Modélisation des bases de données multidimensionnelles à agrégations multiples et différentiées

Ali Hassan*, Franck Ravat*, Olivier Teste**, Ronan Tournier*, Gilles Zurfluh*

*Université Toulouse 1 Capitole - IRIT (UMR 5505)
**Université Toulouse 3 Paul Sabatier - IRIT (UMR 5505)
118, Route de Narbonne - 31062 Toulouse cedex 9 (France)
{hassan, ravat, teste, tournier, zurfluh}@irit.fr

Résumé. De nombreux modèles ont été proposés pour représenter les données multidimensionnelles. Ces propositions considèrent généralement une même fonction d'agrégation pour déterminer les valeurs d'une mesure aux différents niveaux de granularité de l'espace multidimensionnel. Nous proposons un nouveau modèle conceptuel plus flexible supportant des agrégations multiples différentiées. L'agrégation multiple permet d'associer à une même mesure, une fonction d'agrégation différente pour chaque dimension. L'agrégation différentiée autorise des agrégations spécifiques à chaque paramètre. Notre modèle repose sur un double formalisme graphique suffisamment expressif pour contrôler la validité des fonctions d'agrégation. Nous étudions également les conséquences de cette modélisation conceptuelle pour la construction efficace des treillis de pré-agrégats dans le contexte R-OLAP.

1 Introduction

Les analyses OLAP consistent à suivre des indicateurs considérés comme des points observés dans un espace défini par différents axes d'analyse. Les données analysées sont hiérarchisées selon différents niveaux de détails et des fonctions d'agrégation sont utilisées pour obtenir une vision synthétique. Les données sont regroupées selon le niveau de détail sélectionné et agrégées avec les fonctions. Les opérations de forage (*roll-up et drill-down*), souvent employées lors des analyses OLAP, font une utilisation intensive de ces fonctions d'agrégation.

Les bases de données multidimensionnelles (BDM) offrent classiquement un cadre adéquat aux analyses décisionnelles, cependant les fonctions d'agrégation fournies par défaut peuvent s'avérer imparfaites. Par exemple, l'analyse de températures d'un territoire peut être réalisée de manière uniforme avec une seule et même fonction d'agrégation : la température maximale annuelle est obtenue à partir des températures maximales mensuelles, elles-mêmes obtenues à partir des températures maximales quotidiennes. Par contre, l'analyse des températures par département repose généralement sur la moyenne des températures par ville, tandis que l'analyse des températures par région peut être effectuée à partir d'une agrégation des températures des départements qui tient compte de la superficie du département (moyenne pondérée). Cette analyse fait donc intervenir différentes fonctions d'agrégation (moyenne ou moyenne pondérée) selon le niveau d'analyse considéré. D'autre part, dans l'analyse classique, l'agrégation à

un niveau de granularité peut être généralement obtenue à partir de l'agrégation de n'importe quel niveau inférieur (la température maximale annuelle est la température maximale soit mensuelle, soit quotidienne). Dans le cas des températures en fonction de la répartition territoriale, le calcul de la température par région ne peut être obtenu directement à partir des températures des villes ; pour obtenir la température d'une région, il est nécessaire de calculer la température moyenne par département pour ensuite faire la moyenne pondérée de ces températures.

Cet article présente un nouveau modèle multidimensionnel suffisamment expressif pour supporter ces différents cas d'agrégation. Nous exploitons ce modèle conceptuel pour étudier les conséquences au niveau logique sur les treillis de pré-agrégats (Gray et al., 1996).

1.1 Positionnement et contributions des travaux

Il existe classiquement deux approches pour la modélisation des BDM : une approche reposant sur la métaphore du cube de données suivant laquelle la BDM est représentée par des cubes, et une approche dite de modélisation multidimensionnelle où la BDM est décrite par un schéma en étoile ou en constellation (Kimball, 1996). Nos travaux s'inscrivent dans cette seconde approche. En effet, la métaphore du cube repose sur une séparation ambigüe des éléments de structures et des valeurs (Torlone, 2003) : modélisation des axes de l'analyse peu expressive notamment en raison de la difficulté à représenter l'organisation hiérarchique des données. Elle s'avère notamment limitée lorsqu'il s'agit de représenter des constellations de faits et de dimensions partagées.

Plusieurs synthèses du domaine (Chaudhuri et Dayal, 1997), (Vassiliadis et Sellis, 1999), (Mazón et al., 2009) et d'études comparatives (Gyssens et Lakshmanan, 1997), (Vassiliadis et Skiadopoulos, 2000), (Pedersen et al., 2001), (Abelló et al., 2006), (Lujàn-Mora et al., 2006), (Ravat et al., 2008), (Prat et Akoka, 2010), (Oliveira et al., 2011), (Boulil et al., 2011) sont disponibles dans la littérature scientifique. La plupart des propositions existantes considèrent qu'une mesure est associée à une fonction d'agrégation qui sera utilisée à tous les niveaux d'agrégation modélisés. Cette fonction calcule la même agrégation pour toutes les combinaisons de tous les paramètres modélisés.

(Gyssens et Lakshmanan, 1997) et (Vassiliadis et Skiadopoulos, 2000) ne précisent pas des fonctions d'agrégation pour les mesures, mais ils laissent la possibilité d'utiliser pour chaque mesure plusieurs fonctions d'agrégation au cours du processus OLAP. Dans les travaux de (Pedersen et al., 2001), on peut lier à une seule mesure plusieurs fonctions d'agrégation. Néanmoins dans tous les travaux précédents, chaque fonction est utilisée uniformément pour toutes les dimensions et tous les niveaux des hiérarchies. Le modèle YAM^2 (Abelló et al., 2006) permet d'utiliser une fonction d'agrégation différente pour chaque dimension. Ce modèle ne donne cependant pas la possibilité de faire évoluer la fonction avec les niveaux de hiérarchies. Le modèle d'agrégation de (Prat et Akoka, 2010) nous permet d'associer à chaque mesure une fonction d'agrégation pour chaque dimension ou chaque hiérarchie ou sous-hiérarchie. Le modèle proposé ne traite que le cas où il y a des fonctions standards (SUM, AVG, MIN, MAX, COUNT). (Boulil et al., 2011) ont levé cette limite. Ils utilisent un modèle d'agrégation diffèrent de celui de (Prat et Akoka, 2010) du fait qu'il utilise une fonction d'agrégation entre deux niveaux d'agrégation au lieu d'une fonction d'agrégation associée à des sous-hiérarchies. Ces travaux souffrent d'une limite qui est de ne pas traiter le cas où les fonctions d'agrégation sont non commutatives (par exemple, AVG et moyenne pondérée).

En ce qui concerne les outils commerciaux, «Business Objects» utilise une seule fonction d'agrégation pour une mesure. En revanche, l'outil «Analysis Services de Microsoft» offre la possibilité d'appliquer un «rollup personnalisé» à une hiérarchie de plusieurs façons (Harinath et al., 2009) :

– par l'utilisation des opérateurs unaires qui sont utilisés pour résoudre le problème de l'agrégation sur un type particulier de hiérarchie (hiérarchie d'attributs parent-enfant). Une hiérarchie parent-enfant est construite à partir d'un seul attribut parent. Un attribut parent décrit une relation de jointure réflexive dans une table de dimension principale.
– par l'utilisation de scripts MDX, soit directement, soit par l'utilisation de la propriété «CustomRollupColumn» qui indique à une colonne où sont stockés les scripts MDX.

Les deux approches représentent des fonctions d'agrégation mais elles ne sont liées ni à une dimension ni à un niveau d'agrégation. Elles sont liées à un membre (une instance) d'un niveau d'agrégation d'une hiérarchie, c'est-à-dire, à une ligne dans la table de la dimension. Donc, pour appliquer ce «rollup personnalisé» à un seul niveau d'agrégation il faut le répéter pour toutes les instances de ce niveau. Cela pose un problème de stockage et diminue la performance (Harinath et al., 2009). D'un autre coté, la liaison de «rollup personnalisé» avec une instance spécifique peut entraîner des difficultés en ce qui concerne la mise-à-jour des données.

Notre objectif est de lever ces limites en développant un modèle conceptuel de représentation des agrégations multidimensionnelles multiples différentiées. Par *multiples* nous signifions qu'une même mesure peut être agrégée selon plusieurs fonctions d'agrégation et par *différentiées* nous indiquons que ces agrégations peuvent varier en fonction du niveau d'agrégation.

Par ailleurs, les fonctions d'agrégation sont classifiées :

– du point de vue du mécanisme d'agrégation, les fonctions d'agrégations appartiennent à trois catégories différentes (Gray et al., 1996). La première correspond aux fonctions **distributives** qui calculent les valeurs agrégées à un niveau de granularité à partir des valeurs déjà agrégées au niveau de granularité directement inférieur (par exemple, la somme - SUM - d'un montant par année peut se calculer à partir de la somme des montants par semestre). La deuxième correspond aux fonctions **algébriques** qui calculent les valeurs agrégées à partir de résultats intermédiaires stockés (par exemple, la moyenne - AVG - d'un montant par année peut se calculer à partir de la somme - SUM - des montants et du nombre - COUNT - des occurrences). Enfin, la troisième correspond aux fonctions **holistiques** qui ne peuvent pas être calculées à partir de résultats intermédiaires. Dans ce cas, il faut calculer les valeurs agrégées à partir des valeurs de base correspondant au niveau de granularité le plus bas (par exemple, RANK).
– du point de vue de la «Summarizability», les fonctions d'agrégation sont classifiées selon deux groupes (Abelló et al., 2006). Le premier contient les fonctions «Transitives» qui garantissent la «Summarizability». Le deuxième contient les fonctions «Non-Transitives» qui impliquent que l'agrégation doit toujours se calculer à partir du niveau de base.
– du point de vue de la mesure (données), les fonctions d'agrégation sont de trois types (Pedersen et al., 2001). Le premier est applicable aux données additives. Le deuxième est applicable aux données qui peuvent être utilisées pour les calculs de moyenne. Le troisième est applicable aux données constantes, c'est-à-dire qu'elles ne peuvent être que dénombrées.

Toutes ces propositions et les classifications des fonctions d'agrégation existantes estiment

que l'on peut calculer l'agrégation d'une mesure à partir du niveau de base. Notre but est d'ajouter le moyen de traiter le cas contraire (quand on ne peut pas agréger la mesure à partir du niveau de base) en utilisant des *contraintes d'agrégation*.

Cet article est organisé comme suit. La section 2 présente le modèle conceptuel multidimensionnel classique (Ravat et al., 2008) avant de présenter nos extensions pour les agrégations multiples différentiées. Puis nous présentons le formalisme graphique de ces extensions. La section 3 décrit le modèle logique R-OLAP avec ses relations d'optimisation et les impacts de nos extensions sur ce modèle.

2 Modèle conceptuel de données

2.1 Concepts classiques

Soient $\mathcal{N}$= $\{n_1, n_2, ...\}$ un ensemble fini de noms non redondants, F = $\{F_1, ..., F_n\}$ est un ensemble fini de faits, $n \geq 1$, D = $\{D_1, ..., D_m\}$ est un ensemble fini de dimensions, $m \geq 2$.

Définition 1. Un *fait*, noté F_i,$\forall$ i$\in [1..n]$, est défini par (n^{Fi}, M^{Fi}).

- n^{Fi} $\in\mathcal{N}$est le nom identifiant le fait,
- M^{Fi} $= \{m_1, ..., m_{pi}\}$ est un ensemble de *mesures*.

On pose $M = \bigcup_{i=1}^{n} M^{Fi}$.

Définition 2. Une *dimension*, notée D_i,$\forall$ i$\in [1..m]$, est définie par (n^{Di}, A^{Di}, H^{Di}).

- n^{Di} $\in\mathcal{N}$est le nom identifiant la dimension,
- A^{Di} $= \{a_1^{Di}, ..., a_{ri}^{Di}\}$ est l'ensemble des *attributs de la dimension*,
- H^{Di} $= \{H_1^{Di}, ..., H_{si}^{Di}\}$ est un ensemble de *hiérarchies*.

Les hiérarchies organisent les attributs d'une dimension, appelés paramètres, de la graduation la plus fine jusqu'à la graduation la plus générale. Ainsi une hiérarchie définit les chemins de navigation valides sur un axe d'analyse.

On pose $A = \bigcup_{i=1}^{m} A^{Di}$ et $H = \bigcup_{i=1}^{m} H^{Di}$.

Définition 3. Une *hiérarchie*, notée H_j (notation abusive de H_j^{Di},$\forall$ i$\in$ $[1..m]$,$\forall$ j $\in$ $[1..s_i]$), est définie par (n^{Hj}, P^{Hj}, $\prec^{Hj}$,Weak^{Hj}).

- n^{Hj} $\in\mathcal{N}$est le nom identifiant la hiérarchie,
- P^{Hj} $= \{p_1^{Hj}, ..., p_{qj}^{Hj}\}$ est un ensemble d'attributs de la dimension appelés *paramètres*, P^{Hj} $\subseteq$ A^{Di},
- $\prec^{Hj}$= $\{(p_x^{Hj}, p_y^{Hj}) \mid p_x^{Hj} \in P^{Hj} \wedge p_y^{Hj} \in P^{Hj}\}$ est une relation binaire antisymétrique et transitive. Rappelons que l'antisymétrie signifie que $(p_{k1}^{Hj} \prec^{Hj} p_{k2}^{Hj}) \wedge (p_{k2}^{Hj} \prec^{Hj} p_{k1}^{Hj}) \Rightarrow p_{k1}^{Hj} = p_{k2}^{Hj}$ tandis que la transitivité signifie que $(p_{k1}^{Hj} \prec^{Hj} p_{k2}^{Hj}) \wedge (p_{k2}^{Hj} \prec^{Hj} p_{k3}^{Hj}) \Rightarrow p_{k1}^{Hj} \prec^{Hj} p_{k3}^{Hj}$.
- $Weak^{Hj} : P^{Hj} \rightarrow 2^{A^{Di} \setminus P^{Hj}}$ est une application qui associe à chaque paramètre un ensemble d'attributs de dimension, appelés *attributs faibles* (2^E représente toute combinaison de l'ensemble E).

On pose $P^{Di} = \bigcup_{j=1}^{s_i} P^{Hj}$ et $P = \bigcup_{i=1}^{m} P^{Di} = \bigcup_{i=1}^{m} \bigcup_{j=1}^{s_i} P^{Hj}$.

Lemme 1. Pour chaque dimension D_i, un *paramètre racine*, noté $Id^{Di} \in P^{Di}$, existe. Il est défini comme suit : $\forall j \in [1..s_i], \forall p_k^{Hj} \in P^{Di}, Id^{Di} \neq p_k^{Hj} \mid Id^{Di} \prec^{Hj} p_k^{Hj}$.

Lemme 2. Pour chaque dimension D_i, un *paramètre extrémité*, noté $All^{Di} \in P^{Di}$, existe. Il est défini comme suit : $\forall j \in [1..s_i], \forall p_k^{Hj} \in P^{Di}, All^{Di} \neq p_k^{Hj} \mid p_k^{Hj} \prec^{Hj} All^{Di}$.

On pose $W^{Di} = \bigcup_{\forall j \in [1..s_i], \forall k \in [1..q_j]} Weak^{Hj}(p_k^{Hj})$ et

$W = \bigcup_{i=1}^{m} W^{Di} = \bigcup_{i=1}^{m} \bigcup_{\forall j \in [1..s_i], \forall k \in [1..q_j]} Weak^{Hj}(p_k^{Hj})$.

Lemme 3. Pour chaque dimension D_i, ses attributs de dimension sont de manière exclusive soit des paramètres, soit des attributs faibles, $P^{Di} \cap W^{Di} = \emptyset$ and $P^{Di} \cup W^{Di} = A^{Di}$.

Exemple 1. Nous illustrons nos propos à l'aide d'un exemple d'analyse concernant la météo. Dans cet exemple, les analystes étudient les températures maximales, minimales et moyennes ainsi que la vitesse du vent.

Pour supporter ces analyses, nous établissons une BDM comportant deux faits définis de la manière suivante :

- $F_{Température}$ = ('*Température*', {Tem_Moy, Tem_Max, Tem_Min}) ;
- F_{Vent} = ('*Vent*', {Vitesse}).

Ces analyses sont effectuées en fonction d'informations géographiques, temporelles et météorologiques (direction du vent). Nous considérons que chaque pays se compose de plusieurs régions. Chaque région comprend des départements, qui regroupent des villes. Nous considérons également que chaque ville a un niveau administratif selon lequel elle peut être une capitale d'un pays, une préfecture d'un département ou une capitale régionale. A proximité des villes, il y a des stations météorologiques qui mesurent la température et la vitesse du vent plusieurs fois par jour.

Pour simplifier, nous présentons les définitions formelles de deux dimensions seulement relatives aux informations géographiques et aux informations horaires :

- D_{Temps} = ('*Temps*', $\{a_{Toute\_les\_3\_heures}, a_{quart\text{-}jour}, a_{demi\text{-}journée}, ALL^{Temps}\}$, $\{H_{HTemps}\}$) avec
  - H_{HTemps}= ('*HTemps*', $\{a_{Toute\_les\_3\_heures}, a_{quart\text{-}jour}, a_{demi\text{-}journée}, ALL^{Temps}\}$, $\{(a_{Toute\_les\_3\_heures}, a_{quart\text{-}jour}), (a_{quart\text{-}jour}, a_{demi\text{-}journée}), (a_{demi\text{-}journée}, ALL^{Temps})\}$).
- $D_{Géographie}$ = ('*Géographie*', $\{a_{Ville}, a_{Niveau\_Administratif}, a_{Département}, a_{D\_Superficie}, a_{Région}, a_{R\_Superficie}, a_{Pays}, a_{P\_Superficie}, ALL^{Géographie}\}$, $\{H_{Hgéo\_scien}, H_{Hgéo\_simp}\}$) avec
  - $H_{Hgéo\_scien}$= ('*Hgéo_scien*', $\{a_{Ville}, a_{Département}, a_{Région}, a_{Pays}, ALL^{Géographie}\}$, $\{(a_{Ville}, a_{Département}), (a_{Département}, a_{Région}), (a_{Région}, a_{Pays}), (a_{Pays}, ALL^{Géographie})\}$, $\{(a_{Ville}, \{a_{Niveau\_Administratif}\}), (a_{Département}, \{a_{D\_Superficie}\}), (a_{Région}, \{a_{R\_Superficie}\}), (a_{Pays}, \{a_{P\_Superficie}\})\}$) et
  - $H_{Hgéo\_simp}$= ('*Hgéo_simp*', $\{a_{Ville}, a_{Département}, a_{Région}, a_{Pays}, ALL^{Géographie}\}$, $\{(a_{Ville}, a_{Département}), (a_{Département}, a_{Région}), (a_{Région}, a_{Pays}), (a_{Pays}, ALL^{Géographie})\}$, $\{(a_{Ville}, \{a_{Niveau\_Administratif}\})\}$).

La dimension géographique, notée $D_{Géographie}$, comporte deux hiérarchies. Ces deux hiérarchies permettent de tenir compte de deux façons d'observer les températures : simple ou scientifique. L'agrégation **simple** adopte la même méthode utilisée par les présentateurs de la météo. Quand on parle des températures dans un pays (même la maximale ou la minimale), on fait référence à la température de la capitale. De la même manière, lorsque nous parlons des températures dans une région ou d'un département, il s'agit des températures d'une ville considérée significative (capitale régionale ou préfecture). L'agrégation **scientifique** prend en compte les températures de toutes les régions pour calculer celle d'un pays. De même, lorsqu'on calcule les températures dans une région ou dans un département, on prend en compte respectivement tous ses départements ou toutes ses villes.

2.2 Extensions pour les agrégations multiples différentiées

Afin que le modèle multidimensionnel réponde à notre problématique, nous étendons les définitions précédentes à l'aide des extensions suivantes :

- La première extension concerne les agrégations. L'**agrégation générale** consiste à agréger les valeurs d'une mesure entre n'importe quel niveau des hiérarchies en utilisant toujours la même fonction. L'**agrégation multiple** consiste à agréger les valeurs d'une mesure en utilisant différentes fonctions d'agrégation selon les dimensions. L'**agrégation différentiée** consiste à agréger les valeurs d'une mesure en changeant de fonction d'agrégation entre chaque niveau d'agrégation. D'autre part, toutes ces agrégations ne s'effectuent pas nécessairement de manière uniforme à partir de tous les niveaux inférieurs (contrairement au mécanisme d'agrégation prévu dans les modèles multidimensionnels classiques). Par conséquent, nous introduisons un **mécanisme de contrainte** sur l'agrégation pour fixer le niveau d'agrégation valide permettant d'obtenir une agrégation supérieure.
- La seconde extension concerne l'**ordre d'exécution** des fonctions d'agrégation entre les différentes dimensions impliquées dans l'analyse. Il est possible d'avoir deux fonctions d'agrégation différentes pour chaque dimension considérée. Ces fonctions sont généralement non commutatives. Il faut donc prévoir un ordre d'exécution.

Soit $\mathcal{F} = \{f_1, f_2, ...\}$ un ensemble fini de fonctions d'agrégation.

Définition 4. Un *schéma multidimensionnel*, noté S, est défini par (F, D, Star, Order, Aggregate).

- F = $\{F_1, ..., F_n\}$ est l'ensemble des faits, si $\mid F \mid$ =1 alors le schéma multidimensionnel est appelé schéma en étoile tandis que si $\mid F \mid$ >1 alors le schéma est appelé schéma en constellation,
- D = $\{D_1, ..., D_m\}$ est l'ensemble des dimensions,
- Star : F $\rightarrow 2^D$ est une fonction qui associe chaque fait à un ensemble de dimensions en fonction desquelles il peut être analysé.
- Order : M $\rightarrow 2^{D \times \mathbb{N}^*}$ est une fonction qui lie à chaque dimension un ordre de priorité d'exécution utilisé pour les calculs d'agrégation par rapport à chaque mesure. La fonction d'agrégation de la dimension ayant l'ordre le plus petit est considérée comme prioritaire. Si les fonctions d'agrégation des deux dimensions sont commutatives, les deux dimensions auront le même numéro d'ordre,
- Aggregate : M $\rightarrow \mathcal{F} \times 2^D \times 2^{H \times P} \times \mathbb{N}^-$ associe chaque mesure à une fonction d'agrégation et à un niveau précis d'agrégation. Elle permet de définir les différents types de fonction d'agrégation supportés par notre modèle (générale, multiple, différentiée) :
  - Dans le cas où 2^D et $2^{H \times P}$ ne sont pas utilisés (ensembles vides), la fonction est une fonction d'agrégation générale.
  - Dans le cas où seul $2^{H \times P}$ n'est pas utilisé, la fonction est une fonction d'agrégation multiple utilisée pour agréger la mesure sur toute la dimension considérée.
  - Dans le cas où les deux ensembles 2^D et $2^{H \times P}$ sont utilisés, la fonction est une fonction d'agrégation différentiée utilisée pour agréger la mesure entre le paramètre considéré et le paramètre directement supérieur.

$\mathbb{N}^-$ sert à contraindre une agrégation en indiquant un niveau d'agrégation spécifique à partir duquel l'agrégation considérée doit se calculer. Une agrégation non contrainte sera associée à

0 tandis qu'une agrégation contrainte sera associée à une valeur négative pour forcer le calcul à partir d'un niveau inférieur choisi par rapport au niveau considéré.

Exemple 2. Nous reprenons le cas d'étude présenté en exemple 1. Les analystes veulent étudier la vitesse du vent selon sa direction, la géographie et la date tandis qu'ils souhaitent étudier les températures selon le temps, la date et la géographie. Nous définissons formellement une telle BDM par (F, D, Star, Order, Aggregate) où

- F = $\{F_{Température}, F_{Vent}\}$
- D = $\{D_{Géographie}, D_{Temps}, D_{Dates}, D_{Direction}\}$
- Star : F $\rightarrow 2^D$ | Star($F_{Température}$)= $\{D_{Géographie}, D_{Temps}, D_{Dates}\}$
  Star(F_{Vent})= $\{D_{Géographie}, D_{Dates}, D_{Direction}\}$
- Order : M $\rightarrow 2^{D \times \mathbb{N}^*}$ | Order(Vitesse)= {(Géographie, 1), (Direction, 2), (Dates, 2)}
  Order(Tem_Moy)= {(Géographie, 1), (Temps, 2), (Dates, 2)}
  Order(Tem_Max)= {(Géographie, 1), (Temps, 2), (Dates, 2)}
  Order(Tem_Min)= {(Géographie, 1), (Temps, 2), (Dates, 2)}
- Aggregate : M $\rightarrow \mathcal{F} \times 2^D \times 2^{H \times P} \times \mathbb{N}^-$ |
  Aggregate(Vitesse) = { (AVG(Vitesse), {}, {} , 0),
  (Select_center(Niveau_Administratif, Vitesse), {Géographie}, {}, 0),
  (AVG(Vitesse), {Géographie}, {(H_geo_simp, Pays)}, 0),
  (AVG(Vitesse), {Géographie}, {(H_geo_scien, Pays)}, 0)}
  Aggregate(Tem_Moy) = { (AVG(Tem_Moy), {}, {} , 0),
  (Select_center(Niveau_Administratif, Tem_Moy), {Géographie}, {}, 0),
  (AVG(Tem_Moy), {Géographie}, {(H_geo_simp, Pays)}, -1) [1],
  (AVG(Tem_Moy), {Géographie}, {(H_geo_scien, Ville)}, 0),
  (Moyenne(Tem_Moy,D_superficie),{Géographie},{(H_geo_scien,Département)},-1),
  (Moyenne(Tem_Moy, R_superficie), {Géographie},{(H_geo_scien, Région)}, -1),
  (Moyenne(Tem_Moy, P_superficie), {Géographie}, {(H_geo_scien, Pays)}, -1)}
  Aggregate(Tem_Min) = { (MIN(Tem_Min), {}, {} , 0),
  (Select_center(Niveau_Administratif, Tem_Min), {Géographie},
  {(H_geo_simp, Ville)}, 0),
  (Select_center(Niveau_Administratif, Tem_Min), {Géographie},
  {(H_geo_simp, Département)}, 0),
  (Select_center(Niveau_Administratif, Tem_Min), {Géographie},
  {(H_geo_simp, Région)}, 0)} [2]

La fonction Select_center(I, M) prend deux paramètres numériques. Elle rend la valeur M qui correspond au Max(I). S'il y a plusieurs Max(I) alors la fonction rend la moyenne des valeurs M correspondantes. Par exemple, si on applique Select_center(Niveau_Administratif, Temp_Moy) au niveau pays, elle rend la température de la capitale (ville ayant le niveau administratif maximal). De la même manière, si on l'applique au niveau région ou département, elle rend les températures de la capitale régionale ou de la préfecture.

Lemme 4. Les fonctions d'agrégation assurent la *couverture* du schéma multidimensionnel, c'est-à-dire qu'il ne doit pas exister de paramètre (niveaux d'agrégation) pour lequel nous ne connaissons pas la fonction d'agrégation à appliquer.

1. Les valeurs sont agrégées à partir des valeurs agrégées au niveau directement inférieur de celui considéré.
2. Aggregate(Tem_Max) ressemble Aggregate(Tem_Min) sauf qu'il utilise la fonction MAX au lieu de MIN.

$$\forall i \in [1..n], \forall m_k \in M^{Fi}, \begin{cases} \exists f \in \mathcal{F}, \exists x \in \mathbb{N}^- \mid \\ (f, \{\}, \{\}, x) \in \text{Aggregate } (m_k) \\ \forall D_j \in Star(F_i), \exists f \in \mathcal{F}, \exists x \in \mathbb{N}^- \mid \\ (f, \{D_j\}, \{\}, x) \in \text{Aggregate } (m_k) \\ \forall H_s \in H^{Dj}, \forall p_q \in P^{Dj} \setminus \{All^{Dj}\}, \exists f \in \mathcal{F}, \exists x \in \mathbb{N}^- \mid \\ (f, \{D_j\}, \{(H_s, P_q)\}, x) \in \text{Aggregate } (m_k) \end{cases}$$

De manière moins formelle, la couverture du schéma est réalisée de plusieurs façons :
- par l'utilisation d'une fonction d'agrégation générale,
- par l'utilisation d'une fonction d'agrégation multiple pour chaque dimension,
- par l'utilisation d'une fonction d'agrégation différentiée pour chaque paramètre,
- par combinaison de fonctions d'agrégation multiple et différentiée où la dimension, qui n'a pas de fonction multiple doit avoir une fonction différentiée pour chaque paramètre.

2.3 Formalismes graphiques

Associés aux définitions formelles, nous introduisons un formalisme graphique facilitant la compréhension du schéma de la BDM. Ces représentations graphiques sont de deux niveaux.

2.3.1 Schéma structurel

Le schéma structurel permet de visualiser globalement le schéma multidimensionnel de la BDM en masquant les mécanismes d'agrégation. Cette vue globale est obtenue à partir de la fonction Star.

Exemple 3. La BDM décrite formellement dans les exemples précédents, se représente graphiquement conformément à la figure 1. Le formalisme graphique que nous utilisons repose sur les propositions de (Golfarelli et al., 1998) et (Ravat et al., 2001). Cette BDM permet d'analyser les indicateurs (mesures) de météo : températures moyenne, maximale et minimale ainsi que la vitesse du vent. Ces mesures sont organisées dans deux faits : 'Vent' et 'Température'. Chaque fait est associé à trois dimensions parmi quatre dimensions : 'Géographie', 'Dates', 'Temps' et 'Direction'. La dimension 'Géographie' se compose de deux hiérarchies correspondant à la manière d'agréger les données : 'Simple' et 'Scientifique'. La dimension 'Temps' a une seule hiérarchie qui ordonne les granularités horaires auxquelles les températures ont été mesurées pendant la journée. La dimension 'Direction' a une hiérarchie qui n'a que le niveau racine et le niveau 'ALL'. La dimension 'Date' comprend plusieurs hiérarchies qui organisent les niveaux des granularités de la journée.

2.3.2 Schéma d'agrégation

Pour chaque mesure $m_k \in F_i$, un schéma d'agrégation peut être obtenu grâce aux fonctions Order et Aggregate. Cette vision détaille les mécanismes d'agrégation impliqués durant une analyse faisant intervenir la mesure considérée en ne faisant apparaître que les éléments structurels directement liés à la mesure (F_i et Star(F_i))

Exemple 4. A partir du schéma structurel présenté dans l'exemple précédent, il est possible d'obtenir les schémas d'agrégation des différentes mesures. La figure 2 décrit les trois

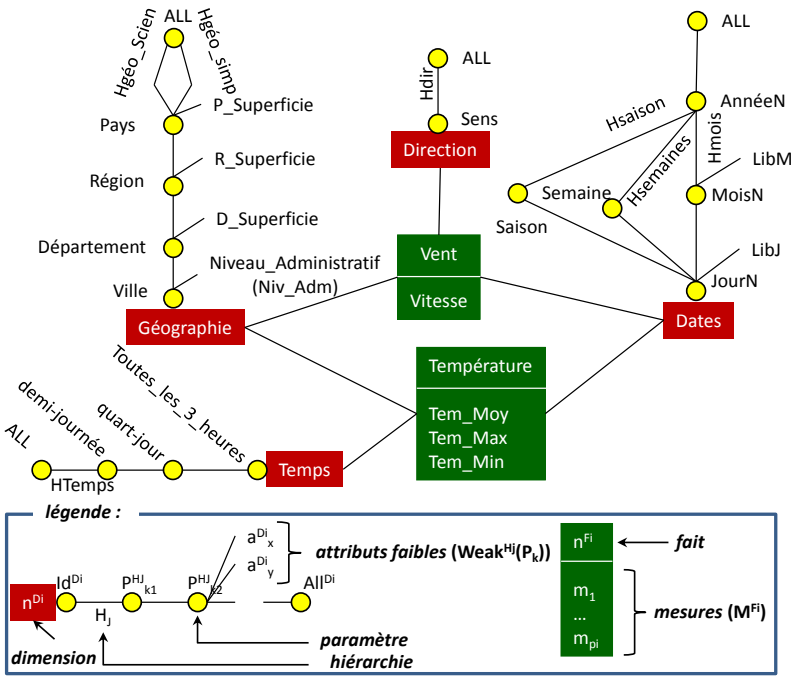

FIG. 1 – *Schéma structurel.*

schémas d'agrégation correspondant aux mesures 'Vitesse', 'Tem_Moy' et 'Tem_Min' (nous ne présentons pas celui de la mesure 'Tem_Max').

Comme la figure 2 l'illustre, l'ordre d'exécution est symbolisé par des chiffres dans des cercles sur les arcs reliant le fait aux dimensions et les fonctions d'agrégation sont modélisées par des losanges. Chaque losange fait apparaître la fonction d'agrégation associée à la mesure et l'éventuelle contrainte d'agrégation. Les positions des losanges dépendent du type de fonction :

- la fonction générale est représentée par un losange sur le bord du fait,
- la fonction d'agrégation multiple est localisée sur l'arc reliant le fait à la dimension,
- la fonction d'agrégation différentiée étiquette l'arc reliant deux paramètres.

Les agrégations contraintes sont calculées à partir du niveau directement inférieur (contrainte fixée à -1) ; par exemple, la température moyenne par pays est calculée à partir des températures par région. Dans l'hypothèse où nous aurions choisi de calculer cette température par pays à partir des températures par villes, la contrainte aurait été fixée à -3.

En résumé, nous proposons de représenter graphiquement une BDM à deux niveaux : le schéma structurel donne une vue complète des éléments structurels multidimensionnels (faits, dimensions, hiérarchies) en masquant la complexité due aux agrégations tandis que les schémas d'agrégation présentent de manière détaillée les mécanismes d'agrégation liés à une mesure (agrégations multiples, différentiées et générales, contraintes d'agrégation et ordre d'agrégation) en minimisant la complexité induite par le schéma multidimensionnel en constellation.

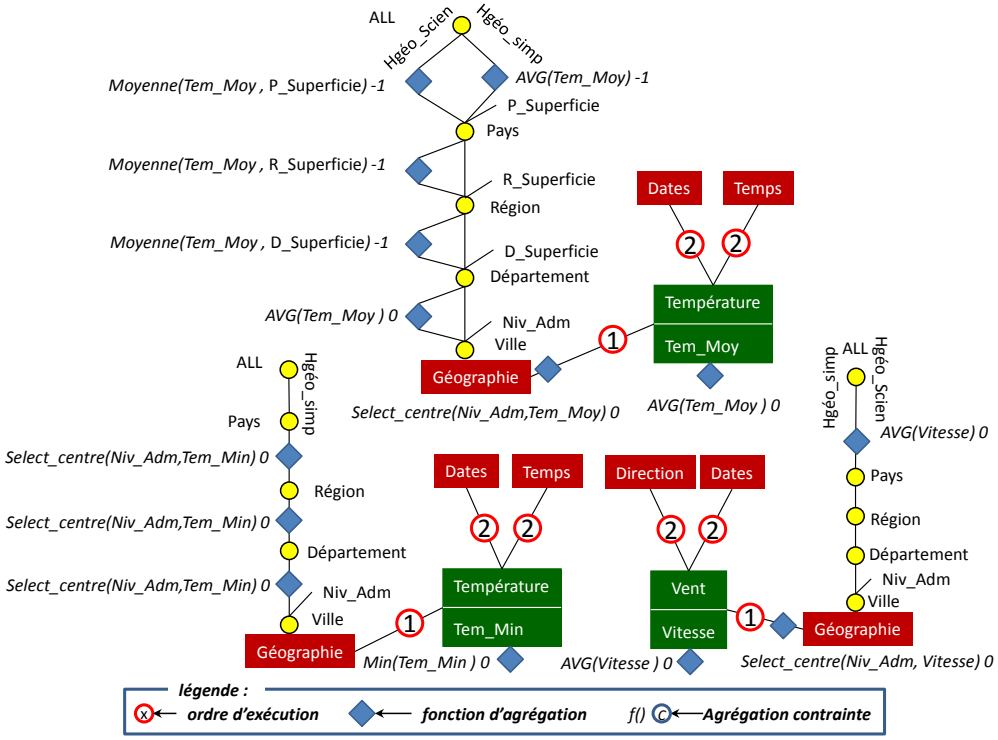

FIG. 2 – *Schémas d'agrégation.* [3]

3 Modèle logique R-OLAP

3.1 Approche classique

L'implantation courante repose sur l'approche dite R-OLAP. Elle consiste à utiliser l'approche relationnelle pour implanter les schémas multidimensionnels (Kimball, 1996). Cette approche procure de nombreux avantages dont notamment la réutilisation des mécanismes de gestion des données éprouvés depuis des décennies et la capacité de ces systèmes à gérer des volumes de données importants.

Les structures multidimensionnelles conceptuelles (*faits, dimensions*) sont donc traduites au niveau logique sous la forme de relations (Kimball, 1996). La structuration hiérarchique des attributs de dimension (*hiérarchies*) au niveau conceptuel est exploitée pour optimiser la BDM. Cette optimisation consiste à pré-calculer les agrégations nécessaires aux décideurs lors de leurs interrogations et analyses OLAP au sein de l'espace multidimensionnel (Gray et al., 1996). Ces pré-agrégations peuvent se modéliser par un treillis de pré-agrégats (Chaudhuri et Dayal, 1997) où chaque nœud représente un pré-agrégat et chaque arc représente le calcul d'agrégation. Lorsque la fonction d'agrégation utilisée est distributive ou algébrique, un agrégat est calculable à partir de l'agrégat directement inférieur, tandis que dans le cas d'une agrégation holistique, l'agrégat se calcule en cheminant jusqu'aux relations de base.

3. Ici nous utilisons l'abréviation (Niv_Adm) pour (Niveau_Administratif)

Exemple 5. Pour illustrer l'implantation R-OLAP classique, nous utilisons un exemple simplifié de BDM (figure 1). Le schéma structurel que nous utilisons dans cet exemple est décrit dans la figure 3 (a).

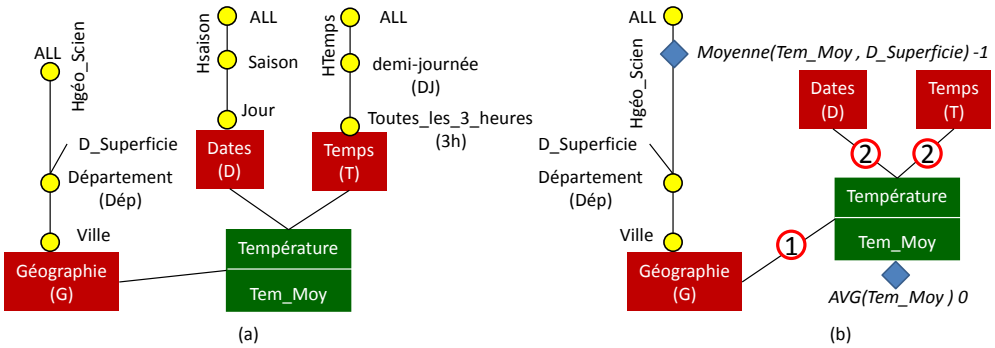

FIG. 3 – *Schémas structurel et d'agrégation de l'exemple simplifié.*

A partir de ce schéma conceptuel, sont construites les structures relationnelles suivantes :

1. Le fait et les dimensions donnent lieu aux relations suivantes :
   Géographie(**Ville**, Département, D_Superficie)
   Dates(**Jour**, Saison)
   Temps(**Toutes_les_3_heures**, demi_journée)
   Température(**Ville#, Jour#, Toutes_les_3_heures#**, Tem_Moy)
2. Les hiérarchies sont exploitées pour compléter le schéma par un ensemble de relations pré-calculant les agrégats potentiellement nécessaires aux calculs des requêtes utilisateurs. Le treillis de pré-agrégats dans la figure 4 représente l'ensemble de ces relations.

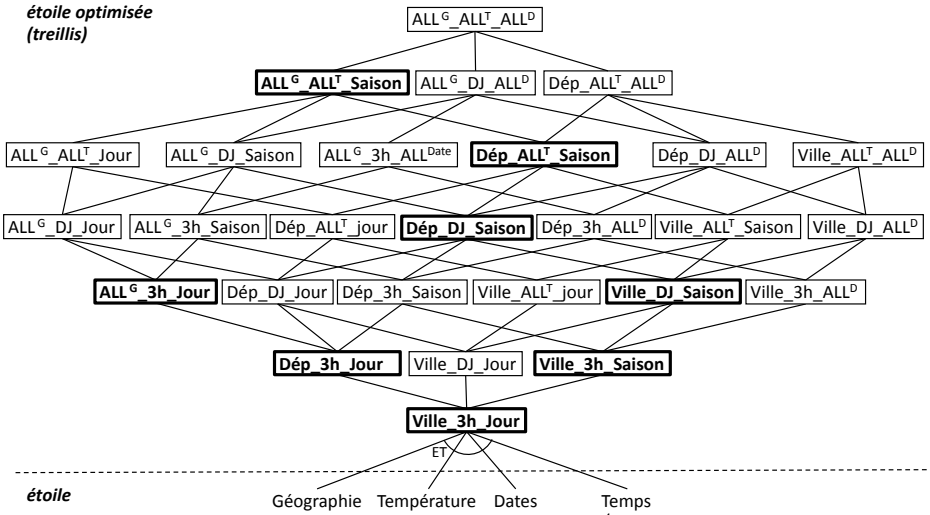

FIG. 4 – *Treillis classique.* [4]

4. Ici, nous avons utilisé les abréviations (qui sont entre parenthèses dans la figure 3)

Chaque nœud représente une relation. Par exemple, les nœuds 'Dép_3h_Jour' et 'ALL^G_ 3h_Jour' correspondent aux relations suivantes :

Dép_3h_Jour ⇒ Dép_3h_Jour (**Département, Jour, Toutes_les_3_heures**, Tem_Moy)

ALL^G_3h_Jour ⇒ ALL^G_3h_Jour (**Jour, Toutes_les_3_heures**, Tem_Moy)

Dans ces relations, l'attribut 'Tem_Moy' représente la moyenne des températures calculée par la fonction d'agrégation AVG. Il s'agit, ici, d'un cas de fonction algébrique. Dans cette approche classique, contrairement à notre proposition, une fonction d'agrégation unique est utilisée dans l'ensemble du treillis pour la mesure 'Tem_Moy'.

3.2 Extensions par les agrégations multiples et différentiées

Les extensions que nous avons introduites dans le modèle conceptuel impactent le treillis.

Typage des arcs. Les fonctions d'agrégation multiples et différentiées impliquent l'utilisation d'agrégations différentes sur chaque arc du treillis, contrairement à l'approche classique qui considère une fonction d'agrégation unique. Les différentes fonctions d'agrégation sont décrites dans la BDM au travers d'un méta-schéma que nous ne présentons pas dans cet article en raison du manque de place. Ce méta-schéma décrit les structures multidimensionnelles (*faits, dimensions, hiérarchies*) correspondant aux relations ROLAP. Il décrit également les différentes fonctions d'agrégation et les contraintes de calcul.

La possibilité pour une même mesure d'utiliser différentes fonctions d'agrégation selon les paramètres nécessite de typer les arcs du treillis. Ce typage permet d'indiquer entre deux nœuds la fonction d'agrégation correspondante.

Exemple 6. Nous complétons le schéma structurel de l'exemple simplifié (exemple 5) par son schéma d'agrégation (figure 3 (b)). Nous utilisons deux fonctions d'agrégation pour calculer la moyenne des températures : pour le calcul de la moyenne par département nous utilisons la moyenne classique comme dans tout le schéma (AVG) tandis que le calcul de la moyenne générale sur la dimension 'Géographie' se fait en pondérant par la superficie des départements. Dans le treillis, il faut donc distinguer les arcs qui relient les nœuds faisant intervenir le paramètre 'Ville' aux nœuds faisant intervenir le paramètre 'Département' (qui utilisent la fonction AVG), de ceux qui relient les nœuds faisant intervenir le paramètre 'Département' aux nœuds faisant intervenir le paramètre 'ALL' (qui utilisent la fonction moyenne pondérée). Dans la figure 5, les arcs correspondant à la fonction d'agrégation moyenne (AVG) apparaissent en trait simple tandis que les arcs correspondant à la fonction moyenne pondérée apparaissent en double trait.

Elagage du treillis. Dans le modèle conceptuel, nous avons introduit un ordonnancement des dimensions. Cet ordre spécifie l'ordre avec lequel doivent être exécutées les fonctions d'agrégation entre les dimensions. Cette extension de la modélisation multidimensionnelle induit des chemins invalides dans le treillis classique. Nous exploitons ainsi ce mécanisme d'ordonnancement pour effectuer un élagage du treillis (simplification par suppression d'arcs).

Exemple 7. Dans notre exemple (figure 3 (b)), nous ne pouvons appliquer la fonction de la moyenne pondérée (Moyenne (Tem_Moy, D_Superficie)) sur la dimension 'Géographie' après l'application de la fonction (AVG(Tem_Moy)) sur la dimension 'Dates' car cela donnerait un

résultat erroné. Autrement dit, nous ne pouvons pas calculer la moyenne générale des températures par saison (nœud 'ALL^G_ALL^T_Saison') à partir de la moyenne des températures des départements par saison (nœud 'Dép_ALL^T_Saison'). Donc, l'arc entre 'Dép_ALL^T_Saison' et 'ALL^G_ALL^T_Saison' peut être supprimé.

Modification d'arcs. Dans notre modèle, nous avons proposé également un mécanisme de contrainte sur l'agrégation pour fixer le niveau d'agrégation valide à partir duquel se calcule une agrégation supérieure. Ce niveau valide n'est pas forcément le niveau directement inférieur. Nous exprimons ce cas lorsque nous utilisons une valeur de la contrainte différente de 0 (l'agrégation est calculable à partir de n'importe quel niveau inférieur) ou -1 (l'agrégation est calculable uniquement à partir du niveau directement inférieur). Les contraintes différentes 0 et -1 induisent potentiellement des changements de chemins dans le treillis.

Blocage de la transitivité. Les contraintes associées aux agrégations ont une autre répercussion sur le treillis. Les arcs obtenus à partir de ces contraintes (arcs contraints) imposent de calculer un nœud à partir d'un nœud précis. Il est alors interdit de calculer un nœud supérieur par transitivité des nœuds inférieurs comme cela est classiquement possible. Ainsi, les chemins de calcul sont bloqués dès qu'un arc contraint intervient.

Exemple 8. Selon le schéma d'agrégation de l'exemple simplifié (exemple 5) le nœud 'Ville_DJ_Saison' est calculable à partir du nœud inférieur direct 'Ville_3h_Saison', par transitivité, il est également calculable à partir du nœud inférieur 'Ville_3h_Jour'. Par contre, l'arc issu de la contrainte de la fonction 'Moyenne (Tem_Moy, D_Superficie)', qui associe le nœud 'Dép_3h_Jour' au nœud 'ALL^G_3h_Jour', bloque la transitivité des calculs. Donc, le nœud 'ALL^G_3h_Jour' est calculable à partir du nœud inférieur direct 'Dép_3h_Jour' mais il n'est pas calculable par transitivité à partir du nœud inférieur, c'est-à-dire, il n'est pas calculable à partir du nœud 'Ville_3h_Jour'.

De la même manière, les ordres d'exécution entre dimensions provoquent un blocage de la transitivité. Lorsque deux dimensions possèdent un ordre différent, cela induit dans le treillis des arcs non transitifs, c'est-à-dire des arcs à partir desquels il n'est possible de calculer l'agrégation qu'avec le nœud inférieur directement lié.

Exemple 9. Selon le schéma d'agrégation de l'exemple simplifié (exemple 5) le nœud 'Dép_DJ_Saison' est calculable par transitivité à partir du nœud 'Dép_3h_Jour' mais il n'est pas calculable par transitivité à partir du nœud 'Ville_3h_Jour', parce que le schéma d'agrégation (figure 5(b)) impose par l'ordre d'exécution de calculer d'abord les températures en fonction de la dimension 'Géographie' (nœud 'Dép_3h_Jour') pour pouvoir calculer ensuite les températures en fonction des dimensions 'Dates' et 'Temps' ('Dép_DJ_Saison').

La figure 5 décrit un treillis de pré-agrégats contrôlés dans lequel sont répercutées toutes les conséquences de notre modèle conceptuel. Les arcs étiquetés par un cercle barré sont obtenus soit à partir des contraintes d'agrégation, soit à partir de l'ordre d'exécution imposé entre les dimensions.

4 Conclusion

Dans cet article nous proposons un modèle conceptuel de données multidimensionnelles permettant d'associer à une seule mesure plusieurs fonctions d'agrégation suivant les axes

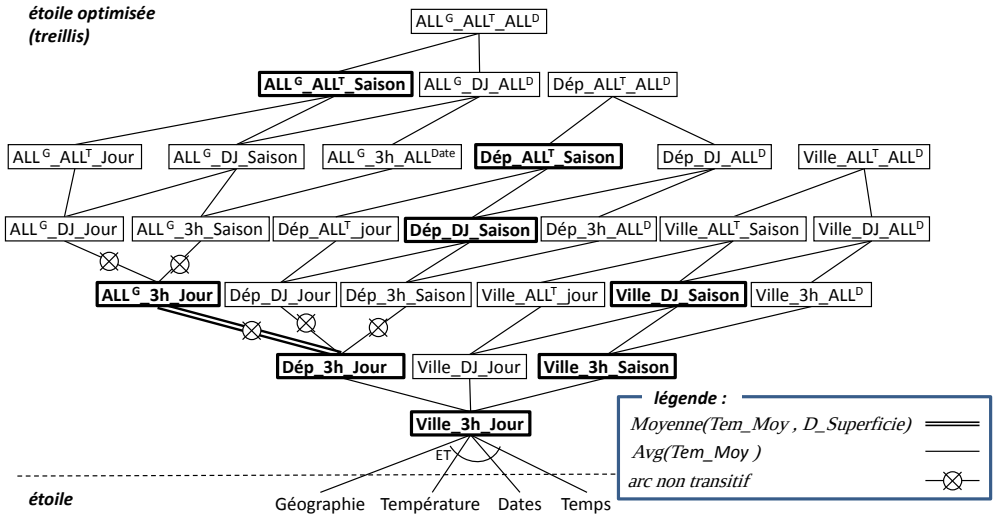

FIG. 5 – *Treillis de pré-agrégats contrôlés.*

d'analyse (fonctions multiples) et les niveaux de granularité (fonction différentiées). Le modèle proposé est suffisamment expressif pour contrôler la validité de ces fonctions : les contraintes d'agrégation fixent le niveau avec lequel une agrégation doit se calculer tandis que l'ordre d'exécution définit l'ordre de calcul entre les dimensions. Ce modèle s'appuie sur des formalismes graphiques à deux niveaux : le schéma structurel qui décrit les structures multidimensionnelles en masquant la complexité des agrégations et les schémas d'agrégation qui décrivent précisément les mécanismes d'agrégation liés à une mesure. En outre, au niveau logique (R-OLAP), nous montrons comment exploiter notre modélisation conceptuelle au sein de treillis de pré-agrégats contrôlés.

Nous envisageons de poursuivre nos travaux en revisitant les algorithmes de calcul des pré-agrégats adaptés à notre modélisation et en définissant une fonction de coût pour l'implantation et la maintenance d'une BDM optimisée. Nous envisageons également de poursuivre ces travaux par l'étude des opérateurs OLAP appliqués à notre modèle.

Références

Abelló, A., J. Samos, et F. Saltor (2006). Yam2 : A multidimensional conceptual model extending uml. *Information Systems 31*, 541–567.

Boulil, K., S. Bimonte, et F. Pinet (2011). Un modèle uml et des contraintes ocl pour les entrepôts de données spatiales. de la représentation conceptuelle à l'implémentation. *Ingénierie des Systèmes d'Information (ISI) 16(6)*, 11–39.

Chaudhuri, S. et U. Dayal (1997). An overview of data warehousing and olap technology. *SIGMOD Record 26*, 65–74.

Golfarelli, M., D. Maio, et S. Rizzi (1998). Conceptual design of data warehouses from e/r schemes. *Intl. Conf. HICSS'98 7*, 334–343.

Gray, J., A. Bosworth, A. Layman, et H. Pirahesh (1996). Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-total. *Intl. Conf. ICDE 96*, 152–159.

Gyssens, M. et L. V. S. Lakshmanan (1997). A foundation for multi-dimensional databases. *Intl. Conf. VLDB 97*, 106–115.

Harinath, S., R. Zare, S. Meenakshisundaram, M. Carroll, et G.-Y. L. D. (2009). *Professional Microsoft SQL Server Analysis Services 2008 with MDX,* .

Kimball, R. (1996). *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Wiley & Sons, USA (1996).* USA: John Wiley & Sons.

Lujàn-Mora, S., J. Trujillo, et I. Y. Song (2006). A uml profile for multidimensional modeling in data warehouses. *Data & knowledge Engineering 59*, 725–769.

Mazón, J. N., J. Lechtenbörger, et J. Trujillo (2009). A survey on summarizability issues in multidimensional modelling. *Data & Knowledge Engineering 68*, 1452–1469.

Oliveira, R., F. Rodrigues, P. Martins, et J. P. Moura (2011). Extending the dimensional templates approach to integrate complex multidimensional design concepts. *Intl. Conf DaWaK LNCS 6862*, 26–38.

Pedersen, T., C. Jensen, et C. Dyreson (2001). A foundation for capturing and querying complex multidimensional data. *Information Systems 26, n 5*, 383–423.

Prat, N.and Wattiau, I. et J. Akoka (2010). Representation of aggregation knowledge in olap systems. *The 18th European Conference on Information Systems(ECIS).*

Ravat, F., O. Teste, R. Tournier, et G. Zurfluh (2008). Algebraic and graphic languages for olap manipulations. *Intl. Journal of Data Warehousing and Mining, IGI Publishing 4(1)*, 17–46.

Ravat, F., O. Teste, et G. Zurfluh (2001). Modélisation multidimensionnelle des systèmes décisionnels. *1ère journée Extraction et Gestion des Connaissances(EGC) n 1-2*, 201–212.

Torlone, R. (2003). *Conceptual Multidimensional Models, Chapitre 3 dans Multidimensional Databases : Problems and Solutions,*. IGI Publishing Group.

Vassiliadis, P. et T. K. Sellis (1999). A survey of logical models for olap databases. *SIGMOD Record 28(4)*, 64–69.

Vassiliadis, P. et S. Skiadopoulos (2000). Modelling and optimisation issues for multidimensional databases. *Intl.Conf CAISE 2000 LNCS 1789*, 482–497.

Summary

Many models have been proposed for multidimensional data base modeling. These propositions consider the same aggregate function to determine the values of a measure with different levels of granularity of the multidimensional space. We propose a new conceptual model more flexible supporting multiple differentiated aggregations. Multiple aggregation allows to associate to the same measure, a different aggregation function for each dimension. Differentiated aggregation allows specific aggregations at each parameter. Our model is based on a double graphical formalism expressive enough to control the validity of aggregate functions. We also study the consequences of this conceptual modeling for effective building of lattices of pre-aggregates in the context of R-OLAP.