

N° d'ordre :01/2007-M/IN

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE
SCIENTIFIQUE
UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE
« HOUARI BOUMEDIENE »
FACULTE D'ELECTRONIQUE ET D'INFORMATIQUE



Mémoire

Présenté pour l'obtention du diplôme de **Magister**

En : INFORMATIQUE

Spécialité : Bases de Données Avancées et Intelligence Artificielle

Par AMAROUCHE Idir Amine

Sujet

**Démarche de conception ascendante
d'un entrepôt de données distribué
dans un environnement inter-organisationnel.**

Soutenu le 31 /01 /2007, devant le jury composé de :

Mr. AHMED Nacer mohamed	Professeur, U.S.T.H.B	Président
Mme. ALIMAZIGHI Zaia	Maître de Conférence, U.S.T.H.B.	Directeur de thèse
Mr. ZEGGOUR Djamel Eddine	Professeur, I.N.I	Examineur
Mr. LARABI Slimane	Maître de Conférence, U.S.T.H.B	Examineur
Melle. BOUKHEDOUMA Saida	Chargé de cours, U.S.T.H.B	Invité

Résumé :

Les systèmes d'information inter-organisationnels (SIIO) sont de plus en plus présents dans les organisations car la façon de gérer les entreprises change. En outre, certaines activités d'un processus d'affaires d'une organisation dépendent d'une activité effectuée à l'extérieur de ses frontières. Dès lors, ces entités fournissent des efforts considérables afin d'exploiter le patrimoine informationnel interne et externe à l'entreprise et ce, dans le but de rendre leur processus de prise de décision plus efficient par la réalisation d'un système d'aide à la prise de décision dirigé par les données et basé sur le concept des entrepôts de données qui :

font l'objet d'une attention particulière de la part des utilisateurs de systèmes d'information décisionnels,

et concourent à une meilleure capitalisation du patrimoine informationnel des entreprises et à une production effective des connaissances nécessaires à une prise de décision.

L'étude de stratégies d'implémentation des entrepôts de données dans un environnement distribué, nous a conduit à constater que plusieurs architectures et méthodes d'élaboration des entrepôts de données sont étayées par des démarches plus ou moins formalisées, suivant des principes ascendant ou descendant, en fonction des environnements dans lesquels se place le chercheur.

Néanmoins, le souci d'intégration des structures analytiques à évolution indépendante appartenant à des organisations autonomes formant un système d'information inter-organisationnel, n'est pris en charge qu'à travers quelques travaux ponctuels non détaillés et non formalisés.

Pour cela, notre travail s'inscrit dans ce contexte et vise la mise en oeuvre d'une démarche de conception ascendante d'un entrepôt de données distribué dans un environnement inter-organisationnel et aura pour thèmes majeurs :

les systèmes d'information inter-organisationnels (SIIO);

L'environnement des entrepôts de données : les concepts de base, la modélisation multidimensionnelle, les méthodologies d'élaboration et les architectures en général et celles qui sont distribuées en particulier;

- La proposition d'une démarche de conception ascendante des entrepôts de données distribués dans un SIIO basée sur la définition des objectifs stratégiques et modélisée à l'aide de la notation UML associé au Processus Unifié sous-jacent.

Mots clés : Système d'information inter-organisationnel, Entrepôts de données distribués, Intégration des données multidimensionnelles, notation UML, UP.

Remerciements

J'exprime ma profonde reconnaissance et mes vifs remerciements à mon directeur de thèse Mme Z.Alimazighi pour m'avoir encadré, et accueilli au sein de son équipe et de m'avoir donné la possibilité de mener à bien les travaux.

Je remercie également Mr N.Selmoune et Mme S.Boukhadouma pour leur disponibilité, leurs lectures attentives de mes rapports et dont les critiques et suggestions étaient d'un grand apport pour mener à terme le mémoire.

Je remercie également les membres du jury : Pr. AHMED Nacer mohamed de m'avoir fait l'honneur de présider mon jury de thèse, Pr D.ZEGGOUR d'avoir accepté de juger ce travail, M.C. S. Larabi d'avoir accepté de faire partie de mon jury.

Je remercie particulièrement mon père pour ses relectures attentives du mémoire.

Je voudrai aussi remercier le Col. A.Djaffer et le Cdt M. Merbah ainsi que mes collègues de la division informatique de l'Hôpital Central de l'Armée, qui par leur disponibilité et leur soutien logistique ont contribué à l'aboutissement de ce travail.

Je remercie surtout, mes parents, mes frères et ma soeur qui mon ont soutenu pendant les moments les plus difficiles et particulièrement ma femme qui m'a amplement supporté.

Tables des matières

INTRODUCTION GENERALE	07
CHAPITRE 1 : Systèmes d'information inter-organisationnels.....	10
1.1. Introduction.....	10
1.2. Système d'information.....	10
1.3. Système d'information inter-organisationnel (SIIO).....	11
1.4. Configuration et typologie des SIIO.....	12
1.5. Technologies des SIIO.....	15
1.6. Synthèse.....	17
CHAPITRE 2 : Entrepôts de données : concepts et méthodologies.....	18
2.1. Introduction.....	18
2.2. Entrepôt de données, concepts et composants.....	18
2.3. La modélisation multidimensionnelle.....	19
2.4. Architectures des entrepôts de données.....	21
2.5. Méthodologies, approches d'analyse et stratégies de développement d'un ED.....	21
2.6. Architectures distribuées des entrepôts de données.....	29
2.7. Conclusion	31
CHAPITRE 3 : Intégration des entrepôts de données distribués et SIIO...	33
3.1. Introduction.....	33
3.2. Intégration, concepts et processus.....	33
3.3. Démarche conceptuelle d'intégration générique des ED distribués.....	35
3.4. Critères régissant la démarche d'intégration des ED distribués.	36
3.5. Intégration de schémas.....	42
3.6. Correspondance de schémas des ED.....	48
3.7. Synthèse.....	52
Chapitre 4 : Conception bottom up des ED distribués dans un environnement SIIO : Cycle de vie.....	55
4.1. Introduction.....	55
4.2. Démarche d'intégration des ED distribués dans un SIIO.....	56
4.3. Formalisation	66
4.4. Etude de cas	68
4.5. Conclusion.....	78
Conclusion générale	79
Bibliographie	81

Abréviation

CSCW	: Computer-Supported Cooperative Work
MD	: Magasin de Données (Data Mart)
ED	: Entrepôt de données (Data Warehouse)
EDI	: Electronique Data Interchange (Echange des Données Informatisées)
EIS	: Entreprise Information System (Système d'Information de l'Entreprise)
ERP	: Entreprise Ressource Planning (Progiciel de Gestion Intégrée)
ERM	: Entity Relationship Model (Modèle Entité Association)
ETL	: Extraction, Transformation and Load (Extraction, transformation et chargement)
HOLAP	: Hybrid OLAP
MOLAP	: Multidimensional OLAP
NTIC	: Nouvelles Technologie de l'information et de la communication
OLAP	: On line Analytical Processing
OLTP	: On line Transactional processing
ROLAP	: Relational OLAP (OLAP relationnel)
ROI	: Return of Investment (Retour sur Investissement)
SGBD	: Système de gestion des bases de données
SGDT	: Système de gestion des données techniques
SERM	: Structured Entity Relationship Model (Modèle Entité Association Structuré)
SCM	: Supply Chain Management (la gestion de la chaîne logistique)
SI	: Système d'information
SIIO	: Système d'Information Inter organisationnel
TPS	: Transaction Processing System (Système de traitement transactionnel)
XML	: Extended Markup Language
UML	: Unified Modelisation language (Langage de modélisation Unifié)
UP	: Unified Process (Processus Unifié)

Liste des figures

Figure 1.1	: Classification des SI selon le niveau organisationnel [IT02].....
Figure 1.2	: Lien vertical et horizontal entre organisations [EOM05].....
Figure 1.3	: Classification des SIIO selon Hong [EOM05].....
Figure 1.4	: Classification des technologies en support à chaîne de valeur [BA02].....
Figure 2.1	: Architecture type d'un entrepôt de données [SC02].....
Figure 2.2	: Exemple de schéma conceptuel en flocon de neige d'un ED/MD [LR04, LR05].....
Figure 2.3	: Stratégie d'implémentation Top-Down [TT98].....
Figure 2.4.A	: ED classique d'entreprise [JFJA00].....
Figure 2.4.B	: Système d'ED classique d'entreprise avec ODS [JFJA00].....
Figure 2.5	: Stratégie d'implémentation Bottom-Up [TT98].....
Figure 2.6	: Système de MD incrémental [DH98].....
Figure 2.7	: Système de legamarts [DH98].....
Figure 2.8	: Stratégie d'implémentation hybride [TT98].....
Figure 2.9	: Stratégie d'implémentation fédérée [PS04].....
Figure 2.10	: Entrepôt de données distribué d'après l'approche de Inmon [AK99].....
Figure 2.11	: Entrepôt de données distribué d'après l'approche de White [AK99].....
Figure 2.12	: Entrepôt de données distribué d'après l'approche de Noaman et Barker [AK99].....
Figure 2.13	: Entrepôt de données fédéré d'après l'approche de White [AK99, CW01].....
Figure 3.1	: Le model à 5 niveaux des multibases hétérogènes [MO00].....
Figure 3.2	: Les entrées et les sorties d'un processus d'intégration.....
Figure 3.3	: Les pré-requis d'une démarche d'intégration des entrepôts de données.....
Figure 3.4	: Stratégie d'intégration des ED/MD [CAVN05]
Figure 3.5	: Paradigme d'intégration [SRUH99].....
Figure 3.6	: Processus d'intégration des schémas multidimensionnel d'après [LRO5].....
Figure 3.7	: Application d'une expression de l'algèbre de dimension [LR05].....
Figure 3.8	: Application d'une autre expression de l'algèbre de dimension [LR05].....
Figure 3.9	: Matching entre deux dimensions [LR05].....
Figure 3.10	: ED/MD des ventes [LR04].....
Figure 3.11	: ED/MD de la situation météo [LR04].....
Figure 3.12	: Matching entre les dimensions temps (t1 et t 2) et localisation (S1 et SM2) [LR04].....
Figure 3.13	: Une vue virtuelle sur la portion commune entre les dimensions temps et localisation géographique [LR04].....
Figure 3.14	: Une vue matérialisée sur la portion unifiée entre les dimensions temps et localisation géographique [LR04].....
Figure 3.15	: Processus de correspondance des schémas d'après [TJ04 , ZH97].....
Figure 3.16	: Facteurs influençant le choix de l'approche d'intégration des ED dans un SIIO.....
Figure 4.1	: Architecture d'intégration des ED hétérogènes.....
Figure 4.2	: Processus générique d'intégration des ED à évolution indépendante.....
Figure 4.3	: Etude préalable.....
Figure 4.4	: Diagramme UML reliant les objectifs et les entités analytiques (dimensions et faits).....
Figure 4.5	: Etude élaboration.....
Figure 4.6	: Modélisation des relations sémantiques entre faits et dimensions appartenant à différents ED. [AJF02].....
Figure 4.7	: Correspondance entre les faits et les dimensions des ED autonomes.....
Figure 4.8	: Processus d'intégration des ED.....
Figure 4.9	: Architecture de UP [JBR00].....
Figure 4.10	: Cas d'utilisation métier.....
Figure 4.11	: Diagramme UML de hiérarchisation des objectifs.....
Figure 4.12	: Schéma dimensionnel descendant réconciliant les objectifs des organisations.....
Figure 4.13	: Schémas conceptuels des ED existants.....

Figure 4.14	: Architecture de référence.....
Figure 4.15	: Cas d'utilisation système.....
Figure 4.16	: Cas d'utilisation système [JF98].....
Figure 4.17	: Schéma dimensionnel descendant réconciliant les objectifs des organisations.....
Figure 4.18	: Mapping des schémas dimensionnels locaux et global.....

Liste des Tableaux

Tableau 1.1	: Classification des SIO selon Kumar van & Han [BA02].....
Tableau 1.2	: Etude comparative entre les SIO d'après Hong [EOM05].....
Tableau 1.3	: Technologie des SIO partagés et en interdépendance séquentielle.....
Tableau 1.4	: Technologie des SIO en réseau (Workflow et Gestion documentaire).....
Tableau 2.1	: Comparaison entre les méthodologies d'élaboration d'ED [WP00].....
Tableau 2.2	: Stratégie d'implémentation, architecture et approche d'analyse des méthodologies d'élaboration des entrepôts de données.....
Tableau 3.1	: Synthèse liant les SIO au processus d'intégration.....
Tableau 4.1	: Analyse des schémas physiques des structures analytiques existantes.....
Tableau 4.2	: Matrice d'analyse des entités analytiques redondantes.....
Tableau 4.3	: Correspondance entre fait et dimension appartenant à deux schémas dimensionnels [AJF02].
Tableau 4.4	: Processus UP d'intégration des ED à évolution indépendante [IT02].....

Introduction générale

Le contexte économique actuel, caractérisé par une concurrence mondialisée, voit se multiplier l'émergence de regroupements, conglomérats ou partenariats entre des entreprises de types bien distincts afin d'apporter une réponse globale aux exigences du marché. De ce fait, les entreprises sont amenées à inter relier leurs processus de production, d'information, de décision, pour mettre en oeuvre des réseaux de processus inter-organisationnels distincts.

Dans ce contexte, les technologies de l'information permettent de créer des plates-formes électroniques pour améliorer la coordination au sein d'un réseau d'entreprises, et envisager de nouvelles relations inter-entreprises à travers de réseaux informatiques.

A cet effet, les systèmes d'information inter-organisationnels (SIIO) sont apparus pour supporter les relations inter-entreprises, assister le travail coopératif, améliorer les échanges et le partage d'informations et de connaissances entre des partenaires soigneusement sélectionnés pour des relations durables.

Cette ouverture du système d'information de l'entreprise nous commande de repenser la façon dont seront intégrés ou interopérables les entrepôts de données distribués, hétérogènes et autonomes, existants au niveau des organisations formant le SIIO, et réfléchir quant à la démarche à adopter pour y parvenir. En effet, la conception de solutions relatives à l'utilisation partagée des données décisionnelles est un des défis majeurs pour les systèmes d'information décisionnels. Ainsi, la nécessité de définir une démarche permettant de rendre les entrepôts de données dans des environnements inter-organisationnels interopérables et intégrés constitue le thème que nous nous proposons d'étudier.

Une revue globale de la littérature scientifique concernant les entrepôts de données nous apprendra que plusieurs travaux et guides d'élaboration et de conception des entrepôts de données ont été réalisés. Cela a suscité un grand intérêt pour les aspects d'ingénierie des ED, de la modélisation conceptuelle, logique et physique.

Toutefois, l'aspect intégration des ED à évolution indépendante existants dans des SIIO n'a pas bénéficié d'une attention particulière, du fait que la majorité des travaux élaborés pour l'exploration de cette problématique insistent sur l'aspect conceptuel qui préconise une uniformisation des entités analytiques communes ou comment faire correspondre ces dernière mais dans un cadre ponctuel tel que l'intégration des ED appartenant à un environnement de gestion de la chaîne de soutien (SCM) ou un environnement du E-Commerce sans inscrire ces solutions dans une démarche d'intégration globale.

De plus, la plupart des travaux porte uniquement sur l'intégration des ED au sein de la même organisation entraînant le moins de complexité, mais dès que le problème s'ouvre sur des environnements inter-organisationnels, c'est un axe de recherche peu exploré auquel nous sommes confrontés, car certains facteurs inhérents à la relation existante entre la nature organisationnelle des entreprises et la topologie des entrepôts de données (ED) sont nécessaires pour aider les concepteurs et les dirigeants dans le développement d'un ED de qualité quant bien même ils demeurent encore non formalisés.

Cette problématique peut être analysée selon plusieurs dimensions, que sont la distribution, l'autonomie et l'hétérogénéité des entrepôts de données, la typologie des environnements inter-organisationnels et le paradigme de la démarche en vue d'aboutir à un mécanisme d'accès aux structures analytiques évolutives, dynamiques, autonomes et hétérogènes d'une façon sécurisée, transparente, homogène, virtuelle ou matérialisée en faisant usage d'un langage de requêtes ou d'un schéma global.

C'est dans cette perspective que s'inscrit ce modeste travail qui consiste à proposer une *démarche de conception bottom up d'un entrepôt de données distribué dans un environnement inter-organisationnel*. Pour ce faire, il se fondera sur le concept des systèmes d'information inter-organisationnels (SIIO) et sur une description de l'état de l'art des concepts, méthodologies d'élaboration et architectures des entrepôts de données en général et spécialement les architectures distribuées reflétant le caractère distribué des nouvelles organisations. Sur la base de ces concepts, notre proposition portera sur ce qui suit :

- La mise en relief d'un cadre de référence embryonnaire alignant les caractéristiques des SIIO et la démarche de conception ascendante des entrepôts de données distribués.

- L'identification des phases et des activités majeures régissant la dite démarche,

Afin de mener à terme ce travail, notre rapport est structuré comme suit :

Chapitre 1 : Dans ce chapitre, nous allons exposer le concept du système d'information en général et les concepts relatifs aux SIIO, les caractéristiques des structures inter-organisationnelles, la taxonomie et la typologie des SIIO, en particulier, en vue de souligner la nécessité de mettre en exergue un dispositif permettant l'intégration des données pertinentes pour l'aide à la décision à travers « les entrepôts de données », au sein d'un environnement inter-organisationnel et identifier les configurations d'interopérabilité entre les ED appartenant aux différentes organisations.

Chapitre 2 : Un entrepôt de données est un moyen pour maintenir les données orientées sujet, intégrées, et stables pour le support de l'aide à la décision. A cet effet une revue de l'état de l'art concernant la définition des concepts, des architectures et des méthodologies d'élaboration des entrepôts de données caractérisées par leur approche d'analyse, par leur stratégie d'implémentation et par leur architecture seront succinctement évoquées. Par conséquent, on peut dresser une synthèse sur l'ingénierie des entrepôts de données en rappelant que les méthodologies d'élaboration des ED et les choix architecturaux sont deux aspects liés.

Chapitre 3 : Ce chapitre qui portera sur le fait que la mise en commun ou le partage des données décisionnelles existantes dans des entrepôts de données appartenant à des organisations formant un SIIO, doit être considéré comme une solution d'intégration de données analytiques. A cet égard, les concepts d'intégration des bases de données dimensionnelles et les stratégies d'intégration des structures analytiques telles que les entrepôts de données ou les magasins de données, doivent être explicités afin de dégager la relation existante entre la démarche ou la stratégie d'intégration des structures analytiques autonomes et la nature de l'environnement inter-organisationnel.

Chapitre 4 : Ce chapitre expose la proposition d'une démarche conceptuelle ascendante d'un ED distribué en faisant usage du formalisme objet d'UML associé au processus unifié sous-jacent pour sa spécification laquelle sera étayée par une étude de cas.

Cette démarche sera identifiée par ses phases et ses activités majeures et dont le but est de concevoir un entrepôt de données global permettant aux organisations du SIIO l'accès aux structures analytiques d'une façon transparente et homogène afin de réaliser les objectifs métiers des organisations formant le SIIO.

En conclusion, L'objectif de ce travail est d'aboutir à la proposition d'une démarche de conception bottom up des entrepôts de données distribués en spécifiant l'influence des environnements inter-organisationnels sur la stratégie d'intégration des entités analytiques autonomes, hétérogènes et distribuées et permettant, en outre, de spécifier les perspectives de recherches possibles à prendre en charge.

CHAPITRE 1 : Systèmes d'information inter-organisationnels

1.1. Introduction

Comme la littérature confère une certaine richesse à ce domaine nous allons tenter de synthétiser :

En premier lieu, les concepts de base relatifs aux systèmes d'information, et présenter une taxonomie des SI selon le niveau hiérarchique de l'organisation en vue de faire ressortir le fait que la façon dont les entreprises s'organisent actuellement sont à l'origine de l'émergence des SIIO.

En deuxième lieu, les concepts de base des SIIO, les caractéristiques des réseaux inter-organisationnels, leurs configurations et les typologies les plus répandues basées sur des caractéristiques de classification distinctives, en plus, une étude des différentes plates formes technologiques sera exposée et classée selon le critère d'interdépendance entre les entités de l'environnement inter organisationnel.

Enfin, les incidences des environnements inter-organisationnels sur l'interopérabilité des entrepôts de données à évolution indépendante existants dans les organisations formant le SIIO.

1.2. Système d'information

« Le SI est un système qui rend possible la mise en commun d'informations appartenant à des sources séparées » [MH98] et d'après [FY00], un SI est un ensemble structuré de données et de moyens.

Les systèmes d'information peuvent être classés selon le niveau hiérarchique de l'organisation [IT02]. Les systèmes d'information typiques basés sur la structure de l'organisation sont :

a. Système d'information fonctionnel (Départemental) : Les systèmes d'information fonctionnels sont organisés autour des traditionnels départements ou services d'une organisation : manufacturier (opérations/production), comptabilité, finance, commercialisation, et ressources humaines. Un système spécial qui traverse plusieurs départements est le système de traitement des transactions (« Transaction Processing System », TPS). Le TPS automatise les tâches routinières et répétitives, spécifiques aux processus métiers de l'organisation.

b. Système d'information d'entreprise : Pendant qu'un système d'information départemental est en rapport avec une région fonctionnelle, habituellement, les autres systèmes d'information servent plusieurs départements ou l'entreprise entière. Ces systèmes d'information avec les applications départementales constituent le système d'information de l'entreprise (« Entreprise Information System », EIS). Une des applications d'entreprise les plus courantes est le progiciel de gestion intégré (« Entreprise Ressource Planning », ERP) permettant aux compagnies d'organiser et de diriger les ressources d'une entreprise entière.

c. Système d'information inter-organisationnel : Quelques systèmes d'information connectent deux ou plusieurs organisations, ils sont connus sous le nom de systèmes d'information inter-organisationnels (SIIO), par exemple, le système de réservation des compagnies aériennes tel que

SABRE, composé de plusieurs systèmes qui appartiennent aux différentes compagnies. Les systèmes d'information inter-organisationnels jouent un rôle majeur dans le E-commerce et dans la gestion de la chaîne logistique (« Supply Chain Management», SCM).

Ces systèmes sont organisés sous forme d'une hiérarchie dont chaque système consiste en plusieurs sous systèmes du niveau inférieur, comme le montre la **figure 1.1**, un système départemental supporte les services et les départements dans chaque compagnie. À un plus haut niveau, le système de l'entreprise supporte la compagnie entière, et le système inter-organisationnel connecte les différentes organisations.

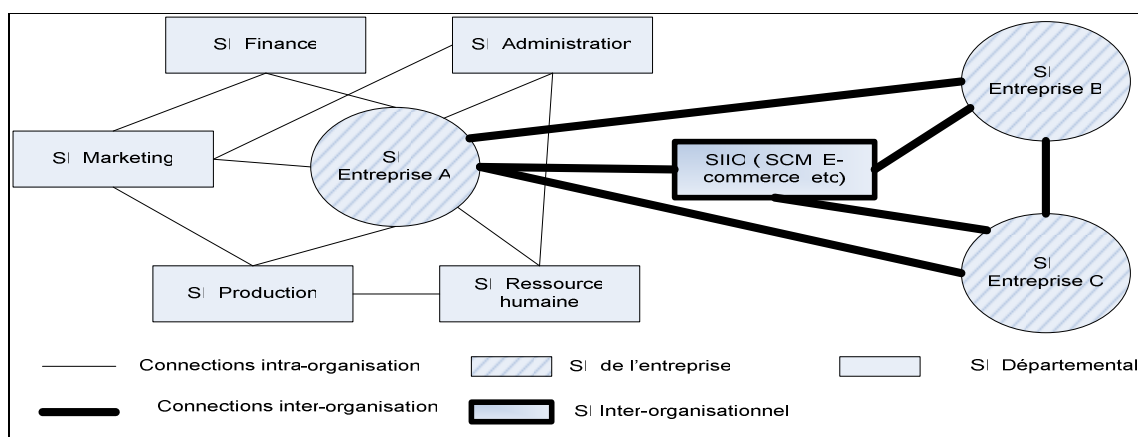


Figure 1.1 : Classification des SI selon le niveau organisationnel [IT02]

En effet, de plus en plus, certaines activités d'un processus d'affaires d'une organisation dépendent d'une activité exercée à l'extérieur de ses frontières, ce qui incite les organisations à adopter des structures plus légères, se concentrant sur les activités proches de leurs compétences de base et confiant à des partenaires le reste de leurs activités [BA02]. Ces entreprises ont dès lors un nombre nettement plus important d'échanges d'information avec leurs partenaires que les entreprises traditionnelles. Cette tendance a permis l'émergence de plusieurs concepts comme « **entreprise agile** », « **entreprise virtuelle** », « **entreprise étendue** », « **réseau d'entreprise** » et « **entreprise numérique** » qui sont supportés par l'utilisation intense de technologies d'information et de communication.

1.3. Système d'information inter-organisationnel (SIIO)

« Un SIIO a pour fonction particulière de supporter des processus qui traversent les frontières d'une organisation. Il peut ainsi supporter des processus étroitement couplés entre deux organisations » [BA02].

« Un SIIO est un système d'information et de gestion qui transcende les frontières d'une organisation via des liaisons électroniques avec ses partenaires commerciaux pour partager des données et des applications métiers, fournir des possibilités de transactions électroniques comprenant les achats et les ventes des marchandises et des services. Il facilite les communications et les prises de décision dans le but d'accroître le rendement, l'efficacité, la compétitivité et la rentabilité pour les différentes organisations. La liaison électronique est établie par les technologies de l'information ,

comprenant l'intranet, l'extranet, l'Internet, le Groupware, l'échange de données électroniques (EDI), le workflow, les technologies de la communication mobile et autres NTIC » [EOM05].

Les systèmes inter-organisationnels peuvent prendre plusieurs formes allant du SIO dédié ou propriétaire aux systèmes complètement ouverts basés sur internet. **La première famille des SIO** caractérisée par la prédétermination des partenaires qui se sont mis d'accord sur des procédures d'échanges électroniques des informations entre eux, communément connues sous le nom d'« **alliances électroniques** ». **La deuxième famille** de SIO est « **le marché électronique** », induite principalement par le développement de l'Internet permettant l'élaboration des relations professionnelles et commerciales entre des membres non prédéterminés [FH99].

1.4. Configuration et typologie des SIO

D'après [EOM05], les configurations des SIO se basent généralement sur :

1. le nombre des organisations interconnectées physiquement ;
2. le sens des flux de données circulant entre les différents systèmes d'information des entreprises qui définit l'interdépendance entre les organisations;
3. la nature similaire ou complémentaire du secteur d'activité des organisations formant le système inter-organisationnel, comme illustré dans la figure 1.2 , le lien entre les chaînes de valeur hétérogènes est *vertical* pour les entreprises qui cherchent à collaborer avec les différents partenaires afin d'améliorer leurs productions ou prestations de services. Par contre le lien entre les entreprises couvrant une même activité économique est *horizontal* car l'objectif est la coopération entre concurrents.

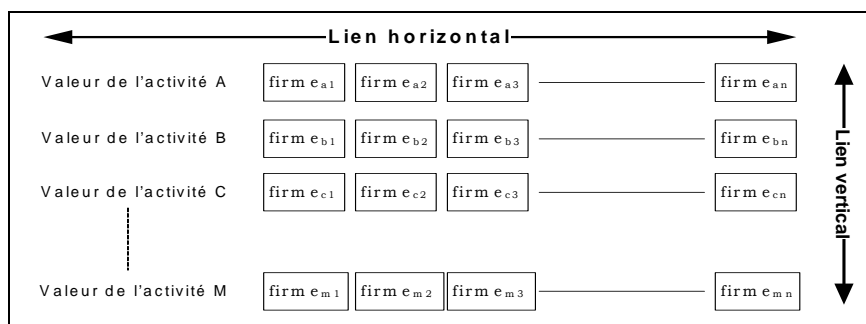


Figure 1.2. : Lien vertical et horizontal entre organisations [EOM05].

La littérature des SIO comprend plusieurs travaux de recherche se focalisant surtout sur l'aspect typologie et classification des environnements inter-organisationnels, ce qui explique la multitude des typologies existantes. Cependant parmi les typologies largement référencées on retrouve :

1.4.1. La classification de Kumar van & Han (1996)

Kumar van & Han (1996) ont dressé une taxonomie basée sur le type d'interdépendance existant entre les différentes organisations utilisant le système. Par "interdépendance", on cherche à

déterminer de quelle façon le travail d'une organisation est dépendant de celui d'une autre [PL03] [KV96] [EA03] [BA02]. Trois formes d'interdépendances et donc de SIIO sont identifiées. Ils sont présentés dans le tableau suivant :

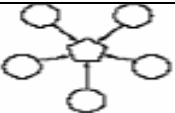
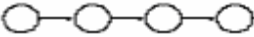
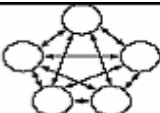
Type de SIIO	SIIO d'information partagée	SIIO de chaîne d'approvisionnement	SIIO réseau
Type d'interdépendance	Groupée	Séquentielle	Réciproque
Configuration			
Exemple de systèmes	<ul style="list-style-type: none"> • Partage d'un Extranet • Place de marché électronique 	<ul style="list-style-type: none"> • Systèmes transactionnels EDI • Systèmes transactionnels XML 	<ul style="list-style-type: none"> • Groupware

Tableau 1.1 : Classification des SIIO selon Kumar van & Han [BA02]

▪ **Interdépendance regroupée :**

La première forme d'interdépendance consiste en un partage collectif d'information (*Hub-and-spoke information sharing*) et sa forme la plus connue est la place de marché électronique ("e-marketplaces") [BA02, EOM05].

▪ **Interdépendance séquentielle :**

La deuxième forme d'interdépendance est séquentielle (*Sequential information sharing*) et correspond aux SIIO où les systèmes sont reliés de manière cohérente avec la séquence d'activités de la chaîne d'approvisionnement tels que l'EDI ou les systèmes de prise de commande [BA02, EOM05].

▪ **Interdépendance réciproque :**

Enfin, la troisième forme correspond à une interdépendance de type réseau et correspond aux SIIO destinés aux travaux collaboratifs (*Reciprocal information sharing*) tels que les *groupwares*. Ces formes utilisent une variété d'outils. Il ne s'agit plus ici d'une relation séquentielle mais d'une relation de type réseau entre les différents partenaires d'affaires. La relation se veut généralement un peu plus aléatoire, moins souvent prédéfinie et, par conséquent, plus difficile à modéliser. Ainsi les flux d'information entre les partenaires sont généralement beaucoup plus dynamiques [BA02, EOM05].

1.4.2. La classification de Hong (2002)

La contribution la plus remarquable, concernant la classification des SIIO, provient des travaux de Hong (2002) [AFG05], car à travers son modèle matriciel, il a permis l'identification de quatre types de SIIO basés sur une analyse bidimensionnelle, présentée comme suit :

1. Le type de la liaison que le SIIO supporte ou le rôle du lien (direction de la coopération) :

- **Horizontale :** Le SIIO connecte des groupes d'organisations homogènes donc des organisations exécutant des activités communes;
- **Verticale :** Le SIIO implique les différents rôles des organisations participant au SIIO le long de la chaîne de valeur, ex : une relation vendeur / acheteur;

2. Le niveau de support que le SIIO fournit ou la pertinence du SIIO:

- **Stratégique** : le SIO implique le lien des activités fondamentales des organisations, le partage d'informations et des risques, la collaboration pour étendre l'accès aux marchés, etc.
- **Opérationnel** : le SIO inclut le support des opérations routinières et quotidiennes du processus métier.

Sur la base du rôle de la liaison et le niveau de support de la coordination, les SIO, d'après Hong [EOM05] sont (Fig.1.3) :



Figure 1.3. : Classification des SIO selon Hong [EOM05].

SIO, Ressource groupée (Resource-Pooling IOS) : Les SIO Appartenant à cette catégorie lient les participants exécutant les mêmes activités de valeurs communes (c.-à-d., entreprises rivales) pour partager le risque ou le coût en regroupant leurs ressources. Ce type, en général, vise à former une coalition pour rivaliser avec les grandes entreprises ou pour l'extension des marchés au profit des parties prenantes de ce système. Les exemples illustrant les systèmes de cette catégorie sont : ACORD (IVANS¹ Insurance Value-Added Network Services), McKerson (ECONOMOST²).

SIO, Coopération complémentaire (Complementary Cooperation IOS) : Cette catégorie représente une forme de coopération entre des entreprises qui jouent des rôles différents dans une chaîne de valeur, elle leur permet d'étendre la capacité métier au-delà de la limite des ressources possédées par une seule entreprise en agrandissant ses ressources. Le but principal de cette catégorie est d'accéder au marché par des avantages complémentaires, l'exemple reflétant le mieux cette catégorie est le système de UAL & SAS³ et de Citibank⁴ ;

SIO, Coopération opérationnelle (Operational Cooperation IOS) : Ce type de SIO réunit des entreprises de chaîne de valeur commune pour améliorer la qualité de service ou pour partager des informations d'intérêt commun. Donc, en étant d'accord avec les rivaux pour collaborer sur quelques opérations communes, les entreprises peuvent créer une organisation virtuelle énorme qui leur permet

¹ **IVANS** : C'est un groupe de compagnies d'assurance autonomes avec des centaines d'agences.

² **McKerson** : McKesson Corporation, un distributeur de produits pharmaceutiques, médicaux et autres produits consommables, a construit ECONOMOST, un système de gestion des inventaires et des commandes, qui interconnecte les magasins autonomes commercialisant exclusivement les produits McKesson.

³ **UAL&SAS** Transporteurs aériens incluant UAL et SAS, ont tenté d'offrir un service de voyage complet comprenant les compagnies aériennes, les services de location de voitures et les prestations hôtelières.

⁴ **Citibank**: une institution financière, Citibank alliée avec American Airlines, les hôtels Marriott et la chaîne de supermarché national pour offrir la carte de paiement CityCard et utilise un réseau informatique pour la capture des données des transactions opérées par la CityCard via les points de service dans une base de données

d'opérer ensemble comme si elles étaient une seule compagnie, à titre indicatif le système de American Airlines (Sabre¹).

SIIO, Coordination opérationnelle (Operational Coordination IOS) : Un SIIO peut être configuré pour interconnecter différents rôles joués par des entreprises dans une chaîne de valeur de l'industrie pour augmenter l'efficacité opérationnelle, comme le cas du système de Nike Inc².

Tous les types de SIIO partagent une caractéristique commune qui est celle de l'interconnexion des processus métiers en permettant la diffusion des informations, néanmoins, quelques divergences subsistent quant aux objectifs, à la nature des relations entre les organisations, aux avantages visés et à la stratégie applicable au niveau des quatre SIIO suscités. Le tableau suivant met en exergue une comparaison effectuée par [HONG02] entre les SIIO :

Type de SIIO	Objectifs	Type de relation entre les organisations	Avantages	Stratégie applicable
Ressource groupée	Pour concurrencer les grandes firmes, en formant un entrepôt virtuel, en partageant les coûts et les risques.	Rivales (horizontal)	Accroître la concurrence et le partage du marché, diminution des charges et des coûts.	Stratégie de croissance
Coopération complémentaire	Fournir un service intégrer, partager les données des transactions de vente pour les analyser.	Un service complémentaire relié aux services/produits (vertical)	Améliorer le service client ; accroître les ventes	Une stratégie de différenciation convergente
Coopération opérationnelle	Création d'un entrepôt virtuel ; partage et échange des informations pour l'aide aux opérations.	Rivales (horizontal)	Diminuer le coût du stockage ; Améliorer le service client ;	Une stratégie de différenciation
Coordination Opérationnelle	Faciliter la communication des processus de coordination ; le traitement efficace des opérations internes	Organisations dans la chaîne de valeur (vertical)	Améliorer la fonction du contrôle ; améliorer l'efficacité opérationnelle	Stratégie du coût leadership

Tableau 1.2 : Etude comparative entre les SIIO d'après Hong [EOM05]

Le cadre d'étude des SIIO est multidisciplinaire, il implique des concepts inhérents aux sciences de l'organisation, du management, sociales et des technologies des SI. Ce dernier point ayant trait aux plates-formes techniques employées dans le contexte inter-organisationnel sera discuté dans la prochaine section.

1.5. Technologies des SIIO

Comme défini préalablement, un SIIO est un système d'information et de gestion transcendant les frontières des organisations via des liaisons électroniques avec leurs partenaires commerciaux. Par conséquent la gestion des échanges est importante, car chaque partenaire de l'entreprise doit avoir accès

¹ **Sabre :** système de réservation créé par la compagnie aérienne American Airlines

² **Nike :** un producteur américain d'articles sportifs.

à l'information utile, en temps voulu, cette information doit être unique pour tous les acteurs de l'échange. Cela signifie concrètement qu'une infrastructure de communication doit être mise en place, que les accès aux données doivent être contrôlés et suivis, les évolutions gérées et diffusées [YP96].

Une infrastructure de système d'information pour le support d'une entreprise étendue ou virtuelle fait appel aux différentes normes et technologies de l'information pour son fonctionnement, allant des solutions de partage de données (SGBD/SGDT, Entrepôt de données,...), via des réseaux distribués (Internet/Intranet, EDI,...), aux solutions de gestion des processus (Groupware, Workflow, ...).

Une synthèse des plateformes technologiques des différents SIIO sera présentée dans cette section selon le type d'interdépendance reliant les organisations.

1.5.1. SIIO en interdépendance groupée et SIIO en interdépendance séquentielle

Les technologies que nous allons présenter ici peuvent être utilisées dans les deux modèles de SIIO, à savoir, SIIO en interdépendance groupée ou SIIO en interdépendance séquentielle. On devra considérer deux critères de sélection pour un système inter-organisationnel [BA02] :

Le mode d'accès à l'information d'une organisation à l'autre : Le mode d'accès peut être basé sur un échange de messages simples entre organisations ou de nature beaucoup plus interactive.

Et le type d'accès permis par le système : Le type d'accès permis peut consister en une simple consultation d'information. Il peut éventuellement permettre des mises à jour ou des modifications dans les bases de données d'une autre organisation.

Comme le montre la figure 1.4, nous pouvons citer quatre technologies inter-organisationnelles soit l'EDI, le XML, les Extranets ainsi que les applications propriétaires qui sont classées en fonction des critères sus cités. Il ne s'agit pas de technologies mutuellement exclusives puisqu'elles peuvent être utilisées en combinaison. On peut remarquer que ces technologies se situent sur un axe en diagonal qui correspond naturellement à une architecture d'intégration inter-organisationnelle allant de très faiblement couplée pour l'EDI à parfaitement intégrée pour les applications propriétaires [BA02].

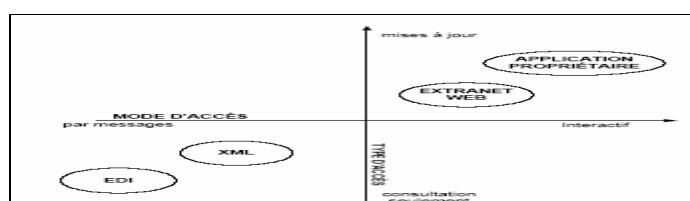


Figure 1.4: classification des technologies en support à chaîne de valeur [BA02].

D'après [BA02], une synthèse de ces technologies permet de déterminer dans quelles conditions elles peuvent soutenir les activités d'un processus inter-organisationnel comme illustré dans le tableau 1.3 qui résume les éléments clés associés à chaque technologie :

	EDI	XML	EXTRANET	APPLICATION PROPRIÉTAIRE
Processus IO supportés	Formels et simples	formels et complexes	ad hoc	ad hoc
Environnements Technolo.	homogènes	hétérogènes	accès Web	nécessite un accès à un réseau privé

Tableau 1.3 : Technologie des SIIO partagés et en interdépendance séquentielle [BA02].

1.5.2. SIIO réseau

Les systèmes d'information inter-organisationnels réseau qui supportent le travail collaboratif sont regroupés sous les termes *Groupware* ou *Computer-Supported Cooperative Work (CSCW)*. Le tableau 1.4 présente une synthèse des deux technologies les plus prépondérantes dans les SIIO de type réseau, qui sont les systèmes de workflow et l'édition commune des documents.

	Système de gestion de workflow et d'intégration	Systèmes de gestion documentaire
Processus supportés	- Processus simples ou complexes formalisés entre partenaires d'affaires.	- Processus ad hoc (non formalisés)
Environnements Technolo.	- Multiples et complexes	- Accès par le Web

Tableau 1.4: Technologie des SIIO en réseau (Workflow et Gestion documentaire) [BA02].

1.6. Synthèse

Faisant face au problème de gestion des informations dans un SIIO, caractérisé par le nombre important des sources de données, autonomes et hétérogènes, la conception et l'élaboration d'un entrepôt de données global impliquant les entités analytiques des organisations formant le SIIO, s'inscrit comme un moyen pour maintenir les données orientées sujet, intégrées et stables pour le support de l'aide à la décision.

Dans ce contexte, le problème d'intégration des entrepôts de données dans un tel environnement revêt un caractère complexe car les stratégies de consolidation et les architectures à adopter en vue de mettre en commun des données analytiques sont intimement liées à des considérations multidisciplinaires, à savoir :

1. La nature des relations entretenues entre les organisations des SIIO ; contractuelles ou conjoncturelles et la prépondérance d'une organisation par rapport aux autres organisations du SIIO selon l'ampleur de ses investissements engagées dans les infrastructures techniques,
2. Le flux informationnel circulant entre les différents SI permettant la mise en relief de trois architectures de données interopérables [JP94] pouvant être résumées comme suit :
 - Pour le SIIO centralisé : l'interopérabilité est configurée en étoile d'où l'élaboration d'un schéma global réconciliant les différentes vues partielles ou totales des schémas des entrepôts de données existants au niveau des organisations formant le SIIO.
 - Pour le SIIO en chaîne : l'interopérabilité est configurée en bus ce qui nécessite la mise en correspondance des différents référentiels de données ou méta-données pour produire un référentiel commun sur la base d'une architecture fédérée.
 - Pour le SIIO en réseau : l'interopérabilité se présente sous forme d'anneaux ou en pair d'où la nécessité des mécanismes de communication bilatéraux.

Ces points vont conditionner le caractère centralisé ou décentralisé de l'architecture de l'entrepôt de données intégré et mis en commun entre les organisations du SIIO. Ce point sera étudié après avoir introduit les concepts et les stratégies de développement des entrepôts de données dans le chapitre 2.

CHAPITRE 2. Entrepôts de données : concepts et méthodologies

2.1. Introduction

Comme l'entrepôt de données s'inscrit dans le cadre d'une solution d'intégration de données pour un usage décisionnel, le but visé dans ce chapitre est de faire ressortir les concepts de bases explicitant les concepts de base de l'ED et de la modélisation multidimensionnelle et la synthèse des différentes variantes méthodologiques d'élaboration des ED.

Le concept d'entrepôt de données est différent de celui d'une base de données transactionnelle conventionnelle. Le dernier supporte le traitement on-line des transactions (*OLTP*), pendant que le premier est dédié au traitement analytique en ligne (*OLAP*), une présentation des concepts inhérents aux ED est donnée ci-après.

2.2. Entrepôt de données, concepts et composants

Bill Inmon dans son ouvrage de référence « Using the Data Warehouse » définit l'entrepôt de données (Data Warehouse) de la façon suivante : « *L'entrepôt de données est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision.* » [TES00]

- *Intégrées* : Les données de l'entrepôt proviennent de différentes sources éventuellement hétérogènes. L'intégration consiste à résoudre les problèmes d'hétérogénéité structurelle et sémantique;
- *Orientées sujet* : Les données de l'entrepôt peuvent être réorganisées autour de thèmes majeurs d'une organisation;
- *Non volatiles* : Les données de l'entrepôt sont essentiellement utilisées en mode de consultation, elles sont très rarement modifiées;
- *Historisées* : La prise en compte de l'évolution des données est primordiale pour la prise de décision et notamment les prédictions ;
- *Résumées* : Les informations issues des sources de données doivent être agrégées et réorganisées afin de faciliter le processus de prise de décision par l'interrogation et l'élaboration des analyses pour les utilisateurs en fonction de leurs privilèges d'accès.

Les composantes majeures d'une architecture d'entrepôt de données sont détaillées dans [KIM96, KIM98, NR02, SC02, KIM02, INM05], comme suit (Fig 2.1) : le système source (« Source System »), la zone de préparation des données (« data staging area »), l'entrepôt de données (« Data Warehouse »), le magasin de données (« Data Mart ») qui est un extrait de l'entrepôt, l'outil ETL (« Extract Transform Load »), le dépôt des métadonnées et les outils d'aide à la prise de décision (« Business Intelligence Tools »).

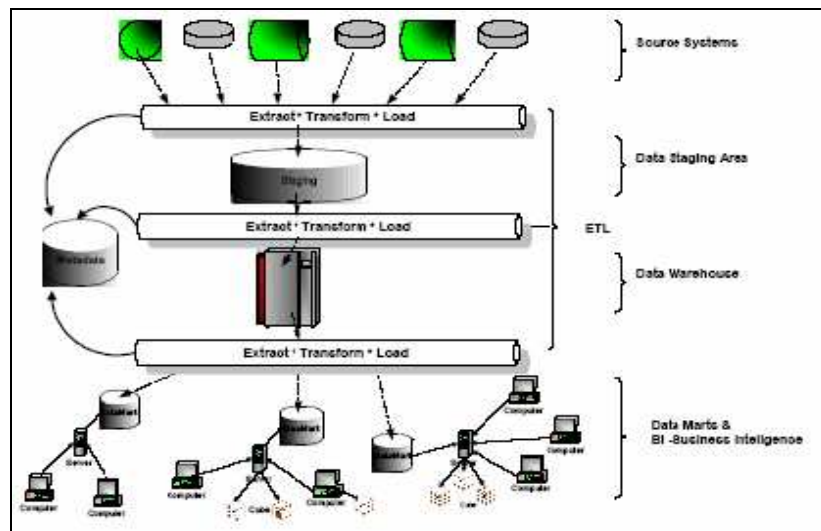


Figure 2.1. : Architecture type d'un entrepôt de données [SC02]

2.3. La modélisation multidimensionnelle

La conception des entrepôts de données est basée sur la modélisation multidimensionnelle [KIM96] comprenant la notion de table de faits (permettant de mesurer l'activité) et d'un ensemble de tables dimensionnelles (stockant les paramètres relatifs aux informations faisant varier les mesures d'activité).

Chacune des tables dimensionnelles possède une clé primaire unique correspondant exactement à l'un des composants de la clé de la table des faits. L'analyse multidimensionnelle consiste à voir les données sous une forme cubique (à trois dimensions ou plus dans le cas d'un hyper cube).

La disposition de la table de faits et de dimensions donne naissance à trois configurations de schémas conceptuelles :

- *Schéma en étoile (star)* : Au milieu, une table de faits connectée à un ensemble de tables de dimensions.
- *Schéma en flocon de neige (snowflake)* : Un raffinement du précédent où certaines tables de dimensions sont normalisées donc décomposées et hiérarchisées.
- *Constellation de faits* : Plusieurs tables de faits partagent quelques tables de dimensions (constellation d'étoiles).

On remarque que plusieurs formalismes de modélisation multidimensionnelle avec leurs langages de requêtes ont été proposés par les communautés de chercheurs où chacun à sa propre vision sur l'analyse multidimensionnelle des besoins ainsi que sur la terminologie et le formalisme à employer¹. Par conséquent, il n'y a pas un consensus autour d'un formalisme de modélisation dimensionnelle. Dans le cadre de ce travail, nous allons adopter le modèle de L. Cabbibo et R. Torlone [LR05],

¹Pour plus de détails sur les formalismes de modélisations dimensionnelles, la référence [BHS98] peut être considérée comme une synthèse dans ce domaine.

comprenant un formalisme de modélisation multidimensionnelle et un langage de requêtes correspondant basé sur le calcul logique et qui offre une solution formelle pour la conception ascendante des entités analytiques. Le modèle multidimensionnel, dans ce cas, est défini comme suit :

Définition 1 (Dimension) : une dimension est définie à partir d' :

- *Un schéma*, défini par un ensemble fini $L(d) = \{l1, \dots, ln\}$ de niveau ; et un ordre partiel $\leq d$ dans $L(d)$; si $l1 \leq d$ $l2$ on dit que $l1$ rolls up à $l2$;
- *Une instance*, définie par une fonction m associant un ensemble de membres avec chaque niveau ; et une famille de fonction ρ incluant une fonction de roll up $\rho_{l1 \rightarrow l2} : m(l1) \rightarrow m(l2)$ pour chaque paire de niveau $l1 \leq l2$.

Définition 2 (Table de fait) : une f_table f sur un ensemble D de dimensions est composée d' :

- *Un schéma* $f[A1 : l1, \dots, An : ln] \rightarrow (M1 : \tau1, \dots, Mm : \tau m)$, ou chaque Ai est un nom d'attribut distinct, chaque li est un niveau d'une dimension dans D , Chaque Mj est un nom d'une mesure distincte, et chaque τj est un type de base ;
- *Une instance*, qui est une fonction partielle de mapping de coordonnées de f aux faits f , où :
 - les coordonnées forment un tuple sur les attributs de f , c'est une fonction de mapping de chaque attribut de nom Ai à un membre de li ;
 - un fait est un tuple sur les mesures de f , c'est une fonction qui mappe chaque mesure Mj à une valeur dans le domaine de type τj .

Une collection de f-table sur les mêmes dimensions compose un ED ou un MD.

Définition 3 : Un ED ou un MD est composé d'un ensemble :

- D de dimensions ; et
- F de f-table sur des dimensions de D .

Exemple : la figure 2.2 illustre des schémas conceptuels en flocon de neige de ED/MD qui consistent en une table de fait. Cet exemple est inspiré d'un cas d'étude illustré dans [KIM02]. L'ED/MD des **ventes** représente les ventes quotidiennes d'un produit dans une chaîne de magasins, et l'ED/MD **inventaire entrepôt** représente l'inventaire et les stocks du même produit dans la même chaîne de magasins par semaine.

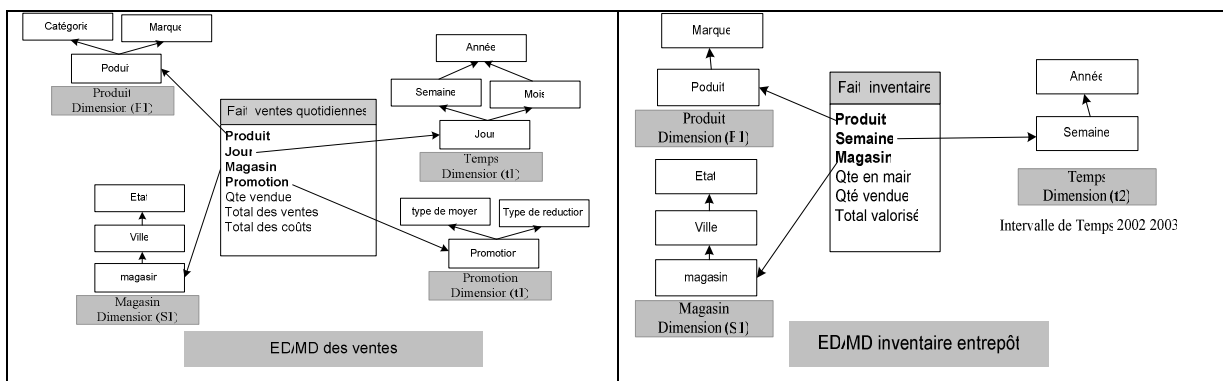


Figure 2.2. : Exemple de schéma conceptuel en flocon de neige d'un ED/MD [LR04, LR05]

Comme on va le spécifier plus loin, il y'a plusieurs méthodologie de conception des entrepôts de données distinctes, toutefois, à un certain niveau d'abstraction leur processus de conception peuvent avoir comme points commun ce qui suit:

- la sélection du domaine métier à analyser à travers l'identification des indicateurs métiers récupérés à partir des sources opérationnelles, ou juste à partir des sources correspondantes à un processus métier ;

- définir la granularité de la table de fait et les dimensions qui vont avec ;
- Identification de la table de fait et de ces mesures ;
- Identification des dimensions et de ces attributs;
- Correspondre les données sources au schéma dimensionnel cible.

Une fois qu'on a abordé brièvement l'aspect conceptuel des ED, il faut mentionner qu'au niveau physique la modélisation multidimensionnelle exige un stockage de données permettant une haute performance de réponse aux requêtes, de scalabilité et des accès de plusieurs utilisateurs. Trois stratégies d'implémentation des structures de stockage d'une base de données multidimensionnelle peuvent être employées pour l'analyse multidimensionnelle :

- ROLAP (« Relational On-Line Analytical Processing ») : Consiste en l'implémentation des cubes comme des tables relationnelles en utilisant un système de gestion de bases de données relationnelles ;
- MOLAP (« *Multidimensional On-Line Analytical Processing* ») : Cette approche consiste à utiliser un système multidimensionnel « *pur* » qui gère des structures multidimensionnelles natives donc l'implémentation des cubes comme des matrices en mémoire;
- HOLAP : (Hybrid-OLAP) est une combinaison du ROLAP et du MOLAP. Une partie des tables de faits et des tables de dimensions sont stockées dans une base relationnelle standard tandis que le reste des données (les calculs) sont stockés dans une base multidimensionnelle.

2.4. Architectures des entrepôts de données

Dans ce qui suit, nous allons exposer les variantes architecturales des ED. Pour implémenter un entrepôt de données, trois d'architectures sont possibles [GC04] :

1. **L'architecture réelle** : Le stockage des données est réalisé dans un SGBD séparé du système de production. L'avantage est de disposer de données préparées pour les besoins de la décision et répondant aux objectifs de l'ED. Les inconvénients sont le coût de stockage supplémentaire et le manque d'accès en temps réel.

2. **L'architecture virtuelle** où 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. Par contre, l'inconvénient est que les données ne sont pas préparées.

3. **L'architecture remote** qui est une combinaison de l'architecture réelle et de l'architecture virtuelle. 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.

2.5. Méthodologies, approches d'analyse et stratégies de développement d'un ED

La communauté de recherche en bases de données s'active pleinement dans le domaine des ED, néanmoins, la problématique liée à l'élaboration d'une méthodologie formelle de construction et de conception d'un ED n'est pas profondément explorée [MGS98]. Tandis que beaucoup de chercheurs

se sont penchés sur la modélisation multidimensionnelle des données et la modélisation physique (la matérialisation des vues, le choix des index,...etc), en revanche, il n'y a pas suffisamment d'efforts significatifs fournis pour le développement d'une méthodologie de conception et d'élaboration d'un ED complète et détaillée. Dans ce contexte, s'inscrivent les travaux de [DJ98], [GS98], [KIM98] et [PVK98] pour la mise en oeuvre d'une méthodologie de réalisation d'un ED.

Parmi ces travaux, [KIM98] et [PVK98] sont considérés comme étant les plus détaillés et complets d'après plusieurs auteurs. Bien que, les phases proposées soient légèrement différentes dans leur essence, ces travaux sont similaires en terme de composants, de fonctionnalités et d'activités.

D'après [JD00] les différences parmi ces méthodologies ou processus sont multiples et variables, mais généralement reliées aux détails du processus. Les points de différenciation communs sont :

- L'organisation : les activités sont structurées, séquentielles ou groupées différemment.
- La terminologie : l'utilisation de différents termes pour le même concept ou le contraire.
- Les livrables : différents ensembles de livrables orientés vers la production du même résultat final un ED ou un MD.
- Le niveau de détail : impliquant des spécificités au niveau du livrable ou de l'activité.

Cependant, d'après [WP00], on peut considérer l'ensemble de ces variantes méthodologiques, comme un cadre structurel qui peut fournir des critères permettant d'élaborer une comparaison entre ces méthodes de réalisation d'un ED comme illustré dans le tableau suivant :

	[D J 9 8]	[G S 9 8]	[K I M 9 8]	[P V K 9 8]
Phase	<ul style="list-style-type: none"> · Etude préalable du M D · planification projet · M D · Implémentation 	<ul style="list-style-type: none"> · Planning · Conception et implémentation · Support et amélioration 	<ul style="list-style-type: none"> · Planification du projet, · Définition des besoins métiers · Conception de l'architecture technique · Sélection des produits et installations · Modélisation dimensionnelle · Conception physique · Conception de la zone de préparation et développement · Spécification des applications des utilisateurs · Développement des applications des utilisateurs · Déploiement · Maintenance et extension 	<ul style="list-style-type: none"> · Planning · Assemblage des données exigées et modélisation · La conception physique de la base de donnée et développement · Sources de données, intégration et correspondance · Alimenter l'ED · Automatisation du processus de management de l'ED · Créé le premier ensemble d'états · Validation et tests de données · Formation · exploitation
Architecture de données	Non	Entreprise DW DM indépendant DM dépendant	Accès aux données (M O L A P et R O L A P)	Générique, DM indépendant, DM dépendant, B D D intégrées pour alimenter un ED
Architecture fonctionnelle	Non	non	Back room, presentation servers, front room, services et metadat	Intégration de données, transformation de données, ED, metadata et architecture de données,
Modélisation dimensionnelle	Non	non	Etoile et flocon de neige	Etoile et flocon de neige, table de fait multiples, tables extérieures et schéma multistar
Conception de la base de données	Non détaillée	Non détaillée	Tres détaillée	Détaillée
Infrastructure technique	Non	non	oui	Oui

Tableau 2.1 – Comparaison entre les méthodologies d'élaboration d'ED [WP00]

A cet effet, il est à noter qu'en plus des différentes projections en phase de ces méthodologies faisant usage de différentes techniques de conception, de construction et de maintenance, elles font appels à des approches d'analyse et des stratégies d'implémentation distinctes que nous allons expliquer dans ce qui suit .

2.5.1. Approches d'analyse

Elles permettent d'orienter la manière par laquelle les besoins sont identifiés et spécifiés pour définir les faits, les dimensions et les attributs d'un schéma multidimensionnel dans un processus de conception d'un ED. Les approches d'analyse suivies pour le développement des ED se scindent en : approche d'analyse orientée processus métier et approche d'analyse orientée données.

A. Approche d'analyse orientée données : Elle est basée sur l'analyse des modèles de données des sources opérationnelles et l'identification des transactions pertinentes, d'où la primauté à la conception et à l'alimentation de l'ED et seulement quand ce dernier serait soumis aux requêtes, alors la réconciliation entre les données, le schéma dimensionnel et les besoins métiers s'achève.

Par exemple [MGSR98] ont développé une méthodologie semi automatique pour la construction d'un modèle dimensionnel d'un ED à partir des schémas entité-relation (ERM) préexistant décrivant le

système d'information opérationnelle, où le modèle d'organisation des données d'après [INM05]. Une autre extension de ce modèle a été introduite par [MA99] qui est le SERM (*Structured Entity Relationship Model*). Le problème majeur avec cette approche est double :

- la restriction du modèle de données du système opérationnel qui spécifie d'une manière incomplète le monde métier des organisations.
- Les besoins des utilisateurs ne sont pris en compte que pendant la phase de maintenance et d'exploitation de l'ED, d'où le risque élevé d'aboutir à une inadéquation entre les besoins et le résultat du développement [MC03].

B. Approche d'analyse orientée processus métiers : Elle se focalise sur le processus métier par qui l'ED s'alimente. L'avantage de cette approche est que le modèle de processus métier comprend une description formelle des informations dont l'utilisateur a besoin ainsi que la possibilité d'identifier des besoins en informations non disponibles et non fournies par le système opérationnel actuel. L'analyse de ces processus peut être dirigée par les besoins des utilisateurs ou par les objectifs à atteindre par ce dernier, et qui peuvent être résumées ainsi :

B.1. Analyse du processus métiers orienté utilisateur : où les mesures des faits correspondent aux facteurs critiques de succès de l'entreprise lesquels dépendent de sa stratégie et de ces objectifs ainsi que des dimensions rattachées à ces facteurs, mais il n'y a pas un cheminement prédéterminé pour leur définition. Cependant, l'approche exploite le concept de guide d'interviewage des différents groupes d'utilisateurs pour le recueil des besoins en information comme illustré par [PVK98]. Dans cette approche, l'ED n'est pas conçu pour supporter un nombre restreint de requêtes mais la prise de décision concerne la totalité du processus métiers.

B.2. Analyse du processus métiers orientée but : assume la question principale du développement de l'ED qui est de décider quel est le processus métier à améliorer et quelle amélioration doit être apportée, à travers l'identification des objectifs à modéliser en faits et de leur rattachement aux besoins en informations contextuelles statiques faisant office de dimensions [MB00].

2.5.2. Stratégies d'implémentation et architectures

Cette section vise un double objectif, en premier lieu, mettre en évidence un critère des méthodologies d'élaboration des ED, à savoir, les stratégies d'implémentation et en deuxième lieu mettre en exergue une observation recueillie lors de nos travaux de recherche [DH98, TT98, WE03, MC03, PS04, HA05] relative à la dépendance existante entre la méthodologie de réalisation de l'ED, la stratégie d'implémentation de l'ED et l'architecture visée. Les quatre approches décrites ci-dessous représentent les tendances dominantes des stratégies d'implémentation des ED.

1^{ère} stratégie : Top down, prônée par **Bill Inmon**, cette stratégie est conduite inversement au cycle de vie habituel de développement classique SDLC¹ (en cascade) car le développement de l'ED est conduit par les données au lieu des besoins des utilisateurs. Cette approche s'appuie sur le

¹ **SDLC** (System Development Life Cycle) : Cycle de vie de développement classique qui commence par l'étude des besoins et abouti à l'implémentation du produit final.

développement d'un ED en un seul projet, et commence par la conception d'un modèle d'entreprise pour l'ED, et le déploie sur une architecture multi tiers comprenant une zone de préparation de données, un ED d'entreprise comprenant des données détaillées, des MD dépendants¹ dédiés aux fonctions spécifiques et comprenant des données agrégées (figure 2.3). Ainsi l'utilisateur peut adresser ses requêtes pour l'ED ou le MD. Cette approche est itérative de nature. L'avantage de ladite approche est qu'il permet de limiter l'étendue de l'analyse aux frontières du système, néanmoins, ses inconvénients sont la longue durée du projet, le coût élevé, la difficulté de la détermination du ROI (le retour sur investissement) qu'à la fin du projet[WE03, MC03].

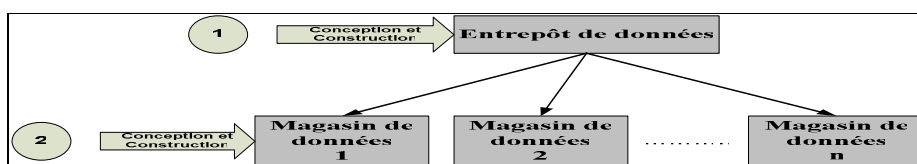


Figure 2.3 : Stratégie d'implémentation Top-Down [TT98].

Les variantes architecturales de cette approche d'implémentation sont l'ED classique de l'entreprise qui est un dépôt commun et unique pour les données de l'entreprise, dédié à la consultation et composé de données détaillées, agrégées, nettoyées et intégrées (Figure 2.4.A) [DH98, SAXY00].

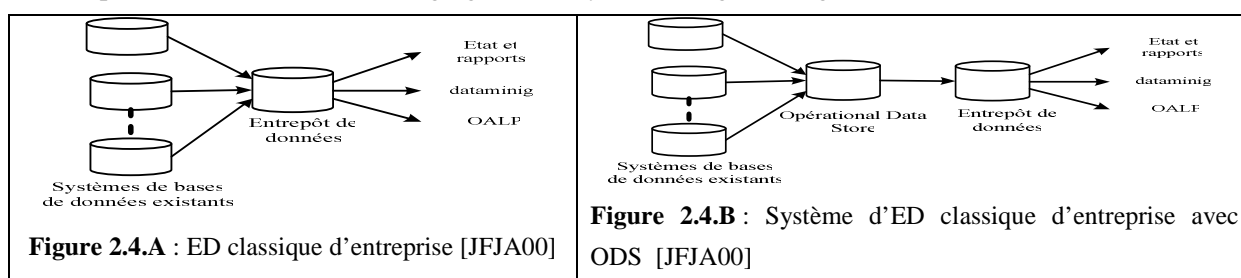


Figure 2.4.A : ED classique d'entreprise [JFJA00]

Figure 2.4.B : Système d'ED classique d'entreprise avec ODS [JFJA00]

Une autre variante de cette architecture est l'architecture Operational Data Store/ ED (ODS) (Fig 2.4.B) qui est une extension de l'architecture centralisée caractérisée par l'ajout d'un espace de stockage des données opérationnelles défini par Bill Inmon comme : « une collection de données orientées sujets, intégrées, volatiles (modifiables), actuelles pour soutenir les analyses décisionnelles quotidiennes tactiques ». A ce niveau, les sources de données sont intégrées dans l'ODS et transférées vers l'entrepôt ensuite [JFJA00].

2^{ème} stratégie : Bottom up, prônée par **Ralph kimball**, elle se focalise sur les besoins d'analyse exprimés par les cadres métiers pour la conception des MD dimensionnels. Le cycle de vie commence avec la planification suivie par la définition des besoins métiers, la conception architecturale fonctionnelle, la modélisation conceptuelle, logique et physique, le déploiement et enfin l'exploitation et la maintenance. D'après cette approche, pour le développement d'un ED d'entreprise, il est irréaliste de cerner tous les besoins métiers à priori, donc l'approche en SDLC (cascade) est non viable. Pour spécifier les besoins, une approche itérative telle que le prototypage, peut être adoptée. La projection en

¹ **MD dépendant** : magasin de données alimenté directement par un entrepôt de données.

phase du développement des MD est plus pratique parce qu'elle se base sur le processus métier dont l'étendu et la complexité est plus restreinte que les besoins globaux de l'entreprise.

En effet, ladite approche consiste en la construction de MD en premier lieu, donc elle débute par la conception d'un modèle dimensionnel généralement en un seul schéma en étoile, logiquement et physiquement déployé d'une façon incrémentale sur une architecture comprenant une zone de préparation et un MD contenant les données détaillées et agrégées répondant à un besoin spécifique au niveau de l'entreprise (figure 2.5), l'architecture résultante de cette approche est illustrée dans la figure 2.6, appelée le **MD architecturé (Incremental Architected Data Marts)**.

Une fois que le MD est opérationnel, il peut être fédéré dans l'ED d'entreprise, donc pour chaque nouveau MD et pour s'intégrer aux autres MD il faut définir des dimensions et des faits conformes¹, afin de faire évoluer le système vers un ED d'entreprise. Cette approche est largement répandue, car elle permet d'atteindre un résultat et une évaluation du ROI immédiat, néanmoins la prolifération incontrôlée des MD créerait des **MD indépendants non architecturés (Leg Marts)** illustrés dans la figure 2.7, engendrant ainsi des problèmes d'intégration pour la construction de l'ED d'entreprise et rend les analyses transversales impossibles à opérer jusqu'au développement complet de l'ED [MC03, WE03].

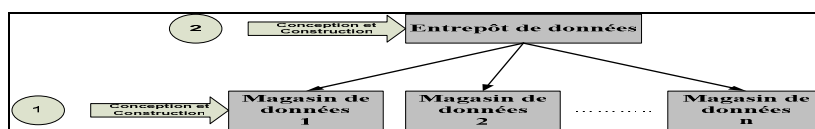
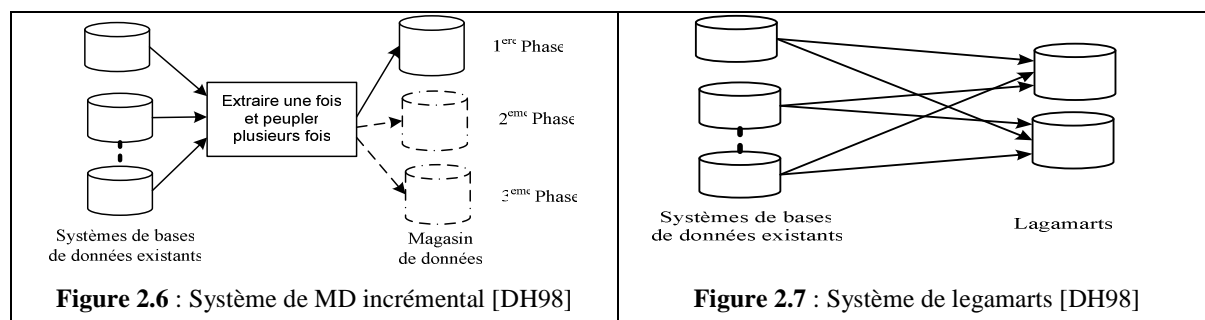


Figure 2.5 : Stratégie d'implémentation Bottom-Up [TT98].



3^{ème} stratégie : Hybride, préconisée par **Tulu Tanrikorur** (1998), comme illustré dans la figure 2.8, elle combine les deux approches suscitées, en se concentrant sur la conception d'un MD en ayant un modèle d'entreprise en tête, elle débute par la conception d'un modèle d'entreprise global et départemental local simultanément. Le modèle de données de l'entreprise est déployé sur un MD initial,

¹ **Dimensions et faits conformes** : Une dimension conforme est une dimension qui a la même signification dans toutes les tables de faits avec lesquelles elle peut être reliée. Une dimension conforme peut être reliée à plus d'une table de faits, ce qui facilite leur croisement. L'identification des faits conformes s'effectue en même temps que celle de dimensions conformes. Les définitions des faits conformes permettent d'employer la même terminologie dans toutes les bases de données multidimensionnelles et elles imposent l'utilisation des mêmes unités de mesure dans toutes les bases de données multidimensionnelles [Kim 00].

modélisé en un seul ou plusieurs schémas en étoile, ensuite, elle peuple le ou les MD par des données atomiques et agrégées via une zone de préparation en utilisant un ETL en synchronisant le contenu des méta-données entre le MD et l'ETL [WE03, TT98], permettant ainsi d'avoir des métadonnées homogènes qui servira dans une phase future pour la construction d'un ED d'entreprise. Cette approche suit un paradigme de développement itératif au même titre que l'approche top down. L'architecture préconisée par cette stratégie est les MD interdépendants.

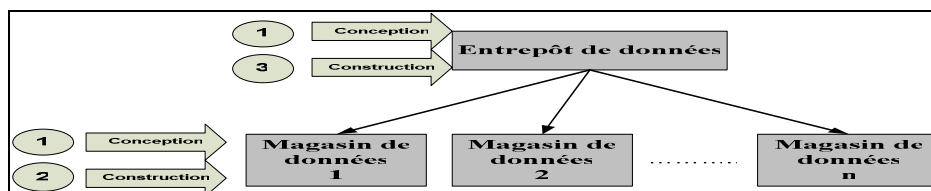


Figure 2.8 : Stratégie d'implémentation hybride [TT98]

4^{re} stratégie : Fédérée, proposée par **Doug Hackney** et **Coline White** [WE03], elle consiste en l'intégration de multiples ED, MD et ODS hétérogènes et distribués d'une façon non matérialisée donc cette architecture laisse les structures analytiques en place et sur la base des besoins métiers, les données sont accessibles à partir des sources, à la volée. Les données peuvent être intégrées physiquement ou logiquement en utilisant des clés partagées, un meta-données global ou des requêtes distribuées.

Selon [CW01], ce modèle d'architecture suppose que la conception et le développement des MD indépendants sont administrés et basés sur un modèle métier partagé et une zone de préparation de données commune. Le modèle métier commun assure la consistance dans l'utilisation des noms de données et la définition des concepts métiers et permet un stockage centralisé de toutes les mesures clés pouvant être partagés entre les différents MD dans l'organisation.

Cette stratégie est préconisée comme une solution pratique pour l'intégration virtuelle des structures analytiques appartenant à différentes organisations ne voulant pas les reconstruire (figure 2.9) [HA05, PS04].



Figure 2.9 : Stratégie d'implémentation fédérée [PS04].

2.5.3. Synthèse sur les méthodologies

En conclusion et en complément du tableau comparatif (en blanc) des caractéristiques importantes des variantes architecturales des ED et des MD de [PO02], nous présentons notre synthèse ayant trait aux dépendances existantes entre les stratégies d'implémentation, les architectures et les approches d'analyse régissant les méthodologies d'élaboration des ED illustrées dans le tableau suivant

Stratégie D'implémentation	Top down	Bottom up	Hybride	fédérée
Auteur	Bill Inmon	Ralph kimball	Tulu Tanrikorur	Doug hackney et Coline White
Architecture De Base	Centralisée	Distribuée	Distribuée	Distribuée
Variants Architecturales	ED d'entreprise centralisé + architecture Operational Data Store/ED + MD dépendants.	MD indépendants : Leg Marts ¹ + Incremental Architected Data Marts ²	MD interdépendants	MD ou ED fédérés
Source De Données	Source opérationnelle	Source opérationnelle	Source opérationnelle	ED ou MD
Paradigme De Développement	Itératif	Itératif et incrémental	Itératif	Itératif
Niveau De Migration Prochain	MD	ED centralisée physiquement ou logiquement	ED centralisée physiquement ou logiquement	ED virtuel intégré
Approche D'analyse	Orientée données	Orientée processus métier	Orientée processus métier et données	Orientée processus métier et données

Tableau 2.2 : stratégie d'implémentation, architecture et approche d'analyse des méthodologies d'élaboration des entrepôts de données.

A travers ce tableau, et en guise de conclusion relative aux méthodologies d'élaboration des ED on peut déduire ce qui suit :

- Il y'a une dépendance entre **la méthodologie adoptée et l'architecture implémentée**, car une stratégie d'implémentation top down s'oriente vers une architecture centralisée ou de MD dépendants, tandis que l'approche bottom up mène vers des MD indépendants (en bus) etc. Néanmoins, on peut dire que toutes les stratégies et les architectures correspondantes sont plus ou moins détaillées et spécifiées hormis l'approche fédérée, car les auteurs, tels que, **Doug Hackney** et **Colin White** ne proposent pas une démarche formelle et détaillée, d'où la spécification d'une méthodologie de consolidation des ED/MD distribués, hétérogènes et autonomes doit être prise en charge.

- Toutes **les approches d'analyse pour le développement d'un ED** sont soit dirigées par les processus (objectifs ou utilisateurs) ou dirigées par les données opérationnelles, aucune ne propose une combinaison des deux approches pour appréhender, à la fois, le changement des objectifs, des besoins métiers et des systèmes opérationnels.

- **Le degré de formalisation des méthodes** utilisées pour l'élaboration des ED varie entre les méthodes dirigées par les guides de recommandation et les méthodes formalisées incluant des notations de méta modélisation. De plus, elles ne formalisent pas uniformément, le passage entre la spécification des besoins et la modélisation conceptuelle de l'ED, la détermination des exigences fonctionnelles et non fonctionnelles de l'ED à travers les besoins des utilisateurs ou à travers les objectifs, etc.

¹ Legamart [DH98]: Magasins de données indépendants non architecturés.

² Incremental Architected Data Marts [DH98]: MD architectures.

- Bien que beaucoup de solutions formelles soient développées pour des problèmes impliquant la structuration, la gestion et le stockage de données, à usage décisionnel, la mise en oeuvre d'une méthodologie de conception et de développement d'un ED relève du concepteur et du praticien qui doivent combiner des solutions partielles, formelles et informelles pour aboutir à un résultat.

- le **paradigme de développement d'un ED** ne peut être qu'un *processus incrémental et itératif*, vu sa complexité qui nécessite des validations continues par l'utilisateur pour atténuer les risques d'échec du projet à travers des prototypes, par exemple, lors de la validation des spécifications des besoins.

Un autre constat peut être fait sur l'architecture d'ED à base de MD laquelle est largement répandue et offre l'occasion de développer un large ED d'entreprise. En effet, une façon réaliste de créer un ED pour une grande entreprise est d'intégrer les MD ou les ED départementaux dans un schéma global. Cela conduit à un genre d'ED distribué ou d'ED fédéré dont le point de différenciation le plus important est le degré d'autonomie des sites.

Cette conclusion nous conduit à explorer les architectures distribuées des ED, qui s'apparente avec le cadre d'étude qui préconise une démarche de conception bottom up d'un entrepôt de données distribué dans un environnement inter-organisationnel où chaque organisation a son propre ED conçu indépendamment des autres.

2.6. Architectures distribuées des entrepôts de données

En se référant à la littérature, *les implémentations des architectures les plus répandues, des entrepôts de données distribués* peuvent être résumées ainsi :

2.6.1. Entrepôt de données distribué

Selon [AK99], les inconvénients d'une architecture centralisée peuvent être surmontés par l'implémentation d'une architecture distribuée qui reflète précisément le caractère distribué des organisations actuelles. Cette approche requiert que l'ED local soit alimenté à partir de l'extraction des informations du système opérationnel local, résultant d'un ED local dans chaque branche. Chaque ED local est un sous-ensemble de l'ED global. [AK99] distingue deux écoles de pensées relatives aux ED distribués, celle de :

- **L'approche de Inmon (1993)** qui repose sur l'existence de l'ED local et l'ED global qui sont mutuellement exclusifs et où l'ED local est alimenté par le système opérationnel local (figure 2.10).

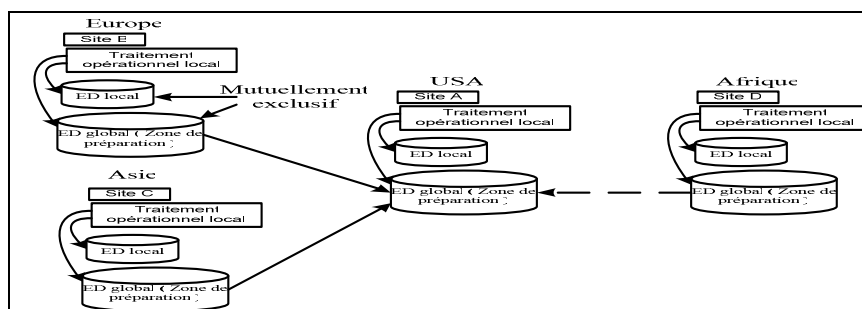


Figure 2.10. : Entrepôt de données distribué d'après l'approche de Inmon [AK99].

• **L'approche de White (1995).** C'est la combinaison d'un ED centralisé et des MD décentralisés dont la mise à jour et l'alimentation sont assurées par l'ED centralisé (figure 2.11).

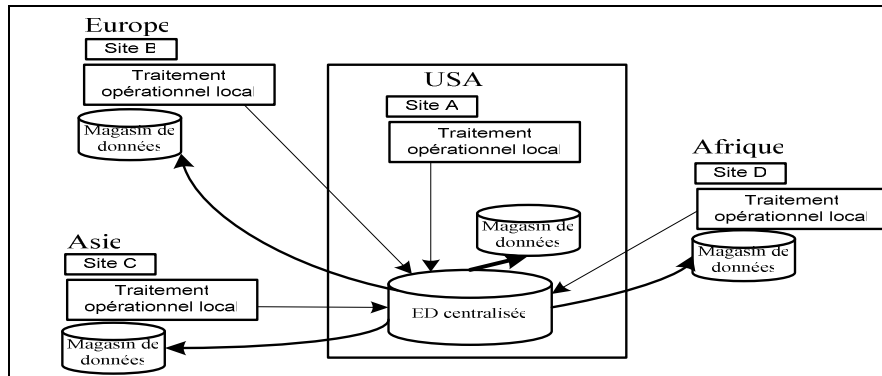


Figure 2.11. : Entrepôt de données distribué d'après l'approche de White [AK99]

[AK99] propose un autre modèle d'architecture distribué (figure 2.12). Dans ce modèle, la prise de décision est divisée en deux catégories différentes, les décisions affectant la globalité de l'organisation et les décisions affectant l'organisation localement. L'architecture proposée reflète ces deux catégories par l'ED détaillé distribué (Distributed Detailed Data Warehouse (DDDW)) et l'ED récapitulatif centralisé (Centralized Summarized Data Warehouse (CSDW)). Ce modèle fournit la transparence de la distribution, car les cadres locaux utilisent l'ED local tandis que les cadres au niveau supérieur (état major) peuvent utiliser l'ED local pour des analyses détaillées ou l'ED récapitulatif centralisé pour avoir une vue générale sur la situation métier.

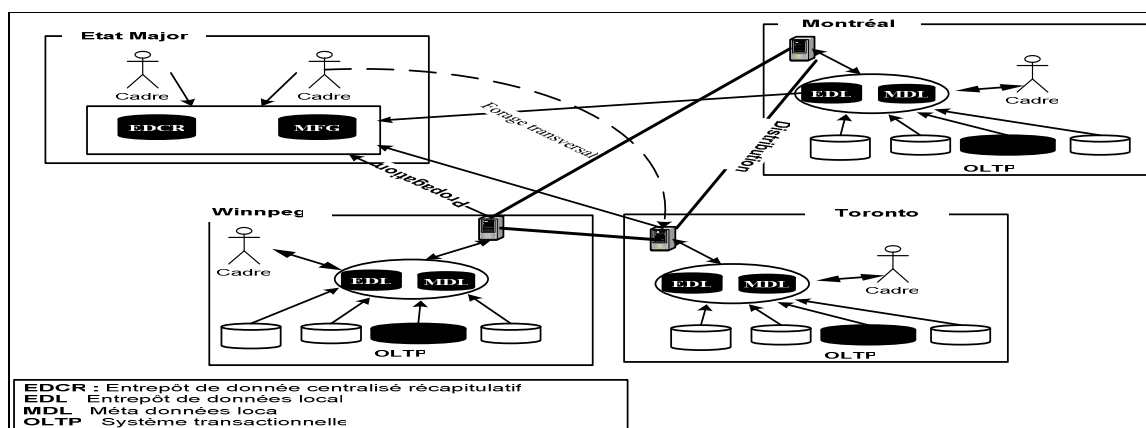


Figure 2.12. : Entrepôt de données distribué d'après l'approche de Noaman et Barker [AK99].

2.6.2. Entrepôt de données fédéré

Plusieurs auteurs [CW01, DH00] ont suggéré le modèle d'implémentation fédérée des ED. Ils présentent l'architecture fédérée de l'ED où les données sont consolidées logiquement mais stockées séparément dans différentes bases de données. Selon Hackney [DH00], l'architecture de l'ED fédérée est en général un système d'architecture qui intègre et consolide de multiples systèmes d'ED/MD ou d'ODS où les ED/MD locaux ne stockent que les informations nécessaires (agrégées et détaillées) pour un département ou un sujet et l'ED est virtuel, contrairement à l'architecture des MD dépendants où

l'ED est matérialisé et chaque MD contient des informations agrégées, et n'inclut pas les données détaillées. [CW01] décrit cette approche en un modèle architectural en spécifiant le *flux de données*, en partant du système opérationnel à la zone de préparation de données impliquant (Figure 2.13) :

- L'utilisation des outils de reingénierie de données,
- L'adoption d'un processus de développement commençant par l'analyse des besoins métiers,
- L'implémentation d'un modèle de données commun pour la zone de préparation de données et pour l'ED. a ce niveau, l'architecture fédérée de l'ED ajoute un détail technique à la conception d'un ED, qui est une zone de préparation de données commune qui fait office de source d'information de l'ED global.

L'idée principale que sous-entend ce modèle est de permettre aux organisations d'avoir un ED centralisé tout en ayant des ED/MD indépendants. La consistance des MD est maintenue et contrôlée par un modèle métier commun et une zone de préparation de données consolidée pour toute l'entreprise.

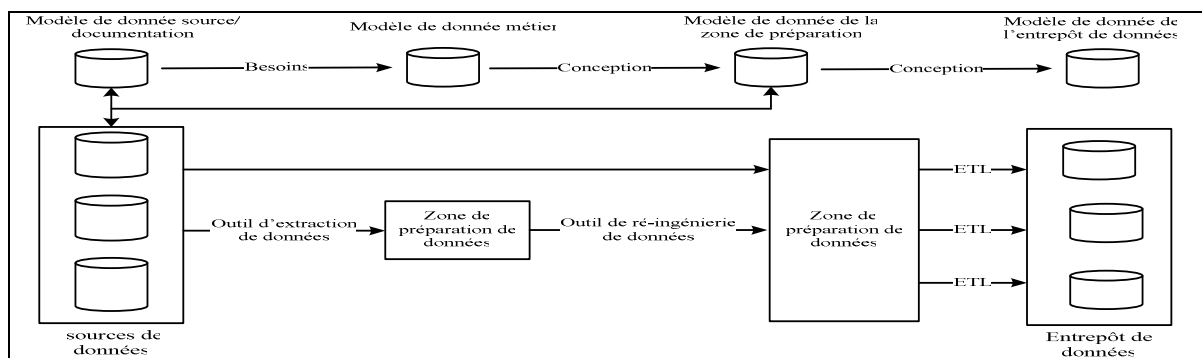


Figure 2.13 : Entrepôt de données fédéré d'après l'approche de White [AK99, CW01]

En résumé, les deux approches présentent quelques similarités. En outre :

- l'approche fédérée implique une zone de préparation de données commune ou un ODS qui réduit la volatilité entre le système opérationnel et l'ED.
- l'approche distribuée implique la reproduction du dépôt de meta-données avec chaque fragment de l'entrepôt, et l'entrepôt entier est administré centralement, contrairement à l'approche fédérée élaborée en vue de remédier à la construction coûteuse d'un seul ED fortement intégré où chaque structure analytique a son propre dépôt de méta données, est administré localement.

2.7. Conclusion

En guise de conclusion, et compte tenu de notre contexte, il est possible de mettre en oeuvre entre les organisations formant le SIIO, cinq architectures d'entrepôts de données [TJ04], à savoir :

- Un entrepôt de données centralisé;
- Un entrepôt de données virtuel;
- Un entrepôt de données distribué;
- Un entrepôt de données fédéré;
- et un entrepôt de données hétérogène.

Dans le cas où les compagnies ne possèdent pas d'ED local avant d'adhérer à un SIIO, donc après leurs intégrations, cela aboutirait à un entrepôt de données centralisé ou virtuel. La conception d'un entrepôt de données centralisé ou virtuel suppose que les compagnies ont des bases de données locales et les utilisent pour construire des entrepôts de données physiques ou virtuels. Force est de constater que les deux approches conceptuelles sont coûteuses et complexes quant à la construction de l'entrepôt de données (l'approche centralisée) et au traitement des requêtes (l'approche virtuelle).

Par contre, si les entrepôts de données locaux existent avant l'adhésion au SIIO, l'approche de conception distribuée ou fédérée peut être rendue effective.

Finalement, si quelques compagnies ont des entrepôts de données, pendant que d'autres disposent des bases de données, l'approche hétérogène peut être utilisée [TJ04].

Par ailleurs, souvent dans le cas des SIIO, les différentes organisations développent leurs ED/MD indépendamment, et il arrive que leur combinaison présente des difficultés dues à l'aspect distribué et hétérogène de ces entrepôts.

A cet effet, la nécessité de les rendre intégrés et interopérables deviendrait rapidement une problématique de taille dans un SIIO, d'où une démarche de conception ascendante qui fait appel à l'intégration des bases de données dimensionnelles s'impose.

Par conséquent, l'intégration des bases de données dimensionnelles ainsi que les critères régissant cette démarche seront évoqués et étayés à travers la méthodologie d'intégration de [LR05] et de [TJ04] - inspirée des travaux de [ZH97]- nous permettant ainsi de dégager l'approche d'analyse, la stratégie d'implémentation et le paradigme de notre démarche à proposer dans un environnement inter-organisationnel et qui sera présentée dans les prochains chapitres.

CHAPITRE 3. Intégration des entrepôts de données distribués et SIIO

3.1. Introduction

Dés qu'on se positionne dans un environnement inter-organisationnel où chaque organisation détient un ED, conçu, élaboré et évoluant indépendamment des autres, la conception d'un ED distribué implique un processus de conception bottom up ou d'intégration.

Au même titre que les bases de données distribuées, un ED distribué peut être soit contrôlé par un seul SGBD il s'agit alors d'un ED réparti, soit distribué parmi un ensemble de SGBD différents qui ont des degrés variables d'autonomie et qui doivent coopérer, dans ce cas, il est question de systèmes multi-ED.

La différence fondamentale entre ces deux types de systèmes concerne le point de départ pour la création de l'entrepôt de données distribué. Dans un ED réparti, la répartition des données est intentionnelle et la conception suit *une démarche ascendante (top down design)*. Par contre, les systèmes multi-ED sont le résultat de l'intégration des ED existants avant la construction du système multi-ED. En d'autres termes, les schémas conceptuels locaux existent et il reste à les unifier dans un schéma conceptuel global permettant ainsi l'accès à l'entrepôt distribué par le niveau local et global, ce qui implique un contrôle décentralisé sur l'entrepôt distribué. Cette démarche suit *une conception ascendante (bottom up design)*.

Cela dit, ce chapitre aura pour objet d'identifier :

- les grandes lignes de la démarche d'intégration des entrepôts de données à travers la spécification de ses différents paramètres de base à savoir l'architecture, le paradigme de développement, le modèle de données, etc.
- les deux approches de conception bottom up des bases de données multidimensionnelles qui sont l'intégration des schémas et la correspondance des schémas.

Ainsi l'ensemble des éléments nécessaires pour la mise en exergue d'un cadre de référence embryonnaire comportant la typologie des SIIO (basée sur les flux d'information) et les paramètres de la démarche d'intégration, seront réunis, souscrivant ainsi les éléments de réponse à la question principale de notre sujet, à savoir, *étant donné un SIIO d'un certain type dont les organisations détiennent des ED autonomes et hétérogènes, comment concevoir un entrepôt intégré et qu'elle est la démarche à adopter pour y parvenir ???*.

3.2. Intégration, concepts et processus

La modélisation des bases de données distribuées et hétérogènes suivant le modèle de référence d'ANSI SPARC est illustrée dans la figure 3.1:

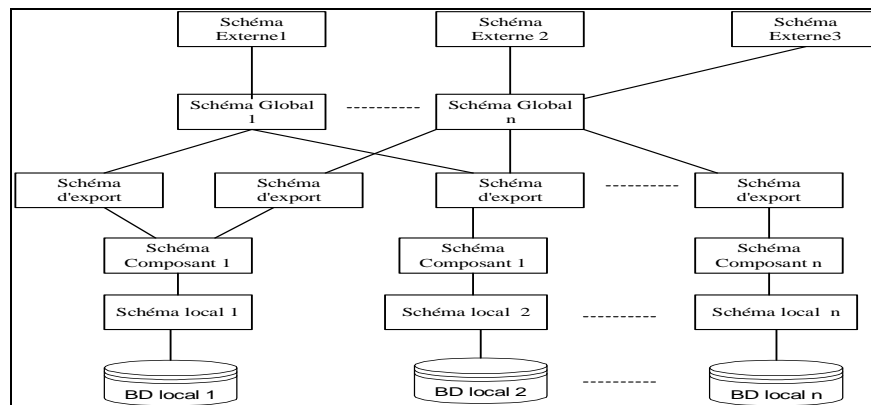


Figure 3.1- Le modèle à 5 niveaux des multibases hétérogènes [MO00].

Le schéma de la base de données distribuée peut en effet concerner des schémas à différents niveaux, local (schéma de chaque SGBD), composant (schéma local traduit dans un modèle commun), exporté (données que chaque SGBD veut partager avec les autres), global (schéma intégré unique ou ensemble de vues sur les schémas exportés) et externe (schéma spécifique pour certaines classes d'applications ou d'utilisateurs) [HGR00].

[MO00, KK90, SL90] modélisent le processus d'intégration des schémas des bases de données hétérogènes, distribuées et autonomes en quatre phases, qui se présentent ainsi : la pré-intégration, la transformation et la comparaison de schémas, la consolidation des schémas et la restructuration des schémas.

Sur la base d'une revue globale de la littérature scientifique relative aux méthodologies d'intégration des schémas des bases de données hétérogènes et distribués, on peut déduire qu'il n'y a pas une méthodologie unique proposée, et nous allons les décrire brièvement dans ce qui suit [PQ05, JP01] :

A. Schéma global : Cette approche consiste à créer une couche constituée d'une vue logique unique pour toutes les bases de données à intégrer.

B. Multi-bases de données : Il s'agit de laisser le plus d'autonomie possible à chacune des bases de données locales. Le système est supposé offrir deux fonctionnalités pour la manipulation et la gestion des données :

- Un langage multi-bases de manipulation de données pour les requêtes et les mises à jour de plus d'une base de données.
- Un langage pour la définition des dépendances de données inter bases.

La combinaison des deux langages est appelée un langage multi-bases de données.

C. Le schéma fédéré : Cette approche, consiste à intégrer plusieurs schémas d'export, contenant les données susceptibles d'être partagées, de chacune des bases de données locales. Ceci permet aux DBA (administrateur de base de données) de créer un schéma importé. Ce schéma décrit les données et les fonctionnalités qui sont accessibles dans la base locale.

L'intégration des schémas de bases de données distribuées exige la comparaison détaillée des schémas. De ce fait la charge de travail dans une comparaison de schéma est proportionnelle au nombre de bases de données qui peuvent atteindre des dizaines ou voire des centaines dans un environnement multi-organisationnel où les bases de données appartiennent aux sites concurrents.

3.3. Démarche conceptuelle d'intégration générique des ED distribués

Intégrer des ED distribués vise principalement à prendre en charge des requêtes de forage transversal ¹ [RA01] exigeant la transformation automatique des requêtes écrites par un ED/MD en une série de transactions, compréhensibles par d'autres ED/MD.

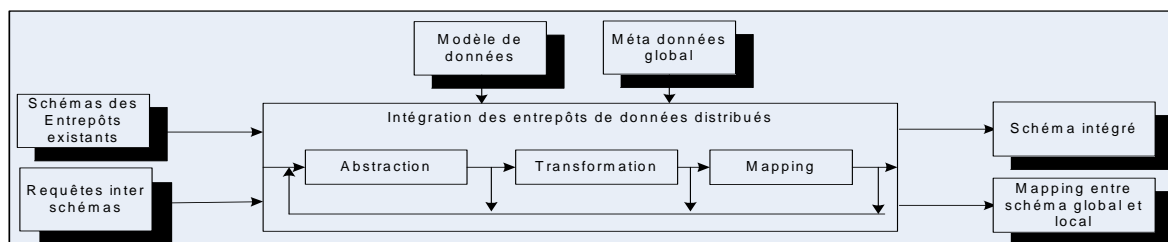


Figure 3.2- Les entrées et les sorties d'un processus d'intégration.

Ce processus exige comme indiqué dans la figure 3.2 une série d'opérations d'abstraction et de transformation de données pour assurer l'uniformité du modèle global unifié.

Abstraction : Elle consiste en une opération de représentation agrégée de relations entre les entités des schémas dimensionnels, ce qui permet de faire ressortir des relations de similitude, de dérivation, de généralisation, de restriction et de spécification entre les dimensions et les faits appartenant aux différents schémas dimensionnels.

Transformation : il s'agit ici de faire appel à plusieurs conversions et transformations pour intégrer les modèles de données des schémas composants, pour l'uniformisation syntaxique et sémantique des schémas dimensionnels existants.

Mapping : après l'abstraction et la transformation, des techniques spéciales de mapping entre les modèles de données hétérogènes existants et le modèle de données cible doivent être exécutées. La composante clé dans cette transformation est le "map" qui indique à l'ED/MD distribué comment les divers ED/MD sont inter-reliés. Cette carte se réfère à un modèle de données global. Il est construit en intégrant les bases de données hétérogènes composantes dans un schéma universel.

Le modèle de données global fournit une représentation virtuelle de toutes les structures de données dans l'entrepôt de données distribué et, avec son propre meta-données, fournit l'épine dorsale pour la transformation des requêtes. Cependant, il convient de noter qu'il contient moins de données que les systèmes locaux qui choisissent de mettre seulement une partie de leurs données disponibles à la disposition de l'entrepôt distribué.

¹ **Requêtes de forage transversales** ou le Drill cross implique l'étude de différentes données appartenant à différents cubes, au même niveau de granularité et contraint par les mêmes conditions pour l'analyse dimensionnelle

Néanmoins, cette démarche générique est soumise à un ensemble de paramètres en raison de la complexité de ce processus dans un environnement inter-organisationnel, car selon [TJ04], il y a deux points importants dans la conception bottom up d'un ED distribué qui préoccupent les partenaires du SIO : **la nature des rapports entre les partenaires** imposant des règles quant à la confidentialité et à la sécurité devant être pris en compte et qui conditionnent **la conception de l'architecture de l'ED global**.

Les dits rapports et d'autres paramètres régissant un processus d'intégration des ED distribué seront présentés dans la prochaine section.

3.4. Critères régissant la démarche d'intégration des ED distribués:

Sur la base des travaux de [CP84, OV99, SRUH99, SL90, TH99] sur l'intégration des systèmes d'information et l'intégration des bases de données, distribuées, hétérogènes et autonomes, nous estimons qu'une méthodologie d'intégration des données structurées en général et dimensionnelles en particulier, dans un environnement SIO, doit prendre en considération les paramètres suivants illustrés dans la figure 3.3 :

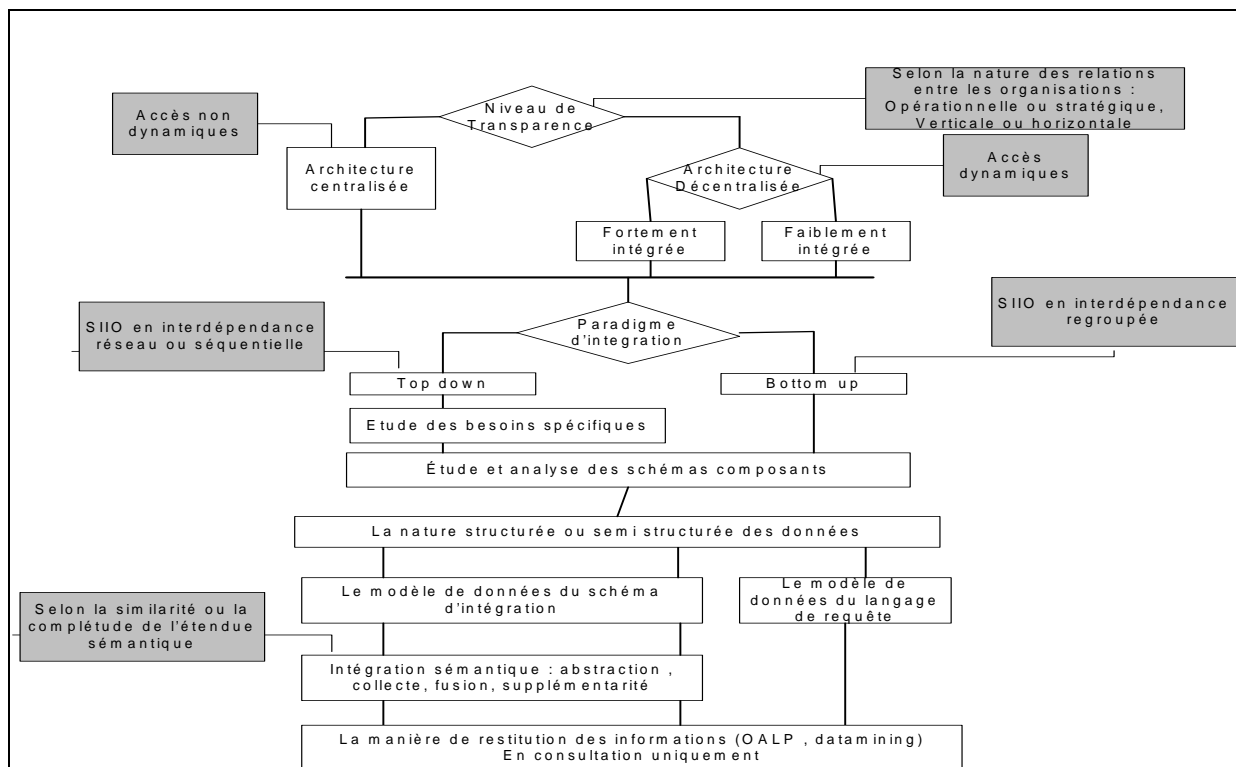


Figure 3.3- Les pré-requis d'une démarche d'intégration des entrepôts de données.

3.4.1. Le niveau de transparence requis

Le but de tout processus d'intégration est de fournir un système intégré pour les utilisateurs qui auront une illusion d'interagir avec un seul système, homogène et consistant. Nous pouvons distinguer les types de transparence suivants :

Transparence de l'emplacement: où l'utilisateur n'a pas besoin de connaître l'emplacement physique des informations. Néanmoins il faut souligner que dans un environnement multi-

organisationnel, les ED/MD sont compétitifs car ils contiennent fréquemment des renseignements semblables, mais servent des intérêts d'affaire séparés. Les utilisateurs des ED compétitifs, ont besoin souvent de connaître les propriétaires des ED pour accéder à des ED sur la base de quelques critères métiers sélectifs. Cette caractéristique « la non transparence de l'emplacement » [ZH97] qui contredit le dogme conventionnel d'un système de gestion de base de données distribué, qui doit rendre l'emplacement de la base de données transparent aux utilisateurs. Donc, le principe de **la transparence de l'emplacement** ne peut pas être appliqué, dans tous les cas.

Transparence du schéma : l'utilisateur n'a pas besoin de connaître sous quel modèle de données sont décrites les différentes structures de données des sources, ce qui permet de masquer l'hétérogénéité logique et qui ne peut être réalisée qu'à travers un schéma intégré. Cependant l'indépendance logique de données est généralement entravée dans un système multi-organisationnel car les ED/MD peuvent changer fréquemment pour deux raisons [TJ04]: (1) les ED sont administrés d'une manière autonome par les différentes organisations, en outre, concurrente, et (2) le nombre des ED est important, et, par conséquent, les changements occasionnés dans chaque ED entraînent plusieurs changements dans les autres ED où une relation métier lie les organisations détentrices d'un ED.

Transparence de langage : l'utilisateur n'aura pas à s'occuper du langage de requête utilisé ainsi que des mécanismes d'accès.

A travers ces trois points on peut déduire ce qui suit :

1. **il y'a une relation évidente entre la résolution du problème de l'hétérogénéité des sources de données et le niveau de transparence que le système intégré doit offrir.** La transparence de schémas masque l'hétérogénéité logique, et la transparence du langage est sollicitée pour masquer l'hétérogénéité d'interface et la transparence de l'emplacement masque l'hétérogénéité technique. Notons que l'achèvement d'une transparence totale n'est pas possible dans tous les cas.
2. **le niveau de transparence requis est intimement lié à la nature des relations entretenues entre les partenaires du SIO.**

Affirmation N°1 : le degré de transparence requis par le processus de consolidation des ED existants au niveau des organisations formant le SIO est à l'origine de la détermination de la nature centralisée ou décentralisée de l'approche d'intégration et de la spécification des méthodes d'accès à l'entrepôt intégré.

3.4.2. Approches d'intégration d'un ED distribué

Le processus de conception bottom up d'un entrepôt de données peut être présenté selon deux approches [OV99] : **décentralisée**, permettant un mode d'accès dynamique aux ED composants et **centralisée** où le mode d'accès est statique. L'intégration peut être différenciée si le stockage des données provenant des schémas composants est persistant au niveau du schéma intégré (matérialisé) ou non (virtuel).

A. L'approche centralisée :

La conception centralisée implique la création d'un entrepôt de données centralisé physique où les données sont maintenues. Les architectures qui matérialisent complètement ou partiellement les sources de données au niveau du schéma intégré ont comme avantages :

- une haute performance des requêtes sur l'ensemble des données matérialisées ;
- le contrôle sur les données matérialisées ;

En contre partie les inconvénients sont:

- la mise à jour des données requiert des procédures complexes ;
- le dépôt du niveau d'intégration doit être muni d'un espace de stockage considérable ;
- elle engendre des difficultés pour atteindre un consensus autour des concepts dans un SIIO.

Comme illustré dans la figure 3.4, cette approche peut être réalisée à travers les stratégies de consolidation suivantes [CAVN05] :

- *Stratégie de migration*, implique le déplacement des structures analytiques existantes sur une seule plateforme remédiant ainsi à l'hétérogénéité technologique uniquement.

- *Recommencer à nouveau (Start from Scratch)* : Construire un nouveau ED/MD au lieu de concevoir ou de fusionner l'existant, cette stratégie est mieux employée quand deux compagnies de taille égale fusionnent.

- *Design un standard et évoluer (Designate and Evolve)* : Désigner un ED/MD existant en tant que standard de corporation et les autres structures analytiques lui sont intégrées, immédiatement ou dans un délais planifié. Cette stratégie est mieux employée quand une plus grande compagnie acquiert une plus petite, donc la fusion ou la coopération stratégique de deux compagnies de taille différente.

- *Alimentation alignée (Backfill)* : Mettre en application une zone de préparation de données commune derrière les ED/MD existants pour consolider les extractions des données. Ainsi l'alimentation des ED/MD s'opère par cette zone. Cette stratégie est conseillée pour les organisations ayant une infrastructure de technologie d'information commune ce qui signifie un environnement de ressource partagé.

- *Synchroniser (Synchronize)* : Synchroniser les ED/MD opérationnels distants à un dépôt central de référence. La synchronisation est conseillée pour de grandes compagnies ayant un nombre d'applications opérationnelles devant être partagées les mêmes données de référence, donc atteindre une gestion et une administration des données de référence centralisée (master data management).

B. L'approche décentralisée

Selon [RZ96], cette approche consiste en la création d'un ED virtuel où les données physiques ne sont pas maintenues dans un entrepôt de données centralisé mais dans les compagnies locales. **L'intégration virtuelle** implique uniquement la matérialisation temporaire des résultats des requêtes pendant le temps de leur exécution. Ce qui implique un mécanisme de translation et de décomposition de requête via le schéma intégré en une ou plusieurs requêtes sémantiquement significatives et

exécutables sur des schémas composants. Les stratégies distribuées se présentent souvent ainsi (figure 3.4) [CAVN05] :

- ED/MD Conforme (Conformed Data Marts). Rendre conforme les modèles de données des ED/MD existants en normalisant les dimensions et les faits communs. Cependant, ladite stratégie peut également être déployée d'une manière centralisée si tous les ED/MD ont des schémas logiques dans une seule instance d'une base de données centrale. Elle est indiquée dans le cas où les organisations adhèrent à un objectif commun ou à une vision globale.
- L'entrepôt des entrepôts (Mart of Marts). Créer une vue d'entreprise à travers des ED/MD en y extrayant des données pour en créer un nouveau ED/MD.
- Requête Distribuée (Distributed Query). Créer une vue d'entreprise en recensant l'ensemble des requêtes fréquentes posées aux multiples ED/MD et la réconciliation s'obtient à la volée.

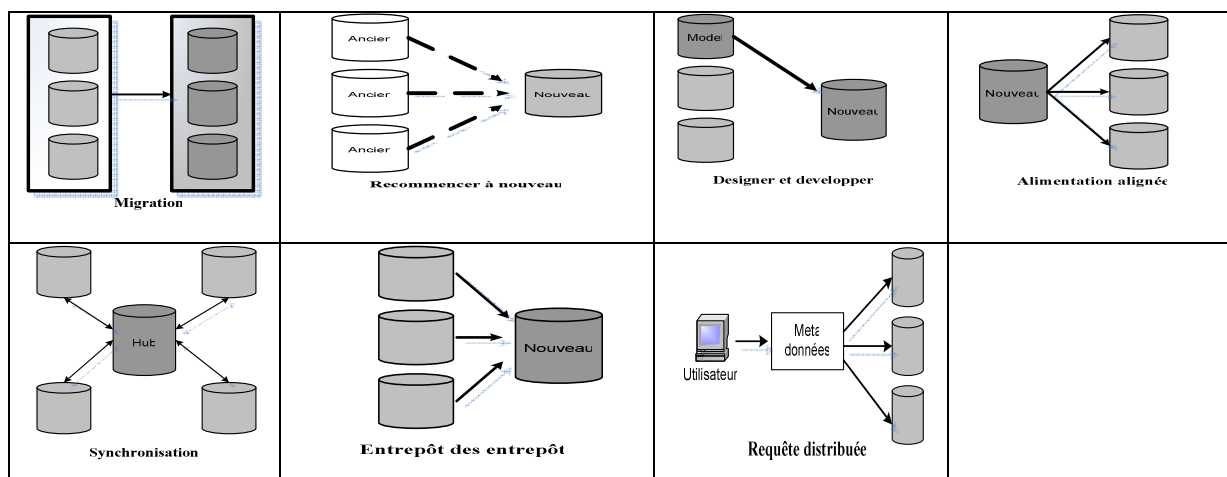


Figure 3.4 - Stratégie d'intégration des ED/MD [CAVN05]

Affirmation N°2 : Il existe une corrélation entre le niveau d'intégration exigé et le temps alloué au projet d'intégration, ce dernier facteur est intimement lié à la pérennité de la relation entretenue entre les organisations formant le SIIO à savoir stratégique ou opérationnel. A cet effet, les stratégies dont le déploiement est rapide -migration, l'entrepôt des entrepôts et les requêtes distribuées- impliquent un niveau d'intégration faiblement couplée. Réciproquement, des stratégies exigeant un niveau d'intégration fortement couplée en faisant usage d'un schéma global particulièrement "Recommencer à nouveau" prennent plus de temps à se déployer. Il pourrait y avoir une troisième approche hybride qui matérialise uniquement un sous-ensemble des données fréquemment consultées et qui correspondent aux données de références des organisations formant le SIIO.

3.4.3. Nature fortement ou faiblement couplée de l'intégration :

Le degré de couplage et de cohésion des schémas composants sélectionnés pour une solution d'intégration est un facteur prédominant dans la conception d'une solution maintenable facilement.

Le couplage est un concept bien connu dans le domaine d'ingénierie de logiciel et concerne le degré d'interdépendance entre les composants logiciels. Donc plus le niveau de dépendance inter-schémas est élevé, plus haut est le niveau de couplage, et par conséquent, plus bas est le niveau

d'autonomie des schémas individuels. La cohésion d'autre part, appartient au degré dans lequel les schémas composants (par exemple, des opérations ou des tâches organisationnelles) sont rapprochés l'un de l'autre [TH99]. A cet effet dans le cas d'une architecture centralisée le degré du couplage est intimement liée à l'intégration des schémas ou des sous schémas (vues) composants et dans le cas d'une architecture décentralisée :

L'intégration décentralisée fortement couplée implique un schéma global unifié couvrant l'étendue sémantique complète ou partielle des schémas composants, et à ce stade là, la résolution de l'hétérogénéité logique des schémas composants, lors de traitement des requêtes, est assurée par un schéma de correspondance (meta données logique pour assuré l'équivalence sémantique).

Par contre **l'intégration décentralisée faiblement couplée** ne requiert qu'un langage unifié (langage de requête multi bases) pour interroger les données des schémas composants, ce qui implique que cette configuration n'est appliquée que pour les schémas composants offrant un langage de requête permettant à l'utilisateur d'exprimer ces requêtes et qui doit prendre en charge la résolution des problèmes d'hétérogénéité schématiques des composants.

Affirmation N°3 : Pour réaliser les niveaux maximaux de réutilisation et de maintenabilité du schéma intégré, les schémas composants devrait être faiblement couplé et cohésif. Cependant si on se base sur le niveau de transparence à fournir, l'intégration décentralisée fortement couplée offre la transparence de schémas, de langage et d'interface et l'intégration décentralisée faiblement couplée n'offre que la dernière option de transparence.

3.4.4. Le paradigme de la démarche d'intégration :

On peut distinguer deux tendances d'ingénierie des systèmes intégrés fortement couplés, (matérialisé ou virtuel) à savoir (figure 3.5):

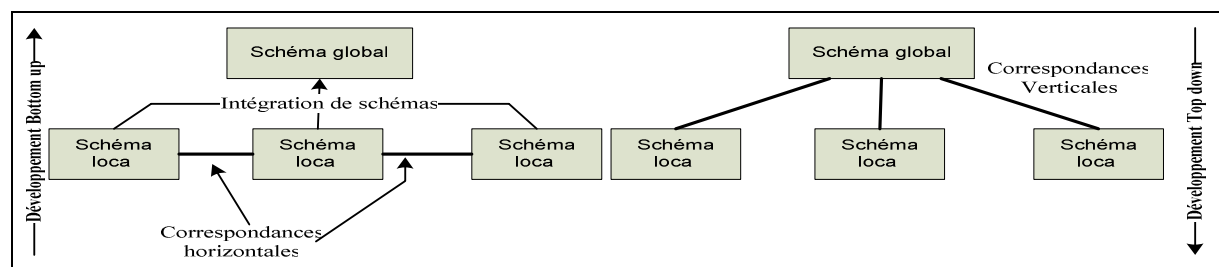


Figure 3.5- Paradigme d'intégration [SRUH99].

Bottom up : Construire un système fortement intégré par cette approche signifie que le besoin important est d'avoir un accès intégré sur un ensemble de sources de données. Un scénario typique, est le besoin d'avoir un accès uniforme et détaillé (exact et complet) sur tous les ED du SIO pour la construction d'une application globale. Dans ce cas, le besoin essentiel est de garantir la complétude et l'exactitude du schéma intégré, impliquant ainsi des techniques d'intégration (semi) formelles. Cette approche conduit à une meilleure intégration sémantique du système car les schémas composants sont analysés et étudiés avant de commencer le processus d'intégration. Le changement de configuration implique le déclenchement de nouveau processus d'intégration [CP84, OV99].

Top down, cette approche commence d'abord par la construction du système intégré selon les besoins globaux ou spécifiques des utilisateurs en informations, ensuite couvrir les sources de données pouvant subvenir à ces exigences. Si on considère les principes d'intégration des schémas qui sont **la complétude, l'exactitude, la compréhension et l'aspect minimal**, deux de ces principes ne seront pas inclus dans cette démarche à savoir :

Premièrement, **la non complétude**, car il n'est pas indispensable d'inclure complètement les schémas si des données spécifiques uniquement sont requises par l'utilisateur.

Deuxièmement, **la non exactitude**, car le besoin de représenter les données au niveau global exactement comme elles le sont au niveau des schémas composants n'est pas nécessaire, puisque seules les données dérivées ou récapitulatives sont requises.

Le schéma global dans cette approche peut être généré d'une façon ad-hoc ou selon un processus d'analyse formel, commençant par la description des cas d'utilisation par exemple et aboutissant aux techniques d'intégration de vues. Dans ce cas, les schémas composants ne sont considérés que lors de la seconde étape où les correspondances entre le schéma global et les schémas sources sont établies afin de permettre la transformation des requêtes [SRUH99].

Affirmation N°4 : Le paradigme de la démarche d'intégration doit prendre en considération les points suivants :

- *Les entrepôts composants évoluent indépendamment ;*
- *La difficulté de l'intégration des schémas dans un contexte inter-organisationnel car ce processus est vulnérable aux changements;*

Cela nous permet d'inscrire l'approche top down comme le paradigme d'intégration parfait - qui est moins fortement couplée par rapport à l'approche bottom up- dans le contexte inter-organisationnel où le partage d'information n'est permis que pour accéder aux informations relatives aux processus métiers transcendant les frontières des organisations formant le SIIO.

Cependant il reste à prendre en charge les besoins des utilisateurs qui changent fréquemment, par conséquent une approche hybride mariant l'aspect top down caractérisé par sa focalisation sur les exigences des utilisateurs et bottom up caractérisé par sa complétude et son exactitude doit être envisagée.

3.4.5. Le modèle de données de l'intégration

Sur la base de la nature des données à intégrer (structurée ou semi structurée) , ladite intégration se base sur un modèle de données canonique ou commun, qui spécifie le modèle du schéma intégré lors de l'intégration fortement couplée, et la nature du langage de requête lors de l'intégration faiblement couplée

Affirmation N°5 : le modèle de données restreint le type de composants à intégrer, à cause des possibilités restreintes de translation entre quelques modèles de données tels que :

- *L'intégration d'un modèle de données dimensionnel semi structuré vers un modèle structuré ;*

- *Intégration d'un modèle sémantiquement riche à un modèle de données sémantiquement pauvre.*

Dans le contexte inter-organisationnel où les plateformes technologiques à base d'Internet ou d'Extranet sont largement exploitées, les modèles de données pivots semi structurés à base d'XML s'inscrivent comme la solution actuelle.

3.4.6. Le procédé d'intégration sémantique

L'intégration de données fortement couplée encapsule les sources de données pour l'utilisateur. Contrairement à l'intégration faiblement couplée qui laisse les sources de données en place, l'intégration fortement couplée induit une meilleure intégration sémantique des données et qui peut être effectuée à travers plusieurs procédés :

La collecte : qui consiste en l'assemblage des données des schémas composants sans apporter des modifications et sans élaboration d'un matching entre les différentes entités des schémas composants.

La fusion : l'intégration des données est faite à travers une extraction des données à partir des sources, exprimées à travers le requêtage des schémas composants et sans élaborer des calculs d'abstraction. Cependant, la fusion des données est accomplie pour identifier l'équivalence sémantique entre les entités des schémas composants, et renforcer la consistance des données.

L'abstraction : à ce niveau, les données intégrées sont extraites à partir des schémas composants. En outre d'autres fonctions doivent être appliquées sur les sources de données pour les agréger ou les détailler vers le niveau d'abstraction du schéma intégré. La nécessité de ce procédé est induite par les conflits sémantiques entre les schémas composants, de plus, il englobe des fonctionnalités d'agrégation de données, de re-classification d'entités, ou d'autres processus complexes, ce qui rend les opérations d'écriture et les opérations rétro impossibles à effectuer.

Affirmation N°6 : la consolidation des entrepôts de données des organisations intégrées horizontalement permettent souvent une intégration par collecte ou par fusion des schémas vue la similarité sémantique de leurs données, à l'encontre des organisations intégrées verticalement ne permettant une intégration qu'à travers une abstraction des schémas en raison de la nature complémentaire de leurs données.

En conclusion, les principes du processus d'intégration des entrepôts de données hétérogènes abordés succinctement indique qu'une démarche de consolidation des entités analytique à évolution indépendante est tributaire des relations métiers et de la complexité des échanges entre les organisation du SIIO, et qui conditionne l'aspect architectural et conceptuel de l'intégration des données, au quelle nous allons mettre en relief à travers la revue d'un processus d'intégration de schémas de [LR03, LR04, LRO5] et de correspondance de schémas de [ZH94, TJ04].

3.5. Intégration de schémas

Ce sujet a été étudié par [KIM96] dans le contexte de la conception de l'entrepôt de données. Dans son livre [KIM02], il a identifié le problème et a introduit le concept informel de dimension et de fait conforme, car selon son approche, les entrepôts de données sont développés selon une modélisation

dimensionnelle, comme une série de MD cohérents. Chaque MD fournit une vue dimensionnelle d'un seul processus d'affaire et un entrepôt de données intégré ne peut être construit qu'à travers des dimensions et des faits conformes (c.-à-d., communs).

Cependant, dans les grandes compagnies, les différents départements développent souvent leurs ED/MD indépendamment, et il arrive que leur combinaison présente des difficultés.

En réalité, le besoin d'intégrer des ED/MD indépendamment développés survient dans d'autres cas, par exemple, quand une compagnie absorbe une autre ou elle est impliquée dans un projet fédéré.

Contrairement au problème général d'intégration des bases de données, l'intégration des ED/MD peut être abordée d'une manière plus systématique pour deux raisons importantes [LR05] :

- En premier lieu, tout ED/MD est structuré dans un formalisme plutôt constant, par les notions de dimension et de fait.
- En second lieu, la qualité des données dans un ED/MD est habituellement plus raffinée que dans une base de données générique, car elles sont obtenues en nettoyant et en réconciliant plusieurs sources de données.

Ces observations suggèrent que le problème d'intégration des ED autonomes peut principalement se restreindre à l'intégration des faits et des dimensions indépendamment développés.

Selon [LRO5] la mise en évidence d'une méthodologie d'intégration des schémas des ED/MD devrait comprendre les phases suivantes (fig3.6) :

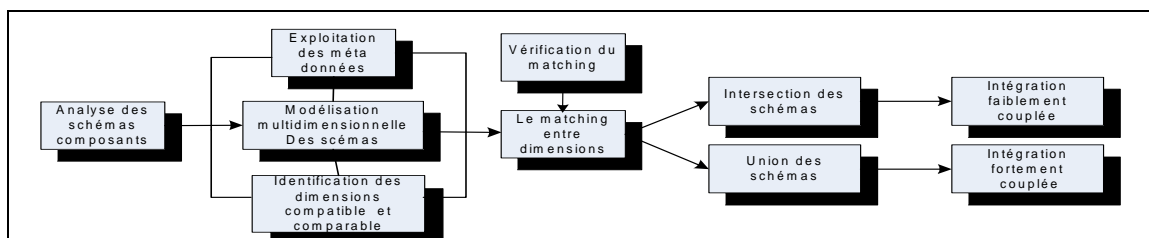


Figure 3.6 - Processus d'intégration des schémas multidimensionnel d'après [LRO5].

- La première phase consiste en l'accès et l'analyse des ED/MD existants et emmagasinés dans les différents systèmes (DB2, Oracle, SQL Server, ...);
- La deuxième phase est l'extraction des métadonnées de description des ED/MD (fait et dimensions) de ces systèmes et transformer cette description en un modèle multidimensionnel pour identifier où et comment leurs dimensions sont compatibles ou conformes [LR03, LRO4, LRO5].
- La spécification et l'identification des matching¹ entre les dimensions autonomes, le matching sémantique des dimensions compatibles est vérifié pour s'assurer que leur jointure est sensée;
- Test des matching pour vérifier l'aspect cohérent, saint et consistant du matching;
- Générer l'intersection entre les dimensions selon une approche faiblement couplée;

¹ Matchnig : une correspondance un à un entre les niveaux des dimensions

- Et fusionner deux dimensions selon une approche d'intégration fortement couplée.

L'intégration des schémas est une méthodologie d'intégration des structures multidimensionnelles autonomes et distribuées à travers l'intégration des vues, basée sur un mécanisme d'extraction des données en utilisant l'algèbre de dimensions permettant de rendre les dimensions compatibles. De ce fait, les notions suivantes doivent être abordées en vue d'introduire la notion de compatibilité des dimensions qui étend le concept de conformité de dimensions de kimball, à savoir

- l'algèbre de dimensions pour l'extraction ou l'abstraction des dimensions en vue de les rendre comparables.
- et le matching entre les dimensions comparables.

3.5.1. Algèbre de dimensions

Notons d une dimension ayant pour schéma $(L, \leq d)$ et instance (m, ρ) . L'algèbre dimensionnelle (AD) peut être utilisée pour la manipulation des dimensions et se base sur trois opérateurs, qui sont [LR05]:

La Sélection : Notons G un sous ensemble des membres basique (granularité la plus détaillée) de d . la sélection $\sigma G(d)$ de G sur d est une dimension d' tel que : (i) le schéma de d' est le même schéma que d et (ii) l'instance de d' contient : les membres basiques dans G , les membres de d pouvant être atteint par l'application d'une fonction de *roll up*¹ dans ρ , des membres de G et la restriction des fonctions *roll up* de d vers les membres de d' .

La Projection : Notons X est le sous ensemble des niveaux de d incluant $\perp d$ (*éléments basiques*). La projection $\pi X(d)$ de d sur X est la dimension d' tel que : (i) le schéma de d' contient X et la restriction de $\leq d$ aux niveaux contenus dans X , (ii) l'instance de d' contient : les membres de d appartenant aux niveaux dans X et les fonctions *roll up* $\rho l_1 \rightarrow l_2$ de d impliquant les niveaux dans X .

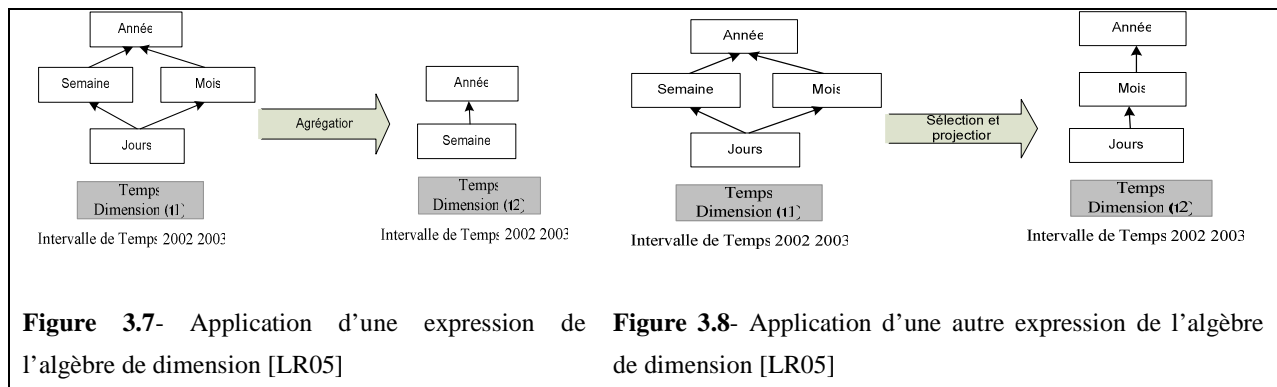
L'agrégation : Notons l un niveau dans L . l'agrégation $\psi l(d)$ de d sur l est la dimension d' tel que : (i) le schéma de d' contient l , les niveaux de d lesquels l rolls up, et la restriction de \leq à ces niveau, et (ii) l'instance de d' contient : les membres de d appartenant aux niveaux de d' et les fonctions *roll-up* $\rho l_1 \rightarrow l_2$ de d impliquant les niveaux dans d' .

Pour une expression E de l'AD et la dimension d , nous la dénotons par $E(d)$ qui est la dimension obtenue par l'application de E sur d .

Exemple sur la base de l'exemple des ED/MD ventes et inventaires sus cités dans le chapitre 2, la dimension temps t_2 du ED/MD inventaire des entrepôts peut être calculée par l'application d'une agrégation de la dimension temps t_1 du ED/MD des ventes comme suit (Figure 3.7): *ψweek* (d).

De même, la dimension temps t_2 du ED/MD de l'inventaire des entrepôts peut être calculée par l'application d'une sélection suivie par une projection de la dimension temps t_1 du ED/MD des ventes, comme suit (Figure 3.8): *Π day, month, year, day_month, month_year* ($\sigma O_{2002-2003}$ (d)), où $2002-2003$ dénote les jours appartenant aux années 2002-2003.

¹ **Roll up** : un opérateur dimensionnel d'agrégation



En effet, une expression de l'AD à pour but de calculer un sous ensemble d'une dimension. Notons que la projection et l'agrégation ont différents buts, pendant que la projection prend toujours le niveau bottom, l'agrégation l'efface.

Expression Lossless : une expression de l'AD sur la dimension d est *lossless* si pour chaque membre o dans $E(d)$, tous les membres qui se roll up à o dans d appartient à $E(d)$. De ce fait les expressions de l'AD impliquant uniquement les projections et les agrégations sont toujours *lossless*.

Nous pouvons à ce stade énoncer le concept d'intégration des dimensions autonomes et identifier un nombre de propriétés du matching entre les dimensions (Une correspondance un à un entre les niveaux des dimensions).

3.5.2. Matching des dimensions autonomes

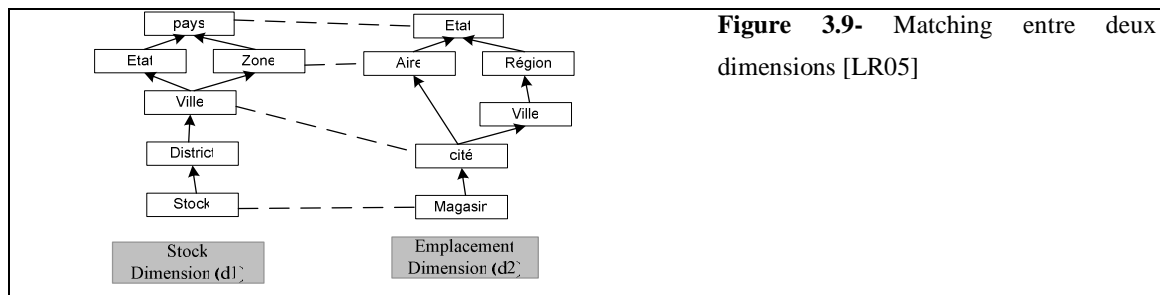
Dans ce qui suit, $d1$ et $d2$ dénotent deux dimensions, appartenant aux différents ED/MD, ayant pour schéma $S(di) = (Li, \leq i)$ et instance $I(di) = (mi, \rho i)$, respectivement [LR05].

Le matching de dimensions : le matching μ entre deux dimensions $d1$ et $d2$ et un mapping injectif partiel¹ entre $L1$ et $L2$.

Les propriétés du matching : Un matching total est cohérent, exact et consistant. Notons que la cohérence signifie la préservation de l'ordre, l'exactitude signifie la préservation de l'ensemble des membres, et la consistance signifie la préservation de la fonction d'agrégation (roll up).

Exemple la figure 3.9 montre un exemple de matching entre deux dimensions géographiques associant un dépôt de stock avec un magasin, la ville avec la cité, la zone avec une aire et le pays avec l'état. Le matching est logique et cohérent. Si les niveaux mappés ont les mêmes membres il est exact .la consistance suit dans le cas de l'équivalence de la fonction d'agrégation entre les niveau. Si les niveaux tracés ont les mêmes membres il est également exact.

¹ La fonction f est dite *injective* si deux éléments x et y de l'ensemble de départ ont toujours deux images distinctes dans l'ensemble d'arrivée et d'ordre partiel (une relation d'ordre est une relation réflexive, transitive et antisymétrique. Si R est une relation d'ordre sur E , on dit que R est une relation d'ordre complet si deux éléments quelconques de E sont comparables sous R . Sinon, on parle de relation d'ordre partiel).



Evidemment le matching total est difficile à effectuer dans la réalité. Néanmoins dans plusieurs cas, les dimensions autonomes actuellement partagent quelques informations. Pour identifier ces informations, nous avons besoin de sélectionner un sous ensemble des deux dimensions, d'où la compatibilité des deux dimensions est [LR05] :

3.5.3. Compatibilité de dimensions

Deux dimensions $d1$ et $d2$ sont compatibles s'il y'a deux expressions de l'AD lossless $E1$ et $E2$ sur $d1$ et $d2$, respectivement, tel que le matching parfait entre $E1(d1)$ et $E2(d2)$ est vérifié dans ce cas nous pouvons dire que $d1$ et $d2$ sont compatible en utilisant $E1$ et $E2$. Le raisonnement sous-tendant la définition de la compatibilité qui est : (i) deux dimensions peuvent avoir des informations communes suivant une approche conceptuelle sous laquelle deux niveaux coïncident s'ils sont peuplés par les mêmes entités réelles du monde. Dans une approche logique cette notion serait basée sur une bijection entre les identifiants des entités. (ii) l'intersection peut être identifiée par des expressions DA; et (iii) les expressions lossless garantissent l'exactitude des opérations OLAP sur l'intersection [LR04].

Exemple : le matching dimensionnel rapporté à la figure 3.9, peut être rendu parfait par l'application des opérations suivantes à $d1$ et $d2$:

- π magasin, cite , aire , état ($\sigma m1(\text{magasin}) \cap m2(\text{stock})(d1)$),
- Π stock, ville, zone, pays ($\sigma m1(\text{magasin}) \cap m2(\text{stock})(d2)$).

Sur la base de la compatibilité entre dimension, on peut procéder à l'intégration des ED/MD selon deux approches différentes.

3.5.4. Intégration faiblement couplée

Le support des requêtes transversales entre les ED/MD, signifie l'identification de l'intersection entre leurs dimensions. Actuellement [LR04] propose de vérifier la qualité de telles intersections par la spécification de la compatibilité de dimensions.

Exemple : si on veut intégrer le ED/MD des ventes dans la figure 3.10 avec le ED/MD de la situation météorologique rapporté dans la figure 3.11, dans le but de corrélérer les ventes avec les conditions météorologiques. L'intégration entre ces sources de données peut être basée sur le matching entre le temps ($t1$ et $t2$) et la dimension de la localisation géographique ($s1$ et $sm2$) comme mentionnée dans la figure 3.12.

Cette démarche vérifie la compatibilité des dimensions en entrée et retourne la paire de fonctions suivantes, les deux premières expressions sélectionnent les membres communs dans la dimension temps

- Π jour, mois, an ($\sigma \text{jour}_{t1} \cap \text{jour}_{t2}(t1), (I)$)

- $\Psi_{jour} (\sigma_{jour_{t1} \cap jour_{t2}} (t2)). (2)$

Où $jour_{t1} \cap jour_{t2}$ dénote les jours en communs permettant d'effectuer le matching parfait entre $t1$ et $t2$. L'autre paire d'expression sélectionne les membres communs dans la dimension localisation :

- $\Psi_{cité} (\sigma_{cité_{s1} \cap cité_{sm2}} (s1)). (3)$
- $\Psi_{cité} (\sigma_{cité_{s1} \cap cité_{sm2}} (sm2)). (4)$

Où $cité_{s1} \cap cité_{sm2}$ dénotes les villes en commun permettant d'effectuer le matching parfait entre $s1$ et $sm2$.

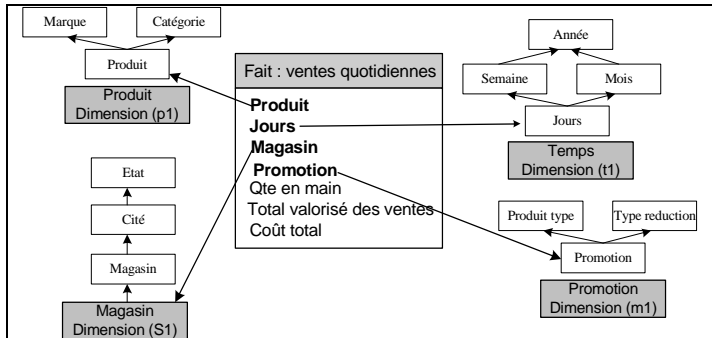


Figure 3.10 ED/MD des ventes [LR04]

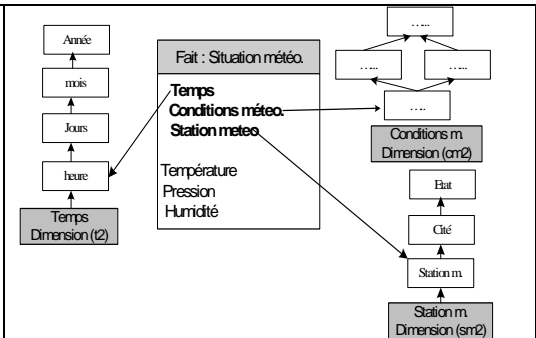


Figure 3.11-ED/MD de la situation météo [LR04]

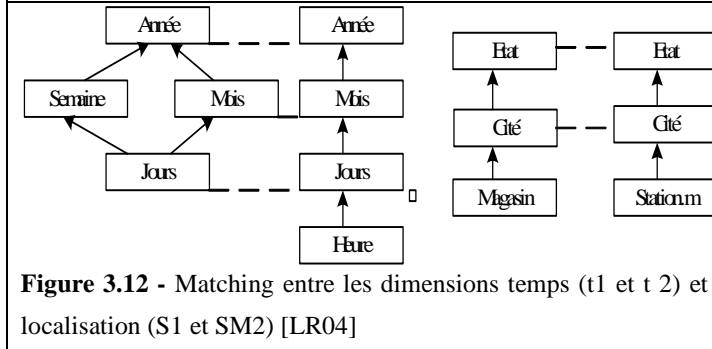
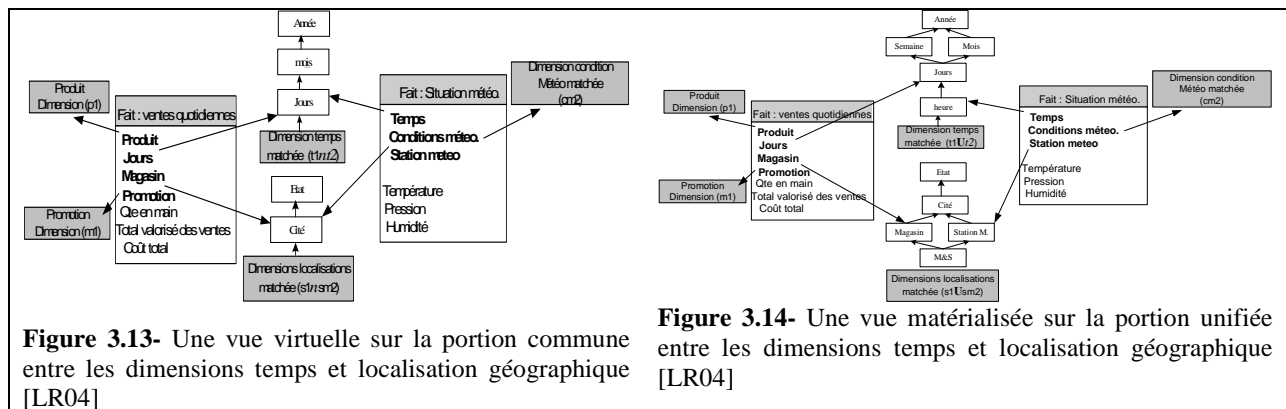


Figure 3.12 - Matching entre les dimensions temps (t1 et t2) et localisation (S1 et SM2) [LR04]

De ce qui précède, il ressort que les deux MD peuvent être reliés pour extraire les données quotidiennes et les données de villes mais les données horaires et les données des magasins ne peuvent pas être calculées. Par conséquent, si nous appliquons les expressions (1,2,3,4) aux dimensions avant l'exécution du drill cross nous empêchons des inconsistances dans le calcul des agrégations sur les résultats de la jointure [LR04].

Il s'en suit que l'opération du drill cross peut être définie sur la vue virtuelle représentée sur le schéma 3.13. *Dans un scénario d'intégration faiblement couplée, nous avons besoin d'identifier les informations communes entre les différentes sources de données pour effectuer des requêtes transversales sur les sources originales. De ce fait, à travers une paire de dimensions d1 et d2 et le matching μ entre elles, l'approche vise à la dérivation de deux expressions rendant μ parfait.*



3.5.5. Intégration fortement couplée

Elle vise à combiner des données appartenant à plusieurs dimensions par le calcul de leurs unions au lieu de leurs intersections et d'établir une vue matérialisée combinant différentes sources de données et exécuter des requêtes au-dessus de cette vue. Le but est la dérivation de nouvelles dimensions obtenues en fusionnant les dimensions des sources originales. Ce qui peut être utile dans le cas de la réconciliation et de la fusion de deux ED/MD indépendamment développés.

En se référant encore à l'exemple (figure 3.10 et 3.11), si nous procédons par l'intégration fortement couplée sur les dimensions temps et localisation et le matching illustré dans la figure 3.10, nous générons deux nouvelles dimensions pouvant être matérialisées et utilisées par les deux ED/MD. Nous pouvons nous référer au schéma de la figure 3.14 pour opérer des requêtes drill across [LR04].

3.5.6. Synthèse :

A travers l'exposé de la démarche d'intégration des schémas, nous pouvons observer la nature bottom up du paradigme et l'aspect complétude de ce processus. Néanmoins, dans le cadre du travail exigé, à savoir l'élaboration du processus d'intégration des schémas dimensionnels autonomes dans un environnement inter-organisationnel, il est nécessaire de tenir compte des faits ci-après :

- Le nombre des ED/MD dans un environnement inter-organisationnel est important et assez large pour qu'on puisse l'intégrer dans un seul schéma ce qui nécessite beaucoup d'efforts.
- Le meta-données des ED/MD est en évolution constante et continue, donc le meta-données global doit être mis à jour fréquemment.
- Plusieurs ED/MD sont similaires et compétitifs et leur identité doit être maintenue et rendue explicite pour l'utilisateur.

Dans ce sens l'autre variante du processus d'intégration des schémas qui permet de remédier au problème de la non transparence de l'emplacement et l'aspect dynamique des schémas est la correspondance de schémas développée pour le support des requêtes coopératives ou transversales dans les systèmes décentralisés fédérés que nous allons exposer par la suite.

3.6. Correspondance de schémas des ED

Zhao [ZH97] propose la coordination des schémas au lieu de l'approche d'intégration de schémas, pour la conception ascendante des bases de données distribuées en établissant une correspondance entre les attributs composants et les attributs fédérés. La solution de Zhao est basée sur une " **matrice de**

correspondance d'attribut (MCA) " qui est une table plate qui fait correspondre tous les attributs dans la fédération à un ensemble d'attributs génériques. En plus de la MCA, les schémas d'export des ED/MD composants sont utilisés pour la spécification logique de leurs structures de données individuellement. La MCA et les schémas d'export représentent les meta-données de la fédération utilisées pour le traitement des requêtes coopératives. Le concept de la MCA utilise la notion de relation universelle¹ comme concept de base (fig3.15).

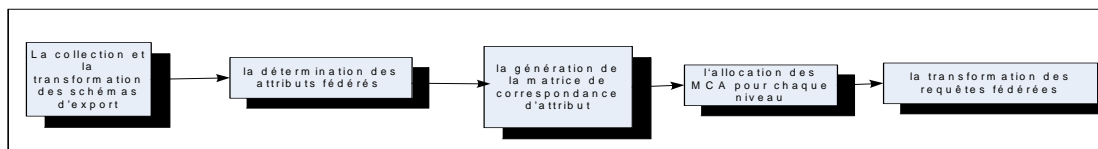


Figure 3.15- Processus de correspondance des schémas d'après [TJ04 , ZH97].

L'entrepôt de données distribué utilise un schéma de coordination multi niveau (MSC) pour intégrer les ED locaux. Avec cette approche, un format d'ED largement accepté a besoin d'être proposé avant la conception de l'ED distribué. Le MSC, qui est une modification de la MCA, est utilisé en considérant les noeuds du SIIO disposés en couches. Par conséquent, le MSC a pour objet de maintenir des relations en couches et la sécurisation des données entre les partenaires.

Les partenaires, tels que les fournisseurs des matières, les fabricants des produits, les grossistes, et les détaillants, forment l'environnement inter-organisationnel, (dans le même niveau) et ils sont supposés être des concurrents. Ces derniers n'ont pas le droit d'accéder aux données mutuellement, et le MSC alloue un ACM aux noeuds locaux. Construire un MSC implique cinq phases [TJ04] :

(1) **la collection et la transformation des schémas d'export**: en premier lieu, chaque ED détermine la portion de son schéma local à exporter pour la fédération². La coordination de schéma n'intègre pas les schémas d'exportation dans un seul schéma intégré; elle maintient tous les schémas d'export dans leur format.

Le schéma d'export a trois colonnes (**attribut, type, et indicateur**): (1) noms d'attributs comme ils apparaissent dans l'ED locale; (2) types d'attribut (clé, non clé et clé étrangère); et (3) pointeurs d'attribut qui représentent la structure logique de l'ED. Pour construire correctement les pointeurs d'attribut, un attribut non principal doit pointer sur son attribut principal ou clé et un attribut clé et principal pointe sur ses clefs étrangères éventuelles.

(2) **la détermination des attributs fédérés** : A ce niveau, La première tâche consiste à déterminer l'ensemble d'attributs qui englobent tous les attributs dans la fédération. L'équivalence

¹ **Relation universelle fédérée**: la relation universelle est un concept et un cadre de référence, comme concept elle se réfère à une super relation comprenant tous les attributs dans une base de données. Et comme cadre de référence d'implémentation la relation universelle fournit un langage d'interrogation de plus haut niveau pour rendre la base de données plus accessible en permettant à l'utilisateur de mentionner seulement des noms d'attribut d'intérêt pour la question

² Cette phase est semblable à l'approche de l'ED virtuel mais le schéma d'export d'une MSC est le schéma d'export des ED locaux au lieu de bases de données locales.

parmi tous les attributs est réalisée en appliquant le concept de l'attribut sémantique¹, une fois que les attributs sémantiques uniques sont identifiés dans les schémas d'export et les attributs fédérés sont nommés. L'ensemble des attributs fédérés constitue la relation universelle fédérée. Tous les attributs des schémas d'export de la fédération sont considérés comme faisant partie du schéma dimensionnel global (en étoile par exemple).

(3) **la génération de la matrice de correspondance d'attribut (MCA)** : Une fois les attributs fédérés sont déterminés, on peut facilement trouver la correspondance entre les attributs fédérés et ceux des ED composants en comparant leur sémantique. La relation universelle fédérée et tous les attributs composants peuvent être placés dans une table plate, qui est désignée sous le nom de la matrice de correspondance d'attribut (MCA) qui est définie par (1) une matrice, dont les lignes sont destinées pour chaque attribut de la fédération et les colonnes pour chaque ED composante et (2) insérer chaque attribut composant dans la cellule résultante de l'intersection de la ligne de l'attribut de la fédération et de la colonne de son ED composant.

Cette liste d'attributs sémantique et de noms fédérés devrait être procurable et basée sur un accord entre des membres de la fédération. Deux approches possibles peuvent être utilisées dans le MSC: une MCA complète (tous les noeuds maintiennent la même MCA) ou une MCA partielle (les noeuds conservent différentes MCA). La MCA contient tous les attributs d'exports d'un ensemble d'ED locaux en correspondance avec le schéma de la fédération (attributs fédéré). Une MCA partielle permet aux différents noeuds d'avoir différentes MCA incluant seulement les attributs des entrepôts de données de niveau supérieurs ou/et inférieurs. Cette conception maintient une MCA de petite taille et empêche la révélation des informations aux concurrents. La conception distribuée adopte une approche de MCA partielle. Dans cette approche les attributs fédérés du schéma global sont définis dans les MCA locaux. La MCA au niveau du nœud *ij* est composée d' :

- attributs étoile fédérés du schéma global et ceux du nœud *ij*;
- attributs de l'ED du nœud *ij*; et
- attributs étoiles fédérés des noeuds connectées au niveau *i-1*.

Ce qui donne, $MCA_{ij} = MCA[F, f_i, ED_{ij}, f_{(i-1)1}, f_{(i-1)2}, \dots,]$, ou $i = 1$, $MCA_{i1} = MCA[F, f_{i1}, ED_{i1}]$, et :

M	Est le niveau supérieur de l'environnement SIIO, l'inférieur est 1
N_i	Le nombre de nœud dans le niveau <i>i</i>
$Node_{ij}$	Est le nœud <i>j</i> au niveau <i>i</i>
$MCA[]$	Est la matrice de correspondance d'attribut
ED_{ij}	Est l'ED du nœud <i>j</i> au niveau <i>i</i>

¹**Attribut Sémantique:** Ceci indique que différents noms d'attribut peuvent être choisis pour le même attribut sémantique. En conséquence, les conflits nommés se produisent dans les noms d'attribut des bases de données hétérogènes, pas sur la sémantique d'attribut. Des conflits de nom peuvent être résolus en examinant la sémantique de l'attribut. Un attribut universel fédéré devrait être nommé selon son attribut sémantique, qui devrait être rendu évident aux utilisateurs de base de données.

ED_{ij}	Est l'attribut de l'ED du noeud j au niveau i
F_{ij}	Est l'attribut de l'étoile fédéré du $noeud_{ij}$ et les noeuds connectés au niveau $i-1$, i.e. $f_{ij} = \{ED_{ij} \cup f_{(i-1)1} \cup f_{(i-1)2} \cup \dots \cup f_{(i-1)n_{(i-1)}}\}$
F	Est l'attribut étoile fédéré du schéma étoile global
ACM_{ij}	Est l'ACM du $noeud_{ij}$

(4) **l'allocation d'une MCA.** La séquence d'allocation des MCA partielles aux noeuds peut être menée de façon ascendante ou descendante, néanmoins la séquence doit être conforme à la direction de passage des requêtes dans le but de maintenir la sécurité des données. L'algorithme mentionné dans la figure suivante illustre une allocation bottom up est: $N11 \rightarrow N12 \rightarrow N21 \rightarrow N22 \rightarrow N31 \rightarrow N32 \rightarrow N41 \rightarrow N41$. (algorithme d'allocation de MCA dans un ED) :

Supposition: les noeuds du même niveau ne peuvent pas accéder aux données des autres noeuds.
Entrée: un environnement SIIO de m niveaux et chaque niveau à n_i noeuds.
Sortie : Une matrice de correspondance d'attribut pour chaque noeuds: MCA_{ij}
<pre> begin for j = 1 to n_1 //création de la matrice de correspondance d'attribut fédéré pour le premier niveau $MCA_{1j} = MCA[F, f_{1j}, ED_{1j}]$ End for for i = 2 to m // création de la matrice de correspondance d'attribut fédéré pour les autres niveaux for j = 1 to n_i $MCA_{ij} = MCA[F, f_i, ED_{ij}, f_{(i-1)1}, f_{(i-1)2}, \dots, f_{(i-1)n_{(i-1)}}]$ endfor endfor end </pre>

Dans cet algorithme, les attributs étoiles fédérés f_{ij} sont obtenus à partir du ED_{ij} et $f_{(i-1)1}, f_{(i-1)2}, \dots, f_{(i-1)n_{(i-1)}}$. ce qui veut dire que les attributs étoiles fédérés des noeud du niveau inférieur doivent être obtenus à l'avance ,donc , $MCA_{ij} = MCA[F, f_i, ED_{ij}, f_{(i-1)1}, f_{(i-1)2}, \dots, f_{(i-1)n_{(i-1)}}]$, $1 < i < m$, $1 < j < n_j$. En d'autres termes, l'algorithme propose que les MCA doivent être construites du niveau inférieur vers le niveau supérieur. Cette conception vise à fournir une sécurisation pour que les données de l'entreprise ne soient pas révélées aux compétiteurs. Par exemple, si l'environnement SIIO est composé de quatre niveaux et chaque niveau contient deux noeuds, les MCA_{ij} sont :

noeud11 $MCA_{11} = MCA[F, f_{11}, ED_{11}]$, noeud 12 $MCA_{12} = MCA[F, f_{12}, ED_{12}]$
 noeud 21 $MCA_{21} = MCA[F, f_{21}, ED_{21}, f_{11}, f_{12}]$, noeud 22 $MCA_{22} = MCA[F, f_{22}, ED_{22}, f_{11}, f_{12}]$
 noeud 31 $MCA_{31} = MCA[F, f_{31}, ED_{31}, f_{21}, f_{22}]$, noeud 32 $MCA_{32} = MCA[F, f_{32}, ED_{32}, f_{21}, f_{22}]$
 noeud 41 $MCA_{41} = MCA[F, f_{41}, ED_{41}, f_{31}, f_{32}]$, noeud 42 $MCA_{42} = MCA[F, f_{42}, ED_{42}, f_{31}, f_{32}]$.

En utilisant le noeud 21 comme un exemple, la MCA du ED_{21} inclut les attributs des niveaux inférieurs (ED_{11}, ED_{12}), et niveau supérieurs (ED_{31}, ED_{32}), et son propre niveau.

(5) **la transformation des requêtes fédérées.** La traduction d'une requête fédérée nécessite quatre étapes pour sa traduction à partir d'une requête locale:

- décomposition de la requête,
- conversion des attributs,
- spécification des conditions de jointures,

- et l'identification des objets source.

La conception distribuée de l'entrepôt de données maintient une MCA partielle pour le traitement des requêtes locales. La requête locale traverse les associés du SIIO pour la recherche des données. Dans la conception, les données peuvent être conçues de sorte qu'elles ne sont pas consultées par des concurrents tout en étant partagées avec des associés. L'inconvénient est que toutes les requêtes font participer des associés du SIIO et ont comme conséquence l'exécution inefficace de la requête, particulièrement quand les associés sont dans des sites éloignés.

3.7. Synthèse

En résumé, le problème d'intégration et d'interopérabilité des entrepôts de données caractérisé par le degré d'autonomie, de distribution et d'hétérogénéité des entrepôts existants, d'avantage, résidants au niveau des organisations formant un SIIO, ne peut pas se fier uniquement aux trois critères suscités en vue d'opter pour une solution de consolidation. Pour cela, on a évoqué les principaux facteurs influençant le choix d'une stratégie d'intégration des ED dans un environnement inter-organisationnel à savoir, Le degré de transparence voulu ; La nature des relations entre ces organisations ; Le paradigme d'intégration à adopter; Le degré du couplage à mettre en valeur entre les schémas composants de l'entrepôt global; La modélisation du schéma intégré , etc..;

En conclusion pour mettre en exergue une stratégie d'intégration et de consolidation des silos analytiques dans un environnement SIIO, un certain nombre de facteurs doivent être évalués à savoir (fig 3.16) :

- **Des critères organisationnels** : la taille des entreprises dans le SIIO, la structure des relations entre les organisations impose un mode de coopération, de coordination et de contrôle et qui se répercute sur le degré d'autonomie de chaque partie prenante, le niveau métier sur lequel ces relations activent (opérationnelle ou stratégique).

- **Des critères informationnelles** : La nature des silos analytiques autonomes et distribués (entrepôts de données, marchés de données, magasins opérationnels de données), la nature sémantique des informations contenues dans ces dits silos, s'ils appartiennent au même secteur d'activité (intégration horizontale) ou à des secteurs complémentaires (intégration verticale), et de la durée allouée pour la prise en charge un besoin opérationnel ou stratégique d'intégration.

- **Des critères économiques**: Le souci stratégique des cadres (réduction des coûts ou des initiatives stratégiques), la nature des objectifs du projet et des économies potentielles;

- **De l'environnement TIC existant** : De la portée géographique du SIIO et des silos analytiques, des plates formes technologiques disponibles (*réseaux, nature des données, standards....*)

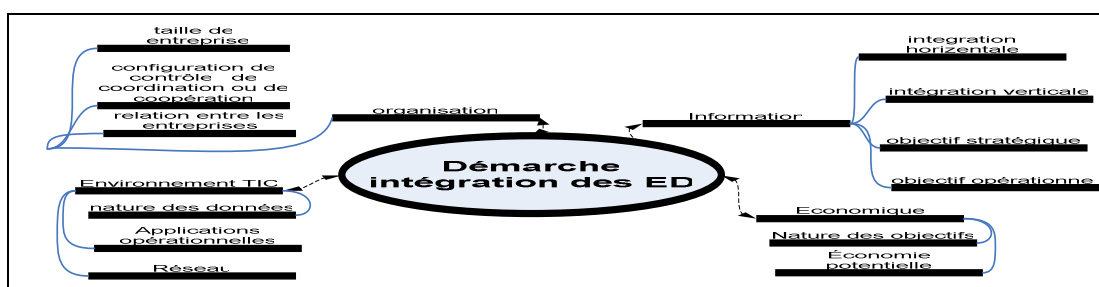


Figure 3.16- facteurs influençant le choix de l'approche d'intégration des ED dans un SIO.

Bien que désirable, nous pensons qu'il n'est pas possible d'élaborer une classification orthogonale des critères régissant la dite démarche. Toutefois une classification basée sur une dimension induit le choix d'une autre dimension, par exemple l'aspect stratégique d'un SIO entraîne une matérialisation du schéma intégré, un SIO dont les organisations sont intégrées horizontalement suppose une intégration à travers la correspondance des schémas au lieu d'élaborer un schéma global.

Ainsi pour répondre aux questions suivantes:

- Quel est le paradigme d'intégration à employer ?
- Devrions nous adopter une approche centralisée ou décentralisée d'intégration ?
- Dans le cas d'un schéma centralisé quelle est la nature du schéma commun ?
- Quel est le degré de couplage des schémas composants ?
- Etc...

La réponse à ces questions et autres nous permet de caractériser notre démarche dans un SIO classés par nature d'interdépendance et qui se présentent, selon notre synthèse, comme suit :

	SIO d'information partagée	SIO de chaîne d'approvisionnement	SIO réseau
Type d'interdépendance	Groupée	Séquentielle	Réciproque
Intégration forte ou faible	Forte	Moins forte	Faible
Hétérogénéités traitées	Toutes	De schéma et de langage	De schéma
Approche d'intégration	Centralisée (matérialisée ou virtuelle)	Décentralisée (virtuelle)	Décentralisée (virtuelle)
Paradigme d'intégration	<ul style="list-style-type: none"> • Bottom up pour un support stratégique • ou top down pour un support opérationnel • ou hybride pour un niveau stratégique entre concurrents. 	Top down ou hybride	Top down
Intégration des schémas ou correspondance des schémas	Intégration en un seul schéma.	Intégration en un ou plusieurs schémas, ou correspondance entre plusieurs schémas	Correspondance entre schémas ou P2P
Stratégies d'intégration	<ul style="list-style-type: none"> • Stratégie de migration, • Recommencer à nouveau (Start from Scratch) • Designer un standard 	<ul style="list-style-type: none"> • ED/MD Conforme (Conformed Data Marts) • L'entrepôt des entrepôts (Mart of Marts) • Requête Distribuée (Distributed Query). 	<ul style="list-style-type: none"> • Requête Distribuée (Distributed Query).

	et évoluez (Designate and Evolve) <ul style="list-style-type: none"> • Alimentation alignée (Backfill) • Synchroniser (Synchronize). 		
Architecture d'ED intégré	Centralisée et matérialisé ou fédérée fortement couplée.	Mono ou multi fédérations fortement ou faiblement couplées.	P2P
Durée du projet	Selon le support du SIIO au niveau stratégique ou opérationnel		
procédé sémantique	Selon la similarité (collecte et fusion) ou la complémentarité (collecte et abstraction) de la sémantiques des données.		

Tableau 3.1-Synthèse liant les SIIO au processus d'intégration.

En conclusion, la prise en charge de la formalisation d'une démarche pour la conception ascendante d'un ED distribué serait d'un intérêt certain et évident, d'autant plus que la littérature scientifique ne comporte sur ce sujet que quelques tentatives de formalisation de développement d'ED distribués impliquant des cas spécifiques du SIIO tels que le E-commerce ou la gestion de la chaîne d'approvisionnement (SCM).

Ainsi nous espérons que notre proposition contribuera à mettre en exergue une démarche de conception en spécifiant les phases majeures, les activités et les artefacts correspondants à chaque phase et qui sera présenté dans le prochain chapitre.

CHAPITRE 4 : Conception bottom up des ED distribués dans un environnement SIIO : Cycle de vie

4.1. Introduction

A travers l'exposé de l'état de l'art relatif à l'intégration des ED, la principale observation recueillie suggère que le problème d'intégration des ED autonomes peut principalement se restreindre à l'intégration des faits et des dimensions indépendamment conçus, concrétisée à travers un schéma déployé sur une architecture centralisée ou décentralisée. L'architecture dudit schéma est tributaire de la nature des relations entre les organisations formant le SIIO, comprenant :

- la nature des interdépendances entre les organisations;
- le support opérationnel ou stratégique assuré par le SIIO ;
- la nature similaire (intégration horizontale) ou complémentaire (intégration verticale) du domaine d'activité des organisations formant le SIIO.

C'est dans ce contexte – sans tenir compte des environnements inter-organisationnel- et d'un point de vue conceptuel que s'inscrivent les travaux de [LR03, LR04, LR05], en vue d'étendre le concept de la conformité des dimensions (faits) de **Ralph Kimball** par la notion de compatibilité de dimensions (faits), stipulant que deux dimensions (faits) sont compatibles si leurs informations communes, extraites à travers une expression algébrique dimensionnelle, sont consistantes¹.

Toutefois, il n'y a pas une méthodologie formelle pour la conception ascendante des ED distribués car les auteurs suscités et autres, tels que, Doug Hackney [DH98], Colin White [CW01], R.A. Moeller [RA01] ou Bill Inmon [INM05] ne proposent pas une démarche détaillée pour la spécification d'une méthodologie de consolidation des ED/MD distribués, hétérogènes et autonomes, et davantage dans un environnement inter-organisationnel problème auquel nous nous sommes intéressés et dont nous avons tenté d'identifier les spécificités majeures par une démarche de conception. A cet effet, ce chapitre va aborder :

- En premier lieu, la démarche d'intégration des ED distribués,
- En deuxième lieu, sur la base de l'aspect itératif et incrémental caractérisant généralement les démarches d'élaboration des entrepôts de données et spécialement cette démarche, dont l'analyse est dirigée principalement par les besoins des utilisateurs, par l'analyse des ED existants, et par les objectifs de la coalition, la démarche d'intégration des ED dans un environnement inter-organisationnel, plaide en faveur d'une démarche de processus unifié (« Unified Process »), impliquant un processus d'analyse et de conception orienté objet en UML.
- En dernier lieu, une étude de cas sera menée en vue de mettre en exergue la dite démarche sur une architecture centralisée dans un SIIO stratégique et vertical caractérisant le cercle de valeur de la santé.

¹ Les informations sont consistantes si leur matching est vérifié.

4.2. Démarche d'intégration des ED distribués dans un SIIO :

Dans cette section, nous allons présenter notre vision sur la démarche d'intégration des ED autonomes et hétérogènes dans un environnement inter-organisationnel et dont le résultat sera déployé sur une architecture modélisée selon l'architecture fédérée à cinq niveau de [SL90], avec quelques simplifications et faisant appel à quatre niveaux de schémas, à savoir :

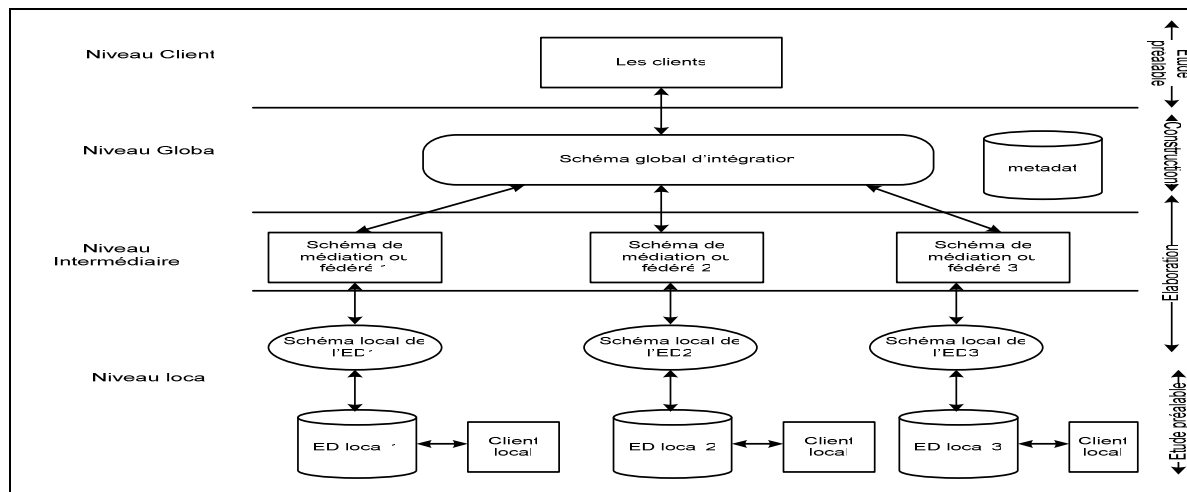


Figure 4.1 - Architecture d'intégration des ED hétérogènes.

– **Le niveau local**, il fournit les ressources et les schémas des ED locaux. l'ED participant à l'intégration utilise le schéma local pour la description de ces données. Ce schéma local, comprenant globalement ou partiellement les données sources, est conçu indépendamment des autres selon un modèle quelconque qui fournit une vue multidimensionnelle sur les données (schéma en étoile, en flocon de neige ou autres) et qui doit être unifié après la résolution des problème d'hétérogénéité pour être intégré.

– **Le niveau intermédiaire**, Ce niveau est utilisé pour assurer la liaison entre les schémas locaux et le schéma global en facilitant les échanges entre les metadonnées des schémas, comportant les données extraites à partir des schémas locaux modélisés uniformément et choisies pour être mises en commun entre les différents ED et dont la fonctionnalité principale est de :

1. Transformer les requêtes globales en un langage de requête compréhensible par les ED locaux dans le cas d'une architecture virtuelle ou assurer la mise à jour périodique du schéma global à partir des ED locaux dans le cas d'une architecture matérialisée ;
2. Transformer les résultats partiels renvoyés par les ED locaux d'un format propriétaire à un format unifié.

– **Le niveau global**, Ce niveau rend transparent l'hétérogénéité structurelle et sémantique des ED locaux et fournit une vue unifiée pour les utilisateurs pour accéder aux ED locaux. A ce niveau le schéma crée, comporte une collection d'attributs, de mesures et de dimensions des schémas intégrés et qui est l'instanciation des ED locaux participant à l'intégration. Pour ce faire, ce niveau comprend :

1. **Un modèle de données canonique** : servant à la description de chaque schéma local incluant, faits, mesures, dimensions, hiérarchies de dimensions et permettant la modélisation multidimensionnelle des données en un schéma en étoile ou autres.

2. **un fichier de mapping** : est un fichier incluant la description des schémas, des ED locaux et des correspondances verticales entre les entités du niveau globales et les entités du niveau locales , à savoir, les faits, les mesures, les attributs et les dimensions.

– **le niveau externe** : A ce niveau le client émet ses requêtes qui sont envoyées au niveau global qui les dirige vers l'analyse des ED locaux dans le cas d'une architecture virtuelle, ou simplement, le niveau global renvoie la réponse au client dans le cas d'une architecture centralisée.

En vue d'atteindre la réalisation de cette architecture, on va adopter un phasage respectant celui du processus unifié -qui sera expliqué par la suite- faisant appel à :

- **une étude préalable** pour l'analyse de l'existant et la sélection des schémas où sous schémas à inclure dans l'intégration à la lumière des objectifs assignés par les membres du SIIO à cette intégration.
- **une phase d'élaboration** où le modèle conceptuel global ascendant reflétant les objectifs du SIIO sera spécifié, la modélisation conceptuelle uniforme des schémas recouverts à partir des ED locaux sera effectuée et leurs correspondances horizontales seront définies, de même que les correspondance verticales entre les schémas locaux et le schéma global seront identifiées,
- ainsi au niveau de la dernière **étape de construction** le schéma consolidé est élaboré et sa modélisation logique et physique sera effectuée.

Compte tenu de la complexité des environnements inter-organisationnels, force est de constater que notre démarche ne peut pas s'appuyer uniquement sur un paradigme bottom up d'intégration, par contre à travers une démarche qui respecte une approche hybride mariant l'orientation bottom up et top down du paradigme d'intégration, et qui se présente ainsi (fig 4.2):

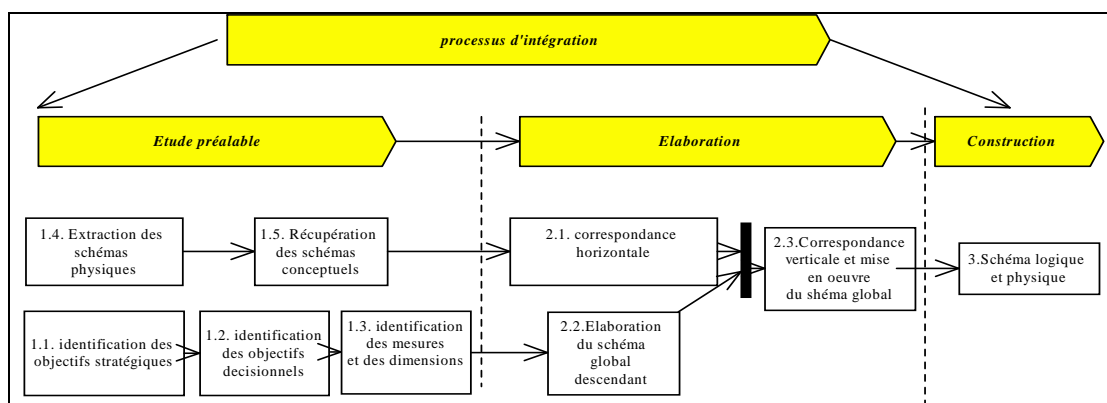


Figure 4.2- Processus générique d'intégration des ED à évolution indépendante.

1. Etude préalable :

Au niveau de cette phase il faut prendre en charge les exigences du niveau client ou les organisations du SIIO dans notre cas, par l'étude des objectifs de cette intégration en vue d'identifier les

données d'analyse (faits, dimensions et mesures) et en parallèle extraire les schémas physiques des ED existants via un processus de rétro-ingénierie de données, après quoi on procède à la récupération des schémas conceptuels des ED locaux selon un modèle multidimensionnel commun (Fig 4.3);

Cela dit, on aura comme résultat les schémas conceptuels des ED existants ainsi que le schéma dimensionnel ascendant¹;

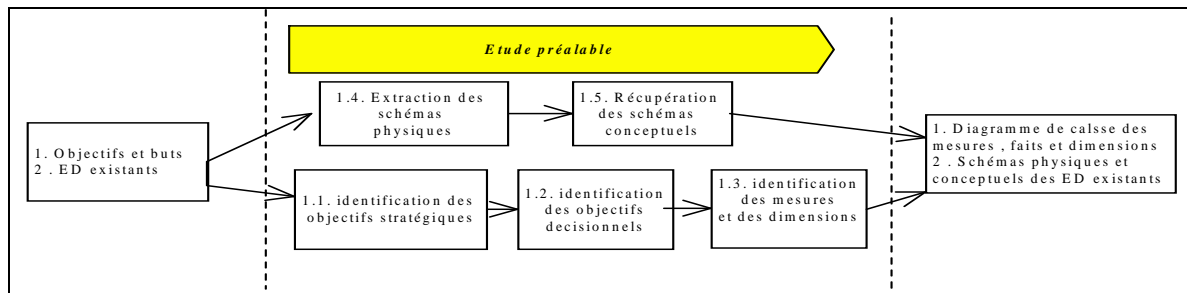


Figure 4.3- Phase d'étude préalable

Compte tenu de la détermination du domaine que l'on désire analyser, ce qui revient la plupart du temps à choisir le processus métier transcendant les organisations du SIIO, il faut spécifier les soucis stratégiques devant être pris en charge par l'ED global et le niveau d'accessibilité souhaité ce qui permet de délimiter la portée de la l'intégration.

La spécification des objectifs du projet d'intégration et de consolidation des ED/MD appartenant à différentes organisations est effectué à travers un outil de modélisation des buts (formalisme i* ou un diagramme de classe en UML) pour définir l'étendue des analyses et les mesures à considérer ainsi que les entités analytiques existantes à manipuler. La description de ces objectifs doit être structurée de façon à permettre d'associer pour chaque processus métier inter-organisationnel (figure 4.4) :

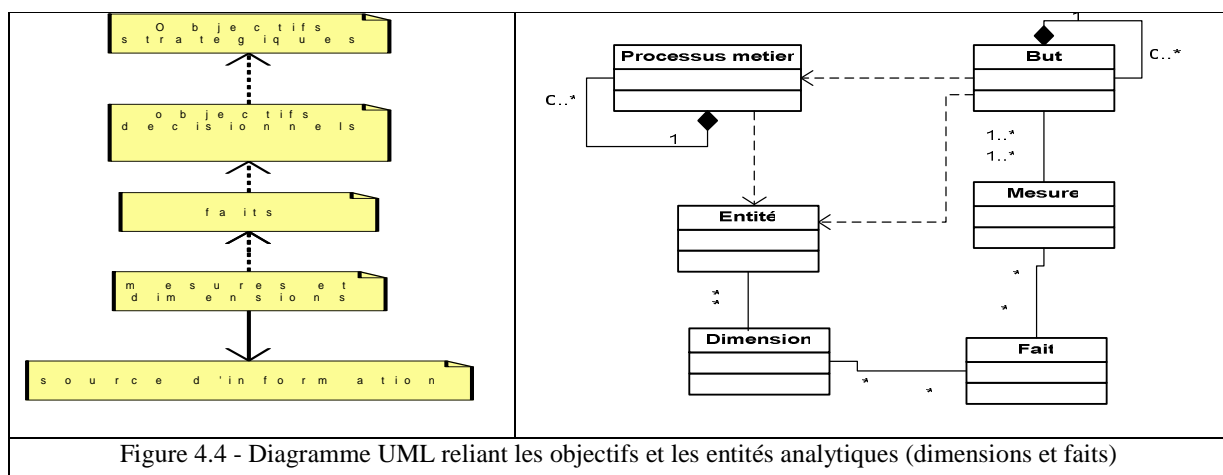


Figure 4.4 - Diagramme UML reliant les objectifs et les entités analytiques (dimensions et faits)

¹ Il faut noter que la numérotation adoptée correspond aux phases et aux activités de la démarche illustrée dans la figure 4.2

1.1. L'objectif stratégique ou le but principal à analyser: l'objectif stratégique de la coalition motivant le souci d'intégration représente le niveau d'abstraction le plus élevé des objectifs du processus métier inter-organisationnel et qui préconise le passage de la situation métier actuelle à une meilleure situation telle que « l'atténuation des coûts de la production » et dont l'achèvement se répercute par un bénéfice sur les organisations du SIIO ;

1.2. Les objectifs décisionnels ou les sous buts correspondants: représentent le niveau intermédiaire d'abstraction des objectifs du processus métier inter-organisationnel toute en étant inscrit dans un contexte d'objectif stratégique. Les objectifs tactiques s'illustrent à travers la question suivante : « comment achever l'objectif stratégique ?? ».

1.3. Les faits, les mesures et les dimensions du schéma dimensionnel global consenti par les organisations formant le SIIO. A ce niveau, il faut aboutir à un consentement sur les faits, les mesures et les dimensions correspondants à la question suivante « comment achever les objectifs tactiques en termes d'information exigées ?? ». A cet effet, on déplore l'absence d'un formalisme de passage des spécifications des buts aux entités conceptuelles analytiques qui ne sont exprimés qu'à travers des guides tels qu'illustrés dans les travaux de [JJMM05] et de [JF98].

Après l'identification des faits, des mesures et des dimensions on peut dresser le schéma dimensionnel global descendant dont l'architecture centralisée ou décentralisée est conditionnée selon le degré de transparence et d'autonomie assigné à cette intégration envisagée par les organisations formant le SIIO.

1.4. Extraction des schémas physiques des ED existants : en parallèle à la phase d'identification des objectifs, cette phase, qui découle d'un processus de retro-ingénierie, vise à récupérer les schémas physiques où sont documentées les entités analytiques et leurs propriétés, c a d, les définitions des structures de données, les métadonnées et les applications de restitution.

A cet effet, l'étude des structures analytiques n'est possible qu'à travers l'étude des métadonnées métiers et techniques, permettant d'illustrer, ce qui suit :

a. Meta-données métiers : aident à identifier et à localiser les informations à usage analytique, incluant :

- le glossaire des termes métiers ;
- les termes métiers et les définitions des tables et les colonnes ;
- la définition des états de restitution des données analytiques;
- la définition des données métiers dans les ED/MD.

b. Meta données technique incluent les aspects techniques tels que les colonnes des tables, type des données et longueur. Il permet d'approfondir la compréhension des structures de données physiques et de leurs relations dans le contexte de l'organisation. Cette analyse doit comprendre :

- les noms des tables physiques et de leurs colonnes,
- le mapping des données et les transformations logiques,
- le détail du système source,

- les clés étrangères et les index,
- Le mode de sécurisation adopté.
- l'analyse linéaire impliquant le suivi des données du portail de restitution vers la source, avec la spécification des transformations impliquées.

Selon [CAVN05], les informations devant être recueillies lors d'un processus de rétro-ingénierie des structures de données analytiques, peuvent être résumées comme suit :

Paramétré	description
Nom	Le nom de la structure analytique, qui peut être : un ED, un MD indépendant, un MD dépendant ou un ODS.
Processus métier	Le processus métier utilisant cette structure.
Granularité	Le niveau de détail des données emmagasinées dans la structure analytique.
Dimensions	Les dimensions définissant la structure analytique.
Faits	Les tables de faits contenus dans ces structures.
Systèmes sources	Les systèmes sources utilisés pour l'alimentation des structures analytiques.
Le propriétaire du système source	L'organisation détentrice du système source.
Le propriétaire de la structure analytique	L'organisation détentrice de la structure analytique.
Les états générés	Les états générés à partir des informations analytiques, leur nombre et leur type.
Le nombre des utilisateurs métiers	Total des utilisateurs accédant aux structures analytiques.
Le nombre des tables et attribut	Pour évaluer la complexité des données.
Le volume des données	Pour l'hébergement de la taille globale des structures analytiques.
Nombre des outils et routines ETL	Pour évaluer la complexité des processus d'alimentation des structures analytiques à intégrer.
Scalabilité de la structure analytique	La scalabilité du système à l'égard de l'accroissement du volume des données et des utilisateurs

Tableau 4.1-Analyse des schémas physiques des structures analytiques existantes.

1.5. Récupération des schémas conceptuels des ED existants : A partir des schémas physiques des ED existants, et en faisant usage d'un procédé de rétro-conception qui permet de fournir une visualisation et une représentation abstraite de haut niveau des bases de données multidimensionnelles existantes, on cherche à enrichir sémantiquement les schémas physiques des ED existants où les structures de données sont incrémentalement identifiées, ainsi, les faits, les dimensions, les attributs et les mesures sont transformés en descriptions sémantiquement standardisées et qui résultent des schémas dimensionnels exprimés par un modèle multidimensionnel commun (Entité Association ou en Orienté Objet).

En résumé, compte tenu des informations recueillies lors de cette phase, impliquant les buts qui délimitent l'étendue du projet et les structures analytiques existantes, on pourrait tirer les conclusions suivantes :

1. **Récupération des schémas conceptuels** des ED existants et les entités analytiques globales (fait et dimensions) décelées lors de l'analyse des objectifs.

2. **Détermination des informations accessibles** et celles qui ne le sont pas à travers la définition des schémas ou sous schémas conceptuels locaux devant être mis en commun entre les organisations formant le SIIO qui sont assujettis à la nature des relations entre les organisations et le degré d'autonomie et de transparence assignée à cette intégration, car si les organisations sont en relation :

➤ **Opérationnelle**, où l'accessibilité n'est permise que sur les informations relatives aux processus métiers opérationnels transcendants les organisations pour une durée limitée correspondante à l'exécution contractuelle du processus métier (intégration verticale) ou à travers un référentiel de base sur les données des activités opérationnelles communes (intégration horizontale).

➤ **Stratégique**, où l'accessibilité n'est permise que sur les informations synthétisées orientant les choix stratégiques dans l'exécution des processus métiers entre organisations exerçant des activités complémentaires (intégration verticale) ou similaires (intégration horizontale).

3. **Sélection d'une stratégie d'intégration**, permettant de mettre en place :

➤ Une architecture centralisée où il faut intégrer les ED/MD au niveau d'une seule plate forme, appartenant généralement à l'organisation qui a investi le plus de ressources, dans la cas de SIIO à ressources partagées ou dans le cas des SIIO horizontaux qui optent pour la centralisation d'un référentiel des données de bases, et qui se traduit conceptuellement par l'union des schémas locaux des ED existants.

➤ Une architecture décentralisée, qui se concrétise conceptuellement par l'intersection des schémas locaux des ED existants à travers :

- le partage des entités analytiques : dimensions et faits des différents ED/MD,
- la standardisation de la zone de préparation de données,
- la standardisation des meta-données.

4. Effectuer une évaluation de la charge de travail nécessaire pour l'intégration à travers la mise en conformité ou compatibilité des dimensions et des faits des schémas dimensionnels existants en vue de supporter les différents objectifs définis par les organisations du SIIO.

2. Elaboration :

Comme illustré dans la figure Fig4.5, sur la base :

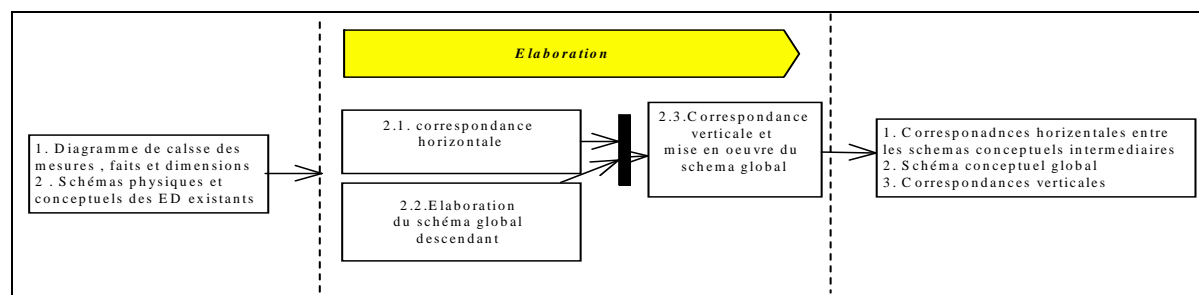


Figure 4.5 - Phase d'élaboration

- du schéma dimensionnel ascendant résultant de l'étude des objectifs d'intégration,
- des schémas conceptuels récupérés à travers l'étude des ED existants,
- l'approche d'intégration adoptée, à savoir :

Approche centralisée, qui se base sur l'union des schémas et pouvant être effectuée de deux façons : par la re-conception où le schéma de l'ED intégré cible (global) doit être conçu de nouveau, ou par la fusion avec un schéma sélectionné parmi les ED/MD considéré comme référence ainsi les autres ED/MD seront fusionnés au schéma primaire. Le résultat de cette approche est un schéma intégré matérialisé.

Approche distribuée : dans cette approche l'intégration est virtuelle. Elle est élaborée par une adaptation des dimensions communes pour les rendre compatibles ou conformes avec les autres en premier lieu, et la coordination des schémas intermédiaires résultants de l'analyse des schémas conceptuels de chaque ED/MD par l'élaboration des matrices de correspondance d'attribut en second lieu.

On peut citer au niveau de cette phase les activités prédominantes qui se résument ainsi :

2.1. Etude des correspondances horizontales entre les schémas conceptuels des ED existants

Cette étape vise à partir des schémas conceptuels intermédiaires résultant de la modélisation uniforme des schémas ou des sous schémas locaux récupérés des ED existants :

- À identifier les dimensions et les faits redondants ;
- A déceler les données inconsistantes ;
- À déduire les relations sémantiques entre les faits et les dimensions des ED existants,
- En plus la standardisation des métadonnées;

Ce qui permet l'identification des correspondances d'un point de vue syntaxique et sémantique entre les schémas dimensionnels intermédiaires. A cet effet, les activités suivantes s'inscrivent dans ce sens :

2.1.1. La redondance des entités analytiques:

Pour identifier la redondance entre les schémas on dresse une matrice (tableau 4.2) où sont listées :

- Horizontalement, les entités analytiques des schémas dimensionnelles (Dim N, Fait N) selon leur appartenance ;
- Verticalement, toutes les entités analytiques ascendantes (faits + dimensions) consenties par les organisations du SIO et impliquées dans l'intégration (Fait G, Dim G).

Cette matrice permet de repérer les dimensions et les faits communs comparables qui pourront faire l'objet d'une mise en conformité ou compatibilité.

	ED1/MD1			ED2/MD2			EDn/MDn		
fait et dimensions universels	Fait1	Dim1.1	Dim1.2	Fait2	Dim2.1	Dim2.2	Fait n	Dim n.1	Dim n.2
Dim G.1	X				X			X	
Dim G.2		X							X
Dim G.3									
Fait G.1									
Autres									

Tableau 4.2 -Matrice d'analyse des entités analytiques redondantes

Cette matrice ne peut déceler que les fait et les dimension communs ou redondants, cependant, il pourrait y'avoir d'autre cas de similarité dissimulée dûes a l'hétérogénéité sémantique des schémas. Pour cela l'identification des relations sémantiques permettrait d'identifier les relations entre les faits et les dimensions manipulés par les ED existant, en vue de mettre en valeur les possibilités d'élaboration des requêtes sur les schémas dimensionnels. Ces relations sémantiques ont pour but de dissiper toute hétérogénéité sémantique par l'uniformisation des schémas dimensionnels et des méta-adonnées .

2.1.2. Identification des relations sémantiques entre les entités des schémas dimensionnels intermédiaires :

La définition des relations sémantiques entre les faits et les dimensions des différents ED autonomes permet à partir d'une instance d'une dimension ou d'un fait dans un schéma dimensionnel de déterminer d'autres instances de faits ou de dimensions dans d'autres schémas, rendant ainsi l'élaboration d'une requête de forage transversal ou de Drill-across sur plusieurs cubes ne comprenant pas forcément des dimensions ou des faits conformes possibles, toutefois, leurs dimensions sont tributaires des mêmes conditions.

Selon [AJF02], les propriétés pertinentes des requêtes de forage transversal sollicitant la navigation entre les MD/ED sont conditionnées par les relations sémantiques entre les entités des schémas dimensionnels. Il a énuméré quatre relations reliant les dimensions et les faits entre eux, à savoir (figure 4.6) :

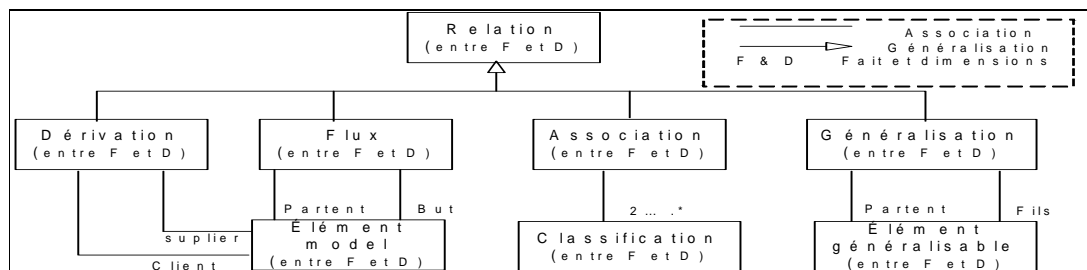


Figure4.6- Modélisation des relations sémantiques entre faits et dimensions appartenant à différents ED[AJF02].

- *la généralisation*, elle peut seulement être retrouvée entre deux dimensions ou deux faits. Les dimensions et les faits peuvent seulement être reliés par dérivation ou association.

- *la dérivation* : une « dimension » peut être une vue d'une autre « dimension » ou un « fait », et un « fait » pourrait être une vue d'un autre « fait ». Un « fait » ne peut pas être une vue d'une « dimension », parce que les faits représentent des données mesurées.

- *l'association* : les associations dépendent de leurs multiplicités. Si la multiplicité est un-à-un ou un-à-plusieurs, ils peuvent facilement être mis en application au moyen d'une contrainte d'intégrité référentielle (FK). Si la multiplicité est de type plusieurs-à-plusieurs, ils peuvent être mis en application en utilisant une « entité passerelle ». Les associations existent entre deux dimensions, deux faits ou un fait et une dimension.

- *le flux*, ce genre de rapport devrait être mis en application au moyen d'une contrainte d'intégrité référentielle (FK) entre de vieilles et de nouvelles versions de faits ou de dimensions. car une dimension ne peut pas par la suite évoluer vers un fait, par contre un fait peut être transformés vers une dimension via une abstraction temporelle.

Par conséquent, pour une

Relation dimension-dimension : deux dimensions sont reliées par la relation de généralisation ou de flux, donc elles seront différentes. Néanmoins, dans la mesure où leurs instances seront reliées avec une relation de type « un à un », alors, l'analyse drill-across serait possible. Dans le cas où les dimensions seront reliées par une association ou une dérivation leurs instances ne coïncideront pas. Cependant l'instance d'une dimension déterminera l'instance d'une autre.

Relation fait-fait : Notons qu'il n'est pas nécessaire que les dimensions dans un fait soient en liaison avec celles d'un autre fait mais il faut que les cellules d'un fait déterminent les cellules dans un autre fait. Donc nous aurons juste à remplacer les mesures d'une cellule par celles des autres faits. De cette manière nous pouvons étudier des données dans deux cubes différents où leurs faits sont reliés sémantiquement.

Relation fait-dimension et dimension-fait : La relation entre les faits et les dimensions ne permet pas une analyse drill cross. Cependant si les cellules sélectionnées pendant l'analyse d'un fait déterminent un ensemble de points dans une dimension cela peut être utilisé dans l'analyse d'un autre fait. Remarquons que ni la relation de généralisation ni la relation de flux entre les faits et les dimensions ne sont considérées, car une transformation temporelle s'impose pour convertir un fait en une dimension et non le contraire.

Le tableau 4.3 [AJF02], résume la nature des relations entre les différentes entités des schémas dimensionnels en étoile pour une analyse drill cross par les outils OLAP. Il montre comment les différents schémas en étoile peuvent être liés par une relation de généralisation, d'association, de flux ou de dérivation (en terminologie UML), et comment utiliser ces relations pour réaliser des requêtes entre les schémas dimensionnels même si les dimensions ne sont pas partagées.

Relation	Dimension-Dimension	Fait-Fait	Fait-Dimension	Dimension-Fait
Dérivation	Oui	Oui	Oui	Non
Généralisation	Oui	Oui	Non	Non
Association	Oui	Oui	Oui	Oui
Flux	Oui	Oui	Non	non

Tableau 4.3 - Correspondance entre fait et dimension appartenant à deux schémas dimensionnels [AJF02]

D'un autre coté, ces relations permettent de définir les matching entre les dimensions et les faits des schémas conceptuels sélectionnés pour être intégrés et les entités métiers analytiques abstraites déduites à travers les objectifs assignés par les organisations formant le SIIO à cette intégration (fig 4.7).

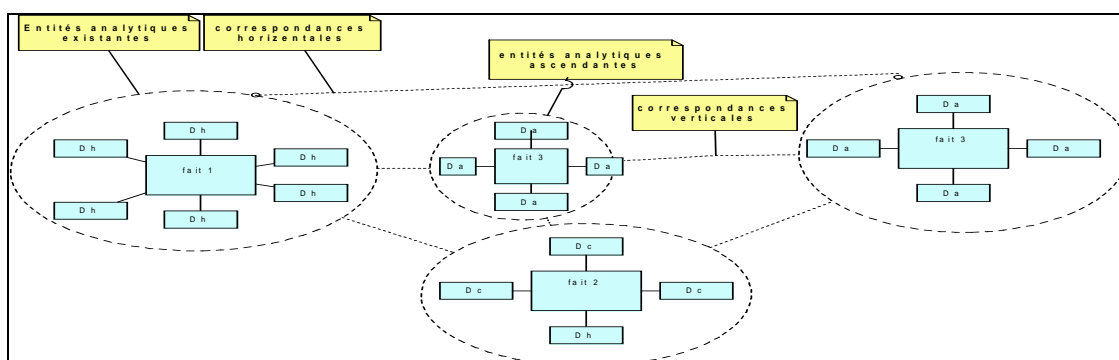


Figure 4.7- correspondance entre les faits et les dimensions des ED autonomes

- 2.2. **Spécification du schéma dimensionnel global descendant** à partir du diagramme de classe représentant les objectifs stratégiques et décisionnels définissant les faits et dimensions du schéma global et selon les entités analytiques résultantes des correspondance horizontales entre les schémas conceptuels récupérés à partir des ED existants on peut mettre en œuvre le schéma global et les correspondances verticales correspondantes, par la suite;
- 2.3. **Elaboration du mapping entre les schémas dimensionnels locaux et le schéma dimensionnel descendant** à travers l'élaboration des correspondances verticales en spécifiant les matching entre les dimensions et faits des schémas locaux et ascendant la standardisation entre les met adonnées des schémas locaux et le schéma global;

Cette phase aura comme résultat le schéma conceptuel de l'ED/MD distribué et le mapping entre les schémas locaux et le schéma global. Ceci va permettre dans le cas d'une architecture centralisée de concevoir les procédures ETL pour l'alimentation du schéma dimensionnel global, ou dans le cas d'une architecture décentralisée de préparer les tables de correspondances globales et intermédiaires prenant en charge les requêtes distribuées.

3. Construction :

Cette phase comprend la matérialisation technique proprement dite du schéma de l'ED distribué et implique la construction logique et physique du schéma visé. Pour le passage du schéma conceptuel vers le schéma logique, il faut prendre en considération les points ci-dessous : les systèmes sources, les restrictions d'accès, les spécifications non fonctionnelles, la correspondance entre le schéma conceptuel

et les ED/MD qui permettent d'identifier les sources de données des faits et des dimensions et des propriétés conceptuelles et les spécifications inhérentes à la contrainte d'espace disque, la fréquence de mise à jour, la matérialisation des vues dans le cas d'une approche centralisée, etc.....

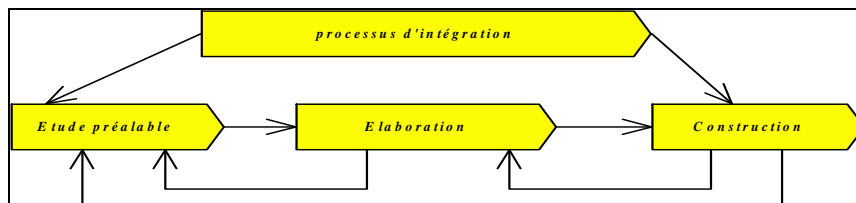


Figure 4.8 – Allure du processus d'intégration des ED.

Enfin, il ne faut pas oublier de mentionner que ce cycle de vie inclus la possibilité d'effectuer un ou plusieurs retours (fig 4.8) en arrière permettant la revue d'une activité ou d'une étape ou d'une phase ou le processus en entier une autre fois.

Le cycle de vie est conçu pour s'accommoder à une variété de scénarios impliquant des configurations organisationnelles et opérationnels, ce qui confère à ce cycle, au même titre que celui des ED en général, le caractère itératif, continu et progressif pour son accomplissement. Ce cycle de vie peut être réalisé à travers plusieurs projets d'intégration, mais il peut être concrétisé aussi à travers de multiples itérations relatives à un seul projet de consolidation. Cela signifie qu'il n'y a aucune exigence d'achever un projet de consolidation dans un effort majeur. Selon [CAVN05], dans le cas où un effort de consolidation majeur est nécessaire sur les opérations continues de l'entreprise, le projet doit être accompagné par des activités de mesures de risque et des procédés de prototypage.

4.3. Formalisation:

La méthodologie la plus couramment utilisée en matière de génie logiciel est basée sur le cycle de « spécification, conception, codage et maintenance ». Cette approche, dite en cascade, nécessite que chaque phase doit être entièrement achevée avant d'entamer la phase suivante. Un retour à une phase précédente étant permis seulement en cas de modification des spécifications initiales. D'autres méthodologies très utilisées dans les environnements objets, et notamment dans le cadre du développement de prototypes, autorisent l'itération sur certaines phases du cycle de conception du logiciel. L'une des caractéristiques essentielles des méthodologies de conception orientées objet réside dans le fait que l'intégralité de la démarche de conception peut entièrement être appliquée de nouveau sur le résultat obtenu par un précédent cycle ou itération. Dans notre cas, nous avons choisi d'utiliser le formalisme objet UML associé au processus unifié sous-jacent pour spécifier notre démarche [JBR00]. Le processus unifié (UP), basé sur la notation UML, permet le développement de logiciels suivants :

- **Une stratégie itérative et incrémentale** : Le caractère itératif correspond à la désignation des étapes de l'enchaînement d'activités, tandis que les incréments correspondent à des stades de développement du produit [JBR00], car dans une démarche de conception ascendante, les composants de niveaux inférieurs sont d'abord définis, ensuite ils sont utilisés pour réaliser ceux du niveau au dessus jusqu'à arriver au niveau supérieur où se trouve le composant principal.

- **Une stratégie pilotée par les cas d'utilisation**, l'aspect dirigé par les cas d'utilisation fonctionnels représentent une façon d'utiliser le système d'information au bénéfice d'un ou de plusieurs acteurs. Il permet une meilleure définition des besoins fonctionnels et une référence régulière à ces spécifications [JBR00].

- **Une stratégie centrée sur l'architecture**, l'architecture signifie les aspects statiques et dynamiques du système, émergeant des besoins de l'entreprise, tels qu'ils sont reflétés par les cas d'utilisation. D'autre part, comme l'ingénierie des entrepôts de données distribués dans les environnements inter-organisationnels est un processus qui rend complexe l'implémentation, la maintenance, la validation et la vérification de toute solution d'intégration, alors les activités de conception et d'architecture se joignent.

UP (processus unifié) est un ensemble de principes génériques adapté en fonction des spécificités du projet dont l'objectif est de maîtriser sa complexité en diminuant les risques [JBR00]. Il gère le processus de développement selon deux axes (voir la figure 4.9):

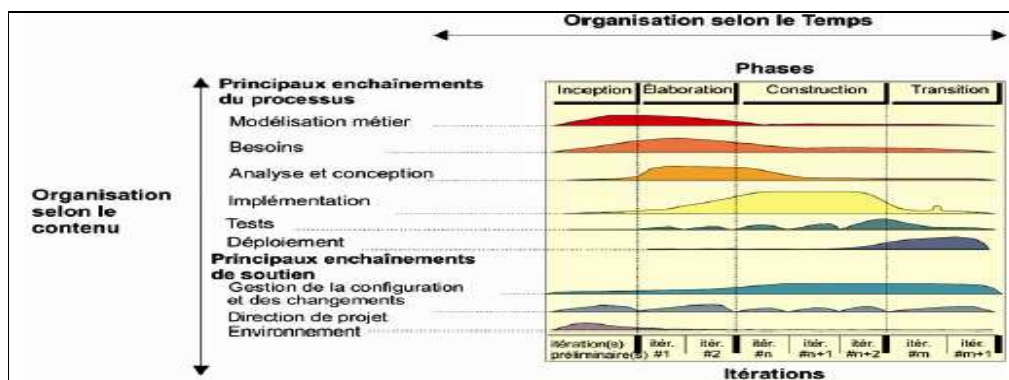


Figure 4.9 - Architecture de UP [JBR00].

L'axe horizontal représente le temps et montre les phases du cycle de vie du procédé logiciel. Il est exprimé en terme de cycles, phases, itérations et jalons. Le processus de développement est perçu comme une activité itérative à quatre phases :

- L'*initialisation* ou l'*inception* : Cette phase répond essentiellement aux questions relatives aux exigences métiers et à l'architecture fonctionnelle du système, exprimées à travers des cas d'utilisations métiers, ainsi que l'organisation de l'activité du développement.

- l'*élaboration*, reprend les éléments de la phase d'initialisation et les précise pour arriver à une spécification détaillée de la solution à mettre en œuvre. L'élaboration permet de préciser la plupart des cas d'utilisation, de concevoir l'architecture du système et surtout de déterminer l'architecture de référence.

- la *construction* : dans cette phase on construit le système,
- et la *transition* déploie le système dans l'environnement du client.

L'axe vertical représente les disciplines du procédé, regroupant les activités par leur nature. Cette dimension représente l'aspect statique du procédé qui spécifie les composants, les activités, le workflow et les artefacts.

A cet effet, le résultat de l'instanciation du processus unifié sur notre démarche entraînerait les phases et les activités suivantes (tableau 4.4):

Phases & activités	Phase d'inception	Phase d'élaboration	Phase de construction
Modélisation métier	Activité (1.1) : à travers un cas d'utilisation métier on identifiera les objectifs fonctionnels et tactiques de cette intégration.		
Besoins	Activité (1-2) : à travers l'identification des objectifs stratégiques et tactiques de l'intégration on va déceler les entités analytiques globales de l'intégration.	Activité(1.3&2.2) : Spécification du cas métier par un cas d'utilisation système en vue de spécifier le schéma dimensionnel global.	
Conception	Activité (1.4) : Etude des schémas physiques des ED existants en vue de récupérer les schémas conceptuels selon un model multidimensionnel commun . Définir l'architecture et la stratégie d'intégration.	Activité (1.5&2.1&2.3) Après récupération des schémas conceptuels des ED existants et identification des correspondances horizontales pour élaborer les correspondances verticales avec le schéma global.	Modélisation logique et physique du schéma global et la standardisation des métadonnées.

Tableau 4.4- Processus UP d'intégration des ED à évolution indépendante

Pour mener efficacement un tel cycle, dans le cadre de notre étude, on fera appel seulement aux diagrammes de cas d'utilisation métier et systèmes où sont identifiés respectivement les objectifs et les fonctionnalités du système ainsi que les diagrammes de classes définissant la structure statique du système.

Afin d'illustrer notre démarche, nous prendrons un exemple représentant un SIIO d'ordre vertical et stratégique d'un système de santé.

4.4. Etude de cas :

Nous considérons un processus métier incluant une **caisse de sécurité sociale** prenant en charge les patients bénéficiaires d'une hospitalisation au sein d'un **hôpital** qui dépend d'une **administration de santé publique**. L'hôpital, l'administration de la santé publique et la caisse de sécurité gèrent des volumes d'informations considérables, dont l'exploitation pose des problèmes, de qualité des données, d'hétérogénéité sémantique, structurelle et technique, etc. Le régime général de ce processus représente des milliers de salariés affiliés à une caisse et engageant des transactions de cotisations par un pourcentage de leur salaire, avec lesquelles la caisse procède aux paiements et aux remboursements

des factures de soins évaluées à des centaines de millions de dinars par an au profit de l'hôpital qui rend compte périodiquement de ces activités à l'administration santé publique.

Par conséquent, la problématique au niveau de ces organisations peut être résumée comme suit : dans le cadre de l'analyse budgétaire et des prestations de soins, **comment maîtriser le coût des prestations de soins**, cette question soulève des interrogations quant à l'utilisations optimale des ressources, la mise à jour des tarifications des actes médicaux et d'analyse budgétaire de l'hôpital.

Les ED/MD antérieurs, au niveau de l'hôpital, de la caisse de sécurité et de l'administration sont le résultat du développement de nombreux ED permettant de répondre à tout ou à une partie des besoins décisionnels exprimés par chaque entité.

C'est pour cela que l'hôpital, la caisse et l'administration éprouve la nécessité de développer un entrepôt de donnée intégré et centralisé permettant la disponibilité des informations soutenant les décisions affectant les analyses budgétaires.

Par conséquent, selon notre démarche, proposée pour la conception ascendante d'un entrepôt de données distribué on peut procéder selon les phases et les activités suivantes :

A. Etude préalable:

Au niveau de cette phase on va traduire une idée en vision de produit fini on répondant aux questions suivantes : « Que va faire le système pour les cadres dirigeants des organisations du SIIO ? » et « A quoi peut ressembler l'architecture d'un tel système ? ».

Au commencement, lors de la **modélisation métier** les organisations se mettent d'accord sur le principal cas d'utilisation métier reflétant l'objectif global du système à concevoir et qui va guider le processus de développement en concert avec l'architecture du système à mettre en place, ensuite, **l'étude des besoins** déterminera les objectifs fonctionnels principaux et secondaires du système et les dimensions et les faits globaux correspondants, et enfin pendant la **phase conceptuelle** l'étude des ED existants sera abordée pour récupérer leurs schémas conceptuels.

Aussitôt cette phase sera conclue, c'est la stratégie d'intégration qui sera définie, les dimensions et les fait du schéma global seront identifiés et les schémas conceptuels des ED récupérés.

A.1. La modélisation métier :

Le processus métier du SIIO décrit à la figure 4.10 à travers un diagramme de cas d'utilisation métiers présente le *système métier intégré* du point de vue de son utilisation et indique la façon dont il rend service à ses acteurs. La modélisation des besoins métiers par le cas d'utilisation est motivée par le faite que cette approche consiste à regarder le système à construire de l'extérieur, du point de vue de l'utilisateur et des fonctionnalités qu'il en attend au lieu de poser la question à l'utilisateur sur ce que le système doit faire. Les cas d'utilisation sont par conséquent très utiles en phase d'analyse des besoins.

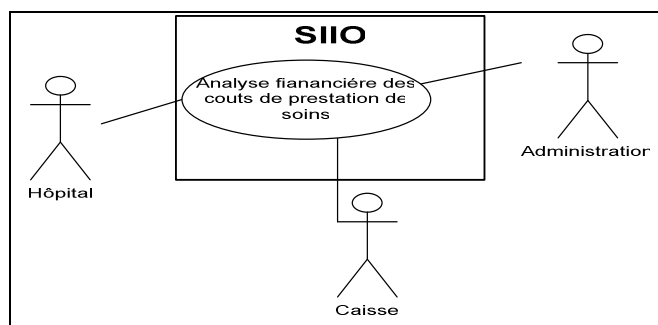


Figure 4.10 - Cas d'utilisation métier.

La spécification du cas d'utilisation métier est illustrée comme suit :

Cas d'utilisation métier : Analyse des coûts des prestations hospitalière.
Portée : SIIO (hôpital -caisse- administration)
Acteurs métier : hôpital, caisse de sécurité, administration de santé publique
<p>Événement déclencheur et déroulement du use case: Suite à une période d'activité d'hospitalisation ou de prestation de soin prodiguées aux profits des patients affiliés à la caisse conventionnée avec l'hôpital, une analyse des activités, regroupant l'hôpital, la caisse et l'administration de la santé publique est menée en vue d'analyser le rendement de l'activité hospitalière et la qualité des soins prodiguées pour proposer une nouvelle nomenclature de tarification des actes médicaux et étudier les protocoles de soin pratiqués par pathologie. Cette situation nécessite des rapports reflétant l'activité de l'exercice et doivent porter sur :</p> <ul style="list-style-type: none"> ▪ La situation financière des factures des hospitalisations par patient, par acte médical pratiqué, par diagnostic, par frais de soins, etc...extraite de l'ED de l'hôpital ▪ La situation financière des prises en charge incluant les remboursements et les cotisations par patient, les factures payées, les factures impayées, etc. extraite de l'ED de la caisse. ▪ La situation des modes d'entrée et de sortie des hospitalisés et leurs pathologies (le nombre de décès, les maladies à déclaration obligatoire, etc.. extraite de l'ED de l'administration a caisse
<p>Parties prenantes et intérêts :</p> <p>L'hôpital : qui prend en charge les patients pour une hospitalisation ou des prestations de soins externes, ses cadres dirigeants veulent avoir pour l'exercice de leurs fonctions, une situation sur les remboursements des frais médicaux, la rentabilité des différents départements par l'analyse des coûts des protocoles médicaux pratiqués au sein des services médicaux, chirurgicaux ou d'exploration fonctionnelle.</p> <p>La caisse : auprès de laquelle le patient est affilié et cotise mensuellement par un pourcentage de son salaire, ainsi elle prend en charge les différents frais de prestations de soins. La caisse veut pouvoir analyser la situation financière des cotisations et des remboursements des frais de soins de ses adhérents, la nature des maladies déclarées par ses adhérents nécessitant une prise en charge.</p> <p>L'administration de la santé publique: veut analyser les indicateurs d'activité de soins sur la population, impliquant les modes d'entrée, les modes de sorties, les maladies à déclaration obligatoire,etc.,..... et mener les actions de prévention devant être prisent en charge pour se prémunir contre des épidémies, par exemple.</p>
<p>Pré conditions (l'état du système avant le commencement du use case): à la fin de chaque exercice budgétaire et dans le cadre de l'élaboration du budget de fonctionnement de l'hôpital, une étude budgétaire regroupant l'hôpital, la caisse, et l'AD sera menée en vue de prévoir le budget nécessaire pour le fonctionnement de l'hôpital, analyser le recouvrement des frais de prestations de soins et l'optimalité des protocole de soins en usage dans l'hôpital.</p>
<p>Post-conditions (l'état du système après le fin du use case) : Les situations des bilans d'activités hospitalières, des remboursements, des cotisations, des modes de sorties et d'entrée, des bénéficières des hospitalisations sont enregistrées et analysées, ainsi les conclusions relatives aux nouvelles tarifications des actes médicaux sont élaborés ainsi que les prévisions du budget de fonctionnement de l'hôpital sont effectués.</p>
<p>Extensions :</p> <ol style="list-style-type: none"> 1. un hôpital peut prendre en charge un patient gratuitement dans le cas d'un projet de recherche scientifique, 2. un hôpital peut avoir plusieurs conventions avec des caisses d'assurance maladies. 3. en plus des remboursements des frais médicaux par la caisse, l'hôpital bénéficie d'une subvention financière annuelle au titre du budget de fonctionnement, évaluée par l'administration de la santé publique, ce qui explique l'intervention de l'administration de

la santé publique pour attester et valider les prévisions financières de l'hôpital.

Règles : Les coûts des soins et des hospitalisations sont calculés selon des normes élaborées conjointement entre l'administration de la santé publique, l'hôpital et la caisse de sécurité sociale.

Ce canevas nous permet d'illustrer quelque élément du processus métiers transcendant les dites organisation qui est la prise en charge sanitaire des patients et la spécification du résultat principal de ce dernier qui est de prodiguer des soins à un coût optimal. Ce canevas a permis en outre d'explicitier les points suivants :

- Les fonctionnalités des systèmes existants, en mentionnant la nature des états et des documents analytiques générés par leurs ED respectifs :
 - ED de l'hôpital : Etat relatif aux hospitalisations;
 - ED de la caisse : Etat des paiements et des remboursements;
 - ED de l'administration : Etat des indicateurs de l'activité hospitalière;
- Les fonctionnalités et les exigences du nouveau système en cour de construction, à travers l'identification de l'objectif principal permettant par la suite d'indiquer les mesures, les dimensions d'analyse, la nature des requêtes à utiliser pour la génération des états périodiques, et qui peut se résumer ainsi : **L'étude des frais des hospitalisations dispensées aux adhérents de la caisse, par mode d'entrée, mode de sorties, diagnostic, acte médical principal, pathologie.**

Sur la base de la modélisation métier, on doit parvenir à délimiter la portée du projet de consolidation afin de saisir l'étendue même de l'architecture.

A.2. Etude des besoins :

L'étude des besoins nécessaires à ces projets, comparée aux méthodologies d'élaboration des ED, fait appel à une approche d'analyse hybride combinant : l'approche dirigée par l'étude des ED/MD existants et l'approche dirigée par le processus métier transcendant les organisations ce qui impliquerait, les besoins des utilisateurs en matière de requêtes d'analyse (Drill cross¹) et les objectifs fixés en commun entre les organisations. Donc il faut spécifier en premier lieu les objectifs de cette intégration en suite étudier les entrepôts de données existants faisant office de source d'alimentation de l'entrepôt global.

Une fois l'objectif stratégique est défini, à savoir « analyse des coûts de prestation hospitalières », les objectifs décisionnels correspondants seront définis ainsi que les critères primaires intéressants dictés par la nature des objectifs décisionnels, qui vont déterminer à leur tour les entités informationnelles à manipuler.

À travers un diagramme de classe on modélise les objectifs stratégiques et décisionnels permettant l'identification des dimensions et des mesures à considérer ainsi que les entités informationnelles existantes dans les ED locaux qui leur correspond (figure 4.11).

¹ Requête de forage transversal, combinant des données provenant de plusieurs tables de faits.

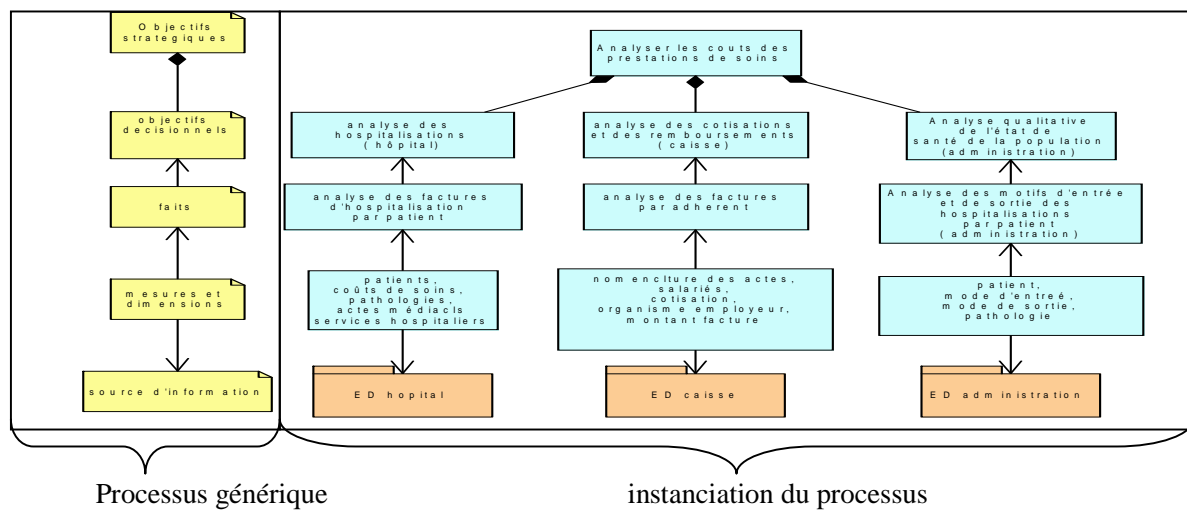


Figure 4.11- Diagramme UML de hiérarchisation des objectifs.

on peut pas dire pour l'instant que les entités analytiques du diagramme récupéré à partir du cas d'utilisation métier sont définitives mais elles seront affinées avec les cas d'utilisation système lors de la phase d'élaboration.

néanmoins vu qu'un processus métier est un classement spécifique d'activités dans le temps et dans l'espace avec un commencement et une fin, et des entrées et des sorties clairement identifiées donc une structure d'action [BJA02], le dit processus métier met en valeur des caractéristiques clés représentant aussi des dimensions de l'entrepôt de données (par exemple patient, acte médical, pathologie, temps, etc.) et des Mesures de l'entrepôt de données (coûts des hospitalisations). Comme ces objets clés du processus métiers peuvent être retrouvés dans tout processus métier, nous proposons un ensemble d'objets dimension qui représentent des dimensions de l'entrepôt des données, le fait qui leur correspond à travers le diagramme des objectifs nous permettra de mettre en exergue les dimensions et la mesure suivantes illustrées dans la figure 4.12;

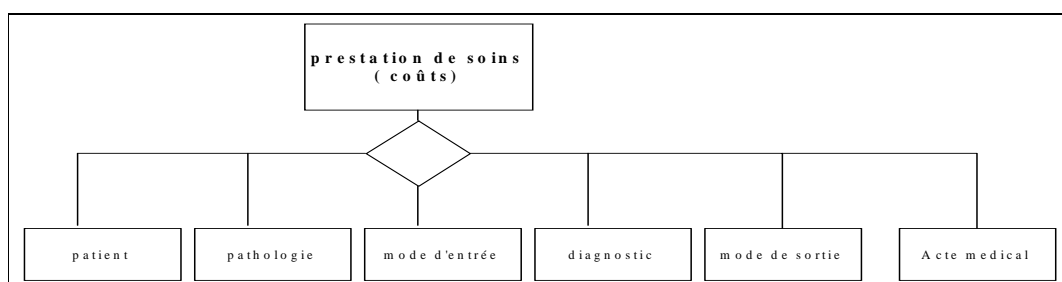


Figure 4.12 - Schéma dimensionnel descendant réconciliant les objectifs des organisations

Il faut mentionner qu'un diagramme de classes fait abstraction des aspects dynamiques et temporels, toutefois pour un modèle complexe, plusieurs diagrammes de classes complémentaires doivent être construits, on peut par exemple se focaliser sur :

- les classes qui participent à un cas d'utilisation,
- les classes associées dans la réalisation d'un scénario précis,
- la structure hiérarchique d'un ensemble de classes.

A.3. Conception :

Durant cette activité on procède à l'extraction des schémas physiques des ED existants en vue de récupérer leur schéma conceptuel dimensionnel à travers un processus de rétro-ingénierie des bases de données multidimensionnelles et dont le résumé de son résultat est illustré dans le tableau 4.5.

Acteurs métiers	Hôpital	Caisse	Administration
Structures analytiques	ED	ED	ED
Fait	Facture d'hospitalisation	Prise en charge financière	Indicateurs de l'activité hospitalière
mesures	Montant de la facture de soins par patient, par acte médical principal, et par diagnostic.	Montant des factures de prestations de soins, des cotisations et des remboursements par adhérent, et par pathologie.	Etat des hospitalisations par patient, par mode d'entrée et de sortie et par pathologie.
Dimension	Patient, acte médical principal, diagnostic, pathologie, date	Adhérent, pathologie, Remboursement, date.	Patient, modes d'entrée (les modes d'entrées correspondent aux diagnostics principaux), mode de sortie, date
Hiérarchies des dimensions	Diagnostic principal et élémentaire, acte médical principal	Pathologie par famille, acte médical principal	Pathologie par famille

Tableau 4.5 -Analyse des schémas physiques des ED existants

Ce processus de rétro-ingénierie doit permettre d'aboutir à la récupération des schémas conceptuels illustrés dans la figure 4.13, et ceci après standardisation des nominations et omission de toute considération des aspects physiques des schémas, à savoir le modèle de données d'implémentation, les indexes, les structures de données, etc...

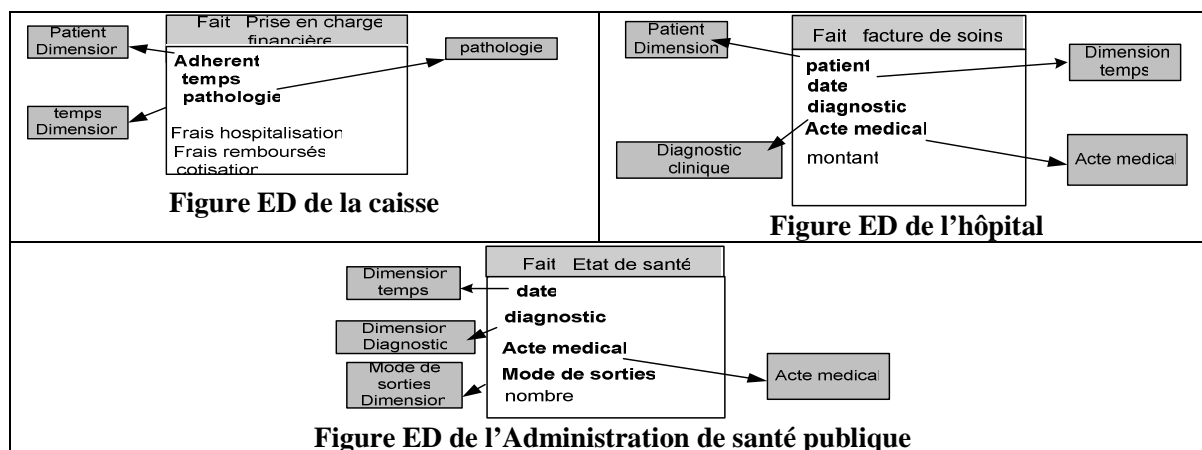


Figure 4.13 -Schémas conceptuels des ED existants.

En réponse, aux objectifs fonctionnelles de l'ED global permettant d'esquisser les entités du schéma dimensionnel intégré et après récupération des schémas conceptuels des ED/MD existants et selon la nature des relations métiers [TJ04] entre les organisation qui spécifient les restrictions d'accès

sur les informations mise en commun, on peut déterminer comment mettre en place la stratégie d'intégration centralisée des schémas dimensionnels locaux (figure 4.14).

A cet effet [CAVN05] [WE04] préconise **la re-conception globale** du schéma intégré on procédant par collection, comparaison et fusion des schémas ou **la désignation d'un ED comme référence** en procédant par la comparaison et la fusion des schémas ou **la synchronisation** des données de base manipulées par les organisation du SIIO.

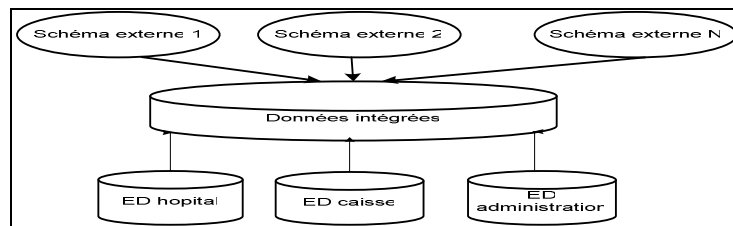


Figure 4.14- Architecture de référence.

B. Elaboration :

Permet de spécifier le cas d'utilisation métier par des cas d'utilisation systèmes choisis par leur degré d'importance permettant de généré en concert avec le diagramme des objectifs déduits lors de la phase d'inception, le **model multidimensionnel objet global**.

Les schémas conceptuels seront confrontés en vue de dissiper tous conflits d'ordre sémantique ou structurelle et confrontés au schéma global pour définir les correspondances verticales qui s'imposent. A l'issue de cette phase, on doit être en mesure d'estimer les ressources nécessaires à l'intégration des ED.

B.1. Modélisation métier et étude des besoins :

Selon la hiérarchie des objectifs et des buts on doit déterminer le (les) cas d'utilisation système(s) dérivé(s) (fig 4.15) (ou le scénario principal du cas d'utilisation système) du cas d'utilisation métier devant être prient comme spécification de notre système intégré et sur lesquels l'ensemble des tâches de spécification des besoins, de conception, d'implémentation et de test se basent. Par ordre de priorité ces cas d'utilisation seront classés et leur développement constituera une itération dans leur la démarche de conception. L'analyse de ces cas d'utilisation permet de faire ressortir le diagramme de classe ou le model dimensionnel objet global représentant les entités analytiques consenties par les organisation du SIIO.

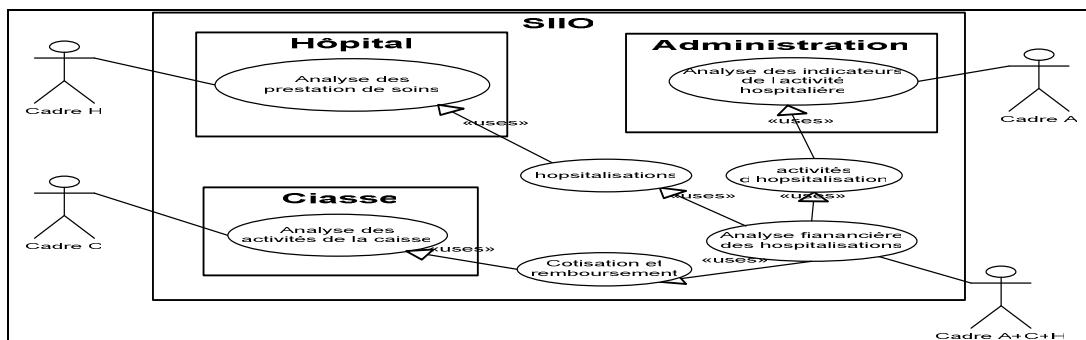


Figure 4.15- Cas d'utilisation systeme.

Selon [JF98] Un cas d'utilisation système est projeté d'accomplir quelques objectifs tactiques pour un acteur, ou aider dans l'accomplissement d'un objectif tactique. Le concept du cas d'utilisation se focalise sur le point de vue de l'utilisateur au sujet de ce que le système est supposé donner en réponse à l'entrée de l'utilisateur. Ce dernier est supposé fournir une réponse qui aura une valeur correspondante à une hiérarchie d'objectifs tactiques et stratégiques. La figure suivante 4.16 illustre le rapport entre un cas d'utilisation système et une hiérarchie de buts et d'objectifs.

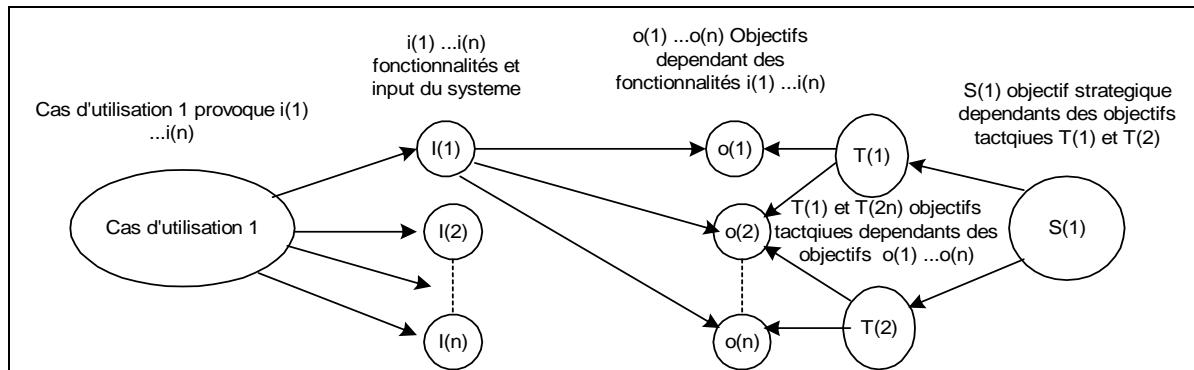


Figure 4.16: Cas d'utilisation système[JF98].

B.2. Conception :

Le cas d'utilisation système permet de mettre en exergue les éléments du schéma dimensionnel descendant à travers la correspondance entre les objectifs et le cas d'utilisation système (fig 4.17) et sur la base du diagramme de classe figure 4.12 .

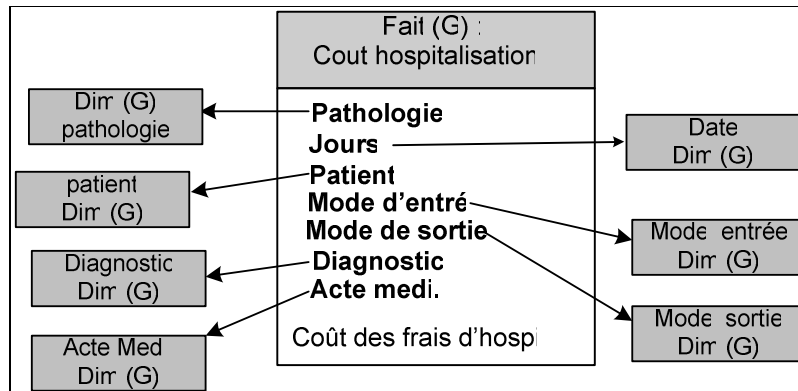


Figure 4.17 - Schéma dimensionnel descendant réconciliant les objectifs des organisations

Après avoir déterminé le schéma conceptuel global à partir du diagramme des but il faut s'orienter vers les schémas conceptuels récupérés auparavant afin de résoudre tous conflits et préparer l'identification des correspondances verticales pour l'alimentation de l'ED intégré.

L'intégration des schémas dans un schéma global exige que les concepts équivalents soient identifiés et fusionnés et les concepts dissemblables soient tenu à part comme des éléments séparés dans le schéma global. L'intégration peut aussi exiger une restructuration des schémas par exemple deux entités dans des schémas individuels peuvent être généralisées dans une entité combinée de haut niveau d'abstraction. D'où

- l'Identification des liaisons horizontales entre les ED permet de :

- Dissiper tous conflits entre les mesures, attributs, dimensions et faits afin de déceler les dimensions et faits communs qui pourront jouer le rôle d'entités à rendre conformes ou compatibles.
- Détecter les redondances et les inconsistances
- les relations sémantiques [AJF02] (généralisation, dérivation, flux, association) dimension-dimension, fait-fait et dimension-fait permettant de spécifier comment utiliser ces relations pour réaliser des requêtes entre les schémas dimensionnels à la lumière des objectifs définis auparavant.

	ED Hôpital			ED Caisse			ED Administration			
fait et dimensions universels	Fait Facture	Dim patient	Dim Diagnostic	Dim Acte médical	Fait facture	Dim adhèrent	Dim pathologie	Fait Hospitalisation	Dim Mode d'entrée	Mode de sortie
Dim G.1										
Dim G.2										
Dim G.3										
Fait G.1										
Autres										

Tableau 4.6- Matrice d'analyse des entités analytiques redondantes

La phase de correspondance est adressée vers l'identification de entités informationnelles existantes dans les ED, qui peuvent être réutilisés dans le schéma global.

- *Identification des liaisons verticales* : spécifier les associations entre les schémas locaux des ED et le schéma global en vue d'établir les correspondances ou le mapping entre les schémas tel qu'il est illustré dans la figure 4.18.

[wp03] démontre le rôle des correspondances verticales et horizontales pour définir les relations entre les schémas intermédiaires et le schéma globales. les **correspondances horizontales** sont un résultat du processus de retro ingénierie. Ils exposent les relations entre les schémas intermédiaires et tombent dans trois catégories possibles : syntactique, sémantique et instance. D'autre part, les **correspondances verticales** spécifient explicitement les différents rapports entre les schémas locaux et global.

A noter que le passage entre les deux niveaux est exécuté par des techniques comme le changement de nom, la traduction, les résolution de conflits qui sont essentiellement des transformations de schémas d'où la nécessité de définir les configurations de la **transformation de schéma** et qui peuvent être formalisées comme une chaîne de transformation.

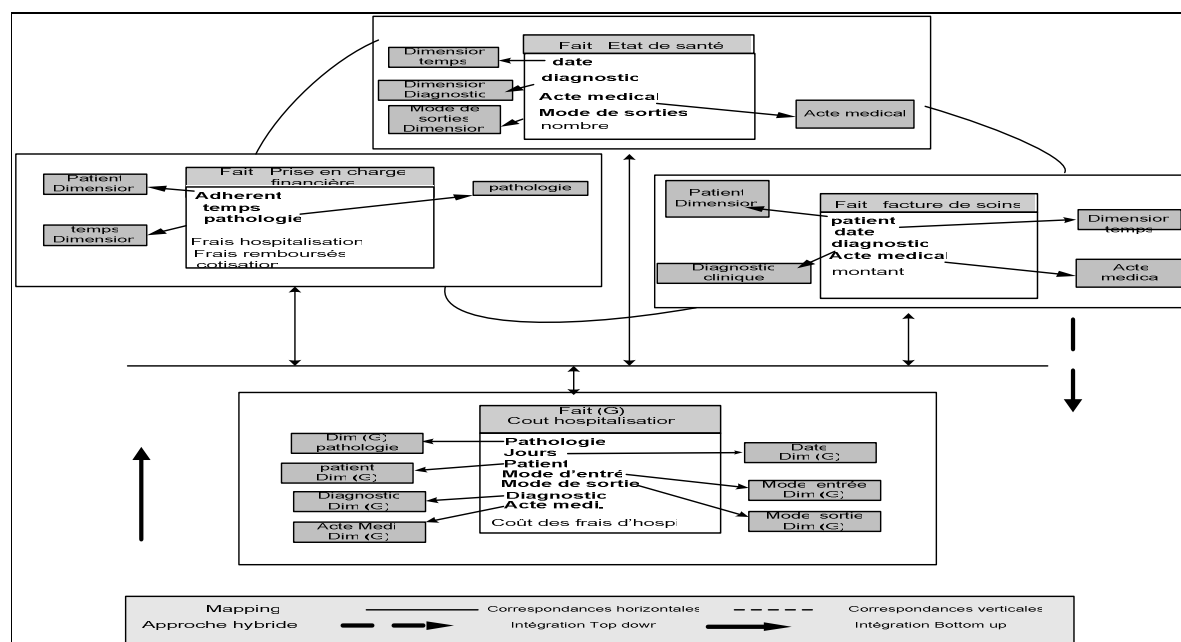


Figure 4.18- Mapping des schémas dimensionnels locaux et global.

La méthodologie la plus communément utilisée pour la transformation des modèles de données est la correspondance entre la description des données locales et la définition des données cibles en accord avec les règles des modèles de données source et cible [RA01]. Ce qui veut dire que le modèle de données local est mappé au Modèle Global de Données un à un avec toutes les transformations et agrégations mentionnées dans le meta-données.

L'état de la base de données dans le modèle de données source et cible sont équivalents s'ils peuvent être mappés en un même état dans un modèle abstrait ou un meta modèle. Les méthodes formelles pour l'établissement de cette équivalence dépendent du consentement d'employer un meta modèle.

Une de ces méthodes est de concevoir le modèle de données globale de telle manière que chaque requête qui peut être effectuée sur un modèle global peut l'être sur un modèle local. Donc la séquence des événements peut être définie, pour transformer les opérateurs du modèle global en opérateurs du modèle de données local, si les deux modèles de données locales et globales sont complets et reliés un à un, ils peuvent donc traiter les mêmes requêtes parce qu'ils sont équivalents.

En conclusion, cette phase aura pour résultat le **schéma conceptuel de l'ED/MD distribué**, la **standardisation des règles et les termes métiers** à travers le processus métiers transcendant les organisations et par conséquent la **standardisation des meta données**.

C. La construction :

La phase d'implémentation implique la construction logique et physique du schéma intégré, le développement du processus ETL et la standardisation de l'environnement de restitution, qui ne fait *pas l'objet de ce travail et qui doit être inscrit comme une perspective devant être prise en charge par la suite*. Néanmoins on peut les résumer à travers ces points :

- **Définition de la correspondance entre le système source et l'ED/MD :** Avant de définir le processus ETL et les spécifications de l'extraction, il importe que les transformations et les réconciliations soient enregistrées dans un tableau de correspondance entre l'ED/MD source vers l'ED/MD cible spécifiant les règles ETL.
- **Conception du processus ETL :** Dans l'approche d'intégration centralisée, la conception du processus ETL suppose un processus ETL est créé impliquant le chargement des données de l'ED/MD source vers l'ED/MD intégré cible. Les autres facteurs clés de ce processus sont : la standardisation du meta données est disponible, les données sont intégrées et leurs consistante et leurs qualité sont améliorées,
- **Etc....**

4.5. Conclusion

Au niveau de cette proposition, une démarche de conception d'un ED distribué dans un environnement inter organisationnel a été présentée en utilisant une modélisation orientée objet et un paradigme d'intégration hybride combinant les objectifs stratégiques et les schémas des ED locaux. Cependant le travail effectué doit être approfondi afin d'appréhender les perspectives suivantes :

- Le cadre de référence qui relie les typologies des SIIO au processus d'intégration des ED autonomes selon une approche centralisée ou décentralisée;
- Les techniques de mapping entre les schémas dimensionnels des ED locaux ou le schéma global.

Conclusion générale

Avec l'avènement des SIIO, le problème d'intégration des entrepôts de données est devenu un thème et une problématique avérée. En plus de la taille et de l'ampleur de ces entrepôts de données qui augmentent continuellement, engendrant des masses d'information hétérogènes et distribuées, le processus d'intégration est tributaire de la nature des relations entretenues entre les organisations, les technologies utilisées et le caractère organisationnel de ces environnements.

A cet égard, nous avons mis en relief une proposition d'une démarche de conception ascendante d'un ED distribué dans un contexte inter-organisationnel en spécifiant les caractéristiques majeures de son cycle de vie, et que nous avons tenté d'expliquer à travers les deux étapes ci-après:

- (i) En premier lieu, synthétiser l'état de l'art lié :
 - aux systèmes d'information inter-organisationnels, en exposant les définitions majeures et les typologies les plus répandues de ce nouveau concept afin d'aboutir aux architectures principales d'interopérabilité de données entre les organisations formant le SIIO et qui se base sur la nature du flux d'information qui circule entre les dites organisations;
 - aux entrepôts de données en élaborant une synthèse sur les concepts et les variantes des méthodologies d'élaboration basées sur les approches d'analyse et les stratégies d'implémentation afin de faire ressortir la relation entre ces méthodologies aux architectures distribuées des ED.
 - aux entrepôts de données distribués en général et les spécificités dictées par l'environnement inter-organisationnel quant aux choix architecturaux et à la conduite du processus de conception ascendante ou d'intégration des bases de données multidimensionnelles hétérogènes et autonomes en spécifiant ces facteurs prédominants.
- (ii) En deuxième lieu, mettre en évidence une démarche pour la prise en charge de la consolidation et de la conception ascendante d'un entrepôt de données distribué dans un environnement inter-organisationnel en utilisant une notation orientée objet à savoir UML et en faisant usage du processus unifié sous-jacent, étayée par une étude de cas regroupant des organisations actives dans le domaine de la santé et formant un SIIO vertical et stratégique ;

Néanmoins, ce travail n'a pas la prétention d'être exhaustif et notre démarche doit être affinée, par la prise en considération, dans des travaux futurs, des points suivants :

- **Cadre de référence reliant la conception ascendante ou l'intégration des schémas multidimensionnel et les SIIO** : nous pouvons affirmer qu'une démarche de conception bottom up des ED autonomes appartenant à des organisations formant un SIIO est conditionnée par des facteurs d'ordre organisationnels, managériaux et technologiques dictés par la nature complexe des environnements multi-organisationnel, imposant ainsi deux types d'approches majeures d'intégration à savoir l'approche centralisée ou décentralisée.

On pourrait aboutir à la même conclusion technique car lors de notre étude des méthodologies d'élaboration d'ED et à un certain niveau d'abstraction, il est apparu que la majorité des méthodologies sont divisées en deux catégories selon les segments d'activités majeurs : l'architecture et l'implémentation, où l'architecture établit une vision de l'ED, planifie la séquence d'implémentation, adresse les besoins en informations métiers, la configuration de l'ED, des standards et des conventions sur un panel d'organisations et l'implémentation est élaborée comme une série de petits projets, chacun délivre un sous-ensemble de l'ED avec des valeurs de mesures métiers. A cause de cette dualité, et dans un contexte inter-organisationnel on peut observer que toute relation stratégique requière une approche centralisée d'intégration ou fortement couplée, par contre, toute relation opérationnelle qui par définition est limitée dans le temps, se traduirait donc par une approche décentralisée d'intégration ou faiblement couplée. Or ce constat nous permet de remarquer que notre problématique, d'un point de vue technique, est similaire à toutes autres problématiques d'intégration intra-organisationnel néanmoins le degré de complexité attribué à l'environnement inter-organisationnel est dû aux aspects caractérisés ces environnements complexes liés à la gestion, à l'organisation et au management, à cet effet, il doit y'avoir un cadre de référence qui relie les typologies des SIO au processus d'intégration des ED.

• **Le support méthodologique de la conception bottom up** : L'un des problèmes majeurs soulevés par ce travail est la définition d'un support méthodologique de la conception ascendante des ED distribués. En perspective, cette démarche devrait accéder au niveau des méthodes formelles caractérisées par les trois points suivants, dont le premier est pris en charge lors de ce travail, à savoir :

1. **La spécification d'une démarche** : définissant les phases, les activités, les tâches de mise en œuvre ainsi que l'allure que prend le cycle de vie de la démarche, et qu'il faut compléter par deux autres aspects ;

2. à **la modélisation** de cette démarche par la spécification des techniques de transformation et de passage entre les différents niveaux commençant par la formulation des besoins aux différents niveaux d'abstraction conceptuel, logique et physique, et ce, par l'utilisation d'un formalisme tel que UML assujéti aux spécifications du Processus Unifié ;

3. et la **spécification des moyens et outils de mise en œuvre.**

Ces trois points réunis donneront naissance dans le cadre de notre problématique à une méthode de conception ascendante des entrepôts de données distribués tenant compte des spécificités des environnements inter-organisationnels.

Bibliographie:

- [AFG05] Aurelio Ravarini, Federico Pigni, Giacomo Buonanno : “Exploring the Role of Inter-Organizational Information Systems within SMEs Aggregations”, 18th Bled eConference eIntegration in Action Bled, Slovenia, June 6 - 8, 2005.
- [AJF02] A. Abelló, J. Samos, and F. Saltor. On relationships Offering, New Drill-across Possibilities. In ACM Fifth Int. Workshop on Data Warehousing and OLAP (DOLAP 2002), pages 7–13, 2002.
- [AK99] Amin Y. Noaman Ken Barker : A Horizontal Fragmentation Algorithm for the Fact Relation in a Distributed Data Warehouse, ACM, 1999.
- [BA02] Benoit A. Aubert et Aymeric Dussart, « Systèmes D’information Inter-Organisationnels », Mars 2002, Centre Interuniversitaire de Recherche en Analyse des Organisations, Quebec, Canada.
- [BHS98] Markus Blaschka, Carsten Sapia, Gabriele Höfling, Barbara Dinter , Finding YourWay through Multidimensional Data Models1, FORWISS (Bavarian Research Center for Knowledge-Based Systems) Orleansstr. 34, D-81667 Munich, Germany, IEEE, 1998.
- [BJA02] Beate List, Josef Schiefer, A Min Tjoa Process-Oriented Requirement Analysis Supporting the Data Warehouse Design Process A Use Case Driven Approach, , Institute of Software Technology (E188) Vienna University of Technology, 2002.
- [CAVN05] Chuck Ballard, Amit Gupta, Vijaya Krishnan, Nelson Pessoa, Olaf Stephan, Data Mart Consolidation: Getting Control of Your Enterprise Information, IBM Redbooks July 2005, International.
- [CP84] CERI, S., AND PELAGATTI, G. Distributed Databases Principles and System. McGraw-Hill, New York, 1984.
- [CW01] Colin White, The Federated Data Warehouse Article published in DM Review Magazine , March 2000 Issue, <http://www.dmreview.com/index.cfm>.
- [DH00] Douglas Hackney, How do we go about designing and building a federated ED architecture? Data Warehouse Delivery : Federated FAQs, Published August 15, 2000, <http://www.dmreview.com/master.cfm?NavID=55&EdID=2077>.
- [DH98] Douglas Hackney: Architectures and Approaches for Successful Data Warehouses Copyright© 1998 Enterprise Group, Ltd.
- [DJ98] Dyché, Jill. Scoping Your Data Mart Implementation. DBMS. Aug; 1998.
- [EA03] Emily (Rong) Liu et Akhil Kumar, “LEVERAGING INFORMATION SHARING TO INCREASE SUPPLY CHAIN CONFIGURABILITY”, Department of Supply Chains and Information Systems, Smeal College of Business, Penn State University Park, USA, 2003, Twenty-Fourth International Conference on Information Systems.
- [EOM05] Eom, S. B., (2005): An Introduction to Inter-Organizational Information Systems with Selected Bibliography, "Inter-Organizational Information Systems in the Internet Age", Idea Group Publishing Hershey.
- [FH99] Feng Li & Howard Williams : “New Collaboration between Firms: The Role of Interorganizational Systems”, Proceedings of the 32nd Hawaii International Conference on System Sciences – 1999.
- [FY00] F.-Y. VILLEMIN, « processus d’informatisation, introduction aux systèmes d’information » , cours du CNAM-CEDRIC,2000.
- [GC04] Georges El Helou et Charbel Abou khalil : « Data Mining Techniques d’extraction des connaissances » Module 4.1 - Management et NTIC Professeur : Mélissa Saadoun Projet soutenu le 16 février 2004.

[GS98] Gardner, Stephen R. Building the Data Warehouse. Communications of the ACM, 41 (9): 52-60. Sept; 1998.

[HA05] [Hugh J. Watson, Thilini Ariyachandra, “ Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures”, July 2005, http://www.terry.uga.edu/~hwatson/DW_Architecture_Report.pdf](http://www.terry.uga.edu/~hwatson/DW_Architecture_Report.pdf)

[HGR00] Helena GRAZZIOTIN RIBEIRO, Un service de règles active pour fédérations de bases de données, thèse de doctorat, juin 2000.

[HONG02] Ilyoo B. Hong , Classifying B2B Inter-Organizational Systems : A Role Linkage Perspective, University of California at Los Angeles, USA, 2002.

[INM05] W. H. Inmon , Building the Data Warehouse, Fourth Edition, Published by Wiley Publishing, Inc. 10475 Crosspoint Boulevard Indianapolis, IN 46256 www.wiley.com, Copyright © 2005 by Wiley Publishing, Inc., Indianapolis, Indiana.

[IT02] IT in the Organization, Information Technologies: Concepts and Management 2002.

[JBR00] Jacobson, G. Booch, J. Rumbaugh, The Unified Software Development Process c/o Addison-Wesley, publié en 1999.

[JD00] Dr. James Thomann and David L. Wells, Evaluating Data Warehousing Methodologies: Objectives and Criteria, published in The Journal of Data Warehousing, January 2000.

[JFJA00] José Samos , Fèlix Saltor , Jaume Sistac, and Agustí Bardés: “Database Architecture for Data Warehousing: An Evolutionary Approach”, 2000.

[JF98] Joseph M. Firestone, Ph.D, Dimensional Object Modeling. Executive Information Systems, Inc., White Paper n. 7, April 1998 (available at <http://www.dkms.com/DOM.htm>).

[JMM05] Designing Data Warehouses: From Business Requirement Analysis to Multidimensional Modeling, Jose-Norberto Mazón¹, Juan Trujillo¹, Manuel Serrano², and Mario Piattini, 2005.

[JP01] Jinsoo Park , Schema Integration Methodology and Toolkit for Heterogeneous and Distributed Geographic Databases, Information and Decision Sciences Department, Carlson School of Management, Working Paper, Last revised on August 2001.

[KIM02] Kimball and M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. John Wiley & Sons, Second edition, 2002.

[KIM96] R. Kimball & L. Reeves & M. Ross & W. Thornthwaite «Concevoir et déployer un Data Warehouse ». Edition Eyrolles, 1996.

[KIM98] Kimball, Ralph; Reeves, Laura; Ross, Margy & Thornthwaite, Warren. The Data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses. New York, John Wiley & Sons, 1998.

[KK90] Kamel Nabil N. Kamel , The Federated Database Management System: An Architecture of Distributed Systems for the 90's, IEEE, 1990, The research reported here was conducted with funds administered by the Naval Postgraduate School Research Council.

[KV96] Kumar, K. et Van Dissel, H.G., “ Sustainable Collaboration : Managing Conflict and Cooperation in Interorganizational Systems ” MIS Quarterly, Sep 1996, 20(3), 279-300.

[LR03] Luca Cabibbo and Riccardo Torlone , Dimension Compatibility for Data Mart Integration, In 15th Int. Conference on Scientific and Statistical Database Management (SSDBM' 03), 2003.

-
- [LR04] L. Cabibbo and R. Torlone. On the Integration of Autonomous Data Marts. In 16th Int. Conference on Scientific and Statistical Database Management (SSDBM' 04), pages 223–234, 2004.
- [LR05] Luca Cabibbo and Riccardo Torlone, Integrating Heterogeneous Multidimensional Databases.
- [MA99] Michael Boehnlein, Achim Ulbrich-vom Ende , Deriving Initial Data Warehouse Structures from the Conceptual Data Models of the Underlying Operational Information Systems, , University of Bamberg, Germany, 1999.
- [MB00] Michael Böhnlein, Achim Ulbrich-vom Ende , Business Process Oriented Development of Data Warehouse Structures University of Bamberg, Feldkirchenstr. 21, D-96045 Bamberg, Germany ,2000.
- [MC03] Mohamed Frendi, Camille Salinesi, Requirements Engineering for Data Warehousing, Centre de Recherche en Informatique Université de Paris 1, Panthéon Sorbonne 90, rue de Tolbiac, 75013 Paris, France, 2003.
- [MGSR98] Matteo Golfarelli , Stefano Rizzi : A Methodological Framework for Data Warehouse Design, Proceedings ACM First International Workshop on Data Warehousing and OLAP (DOLAP 98), Nov. 7, 1998, Washington, D.C., USA.
- [MH98] Mylène Leitzelman, Henri Dou : « Typology of Information Systems, Essai de typologie des Systèmes d'Informations », International Journal of Information Sciences for Decision Making, Page 55, N°2 - April 1998.
- [MO00] Mario Piattini , Oscar Díaz, Advanced Database Technology and Design, Editors Artech House Boston . London, www.artechhouse.com, 2000, Distributed Databases, Peter McBrien and Alexandra Poulouvassilis.
- [NR02] 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, Mémoire présenté par Nathalie RYSER BOLOGNINI en vue de l'obtention du Diplôme post grade en informatique et organisation, Années académiques 2000-2002 Sous la direction du Prof. Thibault ESTIER.
- [OV99] T. Özsu, P. Valduriez: Principles of Distributed Database Systems. Prentice Hall, 2ème édition, 666, pages, 1999.
- [PL03] Ping Wanga, Lihua Huangb : “Business Value of Vertical IOS: An Extended ROI Case Study of a Networking Company in China” , 7th Pacific Asia Conference on Information Systems, 10-13 July 2003, Adelaide, South Australia.
- [PO02] Ponelis.s, chapter 3 data marts as management information delivery mechanisms university of Pretoria, 2002.
- [PQ05] Professeur Patrick Quinlan, Un projet d'intégration de bases de données hétérogènes, mise en pratique dans un environnement professionnel. Haute École de Gestion, Laboratoire de Knowledge Management, Genève, novembre 2005.
- [PS04] Petri I. Salonen, “Evaluation of a Product Platform Strategy for Analytical Application Software”, Helsinki School Of Economics, print 2004, ISBN 951-791-867-4.
- [PVK98] Poe, Vidette; Klauer, Patricia & Brobst, Stephen. Building a data warehouse for decision support. New Jersey, Prentice Hall PTR, 1998.
- [RA01] R.A. Moeller , Distributed Data Warehousing Using Web Technology, ISBN: 0814405886 , AMACOM © 2001, 384 pages.
- [RZ96] R. Hull, G. Zhou, A Framework for Supporting Data Integration Using the Materialized and Virtual Approaches, ACM Sigmod Record, (June, 1996) 481-492.
- [SAXY00] Shuigeng Zhou, Aoying Zhou, Xiaopeng Tao, and Yunfa Hu “Hierarchically Distributed Data Warehouse” Computer Science Department, Fudan University, Shanghai, 2000, China.

-
- [SC02] State of Connecticut, Data Management and Data Warehouse: Domain Technical Architecture June 6, 2002, Version2.0.
- [SL90] Amit P. Sheth, James A. Larson Federated Database Systems For Managing Distributed, Heterogeneous, And Autonomous Databases', *Acm Computing Surveys*, Vol. 22, No. 3, September 1990.
- [SRUH99] Susanne Busse, Ralf-Detlef Kutsche, Ulf Leser, Herbert Weber, Federated Information Systems: Concepts, Terminology and Architectures, *Forschungsberichte des Fachbereichs Informatik*, Bericht Nr. 1999-9, Technische Universität Berlin, Fachbereich 13 Informatik.
- [TES00] O. Teste « Modélisation et manipulation d'entrepôts de données complexes et historisés ». Thèse de doctorat. Institut de recherche en informatique de TOULOUSE. Publication : Laboratoire IRIT- Pôle SIG. 2000.
- [TJ04] Timon C. Du, Jacqueline Wong, "Designing Data Warehouses for Supply Chain Management", *Proceedings of the IEEE International Conference on E-Commerce Technology*, 2004.
- [TH99] Ph. Thiran, J-L. Hainaut, S. Bodart, A. Deflorenne, J-M. Hick, "Interoperation of Independent, Heterogeneous and Distributed Databases. Methodology and CASE Support: the InterDB Approach" in *Proceedings of CoopIS'98*, IEEE, New-York, August 1998.
- [TT98] Tulu Tanrikorur, Enterprise DSS Architecture: A Hybrid Approach, Article published in *DM Review Magazine*, February 1998 Issue, <http://www.dmreview.com/index.cfm>.
- [WE03] Wayne Eckerson, Four Ways to Build a Data Warehouse, Director of Research, TEDI, <http://WWW.tEDI.org/Publications/display.aspx>, 2003/05/12.
- [WE04] Wayne Eckerson, In Search of a Single Version of Truth: Strategies for Consolidating Analytic Silos , TEDI report series , A u g u s t 2 0 0 4.
- [WP00] Walter Pereira Karin Becker: A Methodology Targeted at the Insertion of Data Warehouse Technology in Corporations, 1999.
- [WP03] Willem-Jan Van Den Heuvel, Philippe Thiran, a methodology for designing federated enterprise models with conceptualized legacy wrappers, *ICEIS 2003 - databases and information systems integration*.
- [YP96] Yves Pigneur: Systèmes d'information, entreprise étendue et commerce Electronique, A Framework for New Informations Systems; F. Bodart & al. (eds), *The Future of Information systems: Challenge and Pitfalls*; FUNDP, Namur, October 1996, pp. 61-102.
- [ZH97] J. L. Zhao, Schema Coordination in Federated Database Management: A Comparison with Schema Integration, *Decision Support Systems*, 20(3), (1997), 243-257.