

Résumé :

Avec le développement des sociétés et l'accroissement des informations des bases de données traditionnel ne sont plus suffisante pour gérer et manipuler ses gros volumes de données.

Les entrepôts de données spatiaux ont été proposé comme solution pour gérer le stockage des bases de données géographiques volumineuse, ajouter à cela l'OLAP spatial dit SOLAP est une extension d'OLAP qui permet l'exploration des données spatiales.

D'un autre côté, le data mining spatial permet d'extraire des données ou d'en ressortir de nouvelles à partir d'anciennes tout en prenant compte de l'interaction dans l'espace.

Notre objectif est le développement d'un prototype de system géodécisionnel qui permet l'exploration d'entrepôts de données spatiaux à travers la technologie SOLAP et la découverte de connaissances en utilisant le data mining spatial.

Mots clés : Entrepôts de données spatiaux, informations spatiales, Data Mining, java, OLAP, OLAP SPATIAL.

Abstract:

With the development of society and the increase in traditional databases information is no longer sufficient to manage and handle its big data.

Spatial data warehouses have been proposed as a solution to manage the storage of large geographic databases, add to that the spatial OLAP known as SOLAP is an extension of the OLAP which allows the exploration of spatial data.

Sommaire:

INTRODUCTION GENERALE.....	iv
Introduction :	1
I - Les Systèmes d'Information Géographique :	1
I. 1. Définition :	1
I. 2. La Donnée Géographique :	1
II- Les Entrepôts de Données :	2
II. 1 - Définition :	2
II. 2 – L'intérêt des Entrepôts de Données :	3
II. 3 – Architectures :	3
II. 4 – Les Magasins de Données - Data Marts :	4
II. 5 – Les Schémas :	4
II. 6 – Les Opérations sur les données :	6
II. 7 – Modélisation d'un entrepôt de données :	7
II. 8 – Cube multidimensionnel :	9
Conclusion :	11
Introduction :	12
III – OLAP, OLTP et SOLAP :	12
III. 1. Définition :	12
III. 2 – Organigramme opérationnelle OLTP et OLAP :	14
III. 3 - Les Requêtes Multidimensionnelles :	14
III. 4 - Intégration des SIG avec OLAP :	15
III. 4.1 OLAP Spatial :	16
III. 4.2. Architecture SOLAP :	16
III. 4.3. Topologies :	18

IV.4.2.1.1 – Généralisation spatial :	22
IV.4.3-Analyse explicative :	23
IV.4.4-Technique utiliser :	24
Conclusion :	25
Introduction :	26
V. Réalisation du projet :	26
V. 1. Les outils de réalisation :	26
V. 2. Source des données :	27
V. 3. Préparation de base de données :	27
V. 4. Résultat de l'application :	36
Conclusion :	38
CONCLUSION GENERALE	39
Bibliographie :	40
Webographie:	41

Liste des Tableaux :

Table 1-1 : Les caractéristiques d'OLTP et l'OLAP13

Liste des Figures :

Figure 1-1 : La représentation "raster" de la donnée – point et ligne.....	1
Figure 1-2 : La représentation "vecteur" de la donnée.....	2
Figure 1-3 : Illustration simplifiée d'un system décisionnel – entrepôt de données.....	3
Figure 1-4 : Illustration simplifiée d'un Data Mart.....	4
Figure 1-5 : Exemple de schéma en étoile présenté par Microsoft.....	5
Figure 1-6 : Exemple de schéma en flocon.....	5
Figure 1-7 : Exemple de schéma en faitde constellation.....	6
Figure 1-8 : Exemple de dimensions.....	8
Figure 1-9 : Modélisation par sujets.....	9
Figure 1-10 : Exemple de cube de données.....	10
Figure 2-1 : Simplification de la relation d'OLTP et OLAP	12
Figure 2-2 : OLTP et OLAP en exécution.....	14
Figure 2-3 : Positionnement des outils vis à vis de l'information.....	16
Figure 2-4 : Logo imaginaire d'un SOLAP	16
Figure 2-5 : Possibilités d'architecture SOLAP	17
Figure 2-6 : Processus d'Extraction de Données	19
Figure 2-7 : techniques datamining classique.....	22
Figure 3-1 : La base de données créée.	29
Figure 3-2 : Modèle conceptuel de données.....	31
Figure 3-3 : Modèle Multidimensionnelle.....	32
Figure 3-4 : Création des tables.	33
Figure 3-5 : Diagramme des relations.	34
Figure 3-7 : Diagramme du cube créé.	35
Figure 3-8 : Affichage des dimensions crée.	36
Figure 3-9 : Analyse sur le cube.	37
Figure 3-9.1 : Interface d'application.....	38
Figure 3-9.2 : Résultat de la requête MDX.....	38
Figure 3-10 : Résultat d'application de l'arbre de décision.....	39

Liste des Abréviations :

AFC	Analyse Factorielle des Correspondances
API	Application Programming Interface
BD	Big Data
BI	Business Intelligence
DM	Data Mining Data Mart
DMS	Data Mining Spatial
DW	Data Warehouse
ED	Entrepôt de Données
ETL	Extraction, Transformation, Load
IDE	Integarted developement Environment
KD	Knowledge Discovery
KDD	Knowledge Discovery in Databases
MDX	Multi-Dimensional eXpressions
OLAP	On-Line Analytical Processing
OLTP	On-Line Transactional Processing
SDK	Software Development Kit
SGBDR	Système de Gestion des Bases de données Relationnelles
SIG	Systèmes d'Information Géographique
SOLAP	Spatial OLAP
SSAS	SQL Server Analysis Service
TI	Technologie d'Information

INTRODUCTION GENERALE

Dans le domaine du géo-décisionnel, plusieurs paramètres sont pris en considération : d'abord il faut un support pour la prise de décision qui offre une meilleure exploitation des informations contenue dans le Système d'Information de l'entreprise. Les data warehouses peuvent accomplir ce rôle. Cependant, les données utiliser peuvent avoir une dimension géographique, dans ce cas, l'intégration de l'information géographique dans l'entrepôt de données est inévitable pour la prise de décision.

D'un autre côté, les Systèmes d'Informations Géographiques (SIG) sont chargés de stocker, afficher et analyser les données géographiques. Ils représentent les données géographiques sous un format vecteur (géométrique) ou raster (image), chaque mode de représentation a ses avantages et ses inconvénients et peut être utilisé séparément mais pour les SIG modernes l'exploitation de ces deux types de modèles doit être obligatoire. La force des SIG est l'analyse spatiale, elle permet d'analyser des données sur une carte en exploitant leur dimension spatiale.

D'autre part l'OLAP SPATIAL (SOLAP) qui tient compte de l'interaction dans l'espace, est considéré comme l'approche à suivre pour les nouvelles générations d'outils d'analyse spatiale. Elle combine l'utilisation de l'analyse spatiale issue des SIG et les entrepôts de données spatiales issues du business intelligence.

En somme les nouvelles solutions géo-décisionnelles doivent pouvoir construire de nouvelles connaissances à partir de données stockées et non volatiles, c'est là qu'entre en jeu le DATA MINING issue du business intelligence.

L'objectif de notre travail est l'implémentation d'une solution géo-décisionnelle complète basée sur l'analyse spatiale, le système SOLAP et le datamining spatial.

Dans ce rapport, nous commençons par donner quelques notions de bases sur les entrepôts de données, l'OLAP Spatial et le Data Mining. Nous présenterons ensuite la solution géo-décisionnelle que nous avons développé et nous finirons par une conclusion générale récapitulant notre travail.

Chapitre I

Entrepôts de Données

Introduction :	1
I - Les Systèmes d'Information Géographique :	1
I. 1. Définition :	1
I. 2. La Donnée Géographique :	1
II- Les Entrepôts de Données :	2
II. 1 - Définition :	2
II. 2 – L'intérêt des Entrepôts de Données :	3
II. 3 – Architectures :	3
II. 4 – Les Magasins de Données - Data Marts :	4
II. 5 – Les Schémas :	4
II. 6 – Les Opérations sur les données :	6
II. 7 – Modélisation d'un entrepôt de données :	7
II. 8 – Cube multidimensionnel :	9
Conclusion :	11

Introduction :

Dans la première section de ce chapitre, nous allons introduire le concept des Systèmes d'Informations Géographiques, après nous allons découvrir le terme d'Entrepôt de Données (i.e. Data Warehouse), ses composantes et ses objectifs. Nous allons aller un peu plus loin et parler d'On-Line Analytical Processing (OLAP) et On-Line Transactional Processing (OLTP), la différence entre ces deux processus et les caractéristiques de chacune, en passant par le langage de requête "MDX", puis dans la troisième partie de ce chapitre, nous allons parler des Systèmes d'Informations Géographiques (SIG) et les deux types des données géographiques, raster et vecteur.

Après, nous verrons comment intégrer la technologie OLAP et les SIG et découvrir le concept d'OLAP Spatial (SOLAP). Nous finirons par parler du data mining et ses apports dans les applications géodécisionnelles.

I - Les Systèmes d'Information Géographique :

I. 1. Définition :

Un Système d'Information Géographique (SIG) peut être définie comme un système mis en place sur l'ordinateur qui stocke, manipule, analyse et visualise les données du monde réel à partir de trois aspects :

Les données géographiques, sont définies dans l'espace et représentées généralement par le biais des cartes;

Les caractéristiques, attributaire, sont généralement composé de valeurs alphanumériques stockées dans des tables;

Les relations spatiales entre les éléments, sont appelés "relations topologiques". **[1.0]**

I. 2. La Donnée Géographique :

Les données géographiques sont des données localisées à la surface terrestre. On a généralement deux types de données : Raster et Vecteur.

Raster : Données où l'espace est divisé de manière régulière en ligne et en colonne; à chaque valeur ligne/ colonne (pixel) sont associées une ou plusieurs valeurs décrivant les caractéristiques de

Vecteur : Pour représenter les objets à la surface du globe, les SIG utilisent trois objets géométriques qui sont le point, la ligne et la surface.

La figure suivante illustre cette représentation.

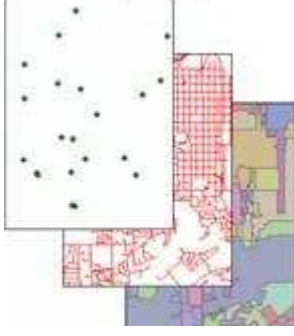


Figure 1-2 : La représentation "vecteur" de la donnée.

II- Les Entrepôts de Données :

II. 1 - Définition :

Une façon courante d'introduire un Entrepôt de Données (ED) est de se référer à la définition présentée par, William H. Inmon.

Selon Inmon, un entrepôt de données est une collection de données **orienté sujet, intégrées, non-volatiles** et **historisées** pour la prise de décisions en entreprise. [1]

Orienté sujet (ou Thématique) : Les entrepôts de données sont conçus pour aider à analyser les données. Par exemple, pour en savoir plus sur les données des ventes d'une entreprise, nous pouvons construire un entrepôt qui se concentre sur les ventes. L'utilisation de cet entrepôt, peut répondre à des questions comme « Qui était le meilleur client pour ce produit l'année dernière? » Cette capacité à définir un ED par sujet (les ventes dans ce cas), fait l'ED orienté sujet.

Intégré : L'intégration est largement liée à l'orientation du sujet. Les entrepôts de données doivent mettre les données provenant de sources différentes dans un format cohérent. Ils doivent résoudre des problèmes tels que les conflits de noms et les incohérences entre les unités de mesure. Quand ils y parviennent, ils sont censés être intégrés.

Non volatiles : Les données ne disparaissent pas et ne changent pas au fil des traitements et du temps (lecture seulement) ce qui est logique car l'objectif d'un ED est de permettre d'analyser tous les actions produites

II. 2 – L'intérêt des Entrepôts de Données :

Un entrepôt de données est un ensemble unique, complet et consistant de données obtenues depuis une variété de sources et rendu disponible à des utilisateurs finaux.

Le but général des ED est de fournir des données agrégées, qui sont dans un format adapté à la prise de décision, ce dernier ne se soucie des détails des opérations et de requêtes, mais ils sont intéressés par les tendances générales afin qu'ils puissent voir si l'effort de l'entreprise peut être mieux utilisé. [2]

II. 3 – Architectures :

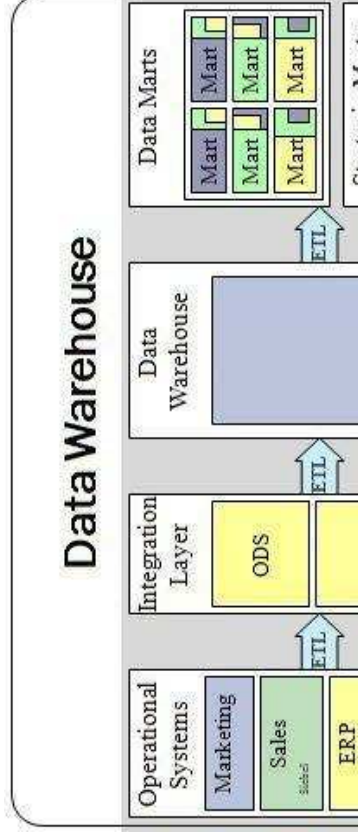
Les entrepôts de données et leurs architectures varient selon les spécificités de la situation de l'organisation. On définit trois architectures courantes :

De Base : Les utilisateurs finaux accèdent directement à des données provenant de plusieurs systèmes sources dans l'entrepôt de données.

Avec une zone de transit (Staging Area) : Dans une architecture de base nous avons besoin d'organiser et traiter nos données avant les mettre dans l'entrepôt, bien que la plupart des entrepôts utilisent une zone de transition à la place. Une zone de transit simplifie la construction des résumés et la gestion de l'entrepôt.

Avec une zone de transit et Data-Marts : Bien que l'architecture avec une zone de transit est assez fréquent, nous pouvons personnaliser notre architecture d'entrepôt pour les différents groupes au sein de notre organisation. Nous pouvons le faire en ajoutant des magasins de données (i.e. Data-Marts), qui sont des systèmes conçus pour une action donnée.

Figure 1-3 illustre un exemple où les achats, les ventes et les stocks sont séparés. [3]



II. 4 – Les Magasins de Données - Data Marts :

Les données d'un entrepôt sont segmentées en composants spécialisés appelés "magasins de données", ces magasins adressent des sujets individuels dans l'organisation, principalement utilisés pour une zone de mise au point spécifique à l'intérieur d'un département de l'entreprise ou d'une fonction d'affaires en particulier.

Par exemple, les données de l'entrepôt d'entreprise peuvent être segmentées en data marts de l'inventaire et l'expédition, finance, géographique, ressources humaines, gestion et ainsi de suite.

Le magasin de données est de portée plus limitée et sert de sand-box pour les analystes et les scientifiques de données pour tester de nouvelles techniques et méthodologies dans une fraction de temps et de coût par rapport à la construction d'un entrepôt de données complet.

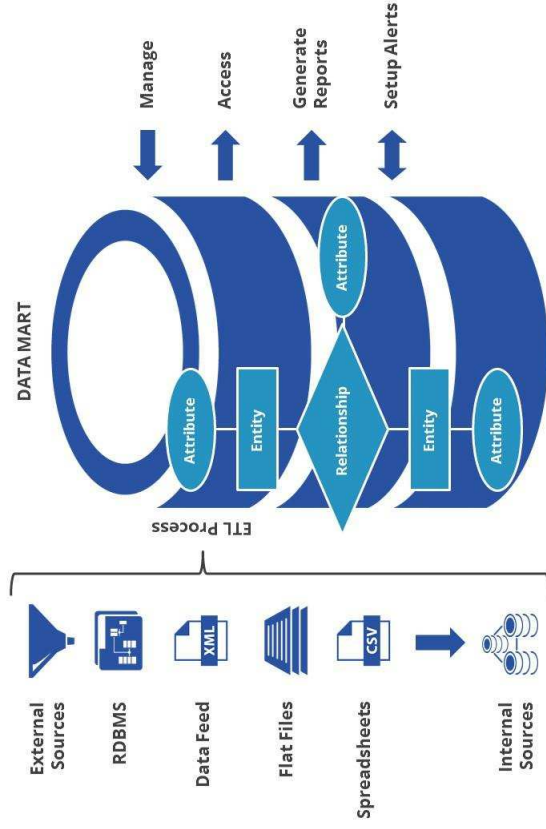


Figure 1-4 : Illustration simplifiée d'un Data Mart.

Pour la création d'un Entrepôt de Données nous devons définir les besoins du business, approuver le cadre d'application et créer un modèle conceptuelle.

II. 5 – Les Schémas :

Cette description présentée par Microsoft est mieux décrite dans la figure suivante.

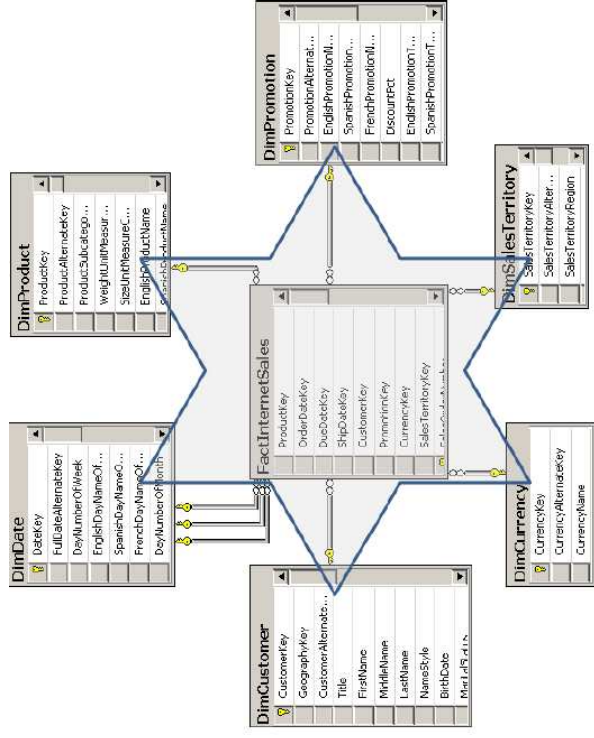
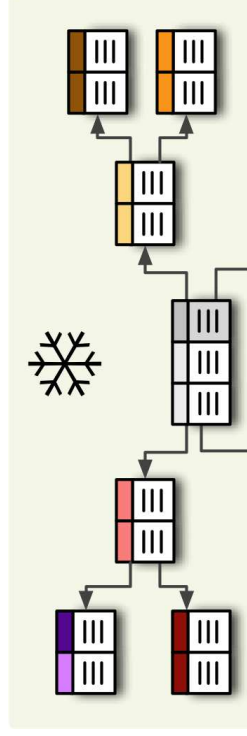


Figure 1-5: Exemple de schéma en étoile présenté par Microsoft.

La table de faits est au milieu du schéma entouré par Les tables de dimension. Cette approche de conception gère efficacement les requêtes simples.

II. 5. 2 - Schéma en Flocon :

Contrairement au schéma en étoile, les tables des dimensions dans un schéma en flocon sont normalisées. La normalisation divise les données dans des tableaux supplémentaires.



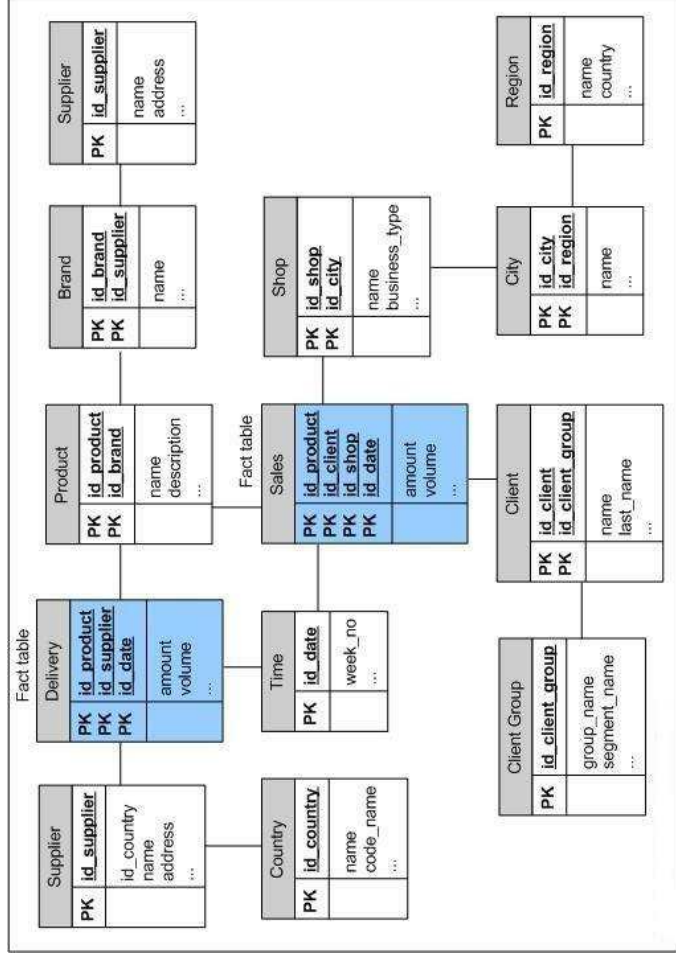


Figure 1-7 : Exemple de schéma en fait de constellation.

II. 6 – Les Opérations sur les données :

II. 6. 1 – Les outils d'ETL :

Les outils d'ETL (extraction, transformation et chargement) automatisent totalement la création, la maintenance et l'extension des entrepôts de données, data marts, micro marts et magasins de données opérationnelles.

- **Extraction** : Extraire les données de leur environnement d'origine (bases de données relationnelles, fichiers plats, ...). Utiliser une technique appropriée pour n'extraire que les données nécessaires. Données créées ou modifiées depuis la dernière opération d'extraction.
- **Transformation** : Une même donnée peut avoir une structure ou une valeur différente en fonction de la base (production, externe, utilisateurs) dont elle provient.

II. 7 – Modélisation d'un entrepôt de données :

Un modèle décisionnel peut être caractérisé par cinq axes : [21]

1. **La lisibilité.** L'utilisateur final doit être capable de comprendre le modèle.
2. **Les performances de chargement.** Le chargement de l'entrepôt de données doit être sans incidence sur le système transactionnel. La période et la fréquence de chargements doivent être choisies avec minutie et en concertation avec les administrateurs du système transactionnel.
3. **Les performances d'exécution des requêtes.** Les tailles phénoménales que peuvent rapidement prendre les entrepôts de données, exigent souvent de leurs administrateurs de calculer des agrégats et des données dérivées les plus susceptibles d'être utilisés. Comme il est difficile de prévoir toutes les requêtes émises par les utilisateurs, il devient difficile d'appliquer les techniques d'optimisation issues du transactionnel. D'autres approches d'optimisation basées sur le chemin d'accès sont utilisées.
4. **L'administration.** L'une des tâches de l'administration est de mettre l'entrepôt de données à la disposition de l'utilisateur en lui offrant les accès souhaitables et en lui interdisant les données auxquelles il n'a pas droit. Aussi, l'administration inclut le repérage et le recensement des requêtes les plus utilisées et la maîtrise et l'industrialisation de tous les processus d'extraction.

5. **L'évolutivité.** Le développement d'un entrepôt de données est incrémental et non itératif. Chaque projet décisionnel vient s'ajouter aux projets de départ pour faire un tout.

II. 7. 1 – Dénormalisation :

Pour une bonne modélisation d'un entrepôt de données, la dénormalisation est nécessaire. Elle résulte principalement de la fusion de plusieurs tables et l'introduction de redondances en dupliquant certains attributs.

Il n'existe pas de technique de dénormalisation standard. L'approche est pragmatique et découle d'une analyse précise des besoins de l'utilisateur. Lors de l'analyse, il est important de déterminer les sujets (patients, visites,...) que l'utilisateur considère comme les plus importants et sur lesquels il prévoirait porter ses études. Sujets que le concepteur de l'entrepôt de données considérera comme les éléments clés lors la construction des nouvelles tables. La

II. 7. 2 – Dimensions hiérarchiques:

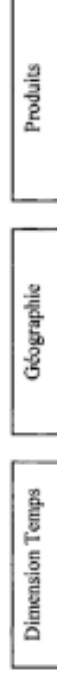
Dans un système décisionnel, une dimension peut être vue comme un système d'unités «contenus, contenus», où l'unité la plus générale est le contenant de toutes les autres unités définies sur le même axe.

Une dimension est un axe d'analyse regroupant des indicateurs de performance et correspondant à un sujet d'intérêt.

La dimension temporelle est présente dans pratiquement toutes les applications décisionnelles. La géographie et la hiérarchie de produits sont des dimensions potentielles. Le chiffre d'affaire d'une entreprise peut être étudié par mois, par trimestre ou par année. Il peut, aussi, être étudié par région, par catégorie de produits ou par client. Tous ces axes d'analyse représentent des dimensions intéressantes pour l'indicateur chiffre d'affaire. Les dimensions les plus fréquemment utilisées sont:

- *La dimension pays*
- *La dimension situation géographique*
- *La dimension structure de l'entreprise*
- *La dimension produits (qui peut être très volumineuse notamment dans le cas de la grande distribution)*
- *La dimension projets*
- *La dimension Clients*
- *La dimension Fournisseurs*
- *La dimension Temps*

Chaque unité peut être mesurée par une unité plus ou moins fine. Lors de l'exploitation, les analystes étendent ou réduisent la représentation des données suivant les axes des dimensions. Certaines dimensions nécessitent pour le besoin de leurs développements l'accès à la base de données. Ce sont généralement les dimensions non denses telles que la dimension géographie ou la dimension produits.



II. 7. 3 – Modélisation par sujets :

Les applications telles que conçues dans les bases de production ne sont pas adaptées pour être exploiter. Une modélisation par sujet s'avère nécessaire. Chaque sujet qui intéresse l'utilisateur est associé à une table gérée par le data warehouse. Pour ce faire il faut:

- *Isoler les données stratégiques.*
- *Déterminer les informations de détail nécessaires.*
- *Déterminer les résumés à conserver.*
- *Déterminer les métadonnées décrivant les tables (nom d'attribut, unité, valeur par défaut,... mémoriser si possible les sources de production pour chaque table, confidentialité.)*

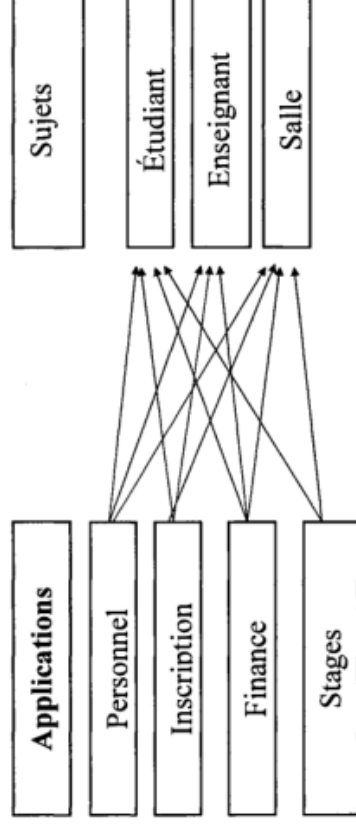


Figure 1-9 : Modélisation par sujets.

II. 7. 4 – Modélisation multidimensionnelle :

La modélisation multidimensionnelle est indépendante du modèle de base de données sous-jacent. Elle s'applique aux bases de données relationnelles, multidimensionnelles ou objet. Cette technique privilégie la performance et la considère comme un objectif central. La performance est déterminée par l'utilisateur.

Le chiffre d'affaire et la rentabilité sont des indicateurs de performance dans le domaine commercial alors que le taux d'occupation d'un lit et la durée de séjour d'un patient sont des indicateurs de performances en milieu hospitalier. L'analyse de ces indicateurs se fait à travers les dimensions.

II. 8. 1 – Définition :

Un cube de données est une représentation de N attributs extraits d'une table sous forme d'un cube. N-k attributs composant les dimensions le long desquelles les groupements sont possible, les k autres étant des mesures résultats de fonctions d'agrégations (généralement k=1). [21.5]

La **Figure 1-10** illustre la modélisation multidimensionnelle. Elle représente un cube de données sur l'attribut quantité selon les axes *N°Pro*, *N°Fou* et *Date* à partir de la table *Ventes* (*N°V*, *N°P*, *N°F*, *Date*, *Quantité*, *Prix*). Par projection, on obtient la table: *VentesCube* (*N°P*, *N°F*, *Date*, *Quantité*). Le cube n'est que la représentation de la vue *VentesCube*.

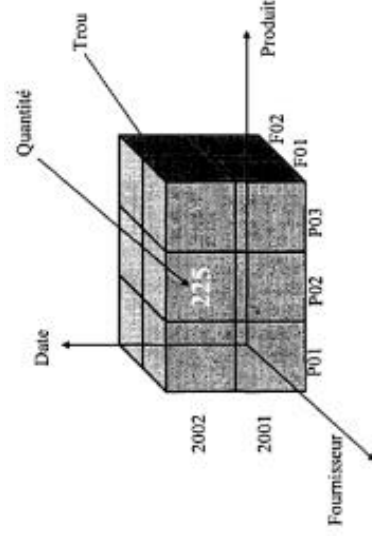


Figure 1-10 : Exemple de cube de données.

La valeur de la variable analysée appelée indicateur, figure dans chaque cube unitaire si elle est définie. Dans notre illustration de **Figure 1-10** à la date 2002, pour le produit P02 et le fournisseur F01, la quantité est 225.

Comme la plupart des valeurs ne sont pas définies le cube est plein de trous correspondant à des valeurs non existantes. En fait le cube n'est que du vide parsemé de quelques valeurs disséminées.

II. 8. 2 – Les opérateurs :

Drilldown (Dépliage) : Extension d'une dimension du cube en la remplaçant par une dimension à grains plus fins, soit en allant du global vers le détails. Dans notre exemple précédent du cube (Fournisseur, Année, Pays), on peut appliquer le Drilldown sur les axes

- Axe géographique: *Rolfup (Pays Région), Rolfup (Pays), Rolfup ()*
- Axe produits : *Rolfup (Fournisseurs)*

Le Drilldown, Rollup ne sont possibles entre deux attributs A et B que s'il y a une dépendance fonctionnelle entre ces deux derniers.

Slice (Coupe) : Sélection de tranches de cube par des prédicats selon une dimension. Il s'agit de sélectionner des portions du cube en filtrant une dimension selon une valeur ou une plage de valeurs. La dimension filtrée est alors translatée et éventuellement tassée si des trous apparaissent, en cas de filtrage sur plusieurs valeurs non successives par exemple.

Dans l'exemple précédent, on peut appliquer le Slice sur le cube (*Produits, Date, Ville*) selon les axes suivant:

- Axe du temps: *Slice (20/09/2001), Slice (>1998 and <1999)*
- Axe Produit: *Slice ('Ordinateurs')*
- Axe Géographique: *Slice ('Trois Rivières'), Slice ('Trois Rivières' or 'Montréal')*

La combinaison des opérateurs Drilldown, Rollup et Slice forme l'algèbre des cubes de données. Il n'y a aucune restriction autre que la nature des données sur les combinaisons possibles. Par exemple, avec le cube (*N°P, Date, Région*) on peut faire le calcul suivant:

Slice (Trois Rivières)
[Drilldown (Ville)
[Rollup (Année, Mois)
[Slice (>1998 and < 1999) [Cube]]]]

Sur le même cube on veut étudier sur l'année 98, par période d'un mois, dans la ville de trois Rivières, les quantités de produits vendues.

Les outils commerciaux définissent d'autres opérateurs de modélisation du cube, opérateurs pas ou peu étudiés par les scientifiques. On peut citer:

Drill through : Le Drill through consiste à visualiser une même information sous l'angle de plusieurs dimensions. On pourrait, par exemple, souhaiter à partir de la visualisation du chiffre d'affaires d'un point de vente par produits et pour un mois déterminé, obtenir la

Chapitre II

OLAP, OLTP, SOLAP et le Data mining

- Introduction :12
- III – OLAP, OLTP et SOLAP :12
 - III. 1. Définition :12
 - III. 2 – Organigramme opérationnelle OLTP et OLAP :14
 - III. 3 - Les Requêtes Multidimensionnelles :14
 - III. 4 - Intégration des SIG avec OLAP :15
- IV- DATA MINING:19
 - IV. 1. Définition :19
 - IV. 2. Processus d'extraction de données :19
 - IV. 3. Méthodes de Data Mining :20
 - IV. 4. Data Mining Spatial:22
- Conclusion :25

Introduction :

Les systèmes informatiques traditionnels ont pour but de traiter instantanément les données qu'ils reçoivent, procédant automatiquement à une mise à jour des fichiers du système de gestion de base de données. Le traitement de l'information est considéré comme une transaction que doit réaliser le système informatique, d'où l'utilisation des termes «transactionnel» et OLTP pour On Line Transaction Processing ou traitement transactionnel en ligne.

Nous allons voir les définitions de l'OLAP et OLTP, les caractéristiques de chacun d'eux, puis la relation entre OLAP et les SIG. Dans le chapitre suivant nous parlerons des techniques de datamining.

III – OLAP, OLTP et SOLAP:

III. 1. Définition :

On peut diviser les systèmes informatiques en transactionnelle (OLTP) et analytique (OLAP). En général, nous pouvons supposer que les systèmes OLTP fournissent des données sources pour les entrepôts, tandis que les systèmes OLAP aident à analyser.

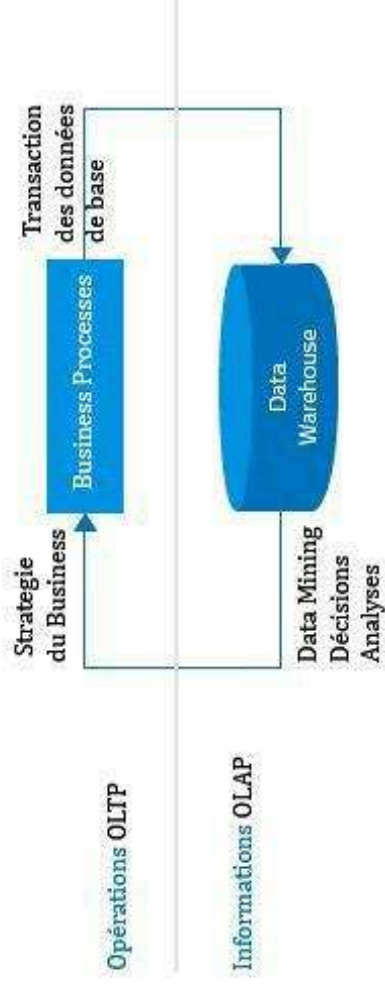


Figure 2-1 : Exemple de schéma en fait de constellation.

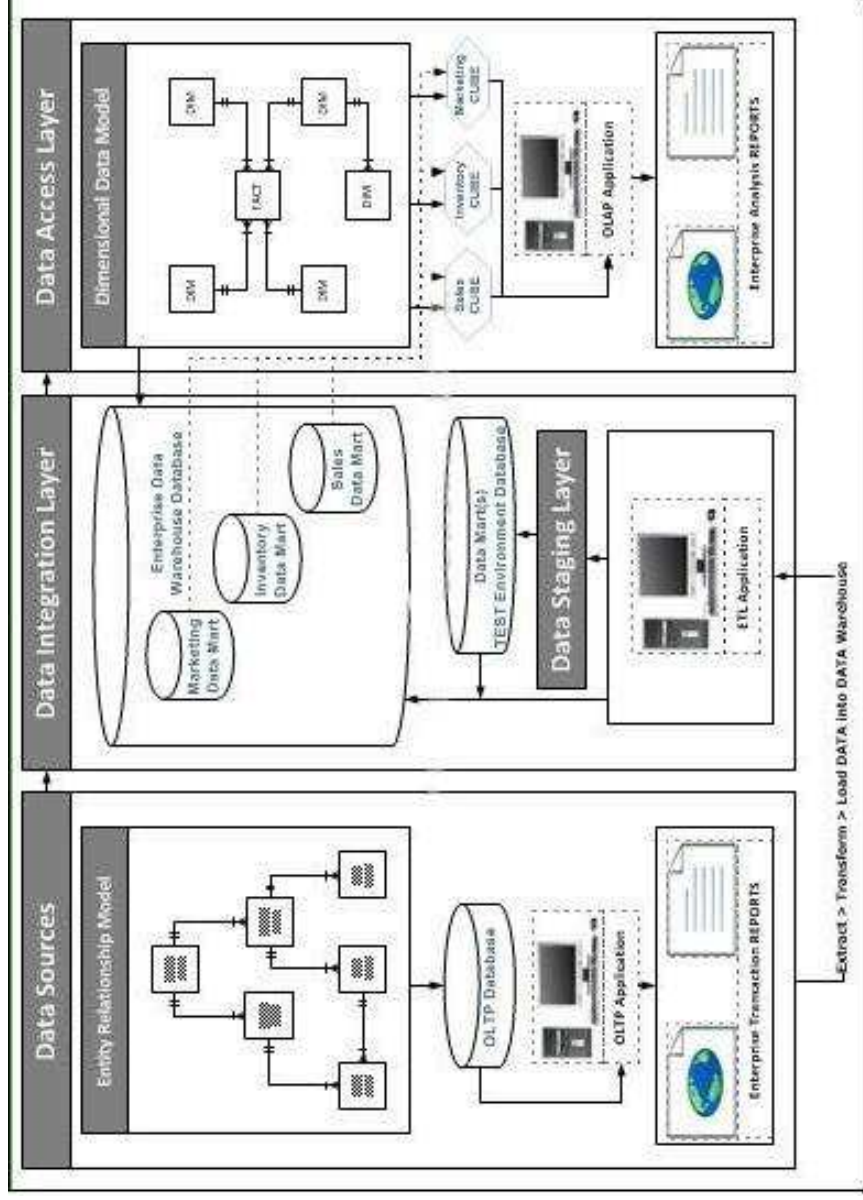
OLTP (On-line Transaction Processing): Caractérisée par un grand nombre de courtes transactions en ligne (INSERT, UPDATE, DELETE). Le point principal pour les systèmes OLTP

Le tableau suivant résume les différences principales entre OLTP et la conception du système OLAP.

	OLTP	OLAP
Source des données	Les données opérationnelles; OLTPs sont la source d'origine des données.	Consolidation de données; données OLAP provient des différentes bases de données OLTP
But des données	Contrôler et exécuter des tâches d'affaires fondamentales.	Aider à la planification, la résolution de problèmes, et aide à la décision
Qu'es ce que les données	Révèle un aperçu des processus d'affaires en cours	Vues multidimensionnelles de toutes sortes d'activités commerciales
Insertions et mises à jour	Rapides et courtes initiées par les utilisateurs finaux	Emplois périodiques de lots de longue durée au cours de l'actualisation les données
Requêtes	Requêtes relativement standardisées et simple renvoyant relativement peu d'enregistrements	Requêtes souvent complexes impliquant agrégations
Vitesse de traitement	Généralement très rapide	Dépend de la quantité de données concernées; vitesse de la requête peut être améliorée par la création des indexes
Espace requis	Peut être relativement faible si les données historiques sont archivées	Plus grande en raison de l'existence de structures d'agrégation et de données historique; nécessite plusieurs indexes que OLTP
La conception de la base de données	Hautement normalisé avec de nombreuses tables	Typiquement dé-normalisés avec moins de tables; l'utilisation des schémas en étoile et / ou des schémas de flocon de neige
Sauvegarde et restauration	Sauvegarde religieusement; données opérationnelles est essentielle pour diriger l'entreprise, la perte de données est susceptible d'entraîner des pertes financières importantes et la responsabilité juridique	Au lieu de sauvegardes régulières, certains environnements peuvent envisager de simplement recharger les données OLTP comme une méthode de récupération

III. 2 – Organigramme opérationnelle OLTP et OLAP :

On résume l'opération du processus et montre la relation entre OLTP et OLAP avec cet organigramme. [4]



Source: www.data-e-education.com

Figure 2-2 : OLTP et OLAP en exécution.

III. 3 - Les Requêtes Multidimensionnelles :

III. 3.1 - Définition :

Mesure: Ca peut être un nombre ou une chaîne de caractères par exemple 12 pour un nombre et "Produit " pour une chaîne sinon on peut utiliser une fonction MDX pour le retourner.

Dimension : Les dimensions organisent les informations dans un cube, ce sont les axes d'analyses des indicateurs, les dimensions sont mutuellement indépendante dans mdx, une dimension peut avoir quelque membre organisé en hiérarchie.

Hiérarchie : Elle peut être spécifiée par son nom unique ou par commande MDX, les hiérarchies sont contenues dans des dimensions.

Niveau : Le niveau dans une hiérarchie de dimension peut être défini par son nom unique ou retourner par une requête MDX.

Membre : Pouvant être spécifié par son nom unique ou par requête MDX, Il appartient à une hiérarchie de dimension.

Tuple: C'est une collection ordonné d'un ou de plusieurs membres de différentes hiérarchie qui peut être défini par une liste de membres ou par une requête MDX.

Exemple de syntaxe d'une requête MDX :

```
SELECT
  {[Measures].[Store Sale]} ON COLUMNS,
  {[Date].[2016],[Date].[2015]} ON ROWS
FROM Sales
WHERE ([Store].[DZD].[AL])
```

III. 4 - Intégration des SIG avec OLAP :

Le concept OLAP répond aux besoins analytiques en s'appuyant sur une conception multidimensionnelle des bases de données, il s'oppose sur ce point aux SGBD dits transactionnels

Nature de l'information

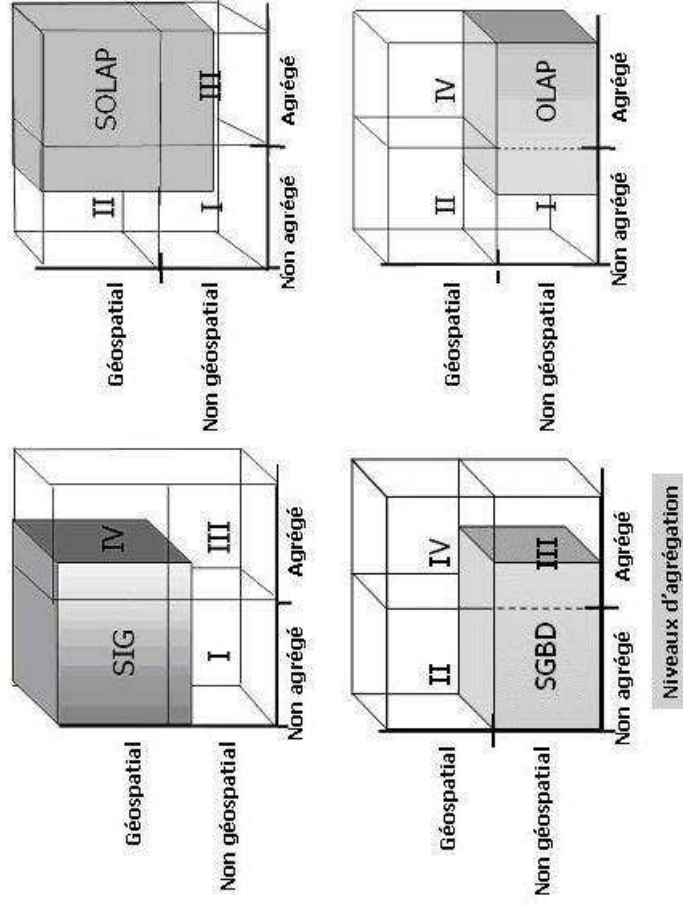


Figure 2-3 : Positionnement des outils vis à vis de l'information.

III. 4.1 OLAP Spatial :

SOLAP vient en quelque sorte combiné (Figure 2-4) la technologie des systèmes d'information géographique pour le traitement des données géographiques et leur visualisation cartographique avec les outils d'exploration et d'analyse OLAP. [8]

Aussi le principe de SOLAP est-il souvent résumé par l'équation symbolique $SIG + OLAP = SOLAP$

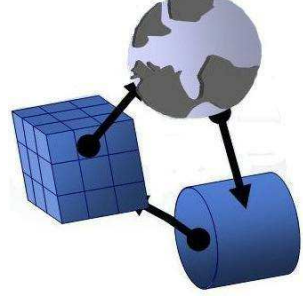


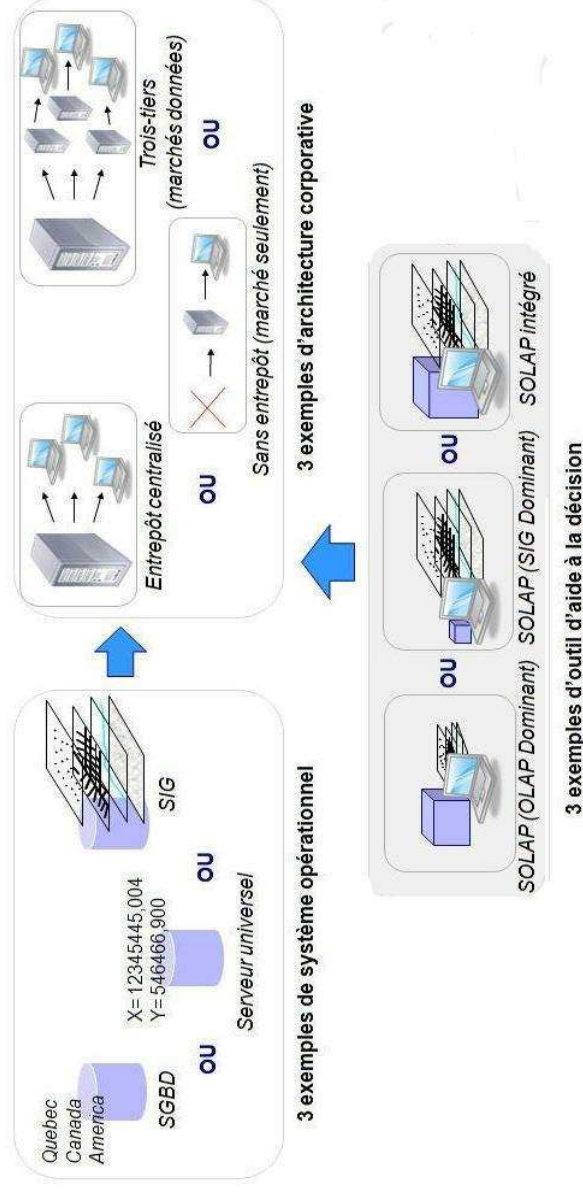
Figure 2-4 : Logo imaginaire d'un SOLAP.

III. 4.2. Architecture SOLAP :

A partir de l'intégration de ces dimensions spatiales dans la base multidimensionnelle, SOLAP va permettre deux types de mesures :

- Une **mesure spatiale** correspondant à un ensemble de coordonnées : intersection ou union d'une sélection de plusieurs formes géométriques parmi plusieurs dimensions choisies.
- Une **mesure topologique** : distance, surface ou "nombre de voisins". Les dimensions spatiales peuvent être organisées de façon à proposer les résultats de ces opérations topologiques suivant des hiérarchies : par exemple une opération d'adjacence pourra être décomposée en deux sous-niveaux, "adjacence par plusieurs points", "adjacence par un point".

Les outils SOLAP représentent donc une nouvelle famille d'outils, exploitant les cubes de données (détaillées et agrégées) construits à partir des entrepôts de données spatiales. L'architecture d'un système SOLAP s'articule de façon analogue à celle d'un système OLAP, suivant trois tiers, comme représentés sur la figure suivante.



III. 4.3. Topologies :

Les solutions SOLAP sont regroupées en trois grandes classes: OLAP dominant, SIG dominant et OLAP-SIG intégré.

- **SIG dominant :**

Comme le nom l'indique les fonctions SIG sont plus dominantes que les fonctions OLAP ce qui est bien avec cette technique c'est qu'elle supporte la cartographie thématique propre aux SIG, elle utilise des capacités d'analyse spatiale, utilisé souvent lorsque les cartes ont une importance principale et que les tableaux et graphiques reste simples. D'un autre coté elle nécessite des analyses complexes pour fournir une interface qui soit claire et net, en plus du temps de programmation que sa prend avec une interaction limité.

- **OLAP dominant :**

Est le contraire du SIG dominant il est utilisé pour des besoins d'exploration et de visualisations cartographiques simples même si ça nécessite des analyses complexes pour avoir une interface qui soit efficace pour l'utilisateur. En plus de la limitation d'interaction et de la dimension spatiales et d'analyses il s'intègre mal au processus de mise à jour des données spatiales.

- **SOLAP intégré :**

Intègre les deux systèmes dans un seul environnement ce qui donne un haut niveau de fonctionnalité pour les vues de données spatiales ou non spatiale. En plus d'une intégration sophistiquée et synchronisée des fonctions OLAP et SIG, et en plus d'avoir une interface qui soit efficace pour l'utilisateur comparé aux 2 autres systèmes, et des fonctions plus riches comme l'exploration synchronisée des cartes et tableaux, il s'intègre efficacement au processus de mise à jour des données spatiales, malheureusement il doit faire une analyses complexe pour pouvoir offrir une interface qui soit efficace.

IV- DATA MINING:

IV.1. Définition :

D'après Han et Kambre, « le terme Data Mining se réfère à l'extraction de connaissance à partir de grandes quantités de données (Big Data). Le Data Mining est un domaine récent qui se situe à l'intersection des statistiques, de l'apprentissage automatique et des bases de données. [6] »

IV. 2. Processus d'extraction de données :

Le datamining est une étape d'un processus plus complet qu'on appelle extraction de connaissance (ECD) ou Knowledge Discovery in Databases (KDD) voici une figure montrant se processus tiré de [1.1]

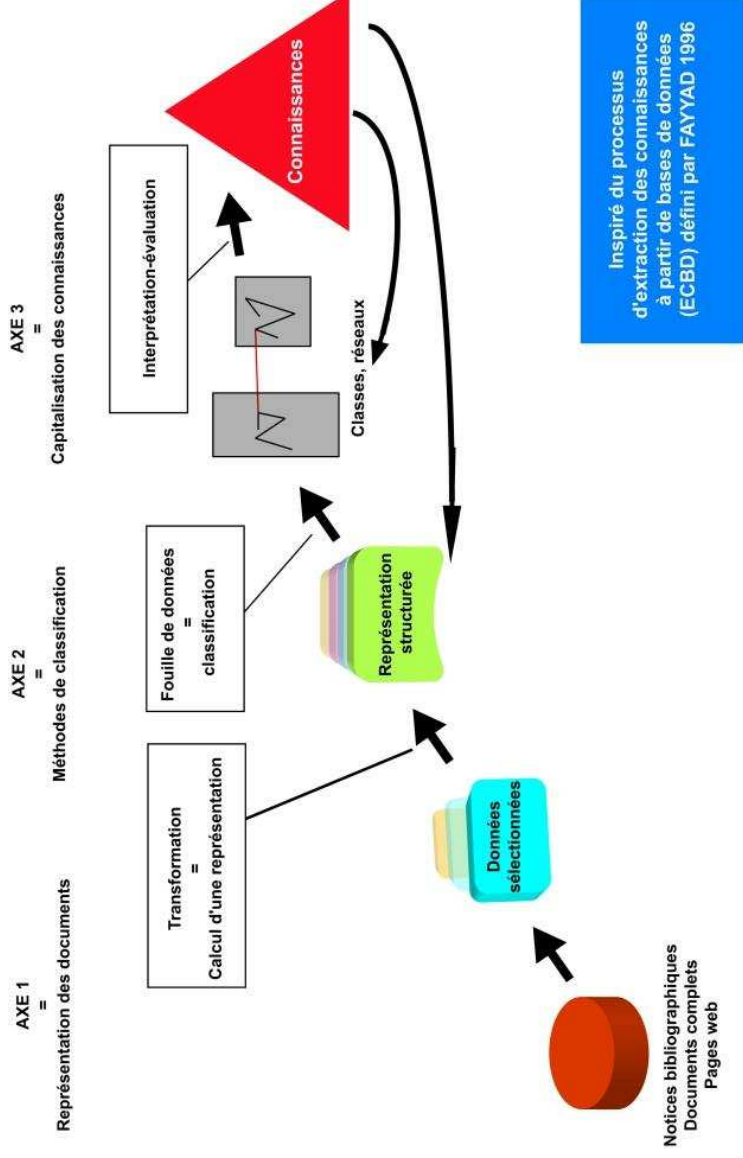


Figure 2-6 : Processus d'Extraction de Données.

formatées d'une façon particulière. Les méthodes de nettoyage sont inhérentes au domaine d'étude et la participation de l'utilisateur final est souvent cruciale.

IV. 2.2 – La phase de Data Mining :

La phase de datamining consiste en l'utilisation d'une ou de plusieurs techniques appliquées aux données, suivi de l'interprétation des résultats. Une phase de post-traitement peut être requise. Bien que la phase de forage des données est la phase la plus importante du processus de KD, rappelons que cette phase ne représente que 20% du temps de tout le processus [21]. Elle est aussi la phase la plus étudiée par les chercheurs parce que, d'une part, les autres phases sont très peu automatisables et d'autre part, elle est la raison d'être des autres étapes et constitue un enjeu majeur d'une nouvelle nature.

IV. 2.3 – La phase de post-traitement :

Les résultats en sortie de l'algorithme de forage de données peuvent être raffinés dans la phase de post-traitement. Celle-ci peut se résumer dans certains cas à l'interprétation des connaissances découvertes ou en des traitements supplémentaires sur les connaissances extraites. Il arrive que le post-traitement soit directement inclus dans la phase de datamining, mais il est préférable de séparer les deux phases. Le but ultime de cette phase est de permettre une meilleure compréhension et interprétation des connaissances générées par les algorithmes de forage de données. Le moyen le plus simple et le plus efficace, dans la plupart du temps, pour faire cela est le recours aux techniques de visualisation graphique. D'autres techniques ont été spécialement développées pour certains algorithmes de forage de données. Des propositions ont été faites pour, par exemple, déduire à partir des poids des interconnexions d'un réseau de neurones des règles d'induction [23], [24]. Aussi, le raffinement des règles associatives issues de la technique du même nom, en supprimant les règles rares, les règles triviales, les règles obsolètes et les règles particulières est une autre forme de post-traitement [25].

IV. 3. Méthodes de Data Mining :

Les modèles de prévision sont généralement développés en utilisant l'apprentissage dit **supervisé**. Cela signifie que les modèles sont développés en utilisant des informations anciennes ou nouvelles comprenant à l'entrée des variables (explicatives) et des résultats (ou cible) variables, et le développement est basé sur un échantillon de formation de données. Dans ce cas, nous

IV. 3.1. Méthodes descriptives :

La description synthétique d'un ensemble d'objets : Qui peut être basé soit sur les statistiques élémentaires ou les statistiques multidimensionnelles et ça dépendant du nombre de variables (s'il est ou non supérieur à 3). Une autre approche c'est l'utilisation de la généralisation conceptuelle en simplifiant les données et cela en réduisant les détails sémantiques. [13]

La recherche de dépendance entre caractéristiques d'objets : Elle permet de détecter si deux variables sont liées soit par l'analyse factorielle des correspondances (AFC), ou on applique l'analyse de correspondance (entre les modalités), sinon récemment une nouvelle technique découverte appliquée dans le domaine commercial, la recherche de règle d'association. [13]

La classification automatique d'objets (Clustering) : Consiste à regrouper les données similaires et à séparer les données éloignées, avec une méthode d'apprentissage non supervisée à partir de données non structurées on décrit une partition de ses données dans des classes. [13]

Détection de tendances (Trends) et de déviations : En utilisant des mesures tel que la médiane on décrit les tendances tout en effaçant l'influence des données extrêmes ou atypique. La déviation peut être détectée avec des tests statistiques sur les écarts. [13]

IV. 3.2. Méthodes prédictives :

Avec l'aide d'anciennes données on arrive à en sortir de nouvelles et pour ça il faut expliquer les données qu'on va utiliser en plus des variables. 2 principales fonctions sont utilisées :

Recherche de règles de classement d'objets : Avec l'aide d'exemples et d'un attribut à prédire ce qui permet une description en intention de classer les données qui vont être déduites. Le résultat peut être un arbre de décision, des règles ou un réseau de neurones. Elle peut être utilisée par exemple dans les prêts bancaires pour déduire les clients qui en bénéficient. [13]

Régression : Grâce à une fonction mathématique de type $y = a_1x_1 + a_2x_2 + \dots + a_px_n + r$ on peut prédire une variable, avec r le résidu et y la variable à rechercher. Elle peut être utilisée pour

Data Mining Techniques

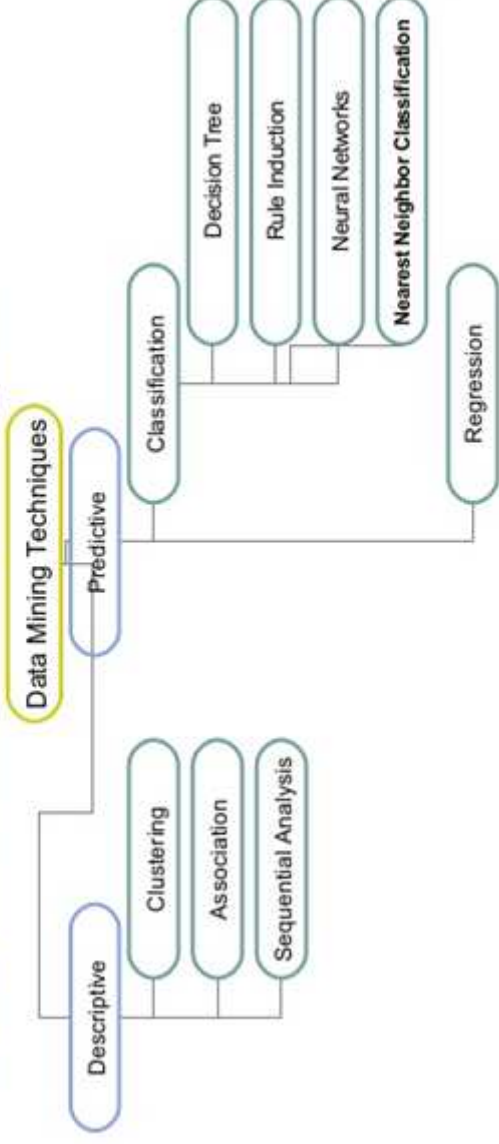


Figure 2-7 : Techniques du datamining classique.

IV. 4. Data Mining Spatial:

IV.4.1-Définition:

D'après [13] « Le Data Mining spatial est défini comme l'extraction de connaissances implicites, de relations spatiales ou d'autres propriétés non explicitement stockées dans la base de données spatiales. Ses avantages sont, d'une part, son aspect exploratoire car, contrairement à l'analyse classique, il génère des hypothèses puis les valide et, d'autre part, il permet l'intégration complète de l'information sur la localisation spatiale et des liens de voisinage. »

IV. 4.2 – Les techniques de data mining spatial:

Les méthodes de data mining spatial passe par 2 phase une phase exploratoire et une phase décisionnel dans les paragraphes suivant on va expliquer les étapes de chaque phases.

IV.4.2.1.1.1-Type de généralisation:

On a deux types de généralisation :

-Généralisation à dominante spatiale : Elle consiste à utiliser une hiérarchie spatiale déjà existante en plus des hiérarchies thématique avec une extraction des attributs qui se fait avec simplification et comptage. [29]

-Généralisation à dominante non spatiale : au contraire de celle citée plus haut elle n'utilise pas de hiérarchie spatiale, elle engendre des localisations moins détaillées par fusion d'objets. Son principe consiste à utiliser une induction orientée attribut avec des hiérarchies thématiques, ce qui nous fait sortir comme résultats des valeurs d'attributs homogènes (doublets) pour plusieurs objets qui sont à la fin fusionnés.[29]

IV.4.2.1.2-Clustering :

Contrairement au clustering pour les fouilles de données classiques le clustering pour les fouilles de données spatiales se base sur une mesure de similarité d'objets localisés suivant leur distance métrique. On essaie pas de former des classes mais de savoir où se trouve la concentration la plus élevée d'un événement par exemple détecter ou il y a le plus de taux de criminalité dans une zone.[29]

IV.4.2.1.3-Analyse globale par lissage :

Elle consiste à remplacer chaque valeur de la matrice de voisinage par le barycentre de ses voisins, ensuite une analyse factorielle est appliquée sur le tableau modifié (tableau lissé). [29]

IV.4.2.1.4-Analyse locale :

Son but est de faire ressortir les particularités pour que la donnée atypique soit mise en évidence qui nous permet de trouver des corrélations qui ne sont pas évidentes quand on les cherche dans un niveau plus général mais qui le sont dans un niveau plus détaillé. Son principe est le même que l'analyse globale sauf que au lieu d'avoir un tableau lissé on a un tableau contrasté ou une analyse factorielle sera appliquée. [29]

IV.4.3-Analyse explicative :

Consistant à voir l'analyste intervenir après découverte de clusters ou de valeurs atypiques par rapport à une tendance, qui focalise son analyse sur un sous-ensemble d'objets, un partie de variables ou une zone géographique. Qui ensuite est analysé pour démontrer sa particularité par des liens avec certaines valeurs ou par des règles caractéristiques. On va voir quelque méthode

IV.4.3.2-Règles d'association :

Elle se résume à trouver des règles d'association entre les propriétés des objets et celle de leur voisins. [29]

IV.4.3.3-Classification :

Qui consiste à classer un ensemble d'objets ayant des propriétés utilisant une partie des attributs communes elle est réalisée par apprentissage supervisé à partir de classe fournie partiellement par extension qui veut dire un échantillon de la base de données qui nous donne comme résultat une description qui permet de classer les données à venir. [29]

IV.4.4-Technique utiliser :

Dans notre cas on a utilisé les arbres de décision qui entre dans la catégorie des algorithmes de classification supervisé qu'on a vu dans les paragraphes d'avant voici une illustration du résultat obtenue

La **Figure 2-11** ci-dessous, illustre un arbre de décision d'une table relationnelle de notre base de données.

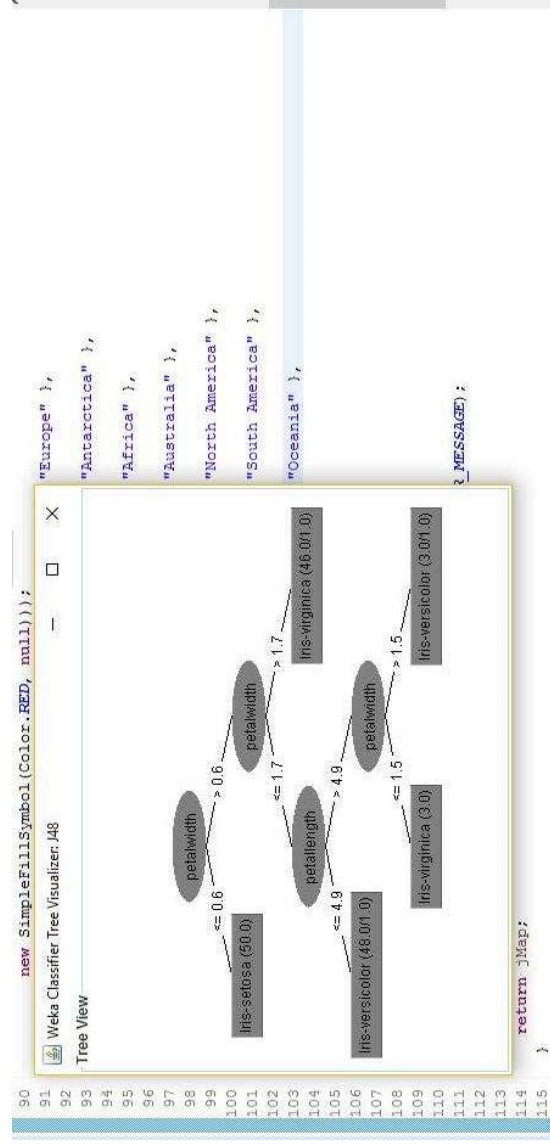


Figure 2-11 : Exemple d'arbre de décision.

Conclusion :

Dans ce chapitre, nous avons commencé par découvrir l'importance des entrepôts de données dans les solutions décisionnelles. Nous avons vu comment un entrepôt est structuré et que se passe-il dans un système décisionnel pendant la phase d'exploration des données.

L'équation SIG + OLAP = SOLAP a été discutée, nous avons aussi parlé des SIGs et des données géographiques, et l'architecture SOLAP a été expliquée brièvement.

Nous avons terminé ce chapitre avec quelques exemples de méthodes du Data Mining tel que le statistiques, logistiques, groupage, segmentation (i.e. arbre de décision)...etc. Et le processus général de cette opération a été décrit.

Chapitre III

Réalisation du projet

Introduction :	26
V. Réalisation du projet :	26
V. 1. Les outils de réalisation :	26
V. 2. Source des données :	27
V. 3. Préparation de base de données :	27
V. 4. Résultat de l'application :	36
Conclusion :	38

Introduction :

La réalisation d'un projet SIG géo-décisionnel pour un domaine assez complexe comme le domaine urbain nécessite une bonne démarche à suivre, et une méthodologie de travail bien précise. Dans ce chapitre on va expliquer la méthode qu'on va suivre pour réaliser notre application, ainsi la description de la zone de notre étude et de proposer un modèle décisionnel pour notre projet.

V. Réalisation du projet :

V. 1. Les outils de réalisation :

Notre solution est constituée de deux parties, d'une part appliquer les techniques SOLAP en combinant les SIG et l'OLAP pour explorer une base de données multidimensionnelle dédiée au suivi des chantiers d'urbanisme. Et d'autre part utiliser les techniques de data mining spatial pour découvrir des connaissances au sein de cette base de données.

Pour commencer à modéliser notre projet nous avons utilisé les outils suivants :

- **Langage Java** : est un langage de programmation et une plate-forme informatique qui ont été créés par Sun Microsystems en 1995. Beaucoup d'applications et de sites Web ne fonctionnent pas si **Java** n'est pas installé et leur nombre ne cesse de croître chaque jour. [27]
- **IDE Eclipse** : Eclipse est un projet libre et open source, décliné et organisé en un ensemble de sous-projets de développements logiciels, de la Fondation Eclipse, visant à développer un environnement de production de logiciels libres qui soit extensible, universel et polyvalent, en s'appuyant principalement sur Java.[28]
- **OLAP4j** : L'API, connu sous le nom olap4j, est l'équivalent OLAP de Java Database Connectivity (JDBC) pour les données relationnelles. Plus précisément, olap4j étend classes de base de JDBC spécifications 3 et 4 afin d'apporter des sources de données OLAP pour la plate-forme Java. Les connexions peuvent être obtenues par le gestionnaire de connexion JDBC.

L'API utilise des instructions, qui sont fournis par les connexions. Les requêtes,

- **ArcGIS SDK pour Java** : Le SDK ArcGIS Runtime pour Java fournit un ensemble d'outils de développement qui permettent de créer des applications ciblées de carte 2D pour répartition sur les plates-formes Windows et Linux. Le SDK (Software Development Kit) permet d'intégrer des cartes, localisateurs, et des modèles de géo-traitement en ligne et locales pour créer des applications riches et des SIG a haute fonctionnalités s qui tirent pleinement parti de la plate-forme Java Desktop.
- **Microsoft SQL Server 2014** : est un produit composé d'une suite de logiciels au même temps, en vus générale est un système de gestion de base de données relationnelles (SGBDR) conçu pour l'environnement de l'entreprise avec **In-Memory OLTP, Columnstore indexes** et **Buffer pool extension** intégrées.

V. 2. Source des données :

La mise en œuvre de notre projet, nécessite en premier lieu la détermination de la source de donnée, qui va être utilisé pour la construction de notre entrepôt, pour cela nous avons suivis le processus ETL et nous nous somme basé sur des données vectoriel et tabulaire, contenu dans le projet réalisé dans ArcGIS pour le suivi des chantiers du projet urbain, à la fin nous avons obtenu des fichier Excel extrait de la géodatabase, qui seront notre source de données pour la suite de la création du projet.

V. 3. Préparation de base de données :

Après avoir défini la source des données nous avons créé notre base de données, pour cela nous avons utilisé Microsoft SQL Server Management Studio qui se base sur le modèle multidimensionnelle en flocon conçu pour la réalisation du projet du suivi des chantiers urbains.

ID	OBJECTID	VILAYA	NOM	CODE	AREA	SHARP	Length
1	1	MOSTAGAN	el thossaine	2732	7683	0.376593	0.000000
2	2	MOSTAGAN	khayma	2733	6272	0.351784	0.000000
3	3	MOSTAGAN	khuda	2704	5420	0.351784	0.000000
4	4	MOSTAGAN	an noussy	2705	3188	0.252718	0.000000
5	5	MOSTAGAN	mazragan	2727	1887	0.194838	0.000000
6	6	MOSTAGAN	mazragan	2723	5733	0.300616	0.000000
7	7	MOSTAGAN	bougouret	2719	8537	0.557782	0.000000
8	8	MOSTAGAN	mebra	2722	4395	0.318543	0.000000
9	9	MOSTAGAN	an sai cheff	2721	2683	0.197327	0.000000
10	10	MOSTAGAN	taouira	2731	2844	0.224646	0.000000
11	11	MOSTAGAN	taouira	2731	2844	0.224646	0.000000

le déroulement du suivi des chantiers, et extraire nos données qui par la suite vont être intégré dans un entrepôt de données. Le résultat obtenu sous format DBase sera transformé en des résultats au format Excel. On est donc passé de la collection des données des différentes sources de données (Tabulaire, Vectoriel), pour ensuite procédé à une transformation de ces données pour les préparé et les traiter. Ce qui va nous permettre de les manipulés et les intégrés dans notre base de données, ceci concerne la phase de l' ETL qui est un processus d'inspection, de nettoyage, et de transformation des données dans le but de mettre en évidence l'information utile, qui permet une prise de décision, cette phase reste importante pour la conception et la réalisation d'un cube de données.

Par la suite nous allons voir comment on a créé notre base de données en se basent sur le modèle multidimensionnel, présenté dans le chapitre précédent, nous allons décrire les étapes et les procédures qu'on a appliqué pour la réalisation de cette tâche.

V. 3.1 – Modèle Conceptuel de Données:

Dans la **Figure 3-2** suivante une illustration de la base de données que nous avons utilisé dans notre projet.

A partir de ce modèle, nous avons établi un modèle pour le suivi qui contient les tableaux suivants :

POS : (ID_POS, Nom_POS, Date_Aprobation).

Zone : (ID_Zone, Superficie_Zone, CES_G, COS_G).

Ilot : (ID_Ilots, Superficie_Ilots).

Parcelle : (ID_Parcelle, Superficie_Parcelle, CES, COS).

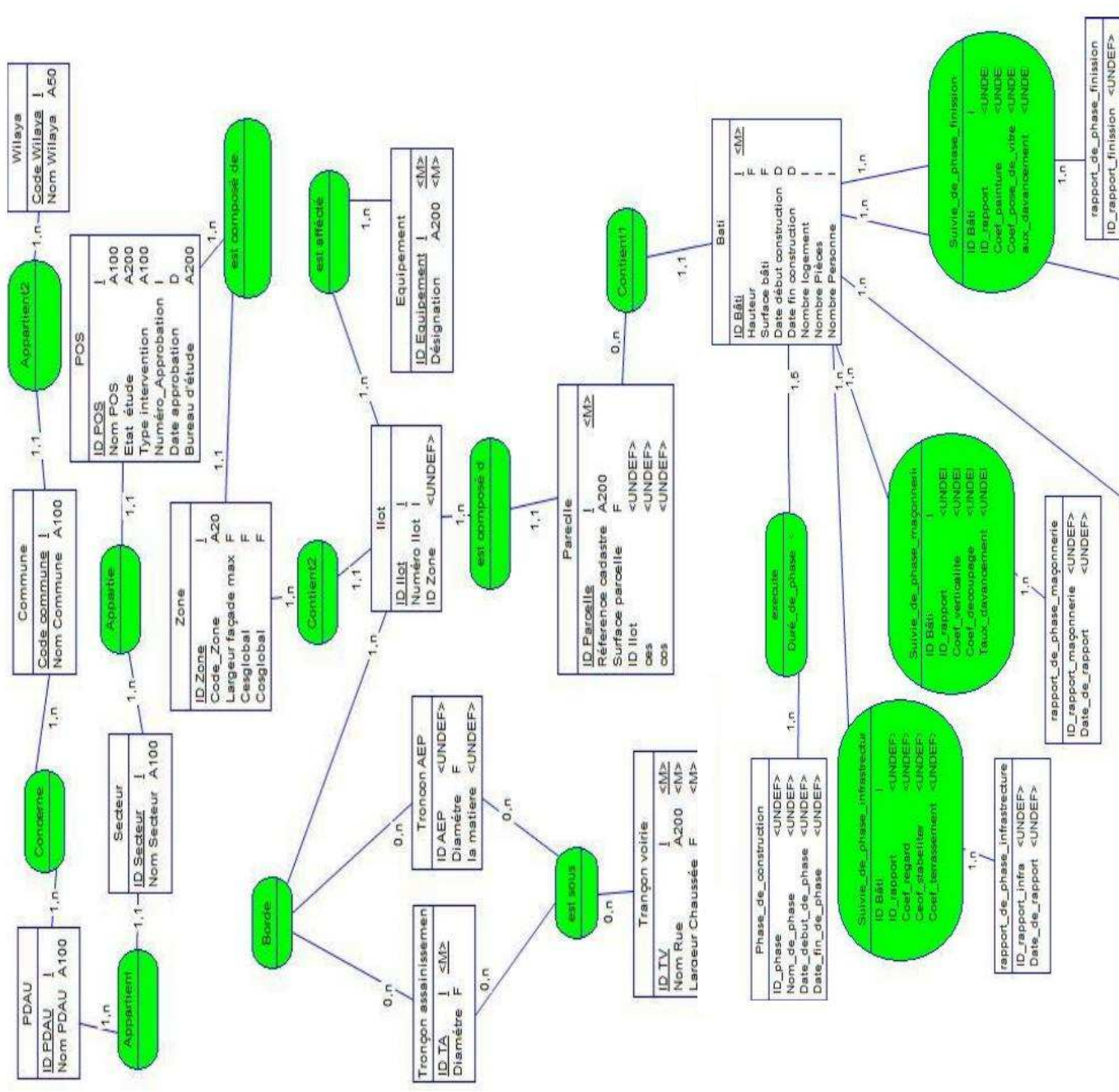
Bati : (ID_Bati, Hauteur_Bati, Surface_Bati, Date_Debut_Construction,

Date_Fin_Construction, Nombre_Logement, Nombre_Pieces, Nom_Bati).

Rapport : (ID_Rapport, Date_Rapport).

Phase : (ID_Phase, Nom_Phase, Date_Debut, Date_Fin).

Ceci nous à permit de construire un modèle multidimensionnelle, en flocon qui contient les tableaux en dessous, **Figure 3-3**:



DIM_POS	
ID POS	<h:1>
Nom POS	
Etat Etude	
Type Intervention	
Numéro Approbation	
Date Approbation	
Espace	<Default>

Dim_Ilot	
ID Ilot	
Nom Ilot	
ID Zone	<h:1>
ID Zone	<Default>

Dim_Parcelle	
ID Parcelle	
Reference cadastre	
Nature Juridique	
Nature d'Occupation	
Taux Consommation	
Numero PC	
Surface Parcelle	
Statut	
ID Ilot	<h:1>
ID Ilot	<Default>

Dim_Zone	
ID Zone	
Code Zone	
Affectation	
Longueur Facade Max	
Longueur Facade Min	
CES Max	
CES Min	
COS Max	
COS Min	
ID POS	<h:1>
ID POS	<Default>

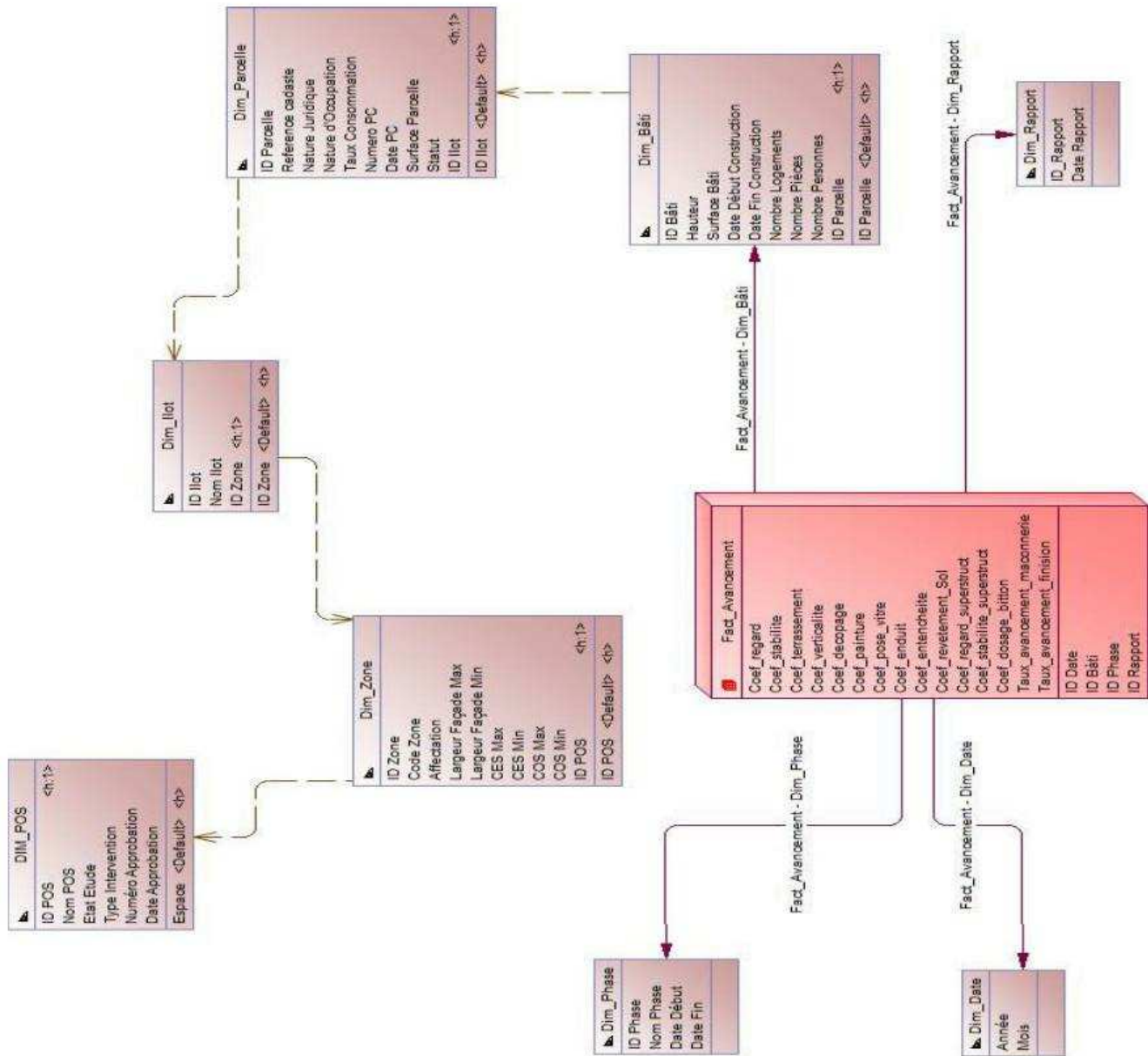
Dim_Phase	
ID Phase	
Nom Phase	
Date Début	
Date Fin	

Dim_Date	
Année	
Mois	

Dim_Bâti	
ID Bâti	
Hauteur	
Surface Bâti	
Date Fin Construction	
Nombre Logements	
Nombre Pièces	
Nombre Personnes	
ID Parcelle	<h:1>
ID Parcelle	<Default>

Dim_Rapport	
ID_Rapport	
Date Rapport	

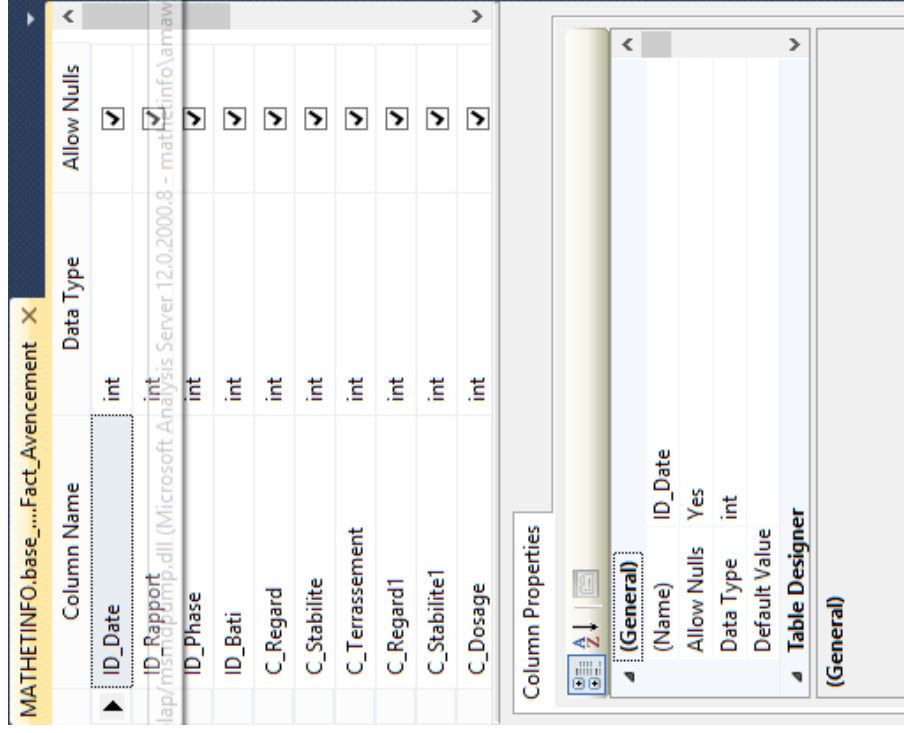
Fac_Avancement	
Coef_regard	
Coef_stabilité	
Coef_terrassement	
Coef_verticlite	
Coef_decoupage	
Coef_painting	
Coef_pose_vitre	
Coef_enduit	
Coef_entretancheite	
Coef_revetement_Sol	
Coef_regard_superstruct	
Coef_stabilité_superstruct	
Coef_dossage_biton	
Taux_avancement_maconnerie	
Taux_avancement_finition	
ID Date	
ID Bâti	
ID Phase	
ID Rapport	



Une fois notre modèle construit, nous sommes passé par une sélection, un nettoyage, et un traitement de ces données, nous avons procédé par la suite à la création de notre entrepôt de données, l'intégration de nos données sélectionné, nettoyé, et traité, pour cela nous avons utilisé des outils permettent de réaliser cette tâche (Microsoft SQL Server 2014).

V. 3.2 – Création des tables :

Pour cela, nous nous sommes basé sur notre modèle multidimensionnel élaboré à partir du MCD conçu pour la réalisation du projet du suivi urbain, qui est un modèle en flocon modélisant les dimensions et la tables des faits qui va contenir les mesures nécessaires pour la bonne modélisation de notre base, et pour faire les bonnes analyses. La **Figure 3-4** illustre la méthode sous management studio.



V. 3.3 – Création des relations :

Cette partie consiste à créer les relations qui existe entre les tables de notre base de données, pour la réaliser il suffit d'un clic droit sur Database Diagrams, choisir New DataBase Diagrams, une fenêtre va s'ouvrir pour choisir parmi les tables de la base de données, les quelles seront concerné per la création des relations, comme il est illustré dans les figures suivantes :

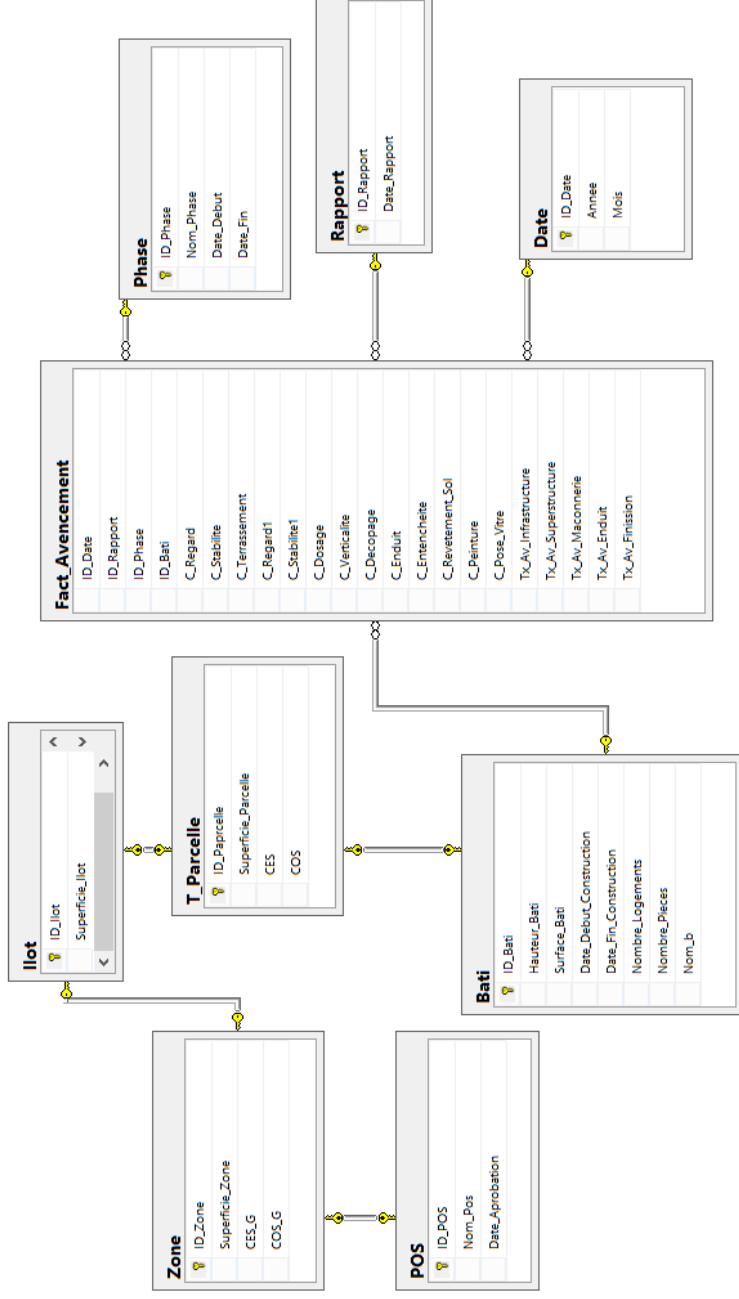


Figure 3-5 : Diagramme des relations.

V. 3.4 – Création du cube de données :

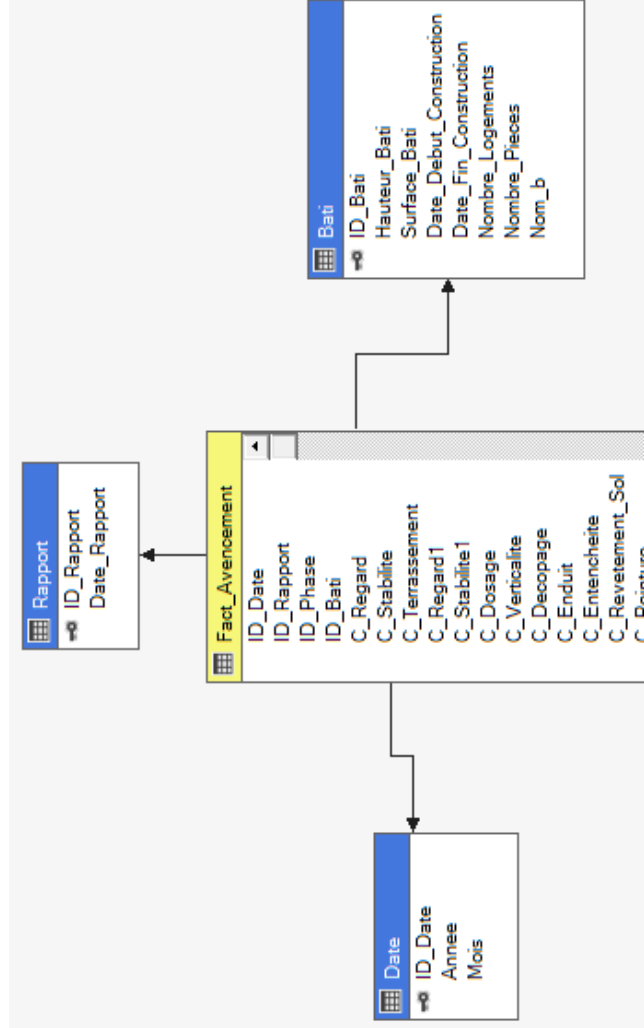
Pour la modélisation de notre cube nous allons utiliser comme outil Analysis Services, il va contenir les dimensions et la table de fait crée précédemment, on va introduire les 4 dimensions suivantes : *Bâti*, *Phase*, *Rapport*, *Date*, qui sont en relation avec la table de fait, ce qui permis de faire une analyse sur le taux d'avancement en fonction de la date, du rapport élaboré de la phase de construction et du bâti choisi, de ce point on commencent à modéliser

offre la possibilité de choisir les dimensions et les tables de fait qui vont être utilisés pour la création du cube, dans notre cas on a choisi les tables de notre base de données.

La troisième étape consiste à la création de des dimensions et leurs hiérarchies, comme expliqué dans les étapes précédentes on a 4 dimensions qu'on va créer dans notre projet, on a choisi pour chaque dimension les attributs nécessaires à la création du cube, et qui vont être utilisés pour la création de la hiérarchie de la dimension, pour faciliter l'analyse.

Une fois qu'on a fini avec la création des dimensions et leur hiérarchie, nous sommes passés à la création du cube, ou on va spécifier la table de fait et les dimensions qui vont contenir les attributs et les mesures qu'on utilisera pour faire notre analyse selon plusieurs dimensions, ce qui nécessite une bonne sélection de ces éléments.

Notre cube sera modélisé sur un modèle en étoile, contenant les 4 dimensions (*Bâti*, *Date*, *Phase*, *Rapport*), qui explique le processus suivi pour la modélisation ou un rapport est généré périodiquement concernant la phase de construction du bâti. La figure suivante illustre le résultat obtenu de la création de notre cube en présentant les dimensions utilisées, la table de fait, les mesures et le modèle obtenu.



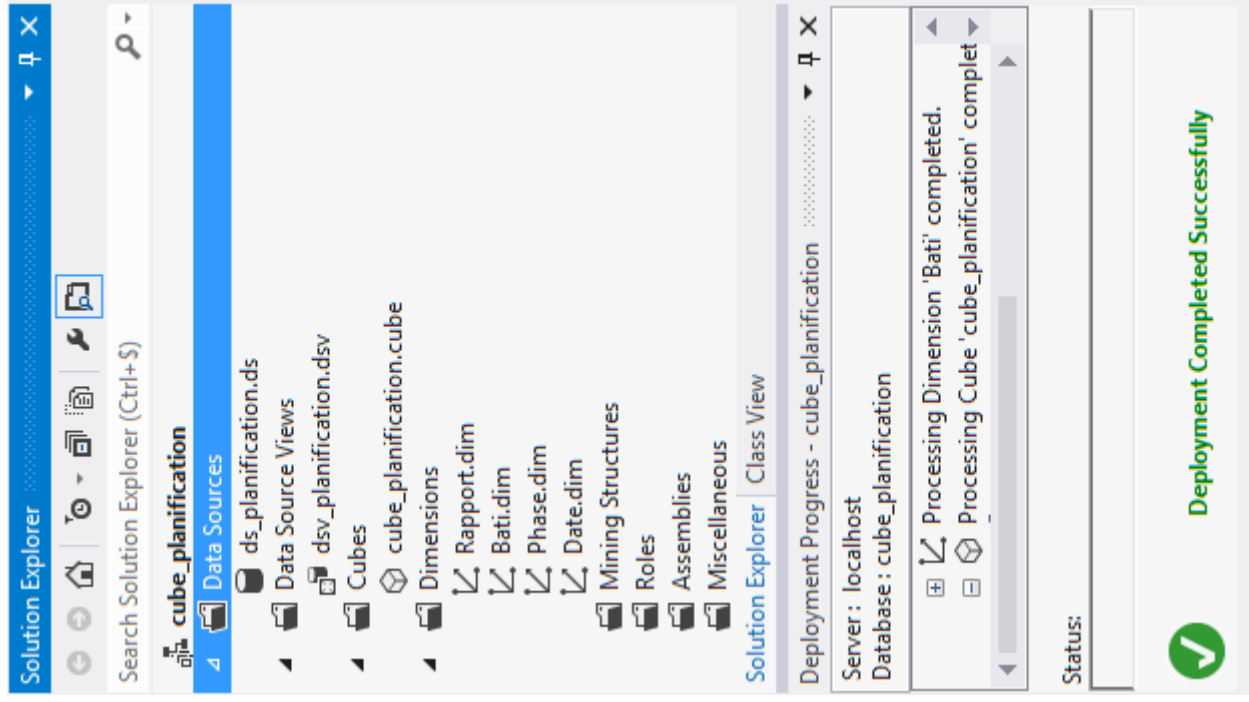


Figure 3-8:
dimensions

Affichage des
créent.

On a créé notre cube en suivant les étapes décrites précédemment, ceci nous permettra de faire des analyses sur le cube en fonction de plusieurs dimensions, nous avons fait quelques

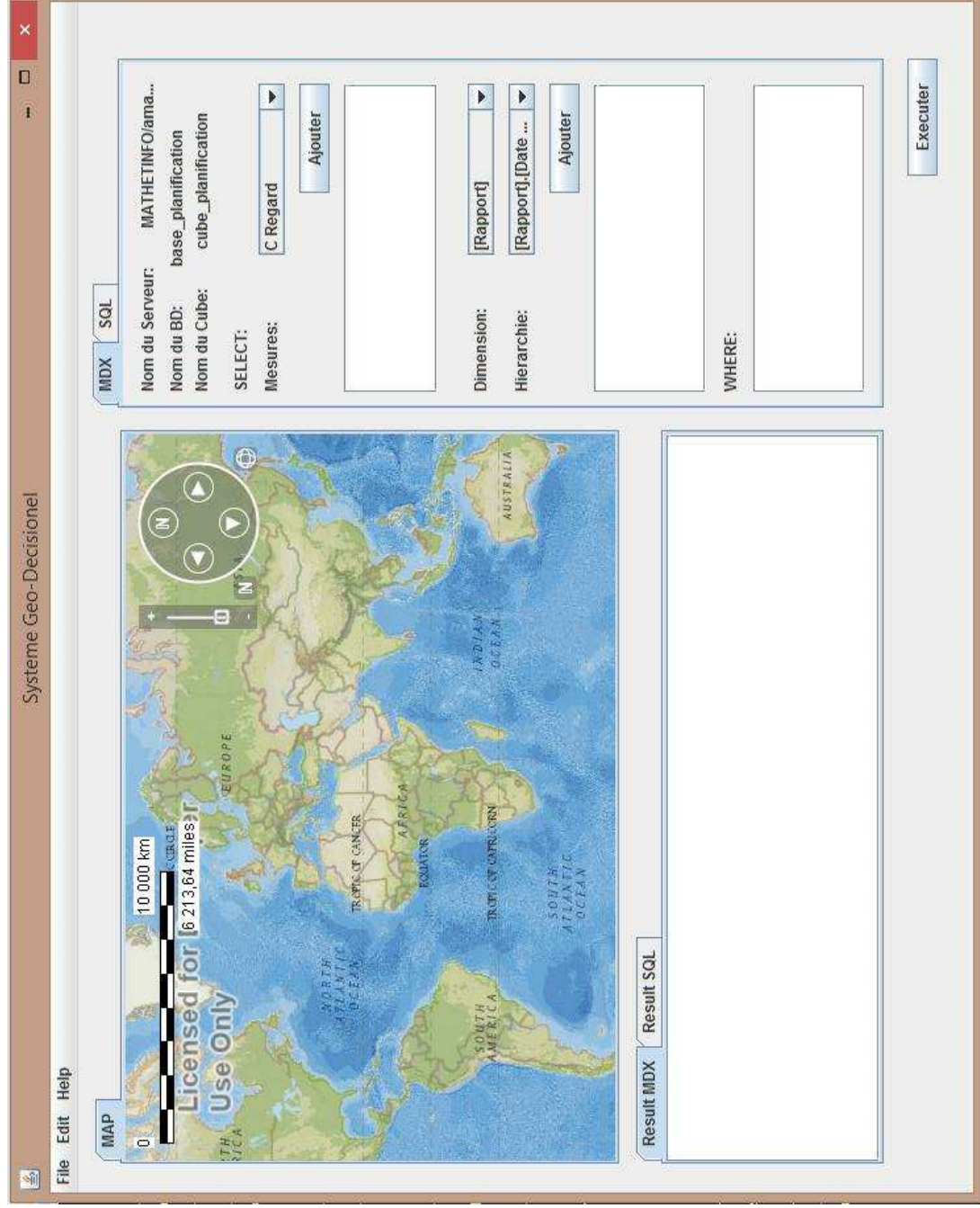
La figure suivante montre le résultat obtenu :

Dimension <Select dimension>	Hierarchy	Operator	Filter Expression	Param...				
Date Debut	Nom B	ID Phase	C Decoupage	C Dosage	C Enduit	C Ententeheite	C Pose Vitre	C Regard
2013-01-0...	Bati_R+5	1	0	0	0	0	0	941
2013-01-0...	Bati_R+9	1	0	1	0	0	0	685
2013-01-0...	Bati_semi_collectif	1	0	0	0	0	0	96
2013-09-0...	Bati_R+5	2	0	1602	0	0	0	2076
2013-09-0...	Bati_R+9	2	0	1229	0	0	0	1560
2013-09-0...	Bati_semi_collectif	2	0	179	0	0	0	228
2014-09-0...	Bati_R+5	3	2422	2595	0	0	0	2595
2014-09-0...	Bati_R+9	3	1818	1950	0	0	0	1950
2014-09-0...	Bati_semi_collectif	3	253	270	0	0	0	285
2016-01-0...	Bati_R+5	4	1384	1384	633	447	0	1384
2016-01-0...	Bati_R+9	4	1040	1040	419	296	0	1040
2016-01-0...	Bati_semi_collectif	4	144	144	50	38	0	152
2016-08-0...	Bati_R+5	5	865	865	525	525	124	865
2016-08-0...	Bati_R+9	5	649	649	360	360	82	649
2016-08-0...	Bati_semi_collectif	5	89	89	44	44	7	94

Figure 3-9 : Analyse sur le cube.

V. 4. Résultat de l'application :

- **Interface d'application** : Ceci est l'interface de notre application, elle est composée d'une partie cartographique et d'une partie de sélection où nous pouvons effectuer des requêtes multidimensionnelles sur notre entrepôt de données



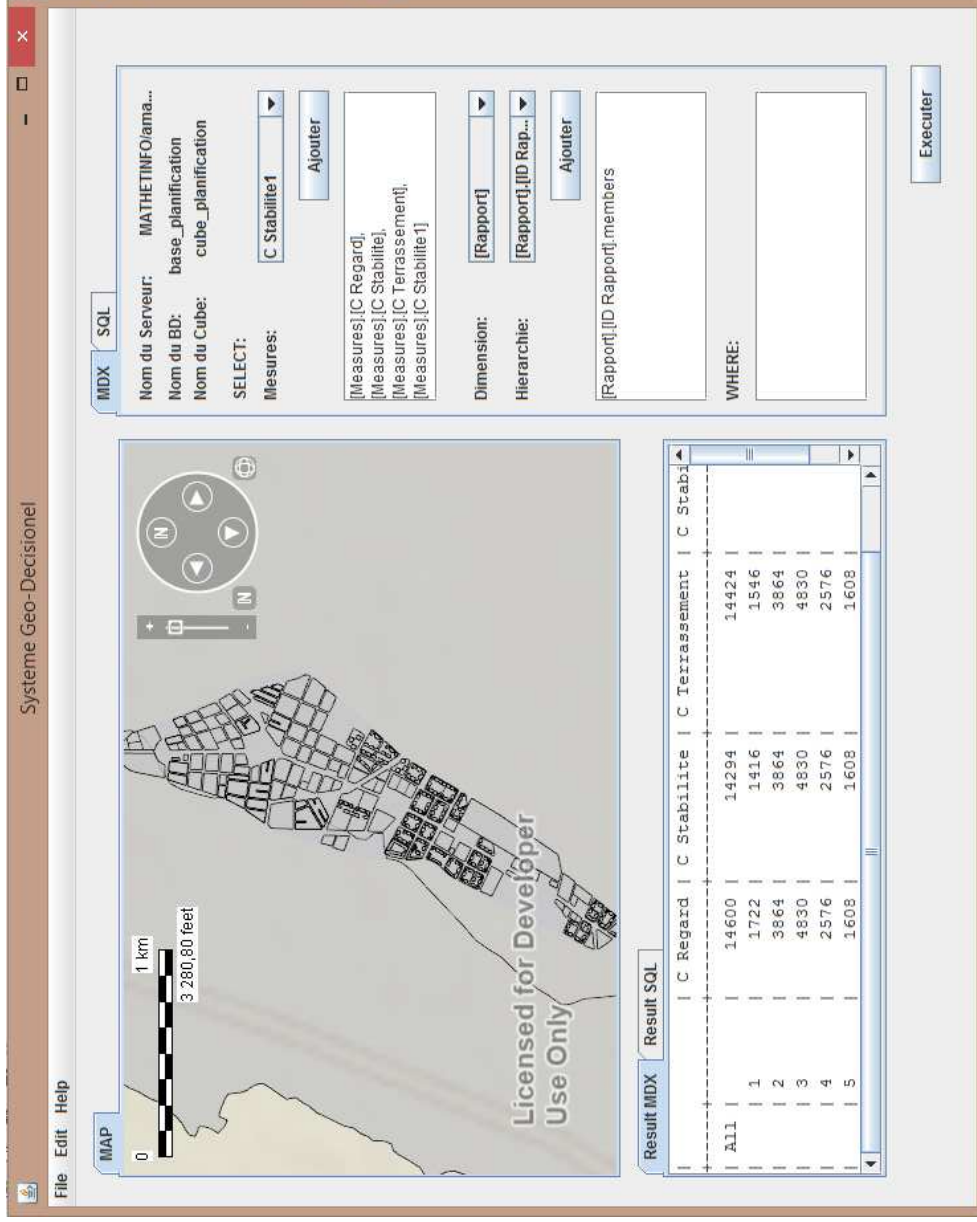


Figure 3-9.2 : Résultat de la requête MDX.

- **Utilisation du data mining sur les données spatiales :** Comme exemple illustratif, nous utilisons l'arbre de décision pour classifier les bâtiments achevés selon le nombre d'étages.

Conclusion :

Dans ce dernier chapitre nous avons vu les composants essentiels de notre application développée sous java et basée sur le concept SOLAP fondé sur l'intégration des outils d'OLAP réalisé sous SQL server 2014 et le SDK d'ArcGIS-java développé par ESRI. Nous avons aussi vu l'application du data mining spatial sur des données spatiales.

Précisément, dans ce chapitre, Nous avons découvert les capacités de SQL server 2014 en termes de création d'entrepôt de données et l'analyse des cubes. Nous avons donc réalisé une application géodécisionnelle complète pouvant combiner l'analyse spatiale et les outils de business intelligence.

Néanmoins, beaucoup de choses restent à faire, comme par exemple l'intégration d'autres techniques de data mining ou l'utilisation générique de notre application à d'autres thématiques.

CONCLUSION GENERALE

Les solutions géo-décisionnel permettent aux décideurs de compléter leurs traditionnels critères de réflexion stratégique par l'intégration et l'analyse des facteurs d'implantation, de proximité, de risques et de mobilité géographique afin de mieux servir les citoyens et les consommateurs.

Le géo-décisionnel réuni deux mondes qui jusqu'alors se côtoyaient peu, celui de l'informatique décisionnelle (aussi appelée Business Intelligence) et celui des Systèmes d'information Géographiques (SIG).

On parle de solutions géo-décisionnelles, qui s'avèrent être bien souvent les clés pour une meilleure efficacité opérationnelle et un vecteur de croissance pour l'organisation. Ces solutions s'articulent généralement autour des bases de données multidimensionnelles qui prennent en charge l'information spatiale sur ses deux modes : raster et vecteur.

Ses solutions, englobent d'une part les outils d'analyse spatiale issues des Systèmes d'information géographique, et d'autre part, les systèmes SOLAP et les techniques de data mining spatial issues du monde de la Business Intelligence et qui permettent de construire une découverte de connaissance spatiale sur des bases de données et des modèles géographiques qui supportent les spécificités de l'information spatiale. Nous avons dans ce rapport décrit les différents concepts liés au géo décisionnel, à savoir les SIG, les entrepôts de données et le data mining spatiale.

Notre objectif a été de mettre en œuvre un système géo décisionnel basé sur la technologie SOLAP intégré et les techniques de data mining spatiale.

Bibliographie :

- [1][2] J.-F. Desnos. (2013) Cours : Entrepôt de Données.
- [3] J.-F. Desnos. Entrepôt de Données 2^{ème} années régime spécial Spé. IHS Master ICA.
- [4] Rémy Choquet. (2008) Entreposage de Données. Processus ETL, Modélisation d'entrepôts, Restitution.
- [7] George Spofford, Sivakumar Harinath, Chris Webb, Dylan Hai Huang, Francesco Civardi. (2006). MDX-Solutions: With Microsoft SQL Server Analysis Services 2005 and Hyperion Essbase. ISBN: 978-0-471-74808-3 John Wiley & Sons, Inc.
- [8] Louis-Etienne Guimond, Marie-Josée Proulx and Dr Yvan Bédard. (2003) Liste détaillées des caractéristiques et fonctionnalités SOLAP.
- [9] Thierry Lallemand. (2008) Le Geodecisionnel: Le Sig Au Service Du Geodecisionnel. Mastère ASIG / Projet Bibliographique.
- [10] Sandro Saïta. (2007) Data Mining Des Données à la Connaissance. Article, EPFL/IMAC
- [11] Stéphane Tuffery. Cours de Data Mining.
- [13] Karine Zeitouni, Laurent Yeh. (1999) Le Data Mining Spatial et les bases de données spatiales. UNIVERSITE DE VERSAILLES -SAINT-QUENTIN
- [14][16] Christof Paar and Jan Pelzl. (2013) Understanding Cryptography. ISBN: 978-3642041006 Springer Berlin Heidelberg New York.
- [15] Mark Stamp. (2011). Information Security: Principals and Practice. ISBN: 978-0-470-62639-9 John Wiley & Sons, Inc.
- [16] Chuck Easttom, (2016). Modern Cryptography: Applied Mathematics for Encryption and Information Security. ISBN: 978-1-25-958809-9 McGraw-Hill Education.

- [21] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, (1996) "from Data Mining to Knowledge Discovery: An Overview", Advances in knowledge discovery and data mining, Usama M. Fayyad & al. eds, page 1-34.
- [22] Jennifer Widom, (November 1995) "Research Problems in Data Warehousing", Proc. 4th Int. Conf. On Information and Knowledge Management (CIKM).
- [23] Vitaly Schetinin, (1996) "Self-Organization of Neuron Collective of Optimal Complexity", Proceedings of Int. Symposium NOL TA'96, Kochi, Japan, page 245-248.
- [24] R. Andrews, J. Diederich, A. Tickle, (1995) "A Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks", Knowledge-Based Systems.
- [25] Nicolas Pasquier, (Janvier 2000) "Data mining: Algorithmes d'extraction et de réduction des règles d'association dans les bases de données" Thèse de doctorat, École Doctorale Sciences pour l'Ingénieur de Clermont-Ferrand, Université de Clermont-Ferrand II.
- [26] Georges Gardarin, (1999) " Internet/intranet et bases de données", Librairie Eyrolles.
- [29] Marie-Aude Aufaure I, Laurent Yeh2, Karine Zeitouni "FOUILLE DE DONNEES SPATIALES " Laboratoire d'Ingénierie des Systèmes d'Information - INSA de Lyon .
- [30] Ester M., Frommelt A., Kriegel H.-P., Sander J., Algorithms for Characterization and Trend Detection in Spatial Databases, Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, New York, NY, (1998).

Webographie:

- [1.0] - <http://bit.ly/1yLR8Bq>
 [1.1] - <http://recherche.inist.fr/spip.php?article6>