

1- Introduction

Dans ce chapitre, nous allons parler de l'entrepôt de données et de son historique, ainsi que de la différence entre les types d'entrepôt de données et les bases de données et les systèmes types d'aide à la décision.

2- définition base de données relationnelle

Une base de données relationnelle est un type de base de données où les données sont liées à d'autres informations au sein des bases de données. Les bases de données relationnelles sont composées d'un ensemble de tables qui peuvent être accessibles et reconstruites de différentes manières, sans qu'il soit nécessaire de réarranger ces tables de quelque façon que ce soit. Le langage de requête structuré (SQL) est l'interface standard pour une base de données relationnelle. Les instructions SQL sont utilisées à la fois pour interroger de façon interactive les données contenues dans la base de données relationnelle et pour collecter les données dans le cadre de rapports.[2]

3- définition l'entrepôt de données

Bill Inmon définit l'entrepôt de données dans son ouvrage "Building Data warehouse " [3] de la façon suivante : L'entrepôt de données est une collection de données orientées sujet, intégrées, non volatiles et historiques, organisées pour support d'un processus d'aide à la décision". Cette définition d'Entrepôt de Données a été conceptualisée en termes de caractéristiques du référentiel des données, qui seront détaillées dans les points suivants: [7]

- **Orientées sujet** : Les données de l'entrepôt sont organisées par thème (autour des sujets majeurs et des métiers de l'entreprise). L'intérêt dans cette organisation est de disposer d'un ensemble d'informations utiles sur un sujet transversal aux structures fonctionnelles et organisationnelles de l'entreprise [3].

- **Intégrées** : Les données dans l'entrepôt proviennent de différentes sources éventuellement hétérogènes. L'intégration est un processus qui consiste à résoudre les problèmes d'hétérogénéité, où les données contenues dans l'Entrepôt de Données sont divisées en grandes subdivisions appelées domaines [4].

- **Non volatiles** : Les données stockées au sein de l'entrepôt sont permanentes et ne peuvent être modifiées, et le rafraîchissement de l'entrepôt de données, consiste seulement à ajouter de nouvelles données sans modifier ou perdre celle qui existent. Ceci pour conserver la traçabilité des informations et des décisions prises [4].

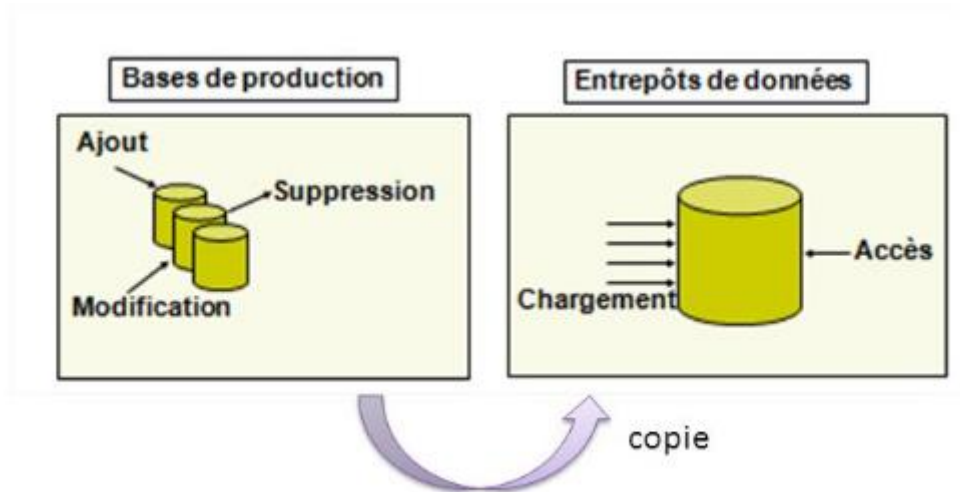


Figure 1.2 :Non volatilité des données [25]

3.1 Historique [7]:

L'origine des ED revient à 1960, où l'entreprise General Mills et l'Université Dartmouth, dans un projet conjoint, créent les termes "faits" et "dimensions". Les dates marquantes de l'histoire des entrepôts de données sont:

- En 1983, Teradata introduit dans son SGBD un système purement décisionnel.
- En 1988, Le terme Data Warehouse est utilisé pour la première fois dans l'article "An architecture for business and information Systems " publié par Barry Devlin et Paul Murphy dans le journal système d'IBM.
- En 1990, Red Brick Systems construit le système " Red Brick Warehouse " dédié à la construction d'entrepôt de données.
- En 1991, Bill Inmon publie le livre " Building the Data warehouse ".

- En 1995, La création de l'organisation " Data Warehousing Institute " pour soutenir et promouvoir la recherche dans le domaine des ED.
- En 1996, Ralph Kimball publie le livre " The Data Warehouse Toolkit ".
- En 1997, Réalisation de " Oracle 8 ", avec la prise en charge des requêtes des schémas en étoiles.

3.2 Objectifs d'un entrepôt de données [5]

L'environnement d'entrepôt de données doit aligner différents ensembles de compétences, fonctionnalités et technologies. Par conséquent, il doit satisfaire les objectifs suivants:

- Regrouper, organiser des informations provenant de sources diverses.
- Intégrer les informations récoltées et les stocker pour donner à l'utilisateur une vue orientée métier.
- Retrouver et analyser l'information selon plusieurs critères.
- Transformer un système d'information (SI) à vocation de production en un SI décisionnel.
- Séparer et combiner les données au moyen de toutes les mesures possibles de l'activité.
- Comporter un ensemble d'outils de requêtes, d'analyse et de présentation de l'information.

4- Différence entre bases de données et entrepôt de données

Le concept d'un entrepôt de données est apparu lors de l'existence de différence entre les systèmes transactionnels en ligne (OLTP) et les systèmes informationnels, dont certaines de ces différences fondamentales sont listées à travers le Tab1.1. Mais d'autres méthodes et techniques de conception et d'implémentation d'ED ont vu le jour, l'une de ces techniques est le modèle dimensionnel de Kim Bail apparue en 1996 [6].

Fonctionnalités	Base de données	Entrepôt de données
Caractéristiques	Basé sur le traitement optionnel	Basé sur le traitement d'information
Données	Actualisation de données stockées	Historisation de données stockées
Fonction	Les opérations quotidiennes	Les besoins d'information à long terme et aide à la décision
Utilisateur	Employés	Analystes, décideurs
Unité de travail	Court et simple transaction	Requêtes complexes
Orientation	L'orientation est sur la transaction	L'orientation est sur l'analyse
Vue	La vue des données est relationnelle plate	La vue des données est multidimensionnelle
Accès	Lecture et écriture	Lecture et rafraîchissement
Taille	Plusieurs gigaoctets	Plusieurs teraoctets
Priorité	Haute performance, haute disponibilité	Grande flexibilité, l'autonomie de l'utilisateur final
Métrique	Mesurer l'efficacité le débit , le débit transactionnel	Mesurer l'efficacité, le débit de la requête et le temps de réponse

Tab. 1.1 – Comparaison entre les bases de données et les entrepôts de données [7].

5- les Modèles des bases données :[26]

Un modèle de base de données illustre la structure logique d'une base de données, y compris les relations et les contraintes qui déterminent comment les données peuvent être stockées et accessibles. Les modèles de base de données individuels sont conçus en fonction des règles et concepts du modèle de données plus général adopté par les concepteurs. La plupart des modèles de données peuvent être représentés par un diagramme de base de données.

Il existe de nombreux types de modèles de bases de données. Parmi les plus courants :

- Modèle de base de données hiérarchique
- Modèle relationnel
- Modèle réseau

- Modèle de base de données orientée objet
- Modèle entité-association
- Modèle document
- Modèle entité-attribut-valeur
- Schéma en étoile
- Modèle multidimensionnel
- Le modèle relationnel-objet, qui associe les deux éléments qui composent son nom

Dans ce qui suit on présente le modèle relationnel ainsi que le modèle multidimensionnel

5.1 Modèle relationnel [27]

Le modèle le plus courant, appelé modèle relationnel, trie les données dans des tables, que l'on appelle aussi des relations, dont chacune se compose de colonnes et de lignes. Chaque colonne contient un attribut de l'entité en question, comme le prix, le code postal ou la date de naissance. L'ensemble des attributs d'une relation est appelé domaine. La clé primaire est constituée par un attribut spécifique ou une combinaison d'attributs. On peut y faire référence dans d'autres tables : elle est alors appelée clé étrangère.

Chaque ligne, également appelée tuple, comprend des données sur une instance spécifique de l'entité en question, comme un employé en particulier. Le modèle tient également compte des types de relations entre ces tables, notamment les relations un-à-un, un-à-plusieurs et plusieurs-à-plusieurs.

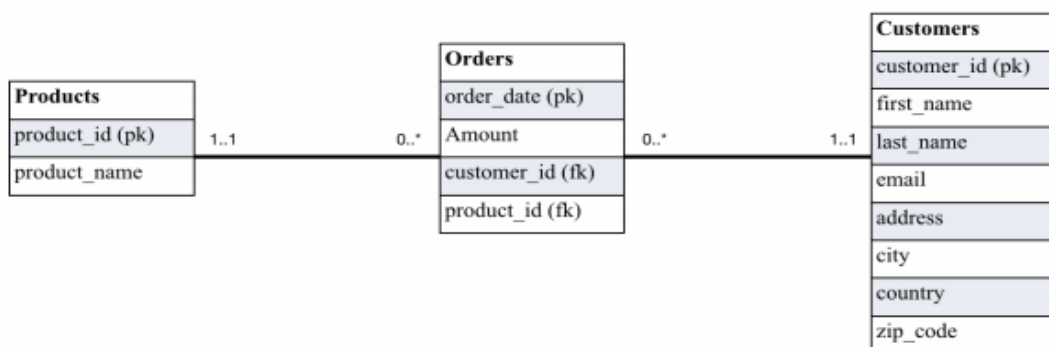


Figure 1.1 Exemple de modèle relationnel [1]

5.1.1 Schéma relationnel en étoile

Pour rendre les SGBDs relationnelles plus utiles pour les applications OLAP, de nouvelles fonctions leur sont rajoutées. Ces caractéristiques, dites super-relationnelles permettent de fournir des temps d'accès rapides aux applications OLAP, les données sont organisées selon un schéma en étoile (star). Ce modèle représente visuellement une étoile, les mesures d'intérêt pour l'OLAP sont stockées dans la table des faits (e.g., ventes, stock). Pour chaque dimension du modèle multidimensionnel il existe une table (e.g., région, produit, temps). Cette table stocke les informations relatives aux dimensions. La figure I.3 illustre ce modèle [10].

Les requêtes généralement appliquées sur ce schéma sont appelées « requêtes de jointure en étoile », et ont les propriétés suivantes [11] : – il y a des jointures multiples entre la table des faits et les tables de dimension. – il n'y a pas de jointure entre les tables de dimensions. – plusieurs prédicats de sélection peuvent être appliqués sur les attributs descriptifs de chaque table de dimension impliquée dans une opération de jointure.

Ce schéma est le plus populaire puisqu'il : – Offre une performance dans la rapidité d'analyse des données en réduisant au maximum le nombre de tables et le nombre de joints entre les tables. – Il offre aussi un maximum de flexibilité des dimensions puisque chaque dimension est représentée par une seule table. – Les modifications des dimensions deviennent plus simples. Sa simplicité peut lui permettre de faire des analyses complexes sans grande difficulté.[9]

Par contre, il y a de la redondance dans les tables de dimensions à cause de la non-normalisation des données [12].

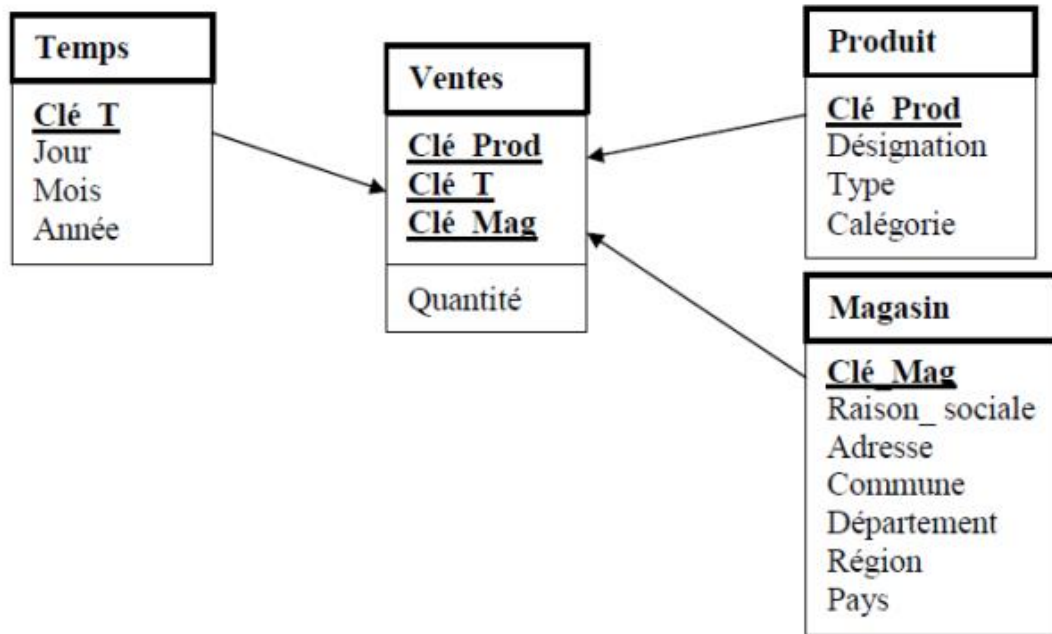


Figure 1.3 Exemple de schéma en étoile [14]

5.1.2 Le schéma en constellation :

de faits consiste à fusionner plusieurs modèles en étoile qui utilisent des dimensions Communes [13]. Un modèle en constellation comprend donc plusieurs faits et des dimensions communes ou non.[9]

5.1.3 Le schéma en flocons de neige (snowflake schema)

Dans le schéma en flocons de neige (snowflake schema) nous éclatons les tables de dimension en sous-tables selon la hiérarchie de cette dimension. Ce qui peut être vu comme une normalisation des tables de dimensions. Ce schéma évite les redondances d'information ce qui permet de gagner de l'espace disque et facilite l'alimentation. Mais peut altérer les performances de l'entrepôt car il nécessite des jointures lors des agrégats sur les dimensions.[9]

5.2 Modèle multidimensionnel [19]

Selon l'architecture à deux niveaux, un entrepôt de données est structuré suivant une modélisation multidimensionnelle. Ceci permet de représenter l'extension d'un entrepôt sous la forme de points dans un espace à plusieurs dimensions avec la métaphore du cube¹ ou de l'hyper-cube de données.

La Figure 1.6 présente un exemple de cube qui permet l'analyse des ventes de matériels informatiques. L'analyse des montants de ventes s'effectue en fonction de trois dimensions : les magasins où ont été effectuées les ventes, les dates de ventes et les produits vendus. Chacune de ces dimensions est associée à des paramètres de granularité différente (pour la dimension Magasin : ville, pays et continent). Ces niveaux hiérarchiques permettent d'obtenir des visions plus ou moins synthétiques lors des analyses OLAP.

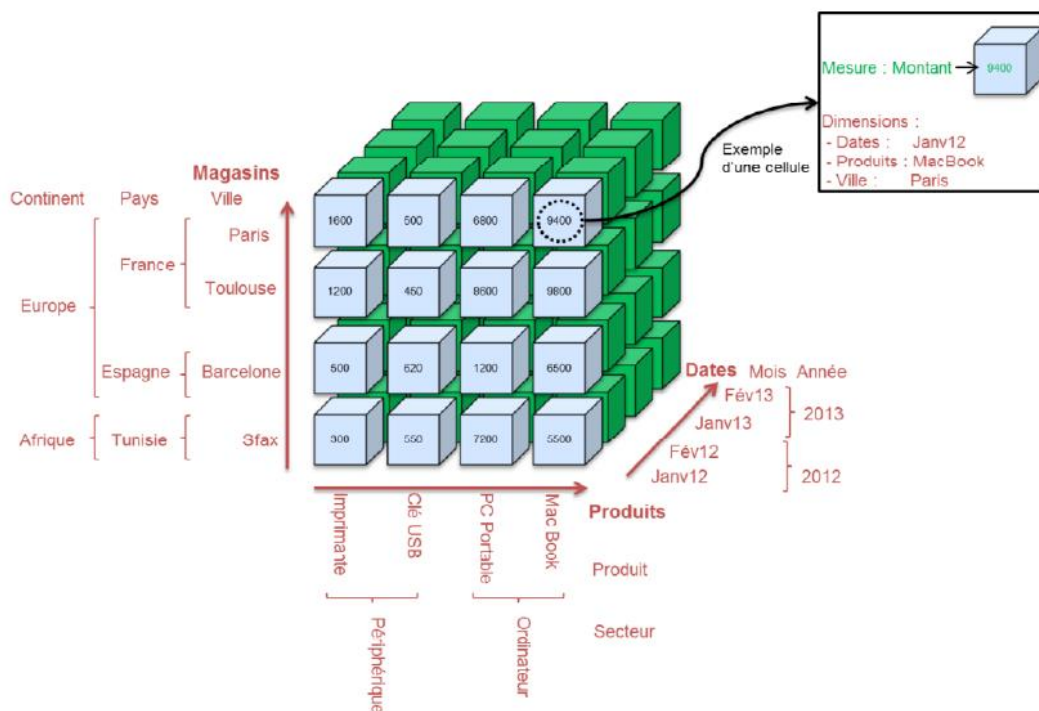


Figure 1.6. Exemple d'un cube représentant les ventes de matériels informatiques [19]

La modélisation d'un entrepôt sous la forme d'un cube s'avère très limitée puisqu'elle se limite à trois dimensions. Pour concevoir des schémas multidimensionnels plus élaborés, des

¹ Un cube est une représentation abstraite d'informations multidimensionnelles exclusivement numérique utilisé par l'approche OLAP (acronyme de On-line Analytical Processing)

structures plus avancées ont été définies ; elles permettent la modélisation de sujets d'analyse appelés faits, et d'axes d'analyse appelés dimensions.

Les faits sont des regroupements d'indicateurs d'analyse appelés mesures. Les dimensions sont composées d'attributs, appelés paramètres, agencés de manière hiérarchique et qui modélisent les différents niveaux de détails des axes d'analyse.

Une analyse multidimensionnelle est une requête partant sur les données d'un entrepôt. Généralement, le résultat d'une requête OLAP est représenté sous la forme d'une table à deux dimensions. La table multidimensionnelle de la Figure 3.1 représente le résultat d'une requête OLAP. Dans cet exemple, la table contient les analyses des montants des ventes en fonction des pays auxquels appartiennent les magasins. La vente est restreinte aux ventes effectuées en janvier 2012.

Ventes SUM(Montant)			Magasins			
			Continent	Europe		Afrique
			Pays	France	Espagne	Tunisie
Produits	Secteur	Produit				
	Ordinateur	MacBook		19200	6500	5500
		PCPortable		15400	1200	7200
	Périphériques	CléUSB		950	620	550
		Imprimante		2800	500	300
Dates = Janv12						

Valeurs cumulées à partir des magasins situés à Paris et à Toulouse

Tab.1.2 Une table multidimensionnelle [19]

5- Les systèmes d'aide a la décision [8]

Généralement, un système d'aide à la décision est constitué de données entreposées. Les architectures classiques reposent sur deux catégories d'espaces de stockage :

- L'entrepôt de données qui héberge les données de manière centralisée et uniforme.

Il constitue un premier niveau de stockage favorisant la collecte et la gestion historisée (conservation de l'évolution des données collectées) des données. (Nous en avons parlé dans le section 2)

- Les magasins de données constituent un second niveau du stockage utilisé à des fins d'analyse. Généralement, un magasin de données est dédié à un domaine métier ou une catégorie d'analyses. Les données sont organisées selon une modélisation multidimensionnelle afin de supporter efficacement les processus d'analyses en ligne (on-line analytic processing, OLAP).

6.1 : Niveaux d'abstraction : Concevoir un système décisionnel nécessite une phase de modélisation des données multidimensionnelles. Plusieurs approches ont été proposées selon trois niveaux d'abstraction :

- **Conceptuel**. Ce niveau d'abstraction consiste à représenter l'espace de données multidimensionnelles indépendamment des techniques informatiques.
- **Logique**. Ce niveau d'abstraction désigne une technique de modélisation (relationnel, objet, NoSQL, etc).
- **Physique**. Ce niveau d'abstraction correspond à un logiciel particulier choisi dans la technologie logique adoptée (Oracle, PostgreSQL, MongoDB, HBase...). Nous détaillons les deux niveaux d'abstraction conceptuel et logique sur lesquels nos travaux de thèse sont focalisés .

6.1.1 : Niveau conceptuel : Différents concepts sont définis pour représenter les données multidimensionnelles. Les sujets d'analyse (appelés faits), regroupent un ensemble d'indicateurs (appelés mesures). Les valeurs de ces indicateurs sont observables selon des axes d'analyse (appelés dimensions). Ces dimensions sont constituées de différents niveaux de détail, eux-mêmes organisés en hiérarchies ; par exemple, nous pourrions analyser le fait ventes au travers d'une mesure montant, ces montants pouvant être observés en fonction d'une dimension temps constituée de trois niveaux de détails (jour, mois, année) organisés au sein d'une hiérarchie définissant le jour comme un niveau de détail inférieur au mois, lui-même inférieur à l'année. Ces différents concepts permettent de concevoir des schémas multidimensionnels, appelés constellation. Les dimensions peuvent ainsi être partagées entre les faits. Un cas particulier consiste à ramener la constellation à un seul fait, on parle alors de schéma en étoile (star schema).

6.1.2 Niveau logique [8]:

Plusieurs modèles logiques ont été proposés pour convertir les schémas en constellation.

– **L’approche R-OLAP** consiste à utiliser le modèle relationnel pour représenter un schéma en constellation .Elle est de loin l’approche la plus utilisée. Ce modèle traduit chaque fait dans une table appelée table de fait. Chaque dimension est traduite dans une table appelée table de dimension. Dans la table de fait on retrouve les attributs représentant les mesures d’activités et les attributs les clés étrangères permettant la relation avec chacune des tables de dimensions. La table de dimension est constituée des paramètres et de la clé primaire (il est possible de normaliser les tables de dimensions constituant ainsi un schéma en flocon).

– **L’approche M-OLAP** consiste à utiliser un système dédié où les données sont organisées sous forme de tableaux multidimensionnels formant des hypercubes de données. Chaque arrête de l’hypercube correspond à une dimension et les cellules correspondent au fait.

– **L’approche H-OLAP** vise à cumuler les avantages des deux approches précédentes. Les données agrégées sont stockées sous formes multidimensionnelles tandis que les données détaillées sont stockées dans des structures relationnelles.

Exemple.

Nous illustrons ci-dessous un exemple basé sur l’approche ROLAP. Dans cet exemple, on observe le fait Tweet décrit selon quatre dimensions (Time, Subject, Location et User). La table Tweet contient des mesures et les clés étrangères pour référencer les tables des dimensions [8].

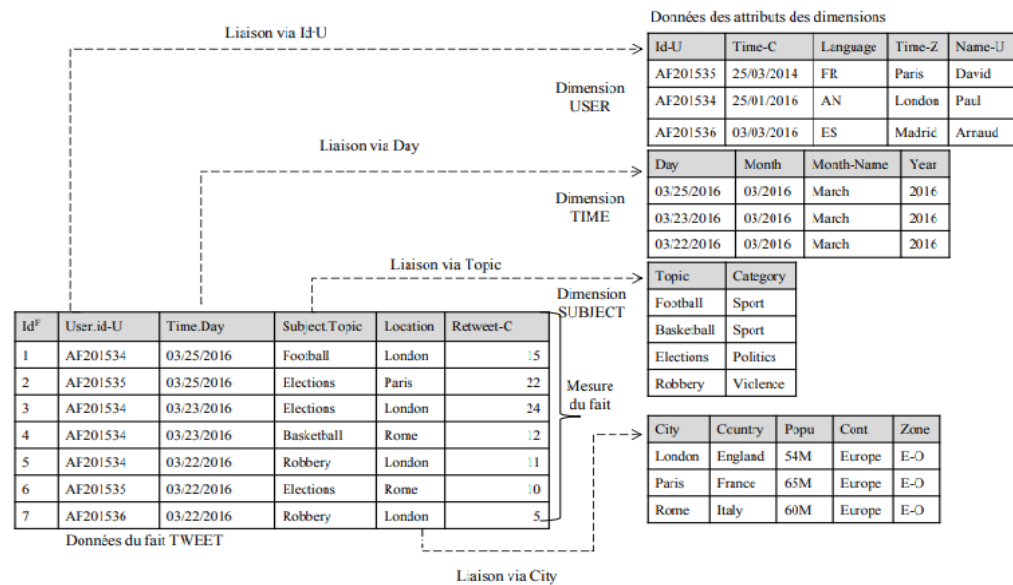


Figure 1.5 Exemple d'entrepôts de données multidimensionnelles R-OLAP concernant des tweets[8]

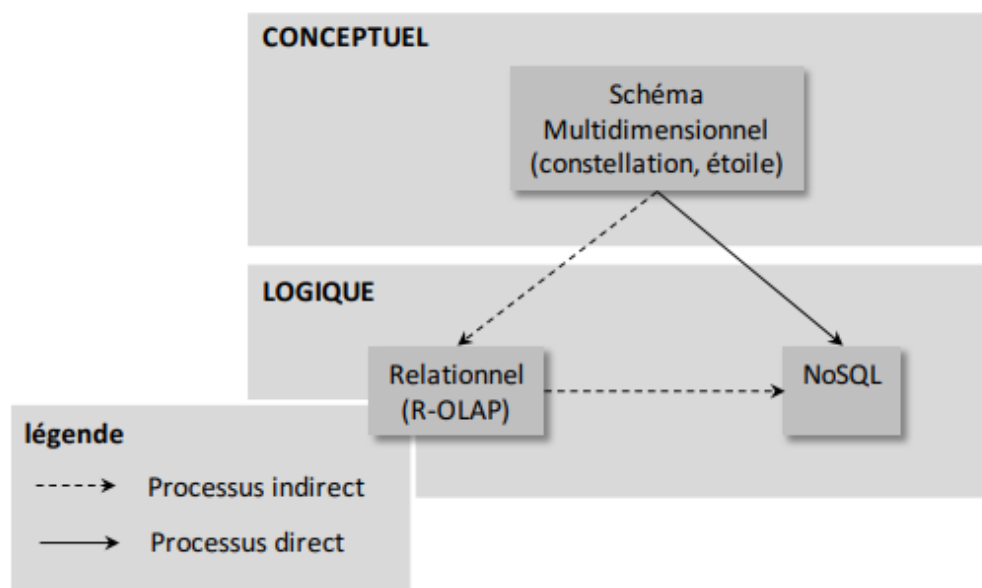


Figure 1.4 Processus de transformation des entrepôts de données multidimensionnelles du niveau conceptuel vers le niveau logique [8]

6-2 OLAP :

OLAP est un acronyme pour Online Analytical Processing. OLAP effectue une analyse multidimensionnelle des données d'entreprise et offre la possibilité de calculs complexes, d'analyses de tendances et de modélisation de données sophistiquées. C'est la base de nombreux types d'applications d'entreprise pour la gestion de la performance d'entreprise, la planification, la budgétisation, les prévisions, les rapports financiers, l'analyse, les modèles de simulation, la découverte des connaissances et les rapports sur les entrepôts de données. OLAP permet aux utilisateurs finaux d'effectuer des analyses de données dans de multiples dimensions, leur fournissant ainsi les informations et la compréhension dont ils ont besoin pour prendre de meilleures décisions.[26]

6-3 OLTP :

OLTP (Online Transactional Processing) est une catégorie de traitement de données axée sur des tâches orientées transactions. OLTP implique généralement l'insertion, la mise à jour et/ou la suppression de petites quantités de données dans une base de données. OLTP traite principalement un grand nombre de transactions effectuées par un grand nombre d'utilisateurs.[26]

7- Conclusion

On peut dire L'entrepôt de données est une collection de données orientées sujet, intégrées, non volatiles et historiées, organisées pour support d'un processus d'aide à la décision , Pour rendre les SGBDs relationnelles plus utiles pour les applications OLAP, de nouvelles fonctions leurs sont rajoutés. Ces caractéristiques, dites super-relationnelles permettent de fournir des temps d'accès rapides aux applications OLAP, les données sont organisées selon un schéma en étoile (star).

les systèmes d'aide a la décision Se compose de Niveaux d'abstraction (Conceptuel - Logique -Physique) Niveau logique se compose Plusieurs modèles ont été proposés (L'approche R-OLAP L'approche M-OLAP – L'approche H-OLAP)