

Vers un entrepôt de données pour la gestion des risques naturels

Hicham Hajji*, Nourdine Badji**, Jean Pierre Asté**
* Université Claude Bernard, LIRIS, Lyon, France
hhajji@bat710.univ-lyon1.fr
**503 Rue du 8 Mai 1945 GIPEA, Caluire, France
gipea@gipea.fr

Résumé. Les entrepôts de données sont l'un des plus importants développements dans le domaine des systèmes d'informations. Ils permettent d'intégrer des données de plusieurs sources, souvent très volumineux, distribuées et hétérogènes. Dans cet article, nous examinons la possibilité d'utiliser la technique d'entrepôt de données dans la gestion des risques naturels. Nous présentons un modèle conceptuel pour l'entrepôt proposé, avec la présence de formats et types variés de données tel que des données géographiques et multimédia. Nous proposons également des opérations OLAP pour la navigation des informations stockées dans le cube de données.

1. Introduction

Les entrepôts de données sont l'un des plus importants développements dans le domaine des systèmes d'informations. Conceptuellement, un entrepôt de données est une base de données qui collecte et stocke des données à partir de plusieurs sources hétérogènes et lointaines. L'approche entrepôt de données présente quelques avantages sur l'approche traditionnelle [Theodoratos, D., et Sellis, T., 1999] :

- Les requêtes peuvent être répondues localement sans avoir à accéder les sources originales.
- Les traitements analytiques en ligne OLAP sont autant que possible séparés des traitements transactionnels en ligne OLTP.

Dans ce papier, nous proposons une architecture pour construire un entrepôt de données spatiales pour la gestion des risques naturels, et nous proposons un outil de requêtes pour assister à extraire les informations et effectuer des opérations analytiques sur un tel environnement. Pour motiver notre étude, nous examinons les exemples suivants :

Exemple 1 : l'évaluation des dommages par rapport à l'unité d'analyse, la parcelle. Un utilisateur peut souhaiter voir l'évaluation des dommages sur une carte par bâtiments, par parcelles, ou même souhaite de descendre ou remonter dynamiquement dans la hiérarchie des dimensions d'analyse pour explorer de nouvelles informations, telles que les zones à risques par rapport à un phénomène naturel particulier.

Exemple 2 : superposer plusieurs cartes thématiques : il existe souvent des cartes thématiques dans des bases de données spatiales, telle que les cartes d'altitudes, les cartes de populations, les cartes d'intensité des inondations, etc.

2. La gestion des risques naturels

En 1997, 530 catastrophes naturelles ont fait 22315 morts ou disparus. Ce nombre est inférieur à celui des années précédentes qui variant entre 580 e 600. Le coût total est estimé à

30 milliards de dollars en 1996 (contre 60 milliards de dollars en 1996), et les compagnies d'assurances ont déboursé 4,5 milliards de dollars.

En France par exemple, la plupart des informations liées à la gestion des risques naturels sont fournies par différents organismes. Dans cet article, nous nous intéressons aux bases de données externes telles que les données de l'INSEE (INstitut des Statistiques et des Etudes Economiques), la base de donnée de DGI (la direction générale des impôts), la base de données de la SIRENE (base de données sur les Entreprises et les Etablissements), des données sur les risques naturels, et les données géographiques fournies par le cadastre (fichiers des parcelles et des bâtiments). Ces sources offrent l'essentiel des données pour mener une évaluation assez détaillée de la vulnérabilité et des dommages.

Dans cet article, nous proposons une architecture d'entrepôts de données pour assister les décideurs dans leurs efforts de prévention et de mitigation dans le contexte des risques naturels.

Organisation de l'article: l'article est organisé comme suit. Dans la section suivante, nous présentons rapidement les concepts clés dans le domaine des entrepôts de données, nous terminons cette section en citant les développements récents dans les entrepôts de données spatiales. Dans la section 3, nous présenterons les modèles des données sources, et nous proposons un modèle pour l'entrepôt de données. Ensuite, nous présentons quelques opérations OLAP intéressantes pour l'analyse des données et l'exploration du cube de données. Nous concluons notre travail dans la section 4.

3. Préliminaires

3.1. Recherches récentes dans les entrepôts de données spatiales

A notre connaissance, il n'y a que peu de travaux qui ont traités les entrepôts de données spatiales EDS. Nous allons brièvement décrire les concepts généraux des entrepôts de données spatiaux en incluant les aspects d'intégration; nous allons aussi mentionner les aspects d'implémentation des outils SLOAP (Olap Spatiaux). Pour plus d'informations sur la fouille de données spatiales, nous référons le lecteur à l'article [Han *et al.*, 1997].

Dans la communauté de la recherche, quand le modèle multidimensionnel a été proposé pour les EDS, généralement la discussion sur les hiérarchies et les agrégations ont eu lieu. Puisque les EDS peuvent inclure des données spatiales et non spatiales qui peuvent nécessiter d'être agrégées, certains auteurs distinguent trois situations générales [Han *et al.*, 1998][Stefanovic *et al.*, 2000][Bédard *et al.*, 2001]: Une dimension spatiale non géométrique, Une dimension spatiale mixte, Une dimension spatiale géométrique.

L'idée d'appliquer des concepts OLAP dans les données spatiales est représentée dans le prototype « Map Cube », c'est-à-dire un cube multidimensionnel qui manipule des cartes [Shekhar *et al.*, 2001]. Il inclut des dimensions spatiales, des mesures numériques, et des agrégations spatiales des polygones continues en utilisant des opérations OLAP spatiales.

4. Un modèle multidimensionnel pour l'analyse des données spatiales dans la gestion des risques naturels

Suivant la tendance du développement des techniques des entrepôts de données et de fouilles de données, nous proposons de construire un entrepôt de données spatiales pour faciliter l'analyse en ligne des données spatiales pour la gestion des risques naturels.

Une meilleure compréhension des risques naturels nécessite plusieurs sources d'informations, telles que celles liées à la population, les données topographiques, les

données hydrologiques (dans le cas du risque d'inondation par exemple), les données historiques sur les risques naturels, les données législatives, etc.

4.1. Les sources de données

Selon leur nature, les données peuvent tomber dans l'une des catégories suivantes: données spatiales, non-spatiales, et les données sur les risques.

4.1.1. Données non-spatiales

Elles sont principalement composées de deux sources: SIRENE et DGI:

Des données pour la description des éléments exposés comme la population et les données sur les bâtiments, Ces données sont utilisées principalement pour la description des éléments présents sur le territoire. Quelques unes de ces informations: la surface du bâtiment, la date de construction, le matériel de construction, etc. L'autre partie des données est celle pour décrire les activités des entreprises. Dans notre cas, elles sont représentées par les données de la SIRENE, qui donnent des informations géographiques et économiques sur les entreprises.

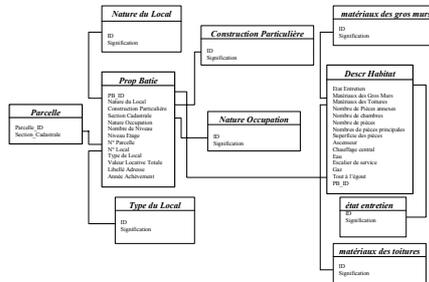


FIG. 1- un exemple des données non spatiales : Source DGI

4.1.2. Les données spatiales

Les données spatiales utilisées dans l'entrepôt de données sont principalement composées des éléments spatiaux de localisation comme les communes, les sections cadastrales, les parcelles, Les bâtiments qui sont décrits par des polygones (voir figure 2). Les autres composantes des données spatiales sont les cartes géologiques et les ortho-photos, et le modèle numérique du terrain MNT.

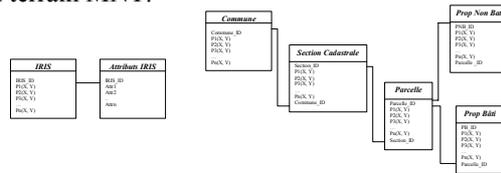


FIG. 2- les données spatiales dans l'entrepôt de données

4.1.3. Les données sur les risques

Les données décrivant les risques ne sont stockées nulle part, mais doivent être calculées par rapport à chaque risque avec des modèles mathématiques. Par exemple, dans le cas du risque d'inondations (figure 3), les données nécessaires sont un modèle numériques du terrain pour décrire les variations altimétriques du terrain, des données hydrologiques et hydrauliques.

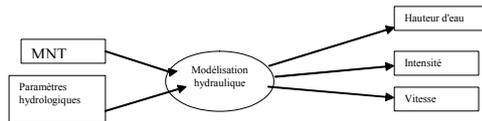


FIG. 3- les étapes pour calculer les données sur le risque d'inondations

Chaque risque naturel peut être représenté par différents paramètres. Dans le cas des inondations, les paramètres montrés par la figure ci-dessus sont la hauteur de l'eau, l'intensité et la vitesse. Nous pouvons utiliser ces paramètres individuellement ou calculer un nouvel indice de synthèse pour renseigner le risque. Dans notre cas, nous calculons un nouveau indicateur appelé "indicateur de risque" à l'intérieur de chaque bâtiment; ainsi nous devons ajouter des attributs aux relations *Prop Bâtis* et *Prop Non Bâtis*.

4.2. Le modèle multidimensionnel de l'entrepôt

Il y a un nombre important d'informations liées à l'analyse des risques qui peuvent être soit stockées ou calculées à partir des données sources comme "l'indicateur du risque", les informations sur les parcelles et les bâtiments, et les données sur les entreprises. Dans ce papier, nous allons utiliser la terminologie généralement utilisée dans la modélisation multidimensionnelle comme les dimensions, les faits, les hiérarchies.

4.2.1. Les mesures

Les mesures sont des attributs représentant les éléments spécifiques de l'analyse, comme la valeur locative, la surface. En général, on peut leur appliquer des calculs de somme et de moyenne pour comprendre des aspects particuliers. Nous distinguons deux types de mesures dans un entrepôt de données spatiales [Han *et al.*, 1997]:

- **Mesure Numérique:** une mesure numérique est une mesure qui contient uniquement des données numériques. Par exemple, une mesure peut être le revenu total d'un bâtiment, une opération de "roll-up" peut donner le revenu total par parcelle, par section cadastrale, etc.
- **Mesure spatiale:** une mesure spatiale est une mesure qui contient une collection de pointeurs aux objets spatiaux. Par exemple, durant la procédure de généralisation, les parcelles avec le même indicateur de risques peuvent être regroupées dans une même cellule, et la mesure ainsi formée contient une collection de pointeurs à ces parcelles.

4.2.2. Les dimensions

Une dimension est un objet qui inclut des attributs permettant à l'utilisateur d'explorer les mesures de plusieurs perspectives d'analyses. La figure 4 présente les dimensions utilisées dans le modèle multidimensionnel. Nous distinguons trois types de dimensions selon elles ont ou non une référence spatiale.

- **Dimension non spatiale:** une dimension non spatiale est une dimension contenant seulement des données spatiales. Dans notre cas, l'indicateur de risque peut être considéré comme une dimension non spatiale. Il contient des données non spatiales correspondant à la valeur du risque, dont la généralisation est aussi non spatiale – zone à faible risque, zone à fort risque.

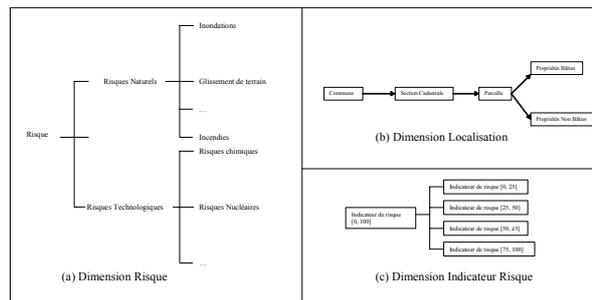


FIG. 4 - les hiérarchies de dimensions of (a) Risque (b) Localisation (c) Indicateur de risque

- **Dimension spatiale à non spatiale** : c'est une dimension dont les données du niveau primitive sont spatiales mais leurs généralisations à un niveau plus élevé, deviennent non spatiales. Nous n'allons pas traiter ce type de dimension dans notre cas.

- **Dimension spatiale** : c'est une dimension dont les données dans le niveau le plus bas et tous les niveaux supérieurs généralisés sont spatiaux. Par exemple, dans notre cas, Bâtiment, Parcelle, et Commune sont tous des éléments spatiaux de la dimension spatiale Localisation.

4.2.3. Le modèle Star

Le modèle logique le plus utilisé pour la conception des entrepôts de données est appelé le schéma Star qui porte dans son centre une table de faits entourée par plusieurs tables de dimensions formant une apparence d'une étoile. Et puisque les entrepôts de données sont orientés sujets, et dans le contexte de la gestion des risques naturels, nous nous intéressons à l'analyse de la vulnérabilité; nous allons construire notre modèle en mettant l'accent sur les mesures de vulnérabilité qui peuvent être obtenues ou calculées à partir des données sources tels que le revenu total, le nombre de salariés, et le chiffre d'affaires. Ainsi, le modèle (voir figure 5) contient une seule table de fait (avec des mesures de vulnérabilité), entourée par des tables de dimensions (spatiales et non spatiales).

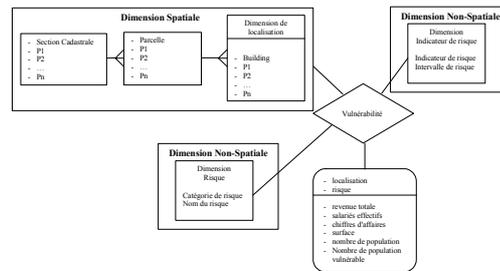


FIG. 5 - le modèle Star pour l'entrepôt de données pour la gestion des risques naturels

4.3. Les opérations OLAP

Avec les dimensions citées ci-dessus, les opérations OLAP peuvent être réalisées en parcourant de haut en bas le long de toutes les dimensions montrées dans la figure 5, nous utiliserons les opérations populaires et nous analyserons comment ils peuvent être réalisés dans un cube de données spatiales.

La figure 6 donne un aperçu des opérations OLAP discutées ci-dessus. En fournissant des opérations en ligne, des résumés flexibles et tabulaires, le cube de données peut être un élément essentiel de la procédure de prise de décision dans la gestion de risques naturels. Les vues multidimensionnelles offertes par l'entrepôt de données permettent aux décideurs d'avoir des vues multiples sur leur données; ainsi, ils peuvent avoir une connaissance à la fois détaillée et synthétique de la vulnérabilité de leur populations face aux risques naturels.

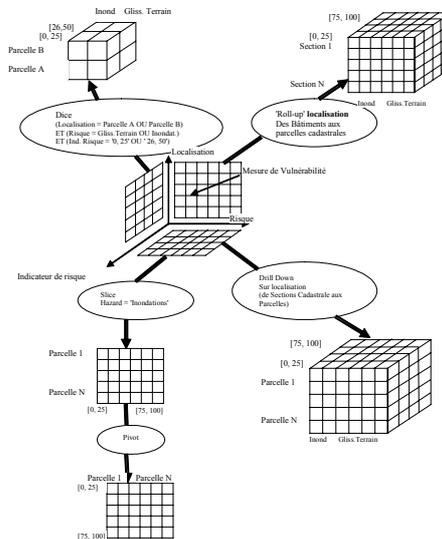


FIG. 6 – Un aperçu sur les opérations Olap sur le Cube de données

5. Conclusion

Dans ce papier, nous avons étudié la construction d'un entrepôt de données spatiales pour la gestion des risques naturels. L'entrepôt de données est basé sur un modèle de cube de données qui consiste à la fois en mesures et dimensions spatiales et non spatiales; le cube de données traite les informations spatiales et non spatiales. Ensuite, nous avons présentés les principales opérations OLAP pour faciliter la navigation et l'exploration du cube de données.

Références

- [Bédard et al., 2001] Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery. Geographic Data Mining and knowledge discovery. Taylor & Francis, 2001.
- [Han et al., 1997]. GeoMiner: A system prototype for spatial data mining. In Proc. 1997 ACM SIGMOD Int Conf Management of Data Tucson Arizona May 1997.
- [Han et al., 1998]. Selective Materialization: An Efficient Method for Spatial Data Cube Construction. In PAKDD, 1998.
- [Shekhar et al., 2001] Map cube: a visualization tool for spatial data warehouses. In Miller H., Han J. Geographic Data Mining and Knowledge Discovery. Taylor & Francis, 2001.
- [Stefanovic et al., 2000] Object-Based Selective Materialization for Efficient Implementation of Spatial Data Cubes. In IEEE TKDE, 12(6), 2000.
- [Theodoratos et al., 1999] Design data warehouse. Data and Knowledge Eng., Vol. 31, 1997

Summary

Data warehousing has emerged as an effective technique for converting data into useful information. This paper examines the possibility of using data warehousing technique in the natural hazards management framework to integrate various functional and operational data which are usually scattered across multiple, dispersed and fragmented systems. We present a conceptual data model for the data warehouse in the presence of various data formats such as geographic and multimedia data. We propose OLAP operations for browsing information in the data warehouse.