.3 Un modèle de processus pour conception de DataWarehouse

Le processus de conception d'entrepôt de données comprend quatre phases séquentielles comme le processus de conception de base de données classique.

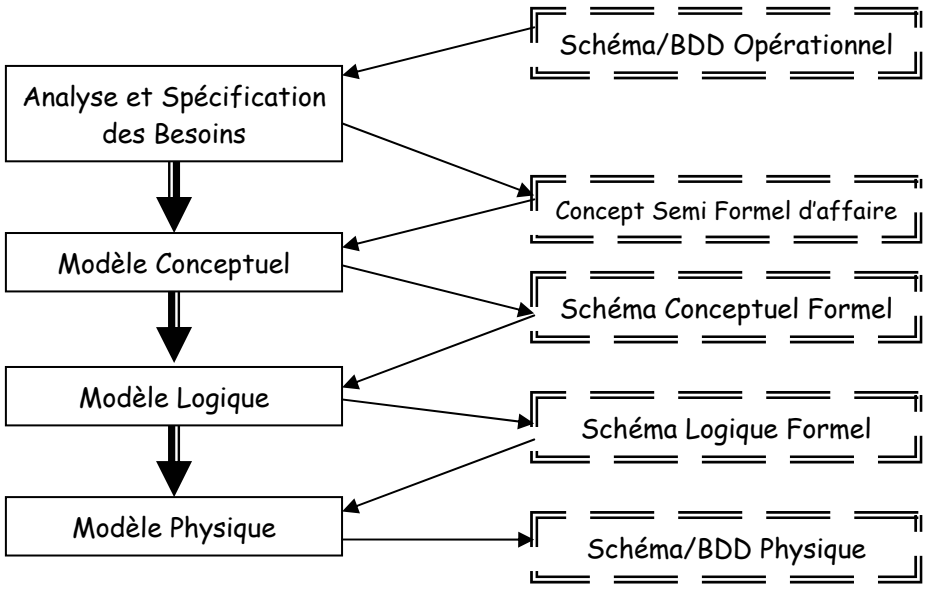


Figure N°31 : Processus de Modélisation pour la Conception d'un DW

√ *L'analyse et spécification de besoins:*

Le schéma E/R opérationnel livre l'information de base pour déterminer l'analyse potentielle multidimensionnelle. Dans cette phase les experts du domaine choisissent des attributs stratégiquement appropriés de la base de données opérationnels et spécifient les propositions de les employer comme des dimensions et/ou des mesures. Pour chaque attribut il est nécessaire de se décider s'il contient des données facultatives ou non.

✓ modèle Conceptuel

La phase du modèle de conception exécute une transformation de la spécification de besoins semi formelle dans un schéma formalisé conceptuel multidimensionnel. La formalisation aboutit à un diagramme graphique multidimensionnel, qui comprend le schéma de fait avec leurs mesures liées et des hiérarchies de dimension. Pour chaque mesure d'un fait les contraintes de la "summarizability " sont formalisées dans une table annexe.

✓ Le modèle logique

La phase de conception logique convertit le schéma conceptuel au schéma logique en respectant le modèle de données logique cible (surtout relationnel ou multidimensionnel). Schéma logique est produit selon des règles de transformation, qui seulement se réfèrent aux diagrammes conceptuels développés et les contraintes de la " summarizability".

✓ Le modèle physique

Le processus de conception d'entrepôt de données aboutit à une mise en oeuvre physique de schéma logique avec respect des propriétés individuelles du système

de base de données cible, incluant des techniques d'optimisation physiques comme des stratégies d'indexation généralement connues, le partitionnement etc.,

IV.3.3.4 Modélisation Conceptuelle du DataWarehouse

Le but de la phase de conception de schéma conceptuel est de produire un schéma graphique multidimensionnel. Exprime pour chaque mesure son contexte multidimensionnel en termes de dimensions appropriées, et leurs hiérarchies, et pour cela ça suppose que :

1- Le schéma E/R global opérationnel initial, qui décrit l'information à la source est disponible.
2- L'analyse des besoins a été effectuée conjointement avec des experts de domaine
3- Le schéma E/R global opérationnel a été analysé pour déterminer des mesures et des dimensions intéressantes, et les initiales questions (requêtes) d'OLAP.

La production de cette phase consiste en :

(a) – **Tables**: qui contient une description informelle pour chaque attribut approprié et indique si l'attribut peut être employé comme une mesure ou attribut dimensionnel et si l'attribut est facultatif ou pas (exemple: l'extrait concernant l'information de *Compte* montrée dans la Table 5)
(b) - Des questions (requêtes) standard multidimensionnelles comme "Montrer le chiffre d'affaires moyen de l'année par groupe de produit", ou "combien de comptes courants sont gérés en ligne."

Il est a noter que la Table 1 contient les attributs complémentaires qui représentent les besoins multidimensionnels qui ne font pas partie du schéma opérationnel (exemple: *mois*, *TypeClient*).

Un processus de modèle conceptuel d'entrepôt de données est subdivisé en trois phases séquentielles :

1. Définition de contexte de mesures,
2. Conception de hiérarchie dimensionnelle.
3. Définition des contraintes "summarizability".

- ✓ *Définition de contexte de mesures*

étant donné l'ensemble **M**= $\{m_1,m_2,.....,m_k\}$ de mesures définies pendant l'analyse de besoins, et l'ensemble **D** d'attributs dimensionnels, chaque fait peut être perçu comme un élément d'un graphe d'ensemble de fonction de: niveaux de dimension vers les mesures. De là, la phase de conception démarre en déterminant des dépendances fonctionnelles (**DFs**) a partir des niveaux de dimension vers (aux) les mesures.

D'abord, nous déterminons une clé (minimale) $D_i \subseteq D$ pour chaque mesure m_i; et soit $\mathbf{F_{clé}}$ l'ensemble de tout les **DFs** de la forme $D_i \rightarrow m_i$ ainsi obtenu. Donc, étant donné une **DF**

$D_i \rightarrow m_i \in \mathbf{F_{clé}}$, les niveaux de dimension dans Di déterminent fonctionnellement la mesure mi, mais ne sont pas fonctionnellement déterminés par aucun autre niveau. De là, on les qualifie comme des niveaux de dimension terminaux qui sont employés comme les racines de hiérarchies de dimension.

Pour chaque niveau de dimension terminale est définie une dimension correspondante. Dans l'exemple, on a la **DF** (Num*Compte, DateEffective*) $\rightarrow$*balance*$\in \mathbf{F_{clé}}$ pour la mesure *balance*. En outre, toutes les mesures mi, mj avec Di =Dj sont groupées dans le même schéma de fait, qu'ils partagent le même contexte dimensionnel.

Attribut	Description	M	D	O
DateEffective	Date Effective	Non	Oui	Non
Mois	Temps d'agrégation	Non	Oui	Non
Trimestre	Temps d'agrégation	Non	Oui	Non
Année	Temps d'agrégation	Non	Oui	Non
NumCompte	N° de Compte	Non	Oui	Non
Balance	Balance d'une Date Effective	Oui	Non	Non
ClasseBalance	Classification de Balance	Non	Oui	Non
ChiffAffaire	Chiffre d'affaire d'une Date	Oui	Non	Non
ClasseChiffAffaire	Effective	Non	Oui	Non
CréditLimit	Classe de Chiffre d'affaire	Oui	Non	Non
intérêt	Limite crédit d'un Compte	Oui	Non	Non
	Taux d'intérêt			
CodeClient	Code Client	Non	Oui	Non
NomClient	Nom de Client	Non	Oui	Non
AgeClient	Age de Client	Non	Oui	Oui
TypeClient	Classification de Client	Non	Oui	Non
Profession	Profession d'un Client	Non	Oui	Oui
Branche	Branche d'un Client	Non	Oui	Oui
CodeProduit	Code Produit	Non	Oui	Non
TypeProduit	Classification de Produit	Non	Oui	Non
CodeOrg	Code d'Organisation	Non	Oui	Non
NomOrg	Nom d'Organisation	Non	Oui	Non
GroupeOrg	Groupe d'Organisation	Non	Oui	Non
TypeOrg	Classification d'Organisation	Non	Oui	Non
SecteurAffaire	Classification d »un GroupeOrg et	Non	Oui	Non
	TypeOrg			

Table 4 : Spécification des Besoins

Schéma de Fait	Mesures	Dimensions	Niveau de Dimension Terminal
FaitCompte	Balance ChiffAffaire CréditLimit intérêt	Compte	NumCompte
		Periode	DateEffective

Table 5 : Dépendance fonctionnelle entre les Mesures et les Niveaux de Dimension Terminal

Table 6 expose le résultat de ce processus appliqué à la mesure *balance, ChiffAffaire*, *CréditLimit* et *Intérêt* figurant dans la Table 5.
Toutes les mesures dépendent fonctionnellement des mêmes niveaux de dimension terminaux, à savoir *NumCompte* et *Dateffective*, qui appartenez à leur tour aux dimensions *Compte* et *Periode* respectivement. De là, toutes les mesures sont groupées dans un Schéma commun de fait *FaitCompte*. A ce point, nous commençons la Conception graphique en modélisant le schéma de fait Aux niveaux de dimension terminaux (voir la Figure 32).

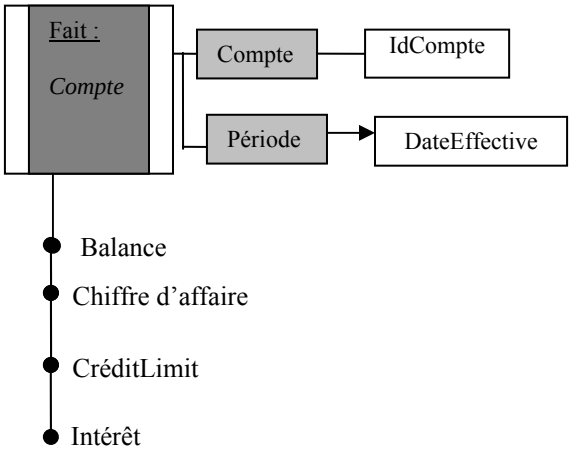


Figure N°32 : Schéma d'un Fait Compte et Dimension Compte et Période

- ✓ Conception de hiérarchie dimensionnelle

Dans l'étape suivante, il y a le développement graduel des hiérarchies de dimension pour chaque dimension. A la fin, on arrive a toutes les **DFs** entre les niveaux de dimension appartenant à une dimension ***Dim*** avec le niveau de dimension terminale $\mathbf{d_j}$ comme suit :

Supposons qu'on nous donne les niveaux de dimension $\mathbf{d_k}$,$\mathbf{d_l} \in \mathbf{D}$ tel que $d_k \rightarrow d_l$ est une **DF** valable et il existe une (potentiellement transitif) dépendance fonctionnelle de d_k sur d_j, alors nous ajoutons $d_k \rightarrow d_l$ à l'ensemble $\mathbf{F_{dim}}$.
Dans notre exemple, le schéma de fait *FaitsCompte* inclut *NumCompte* et *DateEffective* comme Niveaux dimension terminale. Le départ avec le Niveaux

dimension terminale *DateEffective* de la dimension *Temps*, nous déterminons les **DFs** suivant:

F_{time} ={ *DateEffective* → Mois, Mois→Quart, Quart →Annee}

Graphiquement, nous tirons la hiérarchie de dimension simple montré dans la Figure 33.

=

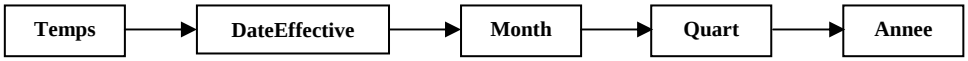


Figure N°33 : Hiérarchie Simple de la Dimension Temps

Les niveaux de dimension de la dimension *Compte* présente les **DFs** suivantes:

F_{Compte} ={ *IdCompte* → IdOrg, *IdCompte* → IdClient, *IdCompte* →*ClasseChiffreAff,*

IdCompte → ClasseBalance, *IdCompte* → IdProduit,

IdProduit → TypeProduit, IdOrg →GroupeOrg, IdOrg →TypeGroupe ,

TypeOrg →SecteurBusiness, GroupeOrg →SecteurBusiness,

IDOrg →NomOrg, IDClient →Profession, IDClient →Branche,

IDClient →Profession, Profession →TypeClient, Branche →TypeClient,

IDClient →NomClient, IDClient →AgeClient}

Dans une **première étape**, les attributs de propriété et les niveaux de dimension doivent être distingués selon les besoins d'analyse (les attributs de propriété sont utilises pour les sélection, mais pas pour les agrégations).

Ensuite, dans une **deuxième étape**, une approximation brute de la hiérarchie de dimension est obtenue en construisant un graphe dirigé dont les noeuds sont des niveaux de dimension. Ce graphe contient Un bord : de niveau de dimension $\mathbf{d_i}$ vers

le niveau $\mathbf{d_j}$, si di ≠dj Et di →dj est une **DF** non transitif, c'est-à-dire, si di →dj est valable et il n'y a aucun niveau de dimension dk
(dk ≠di,dj) tel que di →dk →dj qui soit aussi valable. Le graphe obtenu est jusqu'ici augmenté avec les attributs propriété : l'attribut de propriété $\mathbf{d_p}$ est attaché au niveau de dimension $\mathbf{d_l}$ si la **DF** $d_l \rightarrow d_p$ est non transitif. L'information de savoir si un attribut de propriété est facultatif ou non, peut être recouvre de la spécification des besoins.

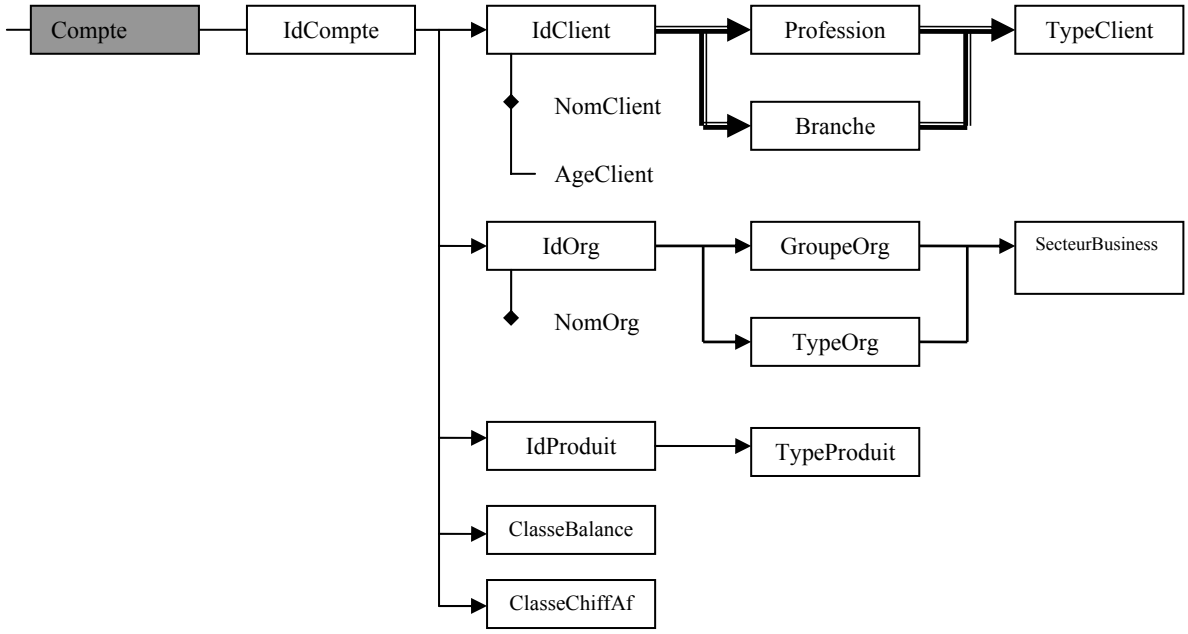


Figure N°34 : Hiérarchie Multiple de la Dimension Compte

- ✓ *Définition de contraintes summarizability*

Un modèle conceptuel doit fournir les moyens de distinguer les agrégations significatives des mesures, le schéma d'entrepôt doit exprimer explicitement quelle mesure peut être agrégée le long de la hiérarchie de dimension et selon quelle fonction d'agrégation. Il y a quatre degrés de niveau de restriction pour les mesures. Etant donné une paire (m, **d**) d'une mesure **m** et d'un niveau de

dimension **d**, on associe la restriction de niveau **1**, si toutes les fonctions d'agrégation peuvent être appliquées pour le *roll_up* **m** à partir de niveau de dimension **d** à chaque dépendance fonctionnellement de niveau le plus haut. La restriction de niveau **2** est associée aux paires (**m, d**), où toutes les fonctions d'agrégation sont autorisées sauf l'opérateur **SUM**. La restriction de niveau **3** représente la limitation la plus haute, où L'agrégation est toujours possible, mais seulement en termes de comptage. Finalement, degré **4** aucune fonction d'agrégation n'est permise.

L'étape suivante doit définir des niveaux de restriction pour toutes les mesures le long des différents chemins d'agrégation dans chaque schéma de fait :

Pour chaque paire de mesures et de niveau de dimension on définie un niveau de restriction tel que chaque question multidimensionnelle basée sur la fonction d'agrégation permise le long de chaque chemin est significative.

Niveau de Restriction	Fonction d'agrégation Applicable
1	{SUM, AVG, MIN, MAX, STDDEV, VAR, COUNT}
2	{AVG, MIN, MAX, STDDEV, VAR, COUNT}
3	{COUNT}
4	{ }

Table 6 Classification des niveaux de restriction

Schéma de Fait	Mesures	Niveaux de dimension	Niveau de Restriction
Fait: Compte	Balance	IdCompte	1
		DateEffective	2
	ChiffreAffaire	IdCompte	1
		DateEffective	1
	LimiteCredit	IdCompte	2
		DateEffective	2
	InterSet	IdCompte	2
		DateEffective	2

Table 7 Annexe de Summarizabilty pour le Schéma de Fait: Compte

PARTIE V APPROCHE PROPOSEE AVEC UNE ETUDE DE CAS:

V.1 - Introduction :

Dans ce chapitre nous aborderons la partie méthodologie ou démarche de la conception multidimensionnelle d'un entrepôt de donnée.

Nous avons opté de consacrer un chapitre à cet effet par rapport à l'importance du choix qui nous ait donné de faire et de l'importance d'une bonne démarche dans le développement de l'entrepôt, pour cela nous utilisons les 03 méthodes :

- La méthode classique de **Ralph Kimball**.
- La méthode de conception à partir d'un modèle d'entreprise de **Mark A.R.Kortink** et **Daniel L. Moody**.
- La méthode de conception de **Bobo Husmann** et **al**.

Chacune des ces méthodes a été présentée dans la partie précédente, pour finalement, nous optons en fonction des avantages de chacune pour une démarche adaptée à notre étude de cas.

V.2. Présentation des méthodes de conception :

Une méthode de conception adaptée à notre problématique retenue est primordiale pour atteindre un haut niveau de performance. Ceci n'est pas chose aisée, car l'inexistence de méthodologie faisant l'unanimité telle que **MERISE** dans la modélisation entité-association fait qu'au moment du choix de la méthode, une réflexion approfondie s'avère indispensable. Nous commençons par identifier les points que nous jugeons faibles pour les 03 méthodes :

- *Les critiques de la méthode de Kimball sont résumées par les points suivants :*

1. Pas de formalisme explicite pour la conception de entrepôt de données
2. L'approche est présentée à travers des exemples plutôt qu'à travers une procédure de conception explicite (la présentation n'est pas formalisée).

3. Kimball ne fourni aucune méthodologie permettant d'exploiter les schémas des systèmes existants et surtout le modèle E-R.

*- Les critiques de la méthode de **Kortink et al** sont résumées par les points suivants :*

1. L'approche ne couvre pas le cycle de vie dimensionnel.
2. Absence des spécifications des besoins décisionnelles.
3. Classification des entités existe, mais en revanche, pas d'indication sur le sort des propriétés du schéma E-R lors de passage au schéma conceptuel dimensionnel.

*- Les critiques de la méthode de **Bobo Husemann** et al sont résumées par les points suivants :*

1. L'exploitation de schéma conceptuel E-R initial n'est pas explicite
2. Beaucoup d'efforts pour la normalisation des schémas multidimensionnels dans le but d'économiser l'espace disque (le gain de la normalisation est de l'ordre 1 /100 **[Kim97]**).

V.3. CHOIX DE L'APPROCHE

Aux vues des caractéristiques des trois méthodes, nous pouvons conclure en disant que les trois méthodes se complètent, la méthode proposée par **Kimball** en ajoutant un formalisme qui permet d'exploiter un système existant.
Notre approche sera inspirée des 03 méthodes précédemment décrites. Ainsi, nous utiliserons l'approche de Kortink et Moody **comme approche de base** pour construire à partir du modèle E/A opérationnel le modèle conceptuel multidimensionnel, tout en prenant en considération les principes de bases de la méthode de Kimball et le formalisme de bobo Husemann et al, car a notre sens **Kortink** possède dans sa démarche les points suivant :

- dans la phase 03 l'opérateur « Réduction de la hiérarchie » avec plusieurs possibilité de schémas (étoile, plat, etc..) donne la possibilité au concepteur le choix de normaliser ou dénormaliser (Etoile et plat contre flocon en neige etc. …)
- Utilisation du schéma E-R opérationnel pour construire le schéma conceptuel multidimensionnel. L'exploitation de schéma conceptuel E-R initial est explicite.
- Une approche facile et efficace.
- Il y a un large choix d'options pour la production des Modèles multidimensionnels à partir d'un modèle Entité Relation, et qui incluent les schémas:

Plat (Flat), Terrasses (Terrace), Etoile (Star), Flocon de neige (Snowflake), Grappe (Star Cluster), Chacune de ces différentes options représente le compromis entre la complexité et la redondance.

V.4. Démarche de l'Approche :

Notre approche doit essentiellement s'articuler à déterminer les faits, les mesures, les dimensions, les hiérarchies, le grain de chaque fait. Pour cela nous proposons une approche à 04 étapes :

V.4.1. Planification du projet

But : se conformer au cycle de vie dimensionnel **[Kim97]**.

V.4.2. Définition et analyse des besoins Décisionnels

But : de bien comprendre les décideurs et les futurs « consommateurs » de la base de données décisionnelles, pour obtenir les attributs stratégiques et les employer soit comme des mesures et/ou dimension.

- Les questions OLAP, qui a un intérêt d'analyse.
- Tableau des spécifications des besoins (Bobo Husemann et al)

V.4.3. Modélisation dimensionnelle des données

• **Identifier les activités:**

But : Un processus d'activité est un processus opérationnel important pour l'organisation, étaye par une application existante à partir de laquelle des données peuvent être collectées au profit de l'entrepôt de données **[KIM00]**.

• **Identifier les sous schémas conceptuels initiales (opérationnel)** :

But : A partir Schéma E-R initial, recenser tout les sous MCD en rapport avec l'activité, afin de pouvoir classifier les entités.

• **Classification des entités, et détermination du grain.**

But : La classification doit faire apparaître les entités composantes, transactionnelles et de classification, les plus en accords avec les indicateurs et les axes identifiés pour l'activité lors de l'analyse des besoins. Pour les entités transactionnelles nous représenterons les faits qu'elle contient et correspond aux spécifications des besoins. , le grain est le niveau de détail fondamental, atomique, des données figurant dans la table des faits pour ce processus. Des grains typiques sont des transactions individuelles, des récapitulations individuelles quotidiennes [**KIM97**] **[Kor00]**

• **Identification des hiérarchies.**

But : Dans cette étape on déterminer toutes les hiérarchies maximales, et les entités minimales et maximales.

La hiérarchie est un concept extrêmement important dans le modèle multidimensionnel, Une hiérarchie est un modèle d'entité-association qui est identifié par des séquences d'entités reliées par des relations (un à plusieurs), toutes alignées dans le même sens.

- **Production du Modèle dimensionnel :**

  But : On utilisant les résultats d'étape (Classification des entités), et l'étape d'identification des hiérarchies, on peut construire notre modèle multidimensionnel. Il y a un large choix d'options pour la production des Modèles multidimensionnels à partir d'un modèle Entité Relation, Chacune de ces différentes options représente le compromis entre la complexité et la redondance et qui incluent les Schémas : Plat (Flat), Terrasses (Terrace), Etoile (Star), Flocon de neige (Snowflake), Grappe (Star Cluster).

- **Définition des contraintes d'agrégations (Summarizabilty) :**

  But : Distinguer les agrégations significatives des mesures, exprimer explicitement quelle mesure peut être agrégée le long de la hiérarchie de dimension et selon quelle fonction d'agrégation. **[HLG00]** définit quatre degrés de niveau de restriction pour les mesures.

V.4.4. Conception et développement de la zone de préparation

But : Définir le schéma et l'architecture de la phase **ETL.**

Pour remplir un entrepôt de données, il faut : **[SQL04]**

– Une étape d'extraction (des données pertinentes des les bases de production).

– Une étape de transformation (nettoyage, formatage, premières agrégations et reconnaissance des membres).

– Une étape de chargement (des données propres dans la base décisionnelle).

En anglais on parle de phase **ETL** pour **E**xtraction, **T**ransformation et **L**oading.

Les sources de données sont généralement multiples et gérées par différents systèmes (géographiquement répartis dans différents sites). Chaque situation rencontrée est très spécifique et l'architecture **ETL** mise en place est souvent dédiée à chaque entreprise.

La fréquence à laquelle les phases **ETL** sont opérées doit être cohérent avec le grain de la dimension temporelle et doit permettre d'historiser les données avant qu'elles ne soient purgées des bases de production.

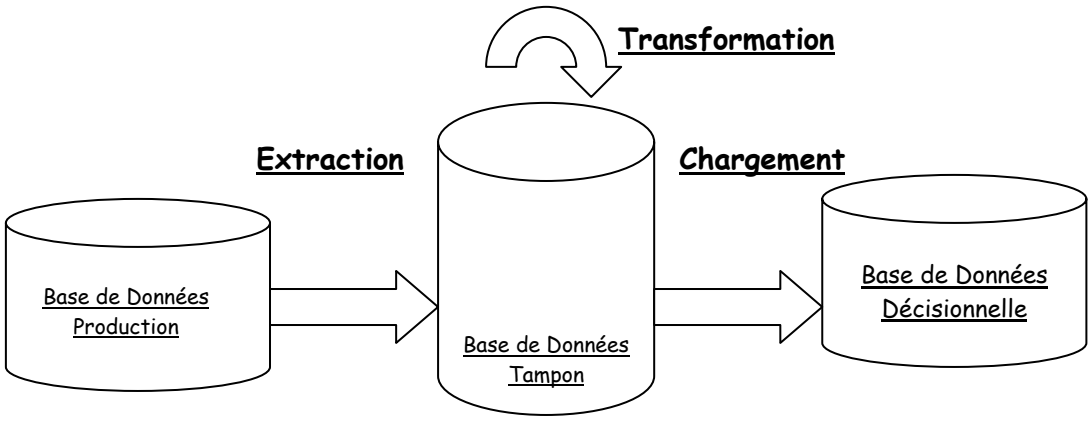


Figure N°35 : Les étapes du processus ETL

Base tampon :

Les données extraites doivent atterrir sur une autre base, appelée base tampon (staging area). Une fois l'étape d'extraction terminée, les transformations nécessaires peuvent être effectuées tranquillement dans la base tampon.

4.4.1- Etapes ETL

Détaillons maintenant les étapes de la figure 35.

a- Extraction

Il est bon que chaque table concernée par l'extraction (clients, produits, etc.) soit munie d'une colonne pour la date de création et une autre pour la date de dernière modification. Avec ces colonnes, l'extraction peut être incrémentale.

b- Transformation

A chaque table de la base décisionnelle correspond une table tampon qui contient :

– Les colonnes de la table de dimension ou de faits correspondante.

– Les clés naturelles et les clés de substitution.

– **U**ne colonne *Valid* de type oui ou non qui dira si le membre existe déjà ou non.

→ Réparation, complétion, synchronisation et formatage

Pendant que les données sont insérées dans les tables tampon, on peut les uniformiser, c'est-à-dire les réparer, les compléter, les synchroniser et les formater.

Exemple de réparation des données : les codes postaux invalides peuvent être corrigés en utilisant un annuaire des codes postaux.

Exemple de complétion des données : déduire la région où est domicilié un propriétaire à partir du numéro d'immatriculation de son véhicule.

Il faut uniformiser les formats avant le chargement, c'est le formatage.

→ Substitution des clés primaires

Une fois que les tables tampons sont remplies, on s'occuper de l'intégrité des données qui vont être chargées dans la base décisionnelle.

Rappelons que la base décisionnelle n'utilise pas les clés naturelles des bases de production car :

– Un produit peut être identifié dans deux bases de production différentes avec des clés distinctes.

– Un numéro de produit peut correspondre à deux produits distincts dans deux bases de production différentes.

Au contraire, pour identifier les membres de manière unique, la base décisionnelle utilise des clés de substitution.

Au cours de l'étape de transformation, il faut donc traduire les clés naturelles en clés de substitution et remplir la colonne *Valid*.

→ Substitution des clés étrangères

Il reste encore à traiter l'intégrité référentielle des données qui vont être chargées dans la base décisionnelle. Pour cela, il faut recalculer les clés

étrangères avec les clés de substitution afin que les relations de la base décisionnelle soient vérifiées lors du chargement.

Remarque : la phase de substitution est plus simple pour un schéma en étoile que pour un schéma en flocon.

c- **C**hargement

Comme les données sont chargées dans la base décisionnelle qui est muni d'un schéma relationnel, il faut charger ses tables dans cet ordre :

– D'abord les tables qui ne contiennent aucune clé étrangère.

– ensuite les tables qui ne contiennent que des clés étrangères vers des tables déjà chargées.

– etc.

Ensuite, pour chaque table, le chargement de décompose en deux requêtes :

– une pour les nouveaux membres ou faits.

– et un pour les membres ou faits modifiés.

V.5.ETUDE DE CAS : *Filiale les Moulins du Hodna M'SILA*

Le passage à l'économie de marche, a crée un environnement concurrentiel très rude engendrant une perte de la part de marche initiale détenue par la filiale.

L'objectif principal de la filiale est double :

- Récupérer une partie de sa part de marche.
- Analyser la situation des créances.

Dans cette optique les décideurs de cette filiale ont optés pour la conception d'un entrepôt de données des activités Commerciales et le suivi des créances,

Pour cela nous possédons un schéma E-R opérationnel de l'activité Commerciale.

Nous appliquons l'approche proposée afin d'exploiter le schéma initial E-R pour aboutir à un schéma multidimensionnel.

Supposons que :

- Le dossier d'application opérationnelle de l'activité commerciale est disponible.
- Le schéma E-R initial qui décrit l'information à la source est disponible.

V.5.1. Définition et analyse des besoins Décisionnels

L'analyse et la définition des besoins décisionnels avec les décideurs et les experts ont fait ressortir :

a- Les questions standard multidimensionnelles :

- Quel est le chiffre d'affaire réaliser par type de client, par type de produit, et par magasin, par région.
- Quel est la quantité vendue par type de client, par type de produit, et par magasin, par région.
- Quelle est la créance contractée par type de client, par région, par magasin.

b- Tableau des spécifications des besoins :

La construction du tableau des données décisionnelles utilise les questions décisionnelles et le schéma E-R initial, afin de faire un tri des données décisionnelles nécessaires à la construction de l'entrepôt de données.

Entité Article :

Attribut	Description	M	D	O
CodeArt	Code Article		X	
DesigArt	Désignation d'article		X	
Colisage	Colisage de l'article		X	

Entité Client :

Attribut	Description	M	D	O
CodeClt	Code Client		X	
NomClt	Nom Client		X	
PreClt	Prénom Client		X	
AdrClt	Adresse Client		X	
NumFax	N°Fax Client			X
NRC	N° Registre de Commerce			X
NimFisc	N° d'imposition Fiscal			X
Sexe	Sexe du Client			X
Age	Age du Client			X

Entité ClasseArticle :

Attribut	Description	M	D	O
ClasseArt	Classe d'article		X	
NomClasseArt	Nom de classe d'article		X	

Entité ClasseClient :

Attribut	Description	M	D	O
ClasseClt	Classe Client		X	
NomClasseClt	Nom de classe Client		X	

Entité Ville :

Attribut	Description	M	D	O
CodeVille	Code Ville		X	
NomVille	Nom de Ville		X	

Entité TypeCréance :

Attribut	Description	M	D	O
CodeTypeCre	Type Créance		X	
NomTypeCre	Nom Type Créance		X	

Entité Magasin :

Attribut	Description	M	D	O
CodeMag	Code Magasin		X	
NomMag	Nom de Magasin		X	

Entité ArticleVendu :

Attribut	Description	M	D	O
NumFact	N° Facture		X	
CodeArt	Code Article		X	
Qte	Quantité	X		
Pu_Art	Prix de Vente	X		
Tva	Taxe TVA	X		

Entité Ventes :

Attribut	Description	M	D	O
NumFact	N° Facture		X	
CodeArt	Code Article		X	
MntHt	Montant Hors Taxe de la Facture	X		
MntTransp	Montant de transport	X		
MntTva	Montant de la Taxe TVA	X		
NumChq	Numéro cheque			

Entité Créance:

Attribut	Description	M	D	O
NumCre	N° Créance		X	
MontCre	Montant De la Créance	X		
LimitCre	Limite de la Créance	X		
NumCont	N° Contrat		X	

V.5.2. Modélisation dimensionnelle des données

Soit le Schéma E-R initial de l'activité Commerciale suivant :

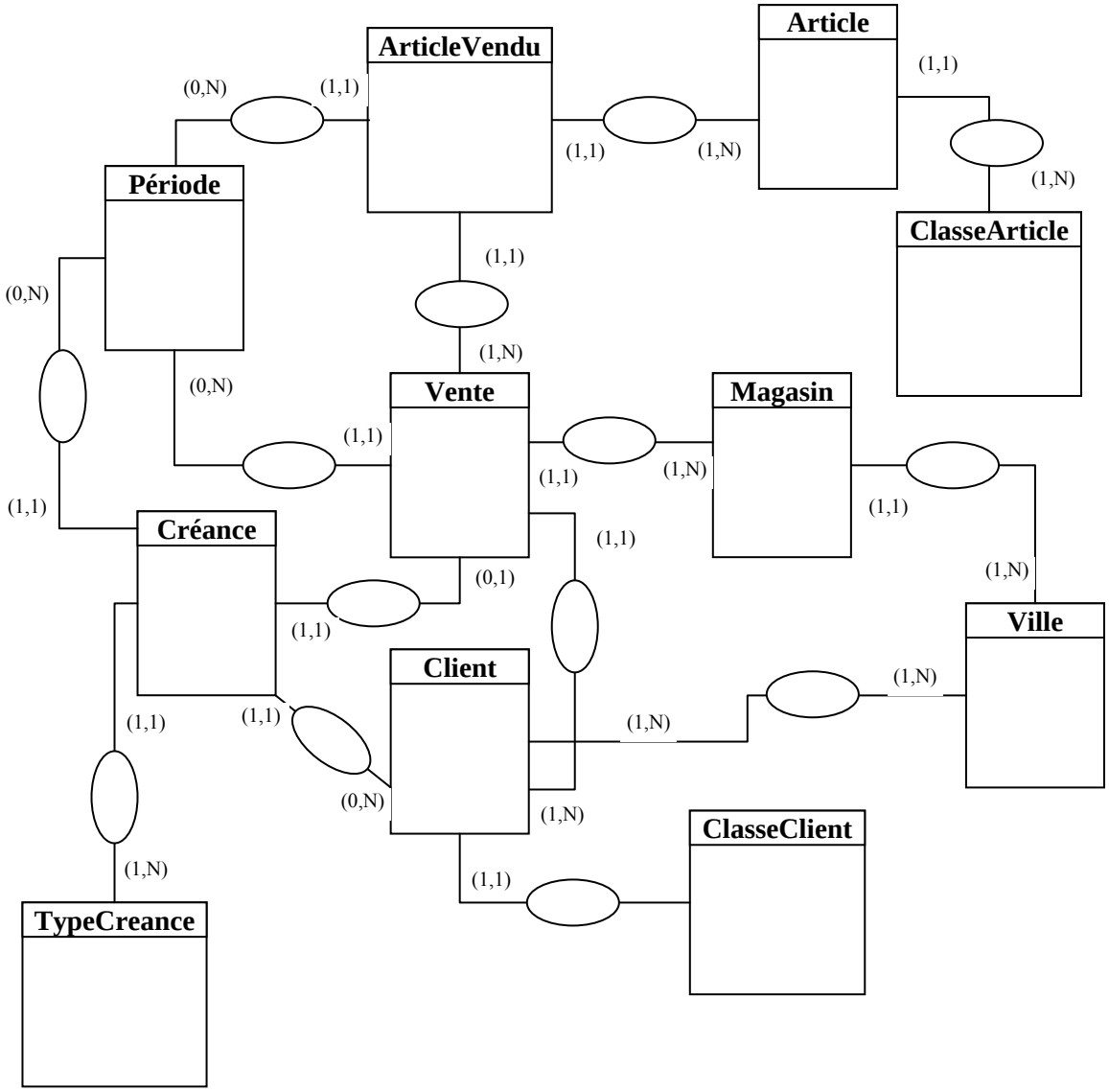


Figure N°36: Schéma E-R conceptuel initial

- **Identifier les activités à modéliser :**

L'analyse et la spécification des besoins fait ressortir les deux activités suivantes :

1- Activité Commerciale.

2- Activité Recouvrement des créances.

- **Identifier les sous schémas conceptuels initiales (opérationnel) :**

On peut recenser les sous MCD en rapport avec l'activité à modéliser, afin de pouvoir classifier les entités a partir Schéma E-R initial, on a les deux sous schémas suivants :

- **Activité Créance :**

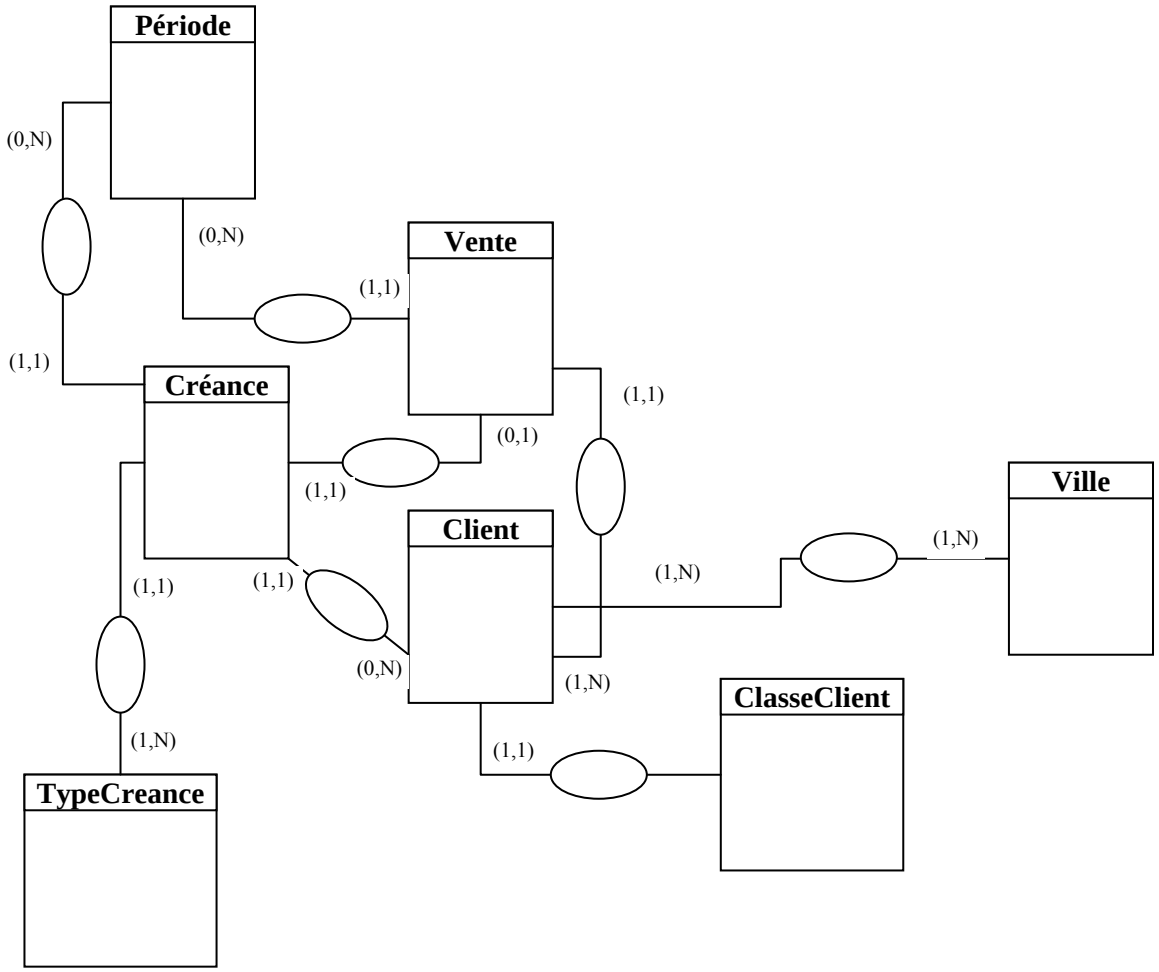


Figure N°37: Sous Schéma E-R conceptuel de l'activité Créance

- **Activité Commerciale :**

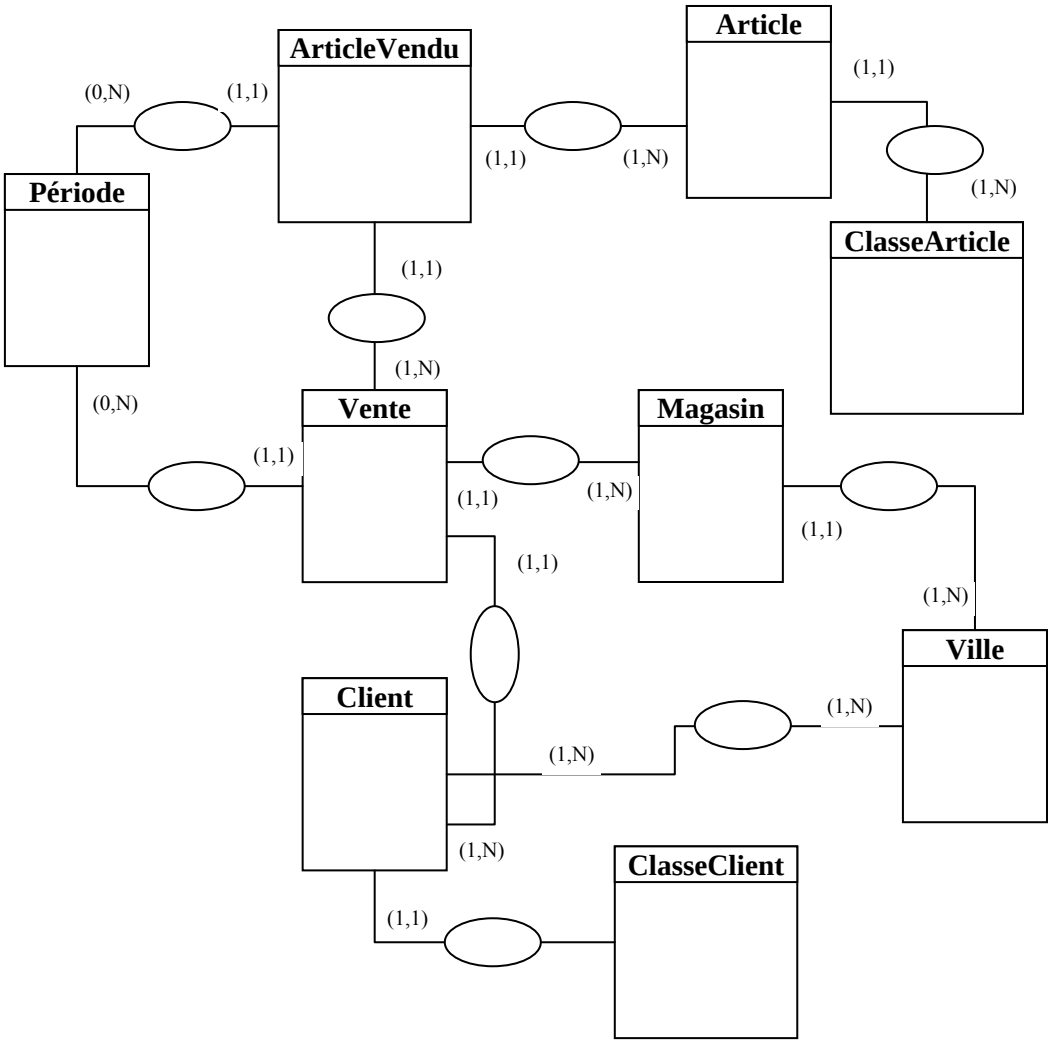


Figure N°38: Sous-Schéma E-R conceptuel de l'activité Commerciale

- **Classification des entités, et détermination du grain**.

La classification des sous schémas obtenus doit faire apparaître les entités composantes, transactionnelles et de classification,

Activité : Commerciale	Entité Transactionnelles	Entité Composantes	Entité Classification
	Vente ArticleVendu	Client Article Magasin Période	ClasseArticle ClasseClient Ville

Activité : Créance	**Entité Transactionnelles**	**Entité Composantes**	**Entité Classification**
	Créance	Client Période ClasseCréance	ClasseClient Ville

- **Identification des hiérarchies**.

On déterminer toutes les hiérarchies maximales ou autres:

- **Activité Commerciale :**

1- ArticleVendu(entité minimale)—Article--ClasseArticle (entité maximale).

2- ArticleVendu(entité minimale)--Ventes—Magasin--Ville(entité maximale).

3- ArticleVendu(entité minimale)--Ventes—Client-- ClasseClient(entité maximale).

4- ArticleVendu(entité minimale)--Ventes—Client-- Ville(entité maximale).

5- ArticleVendu(entité minimale)--Ventes--Période(entité maximale).

6- Ventes—Magasin--Ville(entité maximale).

7- Ventes—Client-- ClasseClient(entité maximale).

8- Ventes--Période(entité maximale).

9- Ventes—Client-- Ville(entité maximale).

- **Activité Créance :**

1- Créance(entité minimale)--Période(entité maximale).

2- Créance(entité minimale)-- Client-- ClasseClient(entité maximale).

3- Créance(entité minimale)-- Client-- Ville(entité maximale).

4- Créance(entité minimale)--ClasseCréance(entité maximale).

- **Production du Modèle dimensionnel :**

Il y a un large choix d'options pour la production des Modèles multidimensionnels, Chacune de ces différentes options représente le compromis entre la complexité et la redondance.

Rappel des Règles de passage pour le schéma en étoile :

- Une table de fait est formée pour chaque entité de transaction. La clef de la table est la combinaison des clefs de ses entités associées composantes.
- Une table de dimension est formée pour chaque entité de composante, par réduction de la hiérarchie liée aux entités de classification dans cette entité composante.
- Quand des relations hiérarchiques existent entre des entités de transaction, l'entité d'enfant hérite toutes les dimensions (et clefs d'attributs) de l'entité parentale, cela fournit la capacité de " Drill Down " entre les niveaux de transaction
- Attributs numériques dans entités de transaction doit être agrégé par des attributs clefs (des dimensions).

→ Nous optons pour un schéma en constellation pour l'activité **Commerciale**, qui consiste en un ensemble de schémas d'étoile avec deux tables de faits hiérarchiquement liées (***Vente*** et **ArticleVendu**), qui fournissent la Capacité de " Drill Down " (forer en bas) entre les niveaux de détail de la table de fait ***Vente*** vers la table de faits **ArticleVendu**.

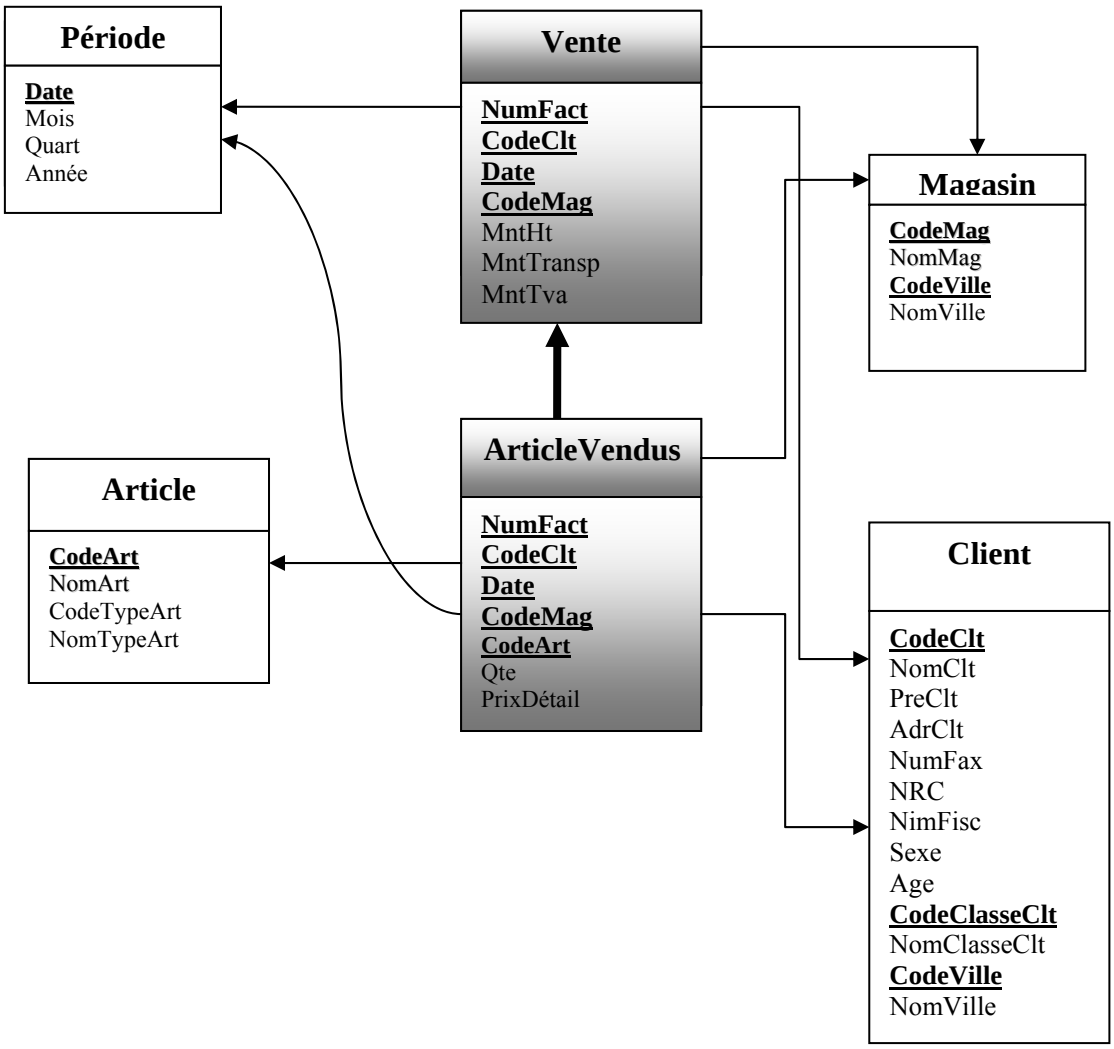


Figure N°39 : Le schéma en Constellation pour l'activité Commerciale

→ Nous optons pour un schéma en étoile pour l'activité **Créance**, ce choix est motivé par :

- L'existence d'une seule entité transactionnelle, donc une seule table de faits
- La taille des tables des dimensions (Clients, ClasseCréance, Période) sont très négligeable par rapport à la table de faits.

Donc on peut dénormaliser notre schéma multidimensionnel pour l'activité **Créance**.

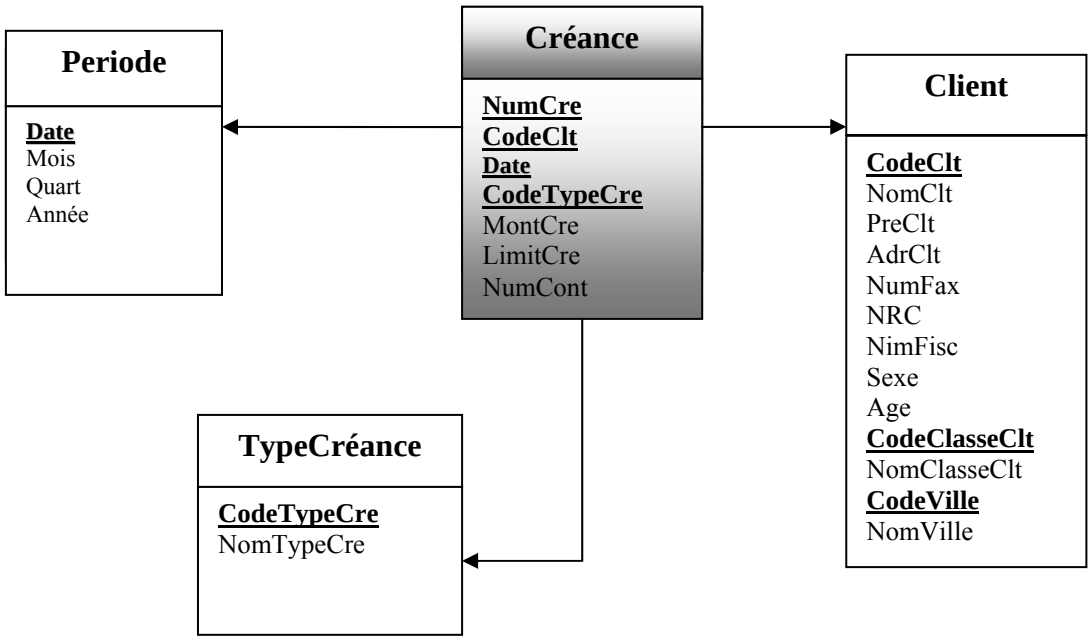


Figure N°40 : Le schéma en étoile pour l'activité Créance

- **Définition des contraintes d'agrégations (Summarizabilty)** :

L'étape suivante doit définir des niveaux de restriction pour toutes les mesures le long des différents chemins d'agrégation dans chaque schéma de fait :

Niveau de Restriction	Fonction d'agrégation Applicable
1	{SUM, AVG, MIN, MAX, STDDEV, VAR, COUNT}
2	{AVG, MIN, MAX, STDDEV, VAR, COUNT}
3	{COUNT}
4	{ }

Table 8 : Classification des niveaux de restriction [HLG00]

Schéma de Fait	Mesures	Niveaux de dimension	Niveau de Restriction
Fait: Créance	MontCre	Client Période ClasseCréance	1 1 1
	LimiteCredit	Client Période ClasseCréance	1 2 1

Table 9: Contraintes d'agrégation (Summarizabilty) pour le Schéma de Fait: Créance

Schéma de Fait	Mesures	Niveaux de dimension	Niveau de Restriction
Fait: Ventes	MntHt	Client Période Magasin	1 1 1
	MntTransp	Client Période Magasin	1 1 1
	MntTva	Client Période Magasin	1 1 1

Table 10 :Contraintes d'agrégation (Summarizabilty) pour le Schéma de Fait: Vente

Schéma de Fait	Mesures	Feuille de dimension	Niveau de Restriction
Fait: ArticleVentdu	Qte	Client Article Magasin Période Vente	1 1 1 1 2
	PrixDétail	Client Article Magasin Période Vente	2 2 2 2 2

Table11 : Contraintes d'agrégation (Summarizabilty) pour le Schéma de Fait: ArticleVendu

CONCLUSION

Dans ce mémoire, nous avons présenté une approche de conception de schéma conceptuel multidimensionnel à partir du schéma E-R opérationnel, inspiré essentiellement de l'approche de Daniel L. **Moddy** et Mark A.R. **Kortink** comme approche de base, ainsi que les travaux de B. **Husemman** et al, sans oublier la démarche de R. **Kimball.** L'objectif principal de cette approche est d'exploiter le schéma E-R initial opérationnel et déduire les faits, les Dimensions en utilisant la méthode de classification des entités du schéma E-R initial en trois classes (Transactionnelle, composante, classification), et puis déterminer les différentes hiérarchies existantes, et en première définir les spécifications des besoins sous forme d'une série des questions ou requêtes pour les futures analyses OLAP, et un tableau de spécification obtenu à partir de l'analyse du schéma E-R initial et les requêtes des décideurs pour classifier les attributs (Mesure, Dimensionnel, Optionnel), et enfin produire le modèle multidimensionnel où nous avons un large choix d'options pour la réalisation de ce modèle. Chacune de ces différentes options (Etoile, Flocon de neige, Galaxie, Plat, etc.) représente le compromis entre la complexité et la redondance, et obéit à des règles de passage du modèle E-R d'entreprise vers le modèle multidimensionnel, et nous avons défini des niveaux de restriction pour toutes les mesures le long des différents chemins d'agrégation dans chaque schéma multidimensionnel de fait.

Perspectives:

- Comment faire le passage des spécifications des besoins décisionnels vers l'identification des FAITS et MESURES et DIMENSIONS, Hiérarchies par l'utilisation d'un formalisme formelle ou semi formelle.

- Fusionner les projets de conception et réalisation des systèmes d'information de production (Opérationnel) avec les systèmes décisionnels, pour une meilleur intégration et une vision plus globale.
- Définir des fourchettes (au point de vue Taille du Data Warehouse) pour décider à quelle niveau il faut normaliser ou dénormaliser.
- Valider les modèles multidimensionnels obtenus par la méthode de Daniel L. **Moddy** et Mark A.R. **Kortink** en utilisant les dépendances fonctionnelles de B. **Husemman.**

Bibliographie

[COD93]	Codd E.F "Providing OLAP to user-analysts : an IT mandate" Technical Report, E.F Codd and Associates, 1993.
[AGR97]	Agrawal R., A. Gupta, S. Sarawagi, "Modeling multidimensional databases," Proc. 13th ICDE 1997, 232–243.
[CT98]	Cabibbo L., R. Torlone, "A logical approach to multidimensional databases," Proc. 6th EDBT 1998, LNCS 1377, 183–197.
[EVO 99]	http://www.prism.uvsq.fr/dataware/coop/evolution.html
[LAW98]	W. Lehner, J. Albrecht, H. Wedekind, "Normal forms for multidimensional databases," Proc. 10th SSDBM 1998, 63–72.
[HLG00]	*B. H¨usemann, J. Lechtenb¨orger, G. Vossen*, "Coceptual Data Warehouse Design" Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000) Stockholm, Sweden, June 5-6, 2000
[KOR00]	Daniel L. Moody & Mark A.R. Kortink « From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design» Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000) Stockholm, Sweden, June 5-6, 2000.
[CNAM98]	Nakache, Didier, Caulier Donneger, Anne, Rivelois Dugresson, Pascale, Dassonville, Philippe et Delebecq, Jean-Louis. "Data warehouse et data mining." C.N.A.M. de Lille, 1998 .
[HES03]	Informaticien de gestion / HES Option d'école/ Base de données informatique décisionnelle. Haute école spécialisée de suisse occidentale. 2003
[GOL98]	Matteo Golfarelli, Dario Maio et Stefano Rizzi « Conceptual design of data warehouses from E/R » *Proceedings ACM First International Workshop on Data Warehousing and OLAP (DOLAP 98),Kona, Hawaii, 1998.*
[TEST00]	Olivier TEST, Thèse Doctorat de l'université Paul Sabatier TOULOUSE, "Modélisation et Manipulation d'entrepôts de Données complexes et historisees", 2000
[KIM97]	Ralph Kimball, "entrepôt des données: Guide pratique du concepteur de Data Warehouse", International Thomson Publishing, France, Paris, 1997
[CHA94]	[CHA 94] CHAWATHE S., GARCIA-MOLINA H., HAMMER J., IRELAND K., PAPAKONSTANTINOU Y., ULLMAN J., WIDOM J., "The TSIMMIS project : Integration of heterogeneous information systems", *Proceedings of IPSI Conference*, Tokyo (Japan), 1994.
[BEL03]	"Techniques d'optimisation des requêtes dans les data warehouse" Laboratoire d'Informatique Scientifique et Industrielle, 2003,
[MAR98]	Marcel P. "Manipulation de Données Multidimensionnelles et Langages de Règles", Thèse de Doctorat de l'institut des Sciences Appliquées de Lyon, 1998.

[KIM02]	Kimball R & Margy Ross "The Data Warehouse Toolkit Second Edition The Complete Guide To Dimensional Modeling ", Wiley Computer Publishing, 2002.
[INM02]	W.H Inmon "Building The Data Warehouse Third Edition ", Wiley Computer Publishing, 2002.
[IGG03]	Claudia Imhoff, Nicholas Galemmo, Jonathan G. Geiger "Mastering Data Warehouse Design, Relational ans De,ensional Thechique ", Wiley Computer Publishing, 2003.
[CHA97]	Chaudhuri S. et Dayal U. "An overview of data warehouseing and olap technology" Sigmod Record, 26 (1): 65-74, Mars 1997.
[MEN97]	Mendelzon A. "Olap: Concepts and products." Talk at University of Toronto, 1997
[BOL02]	Nathalie RYSER BOLOGNINI « Etude pour la création d'un entrepôt de données dans le cadre de l'assurance vie et transformation des données en informations utiles en vue d'une prise de décision », en vue de l'obtention du Diplôme post grade en informatique et organisation, Université de Lausanne école des hautes études commerciales Années académiques 2000-2002 Sous la direction du Prof. Thibault ESTIER
[HES03]	Informaticien de gestion / HES Option d'école/ Base de données informatique décisionnelle. Haute école spécialisée de Suisse occidentale. 2003
[GAR02]	George Gardarin: « Explosion de l'informatique décisionnelle »
[VAN01]	Jean Vanderdonckt et Stéphane Faulkner, Chapitre 3 : Présentation des données dans les systèmes d'information opérationnels et décisionnels, tiré de : « Environnement évolué et évolution de l'IHM »,
[JAM99]	Michel Jambu « Introduction au Data Mining Analyse intelligente des données », Eyrolles, 1999.
[ASK03]	Mlle Askratin O. Thèse de Magistere en Informatique "Une Methode de Conception Des Data Webhouses", 2003
[TRU02]	Trujillo J. et Lujan-Mora S. et Song I. "Extending UML for multidimensial Modeling in proceding language' (UML'02), 2002 Springer Verlag
[AKO01]	AKOKA J., Comyn-Wattiau I. et Prat N. "Dimension Hierarchies Design from UML Generalizations and aggregations', 2001 Springer Verlag
[FRA00]	Franco J.M. Sandrine de Lignerolles 'Piloter l'entreprise grâce au data warehouse' Eyrolles 2000
[KIM00]	Ralph Kimball, "Concevoir et déployer un Data Warehouse", Edition Eyrolles, 2000
[PROB01]	Probatoire, 2000-2001
[TCC99]	« Introduction to Data Mining and Knowledge Discovery»Third Edition, by Two Crows Corporation, 1999.
[ERW01]	Alliaume Erwan et Tetzlaff Franck «The Data mining», 26/03/2001.

[ELH04]	Georges El Helou et Charbel Abou khalil « Data Mining : Techniques d'extraction des connaissances » Professeur : Mélissa Saadoun soutenu le 16 février 2004 Paris II.
[COD70]	CODD, E.F. (1970) A Relational Model of Data for Large Shared Data Banks, *Communications of the ACM*, 13 (6), June: 377-387.

Liste des Figures