# *MODELISATION DE LA REPARTITION DES DONNEES D'UN DATA WAREHOUSE*

**Karima TEKAYA**
Assistante en informatique
Karima.Tekaya@.isi.rnu.tn

**Abdelaziz ABDELLATIF**
Maître-assistant en Informatique
abdelaziz.abdellatif@fst.rnu.tn

**Adresse professionnelle**
Faculté des sciences de Tunis, Département informatique,
Campus universitaire - 2092 Manar II

**Résumé** : Les utilisateurs des data warehouses ne cessent d'augmenter. A l'image des entreprises, ces utilisateurs sont de plus en plus répartis géographiquement sur plusieurs sites. Les data warehouses centralisés ne sont donc plus adaptés à ce genre d'entreprises. Pour répondre à ce nouveau besoin, nous avons proposé une démarche de modélisation de la répartition des données d'un Data Warehouse. Celle-ci, se base essentiellement sur un ensemble de matrices permettant la modélisation de l'intégration logique des données du Data Warehouse d'un côté et leur répartition entre les différents Data Marts de l'organisation d'un autre côté.

**Summary**: The users of Data Warehouses do not cease increasing. With the image of the companies, these users are divided more and more geographically on several sites. Centralized Data Warehouses thus are not adapted more to this kind of companies. To meet this new requirement, we proposed a methodology of modelling the distribution of the data of a Data Warehouse. This one is based primarily on a set of matrices allowing the modelling of the integration of the data in a Data Warehouse. Secondly, their distribution between different Data Marts.

**Mots clés** : Data warehouse, Data mart, Modélisation, Répartition, Intégration.

## 1- INTRODUCTION

Un Data Warehouse (DW) répond aux problèmes de données surabondantes et localisées sur de multiples systèmes hétérogènes. Le DW est un entrepôt de données permettant un stockage intermédiaire des données issues des applications de production, dans lesquelles les utilisateurs finaux puisent avec des outils de restitution et d'analyse.

L'intégration du DW dans une structure unique a pour but d'éviter aux données concernées par plusieurs sujets d'être dupliquées. Le DW est fragmenté en plusieurs bases appelées Data Mart(DM). Un Data Mart est l'implémentation d'un DW pour un domaine bien spécifique. En effet, c'est un sous ensemble d'un DW [1].

On peut avoir plusieurs Data Mart au sein d'une même entreprise [2]. Ces data marts peuvent être répartis par département, les données utilisées sont extraites à partir du DW principal (centralisé).

## 2- PROBLEMATIQUE

Un système d'information est composé d'une composante décisionnelle et d'une composante opérationnelle. Le système d'information opérationnel englobe toutes les informations concernant l'activité de l'entreprise, ces données sont stockées dans une base appelée base de production.

Le système d'information décisionnel englobe des informations provenant de bases de production ou de sources diverses et externes à l'entreprise servant comme support d'aide à la décision. L'ensemble de ces informations est stocké dans le DW.

Le système d'information est en évolution, il fait face aujourd'hui aux problèmes de décentralisation des entreprises, les utilisateurs sont de plus en plus nombreux, ils exercent des activités hétérogènes et appartiennent généralement à des sites éloignés géographiquement. Ceci a eu comme conséquence la décentralisation du système décisionnel.

Les besoins informationnels et les utilisations des données peuvent être différentes d'un site à un autre. De ce fait, une organisation centralisée des données peut être non adéquate à cette nouvelle architecture répartie. Un DW réparti pourra répondre plus efficacement aux besoins des utilisateurs. Les données peuvent être organisées par sujet et une meilleure utilisation du DW est garantie. La répartition d'un DW en plusieurs DM est la solution la plus adéquate pour un système différé puisqu'elle permet de rapprocher les données aux utilisateurs et améliorer l'organisation des données.

Plusieurs contraintes techniques peuvent être rajoutées :

- La communication des informations stratégiques aux différents décideurs s'avère de plus en plus coûteuse de point de vu financier (coût des accès) et temporel (temps d'accès).
- Le DW est centralisé dans une base unique, le stockage des données sur un ordinateur central peut souffrir d'une très longue charge de traitement ce qui peut influer sur sa performance.
- En plus, le volume du DW augmente très rapidement ce qui ralentit les accès et gonfle le stockage [7], [8] et [9].
- D'autre part, la centralisation des données pourrait devenir le point sensible du système informatique.

De ces faits, la centralisation d'un DW peut se refléter négativement sur sa performance et ses fins. Pour faire face à ces différents problèmes, le système opérationnel opte pour l'adaptation des bases de données réparties. Le système d'information décisionnel opte pour la répartition du DW en DM. Plusieurs démarches de modélisation ont été proposées pour modéliser les bases de production réparties. Par contre, aucune

démarche exhaustive n'a été proposée pour la modélisation de la répartition des données d'un DW.

## 3- CONTRIBUTION

La contribution apportée par cet article est de proposer une démarche de modélisation de la répartition des données d'un DW. Celle-ci se base essentiellement sur les niveaux de modélisation classiques, en ajoutant un ensemble de concepts de base, intégrer de nouveaux modèles et proposer un formalisme de présentation.

Dans la section suivante, nous allons citer l'état de l'art. Dans la section 5, nous allons proposer les concepts de base de notre démarche, les modèles nécessaires et le formalisme proposé.

## 4- ETAT DE L'ART

Les méthodologies trouvées dans la littérature ont généralement pour objectif d'intégrer le DW dans une structure unique et ont comme résultat un entrepôt de données centralisé [3] et [4]. Cet entrepôt est appelé DW, s'il est généralisé aux activités de l'entreprise, ou bien DM s'il est spécifique à un département particulier.

On a constaté dans l'état de l'art que tous les travaux concernés par la modélisation de la réparation des données des DW sont orientés vers la modélisation physique [5] et [6]. Des algorithmes de répartition verticale des données ont été proposées dans [14] et [15]. L'idée de répartition des données d'un DW a été évoquée par Noaman, A.Y. et K. Barker dans [7] et [8]. Ils se sont basés sur l'architecture ANSI/SPARC pour la modélisation des données des DW. La démarche proposée par ces auteurs se base essentiellement sur l'approche Top/Down. Ils ont aussi développé un algorithme de fragmentation horizontale des tables de faits dans [9].

Dans [4] une démarche exhaustive a été proposée pour modéliser l'intégration des données d'un DW (Figure 1). Celle-ci se base essentiellement sur l'ajout d'un modèle d'intégration des données permettant de modéliser l'intégration des données (MID) dans le DW. Ce modèle sert à identifier pour les données du modèle logique de données obtenu : leurs sources de données, les transformations éventuelles qu'elles doivent subir, leurs modes de rafraîchissement dans le DW et leurs fréquences d'utilisation. Dans [12] une démarche exhaustive de modélisation de la répartition des données d'une base de production a été bien développée (Figure1). Celle-ci se base essentiellement sur l'ajout d'un modèle de répartition des données (MRD) en tenant compte d'un processus de répartition et en intégrant un programme d'optimisation des différentes allocations en fonction des débits binaires échangés, les fiabilités des échanges et les caractéristiques du réseau. Dans [16], une adaptation du modèle ASM (Abstract State Machines) a été éffectuée pour modéliser un data warehouse réparti.

## 5- SOLUTION PROPOSEE

### 5.1- Concepts de base

Nous visons par cette démarche le côté logique et organisationnel des données qui n'a pas été bien mis en évidence dans l'état de l'art. L'objectif visé est, donc, de modéliser les données contenues dans un DW central et en même temps leur répartition entre plusieurs bases de données distantes qui seront les futurs DM de l'entreprise.

Pour généraliser notre démarche, nous proposons un formalisme que nous pourrons adapter à n'importe quelle approche de conception. Généralement la modélisation d'un système d'information se base sur trois niveaux :

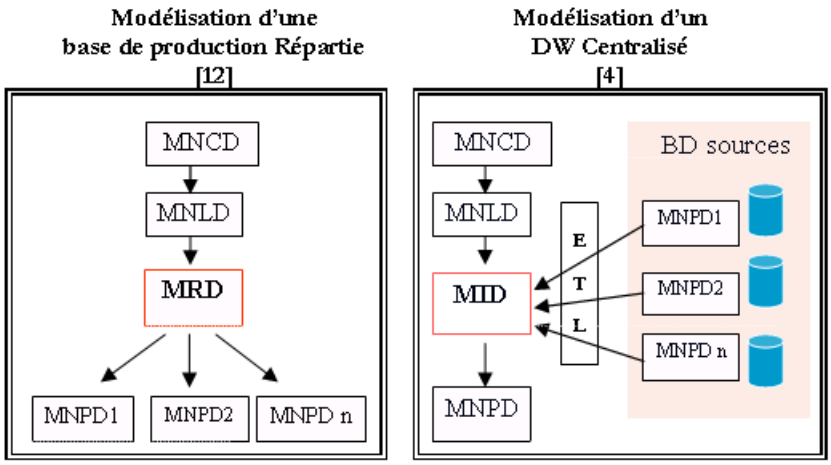


Figure 1 : Démarches de modélisation (Etat de l'art)

1- Modélisation du Niveau Conceptuel des Données (MNCD)
2- Modélisation du Niveau Logique des Données (MNLD)
3- Modélisation du Niveau Physique des Données (MNPD)

En effet, nous allons nous baser sur les deux démarches proposées dans [4] et [12]. Le but est l'adaptation des processus de répartition à la modélisation d'un DW.

Une fois la modélisation du niveau conceptuel des données (MNCD) est réalisée, on entame la modélisation du niveau logique des données, celle-ci peut s'effectuer en deux axes. Deux besoins fondamentaux sont à satisfaire : (1) Il faut tenir compte tout d'abord des besoins d'intégration des données modélisées dans le DW vis-à-vis de leurs sources de données. Elles peuvent subir des transformations pour leur adaptation à la base. (2) Un autre besoin est à satisfaire, c'est celui de la répartition des données entre les différents sites, ceci en tenant compte des débits binaires échangés, des fiabilités des échanges et des besoins de fragmentation des tables. Ces deux axes de modélisation sont indépendants et peuvent être effectués en parallèle. Ainsi, deux équipes de modélisation peuvent travailler en même temps. Une première équipe qui se charge de la répartition des données entre les différents DM et une deuxième équipe qui se charge de l'intégration logique des données sources du DW global. Cette méthode de modélisation permet de garantir un espace de travail partagé, accélérer le rythme du travail et réduire la complexité de la modélisation. Les deux équipes peuvent se réunir ensuite, pour une organisation finale des données intégrées et allouées.

### 5.2- Modèles

Cette méthodologie se base sur six modèles (Figure 2) répartis selon trois niveaux : niveau conceptuel, niveau logique (enrichi) et niveau physique.

Pour choisir l'architecture à mettre en place, nous proposons un modèle introductif appelé Modèle Structurel et Organisationnel de l'Entreprise (MSOE).

Au niveau conceptuel nous gardons le Modèle Conceptuel de Données (MCD) proposé dans les approches classiques.

Au niveau logique, Le modèle logique de Données (MLD) sera généré à partir du MCD.

A ce niveau, nous proposons un enrichissement à travers deux modèles : Un Modèle d'Intégration Logique des Données (MILD) et un Modèle de Répartition Logique des Données (MRLD).

Le MILD permet d'identifier pour chaque donnée du MNLD, la source correspondante et (si nécessaire) les transformations qu'elle doit subir.

Le MRLD permet d'identifier pour chaque donnée du MNLD global le DM auquel elle sera affectée.

Le MILD et le MRLD seront fusionnés pour créer un dernier modèle englobant toutes les informations nécessaires pour la modélisation de la répartition des données d'un DW. Celui-ci est appelé Modèle d'Intégration Logique des Données Réparties (MILDR).

Au niveau physique, plusieurs Modèles Physiques de Données (MPD) seront déduits. Ces derniers, représentent l'organisation physique des différents DM.

### 5.3- FORMALISME PROPOSE

#### 5.3.1- Le MSOE

Pour établir le MSOE, nous proposons le formalisme suivant :

– Un Tableau de Structure (TS)

– Une matrice de liaison inter-site (MLIS)

La MLIS permet de décrire la structure de l'entreprise et son organisation de point de vue géographique. Cette description est nécessaire pour choisir la meilleure architecture à mettre en place.

Pour élaborer le MSOE, nous allons commencer tout d'abord par analyser la structure de l'entreprise. Ceci revient à trouver des réponses aux questions suivantes :

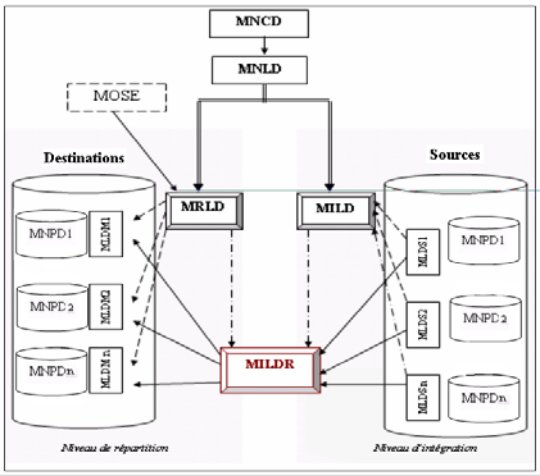


Figure 2 : Modèles proposés

1- L'entreprise est elle répartie géographiquement ?
2- Si oui, quels sont les sites qui la forment ?
3- Comment ces sites sont reliés les uns aux autres ?
4- Quels sont les moyens et les caractéristiques des supports de communication entre les sites (type de réseau, support, protocole, débit binaire de transmission des données, fiabilité…etc.)?

Les réponses à ces questions sont résumées dans le TS (Figure 3). Le tableau est une simple description de la structure de l'entreprise. Il permet de visualiser la liste des groupes de sites de l'entreprise, triée par ordre de priorité décisionnelle. Pour chaque site, on identifie son type et le groupe auquel il appartient. Le tableau de structure est

important pour les étapes suivantes puisqu'il détermine la liste des sites décisionnels de l'entreprise et peut faire l'objet d'une documentation pour l'entreprise. Les sites de priorité 3 seront supprimés de la liste puisqu'ils ne détiennent aucun pouvoir décisionnel.

| Priorité (1, 2,3) | Groupes de sites | Types de sites | Liste des sites |
|---|---|---|---|
| 1 | G1 | T1 | $S_{1.1},...,S_{1.n}$ |
| 2 | G2 | T2 | $S_{2.1},...,S_{2.k}$ |
| 3 | G3 | T3 | $S_{3.1},...,S_{3.j}$ |

Figure 3 : Tableau De Structure

Après avoir décrit la structuration de l'entreprise, il est important d'étudier son infrastructure réseau. Cette étude consiste à identifier le type du réseau, les moyens et les caractéristiques des supports de communication entre les sites. En effet, il s'agit d'identifier pour chaque couple de sites s'il existe une portion du réseau qui les relie. Si cette portion existe, il faut se renseigner sur les caractéristiques des communications entre ces deux sites. Les caractéristiques qui nous semblent les plus importantes à identifier sont la fiabilité et le débit binaire.

Ces deux caractéristiques diffèrent d'une portion de réseau à une autre. En effet, ceci dépend des supports de transmission au sein du réseau, de la distance qui sépare les deux sites, ...etc. Des statistiques sont utilisées afin de déterminer les valeurs de ces caractéristiques.

Pour la présentation de ces caractéristiques nous pouvons utiliser une matrice carrée d'ordre n, où n est le nombre de sites. Cette matrice résume les liaisons entre les différents sites de l'entreprise. Chaque cellule de cette matrice contient le débit binaire échangé et la fiabilité de la portion du réseau liant les deux sites correspondants. Elle représente donc, l'existence d'une liaison entre deux sites quelconques. Une cellule vide indique l'absence de liaison entre les deux sites. Nous pouvons résumer ces différentes données dans la matrice qu'on a choisi d'appeler Matrice des Liaisons Inter-Sites1 (MLIS1) (Figure 4).

| | S1 | S2 | … | Sn |
|---|---|---|---|---|
| S1 | | DB | DB | DB |
| | | F | F | F |
| S2 | DB | | DB | DB |
| | F | | F | F |
| … | DB | DB | | DB |
| | F | F | | F |
| Sn | DB | DB | DB | |
| | F | F | F | |

Figure 4 : La MLIS1

Cette matrice peut être améliorée (Figure 5) par l'ajout des indicateurs caractérisant les débits binaires.

Un débit binaire (DB) peut être : un débit élevé (DE), un débit moyen (DMoy), un débit faible (DF).

Cette classification est basée sur la définition de trois intervalles de débit binaire. Ensuite, suivant l'appartenance du DB à un intervalle parmi ces trois, un

débit peut être classé : (DE), (DMoy) ou bien (DF).

Les modèles logiques des Data Marts décrivent l'allocation logique des différentes données du DW vers les sites correspondants. Cette description ne prend pas en considération les besoins d'intégration des données vis-à-vis de leurs sources. Les informations données par le MSOE sont insuffisantes pour décider l'allocation d'une information vers un site donné. Plusieurs critères sont à prendre en considération pour la répartition des données. Le critère le plus important est celui de la fréquence d'utilisation. On rappelle à cet effet, que les données du DW sont utilisées seulement en consultation.

De ce fait, il faut tout d'abord ressortir les différents traitements possibles qui seront exécutés par les sites de l'entreprise. Pour chaque site, on va énoncer les utilisations possibles des différentes données par les traitements. Les données sont des tables ou bien fragments de tables.

| | S1 | S2 | … | Sn |
|---|---|---|---|---|
| S 1 | | DE/DMoy/DF | DE/DMoy/DF | DE/DMoy/DF |
| | | F | F | F |
| S 2 | DE/DMoy/DF | | DE/DM/DF | DE/DM/DF |
| | F | | F | F |
| … | DE/DMoy/DF | DE/DM/DF | | DE/DM/DF |
| | F | F | | F |
| S n | DE/DMoy/DF | DE/DMoy/DF | DE/DMoy/DF | |
| | F | F | F | |

Figure 5 : MLIS2

### 5.3.2- Le Modèle de Répartition Logique des Données

Pour modéliser la répartition des données d'un DW, la première tâche consiste à identifier pour chaque site mentionné dans le MSOE, les traitements possibles sur les données de la base. Ceci, est visualisé dans une matrice tridimensionnelle car elle englobe les sites de l'entreprise, les traitements à effectuer par site et les données nécessaires. Cette matrice est appelée : « Matrice Utilisation des Données (MUD) ».

La deuxième tâche consiste à déduire à partir de la matrice précédente, les meilleures allocations possibles des données aux sites de l'entreprise. Et ce, par la construction d'une deuxième matrice visualisant le mode d'allocation des données. Cette matrice est appelée : « Matrice d'Allocation des Données (MAD) ».

#### LA MUD

La matrice utilisation des données décrit les différentes utilisations possibles des données par les différents sites décisionnels. Pour établir la MUD nous allons identifier tout d'abord les traitements possibles par site. Il est à noter qu'une donnée $D_u$ peut être soit une table ou bien une portion de table, c'est à dire un fragment de table. Les tables sont extraites directement du MLD, par contre, les fragments de tables ne sont pas facilement identifiables. Pour ce faire, nous allons adapter le formalisme de fragmentation des données de production aux tables du MNLD. Le point de départ est la liste des tables du MLD et les différents traitements ti classés par site. Le résultat est l'ensemble des fragments nécessaires pour les différentes utilisations. Pour ce faire, nous proposons un formalisme qu'on a choisi d'appeler Matrice de fragmentation (MF).

Cette matrice a pour objectif l'identification des critères de fragmentation et les fragments nécessaires.

Elle est tridimensionnelle :

- une dimension pour les tables d'origine (T),
- une dimension pour les sites ($S_i$),
- une troisième pour les traitements susceptibles d'être exécutés sur chaque site ($T_i$).

Il existe deux types de fragments, un fragment vertical et un fragment horizontal : Le fragment vertical est une sélection d'une colonne d'une table. Le fragment horizontal est une sélection d'une ligne d'une table.

Pour établir la MF, il est indispensable d'identifier les différents fragments horizontaux et verticaux. De ce fait, nous proposons deux matrices préliminaires que nous avons choisis d'appeler Matrice de Fragmentation Horizontale (MFH) et la Matrice de Fragmentation Verticale et Mixte (MFVM).

La matrice d'utilisation des données (Figure 6) décrit pour chaque traitement correspondant à un site de l'entreprise, les données nécessaires pour son accomplissement. C'est une matrice tridimensionnelle par ce qu'elle intègre les dimensions suivantes:

- La liste des sites ($S_i$)
- La liste des traitements par site ($T_i$)
- La liste des données nécessaires aux traitements ($D_u$)

Dans la matrice utilisation des données, nous désignons par $D_u$:

- une table non fragmentée,
- un fragment horizontal,
- un fragment vertical,

– ou bien, un fragment mixte.

La construction de cette matrice consiste à identifier, pour chaque traitement $t_ip_i$, les données nécessaires, leur mode et leurs fréquences de consultation par ce traitement. Les opérations de création, de suppression, de modification ne seront pas prises en compte. Ces opérations seront faites par l'administrateur du DW qui s'occupe lui même de toutes les opérations de mise à jour. Chaque case de la matrice indique la fréquence d'utilisation de $D_u$ vis à vis de tipi appartenant à $S_i$.

La matrice d'utilisation des données peut être simplifiée (Figure 7) en indiquant le total des utilisations par site. Ainsi, nous pouvons pour chaque donnée identifier le site le plus prioritaire, en tenant compte du nombre d'utilisations de celle-ci. La simplification aboutit à une deuxième matrice qu'on a choisi d'appeler MUD2.

Cette matrice nous servira de support pour décider l'allocation des données par site. Nous avons choisi d'enrichir cette matrice par les indicateurs de priorité pour chaque site. Chaque cellule contiendra le total des fréquences d'utilisation par site divisé par la priorité de ce dernier.

Le résultat final est un indicateur efficace pour décrire la nécessité ou non d'allocation de la donnée au site correspondant.

A ce niveau, nous disposons d'une liste de tables, de FH, de FV et de FM, nous avons aussi la fréquence d'utilisation de ces données par site ainsi que leurs priorités. Nous pouvons alors, procéder à la construction de la MAD.

**LA MAD**

La matrice d'allocation des données (Figure 8) décrit pour chaque donnée, le site dans lequel elle sera allouée ou bien, elle sera consultée. C'est une matrice bidimensionnelle parce qu'elle englobe : Les données utilisées ($D_u$), Les sites destinataires ($S_i$).

| | | | **TABLES DE FAITS / FRAGMENTS HORIZONTAUX** | | | | |
|---|---|---|---|---|---|---|---|
| | | | **$D_1$** | **…** | **$D_u$** | **…** | **$D_t$** |
| **SITES CIBLES** | **S1** | $t_{1.1}$ | FU | | FU | | FU |
| | | … | | | | | |
| | | $t_{1.p1}$ | FU | | FU | | FU |
| | | … | | | | | |
| | | $t_{1.q1}$ | FU | | FU | | FU |
| | Total des utilisations | | $TU_{1.1}$ | | $TU_{u.1}$ | | $TU_{t.1}$ |
| | **...** | | | | | | |
| | **Si** | $t_{1.i}$ | FU | | FU | | FU |
| | | … | | | | | |
| | | $t_{i.pi}$ | FU | | FU | | FU |
| | | … | | | | | |
| | | $t_{i.qi}$ | FU | | FU | | FU |
| | **...** | | | | | | |
| | Total des utilisations | | $TU_{1.i}$ | | $TU_{u.i}$ | | $TU_{t.i}$ |
| | **…** | | | | | | |
| | **Sn** | $t_{n.1}$ | | | | | x |
| | | … | | | | | |
| | | $t_{n.pn}$ | x | | x | | |
| | | | | | | | |
| | | $t_{n.qn}$ | | | x | | |
| | Total des utilisations | | $TU_{1.n}$ | | $TU_{u.n}$ | | $TU_{t.n}$ |

Figure 6 : MUD1

Chaque cellule de la MUD2 représente un indicateur du besoin d'allocation. Si cet indicateur est faible, ceci signifie la non nécessité d'allocation de la donnée correspondante.

| | | **TABLES DE FAITS / FRAGMENTS HORIZONTAUX** | | | | |
|---|---|---|---|---|---|---|
| | | **$D_1$** | **…** | **$D_u$** | **…** | **$D_t$** |
| **SITES CIBLES** | **$S_1$** | $TU_{11}$ | | $TU_{u1i}$ | | $TU_{t1i}$ |
| | | $P_1$ | | $P_1$ | | $P_1$ |
| | **...** | | | | | |
| | **$S_i$** | $TU_{1.i}$ | | $TU_{u.i}$ | | $TU_{t.i}$ |
| | | $P_i$ | | $P_i$ | | $P_i$ |
| | **...** | | | | | |
| | **$S_n$** | $TU_{1.n}$ | | $TU_{u.n}$ | | $TU_{t.n}$ |
| | | $P_n$ | | $P_n$ | | $P_n$ |

Figure 7 : MUD2

Pour allouer une donnée on cherche le site qui l'utilise le plus c'est à dire dont le total des fréquences d'utilisation est supérieur à tous les autres sites et qui est le plus prioritaire pour son utilisation. Cette méthode d'allocation peut faire l'objet d'automatisation. Ainsi, on pourra décrire l'algorithme correspondant à l'allocation des données du DW vers les différents Data Marts de l'organisation.

Chaque cellule de la matrice d'allocation des données indique si une donnée est une Donnée Persistante (DP) ou bien Donnée Consultée (DC). Une DP signifie qu'elle est allouée au site correspondant, par contre, une DC veut dire qu'elle sera juste consultée par le site correspondant. Une cellule vide indique que la donnée n'est pas consultée par le site et on l'appelle Donnée Absente (DA).

Ayant réalisé le MALD, la modélisation de la répartition logique des Data Mart est achevée. Toutes les données correspondantes à la répartition des données sont identifiées, ces données nous permettront en partie de construire le MILDR. Mais, il faudra tenir compte en parallèle de l'intégration des données sources vis à vis des sources de données.

| | | TABLES DE FAITS / FRAGMENTX HORIZONTAUX | | | | |
|---|---|---|---|---|---|---|
| | | $D_1$ | … | $D_u$ | … | $D_t$ |
| DESTINATIONS | DM1 | DP/DC/DA | | DP/DC/DA | | DP/DC/DA |
| | ... | | | | | |
| | DMi | DP/DC/DA | | DP/DC/DA | | DP/DC/DA |
| | ... | | | | | |
| | DMn | DP/DC/DA | | DP/DC/DA | | DP/DC/DA |

Figure 8 : MALD1

**5.3.3- Modélisation de l'Intégration Logique des Données Sources**

Le modèle d'intégration logique des données sources décrit les sources de données nécessaires pour les besoins d'intégration (Figure 9).

Chaque donnée du MLD est caractérisée par une source correspondante. Elle subit des transformations selon les besoins.

La modélisation de l'intégration logique des données sources n'intègre en aucun cas les besoins de répartition physique. Le formalisme proposé est une Matrice d'Intégration Logique des Données Sources (MILDS) (Figure 9).

Il s'agit de déterminer pour chaque donnée du MLD la source de donnée qui permet de l'alimenter. Cette dernière subit les transformations nécessaires pour l'adapter à la base.

Une source peut être soit :

- *Interne* : c'est l'ensemble des attributs qui se trouvent dans les tables sources des applications fonctionnelles.
- *Externe* : c'est l'ensemble des attributs spécifiques au DW comme les dates, les catégories, les types,…etc., qui ne proviennent pas des données sources.

Les données peuvent subir plusieurs. Une transformation peut faire l'objet :

- D'une transformation élémentaire (TE): formule, expression ou des programmes permettant d'obtenir le contenu d'un attribut (a) à partir d'une source (Sc). Ce type de transformation est fait dans le cas ou l'attribut est obtenu à partir d'une seule source.
- Une transformation composite (TC) : formule, expression ou programme permettant d'obtenir le contenu d'un attribut à partir de deux ou plusieurs sources.

Au niveau de la phase d'intégration, tout attribut de la base doit être caractérisé par la source qui l'alimente et par les transformations nécessaires qu'il doit subir pour son utilisation par les différents sites correspondants. Pour identifier ces transformations nous proposons une matrice que nous avons choisi d'appeler Matrice de Transformations des Données Sources (MTDS).

Celle-ci est tridimensionnelle car elle renferme les dimensions suivantes :

- une dimension pour les attributs des différentes tables du MLD (A)
- une dimension pour les sources internes (SI)
- une dimension pour les sources externes (SE)

Une fois la modélisation terminée, l'équipe de modélisation de la répartition fournit la MAD. La deuxième équipe, s'occupant de la modélisation logique des données sources, fournit la MILDS. Ces deux équipes, peuvent ensuite se réunir pour préparer la modélisation de l'intégration logique des données réparties.

#### 5.3.4- Modélisation de l'Intégration Logique des Données Réparties (MILDR)

La modélisation de l'intégration logique des données réparties (Figure 10) consiste à caractériser chaque donnée allouée à un DM par :

- la source correspondante pour son alimentation (interne ou externe)
- la transformation nécessaire que la donnée source peut subir (élémentaire ou composite) pour son adaptation à la base.
- le DM auquel elle sera allouée.

| **Tables du MNLD** | **Attributs** | **Sources de données** | | | | | | | | | | | | | | | | **TC** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **MLDS1** | | | | | **MLDS r** | | | | | | **MLDS s** | | | | | |
| | | **$A_{11}$** | | **$A_{1z}$** | | **$A_{1x}$** | | **$A_{r1}$** | | **$A_{rz}$** | | **$A_{rx}$** | | **$A_{s1}$** | | **$A_{sz}$** | | **$A_{sx}$** | |
| **TD1** | **$a_{11}$** | TE | | | | | | TE | | | | | | TE | | | | | |
| | **…** | | | | | | | | | | | | | | | | | | |
| | **$a_{l1}$** | x | | | | x | | x | | | | | | x | | | | | **TC** |
| | **...** | | | | | | | | | | | | | | | | | | |
| | **$a_{1m}$** | | | | | | | | | | | | | | | | | | |
| **TDj** | **$a_{j1}$** | | | TE | | | | | | TE | | | | | | TE | | | |
| | **…** | | | | | | | | | | | | | | | | | | |
| | **$a_{j1}$** | | | | | | | | | | | | | | | | | | |
| | **…** | | | | | | | | | | | | | | | | | | |
| | **$a_{jm}$** | | | | | | | | | | | | | | | | | | |
| **TDk** | **$a_{k1}$** | x | | x | | | | x | | x | | | | x | | x | | | **TC** |
| | **…** | | | | | | | | | | | | | | | | | | |
| | **$a_{kl}$** | | | | | TE | | | | | | TE | | | | | | TE | |
| | **…** | | | x | | x | | | | x | | x | | | | x | | x | **TC** |
| | **$a_{km}$** | | | | | | | | | | | | | | | | | | |

Figure 9: MILDS

Le formalisme proposé est une matrice appelée Matrice d'Intégration Logique des Données Réparties (MILDR). La modélisation de l'intégration logique des données réparties consiste à fusionner les deux matrices réalisées au niveau logique (la MILDS et la MRLD). La fusion consiste à remplacer les colonnes de la MILDS par les données réparties entre les différents Data Mart de l'entreprise, ces données sont celles identifiées dans la matrice allocation logique des données. On gardera par contre les mêmes colonnes de la matrice transformation des données. Le résultat de la fusion sera une nouvelle matrice englobant toutes les informations nécessaires pour la répartition des données d'un DW entre plusieurs DM.

La MILDR représente le dernier niveau de modélisation logique, elle garde pour chaque donnée $D_u$ sa traçabilité vis à vis de sa source, les différentes transformations nécessaires pour son adaptation à la base, le type de la transformation voulu et le site correspondant auquel elle sera allouée.

Cette matrice peut subir des modifications selon le besoin, ceci va simplifier les mises à jour et renforcer la flexibilité de la modélisation surtout avec la fluctuation de l'environnement et avec l'extension du besoin informationnel vis à vis du DW.

| Destinations \ Sources | | | $MLDS_1$ | | | $MLDS_r$ | | | $MLDS_s$ | | | TC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $A_{11}$ | $A_{1z}$ | $A_{1x}$ | $A_{r1}$ | $A_{rz}$ | $A_{rx}$ | $A_{s1}$ | $A_{sz}$ | $A_{sx}$ | |
| DM | $MLDM_1$ | $D_1$ | TE | | | TE | | | TE | | | |
| | | … | | | | | | | | | | |
| | | $D_u$ | x | | x | x | | | x | | | TC |
| | | ... | | | | | | | | | | |
| | | $D_t$ | | | | | | | | | | |
| … | | | | | | | | | | | | |
| DM | $MLDM_j$ | D1 | | TE | | | TE | | | TE | | |
| | | … | | | | | | | | | | |
| | | $D_u$ | | | | | | | | | | |
| | | ... | | | | | | | | | | |
| | | $D_t$ | | | | | | | | | | |
| … | | | | | | | | | | | | |
| DM | $MLDM_n$ | $D_1$ | x | x | | x | x | | x | x | | TC |
| | | … | | | | | | | | | | |
| | | $D_u$ | | | TE | | | TE | | | TE | |
| | | ... | | x | x | | x | x | | x | x | TC |
| | | $D_t$ | | | | | | | | | | |

Figure 10 : Matrice de l'Intégration Logique des Données Réparties

## 6- CONCLUSION

Dans cet article nous avons proposé une nouvelle démarche de modélisation de la répartition des données des DW. L'avantage de la démarche proposée est le fait qu'elle constitue une extension au niveau de la modélisation qui peut s'appliquer sur n'importe quelle approche de conception.

L'apport de la démarche proposée est la mise en évidence du coté organisationnel des données d'un DW. Un enrichissement du niveau logique de modélisation est effectué pour garantir une meilleure organisation des données. A ce niveau, un MNLD global du DW est établi. Ensuite, la modélisation logique s'oriente vers deux axes indépendants et qui peuvent se faire en parallèle. Un premier axe consiste à modéliser l'allocation logique des données du DW vers plusieurs DM, le résultat est une MALD. Le deuxième axe consiste à modéliser l'intégration des données vis à vis des sources de données en tenant compte des transformations nécessaires sur les données sources, le résultat est la MILDS. Le résultat donné par le niveau logique est une matrice appelée MILDR du DW et qui n'est autre que la fusion des deux matrices précédentes. Cette dernière permet de visualiser toutes les informations nécessaires pour la répartition des

données d'un DW. Cependant, quelques axes de recherches restent à étudier et à approfondir :

- Jusqu'àprésent il n'y a pas eu de réalisation pour la solution proposée, le travail effectué sera complété ultérieurement par une implémentation.
- Nous envisageons aussi une amélioration du processus d'allocation des données par la prise en compte des caractéristiques du réseau. On pourra intégrer un programme d'optimisation permettant de donner une meilleure allocation possible en tenant compte des caractéristiques du réseau et de sa fiabilité.
- On pourra ensuite, envisager une l'intégration d'une allocation dynamique des données et ceci par l'intégration d'un agent intelligent qui permet de calculer les fréquences d'utilisation des différentes données par les sites de l'organisation.

**BIBLIOGRAPHIE**


[1] Ralph Kimball, «The Data Warehouse has no centre», Volume 2, Nombre 10. (1999).

[2] Bill Inmon, «Data Mart does not equal Data Warehouse», DM Direct. (1999).

[3] Jean-François Goglin; « La construction du data warehouse, du data mart au data web »; Nouvelles Technologies Informatiques; Ed. HERMES.

[4] KOLSI Nader, « Modélisation de l'intégration des données d'un DW ». Institut Supérieur de Gestion, TunisIII, Tunisie (2000).

[5] Ladjel Bellatreche, Kamalakar Karlapalem, «Some Issues in design of Data Warehousing Systems», Department of computer Science & Technology Clear Water Bay Kowloon, Hong Kong.(1999).

[6] GALACSI, « Conception De Bases De Données : Du schéma conceptuel aux schémas physiques » DUNOD informatique.(1989).

[7] Noaman, A.Y. et K. Barker, "Distributed Data warehouse Design", (under revision for) journal Submission. (2000).

[8] Noaman, A.Y. et K. Barker, "Distributed Data warehouse Architecture and design", the Fourteenth International Symposium on computer and Information Sciences (ISCI'99), Kusadasi, Turki. (1999).

[9] Noaman, A.Y. et K. Barker, "A Horizontal Fragmentation Algorithm for the fact relation in a Distributed Data Warehouse", the Eight International Conference on Information and Knowledge Management (CIKM'99), Kansas, Missouri.(1999).

[10] MESSAOUD Saloua, « Modélisation de la répartition et de la réplication des données». Institut Supérieur de Gestion, TunisIII, Tunisie (2000).

[11] Stefano CERI, Giuseppe PELAGATTI, «Distributed Data base: Principles and systems », McGaw-hill. (1984).

[12] George Gardarin, «Bases de données : Objets et relationnelles », Edition EYRLLES (1997).

[13] P.O'Neil and D.Quass. "Improved query performance with variant indexes. Proceedings of the ACM SIGMOD International Conference on Management of Data. (1997).

[14] P.O'Neil et D.Quass. Improved query performance with variant indexes. Proceedings of the ACM SIGMOD International Conference on Management of Data. (1997).

[15] S. Chaudhuri and V.Narasayya. "Index merging". Proceedings of the International Conference on Data Engineering (ICDE). (1999).

[16] Jane Zaho, Klaus-Dieter Schewe ACM International Conference Proceeding Series, Proceedings of the first Asian-Pacific conference on Conceptual modelling, Dunedin, New Zealand. (2004).