Mémoire Master 2 MIAGE classique

Thème

Etude des approches de modélisation des entrepôts de données (DataWarehouse)

Réalisé par :

Zakaria MEDJIR

Encadré par :

Monsieur Reda BENDRAOU

Promotion : 2017-2018

Remerciements

Nous tenons à remercier très sincèrement notre encadrant, Monsieur Reda BENDRAOU, pour avoir dirigé et encadré ce mémoire, pour sa rigueur scientifique, pour ses critiques constructives ainsi que pour ses précieux conseils et encouragements.

Nous adressons nos remerciements les plus sincères à nos camarades ainsi qu'à nos amis pour les moments de joie qu'on a passés ensemble.

Nous remercions les membres du jury pour l'intérêt porté à notre travail. Merci de nous avoir honorés de votre présence.

Nous souhaitions garder le meilleur pour la fin en remerciant nos Familles pour leur soutien tout au long de ce mémoire.

Résumé

La BI est une discipline importante pour les entreprises. Les enjeux auxquels elle est confrontée sont stratégiques. Un concept central dans la BI est celui de l'entrepôt de données qui regroupe un ensemble de données consolidées provenant de sources hétérogènes (généralement des bases de données en 3FN).

Afin de modéliser les entrepôts de données, les deux modèles Inmon et Kimball sont les plus utilisés. Le modèle de Kimball est basé sur les concepts de fait et de dimension tandis que celui d'Inmon est basé sur les schémas traditionnels (modèle Entité-Association). Ils permettent l'intégration de données transformées à partir de sources hétérogènes. Les deux solutions monopolisent le marché de la BI. Cependant, un troisième modèle appelé «Data Vault», de son créateur Linstedt, est en train de gagner du terrain d'année en année.

L'approche de ce modèle est différente : il s'agit de construire un entrepôt de données brutes (sans transformation) à partir de sources hétérogènes.

Le but de ce travail de master est de faire une synthèse sur la BI à travers ces trois approches. D'abord, nous étudions les trois approches à part, ensuite, nous élaborons une étude comparative entre elles. La comparaison entre les trois approches est effectuée selon plusieurs critères tels que : la méthodologie de développement, l'architecture de l'entrepôt et la modélisation de données, ...Et enfin, nous rédigeons une synthèse permettant de désigner la meilleure approche.

Mots clés : BI, Entrepôt de données, Modèle de Kimball, Modèle d'Inmon, Modèle Data Vault.

Abstract

BI is an important discipline for enterprises. The challenges that it faces are strategic. A central concept in BI is the data warehouse, which includes a set of consolidated data from heterogeneous sources (usually databases in 3NF).

To model data warehouses, both Inmon and Kimball models are the most used. Kimball model is based on the concepts of fact and dimension while that of Inmon is based on traditional patterns (Entity-association model). They enable the integration of transformed data from heterogeneous sources. Both solutions monopolize the BI market. However, a third model called "Data Vault" for his creator Linstedt, is gaining position from year to year.

The approach of this model is different : build a data warehouse with raw data (unprocessed) from heterogeneous sources.

The purpose of this master work is to establish synthesis study of the BI, then draw a comparison between the three approaches. First, we study the three approaches apart. Next, we compare them according to several criteria such as : Development methodology, Data warehouse architecture and data modeling. Finally, we synthesizes to designate the best approach.

Keywords : BI, Data warehouse, Inmon model, Kimball model, Data Vault model.

Table des matières

Liste des tableaux

Table des figures

Liste des Abréviations

3NF	3éme Forme Normale
BDM	Base de Données Multidimensionnelles
BI	Business Intelligence
CMMI	Capability Maturity Model Integration
DIS	Data Item Set
DM	Data Mining
DW	Data Warehouse
E/A	Diagramme Entité Association
EDW	Entreprise Data Warehouse
ERD	Diagrammes d'Entité-Association
ERP	Enterprise Resource Planning
ETC	Extraction, transformation et chargement
ETL	Extract, Transform and Load
IT	Information Technologie
KPI	Les indicateurs clés de performance
OLAP	OnLine Analytical Processing
OLTP	OnLine Transaction Processing
SID	Système d'Information Décisionnel

Introduction générale

Aujourd'hui, parmi les défis auxquels font face les entreprises, on trouve l'exploitation et l'analyse de données opérationnelles qu'elles détiennent dans leurs sources de données hétérogènes. Le but ultime de ces tâches est d'obtenir de l'information utile pour la prise de décision.

La Business intelligence ou système décisionnel est l'anneau manquant qui peut transformer ces données brutes en informations utiles et pertinentes qui peuvent supporter les décisions prises par les dirigeants des entreprises. Un concept central dans un tel système est l'entrepôt de données (Data warehouse). Ce dernier est donc un composant principal du système décisionnel qui a pour but de stocker les données opérationnelles, provenant de plusieurs sources, dans une perspective décisionnelle et de les fournir aux utilisateurs sous certaines formes pour des fins d'analyse.

La mise en place d'un entrepôt de données nécessite une approche de modélisation qui prend en considération tous les aspects de développement comme la modélisation de données, la gestion de projet, la gestion des risques, le déploiement et bien d'autres aspects essentiels.

Depuis des années, deux approches s'affrontent quant à la modélisation des entrepôts de données : l'approche de modélisation par sujet d'Inmon et l'approche dimensionnelle de Kimball. Cependant, ces dernières années, une troisième approche est apparue et elle gagne du terrain d'année en année. Cette approche est créée par Linstedt et s'appelle « Data Vault ».

Dans le cadre de ce mémoire de Master, nous élaborons une étude comparative sur ces trois approches de modélisation d'entrepôt de données.

Dans le premier chapitre, nous abordons les systèmes décisionnels et l'architecture

autour de laquelle ils sont construits, en détaillant chacun de ses composants et en mettant l'accent sur l'entrepôt de données.

Dans le second chapitre, nous nous focalisons sur les approches de modélisation des entrepôts de données. Nous présentons donc la définition de l'approche, sa philosophie, son architecture ainsi que sa méthodologie de développement.

Dans le troisième chapitre, nous dressons une analyse comparative entre les trois approches en s'appuyant sur certains critères afin d'aboutir à une synthèse permettant de désigner la meilleure approche.

Dans le quatrième et dernier chapitre, nous choisissons l'approche la plus optimale pour un cas réel en la mettant en place et nous décrivons les résultats par la suite.

Chapitre 1

Les systèmes d'Information décisionnels

Introduction

Aujourd'hui, l'entreprise doit faire face à beaucoup de défis et en particulier, le taux énorme de l'information provenant des différentes sources hétérogènes et sa gestion afin de pouvoir en tirer des profits.

Les sources d'information dans une entreprise sont éclatées et amples. La consolidation et l'analyse de ces sources permettent l'optimisation du patrimoine informationnel de cette entreprise et son pilotage efficace afin de :

— Surmonter la concurrence rude sur le marché.

— Assurer l'innovation continue.

— Fidéliser et être à l'écoute des clients de l'entreprise. . .etc.

Il est donc primordial de doter l'entreprise d'un système d'information décisionnel (SID) qui aura pour mission l'analyse et la consolidation de ses données.

Dans ce chapitre, nous allons aborder les concepts liés aux systèmes décisionnels. Nous commençons par des définitions, nous citons après l'historique et les objectifs des SID et nous nous étalons à la fin sur l'architecture d'un SID.

1.1 Définitions

1.1.1 La notion d'information

Une panoplie de définitions sont données à ce concept. Ces définitions divergent selon le positionnement des chercheurs d'une part et les disciplines concernées d'autre part.

Parmi les nombreuses définitions proposées, nous retiendrons celle de Davis, qui se réfère aux fonctions de l'information, indépendamment de sa forme et de son traitement :

"L'information est une image des objets et des faits ; elle les représente, elle corrige ou confirme l'idée qu'on se faisait. L'information contient une valeur de surprise, en ce sens qu'elle apporte une connaissance que le destinataire ne possédait pas ou qu'il ne pouvait pas prévoir "(1).

L'information a une valeur car elle permet de choisir, de prendre des décisions et d'agir. Sa valeur est donc liée à son emploi dans le contexte de prise de décisions (**?**) . Ainsi pour March, «l'information donne son sens à une situation de décision et modifie donc à la fois la structure des options et les préférences recherchées» (**?**).

1.1.2 La notion de système d'information

La notion de système d'information a donné lieu à différentes interprétations et sa définition est loin de faire l'unanimité (**?**). En effet plusieurs définitions ont été données à ce terme. Dans la première acception, nous retenons celle de Le Moigne qui dit que « Le système d'information est l'ensemble des méthodes et moyens de recueil, de contrôle, et de distribution des informations nécessaires à l'exercice de l'activité en tout point de l'organisation. Il a pour fonction de produire et de mémoriser les informations de l'activité du système opérant, puis de les mettre à disposition du système de décision (système de pilotage) » (**?**). Dans la seconde acception, on retrouve la définition de Reix selon laquelle système d'information est « un système d'interprétation d'un ensemble d'acteurs sociaux qui mémorisent et transforment des représentations via des technologies de l'information et des modes opératoires » (**?**).

1.1.3 Les systèmes d'information décisionnels

Le système d'information décisionnel appelé aussi système décisionnel ou BI (pour Business Intelligence en anglais) est généralement défini comme étant « un système permettant aux décideurs de l'entreprise de disposer d'informations pertinentes et d'outils d'analyse puissants pour aider à prendre les bonnes décisions au bon moment » (**?**)

(**?**) voit qu'un système décisionnel est « l'ensemble des moyens, outils et méthodes qui supportent le processus de collecte, consolidation, modélisation, analyse et restitution des

données issues des systèmes d'information opérationnels dans le but de faciliter la prise de décision » (**?**).

Donc, un système décisionnel manipule des données opérationnelles avec différents moyens de collecte, de stockage et d'analyse pour soutenir le processus d'aide à la décision.

1.2 Historique des Systèmes Décisionnels

Dès les années 60, les données informatisées dans les organisations ont pris une importance qui n'a cessé de croitre. Les systèmes informatiques gérant ces données avaient pour fonction essentielle d'automatiser les processus de production de l'information afin de réduire les ressources consommées en diminuant les tâches redondantes (**?**).

Au début des années 70, les systèmes interactifs d'aide à la décision basés sur des technologies de recherche opérationnelle, simulation et optimisation sont apparus et ont été à l'origine de l'apparition des premiers outils informatiques d'aide à la décision qui allaient principalement s'appliquer, par un dialogue «Homme-Machine», aux processus de décisions exécutés aux niveaux hiérarchiques supérieurs (**?**). Avec l'accroissement des besoins en matière de décision, de nouveaux concepts sont apparus au début des années 90 : l'entrepôt de données (data warehouse) et les magasins de données (data mart). Ces derniers ont révolutionné les outils décisionnels en rassemblant les données dans un référentiel unique, orienté sujet, permettant une grande souplesse et précision. Une nouvelle étape est ainsi franchie dans l'informatique décisionnelle avec ces avancées technologiques : les outils informatiques d'aide à la décision, désormais appelés « Business Intelligence » (**?**).

A partir des années 90, plusieurs éditeurs de logiciels ont commencé à proposer des outils facilitant l'analyse des données pour soutenir les prises de décision comme les tableurs et des outils facilitant l'accès aux données pour les décideurs au travers d'interfaces graphiques dédiées à l'interrogation. Après, les outils ETL (Extract-Transform-Load en français Extraire-Transformer-Charger) destinés à faciliter l'extraction et la transformation de données décisionnelles ont vu la lumière. Dès la fin des années 90, les acteurs

importants tels que Microsoft [1], Oracle [2], IBM [3], SAP [4] sont intervenus sur ce nouveau marché en faisant évoluer leurs outils et en acquérant de nombreux logiciels spécialisés. Cette dernière décennie a été marquée par l'émergence de plusieurs outils de business intelligence issus du monde du logiciel libre (Open Souce), qui ont atteint aujourd'hui une certaine maturité (SpagoBI [5], Talend [6], Jasper [7]) (**?**).

1.3 Objectifs et finalités des systèmes d'informations décisionnels

L'acquisition d'un système d'information décisionnel est un objectif souhaité et partagé par tous les dirigeants des entreprises malgré la variété de leurs champs d'action. Dans (**?**), on recense les objectifs suivants :

- **Faciliter et soutenir la prise de décision.**
- **Améliorer les performances décisionnelles de l'entreprise.**
- **Accessibilité facile et rapide aux informations.**
- **Cohérence des informations :** les données du système sont crédibles et de qualité.
- **Adaptation aux changements :** Les données existantes doivent généralement rester inchangées. Lorsque la technologie ou les besoins changent, les données doivent être changées en tenant au courant tous les utilisateurs du système.
- **Présentation des informations à temps :** Les informations doivent être disponibles au bon moment afin de réagir rapidement.
- **Protection et sécurisation des informations :** Le système doit permettre le contrôle d'accès à ces informations confidentielles.

1. https ://www.microsoft.com/fr-fr/
2. http ://www.oracle.com/fr/index.html
3. http ://www.ibm.com/dz-fr/
4. http ://go.sap.com/index.html
5. http ://www.spagobi.org/
6. https ://fr.talend.com/
7. https ://www.jaspersoft.com/fr

1.4 Système Transactionnel VS Système Décisionnel

Les systèmes transactionnels, d'après (**?**), sont : « Des applications opérationnelles qui capturent les transactions de l'entreprise ».

Le système transactionnel représente donc les tâches, quotidiennes, répétitives, et atomiques effectuées par les employés de l'entreprise.

Les systèmes opérationnels ne peuvent pas répondre aux besoins des décideurs qui veulent des informations synthétisées, et cela à cause du grand volume de données brutes. A cet effet, on a assisté une naissance prématurée des systèmes décisionnels.

Le tableau suivant 1.1 est un comparatif entre les deux systèmes selon les critères d'usage et de données.

Critère	Système Transactionnel	Système Décisionnel
Utilisateurs	Les utilisateurs sont les rouages de l'entreprise.	Les utilisateurs observent les rouages de l'entreprise.
	Beaucoup d'utilisateurs.	Peu d'utilisateurs.
	Niveau des besoins analytiques est bas.	Niveau des besoins analytiques est élevé.
	Exécution d'un grand nombre de fois la même tâche.	Liste uniquement les données qui sont souvent récapitulées.
Données	Orientées applications.	Orienté thèmes et sujets.
	Normalisées.	Agrégées.
	Données opérationnelles : provenant des sources elles-mêmes, exhaustives et détaillées.	Données consolidées : les données de l'OLAP proviennent des différents OLTP, elles sont agrégées et résumées.
	Dynamiques.	Stables.
	Courantes.	Historiques.

TABLE 1.1 – Système transactionnel VS système décisionnel (**?**)

1.5 Architecture d'un système décisionnel

Le processus d'un système décisionnel consiste à récupérer des données brutes issues des différentes sources, internes ou externes, à les transformer en information afin de les diffuser sous forme de rapports ou de tableaux de bord (**?**).

Afin de mettre en œuvre ce processus, l'architecture d'un système décisionnel mise en place est composée en quatre niveaux qui sont : les sources de données, l'entrepôt et magasins de données, la phase ETL et la restitution de données.

La figure suivante illustre l'architecture d'un système décisionnel.

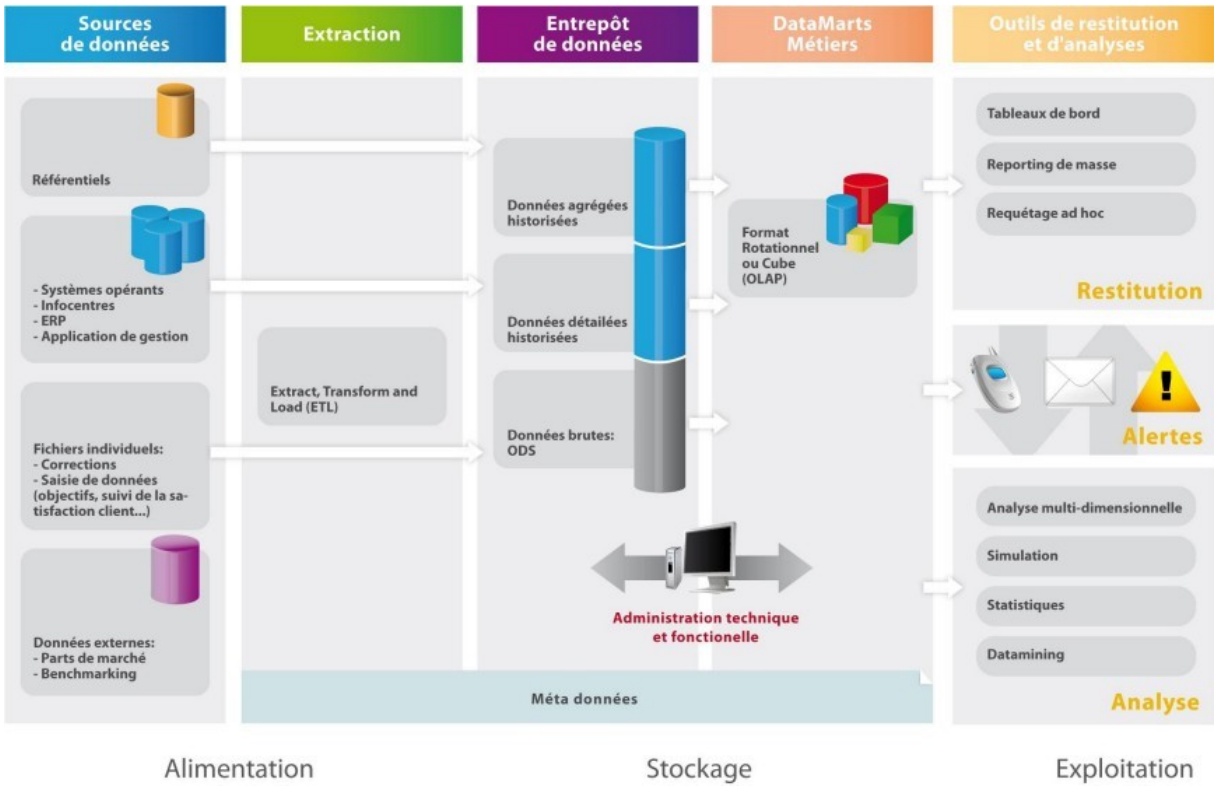

FIGURE 1.1 – Architecture d'un SID (**?**)

1.5.1 Sources de données

Les sources de données alimentent l'entrepôt de données. Elles sont regroupées en quatre catégories (**?**) :

- **Les données de production :** ce sont les données provenant des différents systèmes opérationnels, Elles sont contenues dans des bases de données transactionnelles et elles sont caractérisées par leur disparité (plusieurs technologies, environnements différents, ...).
- **Les données internes :** il s'agit des données concernant l'entreprise mais détenues par des employés ou des départements dans leurs fichiers et base de données personnels, leur acquisition est de ce fait complexe.
- **Les données archivées :** les données anciennes que le système opérationnel stocke périodiquement dans des fichiers ou base de données.
- **Les données externes :** données externes à l'entreprise, il peut s'agir de statistiques d'agence externes, d'indicateurs de performance... nécessaires pour se situer par rapport à la concurrence et repérer les tendances du marché.

1.5.2 Le processus ETL (Extraction, Transformation et Chargement)

Afin d'exploiter les données pour la prise de décision, il faut les rassembler dans une même zone. Et comme les données de l'entreprise, qu'elles soient homogènes ou hétérogènes, se trouvent dans plusieurs endroits alors on utilise un outil ETL pour les rassembler.

Selon (**?**), un ETL extrait, nettoie, et importe les données à partir de différentes sources et les charge dans un entrepôt de données.

1.5.3 Les entrepôts et magasins de données

Grâce au processus ETL, les données sont stockées et organisées dans un entrepôt de données. Le concept d'entrepôt de données a été introduit car les bases de données transactionnelles ne répondent pas aux besoins d'analyse.

1.5.3.1 Entrepôts de données

Kimball définit l'entrepôt de données comme étant « une copie des données transactionnelles d'une entreprise structurée de manière spécifique pour l'interrogation et l'analyse » (**?**).

De son coté, Bill Inmon fournit la définition suivante :

« Un entrepôt de données est une collection de données orientées sujet, intégrées, variant selon le temps et non volatiles, qui sert de support au processus de prise de décision des acteurs du management (les décideurs) » (**?**).

Cette définition met l'accent sur les caractéristiques suivantes :

— **Orientées Sujet :** les données sont liées au métier de l'entreprise et organisées par fonction.

— **Intégrée :** signifie que les données obtenues à partir de plusieurs systèmes opérationnels et externes doivent être réunies, ce qui implique la résolution de problèmes en raison de différences de définition de données et de contenu, tels que les différences de format de données et la codification des données.

— **Variant selon le temps :** toutes les données d'un entrepôt sont identifiées par des périodes temporelles spécifiques, cela signifie qu'on garde l'historique de toutes les transactions.

— **Non Volatiles :**les données dans un entrepôt sont utilisées pour les interrogations et ne peuvent pas être modifiées. Donc, les opérations de mise à jour et suppression ne sont pas autorisées, et la lecture est la seule opération permise.

Donc, un entrepôt de données est un référentiel centralisé qui stocke les données opérationnelles d'une manière spécifique et les rend disponibles et exploitables pour des fins d'analyse.

Les entrepôts sont souvent très volumineux et très complexes à concevoir, ils ont été divisés en petits entrepôts, faciles à créer et à entretenir, appelés data marts ou magasins de données.

1.5.3.2 Magasins de données (DataMart)

Un magasin de données est un extrait de l'entrepôt conforme à des besoins d'analyse particuliers et organisé selon un modèle adapté aux outils d'analyse et d'interrogation décisionnelle. Le magasin est généralement stocké au sein d'une base de données multidimensionnelle (BDM) (**?**). Cela signifie qu'un magasin concentre sur les données d'un département dans l'entreprise.

Pour une bonne gestion de données, l'entrepôt de données doit nécessairement disposer de métadonnées (données sur les données).

Les métadonnées d'un entrepôt de données se présentent sous trois catégories (**?**) :

— **Métadonnées opérationnelles :** Elles contiennent toutes les informations sur les sources de données opérationnelles.

— **Métadonnées d'extraction et de transformation :**Elles contiennent les fréquences d'extraction, les méthodes d'extraction, les règles d'extraction, les informations sur toutes les transformations opérées sur les données.

— **Métadonnées de l'utilisateur final :** Elles permettent à l'utilisateur final de retrouver l'information dans l'entrepôt de données.

1.5.3.3 Types de données stockées dans un entrepôt de données

Un entrepôt de données est articulé en quatre catégories de données, organisées selon un axe historique et un axe synthétique.

L'axe synthétique établit une hiérarchie d'agrégation comprenant (**?**) :

— **Les données détaillées :** Elles proviennent des systèmes opérationnels et représentent les évènements les plus récents. Seules les données qui servent au processus décisionnel sont stockées dans l'entrepôt.
— **Les données agrégées :** Elles synthétisent les données détaillées et correspondent à des éléments d'analyse représentatifs des besoins utilisateurs. Elles ont pour but de faciliter la navigation suivant les besoins décisionnels et la restitution d'un résultat d'analyse ou de synthèse.
— **Les données fortement agrégées :** Elles synthétisent les données agrégées à un niveau supérieur.

L'axe historique comprend **les données détaillées historisées** représentant les évènements passés.

1.5.3.4 Modélisation des entrepôts de données

La modélisation d'un entrepôt de données est le processus permettant de mettre en place un entrepôt de données. Il existe plusieurs approches pour mettre en place un entrepôt de données. Cependant, trois approches sont les plus répandues. Il s'agit de l'approche d'Inmon, l'approche de Kimball et l'approche Data Vault.

L'étude et la comparaison de ces trois approches sont abordées dans les deux prochains chapitres.

1.5.4 Outils d'analyse et de restitution de données

Le dernier niveau correspond à la partie visuelle et accessible par les utilisateurs du système décisionnel. Il est l'élément le plus important pour l'utilisateur, car il lui permet d'exploiter, d'analyser et de restituer les données stockées. On distingue à ce niveau plusieurs types d'outils différents parmi eux :

1.5.4.1 Tableau de bord

Plusieurs auteurs ont donné une définition au tableau de bord. Nous citons celle donnée par (**?**) :

Le tableau de bord est « un outil d'aide à la décision et un ensemble d'indicateurs peu nombreux (cinq à dix) conçus pour permettre aux gestionnaires de prendre connaissance

de l'état et de l'évolution des systèmes qu'ils pilotent et d'identifier les tendances qui les influencent sur un horizon cohérent avec leurs fonctions ».

Les décideurs au sein d'une entreprise ont un ou plusieurs objectifs à atteindre. Le décideur à l'aide du tableau de bord adapte ses ordres et mesures de fonctionnement effectués pour réduire au maximum les écarts constatés entre les mesures et les objectifs à atteindre.

1.5.4.2 Reporting

Le reporting est l'ensemble des comptes rendus permettant à une entreprise de suivre son activité. Cela permet à l'entreprise de s'évaluer grâce à la création périodique de rapports et de bilans analytiques sur son activité. Ces rapports sont souvent destinés au manager ou au corps exécutif (**?**). Le but de ces rapports et bilans réguliers est de faire un point ponctuel sur la stratégie de l'entreprise et ainsi permettre d'évaluer les moyens mis en œuvre. Mais ils fournissent également une aide à la décision pour les choix stratégiques et économiques de l'entreprise (**?**).

1.5.4.3 Analyse OLAP

Une caractéristique fondamentale du modèle multidimensionnel de l'entrepôt de données est qu'il permet de visualiser les données à partir de plusieurs points de vue et à plusieurs niveaux de détail.

Les opérations OLAP permettent de matérialiser ces perspectives ainsi que les niveaux de détails en exploitant les dimensions et leurs hiérarchies, fournissant ainsi un environnement d'analyse interactive des données.

D'après (**?**), « les magasins de données reposent sur une modélisation multidimensionnelle des données extraites de l'entrepôt. Ceci permet de les représenter sous la forme de points dans un espace à plusieurs dimensions avec la métaphore de cube ou d'hypercube de données. Cette modélisation permet l'expression d'analyses en ligne (OLAP) multidimensionnelles. »

1.5.4.4 Data Mining

DM est l'exploration de données historiques (généralement de grande taille) à la recherche d'un modèle cohérent et /ou d'une relation systématique entre les variables. Elle

est ensuite utilisée pour valider les résultats en appliquant les modèles détectés à de nouveaux sous-ensembles de données (**?**).

Selon (**?**), l'exploration de données peut être divisé en deux tâches : tâches prédictives et tâches descriptives. Le but ultime de l'exploration de données est la prédiction ; par conséquent, l'extraction de données prédictives est le type le plus commun de l'exploration de données et est celui qui a le plus d'application à des entreprises car les rendements d'extraction de données peuvent ouvrir les yeux d'une entreprise à de nouveaux marchés, de nouvelles façons d'atteindre les clients et de nouvelles façons de faire des affaires.

Conclusion

Les systèmes décisionnels occupent une place importante dans les entreprises vu qu'ils soutiennent les dirigeants pour prendre des décisions au bon moment.

L'entrepôt de données est le cœur de l'architecture décisionnelle. Il est alimenté à partir de données sources grâce aux outils d'ETL. Ces données deviennent alors exploitables à l'aide d'outils d'analyse et de restitution.

Le chapitre suivant est consacré à l'étude des approches de modélisation des entrepôts de données.

Chapitre 2

Approches de modélisation des entrepôts de données

Introduction

Afin de construire un entrepôt de données pour une entreprise, le choix des méthodes et outils de conception et de maintenance est une étape primordiale et très importante.

Depuis des décennies, deux approches classiques étaient en rivalité rude quant à la modélisation des entrepôts de données. L'approche de modélisation par sujet et normalisation préconisée par son inventeur Inmon et l'approche de modélisation dimensionnelle de Kimball.

Ces dernières années, une nouvelle approche est entrée fortement en compétition et attire l'attention des entreprises jour après jour. Elle s'agit de l'approche Data Vault développée par son inventeur Dan Linstedt. Dans ce chapitre, nous allons étudier chaque approche : sa philosophie, son architecture et la méthodologie pour la mettre en place.

2.1 Approche d'Inmon

Cette approche est créée par Bill Inmon [1] dans les années 90 pour répondre au besoin des entreprises et leur permettre de développer leurs systèmes décisionnels. Elle permet le stockage de l'intégralité des évènements de l'entreprise et engage des ressources et moyens importants pour la réaliser.

1. http ://www.inmoncif.com/about/

L'approche d'Inmon est basée sur les schémas Entité-Association des systèmes opérationnels. Les données de l'entreprise sont chargées sans connaitre à priori les besoins des utilisateurs (**?**), C'est pourquoi cette approche est qualifiée de « piloter par les données ».

2.1.1 Philosophie de l'approche

Inmon, informaticien américain et inventeur de l'approche, définit l'entrepôt de données comme étant une collection de données orientées sujet, intégrées, variant selon le temps et non volatiles (Voir chapitre 01). Pour la construction d'un entrepôt de données, Inmon propose une méthodologie guidée par les données. Cette méthodologie est présentée en trois parties qui sont :

1. **Meth 1 :** pour le développement des systèmes opérationnels.
2. **Meth 2 :** pour le développement des entrepôts de données.
3. **Meth 3 :** pour la description de l'usage de l'entrepôt de données.

L'architecture de l'entrepôt proposée par Inmon [Figure 2.1] inclut tous les systèmes d'informations de l'entreprise avec leurs bases de données au lieu de ne considérer que des fragments d'information. Inmon divise l'environnement des bases de données de l'entreprise en quatre niveaux à savoir : opérationnel, atomique (Entrepôt de données), départemental (magasins de données) et individuel.

Les trois derniers niveaux constituent l'entrepôt alors que le premier niveau (opérationnel) supporte les opérations quotidiennes et contient les données transactionnelles de l'entreprise. Ces données sont ensuite transformées et chargées dans l'entrepôt de données atomique en utilisant un processus ETC (Extraction, transformation et chargement). Selon le besoin des différents départements de l'entreprise, les données détenues par le niveau départemental sont orientées sujet et toujours cohérentes car elles proviennent du même entrepôt.

Inmon considère que l'entrepôt et le magasin sont physiquement dissociés. L'entrepôt possède sa propre existence physique et il est orienté stockage, traçabilité et évolutivité par rapport à de nouveaux besoins. Les magasins possèdent aussi leur propre existence physique et proposent des structures orientées performances de restitution en réponse aux besoins exprimés par l'utilisateur.

Le dernier niveau est créé par les utilisateurs quand ils analysent et exploitent les données chargées dans les magasins de données (Analyse OLAP, Reporting, Tableaux de bord, etc.).

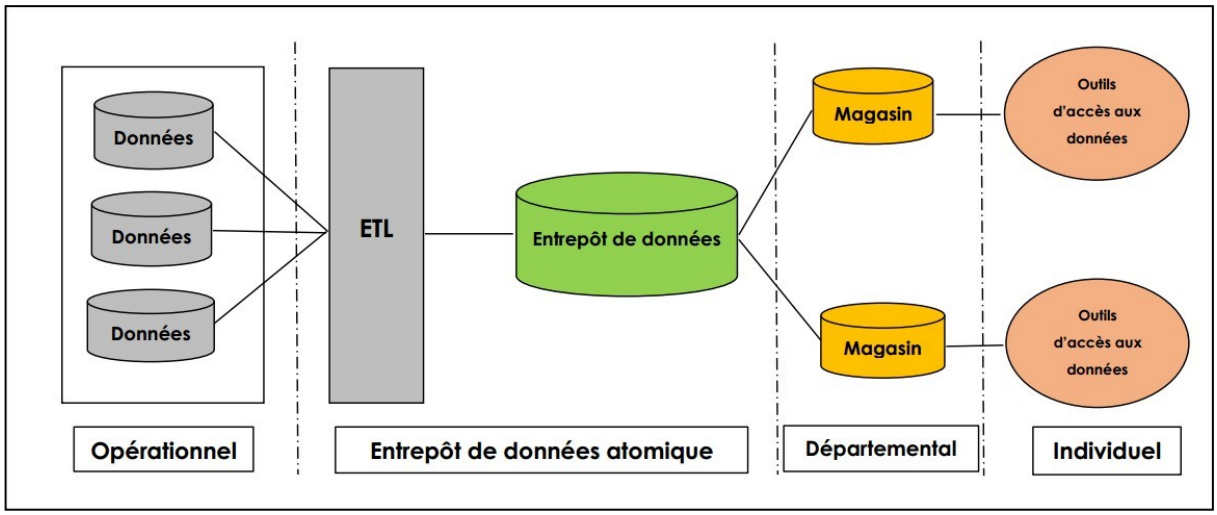

FIGURE 2.1 – Architecture de l'entrepôt de données. Adaptée de (**?**)

L'approche d'Inmon est très complexe et requiert une connaissance approfondie de l'entreprise. Elle maximise la participation des professionnels des technologies de l'information. Cependant, les utilisateurs finaux sont faiblement impliqués dans le processus d'élaboration de l'entrepôt de données.

Dans notre étude, nous nous intéressons à la deuxième partie de la méthodologie (Meth 2) qui présente les étapes de développement d'un entrepôt de données.

2.1.2 Développement de l'entrepôt de données

Pour mettre en œuvre un entrepôt de données, Inmon propose une méthodologie de développement spirale (Meth 2) qui requiert les procédures suivantes :

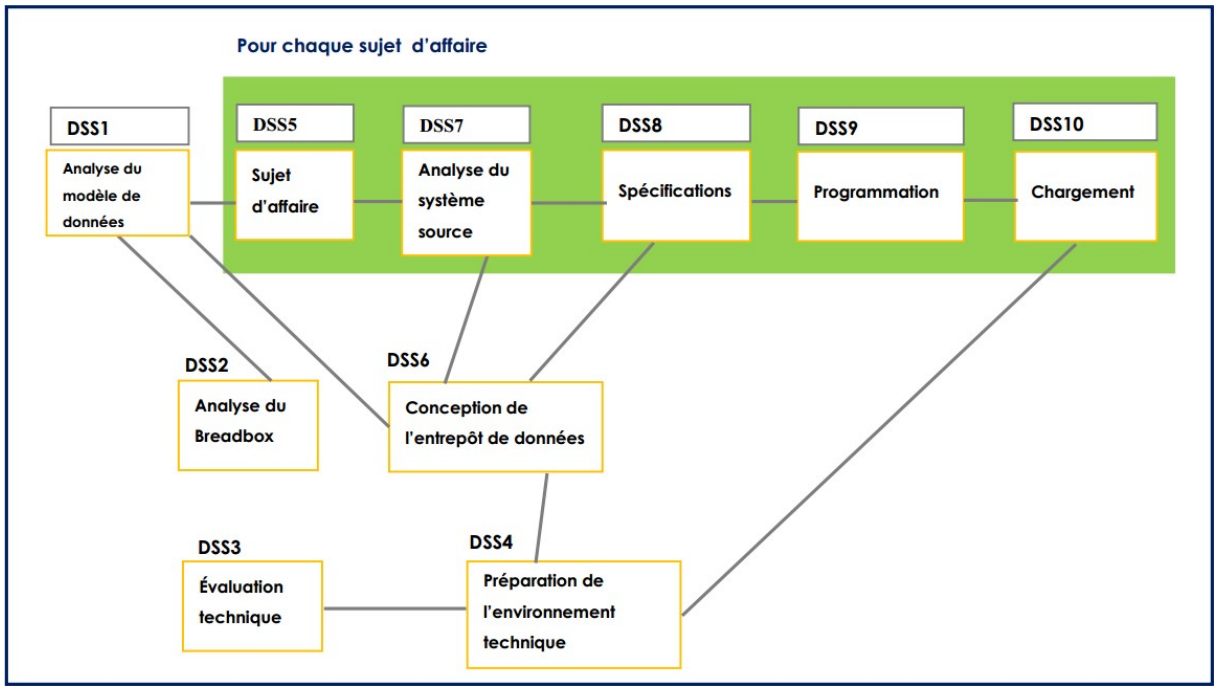

FIGURE 2.2 – Meth 2 d'Inmon. Reproduite (**?**)

2.1.2.1 Analyse du modèle de données

Le modèle de données de l'entreprise est le point de départ pour la mise en œuvre de l'entrepôt de données. A cet effet, il est important qu'il soit complet.

Pour établir ce modèle, Inmon propose trois niveaux de modélisation de données à savoir :

2.1.2.1.1 Diagrammes d'Entité-Association(ERD) : Pour la modélisation à haut niveau d'abstraction de données. Ce modèle, comme dans le développement des bases de données opérationnelles, consiste à identifier dans un premier temps les sujets d'affaires de l'entreprise(les entités) et dans un deuxième temps, identifier les relations qui existent entre ces entités.

Pour chaque département de l'entreprise qui veut utiliser l'entrepôt de données, on crée des ERD et l'ensemble constituera l'ERD de l'entreprise (**?**).

La figure suivante montre un exemple d'un modèle de données de haut niveau.

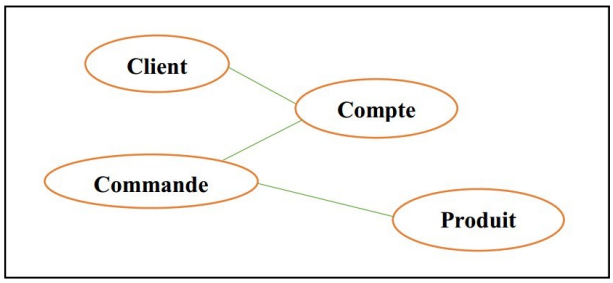

FIGURE 2.3 – Exemple d'un modèle ERD

2.1.2.1.2 Data Item Set : Le deuxième niveau de modélisation représente l'endroit où on trouve le maximum d'informations sur le modèle de données de l'entreprise. Pour chaque sujet d'affaire identifié dans l'ERD de l'entreprise, un DIS (acronyme de Data Item Set) sera créé. Ce niveau de modélisation contient des clés, des attributs, des sous-types, des groupements d'attributs, et des connecteurs.

Un DIS est constitué de quatre parties :

1. **Premier groupement de données :** le premier groupement de données existe seulement pour chaque sujet d'affaire (entité). Il contient les attributs (clé primaire et autres attributs) qui existent une seule fois pour chaque entité, Autrement dit, les caractéristiques qui identifient seulement l'entité qu'on est en train de modéliser. Si l'entité ou notre sujet d'affaire est 'Personne' alors le premier groupement va contenir le numéro de sécurité social de la personne(NSS) et autres attributs en relation directe avec l'entité.
2. **Deuxième groupement de données :** contient les données qui peuvent exister plusieurs fois pour chaque entité ou sujet d'affaire. Supposons que l'entité modélisée est 'Personne' alors il va y avoir multiples ensembles de données concernant son éducation, ses fils, etc. Ces données peuvent être des clés, clés étrangères ou autres attributs, mais la clé à ce niveau peut ne pas être unique.
3. **Connecteur :** le connecteur met en évidence les relations entre les sujets d'affaires.
4. **Types de données :** cette partie montre les différents types de l'entité, c'est-à-dire, pour une entité donnée on détermine les types qui peuvent exister selon des critères de classification. Par exemple, Si l'entité est 'Client' et le critère est la fidélité alors les différents types de données peuvent être client fidèle et client potentiel.

La figure suivante montre un exemple d'un modèle DIS.

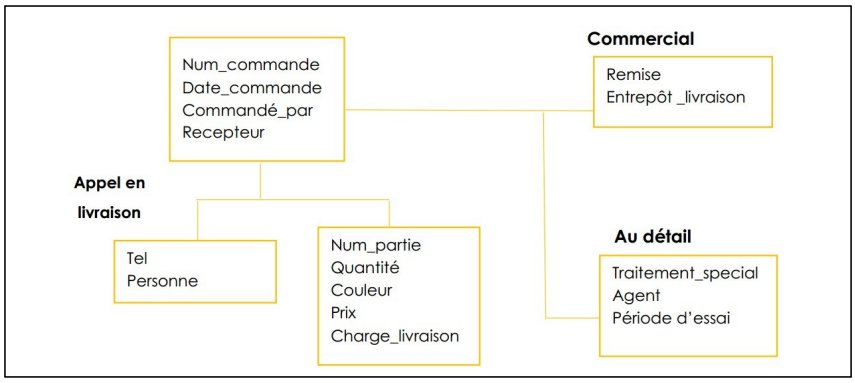

FIGURE 2.4 – Exemple d'un modèle DIS. (**?**)

Nous constatons dans l'exemple le DIS relatif à l'entité Commande. Une commande a un ensemble d'informations basiques comme le numéro de la commande et la date. Il y a différents types de commande comme les commandes commerciales et les commandes au détail. Chaque commande peut avoir multiples occurrences des éléments qui ont été commandés, et peut avoir ou non des informations spécifiques comme la personne à appeler à la livraison.

2.1.2.1.3 Le modèle physique C'est le dernier niveau de modélisation. Il est créé en se basant sur le modèle de données du second niveau. Pour chaque partie du DIS, il y aura un modèle de données physique unique et séparé. Ce modèle ressemble aux tables relationnelles.

2.1.2.2 Analyse du breadbox

L'analyse du breadbox s'inscrit dans le processus des grandes estimations du système décisionnel. Parfois, le volume de l'entrepôt de données pose problème, c'est pourquoi il est préférable de le connaitre dès le début. Cette analyse permet de mettre en lumière ce problème et essaie d'estimer le volume de l'entrepôt de données. Elle considère principalement les niveaux de granularité de données (niveau de détail) car c'est la raison majeure qui influe sur la taille de l'entrepôt.

2.1.2.3 Evaluation technique

« Les exigences techniques pour le management et le traitement de données dans l'environnement opérationnel sont totalement différentes de celles pour manager l'entrepôt de données.» (**?**).

Cette phase consiste à établir une configuration architecturale de l'entrepôt de données. Le format d'évaluation dépend des décisions de l'entreprise et de ce qu'elle détient en termes d'équipements. Aucun format d'évaluation n'est donc prédéfini. Néanmoins pour un entrepôt couronné de succès, ses fonctionnalités doivent respecter les standards suivants :

— Manipulation d'un énorme volume de données.
— Flexibilité d'accès aux données.
— Organisation des données de l'entrepôt en respectant le modèle de données de l'entreprise.
— Capacité d'interaction avec une multitude de technologies.

2.1.2.4 Préparation de l'environnement technique

« Une fois la configuration architecturale définie, on identifie techniquement comment l'architecture peut être accommodée. » (**?**). (**?**) liste les problèmes devant être traités :

— Comment le réseau des technologies d'information de l'entreprise va être affecté par le trafic accru à cause de l'entrepôt de données ?
— Comment minimiser les conflits entre les applications existantes ?
— Quel est le type de connectivité réseau nécessaire ?
— Quel est le volume de traitement prévu ?

2.1.2.5 Analyse des sujets d'affaires

Cette étape est considérée comme la première étape expérimentale dans le processus de développement de l'entrepôt de données.

Le processus d'analyse des sujets d'affaires est fortement lié à la compréhension et à la maitrise du modèle de données de l'entreprise vu qu'on se base principalement sur ce dernier pour dégager les éléments de données qui vont être chargées dans l'entrepôt de données. Il est conseillé que le sujet d'affaire soit suffisamment large pour être significatif et suffisamment petit pour être applicable.

2.1.2.6 Conception de l'entrepôt de données

La conception est la phase la plus importante dans le processus de développement de l'entrepôt de données, c'est pourquoi le modèle de données et l'analyse des sujets d'affaires sont des étapes critiques pour un entrepôt réussi.

Pour construire l'entrepôt de données, un certain nombre de transformations seront faites sur le modèle de données de l'entreprise. Nous abordons dans ce qui suit les transformations qui seront appliquées :

1. **Les données purement opérationnelles sont supprimées entièrement du modèle de données de l'entrepôt :** Il s'agit de toutes les données qui n'ont pas d'intérêt pour les décideurs.

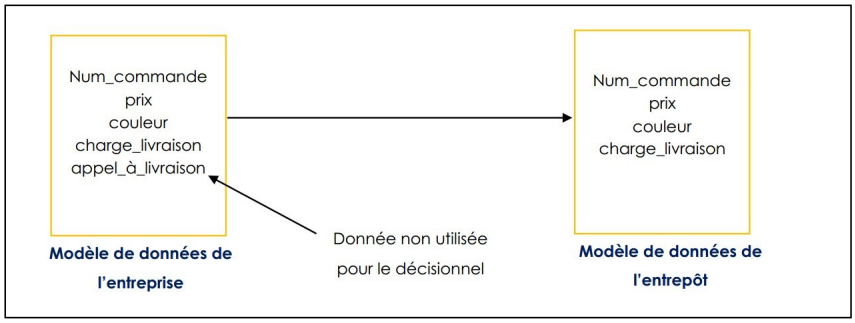

FIGURE 2.5 – Suppression de données non utilisées pour le décisionnel. (**?**)

2. **L'ajout d'un élément de temps** à la clé de l'entrepôt de données s'il n'existe pas.

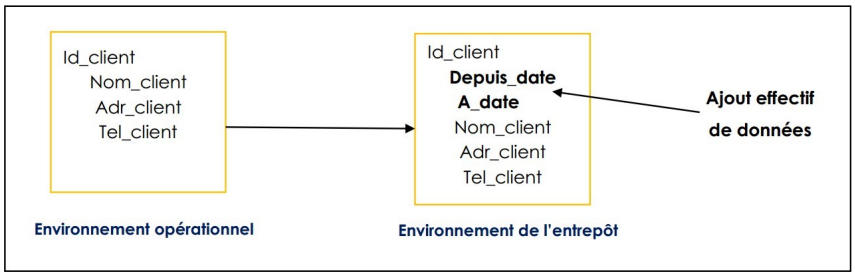

FIGURE 2.6 – Ajout de l'élément temps

3. **L'ajout de données dérivées qui sont populaires et calculées.** La figure suivante montre un exemple d'ajout de données dérivées.

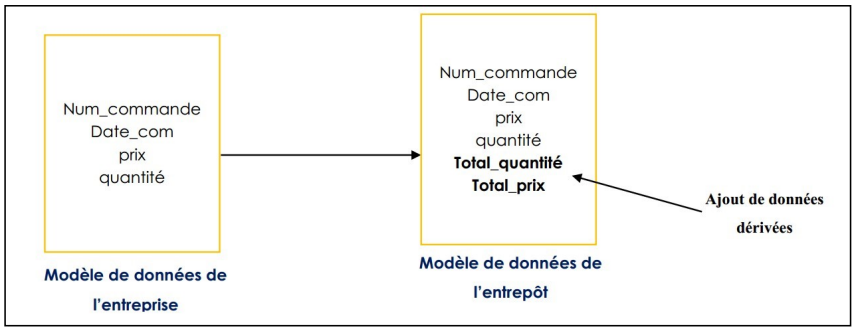

FIGURE 2.7 – Ajout de données dérivées

L'ajout de données dérivées réduit le taux de calcul requis et améliore la crédibilité de données dans l'entrepôt.

4. **Accommodation des différents niveaux de granularité de l'entrepôt de données :** Parmi les caractéristiques principales d'un entrepôt de données la granularité de données. Parfois le niveau de granularité ne change pas quand les données passent de l'environnement opérationnel à l'environnement de l'entrepôt de données. Dans d'autres cas, la granularité change et ce changement doit apparaitre dans le modèle de données de l'entrepôt de données.

   La figure suivante montre un exemple de changement du niveau de granularité.

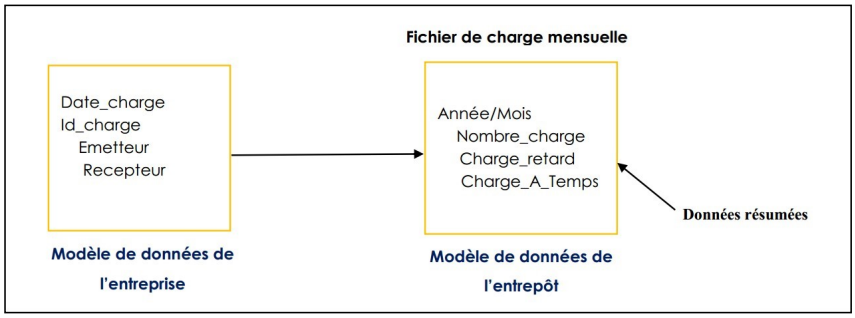

FIGURE 2.8 – Changement du niveau de granularité. (**?**)

5. **Fusion de données semblables de différentes tables en une seule table** : Cela permet de gagner en espace et en performance, mais la fusion ne se fait pas tout le temps et les conditions suivantes sont nécessaires (**?**) :
   - Les tables partagent une clé commune.
   - Les données des différentes tables sont fréquemment utilisées ensemble.
   - Le processus d'insertion de données est presque le même.

Si aucune de ces conditions n'est présente alors aucune fusion de tables ne sera faite.

6. **Création des tableaux de données :** la redondance de données dans le modèle de données de l'entreprise n'est pas permise. Cela est dû au fait que les données sont normalisées. Par contre, l'entrepôt de données peut et devrait contenir des données répétées.

7. **Organisation de données selon leur degré de stabilité :** C'est la dernière transformation à faire sur les données et qui vise à organiser les données dans l'entrepôt selon leur tendance au changement. L'organisation de données dans l'entrepôt est optimale quand les données d'une table changent tous du même rythme. Certaines tables contiennent des données qui changent lentement alors que d'autres contiennent des données qui changent rapidement.

   La figure suivante montre un exemple de cette transformation.

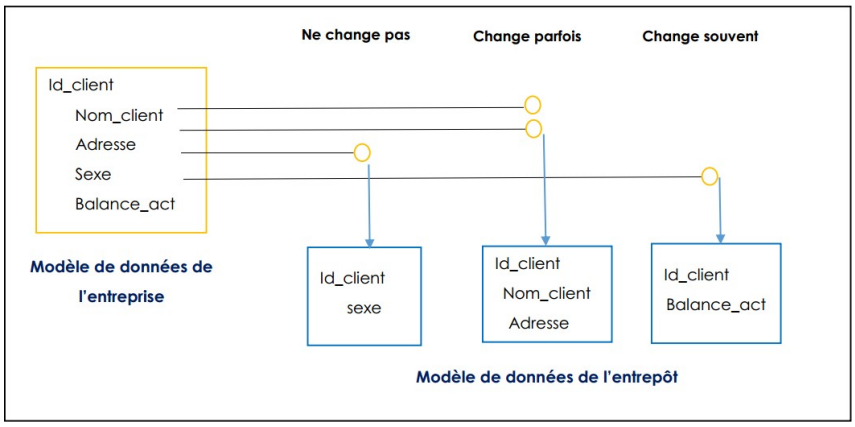

FIGURE 2.9 – Division de données selon leur degré de changement. (**?**)

2.1.2.7 Analyse des systèmes sources

Les sources de données de l'entreprise sont multiples et variées. Dans cette phase, l'équipe de projet est amenée à déterminer les sources de données qui vont alimenter l'entrepôt de données pour chaque sujet d'affaire identifié.

Pour déterminer les meilleures sources de données, Inmon propose les critères suivants :

— Quelles sont les données les plus complètes ?

— Quelles sont les données les plus exactes ?

— Quelles sont les données les mieux appropriées ?

— Quelles sont les données qui suivent la structure du modèle de données ?

Inmon suggère dans l'analyse des sources de données de traiter les problèmes suivants :

— comment choisir à partir de plusieurs sources de données ?

— Comment procéder dans le cas où il n'y a pas de source ?

— Quelles sont les transformations à faire sur les données qui vont alimenter l'entrepôt de données ?

Si on suit ces étapes proprement, alors le système source va être complet, approprié, exact et accessible. Il va aussi suivre la structure de l'entrepôt de données (**?**).

2.1.2.8 Programmation

Cette phase inclut toutes les activités classiques de programmation comme :

— Pseudo-code.

— Codage.

— Compilation.

— Tests.

A l'issue de cette étape, on obtient un code efficace, exact et complet.

2.1.2.9 Chargement

Cette étape correspond au chargement de données dans l'entrepôt de données. Dans un premier temps, on va charger juste la fraction de données dont on a besoin à ce stade.

Le chargement de données dans l'entrepôt ne se fait pas directement, il est précédé par un ensemble d'opérations telles que le nettoyage et le filtrage.

Le chargement d'un petit volume de données permet de faire les changements sur les données rapidement et facilement. Par contre, le chargement d'un volume important diminue considérablement la flexibilité de l'entrepôt.

Une fois les données chargées, les utilisateurs finaux peuvent utiliser les données stockées dans l'entrepôt et communiquer leurs réactions à l'équipe de développement, ce qui permettra le réajustement du système selon les exigences et les besoins de l'organisation. Ce n'est qu'après que le chargement d'un grand volume de données peut être envisagé.

2.2 Approche de Kimball

Kimball [2] a créé son approche dans les années 90 en proposant une nouvelle architecture, nouvelle vision et une modélisation novatrice de l'entrepôt de données.

Cette approche est basée sur le concept de la modélisation dimensionnelle. Kimball oppose la philosophie d'Inmon quant à l'isolation des utilisateurs finaux dans le processus d'élaboration de l'entrepôt. En effet, son approche implique fortement les utilisateurs finaux dès les premières phases du projet, C'est pourquoi cette méthode est appelée « piloter par les besoins utilisateurs ».

2.2.1 Philosophie de l'approche

Ralph Kimball propose une vision complètement différente pour les entrepôts de données. Il considère que l'entrepôt de données peut être vu comme un ensemble de magasins de données cohérents entre eux, s'appuyant sur des dimensions conformes partagées (**?**).

L'intérêt de l'approche de Ralph Kimball est la rapidité de mise en place d'une solution décisionnelle adaptée aux besoins décrits. Ainsi, on gagne en temps de conception, réalisation, et investissements engagés.

Le modèle de Kimball diffère dans plusieurs points importants de l'approche traditionnelle de base de données relationnelle (**?**). Une différence importante est que les entrepôts de données construites avec le modèle Kimball commencent par les tables au lieu des entités comme le modèle E/A. Cette question est examinée dans la section 2.2.3.

Une autre différence importante est que l'architecture globale dispose de plusieurs bases de données qui sont censés être hautement interopérables. Le bus de données est la principale caractéristique de conception qui rend cela possible (**?**).

Plus de détails sur le bus de données et les dimensions conformes, veuillez voir respectivement les deux sections 2.2.3.3.3 et 2.2.3.3.4

La figure suivante illustre l'architecture de l'entrepôt selon la vision de Kimball.

2. http ://www.kimballgroup.com/about-kimball-group/

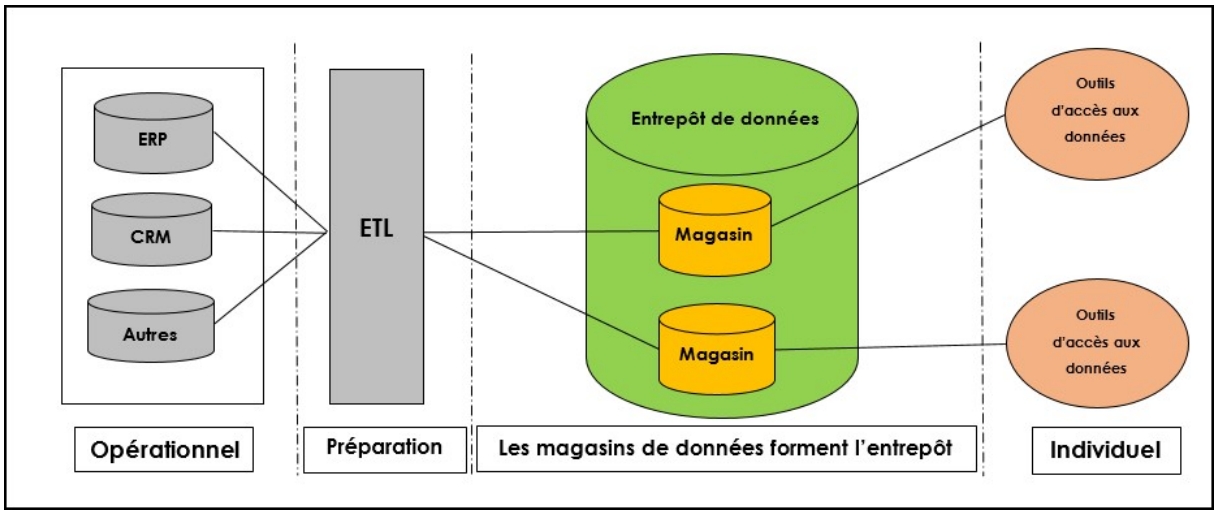

FIGURE 2.10 – L'architecture de l'entrepôt de données.

2.2.2 La modélisation dimensionnelle

La modélisation dimensionnelle de Kimball est une technique de conception logique qui vise à présenter les données dans un cadre normalisé intuitive et permet un accès de haute performance. Elle est fondamentalement basée sur les dimensions et les faits. Chaque modèle dimensionnel est composée d'une table avec une clé, appelé la table de faits, et un ensemble de petites tables appelées tables de dimension.

Pour comprendre toute la démarche de Kimball, nous allons la détailler à travers le cycle de vie Kimball (la figure ci-après), qui représente toutes les étapes garantissant une mise en place d'un entrepôt de données à partir de la planification, la définition de besoins et le passage par la modélisation jusqu'à la maintenance de l'entrepôt.

2.2.3 Présentation du cycle de vie de Kimball

Selon Kimball, la finalité d'un entrepôt de donnée est de « fournir des informations pour soutenir la prise de décisions dans une entreprise» (**?**). De ce fait, l'entrepôt de données est une base de données dédiée et utilisée dans le cadre de la prise de décision et de l'analyse décisionnelle.

Pour pouvoir concevoir un entrepôt de données selon l'approche de kimball, il est primordial de maitriser la modélisation dimensionnelle.

Cette modélisation et d'autres notions nécessaires sont expliquées et détaillées dans le cycle de vie de kimball. La figure suivante représente le cycle de vie dimensionnel de

Kimball illustrant les étapes nécessaires à la modélisation multidimensionnelle.

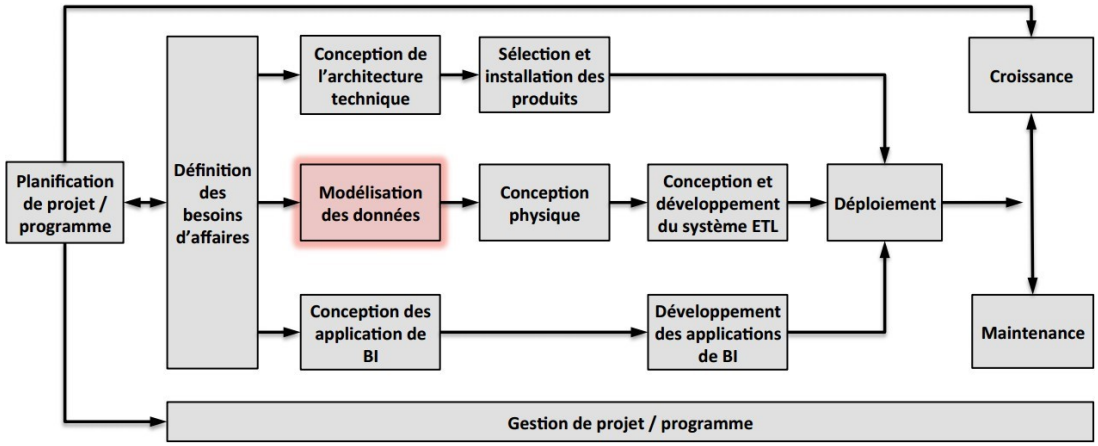

FIGURE 2.11 – L'architecture de l'entrepôt de données.(**?**)

Le cycle contient treize étapes. Nous constatons que la modélisation dimensionnelle est au cœur de ce cycle. Elle vient juste après la planification et la définition de besoins. Dans ce qui suit, nous détaillons les étapes du cycle de vie une par une.

2.2.3.1 Planification du projet

Le cycle de vie commence par la planification du projet. A ce stade on définit plusieurs points critiques qui anticipent l'échec ou le succès du projet (**?**)) :

— l'étendue du projet de l'entrepôt de donnée.

— l'appréciation du niveau de maturité de l'organisation face à ce type d'approche.

— la justification fonctionnelle.

— les coûts associés.

À partir de là, la planification du projet se concentre sur l'affectation des tâches, à leur durée et à leur séquencement. Le planning qui en découle identifie toutes les tâches associées au cycle de vie dimensionnel et mentionne les ressources impliquées. La planification du projet dépend des besoins, comme l'indique la flèche à double sens reliant ces deux activités.

2.2.3.2 Définition des besoins

Toutes les activités du cycle de vie est basée essentiellement sur la définition des besoins (**?**). Cette méthodologie de collecte concernant les besoins de l'entreprise est différente des méthodologies traditionnels pilotée par les données (**?**). Les analystes DW doivent

comprendre les facteurs clés de l'entreprise afin de traduire avec succès les besoins de l'entreprise dans des considérations de conception (**?**).

Voici les étapes suggérées par le cycle de vie Kimball concernant la collecte des besoins de l'entreprise (**?**) :

— Préparer, conduire et résumer les interviews

— Examiner les résultats des interviews

— Etablir des priorités et l'enchaînement des étapes

— Exigences sur le niveau de finalité des projets

— Gestion des risques.

Une fois que tous les besoins de l'entreprise sont finalisés, les différents groupes de l'équipe peuvent commencer le travail sur des différentes parties ; par exemple, un groupe peut commencer le travail sur la conception de l'architecture technique, et un autre groupe peut commencer le travail sur la modélisation dimensionnelle.

2.2.3.3 Modélisation dimensionnelle des données

Popularisée par Ralph Kimball dans les années 90, cette modélisation est aujourd'hui reconnue comme la modélisation la plus appropriée aux besoins d'analyse et de prise de décision (**?**).

La structuration et l'organisation des données offertes par la modélisation dimensionnelle facilitent la conceptualisation, la visualisation et l'analyse des données. Une vision multidimensionnelle des données met en avant le sujet analysé et les différentes perspectives d'analyse (**?**).

Ce modèle multidimensionnel contient plusieurs concepts fondamentaux : la table de fait, les dimensions, les dimensions conformes ainsi que la matrice en bus. Nous les définissons dans cette section.

2.2.3.3.1 La table de fait : Une table de fait est une table qui contient les données observables (les faits) que l'on possède sur un sujet et que l'on veut étudier, selon divers axes d'analyse (les dimensions). Les « faits », dans un entrepôt de données, sont normalement numériques, puisque d'ordre quantitatif. Il peut s'agir du montant en argent des ventes, du nombre d'unités vendues d'un produit,...etc .

FIGURE 2.12 – La table de fait.(**?**)

La figure précédente montre une table de fait possédant des clés étrangères (FK) qui se connectent à des clés primaires de dimension. Par exemple, la clé de produit dans la table de faits correspond toujours à une clé de produit spécifique dans la table de dimension du produit. Le "Montant Ventes" représente le fait à étudier.

2.2.3.3.2 La dimension : Une dimension est une table qui contient les axes d'analyse (les dimensions) selon lesquels on veut étudier des données observables (les faits) qui, soumises à une analyse multidimensionnelle, donnent aux utilisateurs des renseignements nécessaires à la prise de décision (**?**).

On appelle donc « dimension » un axe d'analyse. Il peut s'agir des clients ou des produits d'une entreprise, etc (**?**).

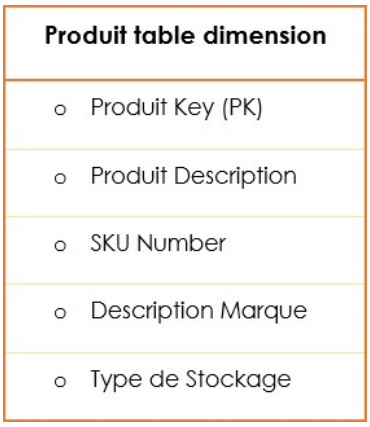

FIGURE 2.13 – La table de dimension Produit.(**?**)

La figure montre les attributs décrivant les lignes de la table Produit. D'autres attributs peuvent être ajoutés à la dimension. Ces attributs sont descriptifs et représentent

l'information utile sur la dimension. Chaque table de dimension devrait contenir les colonnes suivantes :

- **Date effective :** c'est la date à laquelle l'enregistrement est créé.
- **Date retrait :** C'est la date à laquelle l'enregistrement est retiré.
- **Indicateur effectif :** En général est 'O' si l'enregistrement est toujours actif (Date retrait non nulle), 'N' sinon.

2.2.3.3.3 Dimension conforme : On parle de dimension conforme ou partagée (également appelée master dimensions ou common reference dimensions) lorsque la dimension est utilisée par les faits de plus d'un magasin de données. L'exemple le plus courant est la dimension « Produit » qui est utilisée par différents magasins de données «Finance », « Marketing »...(**?**). L'avantage est :

- **la cohérence :** entre les différentes tables de faits.
- **l'intégration :** permet à l'entrepôt de données d'opérer comme un seul bloc uni.
- **La productivité :** favorise l'extension de l'entrepôt d'une itération de développement à l'autre.

2.2.3.3.4 L'architecture en bus : L'architecture en bus de Kimball est un élément clé de l'approche multidimensionnelle de Kimball. Introduit dans les années 1990, la technologie et l'architecture de bus décompose le processus de planification de l'entrepôt de donnée en éléments gérables en se concentrant sur les processus métier de l'organisation, ainsi que les dimensions conformes.

L'architecture matrice en bus, illustré ci-dessous, est un outil de conception clé représentant les processus métier de base de l'organisation et les dimensions conformes associés.

Dimensions Conformes / Processus Métier	Date	Product	Warehouse	Store	Promotion	Customer	Employee
Issues purchase order	X	X	X				
Receive warehouse delivery	X	X	X				X
Warehouse inventory	X	X	X		X		
Receive store delievery	X	X	X	X			X
Store inventory	X	X		X	X		
Retail sales	X	X		X		X	X
Retail sales forecast	X	X		X			

FIGURE 2.14 – Matrice de bus de Kimball (**?**)

D'une autre façon, la matrice de bus identifie les processus métiers clés d'une organisation, ainsi que leurs dimensions associées. L'influence de l'architecture en bus apparait clairement dans :

— **L'efficacité :** une seule copie d'une dimension implique moins d'entretien.

— **La cohérence :** une dimension conforme unique signifie la même chose partout où elle est utilisée.

2.2.3.3.5 Schémas pour modéliser : Le schéma en étoile est conçu avec deux types de tables : tables dimensionnelles et tables de faits. Il existe de nombreuses variantes de schéma en étoile ; par exemple, modèle avec plusieurs tables de faits et schémas de flocon de neige qui se posent lorsque une ou plusieurs dimensions ont une structure hiérarchique (**?**) . Les trois schémas détaillés ici sont les schémas en Etoiles, flocons et constellations.

Le schéma en étoile : Chaque dimension est représentée par une table de dimension et les mesures par une table de faits qui référence les tables de dimension en utilisant une clé étrangère pour chacune d'elles.

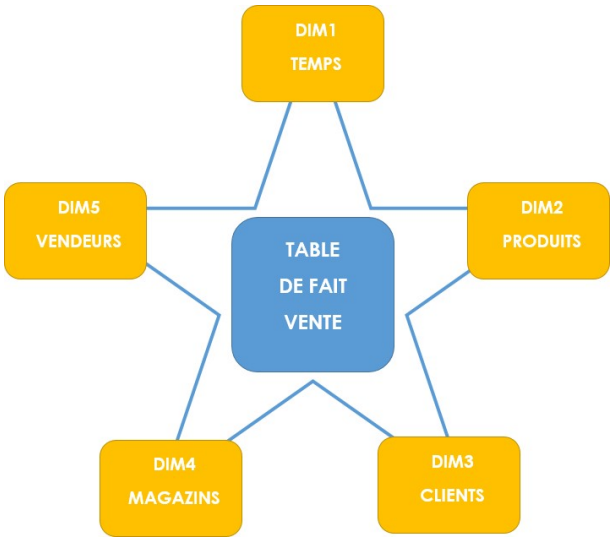

FIGURE 2.15 – Schéma en étoile.(**?**)

Le schéma en flocon de neige : C'est une extension du schéma en étoile. Le schéma en flocon est le résultat de la décomposition d'une ou plusieurs dimensions en plusieurs niveaux formant une hiérarchie. Les tables de dimensions sont ainsi éclatées en plusieurs tables.

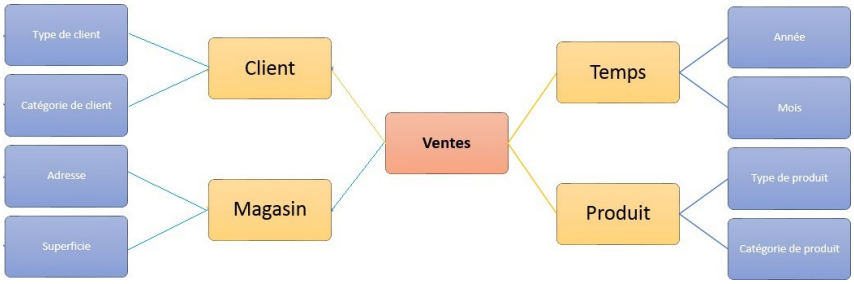

FIGURE 2.16 – Schéma en flocon de neige.(**?**)

Le schéma en constellation : C'est un schéma où plusieurs modèles dimensionnels se partagent les mêmes dimensions. Les tables de dimensions partagées par plusieurs tables de fait doivent être exactement les mêmes.

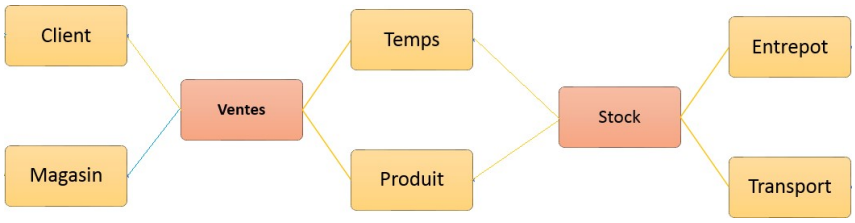

FIGURE 2.17 – Schéma en constellation. (**?**)

2.2.3.4 Processus de modélisation dimensionnelle

Le processus de modélisation dimensionnelle est extrêmement itératif (**?**). Il est nécessaire à plusieurs reprises de tester la série des schémas contre la compréhension des besoins de l'entreprise (**?**).

La modélisation dimensionnelle commence avec un modèle initiale. Cette phase donne également une liste initiale des attributs avant sa validation. Les principaux objectifs de cette phase sont (**?**) :

— créer un modèle qui répond aux besoins décisionnels de l'entreprise.

— vérifier que les données sont disponibles pour remplir le modèle.

— fournir à l'équipe d'ETL un point de départ solide et une orientation claire.

Kimball et al. ont présenté quatre différents processus de conception dimensionnelle (**?**) :

— **Choisir le processus d'affaires :** le concepteur de DW doit d'abord mettre en œuvre un seul magasin de données pour réduire la longue extraction et faciliter la tâche de conception, et ensuite appliquer la dimension conforme aux ensemble des magasins de données afin qu'on puisse plus tard les intégrer dans le bus de l'entrepôt de données.

— **Déclarer le grain :** Lorsqu'on définit le grain des tables de faits, on doit être extrêmement précis.

— **Identifier les dimensions :** Souvent le grain détermine un ensemble minimal de dimensions qui est nécessaire.

— **Identifier les faits :** La dernière étape est la sélection prudente des faits ou des mesures qui sont appropriées pour le processus de l'entreprise.

2.2.3.5 Conception du modèle physique de données

On convertit dans cette phase les données recueillies lors de la phase de conception logique dans des bases de données physiques (**?**) .

La conception physique d'une base de données définit donc les structures physiques nécessaires pour l'implémentation des bases de données logiques.

(**?**) recommande les étapes suivantes pour la conception physique :

— **L'élaboration de normes :** les conventions de nommage, les normes de localisation de fichiers, les clés primaires et les clés étrangères. . .etc.

— **Développer le modèle physique de données :** Définir la structure physique.

— **Instancier base de données relationnelle :** finalisé schéma en étoile ou schéma en flocon.

2.2.3.6 Conception et développement de la zone de préparation des données

Kimball déclare dans son ouvrage que « la plupart des systèmes ont besoin d'un ensemble de tables de transit pour supporter les processus ETL » (**?**). L'extraction, la transformation et le chargement prépare les données sources en vue de leur intégration puis de leur exploitation au sein de l'entrepôt de donnée.

Un ensemble des processus qui nettoient, transforment, combinent, archivent, suppriment les doublons, est exécuté dans cette étape. De cette définition, la zone de préparation des données ne doit offrir ni service des requêtes, ni service de présentation.

Cette étape (**?**) consiste principalement en la normalisation des données, le matching, la fusion des dossiers de la même entité qui sont tirées de sources différentes, la production de données agrégées et le traitement des faits calculés. . .etc.

2.2.3.7 Définition de l'architecture technique

Les environnements des entrepôts de données nécessitent l'intégration de nombreuses technologies. Cette étape de définition donne une vision globale de la structure de l'architecture technique à mettre en œuvre.

Après avoir terminé la phase de conception de l'architecture, un modèle d'architecture technique de haut niveau est élaboré (**?**).Le modèle de Kimball divise logiquement le

système en deux parties (la Figure) : les services de préparation (Back Room) et les services de requête (Front Room).

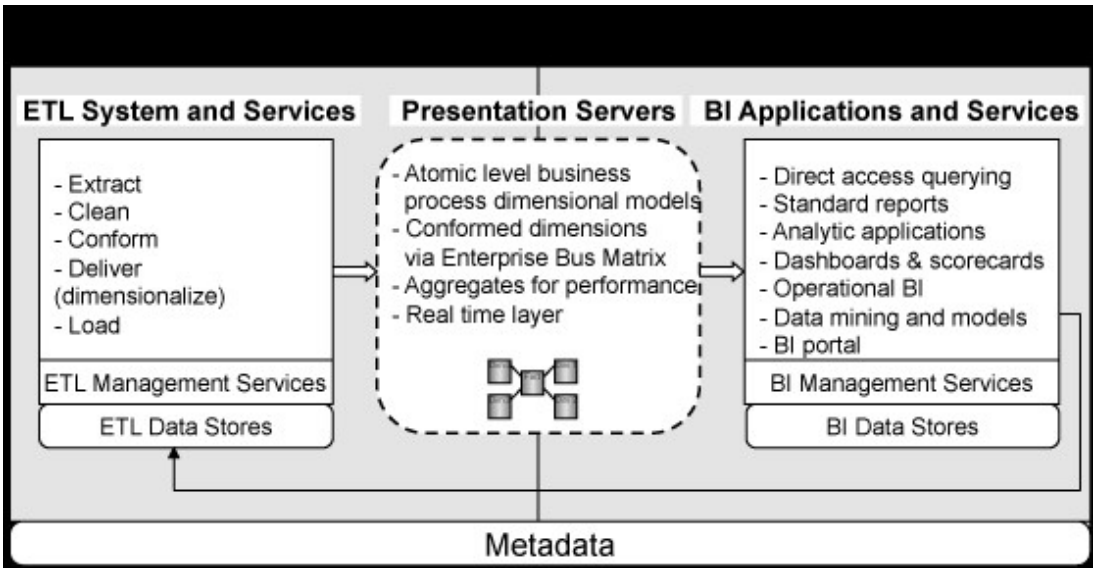

FIGURE 2.18 – L'architecture technique de Kimball.(**?**)

Le service de requête est la région où le processus d'acquisition de données est exécuté. La zone de transit de données est la zone où se passent les travaux d'extraction de données, de transformation, de chargement et de nettoyage.

L'architecture du Back Room architecture est composée de (**?**) :

— systèmes sources opérationnels, systèmes ERP, ordinateurs de bureau des utilisateurs, fournitures extérieures, et fichiers plats

— Un système ETL (Zone de préparation de données) : La zone chargée de contenir les données et d'effectuer le nettoyage et la transformation des données avant le chargement des données dans DW.

La deuxième partie, qui est appelé les services de requête (Front Room), contient des applications de la BI. Le but principal de cette partie est de fournir l'accès à l'entrepôt de données par les utilisateurs pour exploiter le DW (**?**).

L'architecture du Front Room fournit les services suivants (**?**) :

— **Requêtes, standards et rapports :** l'utilisateur peut accéder à l'information par des requêtes et des rapports.

— **Applications analytiques :** En plus de requêtes de base de données, il contient de puissants algorithmes d'analyse.

— **Les tableaux de bord :** Interfaces utilisées pour afficher les indicateurs clés de performance (KPI) textuelle et graphique.
— **BI opérationnelle :** requêtes en temps réel de l'état de fonctionnement.
— **Data mining et modèles :** l'exploitation de données historiques pour tirer des relations entre les variables.

2.2.3.8 Choix technologiques et mise en œuvre

À partir de l'étude de l'architecture technique, on évalue et sélectionne les composants, tels que la plate-forme matérielle, le système de gestion de base de données et les outils de préparation et d'accès aux données.

2.2.3.9 Développement de l'application utilisateur

Kimball recommande de définir une série d'applications standards destinée à l'utilisateur final.

2.2.3.10 Déploiement

Le déploiement est le point de convergence de la technologie, des données et des applications utilisateur accessibles à partir de postes de travail.

2.2.3.11 Maintenance et croissance

Une maintenance d'un DW doit également assurer (**?**) :
— la gestion de l'entrepôt en continu et efficacement.
— la mesure périodique de son acceptation et ses performances.

2.2.3.12 Gestion du projet

La gestion de projet garantit que les activités du cycle de vie dimensionnel restent sur la bonne voie et la bonne synchronisation entre elles. Comme le montre la figure du cycle de vie, les activités de gestion de projet sont étalées tout au long du cycle de vie.

La gestion de projet de Kimball concerne le contrôle de l'état d'avancement du projet, la détection et la résolution des problèmes et le contrôle des changements, afin de rester dans la limite des objectifs et du périmètre.

Après avoir identifié les nouveaux besoins des utilisateurs, on revient au début du cycle de vie, en prenant appui sur ce qui a déjà été mis en place dans l'environnement de l'entrepôt de données et en se penchant sur les nouveaux besoins.

2.3 Approche Data Vault

Au début des années 2000, Dan Linstedt [3] est entré dans la compétition en proposant une troisième approche dite « modélisation Data Vault » (par voûtes de données). Cette modélisation est qualifiée de mitoyenne située entre les deux approches de Kimball et Inmon. Elle est particulièrement adaptée à l'audit de données, à la traçabilité de la donnée et à la résistance au changement de la structure de données.

Cette approche palie plusieurs problèmes auxquels font face les deux approches classiques. Le problème majeur est le re-engineering en cas du changement du business

2.3.1 Définition du Data Vault

Dan Linstedt, informaticien américain et inventeur de l'approche, définit le Data Vault comme étant « un ensemble de tables normalisées orientées détail, suivant l'historique, liées de manière unique et qui supportent un ou plusieurs secteurs fonctionnels de l'entreprise (**?**). Autrement dit, c'est une approche mitoyenne qui prend le meilleur de la 3éme forme normale (3NF) d'Inmon et le schéma en étoile de Kimball.

2.3.2 Philosophie de l'approche

Selon Linstedt, la modélisation 3NF d'Inmon et la modélisation dimensionnelle de Kimball connaissent une grande faiblesse quand le volume de données augmente. De ce fait, une nouvelle approche de modélisation s'impose.

La modélisation d'entrepôt dans les deux approches classiques se base uniquement sur une architecture de données. Data Vault considère la notion d'architecture de processus dans la modélisation. En effet, Les structures de données sont déterminées selon une modélisation relationnelle et selon une notion de processus selon la fonction de la donnée.(**?**)

3. http ://it.toolbox.com/people/dlinstedt/

Data Vault s'intéresse aux changements dans les processus et structures de données plutôt que le changement dans les fonctions d'affaires (**?**).

Les caractéristiques principales du Data Vault sont les suivantes (**?**) :

— Les informations structurelles sont séparées des informations descriptives (attributs) pour des raisons de flexibilité et évitement du re-engineering en cas du changement.

— Le Data Vault permet un chargement parallèle de données.

— Les données ne sont pas transformées ni filtrées. L'approche permet de tracer la source des données.

— Les données ne sont jamais modifiées, elles restent intactes.

— L'accès au Data Vault est restreint, sa structure ne permet pas une exploitation finale de données.

Le Data Vault est une approche riche et solide. Elle est née pour promouvoir la flexibilité, l'évolutivité et la productivité de l'entrepôt de données.

2.3.3 Composants du modèle Data Vault

Dans le modèle Data Vault, on distingue trois types d'entités : les hubs, les liens et les satellites.

2.3.3.1 Hub

2.3.3.1.1 Définition du hub : Les hubs sont définis par une liste unique des clés d'affaires (clés naturelles) qui identifient un concept utilisé par l'organisation tel que Client et Produit. Ces clés sont les piliers de tout le modèle et sont vitales pour les entreprises afin de suivre, localiser et identifier leur information. Les hubs constituent le point de raccordement entre les différents secteurs de l'organisation(**?**).

Parfois ces clés d'affaires se trouvent dans les bases de données des systèmes transactionnels en tant que clés primaires.

Il est impératif que les clés d'affaires aient une unicité historique et générale, c'est-à-dire qu'il ne faut jamais utiliser la clé d'affaire dans différents secteurs de l'organisation pour signifier des choses différentes ou de les réutiliser(**?**).

On peut compléter chaque clé avec une ou plusieurs métadonnées comme son origine et sa date d'extraction.

2.3.3.1.2 Structure du Hub : L'entité hub est composée de ces éléments nécessaires : La clé d'affaire, la date du chargement, la source d'extraction et une clé de substitution (surrogate key en anglais) pour éviter les problèmes de performance liés aux clés complexes (**?**).

La figure suivante montre la structure de l'entité hub et un exemple.

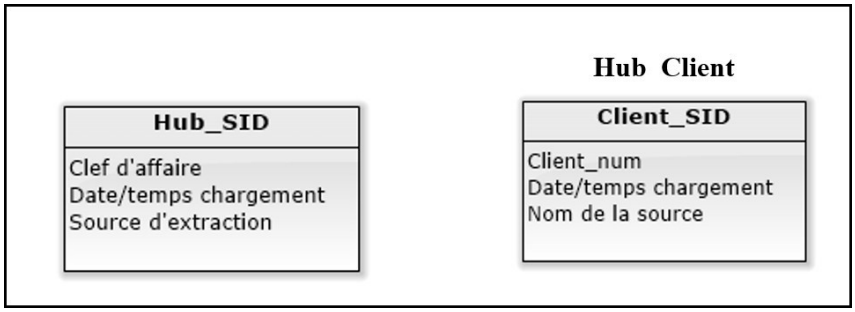

FIGURE 2.19 – Structure de l'entité hub et un exemple (**?**)

La table Hub_Client ci-avant contient :

— **Client_SID :** représente la clé de substitution, elle est généralement générée à partir du système de l'entrepôt de données.

— **Client_num :** Un numéro qui identifie le concept d'affaire Client.

— **Date/temps de chargement :** la date et le temps de la première injection de la clé dans l'entrepôt de données.

— **Source d'extraction :** représente le système source d'où on a pris l'information, elle est utilisée pour la traçabilité des données.

2.3.3.2 Lien(Link)

2.3.3.2.1 Définition du lien Les liens sont des entités associatives. Elles lient ensemble au moins deux hubs, autrement dit, elles mettent en relation des concepts d'affaires(**?**).

Le but de l'entité lien est d'enregistrer continûment toutes les relations qui aient lieu entre les éléments de données et ceci au plus bas niveau de granularité(**?**).

La granularité d'un lien est imposée par les hubs en relation avec le lien, c'est comme dans le cas des faits d'un modèle dimensionnel où la granularité du fait est dictée par les dimensions qui rentrent en relation avec la table de fait.

Les liens, comme les hubs, ne contiennent pas des données descriptives, mais apparaissent dans une autre entité appelée satellite que nous expliquons après.

Selon les besoins d'interrogation de l'entreprise, un ensemble de métadonnées peuvent être ajoutées au lien comme les mises à jour faites, l'évaluation de confidentialité, le cryptage de la clé, etc.

2.3.3.2.2 Structure du lien La structure de l'entité lien contient les éléments nécessaires suivant(**?**) :

- **Clés de substitution des hubs :** les clés d'affaires des différents hubs et liens en relation. Elles sont obligatoires pour l'entité lien.
- **Clé de substitution du lien :** Elle est optionnelle mais peut être une clé primaire obligatoire dans le cas où le lien est associé à d'autres liens.
- **Date de chargement :** une clé obligatoire qui enregistre quand l'association a été introduite la première fois.
- **Source d'extraction :** une clé obligatoire qui référence le système source, elle est utilisée pour des fins de traçabilité et d'intégration.

La figure suivante montre un exemple d'un lien.

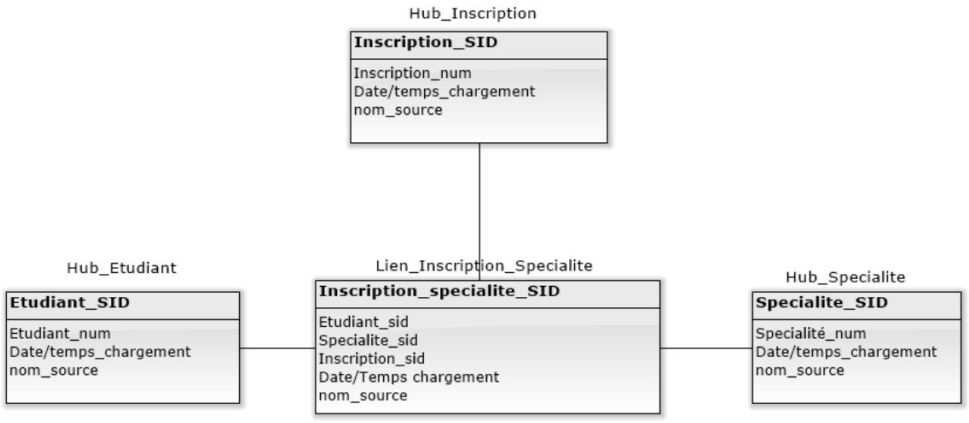

FIGURE 2.20 – Exemple d'un lien

2.3.3.3 Satellite

2.3.3.3.1 Définition du satellite : Les satellites contiennent les données qui décrivent les hubs et les liens à un moment donné et à travers le temps. Ces entités contiennent le contexte (provenant des processus d'affaires) d'un hub ou d'un lien(**?**).

Un satellite ne peut avoir qu'une seule table parent, celle qui le décrit. Par contre un

hub ou un lien peuvent avoir plusieurs satellites. Les satellites capturent les changements des données descriptives de l'entrepôt de données lorsqu'ils surviennent.

2.3.3.3.2 Structure du satellite : La structure de l'entité Satellite se compose des éléments suivants : clé de substitution du hub ou du lien qu'il décrit, date du chargement et source d'extraction. Des métadonnées peuvent être ajoutées au satellite.

La clé primaire du satellite est obtenue en combinant la clé du parent (hub ou lien) et la date du chargement.

La figure suivante montre trois exemples de satellites du hub Employé.

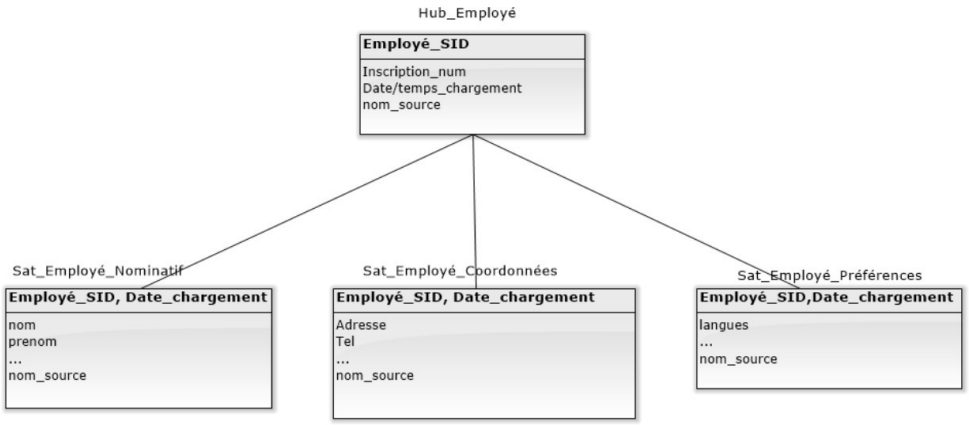

FIGURE 2.21 – Trois satellites du hub Employé

2.3.4 Architecture de l'entrepôt de données du Data Vault

Data Vault repose sur une architecture 3-tiers afin de séparer l'entrepôt de données brut des utilisateurs finaux et les différentes couches d'exploitation de données. Cette isolation permet de réduire les coûts engendrés par les changements qui se produisent dans les processus d'affaires.

Les trois tiers de l'architecture sont : **la zone de préparation de données**, **l'entrepôt de données du Data Vault** et **les magasins de données**.

L'architecture du Data Vault peut marcher avec l'approche d'Inmon dans le sens où le Data vault joue le rôle de l'entrepôt centralisé de l'entreprise (ou EDW en anglais pour Entreprise Data Warehouse)(**?**). Le Data Vault fournit les données aux magasins de données qui à leur tour constituent des référentiels pour l'exploration et l'analyse de données.

a figure suivante montre l'architecture Data Vault de l'entrepôt de données de l'entreprise.

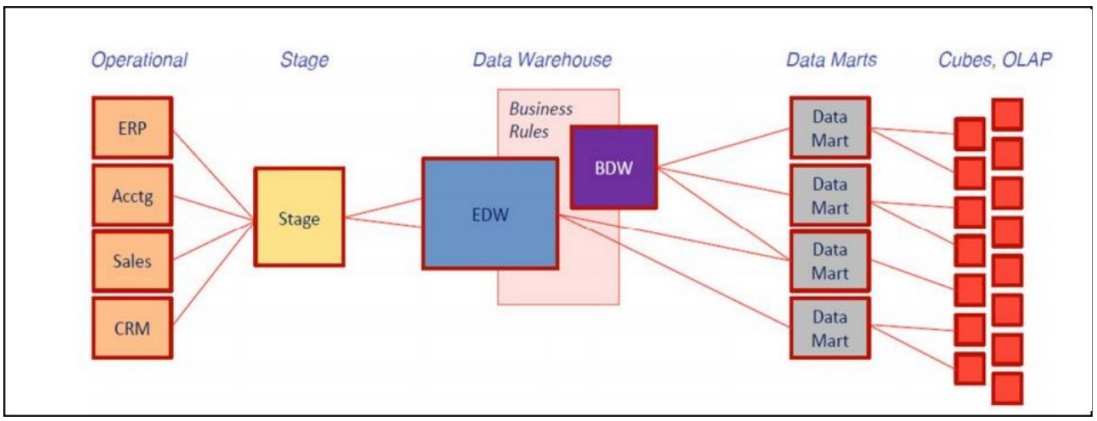

FIGURE 2.22 – Architecture de l'entrepôt Data Vault. (**?**)

En effet, nous constatons les couches suivantes (**?**) :

1. **Zone de préparation :** Cette couche joue le rôle d'un batch et soutient le processus de chargement de données à partir de plusieurs sources au Data Vault. Dans cette couche, aucune règle d'affaire n'est appliquée et les données restent dans leur format d'origine. On note que les données à ce niveau ne sont pas persistantes.
2. **Data Vault :** C'est La couche noyau, elle contient les données provenant des différentes sources sans subir des transformations, c'est-à-dire les données restent brutes. Le fait de laisser les données intactes constitue une différence fondamentale des deux approches classiques.

   Le modèle Data Vault est composé d'un ensemble de hubs, liens et satellites qui sont les éléments principaux du modèle comme nous l'avons expliqué avant dans la partie Composants du Data Vault.
3. **Entrepôt de données orienté métier (en anglais BDW pour Business Data Warehouse) :** Cette couche contient les données résultant de l'application des différentes règles métiers comme l'alignement des données avec les clés d'affaires.
4. **Magasins de données :** c'est la couche de présentation, elle est déployée en utilisant une modélisation dimensionnelle. Les magasins prennent les données à partir de l'entrepôt d'affaires et du Data Vault. On note que les données à ce

niveau ne sont pas persistantes. Les magasins à leur tour fournissent les données aux cubes OLAP pour les exploiter dans le Reporting, Data mining, etc.

2.3.5 Processus de développement du Data Vault

La méthodologie Data Vault se base principalement sur le modèle SEI/CMMI [4] 5 des meilleures pratiques et les combine avec les meilleures pratiques de Six Sigma, TQM, et SDLC. En particulier, il est axé sur une méthodologie agile pour la construction et le déploiement. L'équipe utilisant la méthodologie Data Vault sera automatiquement adaptée pour les projets reproductibles, cohérents et mesurables qui sont inclus dans le CMMI niveau 5

2.3.5.1 Data Vault et SEI/CMMI

CMMI (acronyme de Capability Maturity Model Integration), est un modèle de référence, un ensemble structuré de bonnes pratiques, destiné à évaluer et mesurer la maturité du processus de développement des systèmes informatiques d'une entreprise.

Le modèle CMMI comprend cinq niveaux de maturité qui sont : initial, discipliné, ajusté, géré quantitativement et optimisé.

Chaque niveau du modèle se base sur le niveau qui le précède. Une organisation ne peut pas aboutir à un niveau sans avoir passer avec succès le niveau directement inférieur.

Le dernier niveau sert comme une base pour l'approche Data Vault. L'organisation qui a acquis ce niveau est dans une boucle permanente d'optimisation des processus et des technologies.

L'alignement du modèle de données avec les niveaux de maturité du CMMI est critique et doit inclure : la répétition, la redondance, la fiabilité et la flexibilité (**?**) .

Par exemple, le niveau quatre du CMMI inclut l'optimisation du modèle de données pour des besoins de l'entreprise. Le niveau cinq s'intéresse à l'optimisation de l'architecture du Data Vault. A ce niveau (cinq), les changements sont engagés automatiquement.

L'approche Data Vault met en place une gestion de projet solide. Linstedt incite l'équipe projet à faire le lien entre les composants du projet qui sont : Structure de découpage de projet, structure de découpage organisationnel et structure de découpage de données, et établir un plan pour pouvoir suivre le projet tout le temps (**?**).

4. http ://www.tutorialspoint.com/cmmi/index.htm

2.3.5.2 Construction du Data Vault

Les étapes suivantes sont requises pour construire le Data Vault (**?**) :

- **Identification des concepts d'affaire :** cette étape consiste en l'identification des sujets d'affaires exercés par l'entreprise. Les interviews sont le cœur de cette étape.
- **Modéliser les hubs :** Cela nécessite une compréhension des clés d'affaires et de leur utilisation dans le cadre des concepts d'affaires identifiés.
- **Analyser les relations d'affaires :** cette étape consiste à identifier les relations qui lient les différents concepts d'affaire.
- **Modéliser les liens :** En se basant sur l'analyse des relations d'affaires, il est temps de modéliser les liens qui mettent en relation les hubs déjà modélisés.
- **Modéliser les satellites :** Il s'agit de fournir un contexte aux différents hubs et liens en ajoutant des données descriptives pour chaque entité.

On constate qu'il y a plusieurs règles qui régissent la modélisation du Data Vault. Nous énumérons une liste non exhaustive constituée de règles suivantes(**?**) :

- Les hubs doivent être raccordés par des liens.
- Les liens peuvent être raccordés à d'autres liens.
- Les liens doivent avoir au moins deux hubs associés avec eux afin d'être instanciés.
- Les clés de substitution peuvent être utilisées par les hubs et par les liens.
- Les clés de substitution ne peuvent pas être utilisées pour les satellites.
- Les clés d'affaires des hubs ne change jamais.
- Les satellites peuvent être connectés à un hub ou à un lien.
- Les liens peuvent avoir une clé de substitution.
- Les données sont réparties dans les structures de satellite en se basant sur le type d'information et le taux de variation.

2.3.5.3 Elaboration des magasins de données

Dans cette phase, l'équipe projet procède à définir les besoins des décideurs par différents moyens comme les interviews et les meetings, la rédaction des décideurs d'un document qui comporte leurs besoins, etc.

L'équipe projet peut ensuite développer les magasins de données pour permettre aux utilisateurs d'exploiter leurs données et en tirer de l'information utile.

Le développement des magasins suit la modélisation dimensionnelle de Kimball (Section II.2.5.3) vu sa rigueur en termes de performance d'interrogation.

2.3.5.4 Chargement de données dans le Data Vault

Après avoir modélisé le Data Vault, on procède au chargement de données dans les structures du Data Vault (hubs, liens et satellites).

Le chargement d'un entrepôt de données est réalisé avec des modes de traitement différents. Souvent, les entrepôts de données sont chargés dans un mode discontinu.

Dans ce mode, un nombre de transactions similaires à traiter sont groupées. Ces opérations accumulées sont ensuite présentées à l'entrepôt de données à des intervalles de temps réguliers (**?**).

Une autre façon de traitement des données est le traitement en ligne. Elle implique un système à accès direct. En effet, les données entrant dans le système sont transmises directement à l'utilisateur final sous un format de sortie spécifique (**?**).

Conclusion

A travers ce chapitre, nous avons introduit les trois approches les plus célèbres dans le domaine de modélisation des entrepôts de données à savoir : l'approche d'Inmon, l'approche de Kimball et l'approche Data Vault.

L'approche d'Inmon consiste à construire un grand entrepôt de données à l'échelle de toute l'entreprise. Cet entrepôt va ensuite alimenter des magasins de données. L'approche est basée sur les modèles entité-association de l'entreprise.

L'approche de Kimball consiste à construire des magasins de données indépendants un par un, puis à les regrouper par des niveaux intermédiaires jusqu'à obtention d'un entrepôt. Cette approche est basée sur le concept de modélisation dimensionnelle.

L'approche Data Vault, quant à elle, consiste à créer dans un premier temps le Data Vault qui va stocker les données brutes provenant des différentes sources. Le Data Vault se base sur la modélisation hub-lien-satellite qui sépare les données structurelles des données descriptives.

Le Data Vault alimente ensuite les magasins après avoir appliqué les différentes règles d'affaires. Les magasins sont déployés en utilisant la modélisation dimensionnelle.

Dans le chapitre suivant, nous présenterons une comparaison entre ces trois approches en s'appuyant sur différents critères.

Chapitre 3

Étude Comparative

Introduction

Une entreprise qui veut opter pour un système décisionnel doit réfléchir profondément sur la méthodologie à adopter pour le construire car comme tout projet, un système décisionnel mobilise des gens, génère des coûts et est plein de risques et exposé à l'échec.

Le choix de la méthode repose, d'une part, sur la connaissance profonde de l'entreprise et ses orientations et d'autre part, sur l'analyse critique des méthodes existantes afin d'aboutir à l'approche propice.

Dans ce chapitre, nous effectuons, Dans un premier temps, une comparaison entre les trois méthodes étudiées (Inmon, Kimball et Data Vault) et dans un deuxième temps, nous partageons notre vision personnelle et nous citons des recommandations qui aident à choisir la meilleure méthode dans une situation donnée.

3.1 Analyse comparative

(**?**) constate que les approches de Kimball et Inmon ressemblent du fait qu'ils utilisent des processus ETL pour alimenter l'entrepôt de données.

(**?**)constate que la similarité entre l'approche Data Vault et l'approche de Kimball réside dans l'implémentation itérative de la solution et l'utilisation d'une zone de préparation pour la récupération et la synchronisation.

L'approche Data Vault et l'approche d'Inmon se mettent d'accord sur le fait que l'entrepôt est le plus grand référentiel de l'entreprise (EWD).

Malgré ces petites similarités, les différences entre les trois approches sont plusieurs et profondes.

Les différences importantes apparaissent dans les méthodologies de développement, la modélisation de données, l'architecture de l'entrepôt et le management du cycle de vie.

Ces différences sont résumées dans le tableau ci-dessous, qui est basé sur les travaux de (**?**), (**?**), (**?**), (**?**) et (**?**).

3.1.1 Méthodologie et architecture

La table 3.1 représente une comparaison des trois approches basée sur l'architecture de l'entrepôt, la méthodologie de développement, la complexité générale de l'approche ainsi que le coût et le temps de la mise en place de l'entrepôt de données.

	Relationnelle	**Dimensionnelle**	**Data Vault**
Structure architecturale	L'entrepôt de données atomique (EDW) alimente les magasins de données départementaux.	L'ensemble des magasins de données constitue l'entrepôt de l'entreprise.	Le Data Vault alimente les magasins de données.
Complexité	Assez complexe.	Simple.	Simple.
Méthodologie de développement	Inspirée de la méthodologie spirale.	Basée sur un processus four-step (quatre étapes).	Basée sur une méthodologie agile.
Coûts de déploiement	Les coûts initiaux sont plus élevés et les coûts ultérieurs du développement sont inférieurs.	Les coûts initiaux sont inférieurs. Chaque phase ultérieur coûte presque le même.	Le coût du projet est inférieur par rapport aux deux autres approches.
Temps de déploiement	Longue durée.	Courte durée.	Courte durée

TABLE 3.1 – Comparaison entre les approches selon la Méthodologie et l'Architecture

3.1.2 Modélisation de données

La modélisation de données diffère selon le modèle choisi, L'orientation ainsi que les outils utilisés lors de la modélisation. La table 3.2 présente une comparaison de la manière de modélisation des données selon les trois approches. Cette comparaison est basée sur

les outils, l'orientation des données ainsi que l'implication de l'utilisateur final dans le processus de modélisation.

	Relationnelle	**Dimensionnelle**	**Data Vault**
Orientation de données	Pilotée par les données.	Orientée processus.	Orientée processus.
Outils	Outils classiques de modélisation (ERDs, DISs).	Modélisation dimensionnelle	Modélisation des hubs, liens et satellites.
Implication de l'utilisateur final	Faible.	Forte.	Forte.

TABLE 3.2 – Comparaison entre les approches selon la Modélisation de données

3.1.3 Philosophie

Chaque approche démarre d'une philosophie qui peut être proche ou totalement différentes des deux autres approches. L'acteur primaire visé par l'approche et les objectifs de chacune des trois approches sont présentés dans la table 3.3.

	Relationnelle	**Dimensionnelle**	**Data Vault**
Acteurs primaires	Les professionnels d'IT.	Les utilisateurs finaux.	Les utilisateurs finaux.
Objectifs	Fournir une solution technique complète basée sur des méthodes et technologies prouvées.	Fournir une solution qui facilite l'interrogation directe de données par les utilisateurs finaux.	Fournir une solution technique solide et complète comblant les lacunes des solutions classiques. Elle est basée sur des méthodes prouvées.

TABLE 3.3 – Comparaison entre les approches selon la philosophie

3.1.4 Intégration de données et ETL

Un élément clé dans le choix d'une approche est la possibilité d'une intégration simple et efficace de plusieurs sources de données. Cet élément est expliqué dans la table 3.4 à travers une comparaison entre la possibilité d'une intégration de sources multiples et une mesure de complexité du processus ETL.

	Relationnelle	**Dimensionnelle**	**Data Vault**
Intégration de sources multiples	Les règles de transformation doivent être,implémentées dans les processus ETL.	Les règles de transformation doivent être implémentées dans les processus ETL.	La séparation des satellites et des clés d'affaires diminue fortement la complexité.
Complexité du processus ETL	Les règles de transformations sont simples quand le modèle de données, est similaire aux modèles,des sources de données.	Les transformations entre les modèles OLTP et le modèle dimensionnel sont complexes.	Les règles sont simples pour charger les hubs, liens et satellites.

TABLE 3.4 – Comparaison entre les approches selon l'intégration de données et ETL

3.1.5 Management du cycle de vie

La finalité des trois approches est de concevoir et mettre en place un entrepôt de données qui répond aux besoins et aux attentes des utilisateurs. Ceci n'est pas suffisant pour une entreprise qui a des besoins instables, flexibles et changent au cours du temps. D'autres critères peuvent être rajoutés tels que la facilité du changement, la traçabilité et la performance d'interrogation. La table 3.5 résume les critères cités précédemment.

	Relationnelle	**Dimensionnelle**	**Data Vault**
Flexibilité contre le changement du modèle de données sources	Nécessité de changements dans les tables.	Souvent les changements du modèle source influent sur le modèle de données de l'entrepôt.	Les tables existantes ne sont pas affectées. Le seul changement est l'ajout des satellites appropriés.
Flexibilité contre les nouveaux besoins d'analyse	Le modèle change seulement si les données exigées n'existent pas dans l'entrepôt.	Les nouveaux besoins ont un impact sur le modèle de données.	Aucun changement ne sera fait sur le modèle du Data Vault. Seulement la livraison de données aux magasins qui doit être adaptée.
Facilité de changement du modèle	Les données historiques doivent être migrées dans certains cas.	Le refactoring de certaines tables est nécessaire dans certains cas.	Les tables existantes ne sont pas affectées. Le seul changement est l'ajout des satellites appropriés.
Audit et traçabilité	L'information historique est capturée en insérant un nouvel enregistrement à chaque fois les données sources changent.	Utilise le concept de « Dimension à évolution lente [1] » pour enregistrer le changement d'historique.	L'information historique est capturée en insérant des nouveaux liens et satellites.
Performances d'interrogation	L'interrogation est très lente à cause de la structure 3NF de données. Les magasins dimensionnels sont nécessaires pour l'analyse et le reporting.	Le modèle est conçu pour être très performant en interrogation en dénormalisant les dimensions et laissant les faits en 3NF.	L'interrogation directe est très lente à cause de la haute normalisation de données. Les magasins dimensionnels sont nécessaires pour l'analyse et le reporting.

TABLE 3.5 – Comparaison entre les approches selon le management du cycle de vie

3.2 Appréciation personnelle

Comme la matrice de comparaison précédente le montre, aucune approche ne répond parfaitement à tous les critères. Chaque méthode a ses avantages et ses inconvénients.

Dans ce qui suit et à travers l'étude comparative des trois approches que nous avons effectuée, nous proposons quelques recommandations qui permettent de répondre à la

question « Quelle approche dans quelle situation ? ».

- **Approche d'Inmon :**cette approche est recommandée quand les besoins d'analyse ne sont pas définis ou le but de l'entrepôt est de fournir les informations à plusieurs systèmes BI. Elle est aussi préférée si les structures du système source sont relativement stables.
- **Approche de Kimball :**cette approche est fortement recommandée pour les magasins de données vu que le modèle dimensionnel offre une haute performance d'interrogation et il est compréhensible par les utilisateurs finaux. En outre, elle est aussi appropriée pour développer l'entrepôt si les besoins sont connus et bien définis.
- **Approche Data Vault :**C'est une approche puissante pour développer l'entrepôt quand il y a plusieurs sources de données avec des changements réguliers des structures de ces sources. Elle est efficace dans les environnements de projets agiles. Si la flexibilité, la productivité et l'évolutivité de l'entrepôt sont les préoccupations de l'entreprise alors l'adoption du Data Vault est le meilleur choix.

Conclusion

A l'issue de ce dernier chapitre, nous avons vu les similarités et les différences entre les trois approches en s'appuyant sur certains critères dont les plus importants sont la méthodologie de développement, l'architecture et la modélisation de données.

Nous avons constaté aussi qu'il n'y a pas une approche parfaite, mais avec la connaissance de la situation de l'entreprise on peut opter pour l'une ou l'autre.

Conclusion générale

La concurrence rude entre les entreprises nécessite de plus en plus la mise en place des systèmes décisionnels. Ces derniers permettent aux décideurs de prendre des décisions et des choix dans des temps très courts et ce pour un volume d'information toujours plus important.

A travers ce travail de Master, nous avons vu que les systèmes décisionnels reposent sur un référentiel de stockage centralisé appelé entrepôt de données. Son rôle est d'intégrer et de stocker l'information utile aux décideurs et de conserver l'historique des données pour supporter les analyses effectuées lors de la prise de décision.

Les approches de modélisation des entrepôts de données sont plusieurs et diverses. Cependant, nous avons constaté que trois approches monopolisent le marché des entrepôts. Il s'agit de l'approche de Kimball, l'approche d'Inmon et l'approche Data Vault.

L'approche d'Inmon consiste à construire un grand entrepôt de données à l'échelle de toute l'entreprise. Cet entrepôt va ensuite alimenter des magasins de données. La modélisation de données est basée sur les modèles entité-association de l'entreprise sur lesquels on fait une série de transformations pour aboutir au modèle de données de l'entrepôt de données.

Inmon n'implique pas les utilisateurs finaux dans la construction de l'entrepôt, par contre des équipes de professionnels IT sont exigées.

La vision de Kimball est différente. En effet, il voit qu'un entrepôt de données n'est qu'un ensemble de magasins de données indépendants et on n'a pas besoin d'une séparation physique entre les deux. Cette approche est basée sur les concepts de la modélisation dimensionnelle où on implique fortement l'utilisateur final et l'on incite à exprimer ses besoins, ensuite on tire les faits et dimensions correspondants tout en respectant un schéma

de modélisation selon les besoins de l'entreprise.

L'approche Data Vault, quant à elle, consiste à créer dans un premier temps le Data Vault qui va stocker les données brutes provenant des différentes sources. Le Data Vault se base sur la modélisation hub-lien-satellite qui sépare les données structurelles des données descriptives.

Le Data Vault alimente ensuite les magasins après avoir appliqué les différentes règles d'affaires. Les magasins sont déployés en utilisant la modélisation dimensionnelle.

Nous avons élaboré aussi une analyse comparative entre les trois méthodes et nous avons constaté qu'aucune approche n'est meilleure. En effet, avec la connaissance de la situation de l'entreprise et de ses orientations, nous pouvons choisir une approche plutôt qu'une autre.

L'approche de Kimball est préférée pour modéliser les magasins de données, vu la performance d'interrogation qu'elle fournit, particulièrement quand les besoins sont stables et bien définis. Tandis que, l'approche d'Inmon est recommandée quand les besoins ne sont pas définis ou ils sont très évolutifs.

Les deux approches font face à beaucoup de problèmes surtout quand les sources de données changent souvent. Cela implique un re-engineering de l'entrepôt. Pour pallier ces problèmes, l'approche Data Vault est préconisée. Elle permet une flexibilité et une évolutivité extrêmes. Elle est recommandée quand les sources changent souvent de structures.

Au final, nous avons présenté et comparé les approches de modélisation des entrepôts de données les plus répandues dans la littérature. Pour ce faire, nous nous somme basés sur des travaux traitant ce sujet.

En revanche, nous n'avons pas été en mesure de vérifier de manière pratique les résultats de toutes les approches mais nous envisageons, dans nos travaux futurs, d'inclure ce volet là pour donner nos propres résultats et évaluations.

Bibliographie

[1] John W. Dower 1991.

Bibliographie

[**Weir2008**] Weir : A Configuration Approach for Selecting a Data Warehouse Architecture, Thesis,2008

(Kimball et Ross, 2013) Kimball, Ralph ; Ross, Margy : The data warehouse toolkit : The definitive guide to dimensional modeling. John Wiley & Sons, 2013.

[**Denis 2008**] Denis : Conception et réalisation d'un entrepôt de données institutionnel dans une perspective de support à la prise de décision, Thesis,2008.

[**Teste2009**] Teste, Olivier : Modélisation et manipulation des systèmes OLAP :de l'intégration des documents à l'usager, Université Paul Sabatier-Toulouse III, Dissertaion, 2009

[**Adamson2012**] Adamson, Christopher : Mastering data warehouse aggregates : solutions for star schema performance. JohnWileySons ,2012.

[**Awel2014**] Awel : Data Vault Modelling, Thesis,2014

[**Awel2014**] Mathiews, Diane : Data vault et bi. URL :http ://fr.slideshare.net/dlinstedt/prsentation-data-vault-et-bi-v20120508

[**Orlov2014**] Orlov, Vadim : Data Warehouse Architecture : InmonCIF, Kimball Dimensional or Linstedt DataVault ? 2014. –URL : https ://blog.westmonroepartners.com/data-

warehouse-architecture-inmon-cif-kimball-dimensional-or-linstedt-data-vault/

(Poletto, 2012) Poletto, Maxime : L'informatique décisionnelle, These professionnelle, 2012. – URL http ://news.exia.cesi.fr/wp-content/uploads/2012/06/ Maxime-Poletto-Th%C3%A8se.pd