RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEURE
ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ MOULOUD MAMMERI DE TIZI-OUZOU
FACULTÉ DE GÉNIE ÉLECTRIQUE ET D'INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Mémoire

De fin d'études

Pour l'obtention du Diplôme de Master en Informatique.
***Option** : **I**ngénierie des **S**ystèmes d'**I**nformation.*

Théme

CONCEPTION ET RÉALISATION D'UN ENTREPÔT DE DONNÉES

CAS: DIRECTION INFORMATIQUE DE LA C.N.A.S
BEN AKNOUN ALGER.

Réalisé par :

- Mr KOURABA Abdelmadjid.

Tuteur Entreprise	Tuteur Université
Mr N. KOUADRIA Chef de Département Etudes et Développement Direction Informatique CNAS de Ben Aknoun Alger	Mr Y. CHAIEB Maître assistant à l'UMMTO Université Mouloud Mammeri Tizi-Ouzou

Année Universitaire : 2012-2013

Remerciement

Dieu merci d'avoir été toujours et d'être là, prés de moi, m'éclaircir la vue, me remplir le cœur de foi, de courage et de patience.

Je ne remercierai jamais assez mes parents, sans lesquels, rien n'aurait été possible.

La rédaction de ce travail n'aurait pas été possible sans le concours de certaines personnes que je tiens à remercier très sincèrement ici :

- Mr N. KOUADRIA mon tuteur de stage à la CNAS, pour ses nombreux conseils et sa gestion tout au long du projet.

- Mr Y. CHAIEB mon tuteur de stage à l'université, pour ses précieuses recommandations quant à l'organisation et la rédaction de ce mémoire.

Je remercie l'ensemble des collaborateurs de la CNAS qui m'ont accueilli chaleureusement et qui ont contribué au bon déroulement du stage.

Enfin, je remercie tous mes amis, sans exception et sans distinction, pour leur présence et leur soutien.

Merci à tous.

Dédicace

Je dédie ce modeste travail

A mes parents pour leur esprit de sacrifice
Leur encouragement et leurs conseils
Et toute ma famille

Toute la promotion 2013
Et
Tous ceux qui on contribuer de prêt ou de loin pour la réalisation de ce travail.

Abdelmadjid

Sommaire

Liste des figures

.

Liste des tableaux

INTRODUCTION GÉNÉRALE

Avant l'ère du numérique, les entreprises devaient puiser les informations hétérogène de façon non-automatisée. De plus, les outils de l'époque ne permettaient pas d'effectuer des calculs poussés afin d'analyser les données recueillies. Les décisions d'affaire étaient prises principalement sur la base de l'intuition du corps exécutif. Au fur et à mesure, les compagnies ont commencé à automatiser le processus de collecte de données, et l'information commençait à s'accumuler. Toutefois, l'organisation de ces données était précaire en raison d'un manque d'infrastructure de stockage et d'incompatibilité entre les différents systèmes. L'analyse des données était pénible et demandait un temps considérable, et elle était réservée pour observer les tendances à long terme. Les décisions imminentes reposaient encore une fois sur l'intuition. Avec les progrès informatiques, la collecte des données est devenue abordable, et des entrepôts de données sophistiqués (Data warehouse) sont apparus. Des outils spécialisés (ETL) ont été conçus pour alimenter ces entrepôts. Les techniques de génération de rapports et d'analyse de données sont devenues plus performants. Aujourd'hui, l'art de l'informatique décisionnelle réside dans la manipulation et l'extraction d'information pertinente à partir d'un volume de données gigantesque. Les outils décisionnels actuels, permettent d'analyser, d'extrapoler et de rapporter des données. Certains outils modernes permettent aux utilisateurs d'effectuer des croisements de données et de faire des recherches poussées sur un secteur d'activité particulier.

Les technologies de l'information nous génèrent une multitude de données comme jamais auparavant. Le problème n'est donc plus tant d'acquérir une masse de données, mais de l'exploiter. Pour cela il faut collecter de l'information de qualité, la normaliser, la classer, l'agréger et l'analyser, pour l'exploiter afin d'en extraire la substantifique moelle et donc prendre la bonne décision au bon moment.

Toutes les actions prises par n'importe quel individu ou entreprise est précédé par une décision (préparée ou pas). Dans un environnement concurrentiel, Prendre une bonne ou mauvaise décision ou est équivalente à la vie ou à la mort, gagner ou perdre. Alors les dirigeants de l'entreprise, quelque en soit d'ailleurs le domaine d'activité, doivent être en mesure de mener à bien les missions qui leur incombent en la matière.

Ils devront prendre notamment les décisions les plus opportunes. Ces décisions, qui influeront grandement sur la stratégie de l'entreprise et donc sur son devenir, ne doivent pas être prises ni à la légère, ni de manière trop hâtive, compte tenu de leurs conséquences sur la survie de l'entreprise. Il s'agit de prendre des décisions fondées, basées sur des informations claires, fiables et pertinentes.

Le problème est de savoir donc comment identifier et présenter ces informations à qui de droit, sachant par ailleurs que les entreprises croulent d'une part sous une masse considérable de données et que d'autre part les systèmes opérationnels «transactionnels» s'avèrent limités, voire inaptes à fournir de telles informations et constituer par la même un support appréciable à la prise de décision. C'est dans ce contexte que les «systèmes décisionnels » ont vu le jour. Ils offrent aux décideurs des informations de qualité sur les quelles ils pourront s'appuyer pour arrêter leurs choix décisionnels. Pour se faire, ces systèmes utilisent un large éventail de technologies et de méthodes, dont les «entrepôts de données» (Data Warehouse) et les scripts d'alimentation (ETL) qui représentent les éléments principaux et incontournables pour la mise en place d'un bon système décisionnel.

Un système décisionnel ne remplace pas les systèmes opérationnels qui font fonctionner l'entreprise, mais il vient s'y intégrer, en y extrayant des données, afin d'en diffuser la connaissance, de la manière la plus facilement exploitable par les personnes concernées. Les systèmes décisionnels sont basés sur la construction des entrepôts de données qui contiennent des données issues des différents systèmes opérationnels, après une série de traitement (correction, nettoyage, filtrage,. . .) à l'aide des outils ETL (Extract, transform, load), permettant de rendre les données prêtes à analyser. La structure spécifique des entrepôts de données offre la possibilité de créer des cubes qui sont, à leur tour, très adaptés à des opérations d'analyse souples et rapides assurées par des operateurs d'analyse en ligne OLAP (On-Line Analysis Processing) qui permettent aux décideurs de naviguer librement dans les cubes et d'avoir une vision globale sur leurs entreprises. Et la qualité des données des entrepôts permet de tirer des bonnes connaissances lors de l'application des techniques statistiques. De là, la combinaison des outils OLAP et les techniques statistiques constitue un environnement décisionnel.

Le présent projet tend à la mise en place d'un système en mesure de consolider les données issues des systèmes transactionnels, et d'offrir des informations de qualité pour les décideurs.
Un tel système requiert la mise en place d'un entrent de données fiables contenant les informations nécessaires à l'accomplissement des processus décisionnels.

Chapitre 01

Etude de l’existant

Chapitre 01

Etude de l'existant

Sommaire

1.1 Introduction:

L'étude de l'existant est une étape initiale et nécessaire pour analyser et étudier les systèmes d'information et les bases de données existantes et qui sont opérationnelles, elle permet aussi de prendre connaissance du domaine dont on souhaite améliorer le fonctionnement, et pour atteindre cet objectif nous allons collecter puis représenter l'ensemble des informations qui pourraient se révéler utile aux tâches de conception. Ces informations vont être analysées et diagnostiquées.

1.2 Présentation de l'organisme d'accueil :

La CNAS (Caisse Nationale d'Assurance Sociales des travailleurs salariés) a connu l'informatique depuis le siècle des bobines et la cartes magnétiques, cela reviens à la nature compliqué de son système d'information qui regroupe les assurés de toute la nation d'Algérie, et qui nécessite une structure bien fiable.

Durant les dernières années, la CNAS a connu une révolution dans le domaine informatique en commençant, il y a quelques années par faire un passage (pas encore terminé) de la base de données vers ORACLE pour la sécurité du système et assurer de la mise en œuvre de la carte biométrique CHIFA.

La CNAS est administrée par un Conseil d'Administration, elle est placée sous la tutelle du Ministre du travail, de l'Emploi et de la Sécurité Sociale, son siège est à Alger (BEN AKNOUN), elle a compétence nationale et dispose de services centraux et locaux, parmi ses filières on trouve un centre familial à caractère social à BEN AKNOUN, qui contient une direction informatique où nous faisons notre étude.

1.2.1 L'organigramme de la CNAS :

L'organigramme de la CNAS comprend le siège et les différentes agences de wilayas.

Le siège de la CNAS comprend la direction générale à laquelle sont rattachées 10 directions centrales, son rôle est de gérer toutes les informations provenant des directions centrales à travers les wilayas du pays d'Alger.

Voici l'organigramme général de la CNAS :

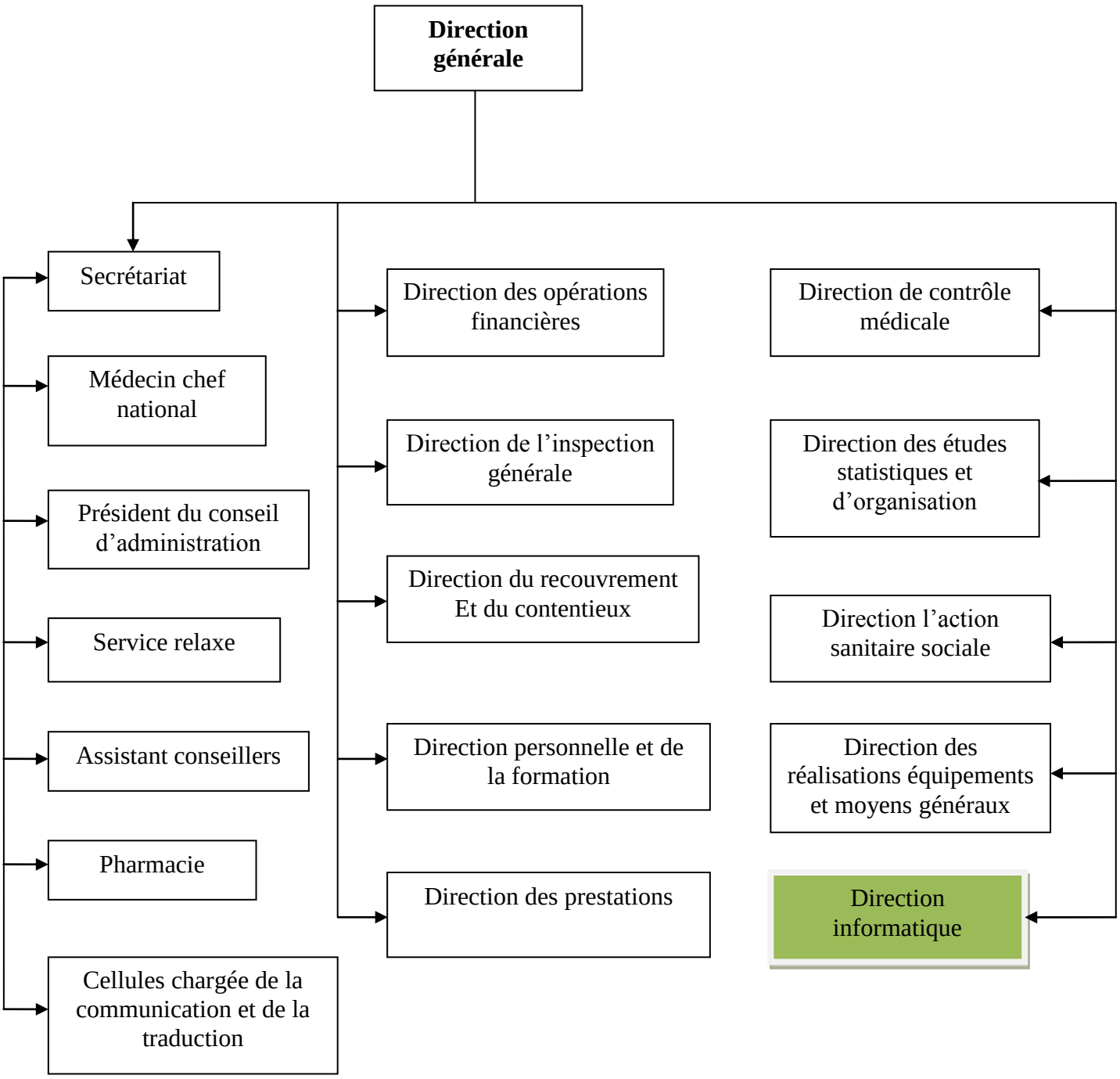

Figure 1.1 : Organigramme général de la CNAS.

Remarque : il se trouve une seule direction informatique qui est rattachée à la direction générale son rôle est d'établir les différents programmes utilisés dans les structures de la CNAS ainsi que la maintenance.

1.2.2 Mission de la CNAS :

La CNAS a pour mission :

- De gérer les prestations familiales.
- D'assurer le recouvrement, le contrôle et les contentieux des recouvrements au financement des prestations.
- De contribuer à promouvoir la politique de prévention des accidents de travail et des maladies professionnelles.
- De gérer les prestations dues aux personnes bénéficiaires de conventions accords internationaux de sécurité sociale.
- D'exercer le contrôle médical des bénéficiaires.
- De faire précéder à l'immatriculation des assurés sociaux, et aux employeurs.
- D'assurer en ce qui concerne l'information des bénéficiaires et des employeurs.
- De rembourser les dépenses occasionnées par le fonctionnement de diverses commissions et juridictions.

1.2.3 Organisme d'accueil :

La direction de l'informatique est chargée de développer et gérer l'ensemble des moyens informatiques de la CNAS sous ses trois aspects :

- Software.
- Hardware.
- Humains.

A ce titre elle conçoit, réalise et met en œuvre les programmes d'informatisations et d'équipement de l'organisme. Elle est en relation constante avec les autres structures pour concrétisation des engagées.

1.2.4 Structure d'accueil :

Pour la réalisation de ses missions, la direction informatique est organisée en quatre départements :

- Le département des études et développement.
- Le département système et réseaux.

- ❖ Le département exploitation.
- ❖ Le département développement CHIFA.

1.2.4.1 Le département étude et développement :

Composé d'ingénieurs et de techniciens ce département a pour missions :

- ✓ De mener les études informatiques.
- ✓ De concevoir les logiciels de gestion.
- ✓ De réaliser et/ou de faire réaliser les logiciels.
- ✓ D'installer les logiciels développés.
- ✓ De former les utilisateurs pour l'exploitation de ces logiciels.
- ✓ D'assurer la maintenance et la mise en conformités des logiciels en exploitation.

Il est organisé par projet.

1.2.4.2 Le département système et réseaux :

Les ingénieurs de ce département ont pour taches :

- ✓ De concevoir et mettre en place le réseau de télétraitement.
- ✓ De rechercher constamment les dernières versions logicielles.
- ✓ De tester les nouvelles plates-formes.
- ✓ De développer les procédures de sécurité.
- ✓ D'être en veille technologique.
- ✓ En relation avec les partenaires, maintenir le site web.
- ✓ Développer et assurer le fonctionnement de l'intranet.

1.2.4.3 Le département exploitation :

Il est composé :

- ✓ Du service matériel.
- ✓ Du service maintenance.

- **Le service matériel :**

  Le magasin est chargé de l'entreposage et de l'inventaire de tous les équipements informatiques.

  Il tient les stocks et enregistre tous les mouvements du matériel.

- **Le service maintenance :**

  Il procède à l'installation et aux réparations de tous les équipements informatiques à l'exception des sites centraux implantés dans les agences de wilaya.

1.2.4.4 Le département développement CHIFA :

Composé d'ingénieur et des techniciens, ce département a pour missions de concevoir et de réaliser les composants logiciels métiers et CNAS en rapport avec le système CHIFA.

1.2.5 L'organigramme d'accueil :

Voici l'Organigramme d'accueil de la Direction Informatique de la CNAS:

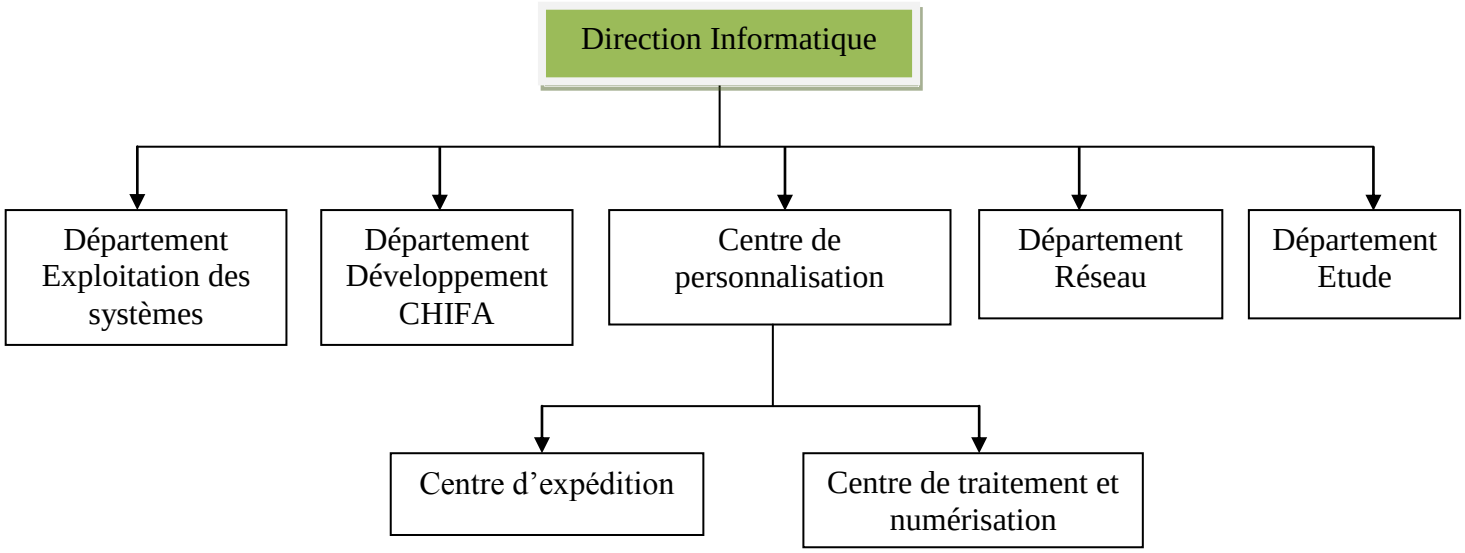

Figure 1.2 : Organigramme d'accueil de la Direction Informatique de la CNAS.

1.3 Le système existant :

Implanté au niveau de la direction informatique, la CNAS dispose de plusieurs bases de données fédérées, chacune concerne un service ou un département. Nous allons nous intéresser dans notre travail au département étude et développement.

1.3.1 Les bases de données :

Dans cette partie, nous allons décrire les différentes tables des deux bases de données concernées par l'étude, et qui sont gérées par un SGBD Oracle.

Base de données des allocations familiales (AF):

- **Table ASSURES :** regroupe tout les employés salariés affiliés a la caisse nationale des assurances pour les travailleurs salariés.

- **Table AYANT_DROIT :** tout les membres reliés a l'assuré et qui ont droit aux prestations.

- **Table ALLOCATAIRE :** c'est un assuré marié et ayant des enfants, qui a le droit aux allocations familiales.

- **Table AF_BENEFICIAIRE :** ce sont des ayants droits, dont l'ascendant est un allocataire.

- **Table AF_DROIT :** regroupe l'ensemble des droits aux allocations familiales de chaque allocataire.

- **Table AF_TP1 :** les montants en trop, attribués à un allocataire.

- **Table AF_PAIEMENT :** tous les paiements effectués par les centres au profit des allocataires ou les bénéficiaires.

- **Table CENTRE :** le centre payeur est l'endroit ou s'effectuent les paiements des allocations familiales.

- **Table AGENCEX :** regroupe les sièges de la CNAS au niveau de chaque wilaya.

- **Table WILAYA :** regroupe l'ensemble des wilayas du territoire national.

La figure 1.3 représente le modèle relationnel de la base de données des allocations familiales (AF).

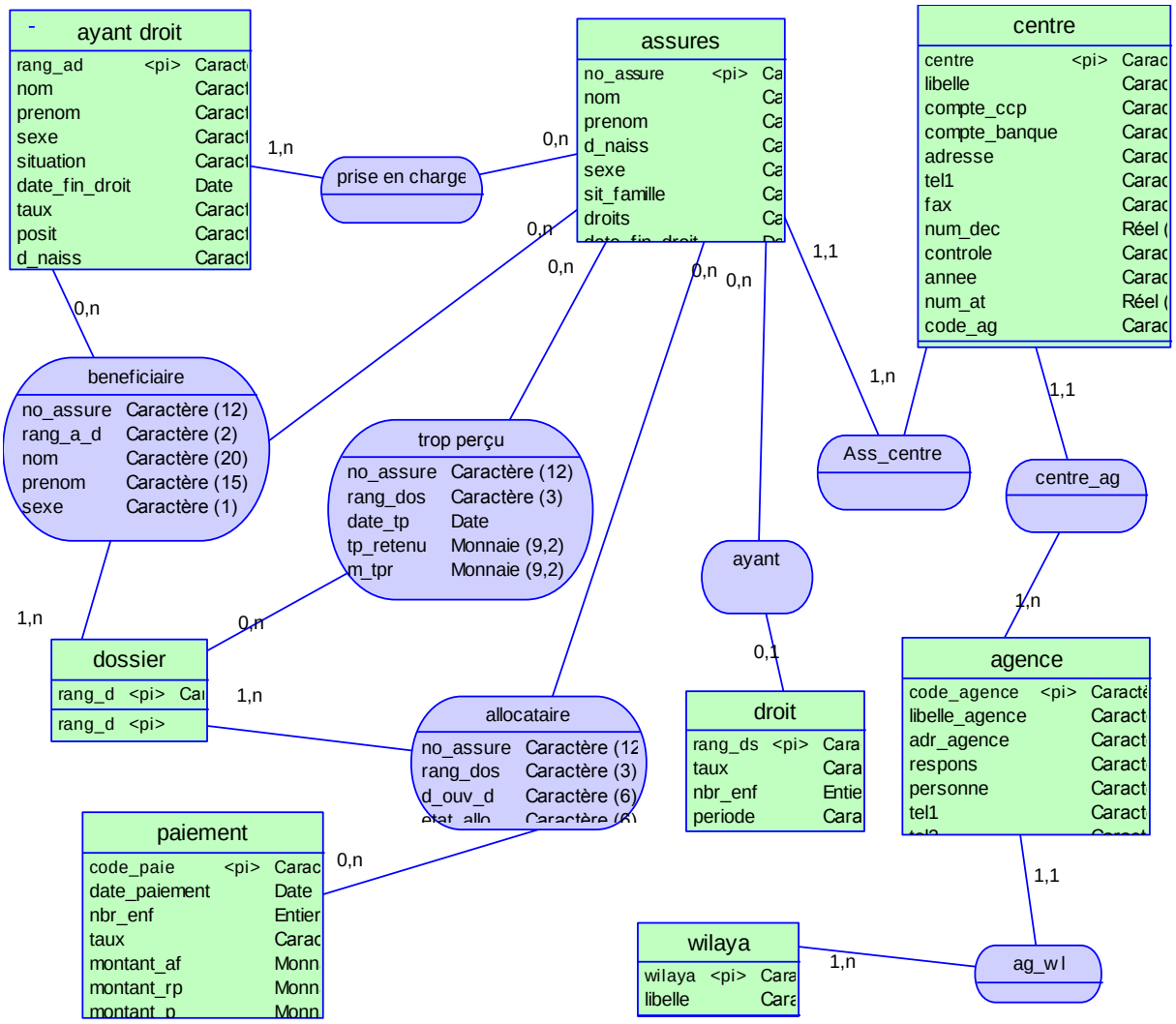

Figure 1.3 : Modèle relationnel de la base de données (AF).

Base de données des rentes :

- **Table ASSURES :** regroupe tout les employés salariés affiliés a la caisse nationale des assurances pour les travailleurs salariés.

- **Table AYANT_DROIT :** tout les membres reliés a l'assuré et qui ont droit aux prestations.

- **Table ACCIDENT_TRAVAIL :** tous les accidents qui se produisent aux assurés sur les lieux de travail.

- **Table RENTE :** contient toutes les informations sur les dédommagements d'un assuré non décédé suite à un accident de travail.

- **Table REVERSION_RENTE :** contient toutes les informations sur les dédommagements d'un assuré décédé suite à un accident de travail.

- **Table PAIEMENT_RENTE :** tous les paiements de rentes effectués aux profits des assurés ayant eu un accident.

- **Table CENTRE :** le centre payeur est l'endroit ou s'effectuent les paiements des rentes et des reversions.

- **Table AGENCEX :** regroupe les sièges de la CNAS au niveau de chaque wilaya.

- **Table WILAYA :** regroupe l'ensemble des wilayas du territoire national.

La figure 1.4 représente le modèle relationnel de la base de données des rentes.

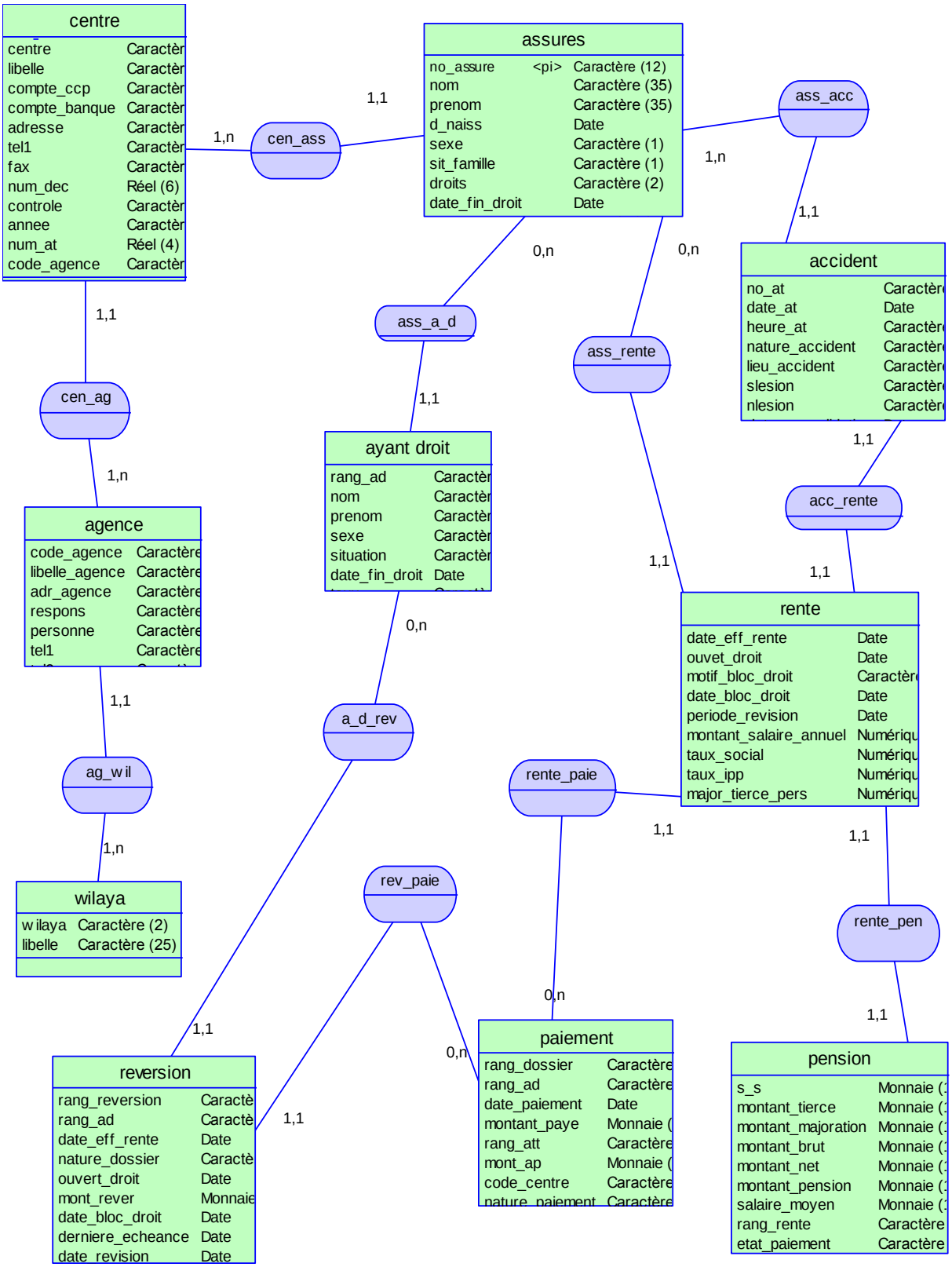

Figure 1.4 : Modèle relationnel de la base de données Rentes.

1.4 Critiques de l'existant :

Après avoir donné une description des différentes tables, cela nous a permis d'enregistrer certaines critiques de ce système :

- L'absence d'interconnexions entre les différents réseaux locaux rend le partage et la récolte des données difficile (duplication des tables ASSURES, AYANT_DROIT, CENTRE, AGENCEX et WILAYA sur les deux bases de données).
- Le système actuel ne permet pas la prise de décision.

1.5 Conclusion :

Dans cette partie, nous avons essayé de présenter l'organisme d'accueil, et le système actuel opérant au niveau de la direction informatique. Cette étude nous sera bénéfique dans le chapitre quatre ou on va modéliser notre entrepôt de données.

Chapitre 02

Introduction à l'informatique décisionnelle

Chapitre 02

Introduction à l'informatique décisionnelle

Sommaire

2.1 Introduction :

L'objectif du présent chapitre est de données une vision générale et simplifiée de l'informatique décisionnelle. Il renferme un bref descriptif de l'architecture d'un système décisionnel, en donnant la fonction de chacune de ses composantes.

2.2 Historique et genèse du business intelligence : [1]

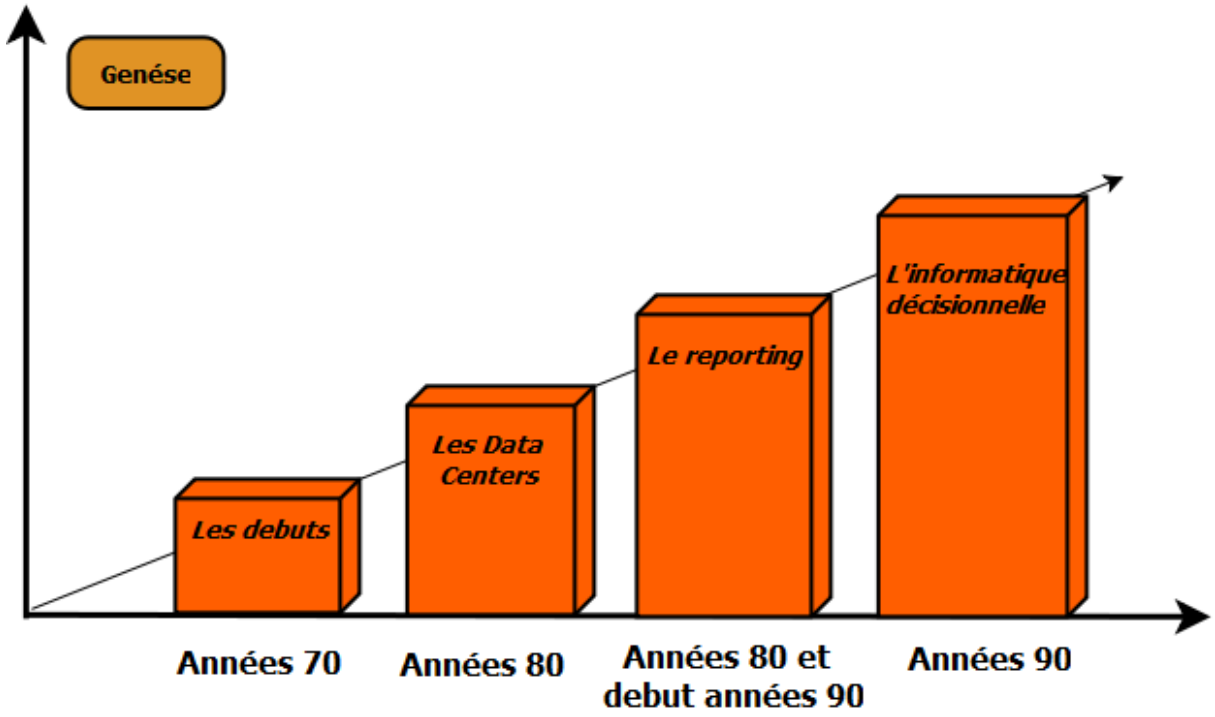

Figure 2.1 : La genèse de l'informatique décisionnelle.

2.2.1 Le commencement:

Début des années 70. L'informatique reste un " petit joujou " que les patrons précurseurs s'offrent. Pas à cause de la rentabilité du produit ou son efficacité, mais parce que sa amuse, sa fait penser au futur… Le besoin en information, à cette époque, commençait à apparaître car la concurrence commence à faire rage. On commence à comprendre : celui qui détient l'information détient le marché.

Toute la gestion des entreprises se faisait à la main, jusqu'au jour ou on entendit parler d'une machine nommée ordinateur et qui pouvait faire des calculs automatiques et les sauvegarder dans leur mémoire. Le boom commence, les entreprises s'informatisent et le besoin en information est assouvi. Les patrons peuvent connaître les résultats de leur activité journalière, et même mensuelle dans certain cas.

2.2.2 Les Data Centers:

Dans les années 80, les entreprises continuent à s'informatiser, mais les plus malignes commencent à accumuler beaucoup de données. Les Data Centers naissent. Des départements informatiques gérant des années et des années de données de production. Mais plus l'entreprise accumule des données, plus les analystes et les patrons veulent faire des analyses dessus. C'est normal, car c'est en fouillant dans les données qu'on peut savoir ce qui peut être amélioré dans l'entreprise. Manque de technologie et de maturité obligent, seul le service informatique peut créer des rapports à partir des sources de données. Un balai incessant entre la direction et le département informatique commence. En effet, le processus de recherche d'information implique fatalement un processus de type question - réponse - question. Chaque réponse entraîne un processus de réflexion qui, à son tour, amène une nouvelle question, et puisqu'à cette époque une question implique une demande de rapport. Nos pauvres informaticiens se retrouvent très rapidement surchargés. Et les systèmes de production aussi.

2.2.3 Le Reporting:

Devant le constat que la demande en information ne pourra jamais être pleinement satisfaite si le département informatique est tout le temps sollicité. Les informaticiens ont pensé des logiciels de génération de rapports. Ces logiciels (principalement à base de menus) contiendraient des rapports paramétrables que les utilisateurs pourront interroger à leur guise. La solution semble régler le problème, mais deux effets de bord vont apparaître suite à la naissance des systèmes de reporting :

- **La demande en information ne cessant de croître, les systèmes se retrouvent surchargés :** après l'apparition des outils de reporting, les utilisateurs se sont sentis plus indépendants.

Ils commencèrent à interroger la base de production sur une base régulière, ce qui entraînât une forte charge de travail sur les serveurs, qui, rappelons le, ne sont pas fait pour créer des rapports complexes, mais pour faire des opérations élémentaires dans la vie d'une entreprise (ajouter un client, une facture, consulter les dernières commandes d'un client, etc.). Cette surcharge fut réparée par des mises à jour matérielles sur les serveurs, mais cela revenait à traiter l'effet et non la cause.

- **La demande en information du marché rendait les décideurs insatisfaits des systèmes de reporting :** en effet, au début des années 90, l'insatisfaction à l'égard des informaticiens était grande. Car ces derniers étaient censés, avec les technologies de l'époque, pouvoir assouvir la soif de connaissance de l'entreprise. Mais les systèmes de reporting donnaient des rapports trop " grand public ", cela ne faisait que titiller encore plus leur curiosité.

2.2.4 Le début de la maturité. L'informatique décisionnelle:

Dans les années 90. Chercheurs en informatique et professionnels se sont penchés sur cette question clé qui est : comment aider les décideurs à prendre des décisions ?
Il fallait un environnement, et non un système, car la seule façon d'assouvir leur soif d'information est de leur permettre de fouiller eux même dans les données pour trouver ce qu'ils cherchent. Car la plupart du temps, les analystes ne savent pas ce qu'ils cherchent, leur travail est d'analyser l'entreprise pour l'améliorer, ils peuvent avoir des pistes, des doutes, des points de départ mais jamais rien de concret. Un processus de input - output ne serait donc pas pertinent pour eux. Il faut un environnement, mais que doit avoir cet environnement pour aider les décideurs à décider :

- **Simple :** les décideurs ne sont pas des gourous en informatique. L'environnement doit donc être assez simple et intuitif pour être manipulé par des non informaticiens.
- **Rapide :** le temps de nos décideurs est précieux. Pas question d'avoir une réponse des jours après l'avoir posé.

- **Gros volume de données :** la prise de décision au niveau des analystes et des patrons se fait à un très haut niveau d'abstraction. On analyse la tendance des ventes sur les trois dernières années pour déterminer des actions à entreprendre. L'environnement doit pouvoir gérer de très gros volumes de données.
- **Indépendant du système de production :** plus question de faire planter le système de production à cause d'une requête faite par un analyste.
- **Pour un membre restreint d'utilisateurs :** en effet, la prise de décision n'est la responsabilité que de quelques personnes dans l'entreprise. Le sommet de la pyramide.
- **Fiable et hétérogène :** l'environnement doit pouvoir compiler toutes les sources de données que possède l'entreprise. La conséquence est qu'un risque d'erreur dans les données peut se produire. Il s'agit de minimiser ce risque. La non fiabilité impliquera forcement le manque de confiance.

À partir de ces caractéristiques, des concepts, outils, logiciels se sont formés et articulés autour de ce nouveau domaine qui est l'informatique décisionnelle. Une nouvelle façon de concevoir les choses était née. On veut maintenant séparer le décisionnel du transactionnel. On a compris que les systèmes d'opération sont fait pour opérer et non pour prendre des décisions stratégiques. Le BI est né.

2.3 L'informatique décisionnelle :

Qu'est ce qu'une décision ?

Dans le Petit Robert, la décision est définie par « La fin de la délibération dans un acte volontaire de faire ou de ne pas faire une chose ». Prendre une décision signifie concevoir et s'engager à une stratégie d'allocation irrévocable de ressources de décision.
Une décision c'est le résultat d'un processus mental qui choisit une parmi plusieurs alternatives mutuellement exclusives. Une décision est prise pour résoudre un problème qui se pose à l'organisation ou à l'individu. Elle peut résulter d'une réponse à une modification de l'environnement ou bien pour saisir une opportunité. [2]

Selon Levine et Pomerol, le but d'une décision est de résoudre un problème qui se pose à l'organisation ou à l'individu. Les sciences cognitives ont perçu la décision comme étant bien plus qu'un choix, elle a été traitée comme un processus de résolution de problèmes.
On introduit donc le temps et le changement en plus des choix pour caractériser la décision.[3]

2.3.1 L'informatique décisionnelle : Business Intelligence : BI

L'informatique décisionnelle, par fois appelée tout simplement « le décisionnel » désigne l'exploitation des données de l'entreprise dans le but d'offrir une aide à la décision, c'est à dire de permettre aux responsables de l'entreprise d'avoir une vue d'ensembles de l'activité traitée, la compréhension du fonctionnement actuel et l'anticipation des actions pour un pilotage éclairé de l'entreprise.
Une définition de la Business Intelligence est donnée par : Howard Dresner « Le Décisionnel est le processus visant à transformer les données en informations et, par l'intermédiaire d'interrogations successives, transformer ces informations en connaissances. » [4]
« L'informatique décisionnelle est un concept désignant les moyens permettant de rassembler, intégrer, analyser les données de l'entreprise afin d'optimiser la prise de décision. Par extension, BI désigne les solutions logicielles combinant à des fins décisionnelle des fonctions d'interrogation de base de données, le reporting d'analyse multidimensionnelle (ou OLAP) de data mining et de visualisation des données. » [5]

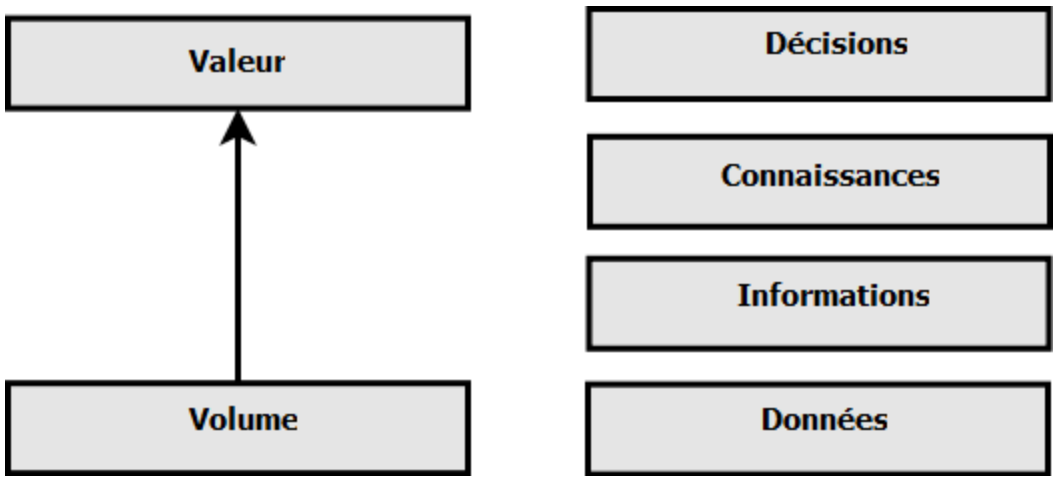

Figure 2.2 : Définition du décisionnel (BI).

2.3.2 Les différents systèmes de l'informatique décisionnelle :

Il existe deux types de systèmes décisionnels : Systèmes d'information pour dirigeant et Systèmes interactifs d'aide à décision.

2.3.2.1 Systèmes d'information pour dirigeant :

(Executive Information System : EIS) : Système d'information informatisé, spécialement conçu pour répondre aux besoins de la haute direction d'une entreprise et qui lui est réservé.
Le système d'information pour dirigeants se doit de fournir une information synthétisée et à jour qui donne un aperçu général continu des activités et des opérations de l'entreprise, à partir des sources externes et internes.
Il s'agit en quelque sorte du « tableau de bord » informatisé des cadres supérieurs, qui sert à la planification stratégique et à partir duquel on peut produire des rapports, des graphiques, faciles à consulter rapidement. Ce système fait partie du système d'information de gestion et diffère d'un système d'aide à la décision, dans la mesure où sa fonction principale est de fournir de l'information, la plupart du temps en temps réel, plutôt que des outils d'analyse et de prise de décision.
Toutefois, l'évolution technique fait que les deux types de systèmes se confondent un peu plus tous les jours. [6]

2.3.2.2 Systèmes interactifs d'aide à décision :

Les SIAD, ont pour but de rationaliser le processus décisionnel en effectuant l'intégration de multiples données, considération et technologie.
Le SIAD ne remplace pas le décideur mais il vient l'appuyer sur une analyse d'information. Il lui permet d'évaluer divers scenarios décisionnels lesquels le conduiront vers une décision raisonnées et justifiable. [7]
P.LEVINE & M.J POMEROL, dans leur livre *« Systèmes interactifs d'aide à la décision et systèmes experts »*, soulignent que la différence entre un SIAD et un autre système informatique est l'interactivité.

Cette interactivité est due au fait qu'une partie de contrôle dans la conduite du processus de décision est laissée à l'utilisateur.

Dans l'expression SIAD :

- Le terme « système » désigne un système de traitement de l'information ayant une architecture plus ou moins complexe.
- Le terme « interactif » implique l'existence d'un dialogue coopératif entre l'utilisateur et le système.
- L'expression « aide à la décision » quant à elle, sous-entend que c'est le décideur qui détient le contrôle dans le processus de décision et que c'est lui qui prend les décisions ultimes.

Dans ce type de systèmes, ce dernier ne fait qu'aider le décideur dans sa démarche. [3]

2.4 Les deux modes, Décisionnel et Opérationnel : [1]

Avant même de commencer à s'intéresser aux concepts du BI, il faut en capter l'essence, la philosophie. Cela ne fera que nous donner une meilleure vision et un meilleur sens de l'analyse. Expliquons d'abord chacun des deux mondes.

2.4.1 Le mode Opérationnel : OLTP

(OLTP pour On_Line Transactional Processing) Un système opérationnel traite des centaines voire des milliers de transactions par jours, chaque transaction est le reflet soit d'une mise à jour, soit d'une suppression ou encore d'un ajout de données nouvelles.
Les systèmes informatiques opérationnels sont faits pour assister les opérations d'une entreprise, ce sont des systèmes de gestion ou de production qui décrivent la vie de l'entreprise dans un environnement informatique, plus restreint, mieux gérable et plus flexible.

Les caractéristiques du système OLTP sont :

- **Grand nombre d'utilisateur :** Les utilisateurs des systèmes OLTP sont des acteurs qui alimentent en permanence et en quotidien les bases de données opérationnelles des organisations, ils sont destinés à tous les employés de l'entreprise. Les décideurs sont exclus du groupe car ils participent à un niveau plus élevé que la gestion quotidienne.

- **Données atomiques** : On entre un produit, une ligne de commande, une facture. Ce sont des éléments avec un grain très fin.

- **Extrêmement rapide :** Temps de réponse rapide.

- **Fermés :** On ne laisse pas la place à l'improvisation dans les OLTP les choix sont restreints, les utilisateurs sont guidés dans le processus.

- **Petite volumétrie des données :** Les systèmes de gestion gèrent des giga octets de données.

- **Transactionnels :** Fonctionnent en utilisant le principe de transaction.

- **Lecture, écriture et modification des données :** On peut ajouter de l'information, en supprimer si elle n'est pas utile pour la production et la modifier s'il existe des erreurs.

- **Projets peu risques :** Les projets du système OLTP sont maintenant bien maitrisés, les fonctionnalités et les besoins évidents, il y'a moins de risque d'échecs.

- **Fragmentés :** On entend par ici décentralisés.

- **Hétérogènes :** Les systèmes OLTP sont souvent des systèmes disparates en termes technologie utilisée. Il n'est pas rare d'avoir dans la même entreprise un système de gestion avec une base de données MYSQL et développé en PHP et un système de production avec une base de données ORACLE et développé en JAVA.

2.4.2 Le mode Décisionnel : OLAP

(OLAP pour On_Line Analytical Processing) Il faut dire que l'exploitation des données contenues dans les systèmes OLTP par les dirigeants de l'entreprise qui désirent améliorer leur prise de décision par une meilleure connaissance de leur propre activité, est devenue une de leur préoccupation essentielle.

Les caractéristiques du système OLAP sont :

- **Petit nombre d'utilisateurs :** Le système décisionnel s'adresse à la population parfaitement identifié, des dirigeants et patrons de l'entreprise.

- **Données générales et détaillés :** On s'intéresse ici aux chiffres par mois par année, par groupe de produit...etc. Les décideurs n'ont pas intérêt à voir la commande de tel ou tel client. Ils veulent voir l'ensemble de l'activité. Par contre, les analystes ont tout intérêt à pouvoir creuser dans les données pour trouver des fraudeurs par exemple.

- **Rapidité suggérée :** Il est clair que plus c'est rapide mieux c'est ! Mais dans la prise de décision stratégique, on ne calcule pas à la seconde. Un décideur peut bien attendre quelques heures pour avoir une information très complexe à produire. Mais dans la plupart des cas, les temps de réponses doivent être calculés en secondes.

- **Ouverts :** Contrairement au monde opérationnel, on laisse cour à la curiosité des utilisateurs, les environnements de BI doivent permettre d'accéder le plus simplement possible aux données et d'en faire tout ce qu'on veut !

- **Gros volume de données :** Les volumes à traiter sont plus importants que ceux gérés en transactionnel. Les environnements de BI doivent regrouper toutes les données de l'entreprise, de la ligne de commande au chiffre d'affaire annuel. Des années et des années d'accumulation de données générant des Terras octets qui doivent être gérés par les environnements de BI.

- **Non transactionnels :** Pas de processus rigide ici. L'utilisateur doit pouvoir commencer une analyse, revenir en arrière, démarrer une autre analyse en parallèle, envoyer un résultat à un collègue pour qu'il puisse creuser une autre piste.

- **Données en lecture seule :** A contraire du transactionnel, le décisionnel ne fera l'objet que d'une seule transaction dans la fréquence est généralement quotidienne. En revanche cette transaction représente des centaines de milliers d'enregistrements de plus, elle s'effectue exclusivement en mode d'ajout de données sans aucune modification ni suppression des données existantes.

- **Projet très risqués :** Les projets du système décisionnel sont très exposés à l'échec. En 2002, 40% des projets de BI ont échoué. Cela du au manque de connaissances dans le domaine.

- **Centralisés :** Joue le rôle de concentrateur des données afin de leur conférer une cohérence globale et partagée par l'ensemble des acteurs de l'entreprise.

Le tableau suivant montre la différence entre les systèmes transactionnels et les systèmes décisionnels du point de vue de leurs usages et des données utilisées.

Décisionnel	Opérationnel
Gros volumes de données à gérer.	Petits volumes de données à gérer.
Nombre d'utilisateur restreint (décideurs, analystes).	Utilisé par toute l'entreprise.
Processus ouverts pour permettre la génération de connaissance.	Processus fermés, transactionnels, le but est de donner le moins de marge de manœuvre possible.
Données en lecture seule.	Données en lecture - Écriture.
Rapidité moyenne comparée aux systèmes opérationnels.	Réponses très rapides.
Niveau de granularité très grand (on peut avoir des résumés sur ce qui c'est passé durant les 10 dernières années par exemple).	Niveau de granularité fin.
Centralisés (on veut avoir toutes les données de l'entreprise dans une seule structure).	Décentralisés.

Tableau 2.1 : Décisionnels & Opérationnels.

2.5 Architecture générale d'un environnement décisionnel :

Une architecture de BI est un ensemble de concepts, outils, méthodes et technologies (logicielles et matérielles) qui, une fois mises en relation, permettant de créer de la connaissance et répondre aux besoins stratégiques de l'entreprise. [5]

L'architecture suivante illustre le processus de construction d'un système d'aide à la décision. Qui consiste à récupérer des données depuis des bases d'informations existantes et à les stockées dans un entrepôt de données. Il faut pour cela mettre en place une procédure d'alimentation qui permettra de rassembler toutes les données utiles, quelle que soit leur origine : comptabilité, gestion commerciale, courrier électronique, serveurs externes ou encore internet.

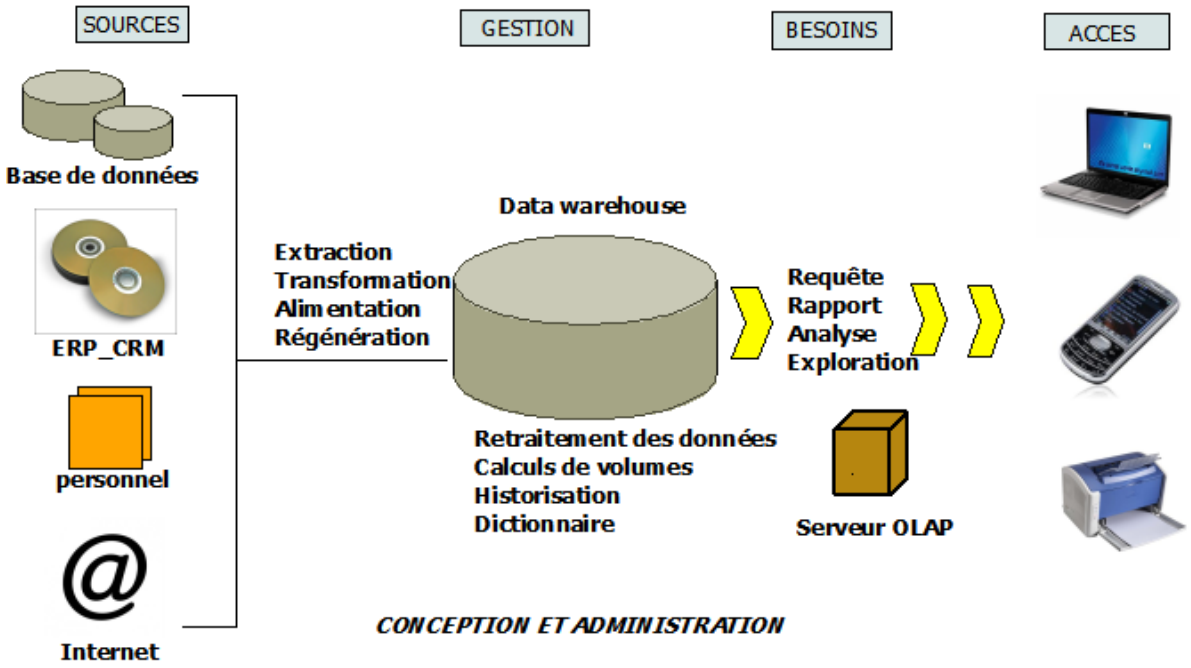

Figure 2.3 : L'architecture d'un environnement décisionnel.

L'architecture d'un environnement décisionnel met en jeu les éléments essentiels suivant :

- **Le système source :**

  Système opérationnel d'enregistrement dont la fonction consiste à capturer les transactions liées à l'activité. Les données de ce système source peuvent être nombreuses, hétérogènes (du point de vue structurelle ou sémantique) distribuées et autonomes. Elles peuvent être interne (base de production) ou externes (internet, base des partenaires). [8]

- **L'ETL (Extract, Transform, Load) :**

  C'est un système qui s'occupe d'analyser et extraire les données à partir des sources hétérogènes, pour ensuite nettoyer les données et en fin les charger dans l'entrepôt de données.

- **L'entrepôt de données : (Data Warehouse : DW)**

  Un entrepôt de données, est une vision centralisée et universelle de toutes les informations de l'entreprise. C'est une structure (comme une base de données) qui à pour but, contrairement aux bases de données, de regrouper les données de l'entreprise pour des fins analytiques et pour aider à la décision stratégique. La décision stratégique étant une action entreprise par les décideurs de l'entreprise et qui vise à améliorer, quantitativement ou qualitativement, la performance de l'entreprise. En gros, c'est un gigantesque tas d'informations épurées, organisées, historiées et provenant de plusieurs sources de données, servant aux analyses et à l'aide à la décision.

- **Serveur d'analyse OLAP : [9] [10]**

  OLAP est un acronyme de « On_line Analytical Processing ». Un serveur d'analyse OLAP et un moyen qui permet une analyse multidimensionnelle sur des bases de données afin de permettre aux analystes et aux décideurs de naviguer, découvrir les données du Data warehouse.

  Grâce à OLAP les utilisateurs peuvent créer des représentations multidimensionnelles (appelées hypercube ou cube «OLAP»).

  Le docteur Edgard Codd a écrit douze règles pour définir ce que signifie OLAP ; ces règles ont été publiées dans un article intitulé « Providing OLAP to user Analysts » aux Etats-Unis.

  Ces règles sont les suivantes :

  1. **Vue multidimensionnelle :**

     La base s'appuie sur un hyper cube (cube à n dimensions). L'administrateur définit une fois pour toute les dimensions qui représentent une façon de trier l'information et regroupent une liste de membres du même type (temps, produit, région, ...). L'analyse pourra ainsi être affinée dans le détail (vision pyramidale). L'utilisateur choisit deux ou trois critères à visualiser sous forme de tableau ou de cube.

Il peut également faire pivoter les axes d'analyse pour projeter les informations sous un angle différent. Ainsi, après avoir examiné les ventes par région, il peut permuter les axes pour une visualisation par mois. Ce critère est certainement le critère-clé du concept OLAP car il reflète la dimensionnalité de l'entreprise telle que la perçoivent ses membres qui ne sont autres que les utilisateurs du système.

2. **Transparence du serveur OLAP à différents types de logiciels:**
   Permet d'implanter le système OLAP sans affecter les fonctionnalités du système central. Ainsi, l'utilisateur doit pouvoir utiliser ses progiciels habituels (tableur, reporting, interface graphique, ...) sans percevoir la présence d'un outil OLAP. L'utilisateur ne doit pas se rendre compte de la provenance des données si celles ci proviennent de sources hétérogènes.

3. **Accessibilité à de nombreuses sources de données :**
   Les outils OLAP ont leur propre schéma logique de stockage de données physiques mais doivent accéder aux données et réaliser n'importe quelle conversion pour présenter une vue simple et cohérente des données. Ils doivent savoir d'où proviennent les données. En fait, par cette règle, le Dr Codd a essentiellement décrit les outils OLAP comme middleware, se plaçant entre des sources de données hétérogènes et une application OLAP.

4. **Performance du système de Reporting :**
   L'augmentation du nombre de dimensions ou du volume de la base de données ne doit pas entraîner de dégradation visible par l'utilisateur.

5. **Architecture Client/serveur :**
   La plupart des données pour OLAP sont stockées sur des gros systèmes et sont accessibles via des PC. Il est donc nécessaire que les produits OLAP soient capables de travailler dans un environnement Client/serveur.

6. **Dimensions Génériques :**

   Toutes les dimensions doivent être équivalentes en structure et en calcul. Il ne doit exister qu'une seule structure logique pour toutes les dimensions. Toute fonction qui s'applique à une dimension doit être aussi capable de s'appliquer à une autre dimension.

7. **Gestion dynamique des matrices creuses :**

   Le schéma physique des outils OLAP doit s'adapter entièrement au modèle d'analyse spécifique crée pour optimiser la gestion des matrices creuses. En effet, dans une analyse à la fois sur les produits et les régions, tous les produits ne sont pas vendus dans toutes les régions.

8. **Support multiutilisateurs :**

   Support des accès concurrents (récupération, mise à jour, ...), garantie de l'intégrité et de la sécurité afin que plusieurs utilisateurs puissent accéder au même modèle d'analyse ou encore créer des modèles d'analyse provenant des mêmes données de l'entreprise.

9. **Calculs à travers les dimensions :**

   Les opérations doivent pouvoir s'effectuer sur toutes les dimensions et ne doivent pas faire intervenir l'utilisateur pour définir un calcul hiérarchique.

10. **Manipulation intuitive des données :**

    L'utilisateur dispose d'une ergonomie de consultation, toute manipulation doit être accomplie via une action directe sur les cellules du modèle sans utiliser des menus ou des chemins multiples à travers l'interface utilisateur, on parle ici de navigation.

11. Souplesse et facilité de constitution des rapports :

L'analyse et la présentation des données sont plus simples lorsque les lignes, colonnes et cellules de données qui doivent être comparées, sont organisées de façon logique, par des regroupements correspondant à la vision de l'entreprise. C'est pour cela que l'élaboration de rapports doit être souple et conviviale.

12. Nombre illimité de niveaux d'agrégation et de dimensions:

Tout outil OLAP doit gérer au moins 5 à 10 dimensions. Le nombre de niveaux d'agrégation est illimité.

- **Zone d'outils d'accès :**

C'est l'ensemble des moyens fournis aux utilisateurs (end user) du Data Warehouse pour exploiter la zone de présentation des données en vue de la prise de décision.

2.6 Conclusion :

Comme nous l'avons déjà cité dans l'introduction de ce chapitre, son objectif était de donner une vision globale de l'informatique décisionnelle. Après une description brève des différents composants de l'environnement décisionnel à savoir : la source de données, le Data warehouse, l'ETL, le serveur OLAP, la zone d'outils d'accès. L'étude approfondie des Data warehouse fera l'objet du chapitre 3.

Chapitre 03

Les Entrepôts de Données

Chapitre 03

Les Entrepôts de Données

Sommaire

3.1 Introduction :

Nous avons vu dans le chapitre précédent ce qu'était le BI, ce que comprenait un environnement décisionnel et qu'il avait comme concept central l'entrepôt de données ou le Data Warehouse. Le concept d'entrepôt de données a pris forme au début des années 90, il est devenu depuis indispensable à la prise de décision dans les entreprises. Un entrepôt de données sert à concentrer les données disséminées dans l'entreprise et à les réunir en une série de structures documentées afin de permettre aux analystes et aux décideurs d'y accéder rapidement sans avoir besoin de connaissances techniques de programmation.

3.2 Présentation et définition :

3.2.1 Présentation :

Pourquoi les entrepôts de données ?

Dans une entreprise, le concept « d'entrepôt » a surgit suite a des besoins d'analyse de données pour faire face à la compétitivité sur le marché. Les données existantes (BD opérationnelles, de type On Line transaction Processing ou OLTP) provenant essentiellement des bases de production conçues pour des fonctions spécifiques de l'entreprise, ne se prêtent a ce type d'analyses. Les données pertinentes pour ces analyses sont disséminées sur diverses bases de données OLTP pas nécessairement compatibles entre elles et sont donc peu structurées pour l'analyse. Comme base de production elles sont focalisées sur les fonctions critiques de l'entreprise, ces systèmes sont donc peu adaptés à la vision à long terme et donc à la prise de décision.

3.2.2 Définition :

Un grand nombre de définitions ont été proposées et leur application concrète varie grandement.

De plus, plusieurs définitions sont biaisées par l'orientation qu'un auteur souhaite donner au concept de l'entrepôt de données afin de favoriser telle ou telle implantation de ce dernier.

Nous évoquerons de qui suit les définitions suivantes :

- Bill Inmon, Considéré comme étant le père fondateur des entrepôts de données et créateur de la : « Corporate Information Factory », Première entreprise évoluant dans ce domaine, définit l'entrepôt de données comme suit : « *L'entrepôt de données est une collection de données orientées sujet, intégrées, non volatiles et historiées, organisées pour le support d'un processus d'aide à la décision* ». [11]

- Ralph Kimball, Désigne un entrepôt de données comme «*Une copie de données correspondant à des transactions, spécialement structurées pour permettre de faire des requêtes et des analyses* ». [12]

- D'après Theodoratos et Bouzeghoub, Un entrepôt de données peut être vu comme « *un ensemble de vues matérialisées définies par des relations sur des sources de données distantes* ». [13]

  Cette définition semble être une simple explication d'une méthode pratique pour réaliser un entrepôt.

  En remarque que les vues Matérialisées ne permettent pas de résoudre tous les problèmes d'implémentation d'un entrepôt, même si elles peuvent faciliter le chargement des données. Cette définition ne tient pas compte de la nature historique d'un entrepôt, elle ne prévoit pas de méthode pour historier les données qui proviennent des sources de données de l'entrepôt. Des tables supplémentaires sont nécessaires pour créer un historique, car une vue matérialisée effectue une copie des données et supprime la version précédente. Les vues matérialisées calculent à l'avance des résultats de requêtes SQL dans une base de données et les conservent physiquement pour accélérer les traitements.

- D'après E. Kerkri, Un entrepôt de données est « *Une collection de technologies décisionnelles formant un environnement permettant aux décideurs de prendre de décisions plus pertinentes et plus rapides* ». [14]

Un entrepôt de données sert à concentrer les données disséminées dans l'entreprise et à les réunir en une série de structures documentées afin de permettre aux analystes et aux décideurs d'y accéder rapidement sans avoir besoin de connaissances techniques de programmation.

3.3 Caractéristiques des données de l'entrepôt :

Les caractéristiques principales des données d'un entrepôt de données peuvent se résumer comme suit :

- **Orientées sujet :**

  «*L'entrepôt de données est organisé autour des sujets majeurs de l'entreprise, contrairement aux données des systèmes opérationnels* ». [15]

  Les données sont structurées par thèmes.
  L'intérêt de cette organisation est de disposer de l'ensemble des informations utiles sur un sujet le plus souvent transversal aux structures fonctionnelles et organisationnelles de l'entreprise.

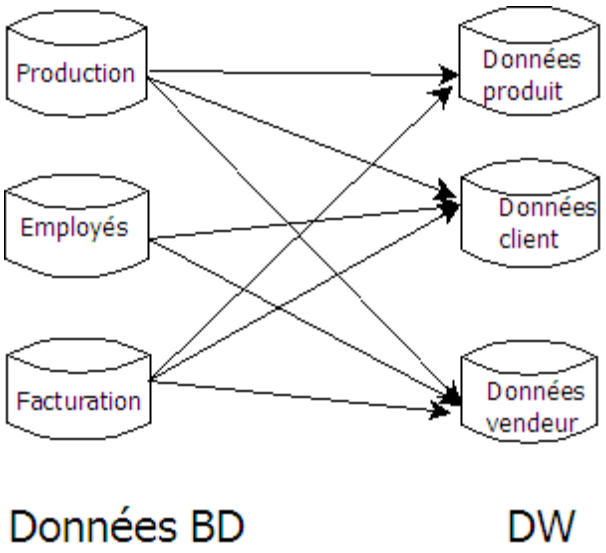

Figure 3.1 : Données orientées sujet.

- **Intégrées :**

  Les données alimentant l'entrepôt de données proviennent de multiples sources hétérogènes et disparates. Avant d'être intégrées dans l'entrepôt. Les données des systèmes de production doivent être converties, reformatées et nettoyées, de façon à avoir une seule vision globale dans l'entrepôt de données. L'intégration des données consiste à contenir leurs hétérogénéités pour donner au contenu de l'entrepôt de données une présentation homogène et pour garantir sa qualité avant d'être intégrées dans l'entrepôt de données. Les données doivent donc être mise en forme et unifiées afin d'en avoir un état cohérent. Cela nécessite un gros travail de normalisation, de gestion des référentiels et de cohérence. Une donnée doit avoir une description et un codage unique. Cette phase d'intégration ou de nettoyage des données est très complexe et représente 60 à 90% de la charge totale d'un projet. Par exemple, la consolidation de l'ensemble des informations concernant un client est nécessaire pour donner une vue homogène de ce client.

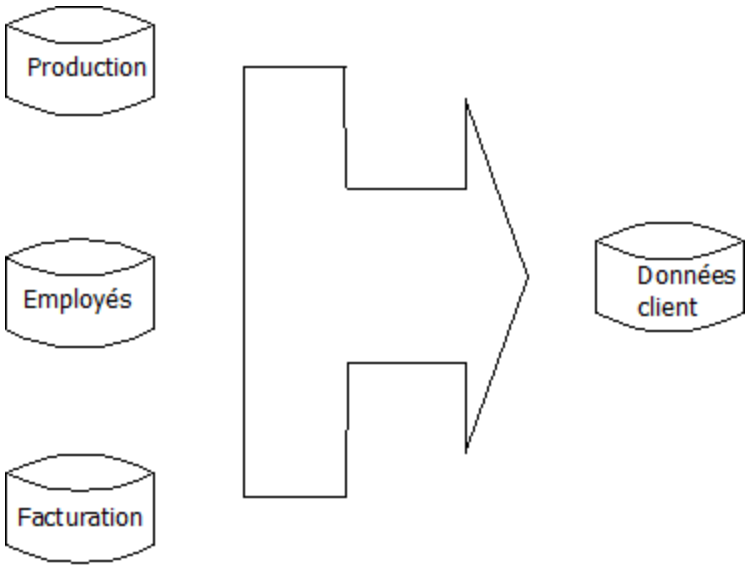

Figure 3.2 : Données intégrées.

- **Non volatiles :**

Les données de l'entrepôt sont essentiellement utilisées en mode consultation, elles sont très rarement modifiées. Un entrepôt de données doit garantir qu'une requête lancée à différentes dates sur les mêmes données donne toujours les mêmes résultats. De plus, les données d'un entrepôt sont mises à jour périodiquement, ce ne sont donc pas des informations en temps réel. La non volatilité est en quelque sorte une conséquence de l'historisation.

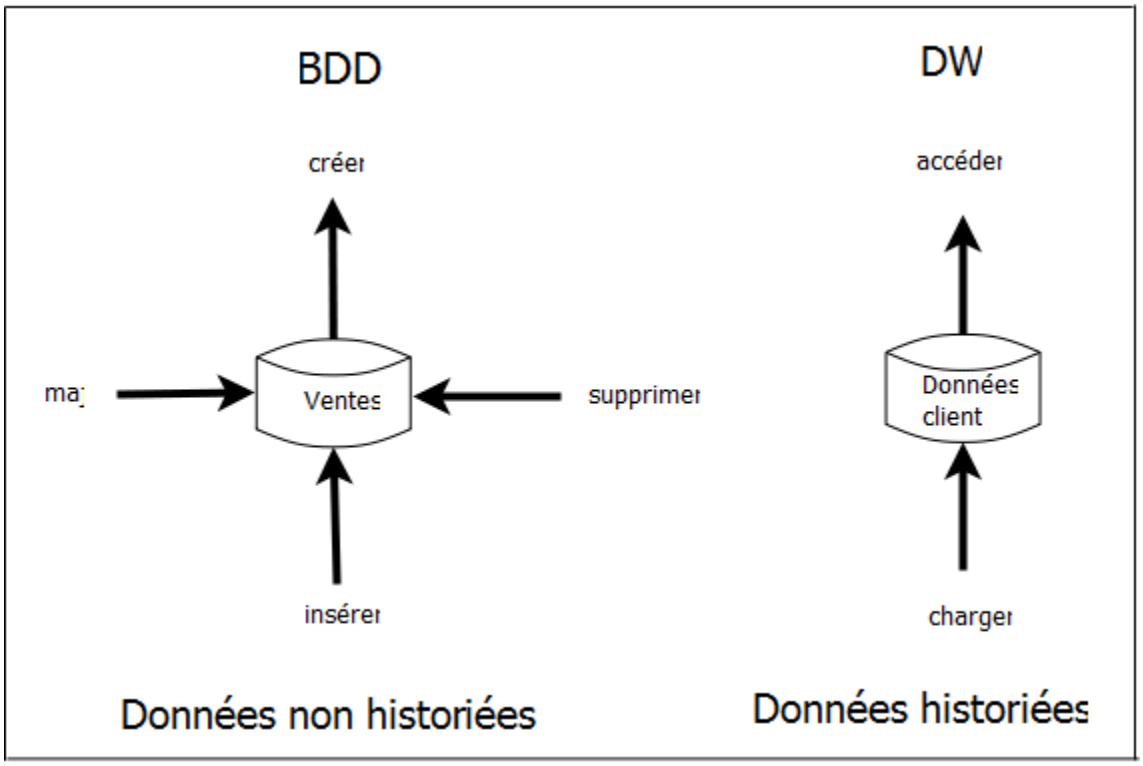

Figure 3.3 : Données non volatiles.

- **Historiées :**

  En générale, dans un système de production, la mise à jour des données se fait lors de chaque nouvelle transaction. Une mise à jour annule et remplace l'ancienne valeur. Dans un entrepôt de données, la donnée ne doit jamais être mise à jour en mode « annule et remplace ». Les données sont historiées et donc datées. L'historisation est nécessaire pour suivre dans le temps l'évolution des différentes valeurs des indicateurs à analyser. Ainsi, un référentiel temps doit être associé aux données afin de permettre l'identification de valeurs précises dans la durée.

- **Aide à la décision :**

  Un entrepôt de données est destiné à l'aide à la décision, ce qui fait que les traitements qui s'y appliquent sont de natures différentes de ceux des systèmes transactionnels, on parle d'OLAP par opposition à OLTP.

La plupart des traitements transactionnels en ligne n'impliquent que quelques données, occasionnent des changements dans la base de données et requièrent une réponse presque instantanée alors que les traitements mis en œuvre pour l'aide à la décision impliquent la lecture de nombreuses données mais n'entraînent pas de changement dans la base de données et ne requièrent pas une réponse instantanée.

3.4 Objectifs de l'entrepôt de données :

- **Construire de l'information utile pour l'aide à la prise de décision :**

  Les entrepôts de données sont des systèmes conçus pour l'aide à la prise de décision, autrement dit transformer un système d'information qui avait une vocation de production en un SI décisionnel. [16]

- **Retrouver et analyser l'information facilement et rapidement:**

  La Gestion et visualisation des données doit être rapide et intuitive. Pour cela, il est nécessaire de retrouver et d'analyser rapidement les données provenant de diverses sources.

- **Regrouper des informations provenant de sources diverses:**

  Le DW permet aux entreprises d'avoir un système qui regroupe au même endroit les informations qui jusqu'à lors étaient éparpillées dans une multitude d'applications ou systèmes différents et souvent non intégrés entre eux.

- **Organiser les données :**

  Le DW est nécessaire Pour gérer une masse de données de plus en plus conséquente, provenant de sources hétérogènes, les intégrer et les stocker pour donner à l'utilisateur une vue orientée métier.

- **Intégrer des différentes Bases de Données :**

  Un data warehouse a pour objectif principal l'intégration, ce qui veut dire qu'il doit constituer un système d'information décisionnel à l'échelon de l'entreprise, donc transversal pour reprendre un terme de l'organisation, résoudre alors le problème d'hétérogénéité des différentes sources.

- **Le stockage et la centralisation de ces données dans un entrepôt constituent un support efficace pour l'analyse.**

- **Supporter un processus d'analyse en ligne.**

3.5 Architecture générale d'un entrepôt de données : [16]

L'architecture d'un entrepôt de données repose souvent sur un SGBD séparé du système de production de l'entreprise qui contient les données de l'entrepôt. Le processus d'extraction des données permet d'alimenter périodiquement ce SGBD. Néanmoins avant d'exécuter ce processus, une phase de transformation est appliquée aux données opérationnelles. Celle-ci consiste à les préparer (mise en correspondance des formats de données), les nettoyer, les filtrer,..., pour finalement aboutir à leur stockage dans l'entrepôt.

L'architecture décisionnelle d'un entrepôt de données se présente comme suit :

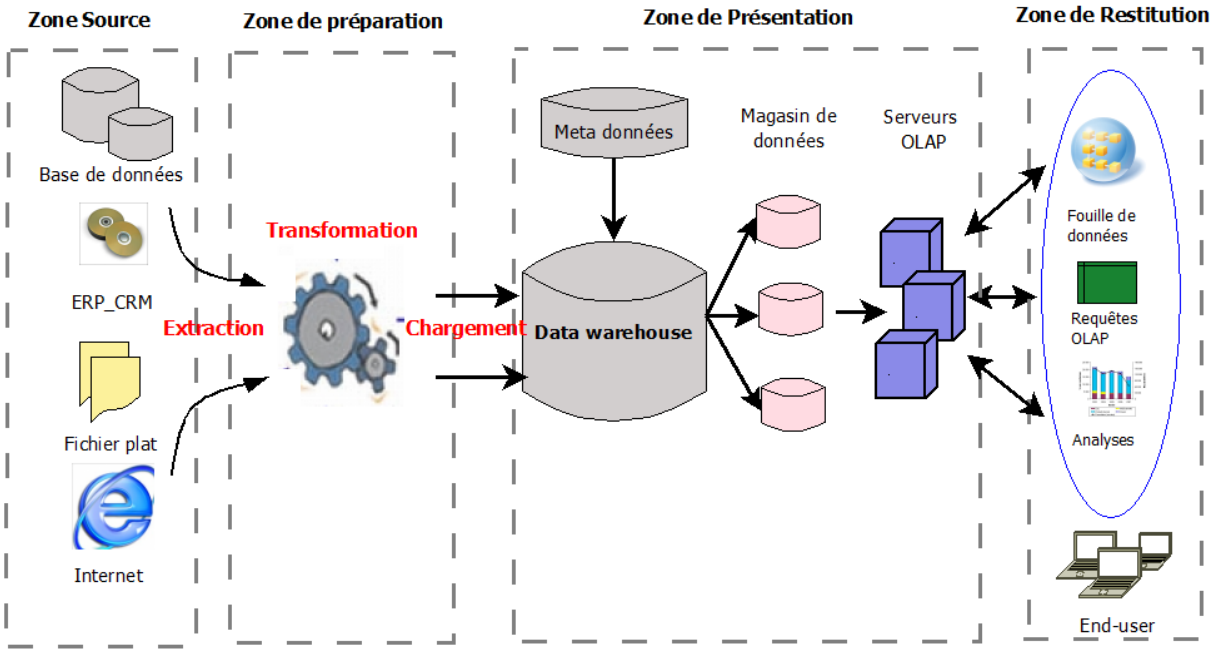

Figure 3.4 : Architecture générale d'un entrepôt de données.

Dans cette partie nous allons étudier les éléments qui composent un environnement complet d'un Data warehouse. Il est utile de comprendre les composants avant de commencer à les combiner pour créer un Data warehouse. Chaque composant sert d'une fonction spécifique. Comme l'illustre la figure 2.4 il ya quatre composants : les sources des données, la zone de préparation de données, la zone de présentation de données et les outils d'accès aux données.

3.5.1 La Sources de données : (Data Sources)

Ce sont les bases de production (relevant d'un système OLTP), les fichiers plats (fichiers Excel...) qui correspondent à des sources internes, mais elles peuvent également être d'origine externe (Internet, bases de partenaires, etc...).

3.5.2 La zone de préparation de données :

« Ensemble de processus qui nettoient, transforment, combinent, archivent, suppriment les doublons, c'est-à-dire prépare les données sources en vue de leur intégration puis de leur exploitation au sein de l'entrepôt de données ». [17]

La zone de préparation des données est analogue à la cuisine d'un restaurant, où les produits alimentaires bruts sont transformés en un bon repas. C'est là que la plupart des opérations de nettoyage et de préparation des données a eu lieu avant leur chargement dans le data warehouse.

Remarque :

« Un point très important, dans l'aménagement d'un entrepôt de données, est d'interdire aux utilisateurs l'accès à la zone de préparation des données, qui ne fournit aucun service de requête ou de présentation ». [18]

3.5.3 L'entrepôt de données :

« L'entrepôt de données correspond à la source de données interrogeable de l'entreprise ».
Il est alimenté par la zone de préparation des données. [18]

3.5.4 Magasin de données :

« Le magasin de données ou Data Mart est défini comme un sous ensemble logique d'un entrepôt de données ».

Les magasins de données sont des extraits de l'entrepôt qui se focalisent sur un sous ensemble de sujets particuliers (orientés métiers, activités, etc.).
Ils sont souvent utilisés comme des éléments supplémentaires et les données extraites sont adaptées à une classe d'utilisateurs (décideurs) ou à un usage particulier (satisfaire les besoins d'analyse au niveau des départements d'une entreprise).
Contenant un volume moindre de données organisées suivant un modèle spécifique qui permet des analyses faciles et rapides à des fins de prise de décision.

L'architecture suivante illustre le positionnement architectural d'un Data Marts et d'un Data Warehouse :

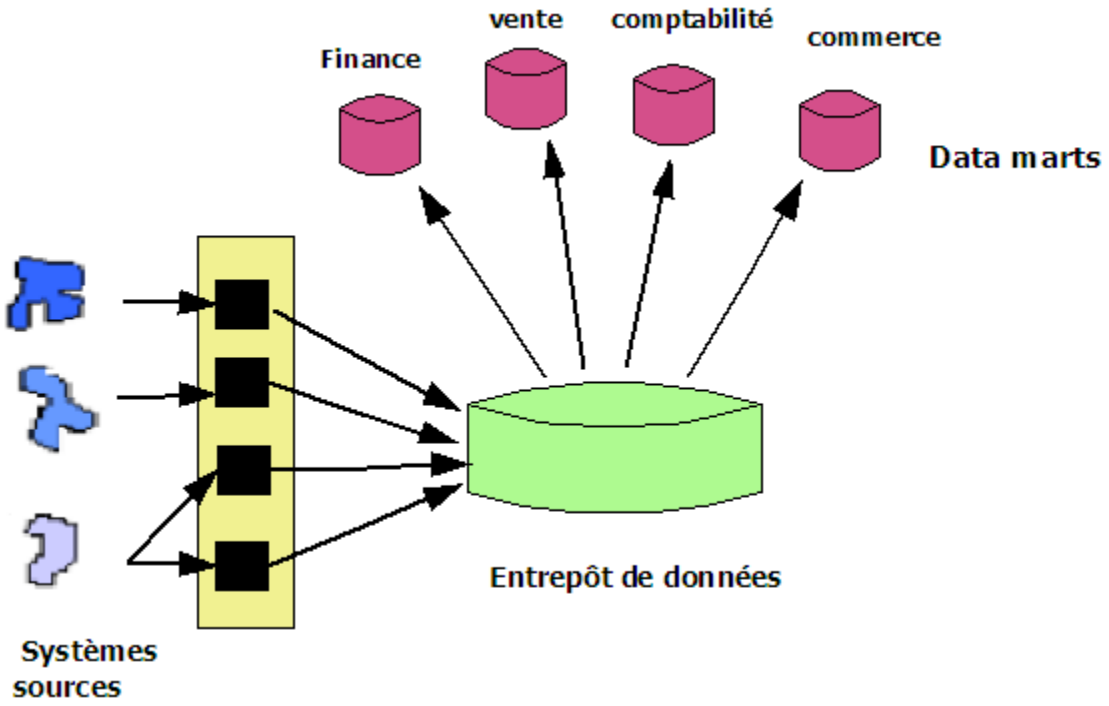

Figure 3.5 : Positionnement architectural d'un data mart et d'un Data warehouse

3.5.5 Les Métadonnées :

Ce sont toutes les informations de l'environnement de l'entrepôt de données qui ne constituent pas les données proprement dites. Ce sont les « données sur les données ».

3.5.6 Le serveur de présentation :

« Machine cible sur laquelle l'entrepôt de données est stocké et organisé pour répondre en accès direct aux requêtes émises par des utilisateurs ». [17]

Sur le serveur de présentation, Les données seront stockées et représentées sous une forme dimensionnelle, de façon à facilité l'accès pour l'utilisateur final.

Dans la majorité des cas le serveur est basé sur une base de données relationnelle, de sorte que les tables y soient organisées sous forme de schéma en étoile ou en flocon.

3.5.7 Portail de restitution :

« C'est la part publique de l'entrepôt de données ». [17]. Il représente ce que voient les utilisateurs, les outils avec lesquels ils travaillent. Les outils d'accès au données c'est l'ensemble des moyens fournis aux utilisateurs du Data warehouse pour exploiter la zone de présentation des données en vue de la prise de décision.

Les outils d'accès aux données comprennent : la navigation dans l'entrepôt et dans les métas données. Il existe sur le marché différents outils pour l'aide à la décision, comme les outils de fouille de données ou Data Mining (pour découvrir des liens sémantiques), outils d'analyse en ligne (pour la synthèse et l'analyse des données multidimensionnelles) exemple «OLAP», outils d'interrogation (pour faciliter l'accès aux données en fournissant une interface conviviale au langage de requêtes).

3.6 Structure des données d'un entrepôt de données :

L'entrepôt de données a une structure bien définie, selon différents niveaux d'agrégation et de détail des données. Cette structure est illustrée dans la figure suivante : [19]

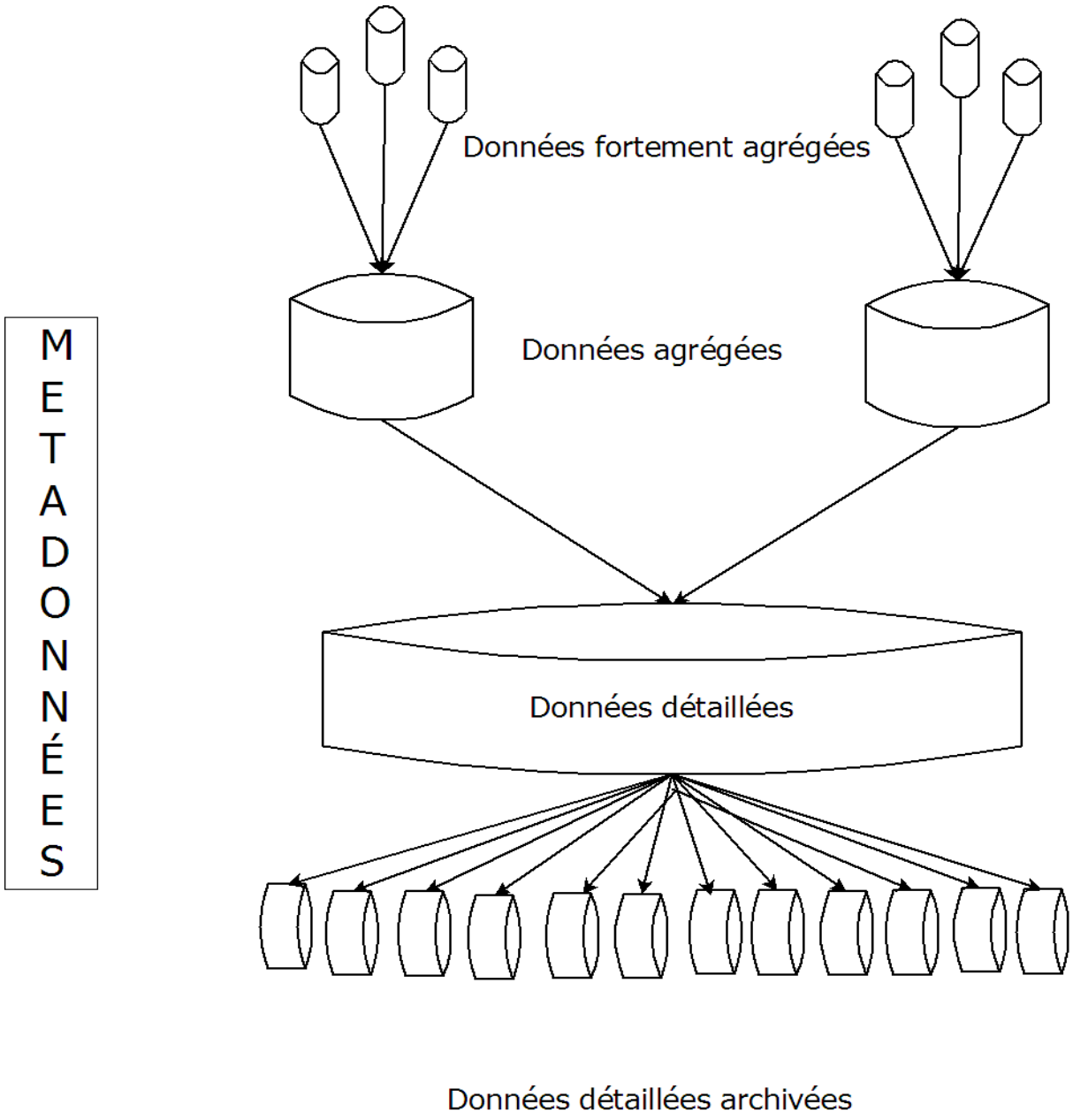

Figure 3.6 : Structure des données d'un Data Warehouse.

3.6.1 Données détaillées :

Elles reflètent les événements les plus récents. Les intégrations régulières des données issues des systèmes de production vont habituellement être réalisées à ce niveau. Généralement volumineuses car elles sont d'un niveau détaillé. Les volumes à traiter sont plus importants que ceux gérés en transactionnel.

Le niveau de détail géré dans l'entrepôt de données n'est pas forcément identique au niveau de détail géré dans les systèmes opérationnels. La donnée insérée dans l'entrepôt de données peut être déjà une agrégation ou une simplification d'informations tirées du système de production. Fréquemment consultées.

3.6.2 Données détaillées archivées :

Ce sont les anciennes données rarement sollicitées, généralement stockées dans un disque de stockage de masse, peu coûteux, à un même niveau de détail que les données détaillées. Un des objectifs de l'entrepôt de données est de conserver en ligne les données historiées.
Chaque nouvelle insertion de données provenant du système de production ne détruit pas les anciennes valeurs, mais créé une nouvelle occurrence de la donnée.

3.6.3 Données agrégées :

Les données agrégées à partir des données détaillées, correspondent à des éléments d'analyse représentatifs des besoins des utilisateurs. Elles constituent déjà un résultat d'analyse et une synthèse de l'information contenue dans le système décisionnel. Ces données doivent être accessibles de façon simple et permettent aux utilisateurs de naviguer suivant une logique intuitive, avec des performances optimales. La définition complète de l'information doit être mise à la disposition de l'utilisateur pour une bonne compréhension. Dans le cas d'un agrégat, l'information est composée du contenu présenté (moyenne de vente) et de l'unité (par mois, par produit).

3.6.4 Données fortement agrégées :

Les données sont agrégées à partir des données détaillées, à un niveau d'agrégation plus élevé que les données agrégées.

3.6.5 Les métadonnées :

Il s'agit « de données sur les données », ce sont les informations relatives à la structure des données, les méthodes d'agrégation et le lien entre les données opérationnelles et celles du Data Warehouse. Les données fédérées dans l'entrepôt de données proviennent de sources très hétérogènes, Les métadonnées constituent un dictionnaire et une véritable aide en ligne permettant de connaitre l'information contenue dans l'entrepôt de données. Elles sont idéalement intégrées dans un référentiel.

Les métadonnées doivent renseigner sur :

- Les données entreposées, leur format, leur signification, leur degrés d'exactitude.
- Les processus de récupération/extraction dans les bases sources.
- La date du dernier chargement du data warehouse.
- L'historique des données sources et de celles du Data warehouse.

Les principales informations sont destinées :

- A l'utilisateur (sémantique, localisation).
- Aux équipes responsables des processus de transformation des données du système de production vers l'entrepôt de données (localisation dans les systèmes de production, description des règles, processus de transformation).
- Aux équipes responsables des processus de création des données agrégées à partir des données détaillées.
- Aux équipes d'administration de la base de données (structure de la base implémentant l'entrepôt de données).
- Aux équipes de production (procédures de changement, historique de mise à jour,. . .).

3.7 Construction d'un entrepôt de données :

L'étape de construction d'un entrepôt de données est précédée d'une étude préalable. Cette étude préalable doit correspondre à la couverture des attentes de l'utilisateur à travers de l'étude des besoins avec la réalité des informations disponible « Après cette étude, le concepteur doit prendre neuf décisions majeures qui jalonnent la conception de l'entrepôt ». Ces décisions portent sur les points suivants : [17]

1 **Choisir les processus d'activité à modéliser** : On va identifier les processus majeurs de l'entreprise dans lesquelles les informations sont collectées au profit de l'entrepôt à partir des applications existantes (par exemple : les commandes, la facturation, les ventes), une fois les processus identifiés, une ou plusieurs tables de faits sont construites.

2 **Choisir le grain de chaque table de faits :** Le grain est la signification précise d'un enregistrement du plus bas niveau dans la table de faits.

3 **Choisir les dimensions de chaque table de faits** : Le choix des dimensions s'accompagne de la définition de tous les attributs textuels (les champs) qui garniront la table de dimension. Exemple de dimension (temps, produit, magasin.).

4 **Choisir les faits mesurés que contiendra chaque enregistrement de la table de faits** : Après le choix des dimensions qu'est le point clé de la conception, on peut le suivre par la définition de tous les faits mesurés de la table de faits.

5 **Choisir les attributs des dimensions, avec des descriptions complètes et la terminologie adéquate :** A ce stade la conception de la structure logique est terminée, et les autres étapes concernent la structure physique.

6 **Comment suivre les dimensions à évolution lente ? :** La dimension client change constamment, car les humains changent de nom, se marient et divorcent et change d'adresse, et dans une compagnie d'assurances il est essentiel de savoir le statut de la personne assurée au moment d'un sinistre passé et non le statut actuel. Ce type de dimensions sont appelées les dimensions à évolution lente. Pour le traitement de ces changements, on a trois solutions :

- Remplacer les valeurs anciennes, et renoncer à les suivre : Elle consiste à ignorer la possibilité de suivre les événements ou situations passées.

- Créer un nouvel enregistrement : Elle consiste à créer un nouvel enregistrement de dimension, ce qui permet de partitionner l'historique selon le temps. Mais le problème rencontré dans cette technique est qu'on ne peut pas faire des comparaisons significatives entre les différentes périodes.

- Ajouter un champ : Elle consiste à ajouter un champ « actuel » dans la dimension, ce qui permet les comparaisons en amont et en aval du changement, mais ne permet plus vraiment le partitionnement de l'historique.

7 **Choix des agrégats :** Le choix d'un agrégat pour chaque fait mesuré nécessite une étude pour toutes les dimensions concernées.

8 **L'étendue historique de la base de données :** On peut permettre à l'historique de s'accumuler au-delà des trois ans envisagés pour n'importe quel entrepôt de données.

9 **L'urgence avec laquelle les données doivent être extraites et chargées dans l'entrepôt :** Généralement le chargement ce fait chaque jour, ou attendre jusqu'à la fin de la semaine.

3.8 Implémentation d'un entrepôt de données :

Il existe trois façons d'implémenter un entrepôt de données :

3.8.1 L'implémentation selon l'architecture réelle :

Elle est généralement retenue pour les systèmes décisionnels. Le stockage des données est réalisé dans un SGBD séparé du système de production. Le SGBD est alimenté par des extractions périodiques.
Avant le chargement, les données subissent d'importants processus d'intégration, de nettoyage, de transformation.
L'avantage est de disposer des données préparées pour les besoins de la décision répondant aux objectifs de l'entrepôt de données.
Les inconvénients sont le coût de stockage supplémentaire et le manque d'accès en temps réel.

3.8.2 L'implémentation selon l'architecture virtuelle :

Cette architecture n'est pratiquement pas utilisée pour l'entrepôt de données. Les données résident dans le système de production. Elles sont rendues visibles par des produits middleware ou par des passerelles.
Il en résulte deux avantages : pas de coût stockage supplémentaire et l'accès se fait en temps réel. L'inconvénient est que les données ne sont pas préparées.

3.8.3 L'implémentation selon l'architecture remote :

C'est une combinaison de l'architecture réelle et de l'architecture virtuelle. L'objectif est d'implémenter physiquement les niveaux agrégés afin d'en faciliter l'accès et de garder le niveau de détail dans le système de production en y donnant l'accès par le biais de middleware ou de passerelle.

	Architecture Réelle	**Architecture Virtuelle**	**Architecture Remote**
Utilisation	Retenue pour les systèmes décisionnels	Rarement utilisée	Rarement utilisée
Stockage	SGBD séparé du système de production, alimenté par des extractions périodiques	Données résident dans le système de production	Combinaison des architectures réelle et virtuelle
Avantage	Données préparées pour les besoins de la décision	Pas de cout de stockage supplémentaire, accès en temps réel	
Inconvénient	Coût de stockage supplémentaire, manque d'accès en temps réel	Données non préparées	

Tableau 3.1 : Synthèse sur les architectures de stockage.

3.9 Modélisation multidimensionnelle des données d'un entrepôt de données :

Au niveau de l'entrepôt, pour pouvoir exploiter facilement les données, le concepteur doit réaliser une classification par sujet fonctionnel plutôt que par application, donc on peut dire qu'un entrepôt de donnée regroupe un ensemble de sujets principaux de l'organisation.

La modélisation dimensionnelle est une méthode de conception logique qui vise à présenter les données sous une forme standardisée intuitive et qui permet des accès hautement performants.

La modélisation multidimensionnelle consiste à considérer un sujet analysé comme un point dans un espace à plusieurs dimensions. Les données sont organisées de manière à mettre en évidence le sujet analysé et les différentes perspectives de l'analyse.

3.9.1 Concepts de modélisation multidimensionnelle :

« La modélisation multidimensionnelle a donné naissance aux concepts de fait et de dimension ». [12]
La modélisation multidimensionnelle se traduit par deux concepts : le concept de fait et le concept de dimension.

- **Concept de fait :**

  « Le fait modélise le sujet de l'analyse. Un fait est formé de mesures correspondant aux informations de l'activité analysée. Les mesures d'un fait sont numériques et généralement valorisées de manière continue». [12]
  Le sujet analysé est représenté par le concept de fait. Une table de faits est la table centrale d'un modèle dimensionnel, elle assure les liens entre les dimensions. Elles comportent des clés étrangères, qui ne sont autres que les clés primaires des tables de dimension.

Exemple :

Prenons l'exemple d'un fait de **vente** constitué des mesures d'activité suivantes : quantité des produits vendus et montant total des ventes :

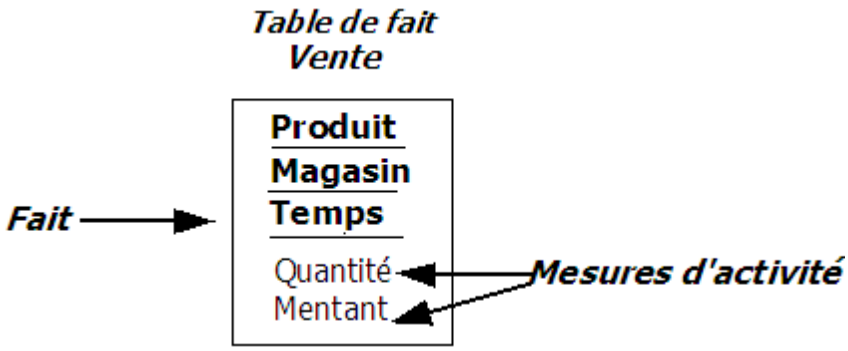

Figure 3.7 : Exemple de table de fait vente.

- **Concept de dimension :**

« Le sujet à analyser, c'est-à-dire le fait, est analysé suivant différentes perspectives. Ces perspectives correspondent à une catégorie utilisée pour caractériser les mesures d'activité analysées ». [20]

Les tables de dimension sont les tables qui raccompagnent une table de faits, elles contiennent les descriptions textuelles de l'activité. Une table de dimension est constituée de nombreuses colonnes qui décrivent une ligne. Une dimension modélise une perspective de l'analyse. C'est grâce à cette table que l'entrepôt de données est compréhensible et utilisable elle permet des analyses en tranches et en dés.

Une dimension est généralement constituée : d'une clé artificielle, une clé naturelle et des attributs. « Une table de dimension établit l'interface homme / entrepôt, elle comporte une clé primaire ». [18]

Exemple :

Le fait à la **Figure 3.7** peut être analysé suivant différentes perspectives correspondant à trois dimensions : temps, produit et magasin.

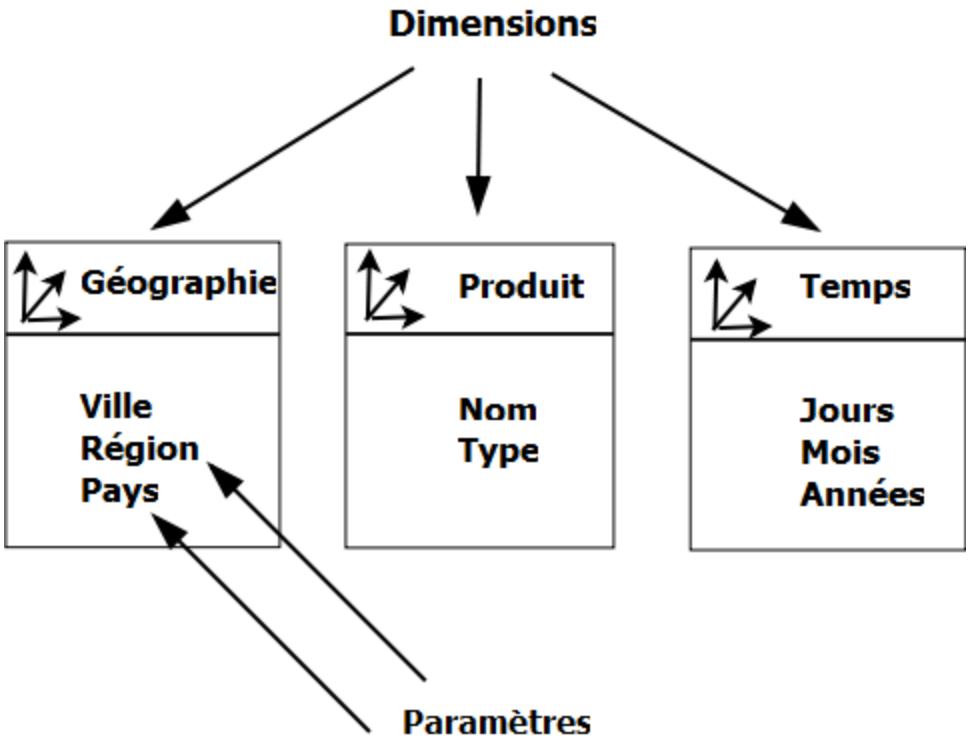

Figure 3.8 : Exemple de table de dimension.

- **Concept hiérarchie :**

  Une hiérarchie organise les paramètres d'une dimension selon une relation "est-plus-fin" conformément à leur niveau de détail. Par exemple, pour la dimension " Géographie ", ces paramètres sont organisés suivant l'hiérarchie suivante : (Ville _ Région _ pays). La hiérarchie sert lors des analyses pour restreindre ou accroître les niveaux de détail de l'analyse.

3.9.2 Les modèles multidimensionnelles : [21]

- **Modèle en étoile :**

Dans le modèle en étoile on trouve au centre la table de faits. L'identifiant de cette table est une clé multiple composée de la concaténation des clés de chacune des dimensions d'analyse. Autour de la table de faits on trouve tous les paramètres qui caractérisent les dimensions d'analyse.

Ces caractéristiques sont regroupées dans des tables de dimension.
La force de ce type de modélisation est sa lisibilité et sa performance.

- **La lisibilité :** La finalité de ce modèle est très évidente et définit clairement les indicateurs d'analyse.

- **La performance** : Les chemins d'accès à la base de données sont prévisibles.

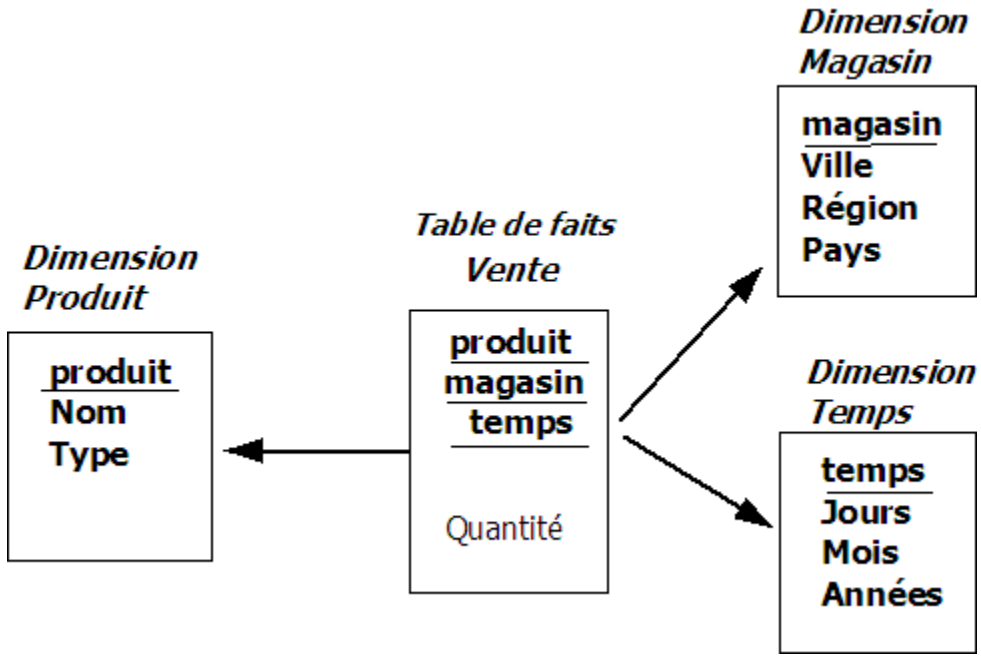

Figure 3.9 : Schéma en étoile.

- **Modèle en flocon :**

  Il correspond à un schéma en étoile dans lequel les dimensions ont été normalisées, réduisant chacune d'elles et faisant ainsi apparaitre des hiérarchies de dimension de façon explicite.

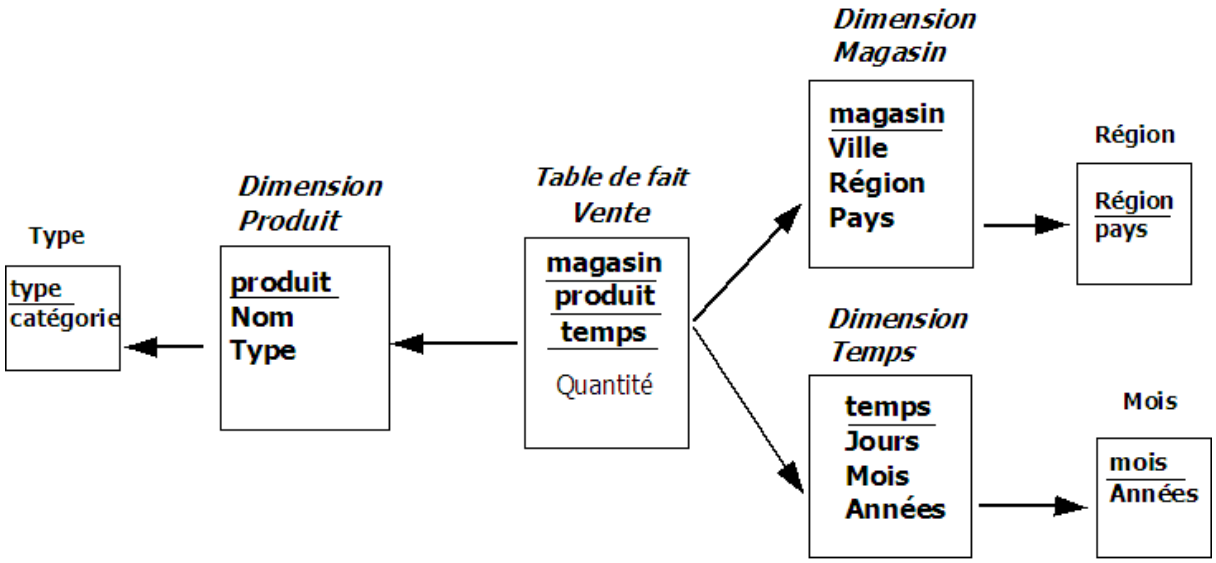

Figure 3.10 : Schéma en flocon.

- **Modèle en constellation :**

  Le principe de la modélisation en constellation est de joindre plusieurs modèles en étoile qui utilisent des dimensions communes. Un modèle en constellation comprend donc plusieurs faits et dimensions communes ou non.

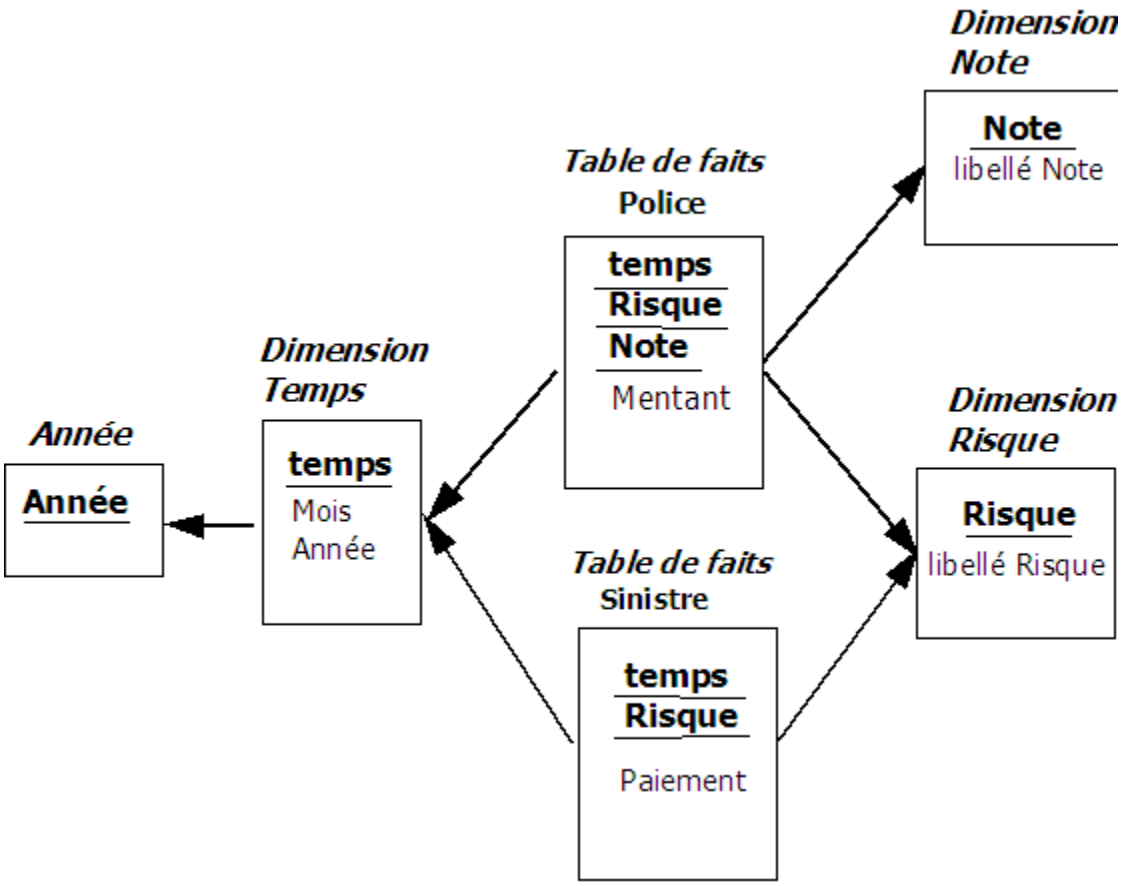

Figure 3.11 : Schéma en constellation.

✓ **Avantages des modèles en étoile et en flocon :**

- Le modèle en flocon offre une vue plus claire de la structure de l'information permettant notamment de déceler une hiérarchie.

- La normalisation de ce modèle permet de plus de diminuer la redondance, en réduisant la taille des tables de dimension. A noter que Kimball a évalué le gain de place disque à 1 % de l'espace disque total.

- Kimball préfère le modèle en étoile sur la base de deux arguments :

1 La dénormalisation permet d'améliorer les performances du système lors de l'exécution des requêtes.

2 Le modèle est plus facile à apprendre par l'utilisateur non informaticien.

3.9.3 Les modèles logiques de données :

Dans cette section, nous décrivons les modèles dimensionnels basés sur un modèle logique de données. Ainsi, nous retrouvons les modèles ROLAP basés sur un modèle relationnel, MOLAP basés sur un modèle dimensionnel, OOLAP basés sur un modèle objet, et HOLAP qui est un modèle hybride.

- **Le modèle de données ROLAP (Relational-OLAP):**
  Les données sont organisées selon des schémas relationnels spécialisés (constellation Figure 2.11, flocon Figure 2.10, ou étoile Figure 2.9). Un outil basé sur ROLAP possède un générateur SQL puissant, qui fournit un mécanisme pour décrire le modèle à travers les méta-données, et utilise les méta-données en temps réel pour construire des requêtes. D'autres fonctionnalités des systèmes relationnels ont été adaptées à cette organisation.

- **Le modèle de données MOLAP (Multidimensionnal-OLAP):**
  Plutôt que de stocker les informations comme des enregistrements, et les enregistrements dans des tables, les bases de données multidimensionnelles basées sur MOLAP stockent les données dans des tableaux, dont l'assemblage forme un cube.
  Les SGBDs qui supportent cette organisation sont capables de fournir des performances remarquables quant au traitement des requêtes par les opérations de pivot, drill down, etc… sans avoir recours à des jointures complexes, à des sous-requêtes, et à des unions.

- **Le modèle de données HOLAP (Hybrid-OLAP):**
  La description hybride vient du fait que ce modèle combine les caractéristiques des deux modèles précédent (ROLAP et MOLAP). Dans ce modèle, les données sont maintenues par un SGBD relationnel alors que les agrégations le sont dans un SGBD multidimensionnel.

- **Le modèle de données OOLAP (Object-OLAP):**
  L'OOLAP vise à combiner les avantages des deux approches ROLAP et MOLAP et à limiter leurs inconvénients. D'un côté, comme le ROLAP, l'OOLAP se base sur un standard, le paradigme objet. De l'autre côté, comme pour le MOLAP, les requêtes objet sont assez flexibles et extensibles pour réaliser facilement les opérations de manipulation OLAP. Buzydlowski propose un modèle dimensionnel comportant trois catégories d'objets : objets de données (les faits et les dimensions), de contrôle (les requêtes et les opérations OLAP) et d'interface (les outils permettant de visualiser les résultats des objets de contrôle). [22]

3.10 Avantages des entrepôts de données :

- Il constitue une collection de données centralisée disponible pour l'aide à la décision (OLAP, datamining,...).

- Les évolutions des données de l'entrepôt sont conservées (historisation des données).

- Il contient un ensemble de données consolidées (données homogènes et fiables).

- Il contient des données agrégées permettant une analyse à différents niveaux de détails.

- Il permet de développer différents thèmes d'analyse (réorganisation en fonction des sujets à analyser).

3.11 Conclusion :

Au terme de ce chapitre on peut conclure que la construction d'un entrepôt de données n'est pas une tache aisée. L'ED comme on vient de le voir est le cœur du système décisionnel. En effet, les décisions sont prises sur la base des données qu'il contient. Il est donc plus que vital que les données du DW soient convenablement préparées pour se prêter aux analyses et donc aider au mieux à la prise de décision.

Ce chapitre nous a permis de présenter plusieurs concepts autour des entrepôts de données, ce qui nous permettra par la suite de modéliser notre entrepôt, en utilisant ces concepts.

Chapitre 04

Conception

Chapitre 04

Conception

Sommaire

4.1 Introduction :

La conception doit passer par la phase d'identifications et analyse des besoins, car ce dernier doit répondre aux attentes des décideurs et des analystes. Dans le cadre de notre travail, on s'intéressera à la Caisse Nationale d'Assurance Sociales des travailleurs salariés et plus particulièrement à son système. Nous commençons ce chapitre par l'identification des besoins, puis identifier les différents utilisateurs du système.

4.2 Définitions des besoins :

Au cours du premier chapitre, nous avons présenté le système existant au niveau de la direction informatique de la CNAS, ainsi qu'une étude qui a porté sur les différents départements et services de la direction informatique concernés par notre projet, et qui nous a permis de définir les différents utilisateurs et les besoins de chacun. Pour mener à bien cette étape importante, nous avons ressenti le besoin d'interviewer certains employés qui occupent les postes des décideurs dans la société, en leur posant des questions qui pouvaient nous éclaircir les besoins et objectifs des décideurs :

Questions :

- Quels sont les objectifs de votre direction ?
- Quels sont les grands domaines de votre direction ?
- Quels sont les axes à prendre en compte lors d'une analyse ?
- Comment souhaitez-vous apercevoir vos résultats d'analyse ?

Cette étape nous permis d'identifier les besoins des décideurs, et de fixer les objectifs qui vont permettre de répondre à leurs attentes. Pour cela, nous avons établi un diagramme des cas d'utilisation UML du système. Un cas d'utilisation (use case) modélise les interactions entre le système à développer et un utilisateur ou acteur interagissant avec le système. Plus précisément, un cas d'utilisation décrit une séquence d'actions réalisées par le système qui produit un résultat observable pour un acteur. Ce dernier représente toutes les personnes qui vont utiliser ce système, dans notre cas les décideurs.

Dans cette étape, nous avons définis les différents domaines sur lesquels nous allons concevoir l'entrepôt, et qui sont les plus sensibles du coté décisionnel au niveau de la direction informatique. Ces domaines sont : le suivi des allocations familiales (AF), le suivi des rentes.

La figure 4.1 représente le diagramme des cas d'utilisations de notre système, et les différents acteurs.

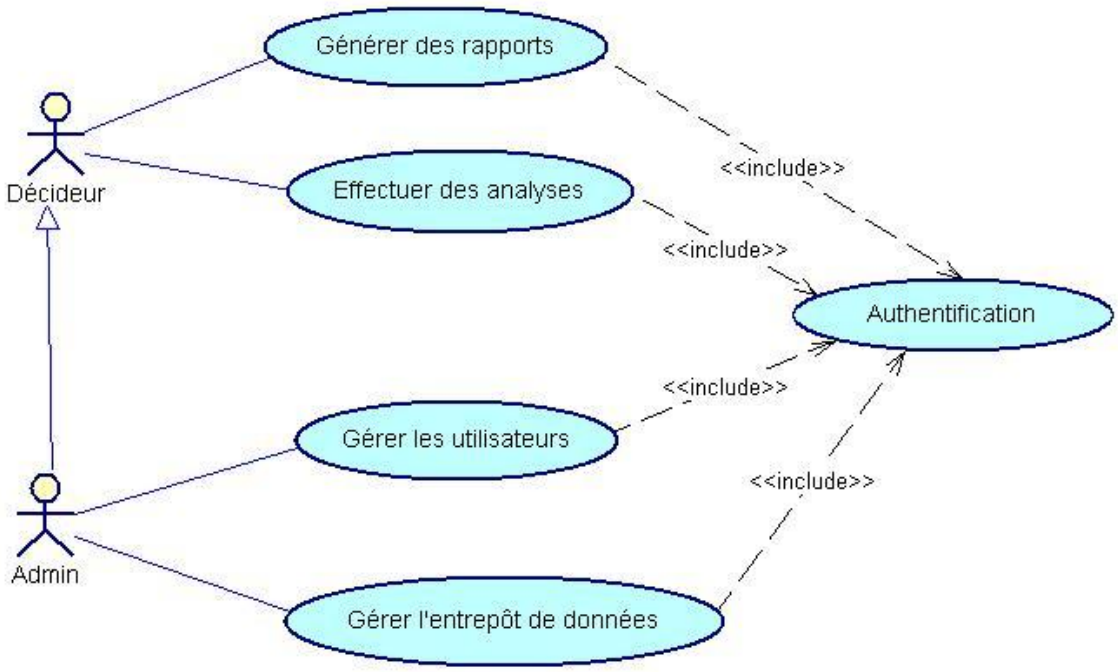

Figure 4.1 : Diagrammes des cas d'utilisations.

Dans notre système, nous avons deux types d'utilisateurs :

L'administrateur : acteur qui gère et attribut les droits d'accès aux décideurs, de plus, il a le rôle de contrôler le bon fonctionnement de l'entrepôt (alimentation, mise a jour des données, utilisateurs, …).

Le décideur : toute personne ayant l'autorisation d'effectuer des opérations sur l'entrepôt de données.

Décrire les cas d'utilisation peut se faire de plusieurs manières, dont la description par tableau. Par la suite, un tableau qui décrit les différents cas d'utilisation de notre système.

Cas d'utilisation	Description
Authentification	Avant toute utilisation du système, chaque utilisateur doit s'identifier via un nom d'utilisateur et un mot de passe.
Effectuer des analyses	Le décideur lance des analyses sur les données du système pour en tirer des informations d'aide à la décision.
Générer des rapports	Le décideur peut exploiter ces résultats d'analyse sous forme de rapports en différents formats (PDF, HTML, XLS, …)
Gérer les utilisateurs	L'administrateur a le droit d'ajouter ou de retirer les utilisateurs au système.
Gérer l'entrepôt de données	L'administrateur s'occupe de contrôler les processus d'alimentation et de mises a jour des données.

Tableau 4.1 : Tableau de description des cas d'utilisation.

4.3 Processus de la modélisation dimensionnelle :

La modélisation dimensionnelle des données permet de fournir aux décideurs un moyen d'exploiter au mieux les informations en temps réels et les aide à la prise de décision. Dans la modélisation dimensionnelle on distingue les faits et les dimensions. Les faits sont ce sur quoi va porter l'analyse, ce sont des tables qui contiennent des informations opérationnelles, les faits que nous traiterons sont de types numériques. Par contre les dimensions sont les axes avec lesquels on veut faire l'analyse.

4.3.1 Domaine suivi des allocations familiales (AF):

Les dimensions ont pour objectif de décrire le fait, donc on essaye de recenser toutes les informations qui décrivent une allocation familiale et qui peuvent intéresser les décideurs. Les principales dimensions identifiées dans la base de données sont :

- **Dimension Temps :**

La dimension temps est selon Ralph Kimball la seule dimension qui figure systématiquement dans tout entrepôt de données, car en pratique tout entrepôt de données est une série temporelle. Le temps est le plus souvent la première dimension dans le classement sous jacent de la base de données.

La dimension temps se présente comme suit :

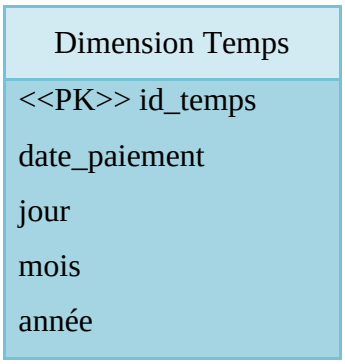

Figure 4.2 : La Dimension Temps du fait suivi AF.

Le niveau de détail le plus bas de cette dimension est la journée. En effet, les utilisateurs ont fait ressortir le besoin de suivre les chiffres au jour le jour et d'en garder l'historique de ces derniers.

- **Dimension Zone Géographique :**

La dimension zone géographique décrit la zone où le fait a eu lieu. Après l'étude des besoins, il parait intéressant de faire des comparaisons par rapport à des zones géographiques.
La dimension Zone se présente comme suit :

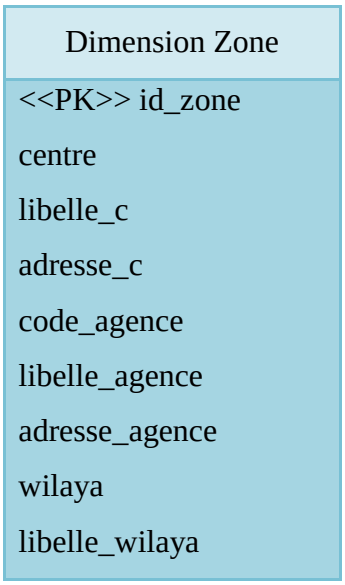

Figure 4.3 : La Dimension Zone du fait suivi AF.

- **Dimension Assuré :**

La dimension Assuré décrit toutes les informations nécessaires pour le fait AF.

La dimension Assure se présente comme suit :

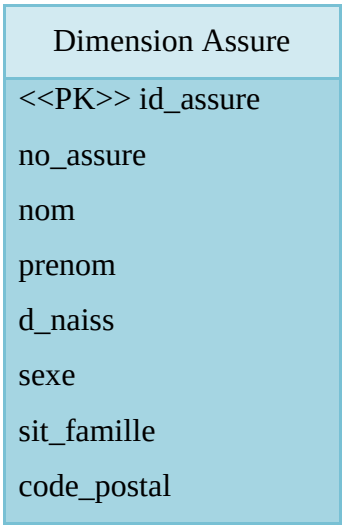

Figure 4.4 : La Dimension Assuré du fait suivi AF.

- **Dimension Paiement :**

La dimension Paiement se présente comme suit :

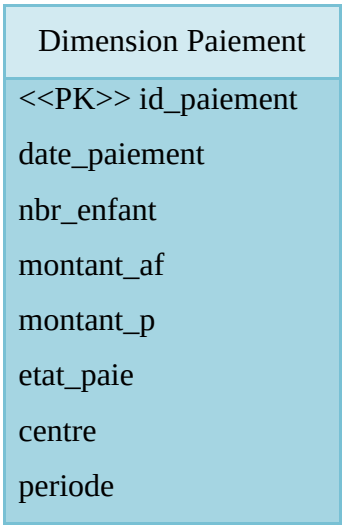

Figure 4.5 : La Dimension Paiement du fait suivi AF.

- **Dimension Allocataire :**

La dimension Allocataire se présente comme suit :

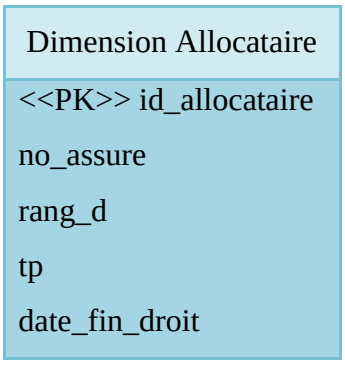

Figure 4.6 : La Dimension Allocataire du fait suivi AF.

- **Les mesurables :**

Les mesurables qui correspondent au domaine suivi des allocations familiales sont le « montant des allocations » et le « montant trop perçu » et le « nombre d'enfant ».
Le fait du domaine suivi des allocations familiales se présente comme suit:

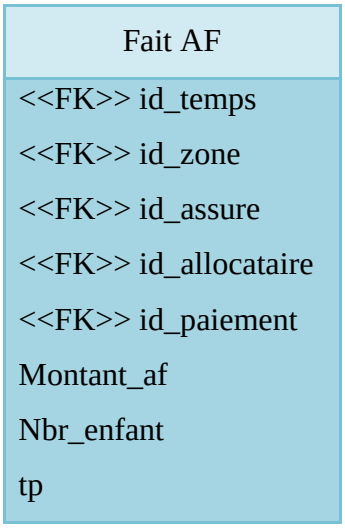

Figure 4.7 : Le fait suivi AF.

Le modèle en étoile du domaine suivi des allocations familiales se présente comme suit:

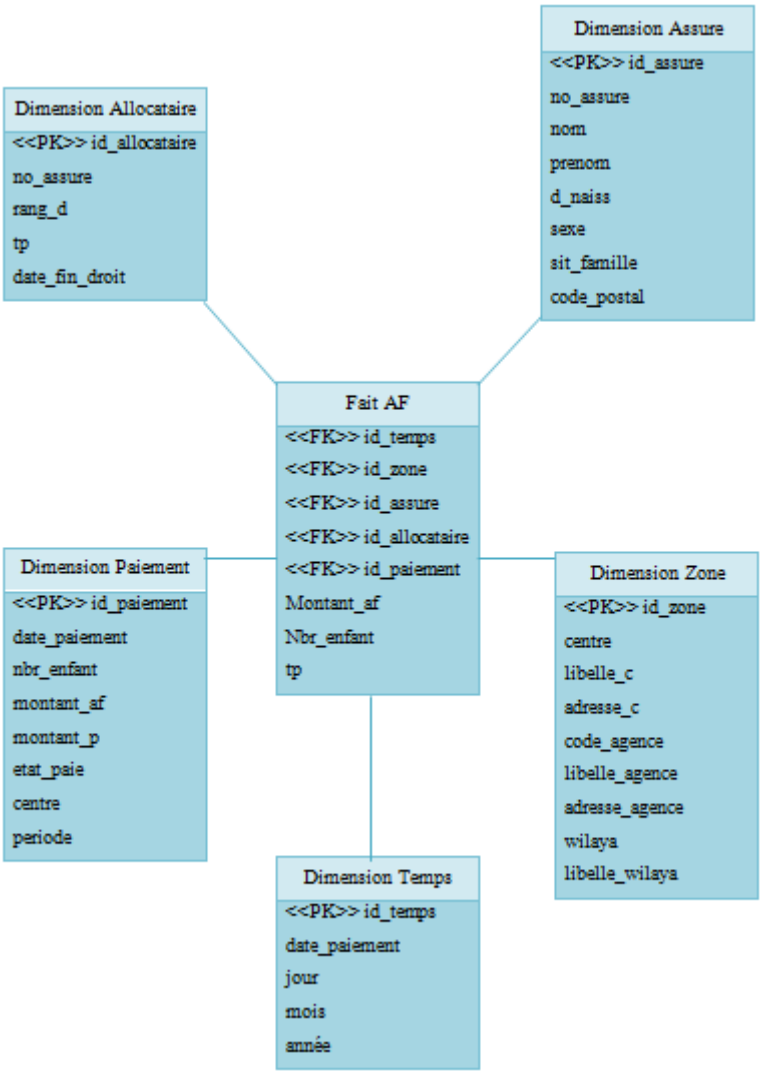

Figure 4.8 : Le modèle en étoile du Domaine suivi AF.

4.3.2 Domaine suivi des Rentes :

Après la détection des dimensions de la nouvelle étoile, on procède à une mise en conformité des dimensions communes. Pour ce faire, on construit un tableau qui croise les étoiles conçues avec leurs dimensions. Le but étant de détecter les dimensions communes pour leurs mises en conformité. Le tableau suivant illustre cela :

Etoile	Allocations familiales	Rentes
Dimension temps	✓	✓
Dimension zone	✓	✓
Dimension assure	✓	✓
Dimension allocataire	✓	
Dimension paiement	✓	
Dimension paiement rente		✓

Tableau 4.2 : Détection des dimensions communes.

A cette étape il existe trois dimensions communes. Ces dimensions étant très détaillées dans la première étoile, il n'y a pas eu nécessité de recourir à une mise en conformité.

- **Dimension paiement rente :**

La dimension paiement rente se présente comme suit :

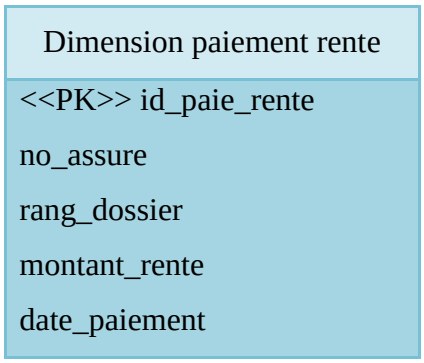

Figure 4.9 : La Dimension paiement rente du fait suivi rentes.

- **Les mesurables :**

La mesure qui correspond au domaine suivi des Rentes est le « montant des rentes ».
Le fait du domaine suivi des rentes se présente comme suit:

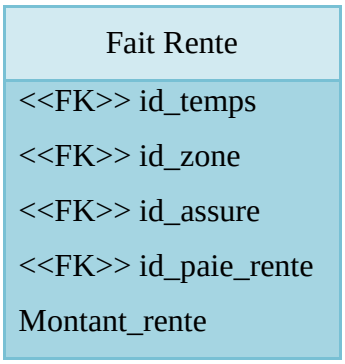

Figure 4.10 : Le fait suivi des Rentes.

Le modèle en étoile du domaine suivi des Rentes se présente comme suit:

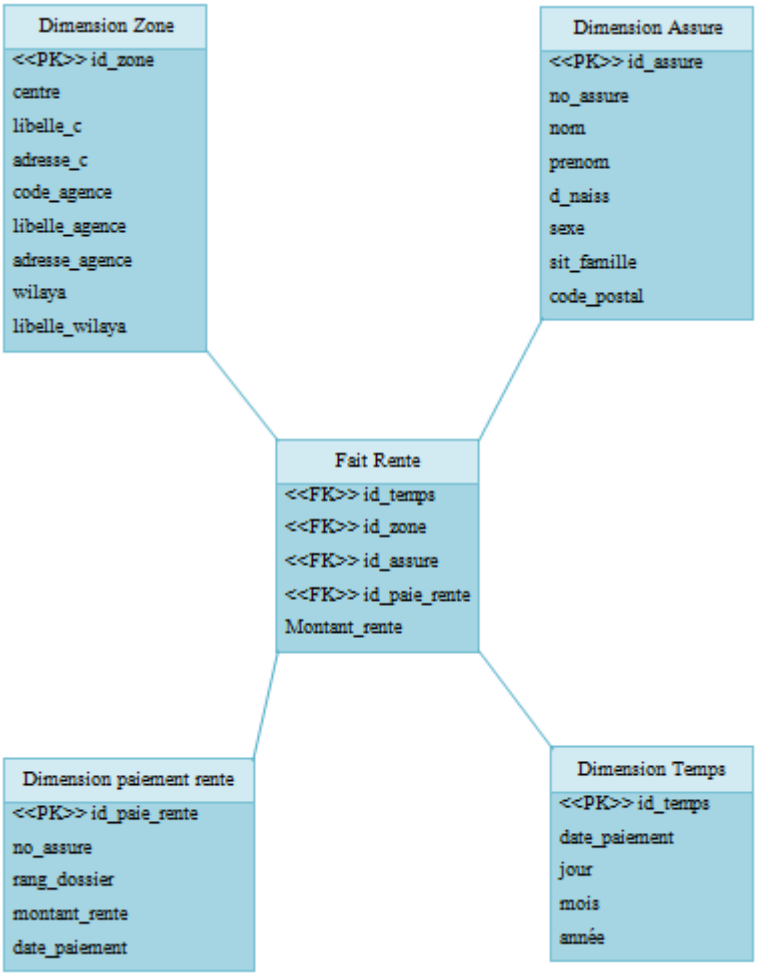

Figure 4.11 : Le modèle en étoile du Domaine suivi Rentes.

4.4 Conception de la zone d'alimentations et de préparation :

Dans cette section, on va définir le travail effectué pour l'élaboration de chacune des trois étapes de l'outil ETL.

4.4.1 Extraction :

Un plan de préparation des données est indispensable pour pouvoir se lancer dans l'extraction de données à partir des tables sources.
L'identification des tables de la base de données source permettra de ciblé les tables des quelles on extraira des données. Nous allons éclairer cette étape par un tableau qui va mettre en évidence chaque dimension et sa/ses table(s) correspondante(s) de la source de données.

Tableaux de correspondance des tables des dimensions :

Tables cibles	Attributs cibles	Attributs sources	Tables sources
Dimension Assuré	ID_ASSURE		Assurés
	NO_ASSURE	NO_ASSURE	
	NOM	NOM	
	PRENOM	PRENOM	
	D_NAISS	D_NAISS	
	SEXE	SEXE	
	SIT_FAMILLE	SIT_FAMILLE	
	CODE_POSTAL	CODE_POSTAL	
	CENTRE	CENTRE	
Dimension Allocataire	ID_ALLOCATAIRE		Allocataire
	NO_ASSURE	NO_ASSURE	
	RANG_D	RANG_D	
	DATE_FIN_DROITAF	DATE_FIN_DROITAF	
	TP	TP	

Dimension Paiement	ID_PAIEMENT		Paiement
	NO_ASSURE	NO_ASSURE	
	RANG_D	RANG_D	
	DATE_PAIEMENT	DATE_PAIEMENT	
	MONTANT_AF	MONTANT_AF	
	MONTANT_RP	MONTANT_RP	
	MONTANT_P	MONTANT_P	
	PERIODE	PERIODE	
	CENTRE	CENTRE	
	ETAT_PAIE	ETAT_PAIE	
Dimension Paiement Rente	ID_PAIE_RENTE		Paiement Rente
	NO_ASSURE	NO_ASSURE	
	RANG_DOSSIER	RANG_DOSSIER	
	MONTANT_RENTE	MONTANT_RENTE	
	DATE_PAIEMENT	DATE_PAIE_RENTE	

Tableau 4.3 : Table de correspondance des tables des dimensions.

Un autre tableau viendra appuyer la table de correspondance des tables des dimensions, et qui a pour rôle de mettre en évidence chaque tables de fait de nos différents magasins de données, ainsi que ses tables correspondantes de la source de données.

Tables cibles	Attributs cibles	Attributs sources	Tables sources
Dimension Zone Géographique	ID_ZONE		Centre
	CENTRE	CENTRE	
	LIBELLE_C	LIBELLE	
	ADRESSE_C	ADRESSE	
	CODE_AGENCE	CODE_AGENCE	Agence
	LIBELLE_AGENCE	LIBELLE_AGENCE	
	ADRESSE_AGENCE	ADRESSE_AGENCE	
	WILAYA	WILAYA	Wilaya
	LIBELLE_W	LIBELLE	
Dimension Temps	ID_TEMPS		Paiement
	DATE_PAIEMENT	DATE_PAIEMENT	
	JOUR		
	MOIS		
	ANNEE		
	DATE_PAIEMENT_R	DATE_PAIE_RENTE	Paiement Rente

Tableau 4.4 : Table de correspondance des tables des dimensions (suite).

Tableaux de correspondance des tables des faits :

Tables cibles	Attributs cibles	Attributs sources	Tables sources
Fait suivi AF	MONTANT_AF	MONTANT_AF	Paiement
	NBR_ENFANT	NBR_ENFANT	
	TP	TP	
Fait suivi Rentes	MONTANT_RENTE	MONTANT_RENTE	Paiement Rente

Tableau 4.5 : Table de correspondance des tables de faits.

4.4.2 Transformation :

Les données extraites doivent atterrir sur une autre base (l'utilisation d'un outil ETL rend invisible cette base) appelée base tampon (Staging area). Une fois l'étape d'extraction terminée, les transformations nécessaires peuvent êtres effectuées tranquillement dans la base tampon.

Pour chaque table de la base décisionnelle on doit :

- **Substitution des clés primaires :** Une fois que les tables tampons sont remplies, on s'occupe de l'intégrité des données qui vont être chargées dans la base décisionnelle on doit garder trace des clé source pour un chargement ultérieur. Exemple dans la table Dimension Zone Géographique on doit inclure les clés primaires de table source Centre, Agence et Wilaya (CENTRE, CODE_AGENCE, WILAYA).
- **Substitution des clés étrangères :** Pour cela, il faut recalculer les clés étrangères avec les clés de substitution afin que les relations de la base décisionnelle soient vérifiées lors du chargement.

4.4.3 Chargement :

Le chargement est une étape assez complexe, vu qu'il faudra planifier l'ordre des extractions des données depuis les tables sources pour leurs chargements vers les dimensions et il faut aussi planifier les éventuelles jointures pour produire une information.
Comme les données sont chargées dans la base décisionnelle qui est muni d'un schéma relationnel, il faut charger les tables dans cet ordre :

- D'abord les tables qui ne contiennent aucune clé étrangère.
- Ensuite les tables qui ne contiennent que des clés étrangères vers des tables déjà chargées.
- Enfin, on termine par le chargement des tables de fait (suivi des allocations familiales (AF), suivi des Rentes).

4.4.4 Construction du CUBE OLAP :

Dans cette étape, nous allons concevoir notre couche multidimensionnelle (cubes OLAP) pour faciliter l'analyse. Nous allons transformer les données stockées dans une base de données relationnelle en une base de données multidimensionnelle afin de rendre l'analyse pertinente et facile. Avant la création des cubes on va définir les niveaux et les hiérarchies.

Dimension	Niveau	Attributs	Hiérarchie
Dimension Zone Géographique	N1	CENTRE	
		LIBELLE	
		ADRESSE	
	N2	CODE_AGENCE	Hiérarchie : N1>N2>N3
		LIBELLE_AGENCE	
		ADRESSE_AGENCE	
	N3	WILAYA	
		LIBELLE	

Tableau 4.6 : Table d'identification des niveaux et des hiérarchies.

4.5 Conclusion :

Au cours de ce chapitre, nous avons tenu à présenter l'étude conceptuelle de l'entrepôt de données qui est constitué de deux magasins de données. Ce qui nous a permis d'identifier les axes d'analyses et les différentes tables de fait sous un schéma en étoile.

Dans le chapitre suivant nous allons proposer une implémentation de notre solution en décrivant les outils utilisés pour réaliser l'application.

Chapitre 05

Réalisation

Chapitre 05

Réalisation

Sommaire

5.1 Introduction :

Dans ce présent chapitre nous passerons à la phase de déploiement et de réalisation pour accomplir les étapes précédemment évoquées dans la partie conception. Nous commencerons d'abord par présenter l'architecture technique, puis une présentation des outils utilisés sur cette architecture, en l'occurrence le SGBD PostgreSQL et le logiciel Pentaho. En suite, nous présentons la configuration du système, pour montrer le rôle de chaque outil dans le déploiement de notre solution. Enfin, une présentation de l'interface utilisateur, et l'aspect sécurité de notre solution.

5.2 Architecture du système :

Notre système décisionnel est destiné aux décideurs de la direction informatique de la Caisse Nationale d'Assurance Sociales des travailleurs salariés, pour avoir une meilleure vue sur les données de la CNAS. L'architecture de notre solution est une architecture en trois tiers,
Architecture trois tiers est une architecture d'application dans laquelle on sépare la présentation, les traitements, et les bases de données. L'objectif ciblé est de permettre l'évolution de l'un de ces trois tiers de façon relativement indépendante des deux autres. L'implémentation physique de ces architectures est souvent soumise à plusieurs contraintes, car elles sont parfois mise en œuvre à travers des plateformes différentes (hétérogène).

Les outils choisis sont :

1. Le SGDB PostgreSQL pour l'intégration de la base de données de notre entrepôt de données.
2. Pentaho data intégration (PDI) qui est un outil ETL intégré dans la solution pentaho.
3. Une installation pré-configurée de pentaho Tomcat sur le serveur d'application.
4. Construction du cube: Schéma Work bench.
5. Analyse multidimensionnelle : Pentaho Analysis Mondrian.
6. Restitution : Pentaho Analysis et Pentaho Report Design pour les rapports.

5.3 Présentation :

Dans cette partie nous présenterons les différents outils qui sont utilisés pour le déploiement de l'entrepôt de données.

5.3.1 Présentation du SGBD PostgreSQL :

PostgreSQL (prononcé postgrécecuelle ou postgrece) est un SGBDR (systèmes de gestion de base de donnée relationnelles) fondé sur postgres. Il a été développé à l'université de Californie au département des sciences informatiques de **Berkeley**. [23]

PostgreSQL est un descendant OpenSource du code original de Berkeley. Il supporte une grande partie du standard SQL tout en offrant de nombreuses fonctionnalités modernes :

- Requêtes complexes ;
- Clés étrangères ;
- Triggers ;
- Vues ;
- Intégrité transactionnelle ;

De plus, PostgreSQL peut être étendu par l'utilisateur de multiples façons. En ajoutant, par exemple :

- Nouveaux types de données ;
- Nouvelles fonctions ;
- Nouveaux opérateurs ;
- Nouvelles fonctions d'agrégat ;
- Nouvelles méthodes d'indexage ;
- Nouveaux langages de procédure.

1. **Historique:**

Le système de gestion de bases de données relationnel objet PostgreSQL est issu de POSTGRES, programme écrit à l'université de Californie à **Berkeley**. Après plus d'une vingtaine d'années de développement, PostgreSQL annonce être devenu la base de données Open Source de référence.

Le projet POSTGRES mené par le professeur **Michael Stonebraker** était sponsorisé par le DARPA (Defense Advanced Research Projects Agency), la NSF (Nationnal Science

Foundation) et ESL, Inc. Le développement de POTGRES à débuté en 1986. Depuis plusieurs versions majeures ont vu le jour (Démo, la version1, la version2…). [23]
POSTGRES a aussi été utilisé comme support de formation dans plusieurs universités.

2. Les différentes plates formes supportées par PostgreSQL :

PostgreSQL fonctionne sur de nombreuses plates formes :

- Unix ;
- Linux ;
- FreeBSD ;
- Aix (IBM) ;
- HP-UX (Hewlet Packard) ;
- IRIX ;
- Solaris (Sun)…

3. Caractéristique de PostgreSQL :

PostgreSQL possède de nombreuses caractéristiques en faisant un SGBDR robuste et puissant digne des SGBD commerciaux: [23]

- Des interfaces graphiques (x-window est donc nécessaire) pour gérer les tables.
- Des bibliothèques pour de nombreux langages (appelés *frontaux*) afin d'accéder aux enregistrements à partir de programmes écrits en : java, langage C/C++, Perl, Tcl/TK.
- PostgreSQL fonctionne selon une architecture client/serveur, il est ainsi constitué :
    - D'une partie serveur, capable de traiter les requêtes des clients, il s'agit dans le cas de PostgreSQL d'un programme résident en mémoire appelé ***postmaster.***
    - D'une partie cliente qui interroge le serveur de base de données à l'aide de requêtes SQL.

4. Installation de PostgreSQL :

La procédure d'installation de PostgreSQL est la suivante :

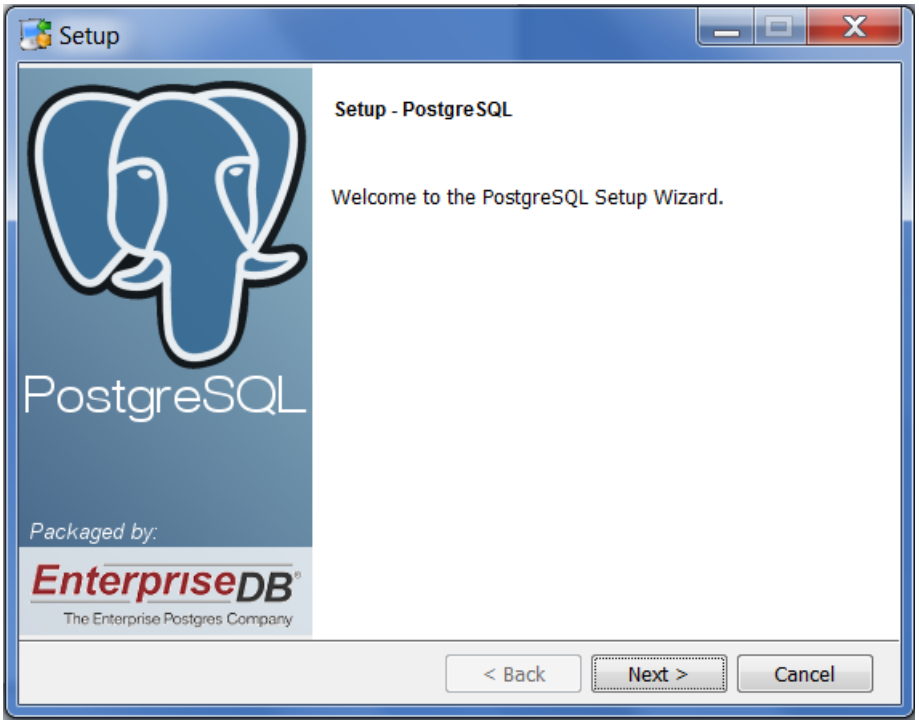

Figure 5.1 : confirmer l'installation.

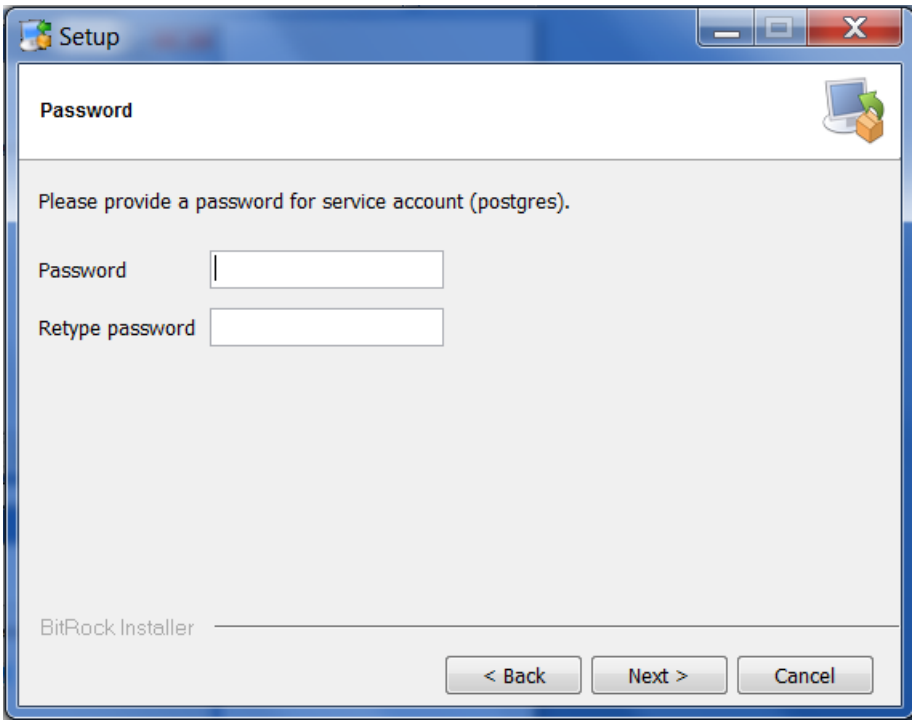

Figure 5.2: Création du mot de passe.

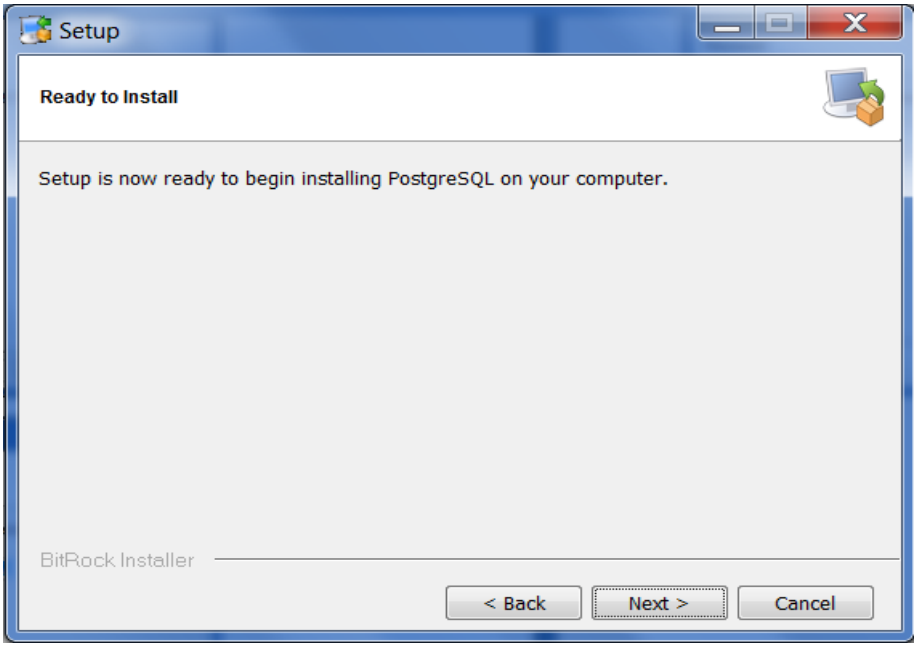

Figure 5.3: Fin de l'installation.

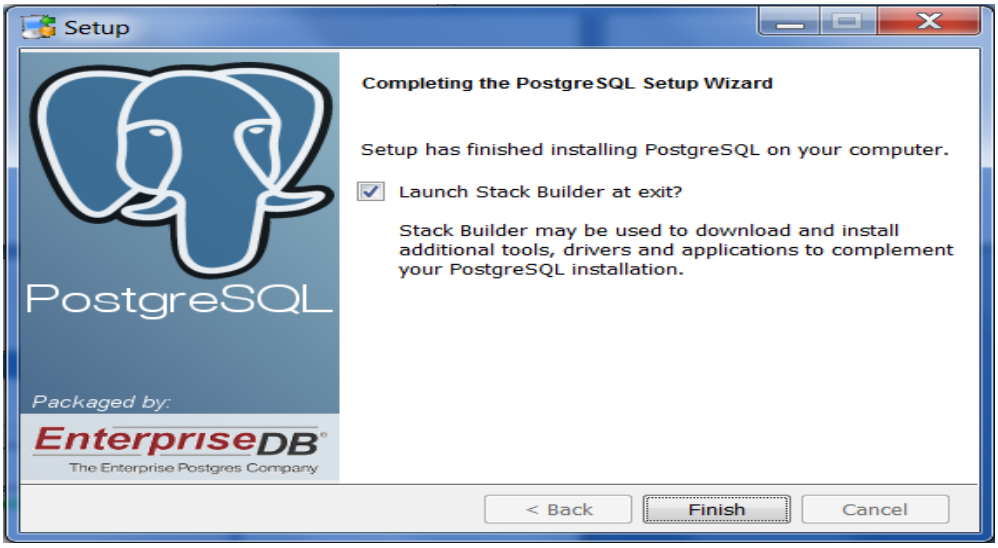

Figure 5.4: Lancement du postgreSql.

✓ L'installation commence par le choix de la langue comme le montre la figure suivante :

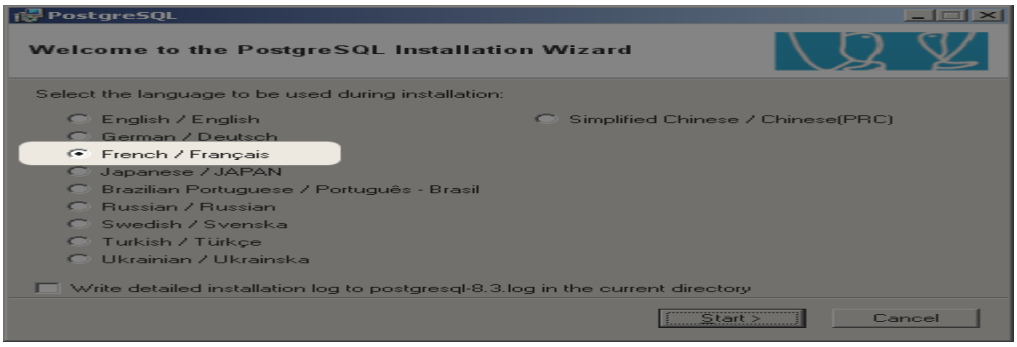

Figure 5.5: choix de langue.

✓ Par la suite il vous demanderez de saisir un mot de passe pour l'utilisateur

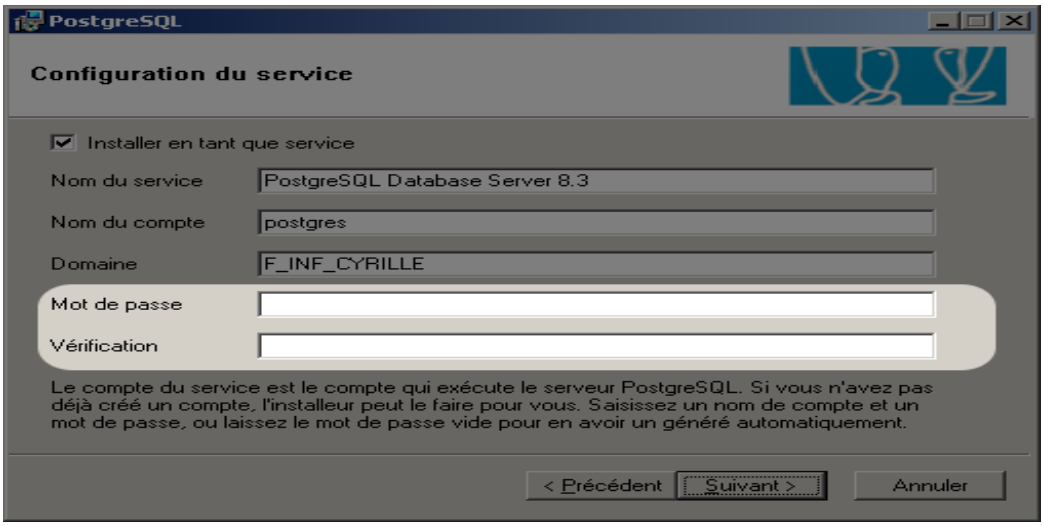

Figure 5.6: interface de saisie de mot de passe utilisateur.

✓ Et un autre pour l'administrateur

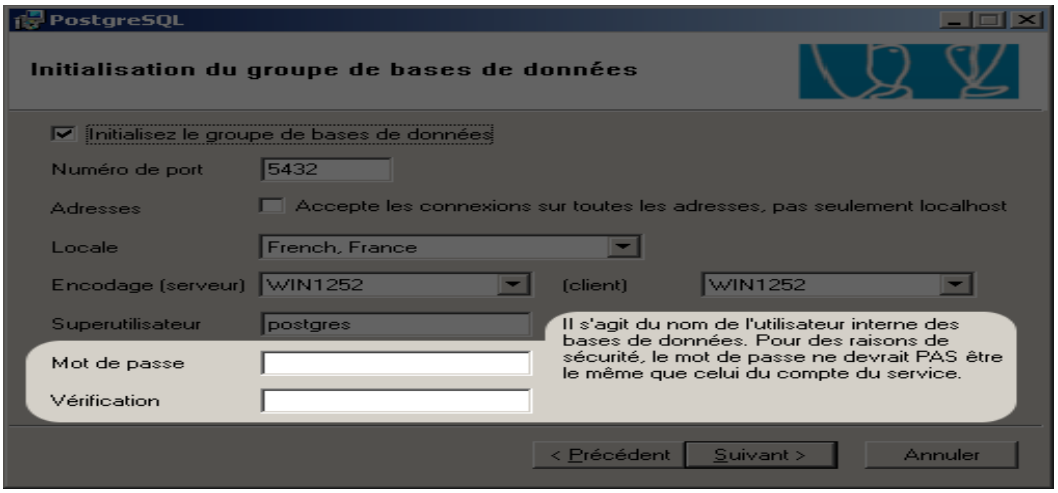

Figure 5.7: interface de saisie de mot de passe administrateur.

✓ A la fin de l'installation il vous demande si vous voulez exécuter le Stack Builder (il permet d'installer d'autre logiciel en rapport avec postgreSQL).

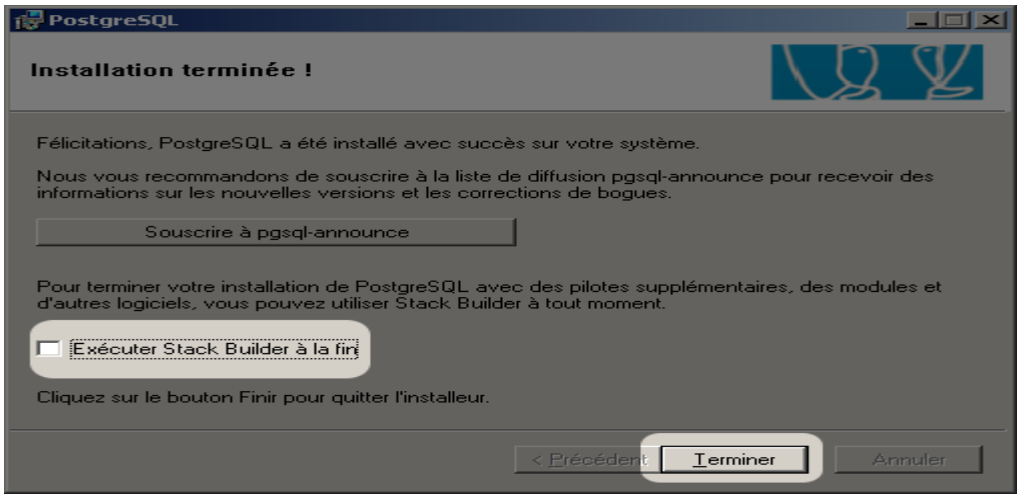

Figure 5.8: fin d'installation.

✓ Regardez dans le menu "**Démarrer**" et allez dans "**Tous les programmes**", vous devriez avoir ceci dans l'encart "**PostgreSQL 8.3**" :

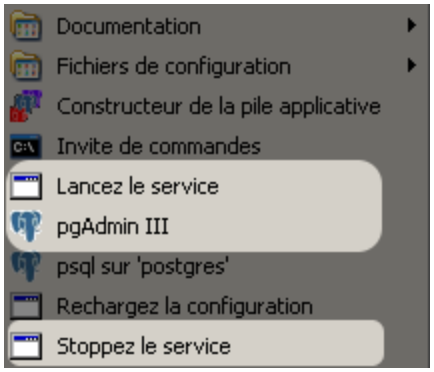

Figure 5.9: Après la fin de l'installation.

I.5.1 : Configuration du serveur :

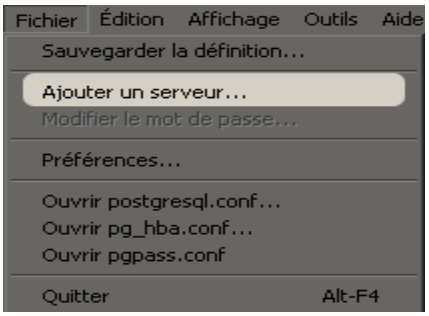

Figure 5.10: Exemple d'ajout d'un serveur.

✓ Ce qui vous amène à ceci :

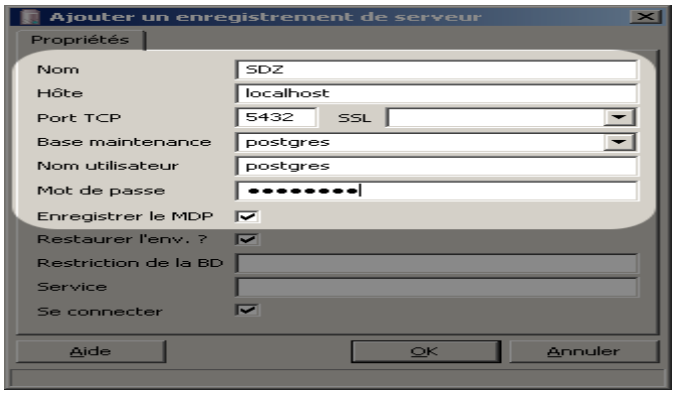

Figure 5.11: Exemple d'enregistrement du serveur.

✓ **Nom** : correspond au nom de la base de données ;

✓ **Hôte :** correspond à l'adresse du serveur sur le réseau.

✓ ensuite, le nom de l'utilisateur et son mot de passe déjà défini.

✓ Après la configuration vous devez avoir cette interface :

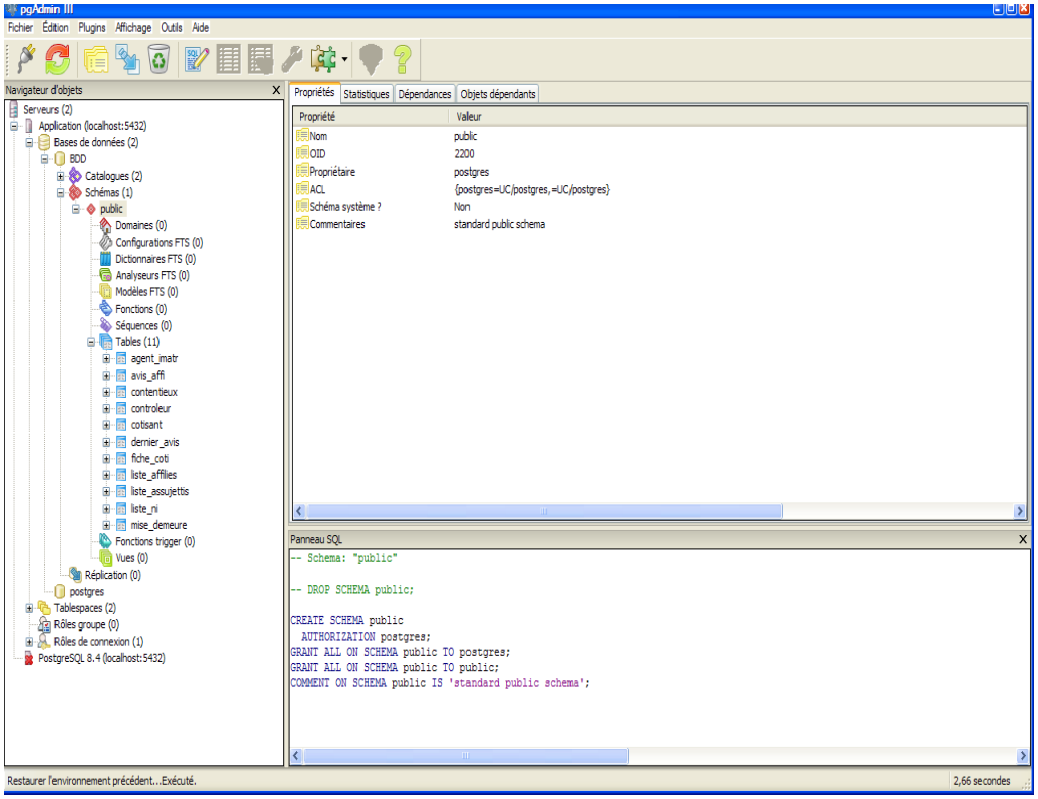

Figure 5.12: Page d'accueil de pgAdminIII

5.3.2 Présentation de Pentaho : [24]

1 Présentation :

Pentaho est une plate-forme décisionnelle open source complète possédant les caractéristiques suivantes :

Une couverture globale des fonctionnalités de la Business Intelligence :

- ✓ ETL (intégration de données)
- ✓ Reporting
- ✓ Tableaux de bords (*"Dashboards"*)
- ✓ Analyse ad-hoc (requêtes à la demande)
- ✓ Analyse multidimensionnelle (OLAP)

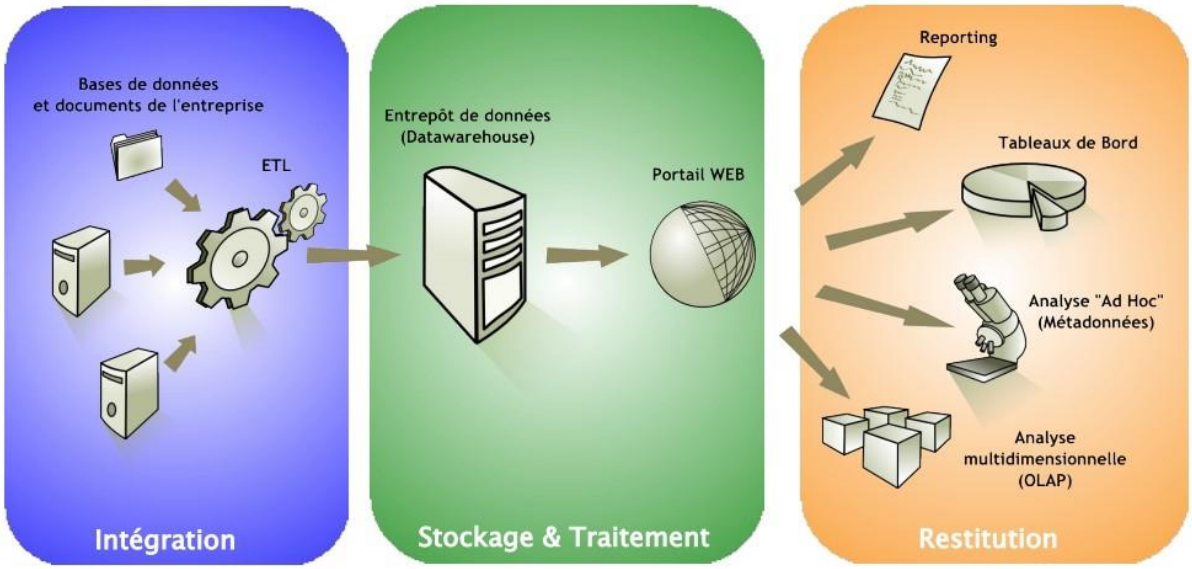

Figure 5.13 : Les fonctionnalités de Pentaho.

Pentaho permet d'adresser 2 typologies d'utilisateurs:

- Les « one-clic users », utilisateurs de base, consommateurs d'indicateurs prédéfinis
- Les utilisateurs avancés, qui ont besoin d'outils d'analyse et d'exploration avancés

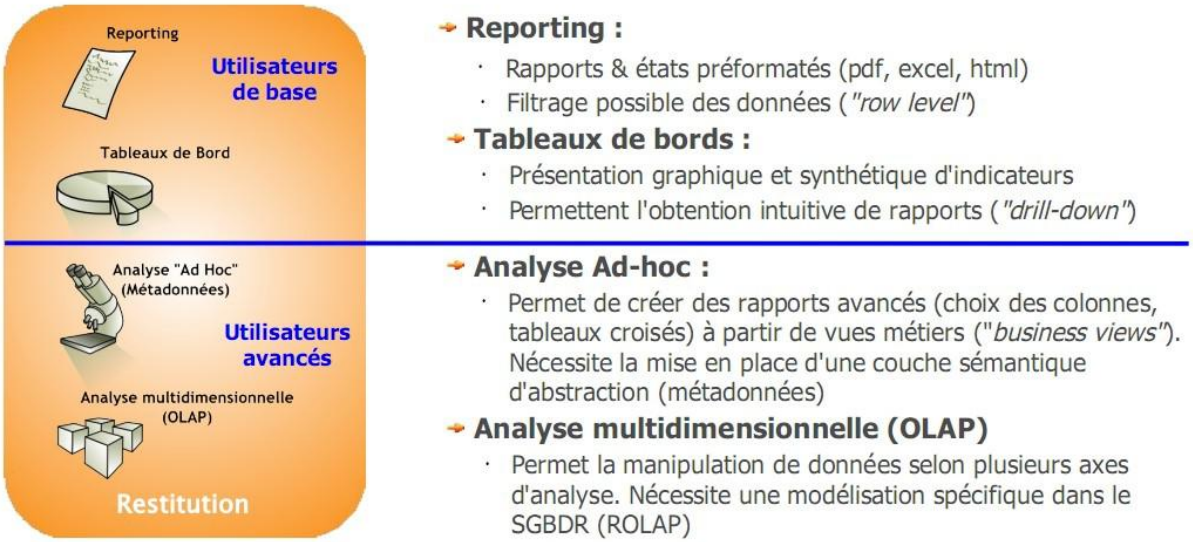

Figure 5.14 : Les utilisateurs de Pentaho.

Une architecture web 2.0 qui se compose :

- D'un serveur web J2EE permettant de mettre à disposition l'ensemble des ressources décisionnelles et ceci au travers d'urls web uniques et standardisées.
  Le serveur est dénommé "Pentaho User Console" (PUC)
- Plusieurs clients riches permettant la conception et la publication des ressources.
  Ces derniers sont librement téléchargeables et peuvent être installés sous des environnements Windows, Linux ou Mac-OS (clients Java):

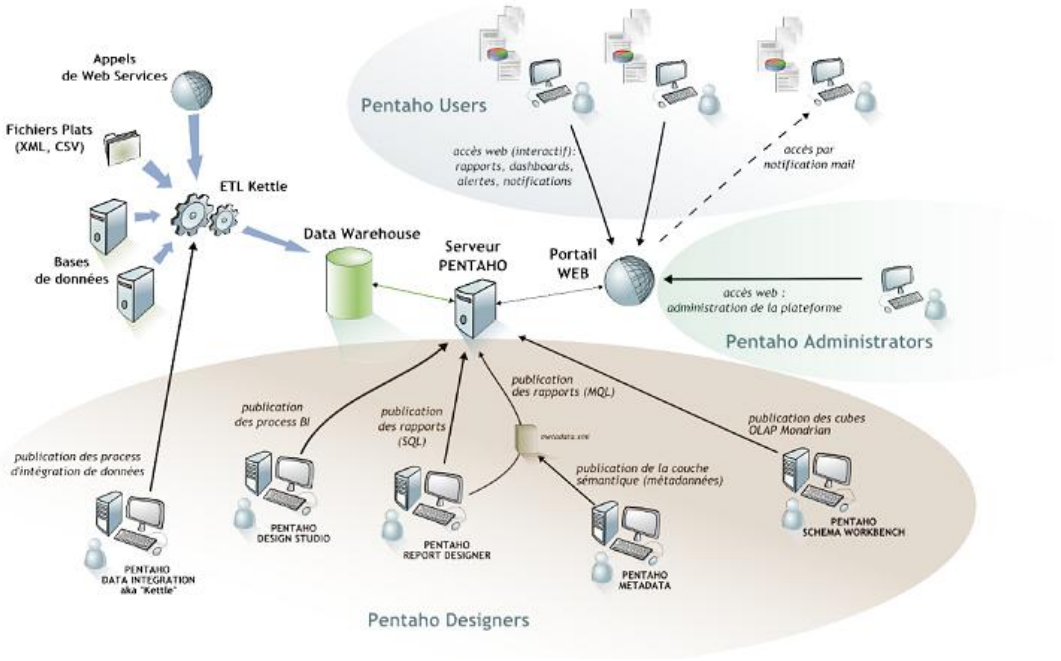

Figure 5.15 : Architecture de Pentaho.

- le serveur Web Pentaho comporte également une plate-forme d'administration (Pentaho Administration Console) pour la gestion des droits d'accès, la planification d'évènements, la gestion centralisée des sources de données... ;
- Pentaho est reconnue pour être une solution d'une grande qualité conceptuelle et technique. La plate-forme est orientée « processus » : au travers de « séquences d'actions » on peut ainsi modéliser avec Pentaho des workflows BI avancés ;
- il n'est pas besoin de connaître JAVA pour travailler avec Pentaho : seule la maîtrise du langage SQL est nécessaire, ainsi que des connaissances de base en XML, HTML et JavaScript. Il faut bien sûr s'autoformer (ou être formé) aux clients de conception ;
- une communauté importante et très active s'anime autour de Pentaho. Celle-ci contribue au codage de nombreux plugins et de projets communautaires : plugins Kettle, Pentaho Analysis Tool, Pentaho Community Dashboard Framework, etc. ;
- Pentaho est une suite décisionnelle open source commerciale qui reste très « ouverte ». Les différences fonctionnelles entre la version libre (community edition) et la version payante (enterprise edition) restent limitées. La version libre de Pentaho permet d'installer une plate-forme décisionnelle complète.

2 Téléchargement : [25]

Pour débuter avec Pentaho, il est conseillé de télécharger la version community, gratuite et libre d'utilisation. Cette version communautaire peut-être téléchargée sur SourceForge:
http://sourceforge.net/projects/pentaho/files

3 Installation de Pentaho (en local) :

Le serveur Pentaho (biserver-ce) est un serveur de démonstration prêt à l'emploi, complètement autonome et pouvant être installé sur un PC bureautique disposant au moins de 1 Go de RAM.
Une fois l'archive téléchargée, il suffit de décompresser celle-ci dans un répertoire préalablement créé, par exemple « C:\Pentaho-3.5.2 » (Windows).
Le répertoire d'installation sera désigné {PENTAHO-HOME} dans la suite de ce document.
Deux répertoires sont créés dans {PENTAHO-HOME} :

- **\biserver-ce :** la console Web d'utilisation (Pentaho User Console) ;
- **\administration-console :** la console Web pour l'administration de la plate-forme (Pentaho Administration Console).

4 Démarrer et arrêter les serveurs Pentaho :

- **Pentaho User Console :**

Les commandes suivantes permettent de lancer et stopper la console d'utilisation Web :

Action	Commande
Démarrage (Windows)	{PENTAHO-HOME}\biserver-ce\start-pentaho.bat
Arrêt (Windows)	{PENTAHO-HOME}\biserver-ce\stop-pentaho.bat
Démarrage (Linux)	{PENTAHO-HOME}\biserver-ce\start-pentaho.sh
Arrêt (Linux)	{PENTAHO-HOME}\biserver-ce\stop-pentaho.sh

Tableau 5.1 : Démarrage et arrêt de Pentaho User Console.

On accède à la console d'utilisation Pentaho en saisissant l'URL suivante dans un navigateur Web : http:\\localhost:8080\pentaho
Puis en saisissant l'identifiant et mot de passe ci-dessous :
login : joe ;
password : password.

- **Pentaho Administration Console :**

Les commandes suivantes permettent de lancer et stopper la console d'utilisation Web :

Action	Commande
Démarrage (Windows)	{PENTAHO-HOME}\administration-console\start-pac.bat
Arrêt (Windows)	{PENTAHO-HOME}\administration-console\stop-pac.bat
Démarrage (Linux)	{PENTAHO-HOME}\administration-console\start-pac.sh
Arrêt (Linux)	{PENTAHO-HOME}\administration-console\stop-pac.sh

Tableau 5.2 : Démarrage et arrêt de Pentaho Administration Console.

On accède à la console d'administration Pentaho en saisissant l'URL suivante dans un navigateur Web : http:\\localhost:8099
Puis en saisissant l'identifiant et mot de passe ci-dessous :
login : admin ;
password : password.

5.4 Configuration du système :

Dans cette partie, nous allons illustrer le déploiement des différents outils.

5.4.1 Pentaho data intégration (PDI) :

Le processus d'ETL est un processus primordial dans l'entrepôt de données, il permet de mettre en place un lien entre le système source et le magasin de données, en faisant le transfert de données. Après le transfert, il entame l'épuration de données afin d'extraire les données utiles de chargement dans la base de notre magasin. Ce processus est réalisé grâce à Pentaho Data Intégration aussi connu sous le nom de KETTLE.
Pentaho Data Intégration est l'ETL de la suite décisionnelle libre Pentaho. Cet ETL est un «moteur de Transformation»: les données et les traitements à effectuer sont parfaitement séparés, on parle d'ETL.

Les traitements sont stockés dans un référentiel (repository) qui peut être soit au format XML (fichiers plats), soit dans une base de données (ce qui permet notamment le partage entre plusieurs designers). De nombreux types de SGBD sont supportés (Pentaho assure la connexion à n'importe quelle base de données au travers d'un driver JDBC) ainsi que tous les types de fichiers plats (CSV, Excel, XML).

Pentaho Data Intégration dispose d'une interface graphique « Spoon », depuis laquelle on peut créer deux types de traitements :

- **Des transformations :** celles-ci constituent les traitements de base d'intégration de données avec toutes les étapes nécessaires à l'extraction, la transformation, et le chargement des données.
- **Des tâches :** celles-ci permettent l'ordonnancement de plusieurs transformations. Plusieurs types de tâches sont disponibles : gestion des erreurs, envoi de mails de notification, transfert FTP, exécution de scripts Shell ou SQL.

La figure suivante illustre une transformation.

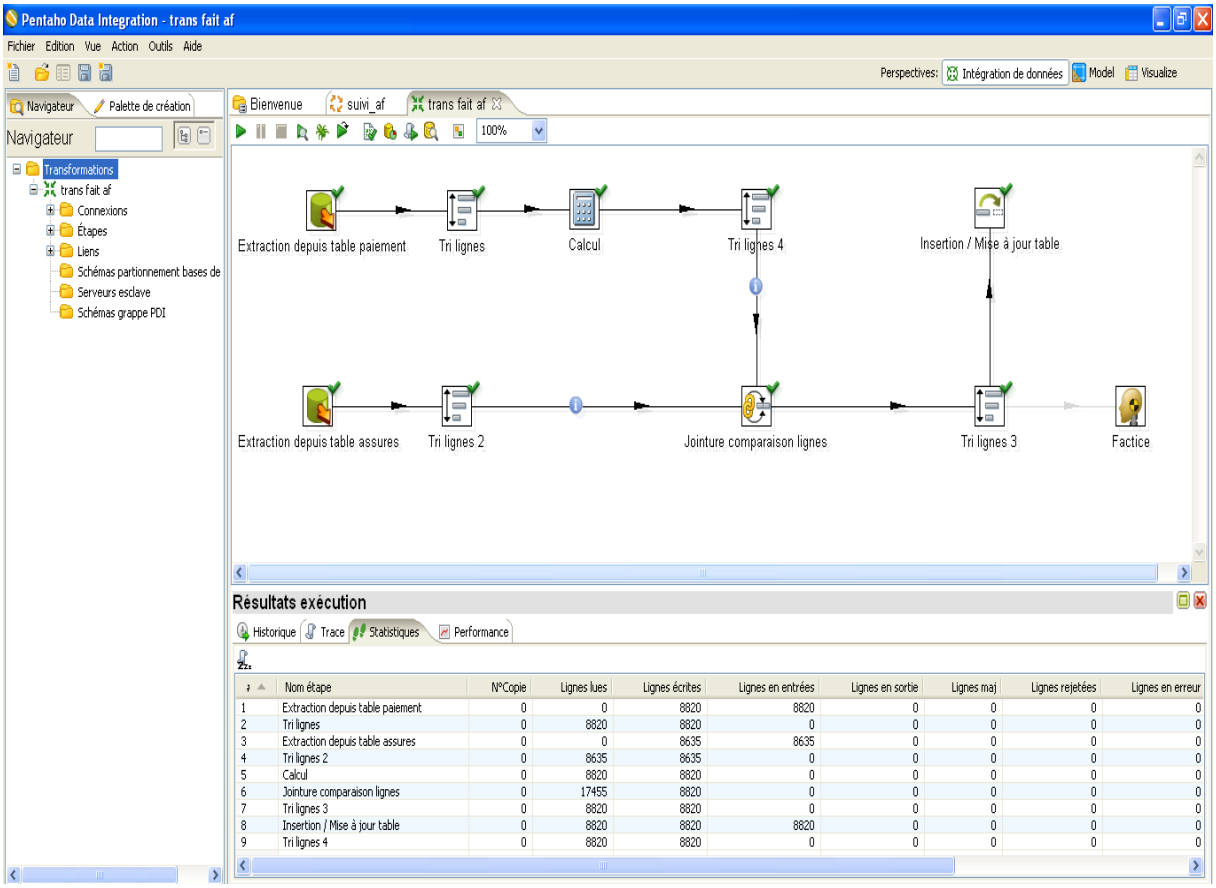

Figure 5.16 : Etapes de la transformation de la table fait.

Une fois les transformations réalisées, une tâche est nécessaire pour ordonnancer les différentes transformations.

La figure suivante illustre une tâche qui est composée des différentes tâches nécessaires pour le chargement des dimensions suivies par l'exécution de la transformation de chargement du fait.

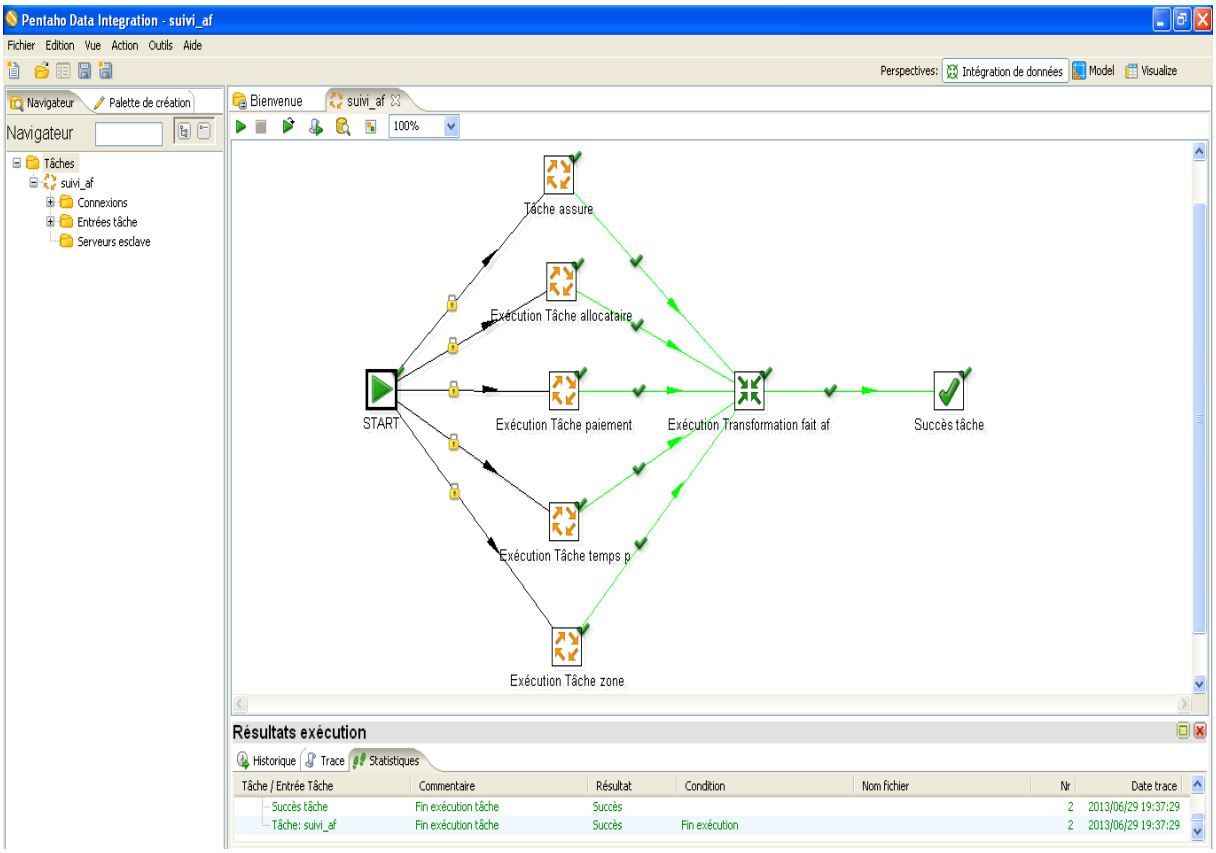

Figure 5.17 : Etapes de la tache chargement de la table fait.

5.4.2 Serveur d'application Tomcat :

Il s'agit en fait d'un serveur d'application sur lequel est déployée l'application web pentaho permettant la diffusion des vues d'analyse. Une JRE (Java Runtime Environement «environnement d'exécution java») est intégrée à cette version.

Tomcat est un serveur libre de servlet J2EE, il est nécessaire de l'installer pour pouvoir utiliser pentaho analysis. Il a été créé par la fondation Apache, il implémente les spécifications des servlet Java et des JSP.

5.4.3 Schéma Workbench :

Schéma workbench est un outil graphique recommandé par l'entreprise Pentaho pour construire des cubes très rapidement. Le cube de notre magasin de données sera enregistré dans un fichier schéma workbench mondrian sous format XML. Workbench permet de créer la table de fait du cube, ses mesures et ses tables de dimensions, ainsi que les différentes hiérarchies du cube. Après sa création, le cube est publié pour permettre aux utilisateurs de le consulter.

La figure suivante montre le schéma de notre cube créé avec le workbench qui sera publié dans notre application.

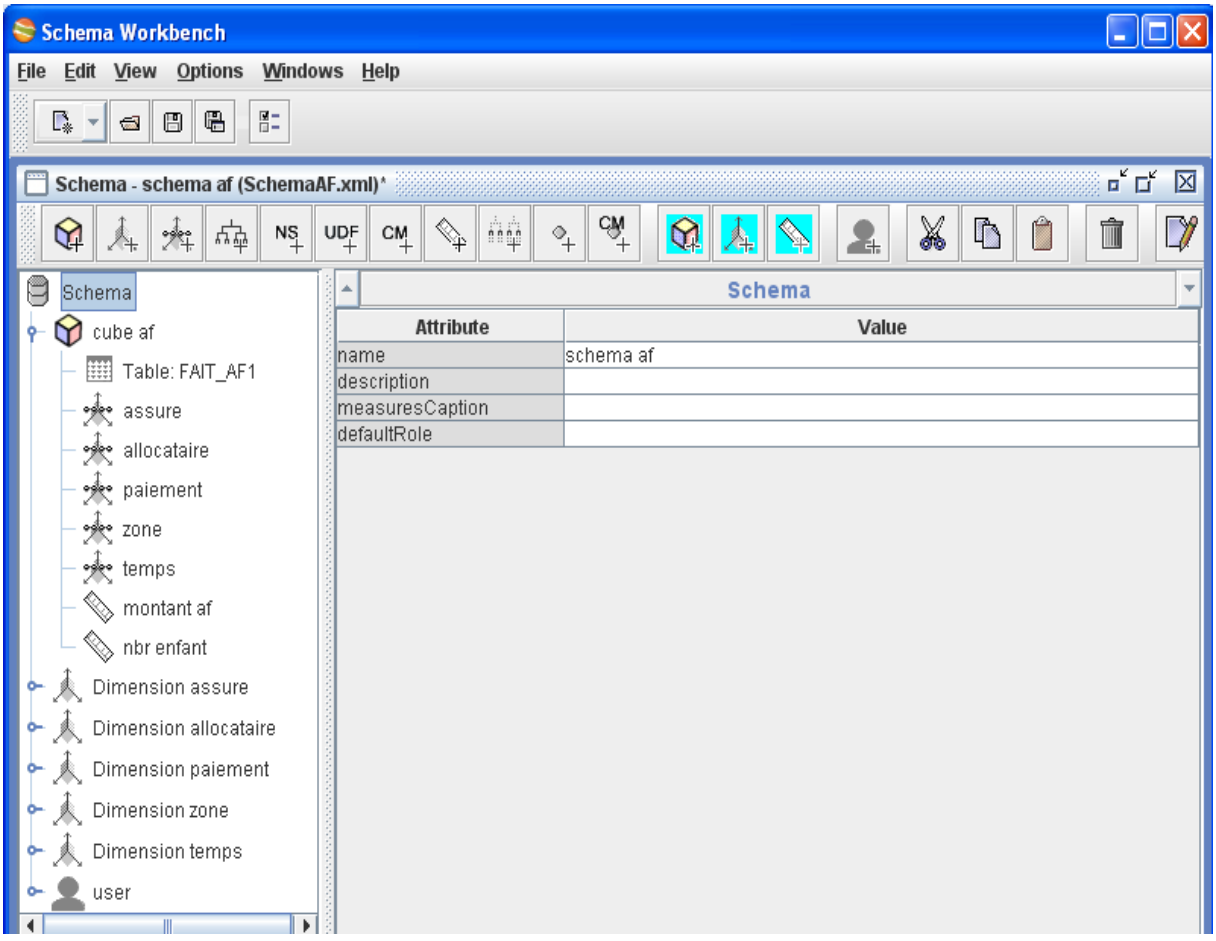

Figure 5.18 : Schéma workbench du cube allocation familiale.

5.4.4 Pentaho Analysis Mondrian :

Mondrian est un moteur OLAP écrit en java, il est fourni sous la forme d'une application web qui s'intègre au serveur d'application. Il fournie :

- La connexion à la base de données via JDBC.
- Le moteur OLAP d'interprétation des cubes.
- La présentation des données via un navigateur web.

Mondrian est composé de 4 couches :

- **Couche de présentation via un navigateur web :** elle permet à l'utilisateur d'observer les résultats des requêtes et d'interagir sur les données présentées (application de filtres, visualisation avancées, export, …)
- **Couche dimension :** c'est la couche qui valide et effectue le calcul de la requête MDX (langage de requête crée pour la manipulation des données multidimensionnelles).
- **Couche d'agrégation :** c'est une couche qui maintient en cache les données agrégées, dans le but d'éviter ainsi de multiples requêtes dans la base de données à chaque action de l'utilisateur.
- **Couche stockage :** c'est une couche représentative de la base de données. L'accès aux données s'effectue via le composant JDBC spécifique au type de base de données.

5.5 Interfaces utilisateur :

Après avoir présenté la plate forme de notre application (zone de préparation des données), nous allons présenter les fonctionnalités offertes à l'utilisateur (zone de restitution des données et de génération des rapports).

5.5.1 Interface administrateur :

Cette interface dédiée uniquement à l'administrateur de données permet de créer, modifier ou supprimer des utilisateurs.

On peut aussi attribuer à chaque utilisateur un ensemble de droit, avec les quelles l'utilisateur peut consulter, modifier ou supprimer des vues (vue d'analyse) ou des rapports publiés par d'autre utilisateurs. On peut aussi créer de nouvelles connections (ODBC, JDBC, ONDI) vers des bases de données ou d'autres solutions pentaho.

Cette interface est dédiée aussi à la gestion des services (déclenchement, rafraîchissement ou arrêter).

La figure suivante représente l'interface administration de Pentaho Suite BI.

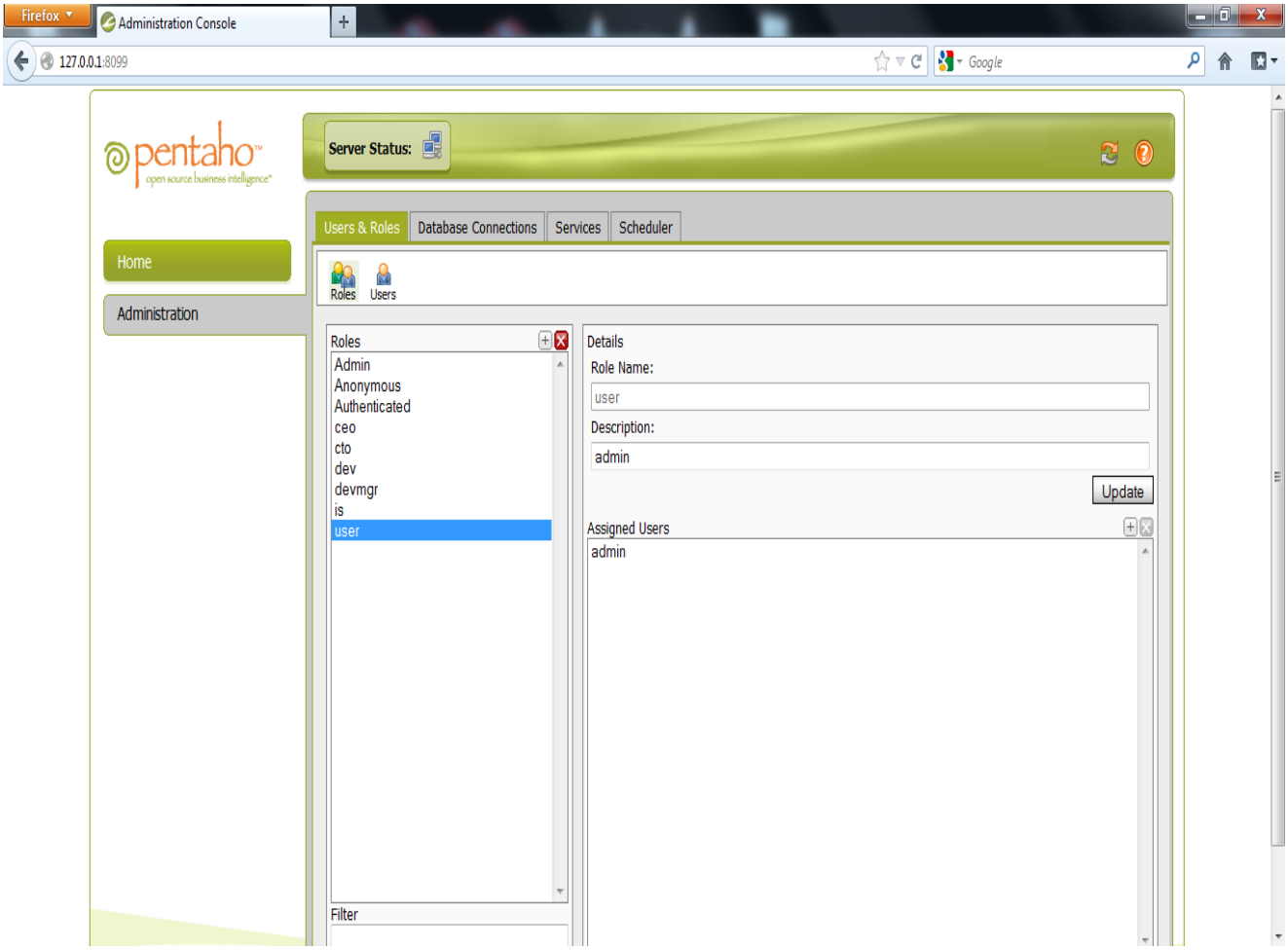

Figure 5.19 : Console administrateur.

5.5.2 Interface décideur :

La page d'accueil donne accès aux utilisateurs finaux suite à une authentification de l'utilisateur, dont le login et mot de passe ont été définis par l'administrateur.

La figure suivante représente la page d'accueil des utilisateurs finaux (décideurs).

La figure 5.21 représente une vue d'une analyse multidimensionnelle, élaborée par un utilisateur.

Figure 5.20 : Console utilisateur.

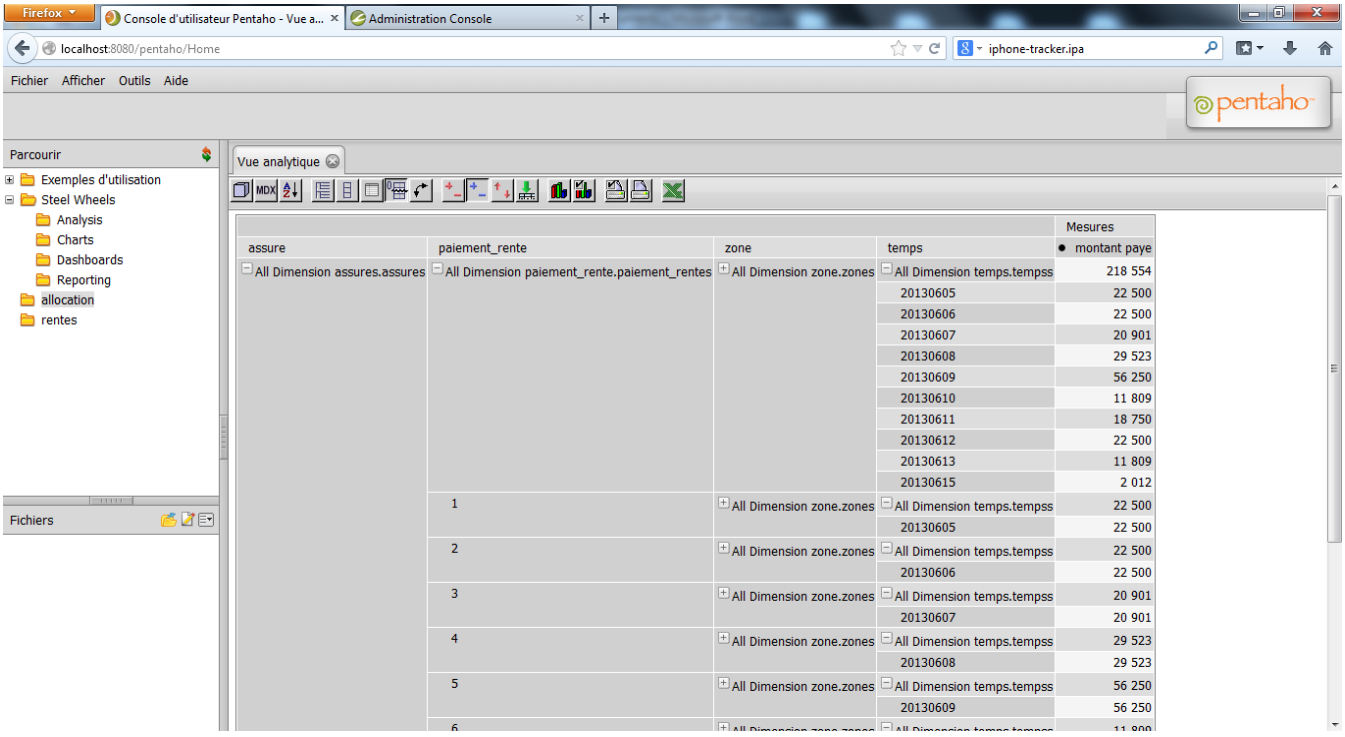

Figure 5.21 : Vue d'une analyse multidimensionnelle.

5.5.3 Les rapports :

L'utilisateur final peut aussi à travers son interface consulter les rapports élaborés auparavant par l'administrateur en utilisant le 'pentaho report designer' (rapport dit statique), ou des rapports édités par d'autres utilisateurs (rapport dit dynamique) en utilisant l'interface 'user' elle même. Pour éditer ces rapports on a utilisé Pentaho Report Designer.

Pentaho Report Designer permet de développer des rapports complexes en association avec la plateforme Pentaho et de les publier directement sur le serveur décisionnel. Il supporte les formats PDF, HTML, csv, Excel et Texte. Très modulaire, il permet aux utilisateurs de développer en Java de nouveaux types de contenus suivant leurs besoins. Ce moteur se base également sur des définitions d'états en XML et peut être intégré dans toute application Java. Pentaho report designer est un outil simple à manipuler, bien intégré à la suite décisionnelle Pentaho BI pour la gestion des paramètres ou la publication sur la plateforme web.

5.5.4 Sécurité du système :

La qualité d'un système d'information suppose au préalable une protection suffisante et cohérente vis-à-vis des risques possibles. Le système mis en place doit être capable de se défendre contre toute attaque. Ceci est possible par l'élaboration d'un plan de sécurité rigoureux qui protégera la capitale information de l'entreprise contre tout individu malsain.

- **Niveau utilisateur :** Pour sécuriser l'accès utilisateurs, la direction informatique de la CNAS dispose d'un mécanisme d'autorisation et d'authentification. En ce qui concerne la sécurité des machines, des antivirus sont installés pour pallier aux attaques de virus et autres logiciels espions, des mises à jour sont effectuées régulièrement.

- **Niveau application :** Le mécanisme de sécurité de notre application et de notre page web est basé sur le mécanisme d'authentification (login et password) et d'autorisation (les droits de publication et de modification des vues et des rapports).

- **Niveau base de données :** C'est la sécurité du serveur de la base de données basée aussi sur le mécanisme d'autorisation et d'authentification (l'accès à la base) et sur la sauvegarde périodique des données au niveau de l'entrepôt de données.

5.6 Conclusion :

Dans ce dernier chapitre, nous avons abordé la partie de réalisation de notre outil. Nous avons commencé par décrire l'architecture technique que nous avons déployée. En suite nous avons décrit Pentaho BI solution ainsi que ses différents outils : Pentaho data intégration pour l'implémentation de l'ETL, Schéma workbench pour la construction du cube, Mondrian et tomcat pour le serveur d'application OLAP, et enfin Pentaho report designer pour les rapports et les états de sortie.

CONCLUSION GÉNÉRALE

L'objectif de ce travail est de développer un outil ETL (Extraction, Transformation, Load) pour un entrepôt de données.

Dans ce document, nous avons commencé par l'étude de l'existant afin de mieux comprendre le système opérationnel et dégager une liste d'objectifs à fixer à partir des besoins exprimés par les décideurs. Pour notre conception nous, avons choisis le modèle en étoile ainsi que la méthodologie a neuf étapes proposée par Kimball.

La base de données a été implémentée sous le SGBD open source PostgreSQL. Cette distribution de PostgreSQL, connu pour ses performances par rapport aux bases de données volumineuses, intègre un ensemble d'outils d'administration et de configuration. Aussi ce SGBD est pré configuré pour la mise en place d'un Data Warehouse.

Une collection entrepôt de données a été créée, ainsi qu'une autre collection source Oracle a été importée pour servir d'environnement de travail.

Des méthodes ont été mises en œuvre afin d'extraire, modifier et insérer les données provenant des sources vers l'entrepôt de données. Ces méthodes ont été appuyées par un ensemble de taches réalisées avec les outils proposés par Pentaho, qui à l'avantage d'être gratuit et open source mais qui s'avère difficile à manier.

Une interface utilisateur a été mise au point afin de permettre aux décideurs d'élaborer des rapports pour des analyses.

Ce mémoire a été une expérience enrichissante, notamment en nouvelles connaissances. En effet, le développement d'un outil informatique qui entre dans le domaine de la recherche laisse la voie libre à la découverte de nouvelles technologies, tout en permettant de mieux mettre en pratique les connaissances acquises, et de comprendre les exigences qu'implique le fait de travailler sur des plates-formes déjà développées.

Bibliographie

[1] Passer en mode BI, Yazid Grim: www.developpez.com

[2] holtzman s. intelligent decision systems reading, massachusetts": addisonwesley. 1989.

[3] P.levine et M.J Pomerol 1989, Systèmes interactifs d'aide à la décision et systèmes experts. Hermes, 1989.

[4] Howard Dresner, « BI: Making the Data Make Sense, » Gartner group, May 2001.

[5] SQL Server 2000, Analysis services et DTS Cyril Guau mars 2004.

[6] Jörgen Fredman, Mathias Horndahl, Lars Tong Strömberg 1999 : a framework for the design of organizational decision support systems.

[7] R. Kimball et J. Caserta ; « The Data warehouse ETL Toolkit» ;Wiley Publisshing, INC 2004

[8] Ralph Kimball, « Le Data Warehouse Guide de conduite de projet », Groupe Eyroles, 2005 pour la nouvelle présentation.

[9] E. F. Codd ; « Providing OLAP (On-Line Analytical Processing) to User- Analysts : an IT mandate. » ; Technical report ; E.F. Codd & Associates; 1993.

[10] Business Intelligence avec SQL serveur 2005 : mise en Œuvre d'un projet décisionnel, Burquier Bertrant.

[11] W H Inmon. Building the Data warehouse.Qed Technique Publishing Groupe, Wellesley, Massachusetts, U.S.A, 1992.

[12] Ralph Kimball. The Data Warehouse Toolkit. John Wiley, U.S.A, 1996.

[13] Theodotatos D. Bouzeghoub M. (2000). A general framework for the view selection problem for data warehouse design and evolution. DOLAP 00, ACM Third International Workshop on data warehousing and OLAP.

[14] E.Kerki, «Processus de mise en oeuvre d'entrepôt de données 'Approche sémantique' » Rapport de recherche entre l'hôpital du Bocage et l'université de Bourgogne, Dijon, France 2001.

[15] Didier Nakache, " Data warehouse et Data mining " , Conservatoire nationale des arts et métiers de Lille, version 1.1, 15 juin 2008.

[16] Cécile Favre. "Evolution de schémas dans les entrepôts de données mise à jour de hiérarchies de dimension pour la personnalisation des analyses''. Thèse de doctorat, Université Lumière-Lyon2, Lyon, France, 2007.

[17] Ralph Kimball: « Entrepôt de données », International Thomson Publishing France, 1997.

[18] R. Kimball et M. Ross ; « Entrepôts de Données : Guide Pratique de Modélisation Dimensionnelle 2 ème édition » ; Vuibert 2002.

[19] B. Inmon; What is a Data Warehouse; Article; http://www.billinmon.com; 2000.

[20] P. Marcel, « Manipulation de Données Multidimensionnelles et Langages de Règles », Thèse de Doctorat de l'institut des Sciences Appliquées de Lyon, 1998.

[21] Edgard Ivan Benitez Guerrero. Infrastructure adaptable pour l'évolution des entrepôts de données. Thèse de doctorat, Université Joseph Fourrier – Grenoble1, Grenoble, France, 2002.

[22] Buzydlowski et al, 1998; J.W. Buzydlowski, I.Y. Song, L. Hassell, « A Framework for Object-Oriented On-line Analytical Processing », 1st International Workshop on Data Warehousing and OLAP (DOLAP'98), pp.10-15, Bethesda (Maryland, USA), Novembre 1998.

[23] S.Boukhdouma, N.Selmoune: «BASE DE DONNEES ET SGBD». Edition Pages Blues;Septembre 2007.

[24] Présentation et installation de Pentaho : www.osbi.fr

[25] Téléchargement de Pentaho : http://sourceforge.net/projects/pentaho/files