

Chapitre 3

Généralités sur les Systèmes d'Information Décisionnels

Ce chapitre a pour objectif de présenter les concepts généraux d'un Système d'Information Décisionnel (SID) que nous allons utiliser pour réaliser notre projet. Ainsi dans la suite de ce chapitre nous allons expliquer les concepts généraux qui existent.

3.1 Concepts généraux

3.1.1 Modélisation multidimensionnelle

C'est une nouvelle méthode de conception autour des concepts métiers. Les analyses décisionnelles sont basées sur des traitements OLAP directement reliés à la modélisation de l'information sous une forme conceptuelle proche de la perception qu'en a l'analyste. Cette perception de l'information est basée sur une vision multidimensionnelle des données. La modélisation multidimensionnelle consiste à considérer un sujet analysé comme un point dans un espace à plusieurs dimensions. Les données sont organisées de manière à mettre en évidence le sujet analysé et les différentes perspectives de l'analyse. [22]

3.1.1.1 Présentation des concepts

Nous commencerons par expliquer les concepts fondamentaux de l'informatique décisionnelle nécessaire pour la mise en place de notre solution.

3.1.1.1.1 Concept de Faits

- **Fait** : modélise le sujet de l'analyse (ce que nous souhaitons mesurer) selon divers axes d'analyse. Il est formé de mesures correspondant aux informations de l'activité analysée et de clés étrangères des axes d'analyses (dimension). [22]

Les faits sont généralement des valeurs numériques provenant des processus d'affaires.

Il existe trois types de fait : fait additif, fait semi-additif et fait non additif.

- ◊ **Fait additif** : un fait est additif s'il est additionnable suivant toutes les dimensions.
- ◊ **Fait semi-additif** : un fait est semi-additif s'il n'est additionnable que suivant certaines dimensions.
- ◊ **Fait non additif** : un fait est non additif s'il n'est pas additionnable quelques soient les dimensions.
- **Table de faits** : c'est la table principale du modèle dimensionnel. Elle contient les données observables (les faits) que nous possédons sur un sujet et que nous voulons étudier, selon divers axes d'analyse (les dimensions). Elle contient un ou plusieurs faits numériques qui se produisent par la combinaison de clés définissant chaque enregistrement. En d'autres termes, c'est un ensemble de données structuré, composé de champs de type dimension et de champs de type mesure. Elle contient les clés étrangères des axes d'analyse. [22]

3.1.1.1.2 Concept de Dimension

- **Dimension** : c'est un axe à partir duquel nous observons les faits. Elle se compose de paramètres correspondant aux informations faisant varier les mesures de l'activité. Les dimensions fournissent le contexte (qui, quoi, quand, où, pourquoi et comment) des faits. [22]
- **Table de dimension** : contient les axes d'analyse selon lesquels nous voulons étudier des données observables qui, soumises à une analyse multidimensionnelle, donnent aux utilisateurs des renseignements nécessaires à la prise de décision. Elle contient l'ensemble des informations descriptives des faits. Et elle contient en général beaucoup moins d'enregistrements qu'une table de faits. [22]

3.1.1.1.3 Concept de Granularité

La granularité est le niveau de détail représenté par les données dans une table de faits ou de dimensions au niveau de l'entrepôt de données.

3.1.1.1.4 Concept d'hiérarchie

- **Hiérarchie** : elle sert, lors des analyses, à restreindre ou accroître les niveaux de détail de l'analyse. Elle organise les paramètres d'une dimension conformément à leur niveau de détail. [22]
- **Niveau d'Hiérarchie** : un niveau d'hiérarchie se définit au niveau des tables de dimensions.

Cela permet d'agréger les données.

3.1.1.1.5 Concept de Mesure

Une mesure est une quantité présente dans la table de faits et qui permet de mesurer les faits. C'est l'élément de données sur lequel portent les analyses, en fonction des différentes dimensions. [22]

Par exemple : nombre de personnes dans une chambre est un exemple de mesure.

3.1.2 Types de modèles

Il existe plusieurs types de modèles qui peuvent être utilisés pour mettre en place un datawarehouse et qui utilisent ces notions de faits et de dimensions. Dans la pratique, pour construire un modèle approprié pour un datawarehouse, nous avons deux différents types de modèles : le modèle en étoile et le modèle en flocon. Nous allons donc présenter chacun de ces modèles.

3.1.2.1 Modèle en étoile

Il est composé d'une table de fait centrale et des tables de dimensions. Ces tables de dimensions n'ont pas de liaison entre elles et qu'elles sont directement reliées à la table de faits qui contient les données à analyser. Ce modèle est basé sur des faits qui ont lieu dans le temps et qui ne se modifient pas.

- **Avantages**

- Facilité de navigation.
- Nombre de jointures limité.

- **Inconvénient**

- Redondance dans les dimensions. [22]

3.1.2.2 Modèle en Flocon

Ce modèle est une variante du modèle en étoile. Il est composé d'une table de fait qui est au cœur du modèle et des dimensions. Le principe étant qu'il peut exister des hiérarchies de ces dimensions qui gravitent autour de la table centrale (table de fait). Dans une table de dimension, nous n'avons qu'un seul niveau hiérarchique. La table de dimension de niveau hiérarchique le plus bas est reliée à la table de fait. Nous disons qu'elle a la granularité la plus fine. L'origine de la modélisation en flocon est de créer des hiérarchies de dimensions, de telle manière à avoir moins de lignes par dimensions.

- **Avantages**

- Normalisation des dimensions.
- Économie d'espace disque.

- **Inconvénients**

- Modèle plus complexe (jointure).
- Requêtes moins performantes.
- Grande complexité en termes de lisibilité et de gestion : navigation difficile. [22]

3.1.3 Stockage des données

Le COUD a une masse de données en croissance constante. Ces données sont d'abord stockées dans un datawarehouse puis dans des cubes ou datamarts selon les traitements métiers que nous aurons à faire.

3.1.3.1 Datawarehouse

Un entrepôt de données, ou dataWarehouse est une vision centralisée et universelle de toutes les informations de l'entreprise. C'est une structure, c'est-à-dire comme une base de données, qui a pour but, contrairement aux bases de données, de regrouper les données de l'entreprise pour des fins analytiques et pour aider à la prise de décision. Cette décision est une action entreprise par les décideurs de l'entreprise et vise à améliorer, quantitativement ou qualitativement, la performance de l'entreprise. En gros, c'est un gros volume d'informations épurées, organisées selon les métiers de l'entreprise, historisées et provenant de plusieurs sources de données, servant aux analyses et à l'aide à la décision. [22]

3.1.3.1.1 Caractéristiques

Les données du datawarehouse ont un ensemble de caractères, qui sont au nombre de quatre (4), qui leur est spécifique. Elles doivent alors être :

- **orientées sujet** : au cœur du datawarehouse, les données sont organisées par thème. Les données propres à un thème, par exemple les données de la scolarité, contiennent seulement l'information utile à la scolarité pour la prise de décision ;
- **intégrées** : les données proviennent de sources hétérogènes utilisant chacune un type de format. Avant d'être intégrées au sein du datawarehouse, elles doivent être mises en forme et unifiées afin d'en assurer la cohérence ;
- **non volatiles** : les données ne disparaissent pas et ne changent pas. Un datawarehouse devrait conserver la traçabilité des informations et des décisions prises. Une requête émise sur les mêmes données à un temps d'intervalle doit donner le même résultat ;
- **historisées** : contrairement au système de production les données ne sont jamais mises à jour. Les données non volatiles sont aussi horodatées. Chaque nouvelle donnée est insérée.

Nous devons pouvoir visualiser l'évolution dans le temps d'une valeur donnée. Le degré de détail de l'archivage est bien entendu relatif à la nature des données. Toutes les données ne méritent pas d'être archivées. [22]

3.1.3.1.2 Structure des données

Le datawarehouse peut se structurer en quatre classes de données organisées selon un axe historique et un axe synthétique.

- **Les données détaillées** : elles reflètent les événements les plus récents. Les intégrations des données issues des systèmes de production vont être réalisées à ce stade.
- **Les données agrégées** : elles correspondent à des éléments d'analyses représentatives des besoins des utilisateurs. Ce sont donc des données déjà traitées par le système et elles constituent un premier résultat d'analyse et de synthèse des informations contenues dans le système décisionnel. Elles doivent être facilement accessibles et compréhensibles.
- **Les données historisées** : chaque nouvelle insertion dans le Datawarehouse ne détruit pas les anciennes valeurs mais crée une nouvelle occurrence de la donnée.
- **Les méta-données** : constituent l'ensemble des données qui décrivent des règles ou processus attachés à d'autres données. Elles permettront notamment de connaître :
 - quelles sont les données entreposées, leur format, leur signification, leur degrés d'exactitude.
 - les processus de récupération/extraction dans les bases sources.
 - la date du dernier chargement du datawarehouse.
 - l'historique des données sources et de celles du datawarehouse. [7]

3.1.3.2 Datamart

Un datamart ou encore magasin de donnée est une vue partielle de l'entrepôt de données mais qui est orientée métier. C'est un sous-ensemble du datawarehouse contenant des informations se rapportant à un secteur d'activité particulier de l'entreprise ou à un métier qui y est exercé. Il se situe en aval du datawarehouse et est alimenté par celui-ci. Nous pouvons donc créer plusieurs datamarts correspondant aux différents besoins des utilisateurs. Cela permet de réduire le nombre d'opérations sur les bases de production. De plus cela permet d'offrir aux utilisateurs un outil spécifiquement adapté à leurs besoins. Cet outil sera plus petit et permettra donc un accès plus rapide aux données. [8]

3.1.3.3 Cube

Il est nécessaire de disposer d'une structure de stockage permettant de visualiser les données dans plusieurs dimensions, de pouvoir définir et ajouter des dimensions facilement, de manipuler les données ainsi représentées facilement et efficacement ; et un cube nous permet de le faire. Un cube est une base de données multidimensionnelle qui contient des informations à des fins analytiques. Il est principalement constitué de dimensions et de mesures.

- **Mesure** : valeur contenue dans une cellule du cube associée aux valeurs (ou positions) prises sur les dimensions composant le cube.
- **Dimension** : détermine la manière dont nous regardons les données pour les analyser. Les dimensions définissent la structure du cube.
- **Manipulation d'un cube** : quatre (4) opérations sont définies sur les cubes OLAP. Les deux opérations de base pouvant être appliquées sur un hypercube sont :
 - ◇ Slicing : sélection de tranches du cube par des prédicats selon une dimension.
 - ★ filtre d'une dimension selon une valeur.
 - ◇ Dicing : extraction d'un bloc de données ou extraction d'un sous-cube en faisant une sélection sur les dimensions. Sur les hiérarchies des dimensions, il nécessite deux opérateurs supplémentaires :
 - ◇ Roll up ou forage vers le haut : consiste à représenter les données du cube à un niveau de granularité supérieur conformément à la hiérarchie définie sur la dimension.
 - ★ Utilisation de la fonction d'agrégation (somme, moyenne, etc.) spécifiée pour la mesure et la dimension.
 - ◇ Drill down ou forage vers le bas : consiste à représenter les données du cube à un niveau de granularité de niveau inférieur, donc sous une forme plus détaillée. [9]

3.1.4 Architecture OLAP

OLAP (Traitement analytique en ligne) est un processus informatique qui permet d'extraire et d'analyser des données sous plusieurs angles. Il a pour but d'organiser les données à analyser par domaine ou thème et d'en ressortir des résultats pertinents pour le décideur. C'est un outil de Reporting dont la couche d'analyse permet de générer les indicateurs de résultats en fonction du contenu d'un entrepôt de données. Le concept OLAP est une méthode d'analyse apparue pour la première fois en 1993 dans le livre blanc intitulé *Providing OLAP to User-Analysts : An IT Mandate*, réalisé par Edgar Frank Codd qui est l'un des concepteurs des bases de données relationnelles, à la demande d'Arbor Software (devenue Hypérion Solutions). [10]

3.1.4.1 Règles de base

Le modèle de base, attribué à Codd, se repose sur douze (12) règles de base. Elles représentent la définition première du concept OLAP. En effet, elles ont été énumérées comme étant les critères décrivant ce que doit être un outil OLAP. En outre elles sont conçues pour définir ce qui est exigé d'un Système de Gestion de Bases de Données (SGBD) afin qu'il puisse être considéré comme relationnel (SGBDR). Ces 12 règles de base sont :

- **Multidirectionnalité** : le Modèle OLAP est multidimensionnel par nature. Une base OLAP offre une vue multidimensionnelle des données ;
- **Transparence** : l'emplacement physique du serveur OLAP c'est à dire les éléments techniques est transparent pour l'utilisateur ;
- **Accessibilité** : l'utilisateur OLAP dispose de l'accessibilité à toutes les données nécessaires à ses analyses. La complexité et l'hétérogénéité des données sont masquées par les outils OLAP ;
- **Constance des temps de réponses (stabilité)** : la performance des reporting reste stable indépendamment du contexte d'analyse ;
- **Architecture Client-Serveur** : le serveur OLAP s'intègre dans une architecture client-serveur. Le côté serveur a en charge l'homogénéisation des données, les clients se connectent simplement au serveur ;
- **Indépendance des dimensions** : toutes les dimensions doivent être équivalentes en structure et en possibilité de calcul. Toute fonction appliquée à une dimension s'applique aussi aux autres. Il ne doit exister qu'une seule structure logique pour toutes les dimensions ;
- **Gestion des matrices creuses** : le serveur OLAP assure la gestion dynamique de la mémoire physique nécessaire pour stocker les données non nulles. Nous parlons ici d'optimisation des matrices creuses ; les cellules non renseignées ne doivent pas encombrer la mémoire et ainsi ralentir les temps d'accès ;
- **Accès Multi-Utilisateurs** : le serveur OLAP offre un support multi-utilisateurs (gestion des mises à jour, intégrité, sécurité, gestion des accès concurrents aux données) ;
- **Pas de restrictions sur les opérations inter et intra dimensions** : le serveur OLAP permet la réalisation d'opérations inter dimensions sans restriction ;
- **Manipulation aisée des données** : le serveur OLAP permet une manipulation intuitive des données ;
- **Simplicité des rapports** : la flexibilité (ou souplesse) de l'édition des rapports est intrinsèque au modèle ;

- **Nombre illimité de dimensions et nombre illimité d'éléments sur les dimensions** : les nombres de dimensions et de niveau d'agrégation possibles sont suffisants pour autoriser les analyses les plus poussées. Tout outil OLAP doit gérer au moins 15 à 20 dimensions dans le même modèle. Le nombre de niveaux d'agrégation est illimité. [10]

3.1.4.2 Caractéristiques

Une autre définition d'OLAP est résumée dans l'acronyme FASMI (Fast Analysis of Shared Multidimensional Information), ou analyse rapide d'information multidimensionnelle partagée. Cette définition caractérise les systèmes OLAP. Chaque mot a sa signification qui lui est propre :

- **Fast** : le serveur doit pouvoir répondre à la plupart des demandes d'un utilisateur dans les cinq secondes ;
- **Analysis** : faire face à toutes les logiques d'affaire et de statistiques, ainsi que fournir la possibilité aux utilisateurs de construire leurs calculs et leurs analyses sans avoir à programmer ;
- **Shared** : le système implémente toutes les conditions de sécurité du partage de l'information sensible afin de préserver la confidentialité, l'intégrité et la disponibilité des informations même lors d'accès simultanés ;
- **Multidimensional** : caractéristique majeure, les produits OLAP doivent fournir des vues conceptuelles multidimensionnelles des données et supporter des hiérarchies de dimensions ; c'est l'intérêt de OLAP, les données doivent pouvoir être consultées sous différents angles de vue ;
- **Information** : accès à toutes les données et informations, où qu'elles résident et de façon illimitée en terme de volumétrie. [10]

3.1.4.3 Types d'implémentation

Le noyau d'un système OLAP est son serveur. Ces serveurs sont classés selon la politique régissant l'architecture du serveur. L'architecture d'un entrepôt de données s'appuie sur trois services :

- **base de données** : support de données agrégées ou résumées possédant une structure multidimensionnelle (SGBD multidimensionnel ou relationnel) ;
- **serveur OLAP** : gestion de la structure multidimensionnelle et des accès aux données de la part des usagers ;
- **module client** : manipulation, exploration et affichage des données sous différentes formes (statistiques, tableaux, etc.).

Ainsi, nous pouvons trouver des usages d'OLAP adaptés aux besoins que nous avons. Parmi les usagers d'OLAP, nous avons l'architecture MOLAP, l'architecture ROLAP et l'architecture HOLAP.

3.1.4.3.1 Architecture MOLAP

Les données détaillées et agrégées sont stockées dans une base de données multidimensionnelle (hypercube). Le serveur OLAP extrait les données de l'hypercube et les présente directement au module client. MOLAP permet un accès direct aux données dans le cube. La base MOLAP est l'application physique du concept OLAP. Il s'agit réellement d'une structure multidimensionnelle.

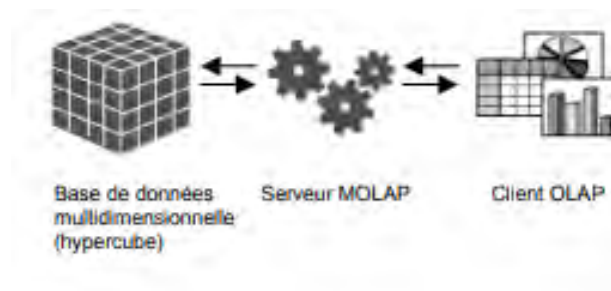


FIGURE 3.1 – Architecture MOLAP [4]

- **Avantages des systèmes MOLAP :**
 - temps de réponses satisfaisants ;
 - rapides et performantes ;
 - conçus exclusivement pour l'analyse multidimensionnelle.
- **Inconvénients des systèmes MOLAP :**
 - difficiles à mettre en place ;
 - formats souvent propriétaires ;
 - peu adaptés pour les gros volumes de données. [4]

3.1.4.3.2 Architecture ROLAP

ROLAP est une technique de modélisation et de stockage des données fondée sur une structure relationnelle. La base ROLAP est en fait une classique base relationnelle organisée pour fonctionner comme une base OLAP. Les données détaillées et agrégées sont stockées sous forme de table dans une base de données relationnelle. La base de données est structurée selon un modèle en étoile ou en flocon. Le serveur extrait les données par les requêtes SQL et les interprète selon une vue multidimensionnelle pour les présenter au module client.

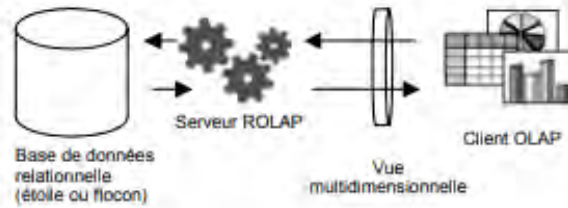


FIGURE 3.2 – Architecture ROLAP [4]

- **Avantages des systèmes ROLAP :**

- souplesse ;
- évolution facile ;
- permettent de stocker de gros volumes de données ;
- peu coûteux.

- **Inconvénients des systèmes ROLAP :**

- peu efficaces pour les calculs complexes ;
- temps de réponse plus lents. [4]

3.1.4.3.3 Architecture HOLAP

HOLAP est une solution hybride de MOLAP et ROLAP. Il propose de cumuler les avantages des deux modèles précédents. Les données multidimensionnelles sont stockées soit dans un SGBD relationnel, soit dans un SGBD multidimensionnel, afin d'éviter les problèmes des systèmes MOLAP et ROLAP. Dans ces systèmes, les données de base c'est à dire les tables de faits et les tables de dimensions sont stockées dans des structures relationnelles et les données agrégées sont stockées sous formes multidimensionnelles. Le serveur accède aux deux bases de données et présente les données, sous forme d'une vue multidimensionnelle, au module client.

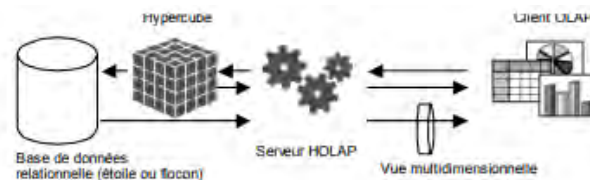


FIGURE 3.3 – Architecture HOLAP [4]

- **Avantages des systèmes HOLAP :**

- bon compromis au niveau des coûts et des performances ;
- permet de gérer de très grandes quantités de données et d'avoir des temps de réponses acceptables en analyse OLAP.

- Inconvénients des systèmes HOLAP :

- difficile de les mettre en œuvre. [4]

3.1.4.3.4 Comparaison des architectures MOLAP, ROLAP et HOLAP

TABLE 3.1 – Résumé des caractéristiques des architectures ROLAP, MOLAP et HOLAP [4]

Critère de comparaison	ROLAP	MOLAP	HOLAP
Stockage des données de base (détaillées)	BD relationnelle	BD multidimensionnelle	BD relationnelle
Stockage des agrégations	BD relationnelle	BD multidimensionnelle	BD multidimensionnelle
Performance des requêtes	1/3	3/3	2/3
Espace de stockage nécessaire	3/3	2/3 (si des technologies de compression sont utilisées)	1/3 (si des technologies de compression sont utilisées)
Maintenance de la structure (en ordre croissant du niveau de maintenance nécessaire)	1/3	2/3	1/3

Certains de ces concepts seront utilisés pour la modélisation de notre projet.