Université Paris Nanterre

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

Table des matières

Chapitre 1

Introduction générale

Aujourd'hui,parmis les défis auxquels font face les entreprises, ont rouve 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(Datawarehouse).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 auxu tilisateurs 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 dedéveloppement 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 lamodélisation 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 comparatives de ces trois approches de modélisation d'entrepôt de données.

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 don-

nées : l'approche de modélisation par sujet **d'Inmon** et l'approche dimensionnelle de **Kimball**. Cependant, 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 »**

Le but de ce travail est de faire une synthèse sur la BI à travers ces trois approches. D'abord, je ferai une étude approfondie sur 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. Ensuite je vais me focaliser sur la présentation de chaque approche : sa définition, sa philosophie, son architecture ainsi que sa méthodologie de développement. Enfin, je dresserai une analyse comparative entre les trois approches en s'appuyant sur des critères tels que :

- La méthodologie et l'architecture : structure & architecture, méthodologie de développement , coût et temps de déploiement... etc
- La modélisation de données : orientation des données, les outils utilisés, l'implication de l'utilisateur final ... etc
- Intégration de données : intégration des sources multiples, complexité de processus ETL ... etc
- Management du cycle du vie : facilité du changement du modèle, performances d'interrogation . . . etc

Cette étude comparative sera accompagnée par des tests sur des données réelles (au sein de l'entreprise de mon contrat professionalisation) avec un groupe témoin afin d'aboutir à une synthèse permettant de répondre à la question : **« Quelle approche dans quelle situation ?».**

Chapitre 2

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 a fin de pouvoir entirer 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,je vais aborder les concepts liés aux systèmes décisionnels. je commence par des définitions, je cite après l'historique et les objectifs des SID etje m'étalons à la fin sur l'architecture d'un SID.

2.1 Définitions

2.1.1 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 es tune 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 (Davis and Olson, 1986)."

" 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 (Haouet, 2008)"

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.(March, 1991)."

2.1.2 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 (Haouet, 2008). En effet plusieurs définitions ont été données à ce terme. Dans la première acception, je retiens 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 depilotage)» (Le Moigne, 1977). Dans la seconde acception,on retrouve la définition de Reix selon laquelle le 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 » (Reix andRowe, 2002).

2.1.3 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» (Devisy,2002).

(LAU et al., 2009) 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, unsystème décisionnel manipule des données opérationnelles avec différents moyens de collecte, destockage et d'analyse pour soutenir le processus d'aide à la décision.

2.2 Historique des Systèmes Décisionnels

Dès les années 60, les données informatisées dans les organisations ont pris une imortance qui n'acessé 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 (Teste, 2009).

Avec l'accroissement des besoins en matière de décision, de nouveaux concepts sont apparus au début des années 90 :l'entrepôt de données (datawarehouse) et les magasins de données (datamart). 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 dansl 'informatique décisionnelle avec ces avancées technologiques :les outils informatiques d'aide à la décision, désormais appelés «Business ntelligence» (Doucet and De La Villarmois, 2007).

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, Oracle, IBM, SA 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 (OpenSouce), qui ont atteint aujourd'hui une certaine maturité (SpagoBI, Talend , Jasper) (Teste, 2009).

2.3 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 (Kimball and Ross, 2013), 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. ...etc.

— 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.

2.4 Système Transactionnel VS Système Décisionnel

Les systèmes transactionnels, d'après (Kimball et al., 2002), 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.

2.5 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(LAU et al., 2009). Afin de mettre en oeuvre 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.

2.5.1 Sources de données

Les sources de données alimentent l'entrepôt de données. Elles sont regroupées en quatre catégories (Ponniah, 2001) :

— Les données de production
— Les données internes
— Les données archivées
— Les données externes

2.5.2 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 (Poletto, 2012), un ETL extrait, nettoie, et importe les données à partir de différentes sources et les charge dans un entrepôt de données.

2.5.3 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.

2.5.3.1 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 »(Kimball and Ross, 1996).

« 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) »(Inmon, 1996).

2.5.3.2 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 etd 'interrogation décisionnelle. Le magasin est généralement stocké au sein d'une base de données multidimensionnelle (BDM)(Tournier, 2007). 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 méta données d'un entrepôt de données se présentent sous trois catégories (Ponniah, 2001) :

— 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.

2.5.3.3 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 (Inmon, 2002) :

— 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.

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

Conclusion

Les systèmes décisionnels occupent une place importante dans les entreprises vu qu'ils soutiennent les dirigeants pour prendre des décisions au bonmoment. L'entrepôt de données est le coeur 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.

Chapitre 3

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 DataVault développée par son inventeur DanLinstedt. Dans ce chapitre,nous allons étudier chaque approche : sa philosophie, son architecture et la méthodologie pour la mettre en place.

3.1 Approche d'Inmon

Cette approche est créée par Bill Inmon 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.

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 (Denis, 2008), C'est pourquoi cette approche est qualifiée de «piloter par les données ».

3.2 Approchede Kimball

Kimball 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 lap hilosophie 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 finaaux dès les premières phases du projet, c'est pourquoi cette méthode est appelée « piloter par les besoins utilisateurs».

3.3 Approchede DataVault

Au début des années 2000, Dan Linstedt est entré dans la compétition en proposant une troisième approche dite «modélisation DataVault»(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

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.pdf.