REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique Université Mouloud MAMMERI de Tizi-Ouzou

Faculté de Génie Electrique et d'Informatique

Département d'informatique

MEMOIRE

DE FIN D'ETUDE

En vue de l'obtention d'un diplôme de Master en Inf ormatique.

Option : ingénierie des systèmes d'information.

Mise en place d'un système d'information décisionnel

Cas : CNR Alger

Encadré par :

- *Mme AMIROUCHE Fatiha.*

Co-encadré par :

- M.*BELLEMOU Rachid.*

Réalisé par :

- *KHELIFI Sabrina*
- LOUGGAR Imene

Promotion 2017-2018

Remerciments

Nos remerciements vont tout spécialement à nos familles, qui ont sus nous supporter et encourager tout au long de notre vie, ainsi que pour leur aide inestimable, leur patience et leur soutien indéfectible.

Nous tenons aussi, à remercier tous les enseignants qui ont contribué de près ou de loinà notre formation.

Nous remercions ***Mme Amirouche*** *pour avoir assuré l'encadrement de ce projet et pour nous avoir accompagnées dès le début de ce projet. Nous tenons à lui exprimer notre gratitude pour ses valeureux conseils.*

On remercie également ***Monsieur Bellemou*** *sous directeur de l'organisation de la CNR pour nos séances de travail agréables et fructueuses, ses remarques pertinentes,*
mais aussi pour son écoute et son discours bienveillant.

On remercie vivement Mesdames et Messieurs les membres du jury d'avoir accepter
d'évaluer ce travail.

Imene et Sabrina

Dédicaces

On dédie ce travail à :

Nos très chers parents

Qui nous ont toujours fait confiance et n'ont jamais cessé

De nous encourager et de nous soutenir, Nos famille sans exception, Nos amis sans exception Et nos camarades.

Imene et sabrina

Sommaire

Table des figures

Table des tableaux

Résumé :

La CNR (Caisse Nationale des Retraites) est un établissement public faisant partie des caisses de la sécurité sociale et dont la principale mission est d'assurer le service des prestations des retraités (3.1 millions de retraités actuellement) dans les meilleures conditions possibles. Elle est constituée d'une direction générale et de 51 agences au niveau national.

Les systèmes opérationnels actuellement utilisés dans la CNR sont décentralisés. Chaque agence manipule les données concernant les retraités qui lui appartiennent uniquement. Ceci rend très difficile l'obtention d'une information globale par la direction générale, et nécessite une consolidation manuelle des données provenant des différentes agences pour l'établissement des rapports d'analyse et de synthèse.

Cependant, la CNR a dernièrement connu un certain nombre d'événements qui ont engendré la nécessité de prise en charge de nouveaux besoins en termes d'analyse. Ainsi, afin de répondre aux requêtes et aux questionnements des citoyens, de la presse et du ministère de tutelle ; elle a décidé de mettre en place un système décisionnel permettant d'assurer un suivi global de son activité au niveau des 51 agences et de fournir une information fiable et agrégée facilitant l'analyse et permettant d'améliorer le processus de prise de décisions.

La direction générale de la CNR nous a donc confié le projet de réalisation d'un Data Warehouse pour la mise en place d'un système en exploitant les fichiers log de l'entreprise en vue d'optimiser certaines de ces ressources.

La réalisation de ce projet passe par plusieurs étapes, à commencer par la collecte et l'analyse des besoins des utilisateurs du nouveau système, la conception et la réalisation du Data Warehouse et enfin la présentation des informations aux utilisateurs.

Mot clés : Data Warehouse, Système décisionnel, CNR, Tableau de bord, OLAP.

I. Introduction générale :

Les dirigeants d'entreprises, quelque en soit le domaine d'activité, doivent être en mesure de mener à bien leurs missions. Ils doivent notamment prendre les décisions les plus opportunes en temps voulu.

Les décisions, fondées et pertinents, sont basées urs des informations claires, fiables et pertinentes. Le problème est de savoir donc comment identifier et présenter ces informations à qui de droit, sachant que d'une part les entreprises croulent 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 unsupport appréciable à la prise de décision.

C'est dans ce contexte que les « systèmes décisionnels » ont vu le jour. Ces systèmes sont sensés offrir aux décideurs outre des informations de qualité sur lesquelles ils pourront s'appuyer pour arrêter leurs choix décisionnels, des outils et méthodes pour une prise de décision efficace.

Notre travail dans le cadre de ce mémoire se situe dans le contexte des systèmes décisionnels et consiste à mettre en place une solution décisionnelle pour la gestion des retraites au niveau de la caisse nationale des retraites (CNR), agence d'Alger.

L'activité journalière de la CNR génère des données complexes et volumineuses. Ces données représentent une source précieuse d'informations, qui serait à même d'améliorer de façon significative le processus de prise de décision. Cependant, ces données ne sont pas exploitées de manière satisfaisante.

Notre présent projet consiste en la mise en place d'un système décisionnel en mesure de consolider les données issues des systèmes opérationnels, et d'offrir des informations de qualité pour les décideurs. En particulier, il s'agira de traiter des fichiers journaux provenant du serveur web de la CNR contenant des pages web statiques et dynamiques, de formater le contenu et d'en analyser les données puis d'établir un rapport qui aidera à la prise de décision.

Chapitre 1 : les systèmes d'information décisionnel.

Introduction :

Toutes les entreprises disposent d'une masse de données plus ou moins considérable. Ces informations proviennent soit de sources internes ou de sources externes, c'est pourquoi les entreprises s'intéressent de plus en plus au management de leur capital informationnel.

Pour un pilotage efficace, se doter d'un système d'information décisionnel est nécessaire afin d'aider les décideurs à prendre les bonnes décision au moment opportun.

A travers ce chapitre, nous allons aborder la notion de système d'information décisionnel et son architecture.

1. Définition et objectifs

Un système d'information décisionnel est : « L'ensemble des outils informatiques (matériels et logiciels) qui permettent l'analyse des données opérationnelles issues du système d'information des entreprises. Ces données sont transformées en une isionv orientée décideur puis analysées au moyen de manipulations et restitutions adaptées.»**[TOURNIER, 2007].**

2. Historique :

Les systèmes d'information ont connu une évolution considérable à travers le temps comme le montre le découpage temporel suivant :

Année 70-90: débuts de l'informatique décisionnelle

Les prémisses de l'informatique décisionnelle apparaissent à la fin des années 70 sous le nom de *Business intelligence* (BI), appellation originellement utilisée en 1958 par Hans Peter, un analyste d'IBM, auteur d'un article intitulé « *A Business Intelligence System* ».

Les premiers **infocentres** apparaissent au début des années 1980 : Un infocentre consistait dans les années 70 à mettre à la disposition d'utilisateurs finaux toute la puissance de calcul d'un ordinateur en temps partagé au moyen de terminaux ntelligence, de banques de données, de langages (BASIC, FORTRAN, APL…), d'une aide en ligne et d'un e équipe d'assistance technique. L'infocentre, précurseur de l'informatique décisionelle, permettait alors aux utilisateurs finaux de prendre des décisions opérationnelles basées sur des valeurs courantes.

Année 90:- essor de l'informatique décisionnelle [WEB02] :

En 1989 Howard Dresner définit le cadre moderne de la BI: des concepts et méthodes pour améliorer la prise de décision grâce à des systèmes d'analyse de données factuelles. Portée par l'avènement des ordinateurs, l'informatique décisionnelle se normalise. La normalisation s'opère dans deux directions : la *récupération et le stockage des données dans des unités spécifiques*d'une part, et *l'amélioration des capacités d'analyse*d'autre part. C'est ainsi que se mettent en place les **entrepôts de données** (ou **datawarehouse**) qui, grâce à des outils informatiques baptisés ET L (*Extract, Transform and Load*), collectent et organisent les informations issues des différentes applications des entreprises ainsi que des donnéeséconomiques générales du secteur d'activité. Des analystes spécialistes sont chargés d'exécuter lesrequêtes grâce à des **applications de type OLAP** (*On Line Analytical Processing*) qui permettent un traitement analytique en ligne. Ces spécialistes

éditent tableaux et rapports de synthèse régulierset offrent aux dirigeants une connaissance complète de leur entreprise et de son contexte.

3. Le Processus décisionnel :

Le processus décisionnel est pyramidal : il consiste à transformer le volume de données d'une entreprise en informations pertinentes à partir des quelles les décideurs peuvent tirer des connaissances afin d'aboutir à de bonnes décisions touchant tous les niveaux de l'entreprise.

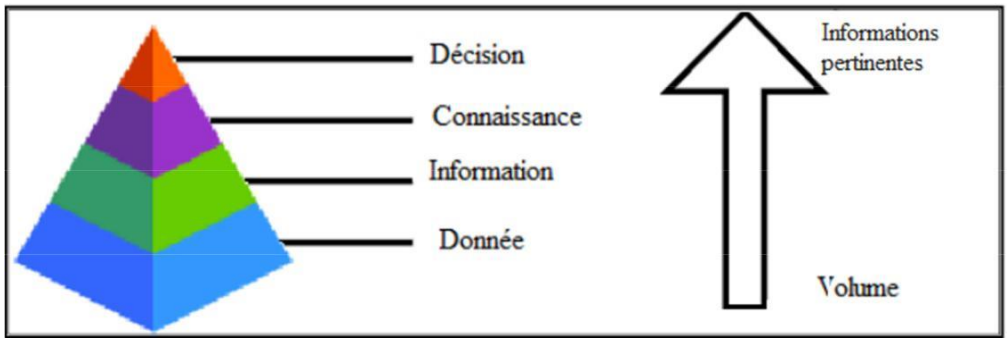

***Figure 1**: processus décisionnel.*

La figure 1 schématise le processus pyramidal d'un processus décisionnel, nous allons dans ce qui suit expliquer chacun de ses composants :

- **Donnée :**

« Par définition, une donnée est un élément brut,uiq n'a pas encore été interprété ou mis en contexte » **[BRUNOCHAUDET 2009]**

- **Information :**

« Une information est par définition une donnée interprétée. En d'autres termes, la mise en contexte d'une donnée crée de la valeur ajoutée pour constituer une information » **[BRUNOCHAUDET 2009]**

- **Connaissance :**

Selon Jean Louis Levet économiste spécialiste**[LEVET 2001]**: « *la connaissance est d'abord une capacité d'apprentissage … la propriété essentielle de la connaissance est de pouvoir par elle-même engendrer de nouvelles connaissances… la connaissan ce est composée non seulement d'information* à *caractère publique, mais aussi de savoir-faire in exprimable formellement et donc difficilement transférable*».

- **Décision :**

La décision est un acte par lequel un décideur opère un choix entre plusieurs options permettant d'apporter une solution satisfaisante à un problème donné, ou d'exécuter d'une action ou un projet, avec toutes les conséquences que cette décision pourrait engendrer **[WEB 01].**

4. La place du décisionnel dans l'entreprise

La mise en place d'un système d'information décisionnel se justifie par la volonté du chef d'entreprise à assurer une bonne gouvernance et per formance de l'entreprise.

Pour ce faire, comme l'illustre le schéma de la figure 2, le chef d'entreprise doit définir, après une analyse des besoins une stratégie adaptéeà l'entreprise, décider des objectifs et priorités et nommer les responsables qui traduiront les objectifs en indicateurs de performance et en tableaux de bord. Ces responsables piloteront ensuite le système qui a été réalisé par des informaticiens spécialisés dans le temps qui leur était imparti**[WEB12]**.

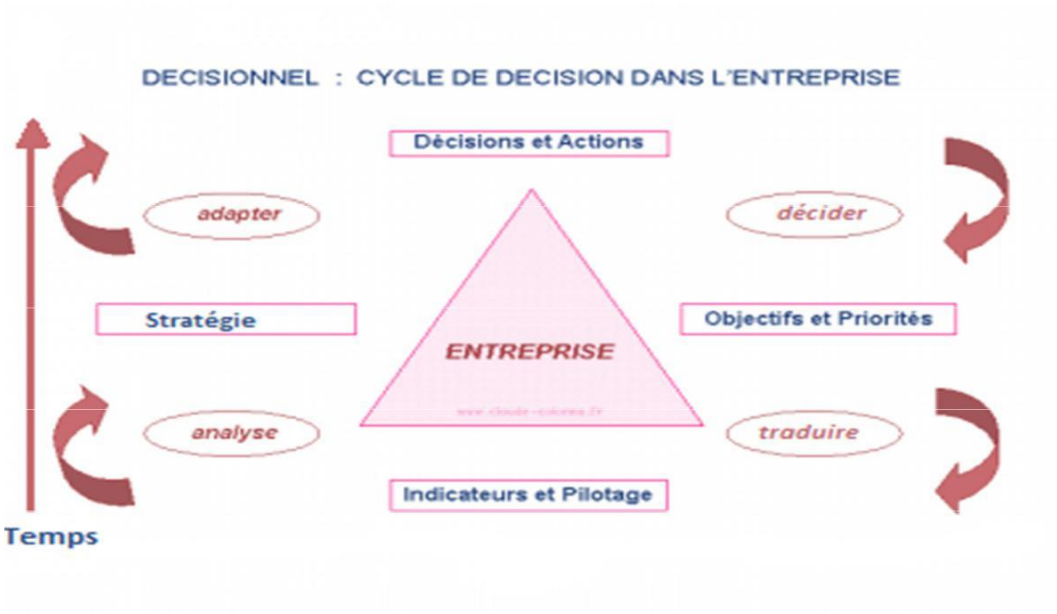

Figure 2: *cycle de décision dans l'entreprise.*

Un **système** décisionnel, c'est aussi la collecte et le stockage des données de l'entreprise provenant de différentes sources et leurs restitutions pour aider les managers à prendre de bonnes décisions.

Les Objectifs d'un **système d'information décisionnel**sont : **[WEB03]**

1 Rapidité: pour que des décisions soient prises à temps, on sera pragmatique et on adoptera une démarche itérative pour parcourir le plus vite possible le chemin entre stratégie, expression de besoins et utilisation.
2 Justesse: pour prendre des décisions, on a besoin d'informations justes, adaptées aux objectifs et rapidement accessibles. On se fabrique donc un entrepôt de données sur-mesure (*Datawarehouse*H) ;
3 Efficacité: pour que les décideurs soient en mesured'analyser et de traduire les faits en décisions, il existe des outils spécialisés en reporting décisionnel.
4 Pertinence : Transformer les données recueillies en informations pertinentes voire en connaissances.
5 Optimalité : Répondre de manière optimisée aux requêtes d'outils de reporting et tableaux de bord d'indicateurs, et mis à la disposition des responsables opérationnels.**[WEB04]**

5. Architecture d'un Système d'information décisionnel:

Le processus des systèmes d'information décisionnel vise à **collecter** puis **stocker** des données brutes et à les transformer en informations utiles à la décision qui seront ensuite **diffusées**sous forme de tableau de bord. **[Stephan LAU, 2009]**

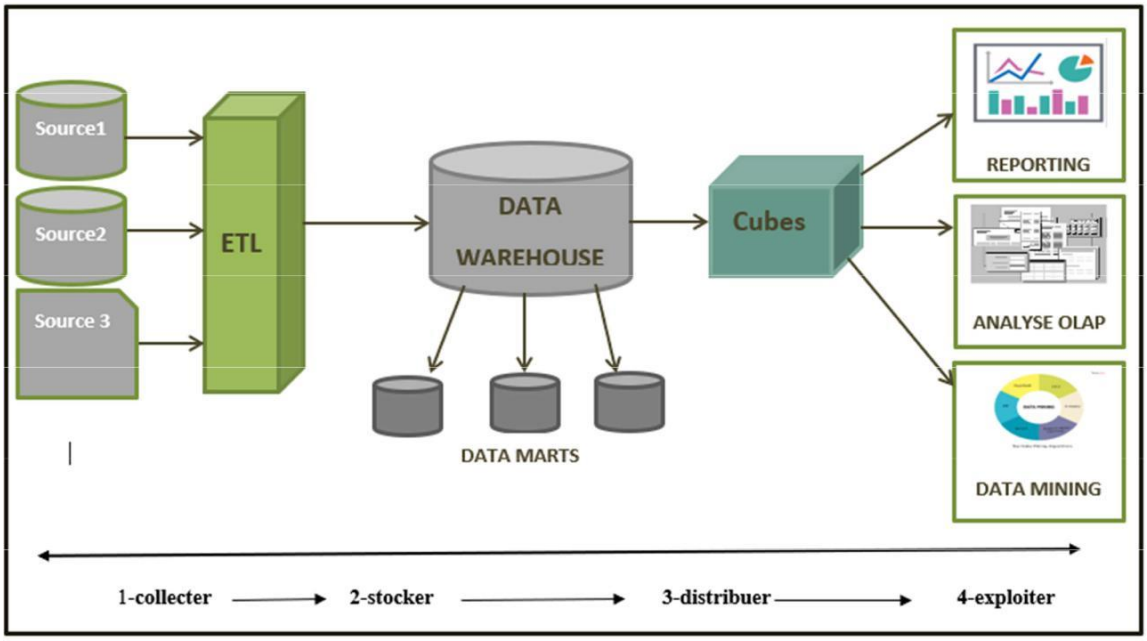

Figure 3: *architecture d'un système d'information décisionnel. [C.Vangenot]*

L'architecture d'un système décisionnel (représentée en figure 3) est articulé autour des étapes suivantes :

1) **Collecte les données :**Afin d'alimenter les entrepôts, les données doivent être identifiées et extraites de leurs emplacements originels. Il s'agit majoritairement de données internes à l'entreprise. Mais il peut aussi s'agir de données de sources externes, récupérées via des services distants. Pour réaliser cette phase, des outils fondamentaux sont utilisés, plus connus sous le nom d'ETL. Les données collectées via un ETL sont fournies ainsi dans un format permettant leur stockage immédiat dans les entrepôts, et ultérieurement exploitables.
2) **Stocker les données:** Durant cette phase, les données précédemment centralisées seront unifiées et structurées au sein d'un entrepôt de données à l'aide de l'ETL, déjà utilisé lors de la collecte, et ce grâce à un connecteur permettant l'écriture dans le futur *datawarehouse*. Ce dernier est l'élément central du dispositif dans el sens où il permet aux applications d'aide à la décision de bénéficier d'une source d'informatio homogène, commune, normalisée et fiable.

3) **Distribuer les données:**Après une structuration multidimensionnelle des données grâce aux cubes OLAP selon les besoins d'analyse, une stimulation de l'activité globale du système est nécessaire, cette étape de diffusion met les informations à disposition de l'ensemble de ses utilisateurs. Cette fonction permet la gestion des droits d'accès en respectant la hiérarchisation des métiers. De ce fait, Il y a deux types de stockage: les *datawarehouses* (qui concentrent l'essentiel des données collectées) etles *datamart* (qui se focalisent sur une partie du métier)

4) **Exploiter les données:**Une fois les données stockées, nettoyées, consolidées et accessibles, les outils d'analyse se chargent de présenter les nformations à valeur ajoutée de la manière la plus visible pour l'utilisateur. On distingue plusieurs types d'outils différents:
   Les outils OLAP (*On-line Analytical Processing*) : pour les analyses multidimensionnelles.
   Le *Data mining* (ou fouille de données): pour la recherche des corélations et des tendances entre les données.
   Les Tableaux de bord : qui présentent les indicateurs clés de l'activité, utiles pour le pilotage de la performance et l'aide à la décision.
   Le *Reporting* : qui consiste à faire le rapport d'activité de l'entreprise.

6. L'entrepôt de données (ou *Datawarehouse*):

6.1 Définition :

Selon BILL Inmon [Jean-François Desnos], « un *datawarehouse* est une collection de données orientées sujet, intégrées, non volatiles, historisées, organisées pour la prise de décision » :

- Orientées sujet:
  - Organisées autour de sujets majeurs de l'entreprise.
  - Données pour l'analyse et la modélisation en vue del'aide à la décision, et non pas pour les opérations et transactions journalières.
  - Vue synthétique des données selon les sujets qui intéressent les décideurs.
- Intégrées:
  - Construit en intégrant des sources de données multiples et hétérogènes.
  - Les données doivent être mises en forme et unifiéesafin d'avoir un état cohérent.
- Non volatiles:
  - ne pas supprimer les données du*Datawarehouse*.
- Historisées:
  - trace des données, suivre l'évolution des indicateurs.
  - Stockage de l'historique des données, pas de mise à jour.

6.2 Différence entre les données opérationnelles les données décisionnelles :

Les entrepôts de données sont créés pour des fins écisionnellesd. Ils permettent de charger des données provenant de différentes sources hétérogènes pour faciliter le processus de prise de décision. Tandis que les BDD relationnelles sont conçues pour supporter des traitements opérationnels journaliers. **[KHOURI 2008]**

Le tableau suivant résume les principales différences entre une BDD opérationnelle et un *Datawarehouse*:

Données opérationnelles	Données décisionnelles
Orientées application	Orientées activité (sujet)
Mise à jour interactive possible	Pas de mise à jour interactive
Accès par une personne à la fois	Utilisées par l'ensemble des analystes
Petite quantité de données utilisées par un traitement	Grande quantité de données utilisée par des traitements
forte probabilité d'accès	faible probabilité d'accès

Tableau 1: *données opérationnelles vs données décisionnelles.*

6.3 Les DataMarts :

Un ***Datamart*** ou **magasin de données**est un sous élément du*dataWarehouse* qui organise les données selon des usages métiers ou des domaines ciblés. Le *Datamart* rassemble un ensemble de données organisées, ciblées, agrégées et regroupéesdans le but de répondre aux besoins des métiers.

D'après bill imno **[WEB13]**

« Le *Datamart* est un **flux de données**en provenance du *DataWarehouse*. Il regroupe de **manière fonctionnelle** les données spécialisées, agrégées pour un métierenparticulier. Le *Datamart* (Dans cette approche) n'est pas au cœur de l'entrepôt de données, mais en périphérie de ce dernier. »

6.4 Architectures des entrepôts de données :

Il existe différentes manières de combiner les différentes composants du *Datawarehouse*, On considère les architectures possibles suivantes : **[S. Chafki, C. Desrosiers]**

6.4.1 Architecture à base de magasins de données indépendants :

Les *Datamarts* sont développés et opèrent de manière indépendante.Cette architecture peut être adoptée dans le cas où les divisions de l'entreprise sont faiblement couplées.

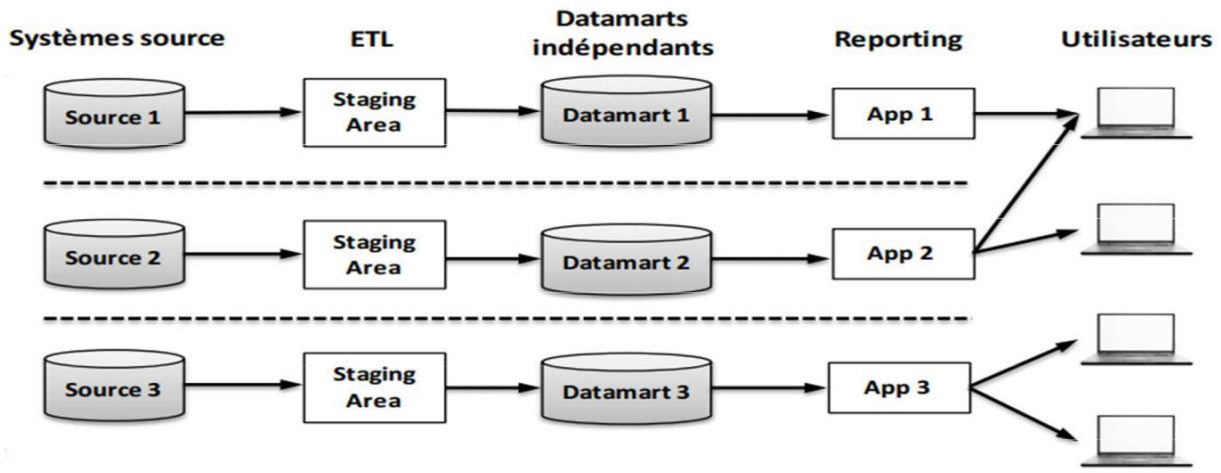

Figure 4: *Architecture à base de Datamarts indépendants [S. Chafki, C. Desrosiers]*

6.4.2 Architecture en bus de datamart :

L'architecture en bus de datamart est constituée d'un ensemble de data marts étroitement intégrés, dont la « source d'alimentation » est un ensemble de dimensions et tables de faits mises en conformité. Une dimension mise en conformité est définie et implémentée une seule fois, et utilisée dans différents schémas en étoile qui forment le datamart d'entreprise, ce qui garantit une intégration logique des *datamarts* et une vue sur l'ensemble des informations de l'en treprise.

Staging area[1] : « STAGING AREA » zone temporaire d'acquisition de données avant leur transformation et leur chargement dans l'entrepôt de données.

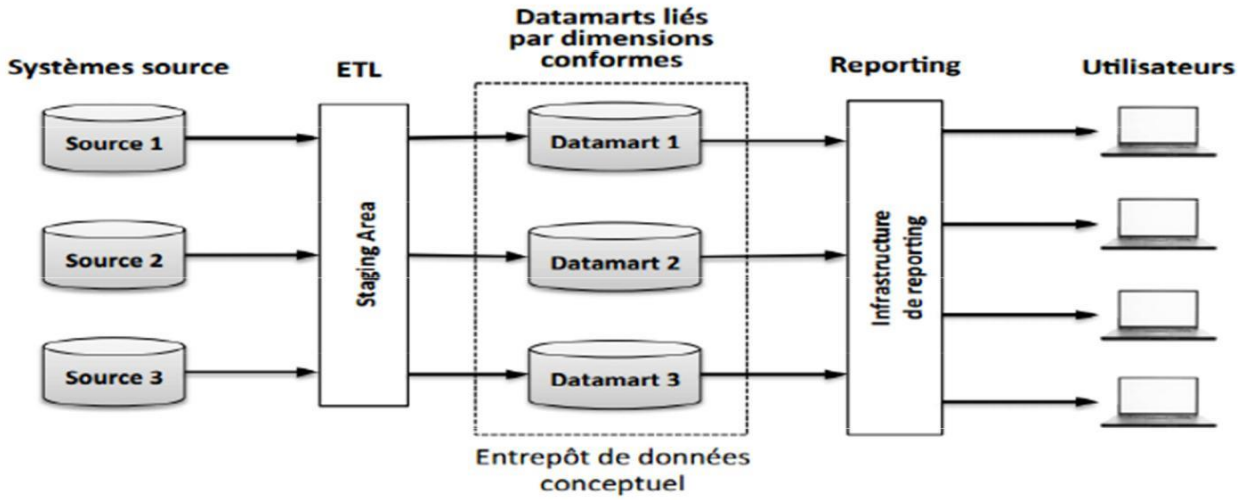

***Figure 5**: Architecture en bus de datamart*

6.4.3 Architecture Hub and Spoke :

Dans l'architecture *hub and spoke* (figure 5), l'entrepôt (qui joue le rôle de **concentrateur** ou ***hub***) contient les données atomiques et normalisées. Les datamarts (jouant le rôle de **rayons** de stockage ou

spokes) reçoivent les données de l'entrepôt. Ces données sont synthétisées (ie. agrégées) et non atomiques. Les utilisateurs accèdent principalement aux *Datamarts*. Ils interrogent rarement l'entrepôt de données.

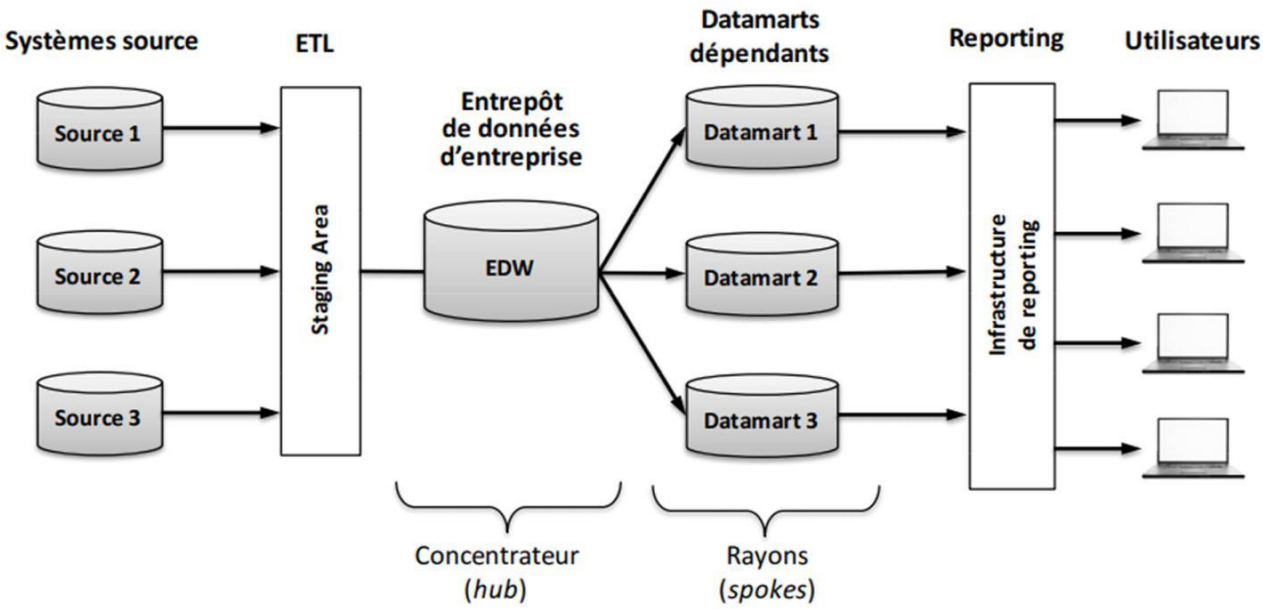

Figure 6: *Architecture hub and spoke*

6.4.4 Architecture centralisée :

Cette architecture, présentée en figure 7, peut être vue comme une implémentation particulière de l'architecture *Hub and Spoke*, où les *Datamarts* sont fusionnés dans l'entrepôt de données.

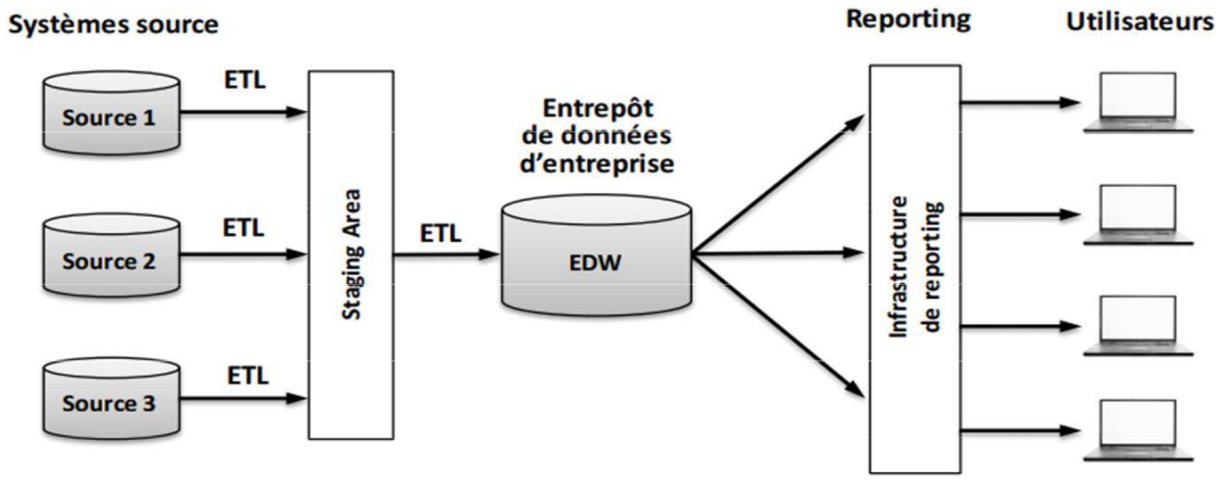

Figure 7: *Architecture centralisée*

6.5 Comparaison entre un DataMart et un Datawarehouse :

DataWarehouse	Datamart
Orienté entreprise contient des données atomiques. utilise un modèle de données normalisé de toute l'entreprise.	orientés processus. contient des données agrégées. utilise des modèles dimensionnels orientés sujet.

Tableau 2: *Data Warehouse vs Datamart*

6.6 Modélisation dimensionnelle d'un Datawarehouse:

La modélisation dimensionnelle, souvent appelée modélisation OLAP (Codd 1993), est une méthode de conception logique qui vise à présenterles données sous une forme standardisée intuitive et qui permet des accès hautement performants et très rapide. C'est une alternative au modèle relationnel, qui permet de représenter les donnéessous forme de cube (centré sur une activité) et non plus sous forme de tables.

La modélisation dimensionnelle repose sur deux concepts fondamentaux : les tables de dimensions et les tables de faits.

6.6.1 Les tables de dimensions :

On entend par dimensions les axes sur lesquels on souhaite faire l'analyse. On peut avoir une dimension client ou produit par exemple.

Les tables de dimension sont un ensemble de tables secondaires possédant chacune une clé de primaire unique correspondant à l'un des composants la clé multiple de la table de faits.

6.6.2 Les tables de faits :

Le « fait» représente le sujet à analyser. Il regroupe un ensemble de mesures (informations opérationnelles). Ces mesures sont stockées dans latable de fait, qui contient aussi une clé multiple composée des clés primaires des tables de dimension qui lui sont associées.

Les mesures numériques dans une table de faits se épartissent en trois catégories (Kimball 2013):

- **Mesures additives :** peuvent être additionnées sur n'importe quelle dimension associée à la table de faits.
- **Mesures semi additives :** peuvent être additionnées sur certaines dimensionsmais pas toutes.
- **Mesures non additives :** ne peuvent être additionnées selon aucune dimensio.

6.6.3 Les cubes de données :

Le cube de données, est une vue restreinte mais intelligente des données de l'entreprise. Composé de cellules qui représentent les mesures (les attributs du fait). Le cube de données permet d'analyser une mesure selon une ou plusieurs dimensions.

Concrètement, un cube se visualise sous la forme d'un tableau croisé dynamique.

Cela permet, dans l'exemple du schéma ci-dessous, d'analyser la répartition de l'indicateur « vente » suivant le temps, les catégories de produit et les régions. En outre, des hiérarchies seront définies pour chaque axe d'analyse (par exemple, l'année, puis la saison, le mois et la semaine, pour l'axe temps).

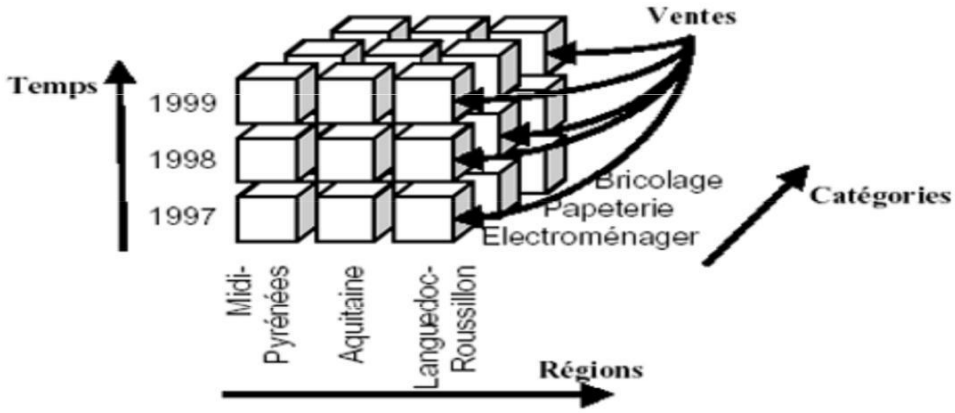

Figure 8: *Représentation d'un hypercube [source : J Detroyes, supinfo]*

6.6.3.1 Opérations sur les cubes de données :

Plusieurs opérations peuvent être utilisées pour viguerna dans les cubes multidimensionnels. Ces opérations sont dites de restructuration. Tout cubeobtenu par une opération de restructuration d'un autre cube contient tout ce qu'il faut pour régénére le cube initial par restructuration réciproque [Bellatreche 2000]. Ces opérations sont :*pivot, switch, split, nest,* et *push*.

- *Pivot* **:** opération de rotation qui permet de tourner le cube pour visualiser une face différente.
- *Switch* **:** opération de permutation qui permet d'interchanger la position des membres d'une dimension.
- *Split* **:** opération qui permet de sélectionner une tranche ud cube selon une de ses trois dimensions.
- *Nest* **:** opération qui permet d'imbriquer des membres issus de dimensions différentes.
- *Push* **:** opération qui permet de combiner les membres d'une dimension aux mesures (les membres deviennent le contenu des cellules).

Il existe deux autres opérations : «*roll-up* » et « *drill-down* » qui sont liées à la granularité et qui permettent d'analyser les données selon les différentes hiérarchies associées à chaque dimension :

- Le ***Roll-up*** **(**ou **forage vers le haut) :** cette opération permet de visualiser les données àun niveau de granularité supérieur conformément à la iérarchie définie sur la dimension.
- Le ***Drill-down*** (ou **forage vers le bas) :** cette opération est l'inverse de l'opération*Roll-up*, elle permet de visualiser les données à un niveau de granularité inférieur conformément à la hiérarchie définie sur la dimension.

6.6.3.2 Implémentations d'un cube de données :

Il existe deux approches fondamentales pour implémenter un modèle de données multidimensionnel : l'approche MOLAP et l'approche ROLAP, ces deux approches se distinguent par la manière de stockage du cube de données.

1) *Approche MOLAP :*

Les systèmes reposant sur l'approche MOLAP (*Multidomensional On-Line Analytical Processing*) stockent le cube de données dans une structure multidimensionnelle : un tableau à *n* dimensions, en utilisant un SGBD multidimensionnel. Chaque dimension du tableau correspond à une dimension du cube.

L'avantage principal de cette approche c'est sa pe rformance en termes de temps de réponse car les différentes agrégations possibles sont pré-calculées et stockées dans le tableau multidimensionnel. Cependant l'approche MOLAP présente quelques limites : [Bellatreche 2000]

- La nécessité de redéfinir les opérations de manipulation de structures multidimensionnelles.
- La difficulté de mettre à jour les données à cause des valeurs du tableau qui doivent être recalculées à chaque opération de mise à jour.
- Ces systèmes sont consommateurs d'espace lorsque les données sont éparses, ce qui nécessite l'utilisation de techniques de compression.

2) *Approche ROLAP :*

Les systèmes reposant sur l'approche ROLAP (*Relational On-Line Analytical Processing*) utilisent l'expérience des SGBD relationnels. Le cube de données est implémenté en utilisant des tables relationnelles. Dans un système ROLAP, chaque fait correspond à une table relationnelle appelée « Table de fait »contient un ensemble d'attributs représentant les mesures à analyser ainsi que les clés étrangères associées aux différentesimensions,d et chaque dimension est représentée à son tour par une table relationnelle appelée « Table de dimension » qui comprend une clé primaire ainsi que les différents niveaux d'agrégation et les propriétés de chaque niveau.

Le principal inconvénient de cette approche est qu'elle présente un temps de réponse trop élevé vu la complexité des requêtes à traiter à travers un schéma relationnel, mais offre par contre la possibilité de stocker de grands volumes de données(Bellatreche 2000).

L'implémentation du cube de données à travers des tables relationnelles peut se faire en utilisant trois schémas :

- Schéma en étoile : *star*(*schema*):

La table de faits est au centre du schéma, et les autres tables des dimensions sont reliées à la table de faits par une seule jointure. C'est la structure de données la plus utilisée et la plus facile pourles utilisateurs de *Datawarehouse*.

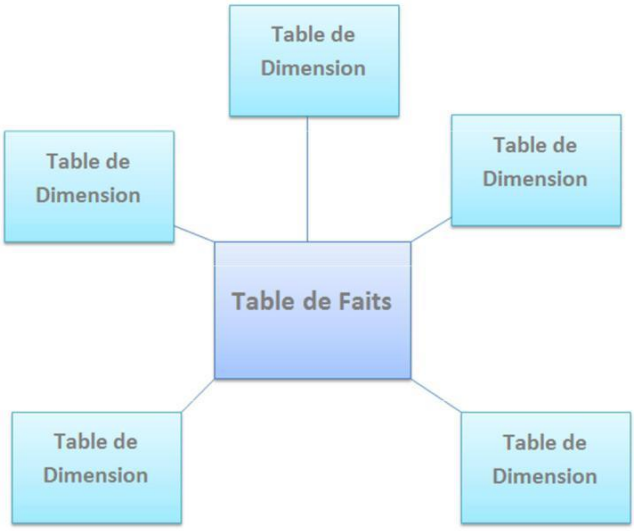

Figure 9: *schéma en étoile*

- Schéma en flocon (*Snow Flakes Schéma*) :

Le modèle en flocon est une variante du modèle en étoile. Il simplifie la normalisation des tables de dimensions. On met les attributs de chaque niveau hiérarchique dans une table de dimension à part.

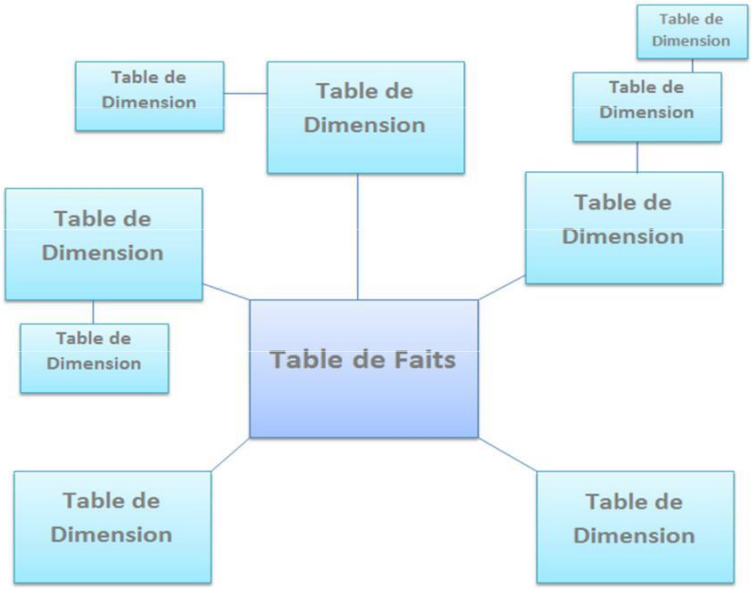

Figure 10: *schéma en flocon*

- Schéma en constellation (constellation schema) :

Ce modèle est la fusion de plusieurs modèles en étoile qui utilisent des dimensions communes.

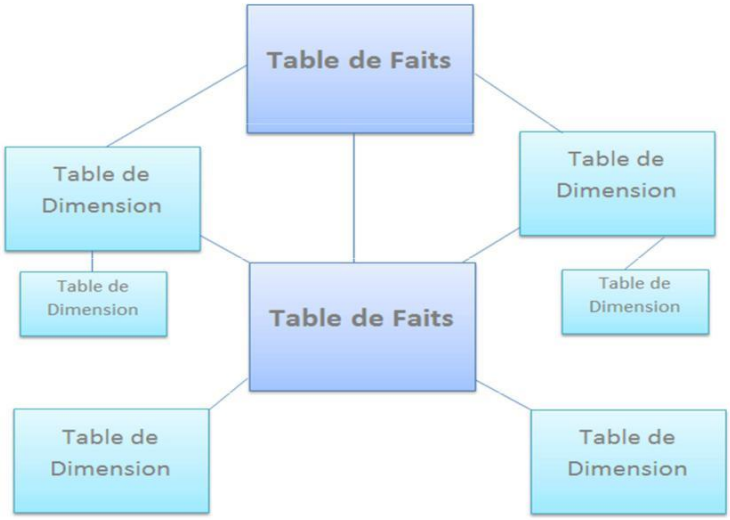

Figure 11: *schéma en constellation*

3) L'approche HOLAP :

L'approche HOLAP (*Hybrid On-Line AnalyticalProcessing*) est une combinaison des deux approches précédentes afin de bénéficier de leursvantages en même temps. Dans cette approche les données fréquemment utilisées sont stockées dans sdetables multidimensionnelles par le biais d'un SGBD multidimensionnel, tandis que les données qui sont rarement utilisées sont stockées dans un SGBD relationnel. La séparation reste transparentepour l'utilisateur. [KHOURI 2008]

6.7 Les approches de conception d'un *DataWarehouse* :

Il existe plusieurs approches de conception de *DataWarehouse*, mais les plus populaires sont: **l'approche orientée sources de données**ou approche « ***Top-Down*** » et **l'approche orientée besoins des utilisateurs** ou approche « ***Bottom-Up*** »**.**

6.7.1 L'approche Top-Down:

Cette approche consiste à commencer par construire tout le *DataWarehouse* de l'entreprise, puis de construire les *Datamarts* par la suite et les alimenter par les données du *Datawarehouse*. Le contenu du *DataWarehouse* est déterminé selon les sources de données et nonselon les besoins des utilisateurs.

6.7.2 L'approche Bottom-Up:

Cette approche consiste en une réalisation incrémentale et indépendante des différents*Datamarts*, qui seront intégrés grâce à une architecture en bus afin de construire l'entrepôt de données.
Le contenu du *DataWarehouse* est déterminé en fonction des besoins des utilisateurs finaux uniquement.

7. Conclusion :

L'informatique décisionnelle est un sujet en pleine évolution et est aujourd'hui l'un des leviers du développement de l'activité des entreprises.

Nous avons, à travers ce chapitre, donné une vue d’ensemble sur l’informatique décisionnelle, ses outils et méthodes, ainsi que son architecture. Parmi les principaux objectifs du décisionnel, essayer de déterminer ou distinguer un comportement à partir d'une source de données, parmi ces sources de données les fichiers logs extraits à partir des serveurs web de l’entreprise, et pour l’exploitation d e ces données on s’intéresse au data mining des pagesweb ou web usage mining que nous allons aborder au prochain chapitre.

Chapitre 2 : Web usage mining

1. Introduction et contexte :

Les entreprises sont inondées de données provenantde différentes sources dont le web. Le Web deviendra bientôt la principale source d'information. Ces masses importantes de données sont inexploitables par les méthodes d'analyse classiques, c'est là que le web mining fait surface afin de développer des approches et des outils, permettantd'extraire des informations pertinentes.

Ces données, en particulier celles relatives à l'usage du Web, sont traitées dans le Web Usage Mining (WUM), que nous abordons dans ce chapitre.

La première utilisation du terme Web mining (WM) revient à Oren Etzioni en 1996, qui a essayé d'appliquer la technologie du data mining sur le Web. Il définit le web mining comme étant " *l'application des techniques du data mining pour l' extraction d'informations pertinentes à partir des ressources disponibles dans le Web*". **[web09]**

Le Web Mining poursuit deux principaux objectifs :

1. ***L'amélioration et la valorisation des sites Web*:** L'analyse et la compréhension du comportement des internautes sur les sites Web permet de valoriser le contenu des sites en améliorant l'organisation et les performances des ites.

2. ***La personnalisation*:** Les techniques de Data Mining appliquées aux données collectées sur le Web permettent d'extraire des informations intéressantes relatives à l'utilisation du site par les internautes. L'analyse de ces informations permet de personnaliser le contenu proposé aux internautes en tenant compte de leurs préférences tede leur profil.

Concrètement, le web mining est divisé en 3 disciplines : le Web Content Mining, le Web Structure Mining et le Web Usage Mining , comme le montre la figure suivante :

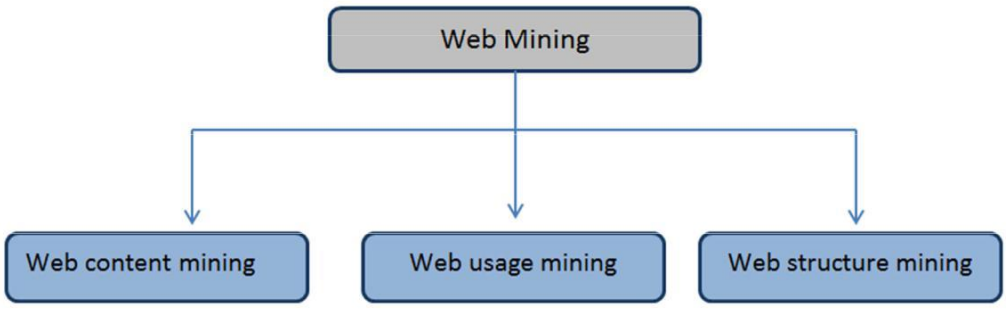

Figure 12: *taxonomie du web mining*

- **Fouille du contenu du web (*web content mining*) :** il s'agit de l'extraction d'informations utiles à partir du contenu de pages Web et de docum ents Web qui sont principalement des

fichiers texte, images et audio/vidéo. Les techniques utilisées dans cette discipline ont été fortement tirées du traitement du langage naturel (TAL) et de la recherche d'information.

- **Fouille de la structure du web (*web structure mining*) :** c'est le processus d'analyse des nœuds et de la structure de connexion d'un site Web à l'aide de la théorie des graphes. On peut en tirer deux choses : la structure d'un site Web en termes de connexion à d'autres sites et la structure du site Web lui-même, ainsi que lamanière dont chaque page est connectée.

- **Fouille de l'usage du web (*web usage mining*) :** il s'agit du processus d'extraction de modèles et d'informations à partir des journaux de serveurs pour obtenir des informations sur l'activité des utilisateurs, sur le nombre de clicssur le site, et sur les activités effectuées surel site.

Nous nous intéressons dans le cadre de notre travail au *web usage mining* que nous détaillons ci-après.

2. Web usage mining :

2.1 Définition :

Le *Web Usage Mining* (WUM) – ou fouille de données d'usage du Web-, désigne l'ensemble de techniques basées sur la fouille de données pour analyser l'usage d'un site Web. En d'autres termes, el WUM correspond au processus d'**E**xtraction de **C**onnaissances à partir des **D**onnées (ECD) issues des fichiers Logs http. L'objectif étant de définirdes modèles comportementaux d'accès au Web en vue de répondre aux besoins des visiteurs de manière spécifique et adaptée, et de faciliter la navigation.

Les profils d'accès à un site Web peuvent être influencés par certains paramètres de nature temporelle (l'heure et le jour de la semaine, des événements saisonniers, etc…). Cependant, la plupart des méthodes consacrées à la WUM prennent en compte dans leur analyse toute la période qui enregistre les traces d'usage: les résultats obtenus sont ainsi ceux qui prédominent sur la totalité de la période.**[ELARBI Nassim 2009]**

2.2 Les données utilisées dans le web usage mining:

a) Les fichiers log :

Un fichier Log (ou fichier journal) est un fichier informatique utilisé pour l'exploitation d'un serveur d'hébergement. Ce fichier comprend les informations (ou logs) enregistrées au niveau des serveurs lorsqu'une requête de chargement de fichiers est effectuée lors d'une visite sur un site web. Les données du fichier journal incluent un enregistrement de l'URL / de la ressource demandée, de l'action effectuée, de l'heure et de la date, de adressel' IP de l'ordinateur d'origine, du type d'utilisateur / navigateur et d'autres informations.

b) Connaissances sur le site web :

Les pages d'un site sont matérialisées par une adresse Internet spécifique, appelée adresse d'allocation de la ressource (*Uniform Resource Locator*). La structure d'un site Internet simple peut être représentée par un arbre dont la racine correspond à la page d'accueil du site. **[ELARBI Nassim, 2009] .** La figure suivante (figure 6) schématise la structure d'un site web:

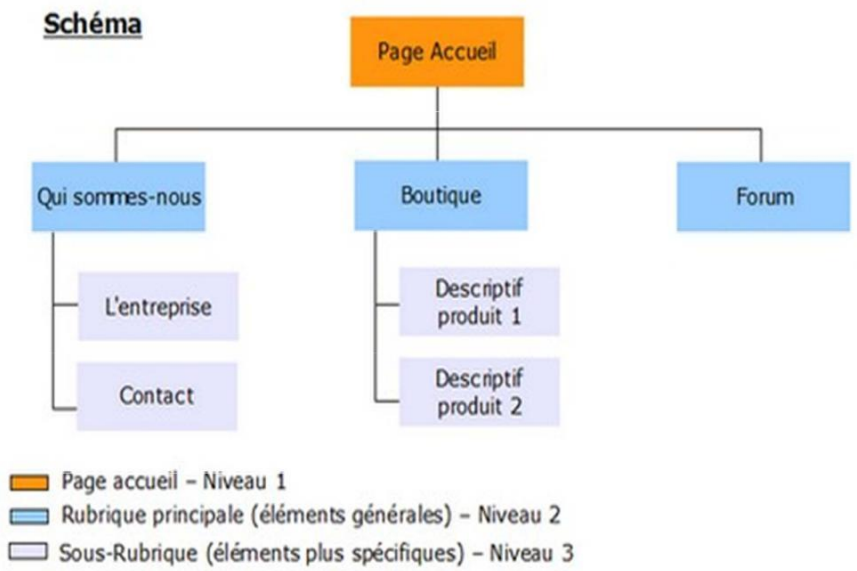

Figure 13: *structure d'un site*

Chaque nœud représente l'adresse d'une page particulière, et les segments reliant ces nœuds indiquent la présence d'un lien hypertexte amenantaux sous-branches immédiates de l'arbre.

c)Connaissances sur les utilisateurs du site :

- Les connaissances sur les utilisateurs d'un site sont généralement obtenues directement auprès des utilisateurs eux-mêmes (âge, sexe, ancienneté urs le Web).
- Dans le cas des sites à base d'inscription, ces connaissances sont recueillies directement à partir du login et du profil-utilisateur donnés parl'internaute au moment de l'inscription.

2.3 Le processus du web Usage Mining :

Web Usage Mining est le processus d'application de techniques de fouille de données à la découverte de modèles d'utilisation à partir de donées Web,

Les trois phases de la fouille de l'utilisation du Web sont les suivantes :

2.3.1 Collecte de données :

La première phase consiste à collecter les données enregistrées au niveau du serveur et les données enregistrées au niveau du client, ainsi queles données enregistrées au niveau du serveur Proxy à analyser. **[Djerbaoui Imad-Eddine Herrouz Hichem 2015]**

- *Données enregistrées au niveau du serveur*: Chaque demande d'affichage d'une page Web de la part d'un utilisateur, peut générer plusieurs requêtes. Des informations sur ces requêtes stockées dans les fichiers Log du serveur Web.
- *Données enregistrées au niveau du client*: Les données sont collectées au niveau du poste client par le biais d'agents incorporés dans les pages Web (sous forme d'appliquettes Java ou en JavaScript) et utilisés pour une collecte directe des informations à partir du poste client. Les informations collectées incluent : le temps d'accès et d'abandon du site, l'historique de navigation, etc.
- *Données enregistrées au niveau du Proxy*: Le serveur Proxy joue le rôle d'intermédiaire entre des clients Web et des serveurs Web. En effet, pour toute requête émise sur une page, le Proxy, après consultation de son disque local, transmet la requête au serveur Web si le document n'est pas disponible à son niveau. Une fois l'information retournée par le serveur, le

Proxy en effectue une copie locale sur son disque puis la transmet à l'initiateur de la requête. Le serveur Proxy garde la trace de toutes les communications établies, dans des fichiers Logs. Ces traces peuvent révéler les requêtes HTTP émisespar plusieurs clients vers plusieurs serveurs Web et servir ainsi de source de données.

2.3.2 Le prétraitement des données (ie. des fichiers Logs) :

Le prétraitement de fichiers logs a comme objectif la structuration et l'amélioration de la qualité des données contenues dans ces fichiers en vue de les préparer à une analyse des usages. Le prétraitement des fichier logs se base sur les étapes suivantes :

a) **Le nettoyage d'un fichier LOG :** consiste à supprimer les :
   - Requêtes échouées et corrompues.
   - Requêtes d'objets multimédias : Les images incluesdans les fichiers HTML sont supprimées car elles ne reflètent pas le comportement de l'internaute. Cependant, ce n'est pas toujours possible d'identifier toutes les images inintéressantes quand le site est volumineux.
   - Requête à l'origine des robots.
   - Les scripts : Le téléchargement d'une page demandéepar un utilisateur est accompagné par le téléchargement automatique des scripts

b) **Transformation d'un fichier LOG :** Cette phase consiste à identifier les utilisateurs, les sessions ainsi que les visites des pages web.

Identification des utilisateurs et des sessions :

Pour identifier les utilisateurs, on se base sur :

- L'adresse IP
- Les cookies (*Client Side Storage*)
- Le mot de passe.

Les identifiants de session permettent à un site en tièrement dynamique d'identifier les internautes individuellement. Ils reposent sur la technologie PHP. Une fois l'utilisateur identifié par l'une de méthodes décrites ci-dessus (à partir de l'adresse ip ,cookies et mot de passe) ,il est possible de reconstituer sa session en regroupant les requêtescontenues dans les fichiers Log et émises par cet utilisateur.

Identification des visites :

Une visite est composée d'une suite de requêtes séquentiellement ordonnées, effectuées pendant la même session et ne présentant pas de rupture deséquence de plus de 30 minutes

L'identification des visites, est effectuée comme suit :

- Déterminer la durée de consultation des pages : Ladurée de consultation d'une page est le temps séparant deux requêtes successives.
- Une fois les visites identifiées, la durée de consultation de la dernière page de chaque visite est obtenue à partir de la moyenne des temps de consultation des pages précédentes appartenant à la même visite.

.

2.3.3 Analyse des résultats :

L'analyse des modèles est la dernière étape dans leprocessus du WUM. Elle consiste à voir comment exploiter toutes les informations qui ont été obtenues. Pour cela plusieurs techniques ont été développées, dont :

- La visualisation des données.
- Un mécanisme de requêtes de bases des données relationnelles. permet de spécifier les conditions à remplir par les données et restreindre l'analyse sur une partie de la base vérifiant certaines conditions.
- Les opérations OLAP (On Line AnalyticalProcessing).
- Les systèmes multi-agents. agents intelligents sont susceptibles d'intervenir aux différentes étapes du traitement de l'information

3. Conclusion

Le web usage mining est la technique pour trouver des informations utiles et intéressantes à partir des données du Web. L'utilisation des données du web, incluent les fichiers logs provenant des journaux(logs) du serveur Web, des journaux du serveur proxy, des journaux du navigateur.

Dans ce chapitre nous avons défini certaines notions relatives au Web Mining et plus particulièrement, au Web Usage Mining ainsi que certaines notions relatives à la structure et au pré-traitement d'un fichier LOG .

Chapitre 3 : Analyse et Conception

1. Introduction :

La CNR souhaite par le biais de ce projet, palier à un manque en matière de décisionnel. Ce manque se caractérise par l'indisponibilité de moyens efficaces en mesure de fournir des informations utiles en temps voulu.

Nous allons à travers ce chapitre présenter une étude de l'existant et une définition des besoins. Une fois les besoins définis, nous présenterons lamodélisation de la solution et la conception de l'entrepôt de données.

2. Etude de l'existant

2.1. Présentation de l'organisme d'accueil:

La Caisse Nationale des Retraites (CNR) a été créépar décret n°:85-223 du 20 août 1985 abrogé et remplacé par le décret N°: 92-07 du 04 janvier 1992 portant statut juridique des Caisses de Sécurité Sociale et organisation administrative etfinancière de la Sécurité Sociale.

La CNR mise en place en 1985 , est chargée de la gestion des différents régimes de retraite existants avant l'institution en 1983 d'un régime national unique de retraite, offrant les mêmes avantages à tous les travailleurs quel que soit leu r secteur d'activité. La CNR est le résultat de la fusion de huit caisses :

1. La Caisse algérienne d'assurance vieillesse**(CAAV):**chargée de la gestion des pensionnés du régime général.
2. Les Centres de gestion des retraites **(CGR):**chargée de la gestion des pensionnés fonctionnaires.
3. La caisse nationale de mutualité agricole **(CNMA) :** chargée de la gestion des pensionnés du régime agricole.
4. La caisse de sécurité sociale des mineurs **CSSM):**(chargée de la gestion des pensionnés du secteur des mines.
5. La caisse d'assurance vieillesse des non-salariés**(CAVNOS) :** chargée de la gestion des pensionnés non-salariés.
6. L'Entreprise de couverture des travailleurs sociaux chargés de la mer et de la fondation d'accorder la retraite aux travailleurs mer **(EPSGM):** chargé de la gestion des pensionnés gens de mer.
7. La caisse d'assurance et de prévoyance des agents de SONELGAZ **(CAPAS):** chargée de la gestion des pensionnés de la SONELGAZ.
8. La Caisse de Retraite des personnels de la SNTF.

2.2 **Structure et Fonctionnement de la CNR**

Les organes essentiels chargés d'assurer le fonctionnement de la caisse sont le conseil d'administration, le Directeur Général et le siègede la caisse :

1. Le Conseil d'Administration : Il administre, contrô le et anime la Caisse. Il est composé de 29 membres répartis comme suit :
   - 18 représentants des travailleurs par les organisations syndicales les plus représentatives
   - 9 représentants des employeurs dont 2 représentantsde la fonction publique,
   - 2 représentants du personnel de la Caisse.
2. Le Directeur Général : Il dirige la Caisse et assure son fonctionnement sous le contrôle du conseil d'administration.
3. Le Siège de la Caisse : est chargé notamment de:
   - organiser, de planifier, de coordonner et de contrô ler : Les activités des agences de wilaya et d'antennes d'administration ou d'entreprise, La gestion des équipements et des moyens humains et matériels de la caisse ; De gérer le budget de la caisse, de coordonner les opérations financières et de centraliser la comptabilité générale ;
   - coordonner le recouvrement des cotisations de retraite ;
   - gérer et de reconstituer les carrières des assurésociaux ;
   - organiser l'information des assurés sociaux et des employeurs ;
   - De suivre l'application des conventions et accords en matière de retraite.

Sous l'autorité du Directeur Général, assisté d'undirecteur général adjoint, le siège de la Caisse comprend :

- La direction des retraites ;
- La direction de la gestion des carrières des assuré sociaux ;
- La direction des finances ou l'agent chargé des opérations financières ;
- La direction de l'informatique et de l'organisation ;
- La direction de l'administration générale ;
- L'inspection générale.

Le directeur général est, en outre, assisté de conseillers et d'assistants pour la prise en charge de dossiers particuliers, et de travaux d'étude, de recherche et d'analyse dictés par la conjoncture. La figure 14 détaille l'organigramme du siège de la caisse.

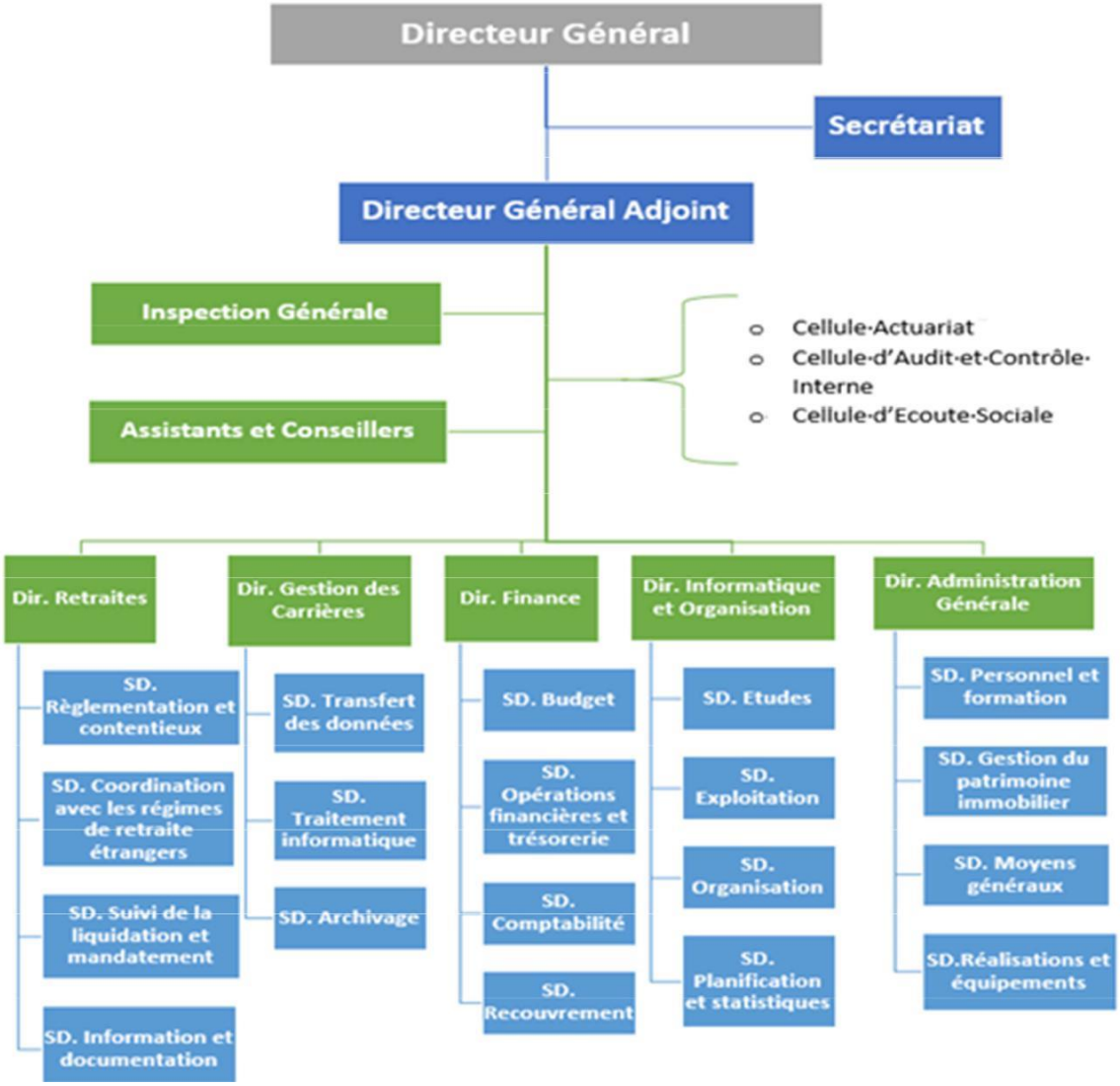

Figure 14: *organigramme du siège CNR*

La CNR est constituée d'une direction générale et ed51 agences au niveau national. Les systèmes opérationnels actuellement utilisés dans la CNR sont décentralisés. Chaque agence manipule les données concernant les retraités qui lui appartiennent uniquement.

Les agences locales de wilayas représentent le noyau de l'activité de gestion de la retraite, elles sont chargées d'assurer le service des pensions deretraite en recevant et en traitant les dossiers des pensionnés, d'effectuer les opérations liées à laeconstitutionr de carrière des assurés sociaux, de enirt la comptabilité, assurer l'exécution des opérationsfinancières et leur coordination, d'assurer la gestion des moyens matériels et humains de l'agence et de veiller à l'écoute et à l'orientation des pensionnés grâce aux cellules d'écoute sociales. Lafigure 10 détaille l'organigramme d'une agence CNR.

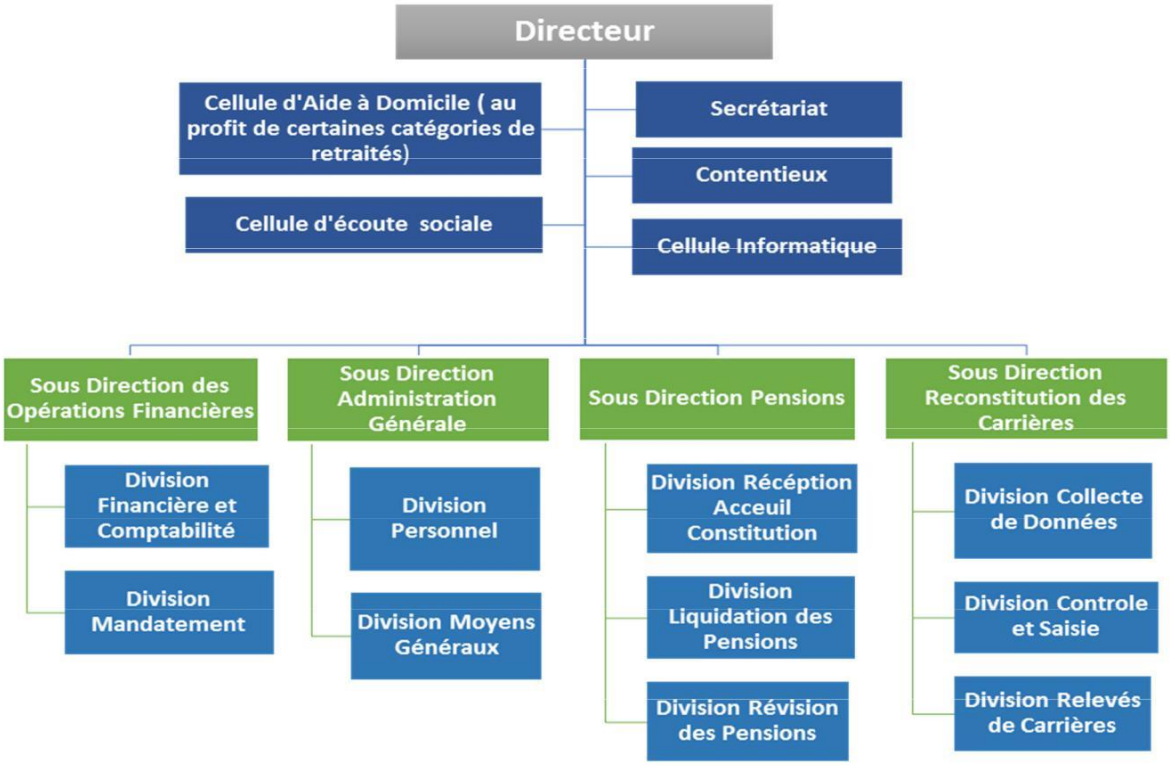

Figure 15: *Organigramme d'une agence CNR*

Cadre du travail :

Nous avons effectué notre stage dans la direction générale de la CNR, au niveau de la direction informatique et organisation, et plus précisément ua sein de la sous-direction organisation, et avons été suivies par Monsieur R.Bellemou, sous-directeurde l'organisation.
La sous-direction de l'organisation est composée de 3 principales structures comme le montre la Figure 16.

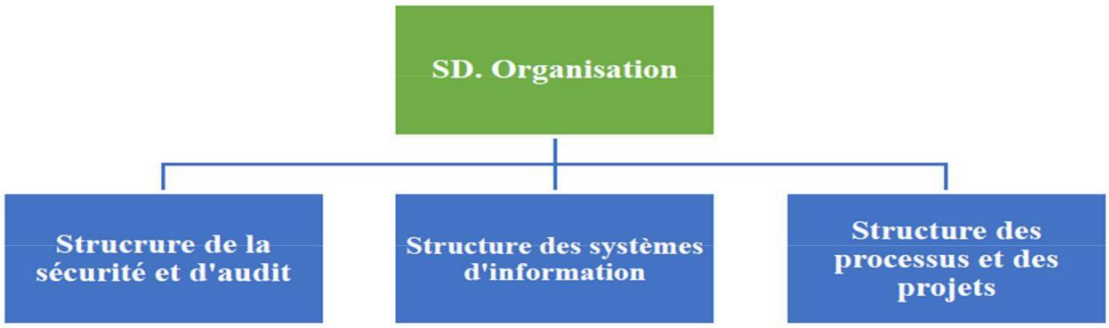

Figure 16: *Organigramme de la sous-direction de l'organisation*

Elle est chargée de concevoir des méthodes d'organisation dans le but d'homogénéiser les procédures et les imprimer et de les mettre en œuvr e, d'élaborer le plan informatique ainsi que le schéma directeur de l'informatisation de la Caisse, elle est également chargée de concevoir et de réaliser les logiciels dont ont besoin les services de la Caisse, d'organiser l'activité des centres de traitement informatique et veiller à leur fonctionn ement selon les normes préalablement définies et

d'assister l'ensemble des utilisateurs de la Caisse dans l'utilisation des logiciels et matériels informatiques.

2.3. Spécification des besoins :

De nos discussions avec le sous-directeur de l'organisation, il est ressorti que la CNR ne possède pas d'outils d'aide à la décision lui permettant d' optimiser ses ressources et ses coûts d'entreprise. Il y a un manque flagrant au niveau de son système décisionnel. En effet, tout processus d'analyse et de prise de décision à tous les niveaux se base essentiellement sur des rapports d'activités et statistiques, dont les données sont extraites et consolidées à partir des systèmes transactionnels. Or la CNR ne possède pas d'outils de *reporting*, ni d'outils d'analyse de ses données d'activité. D'où la nécessité pour l'entreprise, de mettre en place une solution décisionnelle à même de fournir des informations utiles en temps voulu aidant ainsi à la prise de décisions opportunes, en particulier en vue d'optimiser ses ressources matérielles et ses coûts de gestion.

Les indicateurs qui intéressent la CNR dans la prise de décision dans un contexte d'optimisation de ses ressources et ses coûts sont basés sur :

- Le calcul des fréquences des applications pour en déduire leurs importances.
- La définition les périodes d'exploitation des ressources afin de les optimiser.
- Le classement des erreurs par type et l'identification des erreurs courantes en vue de les corriger.
- La détecter des applications qui affichent le plusd'erreurs afin d'y remédier.

Nous proposons dans ce projet, la mise en place d'une solution décisionnelle basée sur le*web usage mining*. L'idée est d'extraire et d'analyser les informations des fichiers logs issus du serveur web de la CNR, dans l'objectif de proposer des rapports d'activités détaillés aux décideurs. La solution sera mise en place d'un système d'information décisionnel avec un datawarehouse dont les sources de données de notre entrepôt sont les fichiers LOGs (access_log et error_log) du serveur web de l'entreprise.au niveau du service informatique qui détient l'ensemble des informations de connexion au niveau de la CNR qui permettra d'afficher les applications les plus fréquentes, leurs périodes d'exploitation ainsi que les erreurs les plus fréquentes de ces application afin d'optimiser les ressources et de corriger les erreurs les plus courantes

2.4 Conditions d'exploitation informatique de la CNR :

En vue de mettre en place un projet décisionnel opportun, une bonne compréhension de l'environnement informatique de la CNR est nécessaire. Il est essentiel de disposer d'informations précises sur l'infrastructure des ressources matérielles (mémoire, processeurs…) et les problèmes qui ont une incidence sur les coûts. En effet, ces informations affectent une grande partie des décisions que nous allons prendre dans le choix de la solution et de son déploiement.

2.4.1 Analyse du parc informatique :

La direction générale de la CNR dispose d'un parc nformatique riche et varié composé de :

- 200 postes de travail : Il s'agit principalement de micro-ordinateurs de type intel Core i3 ou i7, dotés de mémoires de 1 à 8 Giga Octets, et fonctionnant sous WindowsXP ou Windows 10.
- une dizaine de serveurs, avec machines virtuelles, utilisant les systèmes Linux ou Windows Server sous différentes versions : Windows Server 2000, 2012 et 2016.

- Un serveur web apache/2.2.15 sous environnement CentOS release 6.8 (final), doté d'une mémoire de 8 Giga Octets, un disque dur de 100 GigaOctets, et doté d'un processeur de 64 bits.

La CNR assure également des communications internes (au niveau de la direction) et externes (avec les différentes agences et centres de calcul)en utilisant :

- Un réseau interne dédié au transfert de fichiers l'intérieurà du siège, utilisant le protocole FTP.
- Un réseau externe dédié au transfert de fichiers entre le siège, les agences et les centres de calcul à l'aide d'un réseau virtuel privé VPN utilisant le protocole FTP. Ce réseau n'est exploité que par quelques agences ou centres de calcul.
- Une messagerie permettant la communication interne et externe entre les employés de la CNR au niveau du siège, des agences et des centres de calcul.

2.4.2 Les sources de données :

La CNR utilise différentes applications pour assure sa gestion. Ces applications et le serveur web sont utilisés par plusieurs profils d'employés dontles administrateurs applicatifs, les demandeurs, les personnes jouant le rôle de traiteurs des requêtes de demandeurs avec réponses à la requête, et les collecteurs de données pour rapports statistiques.

Les collecteurs de données ont pour mission d'extraire les données de l'entreprise à partir de ses différentes sources, de les analyser, afin de construire des rapports statistiques utiles en matière de décision de l'entreprise. Les données utilisées sont généralement extraites des fichiers logs des différents serveurs de l'entreprise. Ces fichiers logs gardent trace des requêtes reçues et des opérations effectuées par les serveurs.

Pour notre solution informatique nous avons exploité deux types de fichiers logs :

1) Access_log : contient les informations suivantes, tel que le montre l'extrait de la figure 10:
   - Adresse IP de l'utilisateur.
   - La date et l'heure du lancement de la requête.
   - Le type d'opération.
   - Le nom de l'application ainsi que son url.
   - Le navigateur

```
10.10.50.39 - - [08/Apr/2018:07:16:28 +0100] "GET
/doleance/login.php HTTP/1.1" 200 4529 "-" "Mozilla/4.0
(compatible; MSIE 999.1; Unknown)"
10.10.50.39 - - [08/Apr/2018:07:16:30 +0100] "POST
/doleance/controller/autorefresh.php HTTP/1.1" 200 - ""
"Mozilla/4.0 (compatible; MSIE 999.1; Unknown)"
10.10.50.39 - - [08/Apr/2018:07:16:33 +0100] "POST
/doleance/controller/login.php HTTP/1.1" 200 4 "" "Mozilla/4.0
(compatible; MSIE 999.1; Unknown)"
10.10.50.39 - - [08/Apr/2018:07:16:33 +0100] "GET
/doleance/index.php HTTP/1.1" 200 14447 "" "Mozilla/4.0
(compatible; MSIE 999.1; Unknown)"
```

Figure 17: *access_log*

2) Error_log : contient des informations sur les erreurs générées par les applications, dont extrait présenté en figure11 suivante :

```
[Sun Apr 08 03:10:01 2018] [notice] Digest: generating secret
for digest authentication ...
[Sun Apr 08 03:10:01 2018] [notice] Digest: done
PHP Warning:  PHP Startup: Unable to load dynamic library
'/usr/lib64/php/modules/mssql.so' -
/usr/lib64/php/modules/mssql.so: cannot open shared object
file: No such file or directory in Unknown on line 0
PHP Warning:  Module 'PDO' already loaded in Unknown on line 0
[Sun Apr 08 03:10:01 2018] [warn] RSA server certificate
CommonName (CN) `SUPPORT' does NOT match server name!?
[Sun Apr 08 03:10:01 2018] [notice] Apache/2.2.15 (Unix) DAV/2
PHP/5.6.28 mod_ssl/2.2.15 OpenSSL/1.0.1e-fips configured --
resuming normal operations
```

Figure 18: *error_log*

La solution préconisée est la création d'un*Datawarehouse* qui pourra stocker le nombre d'application et le nombre d'erreurs par type et pa r application et ainsi répondre aux besoins cités plus haut.

3. Conception du DATA WAREHOUSE:

Nous allons présenter dans cette partie les différentes étapes de la conception de la solution dont :

La conception de la zone d'entreposage qui représente la structure du datawarehouse.
La conception de la zone d'alimentation du datawarehouse qui représente l'entrepôt avec ses données nettoyées et structurées.
La création des cubes OLAP
Sélection de la démarche de réalisation des rapports.

Avant de présenter les étapes de conception de notre *Datawarehouse* nous allons d'abord présenter un diagramme de cas d'utilisation du système :

3.1 Diagrammes des cas d'utilisation :

Un cas d'utilisation (*use case*) modélise une interaction entre le système informatique à développer et un utilisateur ou acteur interagissan avec le système. Plus précisément, un cas d'utilisation décrit une séquence d'actions réalisées par le système qui produit un résultat observabl pour un acteur.

La Figure 19 représente le formalisme utilisé pource diagramme :

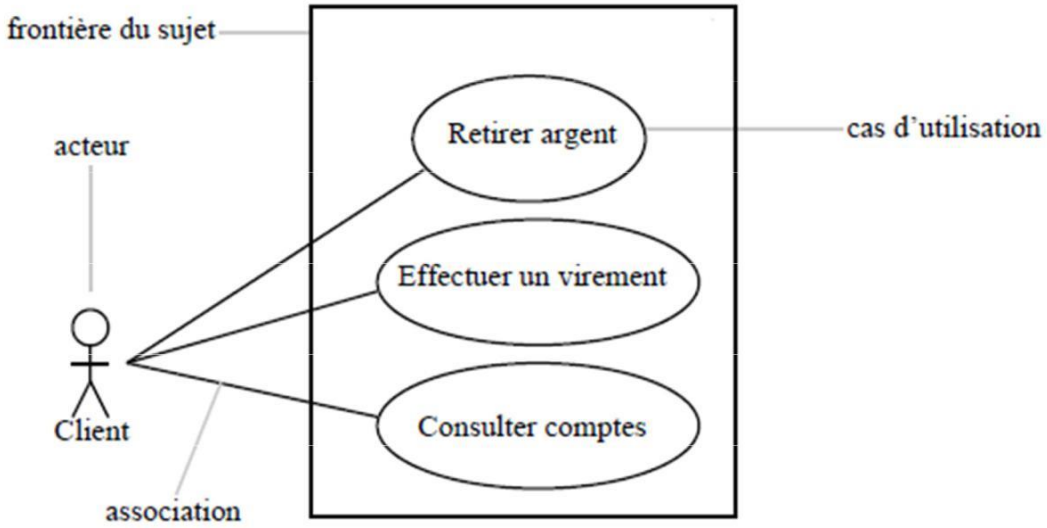

Figure 19: *Formalisme d'un diagramme de cas d'utilisation.*

Un acteur est une entité extérieure au système, quiinteragit directement avec ce dernier.

1. L'informaticien: C'est lui qui possède tous les privilèges et un accès total au système. Il s'occupe de la mise en place du *Datawarehouse* en suivant les différentes étapes dont la création de l'ETL, des cubes OLAP e des rapports ,il choisit les données à afficher et peut consulter les rapports créés
2. Le décideur : cet acteur peut consulter les rapports créés par l'informaticien.

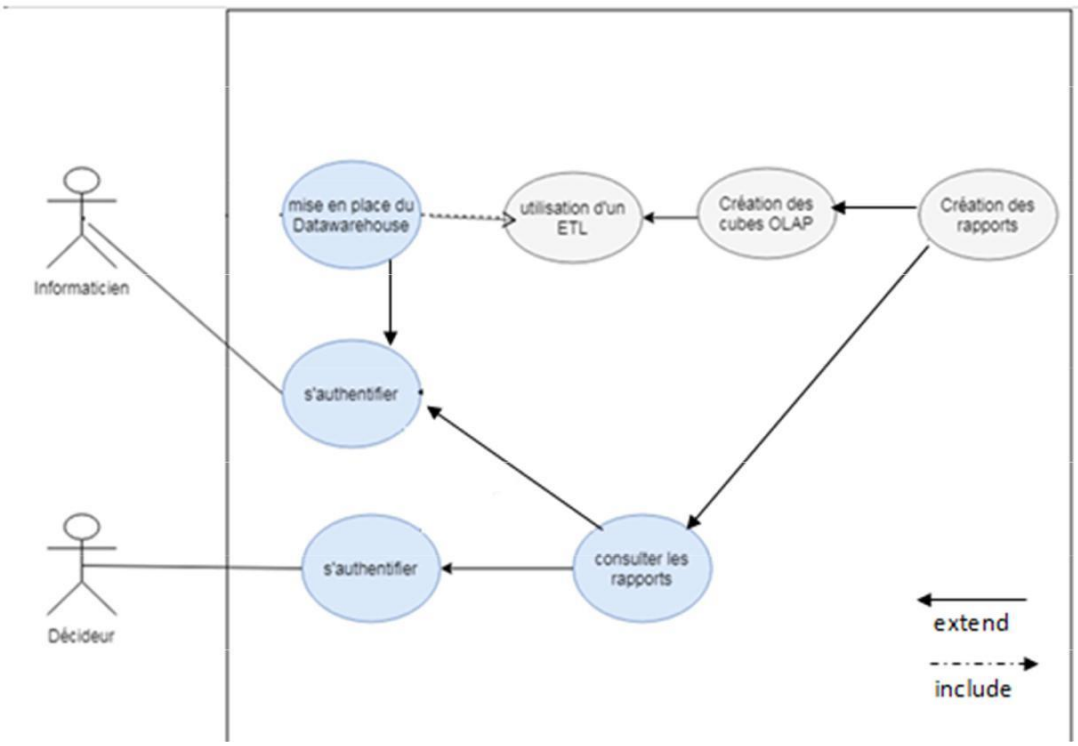

Figure 20: *diagramme de cas d'utilisation*

- Le cas d'utilisation « création du *Datawarehouse* » : C'est à lui que revient la tâche de la création du *Datawarehouse* en suivant les diférentes étapes dont la création edl'ETL des cubes multidimensionnels et des rapports, il peut aussi se connecter à la base de données d'interroger.
- Le cas d'utilisation «Naviguer entre les statistiques» : L'informaticien peut naviguer aisément entre les différentes statistiques et choisir les données à afficher.
- Le cas d'utilisation « s'authentifier » : *l'informaticien :* après son authentification il a accès à tous les privilèges
  Le décideur : après son authentification a des privilèges limitésqui sont définis par l'informaticien.

- Le cas d'utilisation «Consulter le rapport » :L'informaticien peut visualiser le rapport qui rassemble les indicateurs de performance de l'entreprise
- en ce qui concerne le décideur , il fait seulement la consultation des rapports**.**

3.3 Choix de la structure de Base de données :

Dans le cadre de notre travail, nous avions le choix entre deux types de zones de stockage : une base de données relationnelle ou une base de données multidimensionnelle (un datawarehouse). Nous avons choisi d'utiliser une structure multidimensionnelle parce qu'un *Datawarehouse* est une collection de données orientées sujet :

- Intégrées : les différentes données concernant lesapplications dans les fichiers logs nécessitent le regroupement de ces dernières dans un seul *Datawarehouse*.
- Non volatiles : les données que doit contenir notre système doivent être disponibles l'ensemble des utilisateurs, donc on aura des chargements et des consultations. Une même requête, sur la même période, exécutée plusieurs ,fodonnera les mêmes résultats.
- Historisées : le besoin d'analyse requiert l'historisation des données.
- Organisées pour la prise de décision : notre projetest un véritable système d'aide à la décision.

3.3 Choix de l'Architecture du *Datawarehouse* :

Pour la création de l'entrepôt de données, nous avons choisi une architecture centralisée, un modèle utilisé pour les rapports et les analyses spécifiques à un secteur d'activité. Les données sont agrégées et regroupées dans le but de répondre auxbesoins d'un seul type d'utilisateurs, nous n'avons donc pas besoin de les diviser ou les classer par thèmes d'utilisation, il n'est donc pas nécessaire de diviser l'entrepôt en sous-ensembles, d'où le choix de cette architecture.

3.4 Conception de la zone d'entreposage :

D'après Kimball, le schéma en étoile du*Datawarehouse* est construit en analysant les besoins identifiés à partir des processus organisationnels, selon ces quatre étapes :

- Choix de processus d'activité à modéliser : Un processus d'activité est un processus opérationnel important pour l'organisation.
- Choix du grain du processus d'activités : C'est le niveau atomique des données figurant dans la table des faits pour ce processus.

- Choisir les dimensions appliquées pour chaque tablede faits, ou chaque dimension représente un axe d'analyse, Cela revient à répondre à la ques tion suivante : comment décrire les données résultantes du processus modélisé ?
- Identifier les faits pour chaque table de faits : Ce la revient à répondre à la question suivante : qu'allons-nous mesurer ?

En utilisant les besoins collectés et mentionnés dans l'étude des besoins**,** nous pouvons concevoir les schémas en étoile constituant notre*Datawarehouse*. Ces derniers sont interconnectés entre eux grâce à l'ensemble des tables de dimensions commune s entre les tables de fait.

Le Formalisme de représentation de bases de donnéesutilisé :

Avant de faire la conception dimensionnelle des activités, nous allons décrire le formalisme à utiliser. Les schémas suivants illustrent ce formalisme.

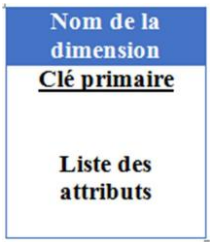

Figure 21: *Formalisme adopté pour la table de dimension*

Figure 22: *Formalisme adopté pour la table de faits*

Figure 23: *Formalisme adopté pour la relation entre les deuxtables*

3.4.1Choix des activités(faits) :

Selon les informations issues des fichiers logs (access_log et error_log) et les besoins de l'entreprise qui ont étaient définis en collaboration avec le décideur, nous avons choisi une activité pour chaque fichier log :

- Suivi_application pour le fichier access_log .
- Suivi_erreur pour le fichier erreur_log

3.4.1.1 suivi_application :

a) Présentation de l'activité suivi_application :

Le serveur web de la CNR contient plusieurs applications utilisées par le personnel, toutes les informations d'accès sont contenues dans le fichier access_log. Afin d'optimiser ses ressources déduire l'importance de chacune en calculant le nombre d'accès aux applications.

b) Le grain :

- Chaque ligne de la table de faits renseigne le nombre d'accès et le nom des applications par date, période de son exécution et par adresse_ip des utilisateurs.
-

c) Présentation des dimensions :

1. La dimension application :

Cette dimension rassemble les informations des applications utilisées par les employés de l'entreprise.

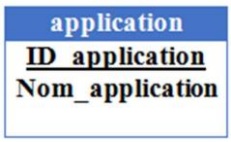

Figure 24: *la dimension « application »*

Dimension :« application »		
Attribut	**Désignation**	**Exemple**
ID_app	Clé de substitution	1
Nom_application	Nom de l'application	Doléance

Tableau 3: *la dimension « application »*

2. La dimension date :

Cette dimension représente la date d'utilisation de l'application.

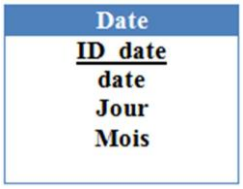

Figure 25: *la dimension « date »*

Dimension : « date »		
Attribut	**Désignation**	**Exemple**
ID_date	Clé de substitution	5698
date	La date entière en précisant le jour, le mois, et l'année.	2018-01-22
Mois	Désignation du mois de l'année	Janvier
jour	Désignation du jour de la semaine	Lundi

Tableau 4: *la dimension « date »*

3. La dimension période :

Cette dimension désigne la période d'utilisation del'application.

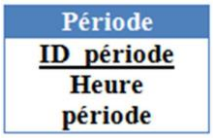

Figure 26: *la dimension « Période »*

Dimension :« période»		
Attribut	**Désignation**	**Exemple**
ID_période	Clé de substitution	15478
Heure	L'heure entière en précisant l'heure, la minute, la seconde.	08 :12 :56
Période	Les différentes périodes d'accès aux applications	Heure_travail

Tableau 5: *la dimension « Période »*

La période est définie comme suit :

Heure travail : entre 8h et 12h et entre 13h et 16h
Pause : entre 12h et 13h
Hors travail : entre 17h et 8h

4. La dimension adresse_ip :

Cette dimension désigne l'adresse ip de l'utilisateur ainsi que le type de l'adresse.

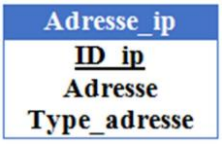

Figure 27: *la dimension «adresse_ip»*

Dimension :«adresse_ip»		
Attribut	**Désignation**	**Exemple**
ID_ip	Clé de substitution	454
Adresse	l'adresse ip de l'utilisateur	10.10.50.39
Type_adresse	Type de l'adresseip	externe

Tableau 6: *la dimension «adresse_ip»*

c) Schéma en étoile :

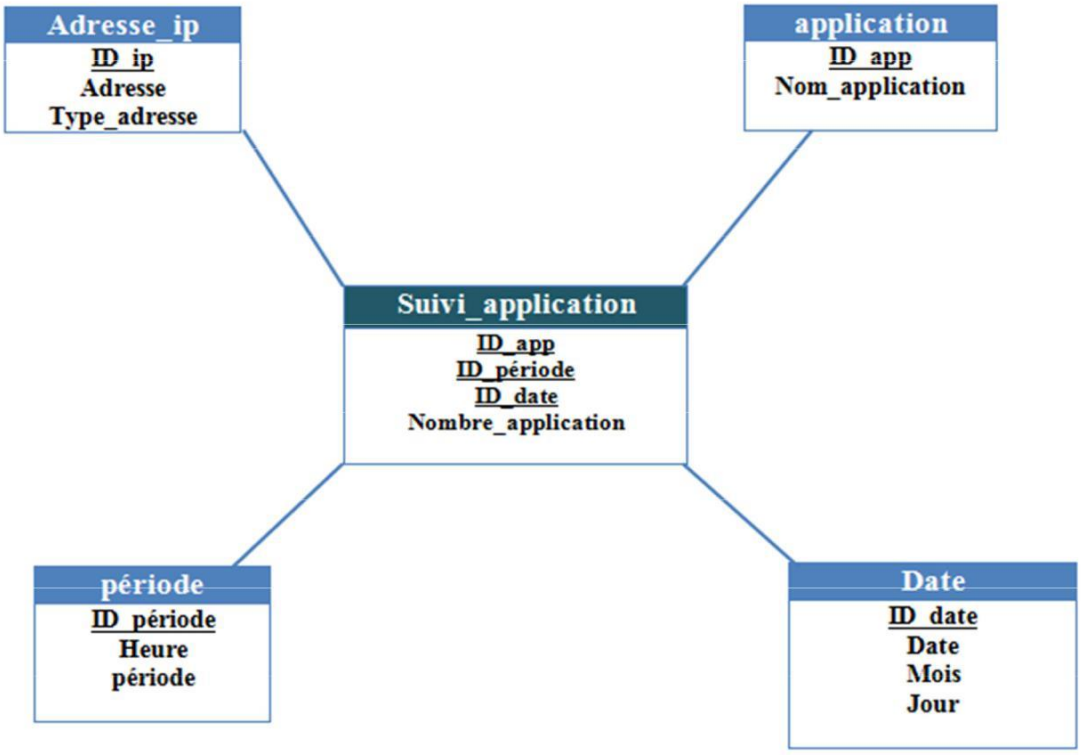

Figure 28: *schéma en étoile « suivi application »*

Suivi_application	
Attribut	**Désignation**
ID_app	Clé étrangère de la dimension application
ID_date	Clé étrangère de la dimension date
ID_période	Clé étrangère de la dimension période
ID_ip	Clé étrangère de la dimension Adresse_ip
Nombre_application	Le nombre de fois ou l'application a été exécutée

Tableau 7: *table des faits « Suivi_application»*

3.4.1.2 Suivi_erreur :

a) présentation de l'activité suivi_erreur:

pour bien gérer un serveur web, il est nécessaire ed disposer d'un retour d'informations a propos de l'activité et des performances. Le fichier error_log contient toutes les informations de diagnostic et toutes les erreurs qui surviennent lors du traitement des requêtes

b) Le grain d'activité

- Chaque ligne de la table de faits renseigne le type d'erreur, ainsi que son contenu par application.

c) Présentation des dimensions :

1. La dimension erreur :

Cette dimension regroupe toutes les informations relatives aux erreurs.

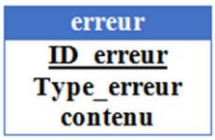

Figure 29: *la dimension «erreur»*

Dimension : « erreur »		
Attribut	**Désignation**	**Exemple**
ID_erreur	Clé de substitution	454
contenu	Contenu de l'erreur	Undefined index first_name in /var/www/html/stat-j/menus.php on line 62

Tableau 8: *la dimension «erreur»*

2. La dimension type_erreur :

Cette dimension représente les types d'erreurs.

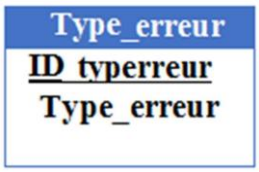

Figure 30 : *la dimension type_erreur*

Dimension : « type_erreur »		
Attribut	**Désignation**	**Exemple**
ID_typerreur	Clé de substitution	454
Type_erreur	Type de l'erreur	PHP Warning

Tableau 9: *la dimension "type_erreur"*

d) Schéma en étoile :

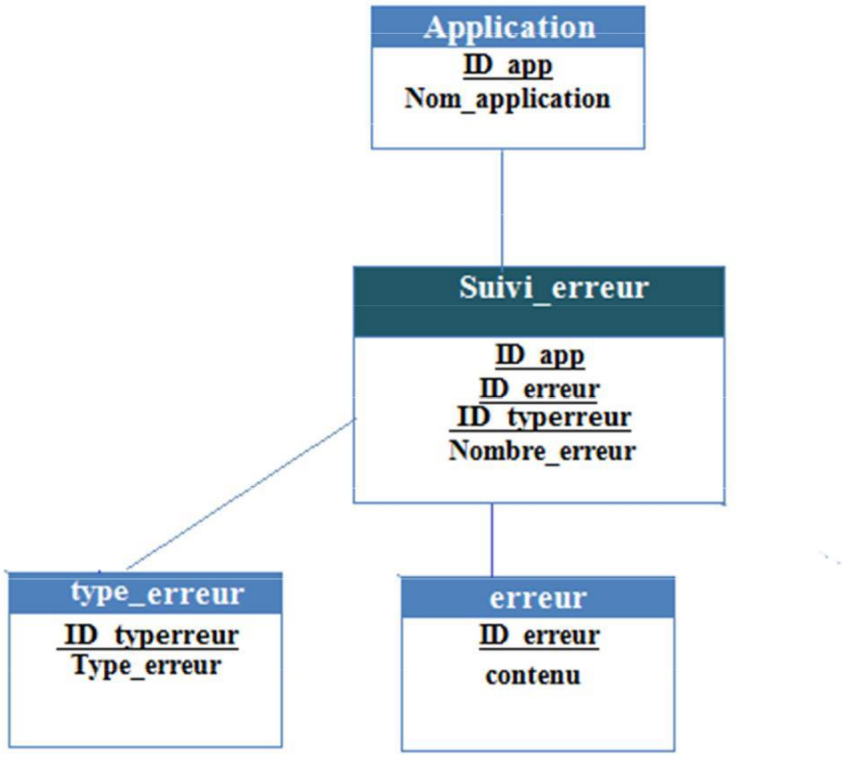

Figure 31: *schéma en étoile « suivi_erreur »*

Suivi_erreur	
Attribut	**Désignation**
ID_erreur	Clé étrangère de la dimension erreur
ID_app	Clé étrangère de la dimension application
Nombre_erreur	Le nombre de fois ou l'erreur est apparue

Tableau 10: *table des faits « Suivi_erreur»*

3.5 Conception de la zone d'alimentation du Datawarehouse :

L'alimentation est la procédure qui permet de transférer des données du système opérationnel vers l'entrepôt de données en les adaptant. La conception de cette opération est une tâche assez complexe (elle constitue 70% d'un projet décisionnel en moyenne). Il est nécessaire de déterminer quelles données seront chargées, quelles transformations et vérifications seront nécessaires, la périodicité et le moment auxquels les transferts auront lieu.

3.5.1 Identification des sources de données

Avant de commencer la conception de l'ETL, il faut bien répondre aux questions : comment sont mes sources ? Et quelles données de production faut-il sélectionner pour alimenter le*Datawarehouse* ?

Les sources de données de notre entrepôt sont les fichiers LOGs (access_log et error_log) du serveur web de l'entreprise.

Afin de déterminer l'emplacement des données à charger dans l'entrepôt, nous avons :

- Listé l'ensemble des informations dont nous avons besoin à partir des modèles en étoile.
- Déterminé leurs emplacements sur les fichiers LOGsUne. fois la source définie, le processus d'alimentation peut commencer.

3.5.2 Processus d'alimentation :

Ce processus passe par trois étapes principales : Extraction, transformation et chargement (ETL).

- L'extraction des données.
- La transformation.
- Le chargement, pour alimenter le Data Warehouse.

3.5.2.1 Extraction des données

L'extraction des données à partir du système source est la première étape du processus ETL. Cette opération peut commencer une fois le plan global depréparation de données établi et les sources de données identifiées précisément.

Dans un premier temps, toutes les données source du système, identifiées comme étant pertinentes, sont extraites et injectées dans la zone de préparation de données *staging*(*area*); c'est là que seront opérés les différents traitements de cesdonnées avant leur chargement. Il s'agit de la première extraction sur le système source; elle concerne par conséquent les données les plus récentes disponibles sur ce système.

Cette technique d'extraction correspond à l'étape de mise en route du DataWarehouse.

3.5.2.2 Préparation du chargement : traitements des données

Les données extraites à partir du système source sont stockées dans la zone préparation de données en vue d'être nettoyées puis chargées dansles dimensions et les tables de faits.

L'objectif de la *staging area* est l'obtention de données prêtes à être chargéesdans une structure de *DataWarehouse*. Ainsi, toutes les étapes du traitement des données se font au sein de l'environnement relationnel de la *staging area*. Une fois la provenance des données précisément établie, il est nécessaire de définir les opérations de transformation des données.

- **Opérateurs de type transformation**

Ces opérateurs permettent de transformer les données des objets sources avant leur chargement dans les objets cibles. Les opérateurs de transformation les plus utilisés sont :

- Filtre: Permet de définir une condition «*Where*» sur les données.
- Convertisseurs: Permet de convertir les types des données.
- Dé-duplicateur: Supprime tous les doublons trouvés.

- Expression personnalisée: Permet à l'utilisateur(informaticien) de définir un code SQL personnalisé pour la transformation.

3.5.2.3 Préparation du chargement des dimensions:1

3.5.2.3.1. Chargement des dimensions

Le chargement des dimensions est une tâche relative ment simple considérant que tous les traitements sur les données ont été faits, il ne stere qu'à les insérer dans les structures représentant les dimensions.

Le chargement des données se fait avec un l'outil de chargement offert par Talend. [https://fr.talend.com/products/talend-open-studio/]

3.5.2.3.2. Préparation du chargement des tables de faits

Chaque enregistrement (ligne) de la table de fait devrait en principe pouvoir être associé à chacune des dimensions du modèle. Pour garantir cela, un nombre de traitements intermédiaires sont nécessaires ; cela va de la constitution de jointures entres les tables du système source à l'ajout de champs calculés.

Le processus de préparation des données représentela partie la plus délicate du projet, car en plus d'extraire les données du système source, il doit assurer le chargement de données correctes et cohérentes.

4 Conception des cubes OLAP :

La création des cubes OLAP consiste à définir un ensemble de vues sur l'entrepôt de données à partir desquelles se fera la sélection des groupes de mesures relatifs au cube en question, ainsi que les dimensions et les hiérarchies à inclure dans le cube qui serviront d'axe d'analyse.

Les tableaux ci-dessous présentent les dimensions, hiérarchies et mesures choisis pour les deux cubes :

a) Cube application :

Tables	dimensions	hiérarchies
Dim_date	Mois	mois
Dim_date	Jour	jour
Application	Application	Nom_application
dim_adresse_ip	dim_adresse_ip	Type_adresse
Dim_période	Dim_période	période

Mesure : nombre d'applications.

b) Cube erreur :

Tables	dimensions	hiérarchies
Erreur	Erreur	Contenu_erreur

type_erreur	type_erreur	Type

Mesure : nombre d'erreurs.

5 Sélection de la démarche de réalisation :

Le système peut être implémenté soit par un développement spécifique ou par le paramétrage d'un outil décisionnel. Le tableau ci-dessous présente une comparaison entre les deux concepts :

Développement spécifique	**Paramétrage**
Mise en œuvre long.	Mise en œuvre rapide
Fonctionnalité sur mesure	Riche en fonctionnalités
Nécessité beaucoup d'implication du client	Nécessité moins d'implication du client
Moins performant	Performant
Coûts supplémentaires de correction et de maintenance	Inexistence de coûts de correction ou de maintenance
Pas de coût de licence ou formation	Coût de licence et formation (sauf pour un outil open source)
Autonomie	Dépendance du fournisseur (sauf pour un outil open source)
Non évolutif	Obtention des mises à jour fonctionnelles automatique gratuitement ou à des coûts maitrisés

Tableau 11: *comparaison entre le développement et le paramétrage d'un outil décisionnel*

Nous optons pour la solution du paramétrage d'un progiciel pour les raisons suivantes :

- Toutes les exigences et les fonctionnalités du notre système sont disponibles au niveau des progiciel décisionnels.
- Rapidité et facilité d'implémentation de la solution.
- Nécessite moins d'implication des utilisateurs.

6 Conclusion :

Dans ce chapitre nous avons détaillé les différentes parties de notre système décisionnel pour la CNR, allant de la conception du *Datawarehouse* à la conception de la zone de restitution de données. Nous avons pu sortir avec un modèle du datawarehouse et un tableau de bord qui couvre au maximum les besoins des décideurs. Et en arrière-plan, un modèle d'extraction, de transformation et de chargement.

L'étape de conception fut en effet une tâche assez complexe qui a nécessité un temps considérable.

Nous présenterons dans le quatrième et dernier chapitre les différentes techniques et technologies que nous avons utilisé dans le but de concrétiser notre conception et d'aboutir au produit final.

Chapitre 4 : Réalisation

Introduction :

Dans ce dernier chapitre, nous allons décrire la mise en place de notre solution, en présentant en détails sa réalisation et son déploiement. Pour cette réalisation, il a été nécessaire de recourir ànu certain nombre d'outils et mettre en place l'enviro nnement d'exécution.

Nous présenterons donc les différents outils utilisés (my sql pour la base de données, talend pour l'etl, pentaho mondiran workbench pour l'implémentation des cubes de données, jaspersSoft pour le reporting) ainsi que la stratégie suivie pour la réalisation de chaque composant du système décisionnel (*Datawarehouse*, ETL, cubes OLAP , reporting).

1. Présentation des outils de développement :

My sql:

Le serveur de base de données MySQL est très rapide, fiable et facile à utiliser. Il dispose aussi de fonctionnalités pratiques, développées en coopération avec ses utilisateurs puisqu'il est Open Source. Le serveur MySQL a été développé à l'origine pour esd grandes bases de données, et a été utilisé avec succès dans des environnements de production très contraints et très exigeant, depuis plusieurs années. Le serveur MySQL offre de nombreuses fonctions puissantes, ses possibilités de connexion, sa rapidité et sa sécurité font du serveur MySQL unserveur hautement adapté au développement des applications.

Talend open studio version 7.0.1:

Talend est une société française créée en 2005, développant une suite de nombreux logiciels Open Source, connue sous le nom de *Talend Open Studio*.

L'objectif de cette suite est de développer, tester, déployer et administrer des projets d'intégration et de gestion de données. *Talend Open Studio for Data Integration* permet de créer un ETL et de visualiser son architecture de façon graphique. La particularité de cet ETL est qu'il génère du code : pour chaque traitement d'intégration de données, uncode spécifique est généré en Java.

Job Designer:

Le *Job Designer* intègre une « *Component Library* »: une palette graphique de composants et connecteurs. Les processus d'intégration sont construits simplement en déposant des composants et

connecteurs sur le diagramme, en dessinant leurs connexions et relations, et en modifiant leurs propriétés. La plupart de ces propriétés peut êtreissue des métadonnées déjà définies.

La *Component Library* inclut plus de 80 composants et connecteurs, fournissant (1) des fonctions basiques telles que des associations, transformations, agrégation et recherches ; et (2) des fonctions spécialisées comme le filtrage de données, le multiplexage de données... Cette librairie supporte les principaux SGBDR, formats de fichiers, annuaires LDAP...

La *Component Library* peut facilement être complétée en utilisant desngagesla standards tels que Perl, Java ou SQL.

Mondrian, Pentaho schema workbench :

Mondrian est un serveur OLAP (On-Line Analytical Processing) écrit en Java par Julian Hyde en 2001. *Mondrian* prend en charge le langage de requêtes MDX (expresions multidimensionnelles) et les spécification XML, sur des entrepôts de donnéess'appuyant sur des SGBDR comme MySQL d'où sa caractérisation de «ROLAP» (*Relational OLAP*).

Mondrian permet d'accéder aux résultats dans un format multidimensionnel (cube de données), car il s'appuie sur une modélisation OLAP standard , et peut donc se connecter à n'importe quel entrepôt de données conçu dans les règles de l'art de la Business Intelligence.

Il est intéressant de noter que*Mondrian* est le composant OLAP utilisé par la plupart des uites de BI Open Source notamment Pentaho, JasperServer et SpagoBI.

Pentaho Schema Workbench est un client riche open source (écrit en Java) qui permet de créer des schémas de cubes sans avoir à connaître la syntaxe XML de *Mondrian*. Cet outil présente une interface avancée qui permet d'effectuer les actions suivantes :

Connexion au Datawarehouse (via JDBC)
Création de schémas et cubes*Mondrian*
Création des mesures, dimensions et hiérarchies d'un
cube Jointures pour les tables floconnées
Tables d'agrégation à utiliser (le cas échéant)
Création de membres calculés, de dimensions partagées, de cubes virtuels Définition des rôles (gestion de la sécurité d'accès à l'intérieur d'un cube) Publication des cubes sur le serveur Pentaho

Jasper Reports Server version 7.1.0:

La suite décisionnelle Jasper inclut *JasperReport Server*, *JasperReports Library*, *iReport Designer*, *Jaspersoft ETL* et *Jaspersoft OLAP* pour constituer un serveur d'applications performant et générer vos rapports.

La version complète de l'application se nomme *Jasper Reports Server* (JRS) depuis la V4 (anciennement JasperServer) et propose un serveur d'application et la création de rapports web.

JasperReports Server est un outil de reporting open source, offert sous forme d'une bibliothèque qui peut être embarquée dans tous types d'applications Java.

2. Réalisation de la solution :

La réalisation de la solution présente plusieurs aspects liés à la mise en œuvre du projet, nous allons décrire dans cette partie, les différentes tapesé de la création du dataWarehouse. Cette partie comprend: la réalisation du dataWarehouse relationnel(zone d'entreposage), du système ETL(zone d'alimentation), et du portail de restitution.

2.1. La réalisation de la zone d'entreposage

Il s'agit de l'implémentation du *Datawarehouse*. Pour l'implémenter physiquement, nous avons créé une base de données relationnelle conforme auxschémas conçus dans la partie conception de la zone d'entreposage à l'aide de mysql. Dans cette BD D, chaque fait du schéma conceptuel correspond à une table appelée table de faits et chaque dimension correspond à une table appelée table de dimension.

La table de dimension contient une clé primaire et un ensemble d'attributs. Tandis que la table de faits possède comme attributs les mesures et les clés étrangères vers les tables de dimension.

2.2. La réalisation de la zone d'alimentation

Une fois la base de données cible prête, nous pouvons procéder au chargement des données. Le chargement de la dimension période se fait à l'aide d'une fonction qui à chaque nouvelle année insère les mois, les trimestres et l'année.

Les autres données sont extraites à partir des bases sources, et ceci après avoir eu recourt aux transformations nécessaires, pour ensuite être chargées dans le Data warehouse.

Comme on a déjà mentionné, nous avons utilisé l'outil *Talend Open Studio*, pour la réalisation des différents programmes ETL. Cet outil présente de grandes capacités de traitement des gros volumes de données.

On vous présente ci-dessous quelques captures du fonctionnement de *Talend* :

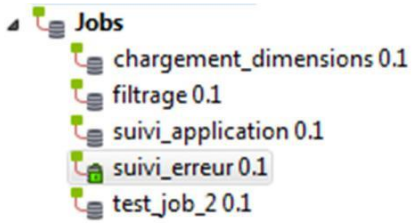

Figure 32: *fenêtre des jobs*

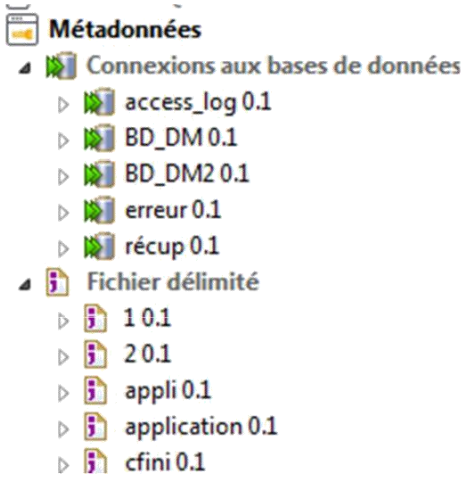

Figure 33: *fenêtre des métadonnées*

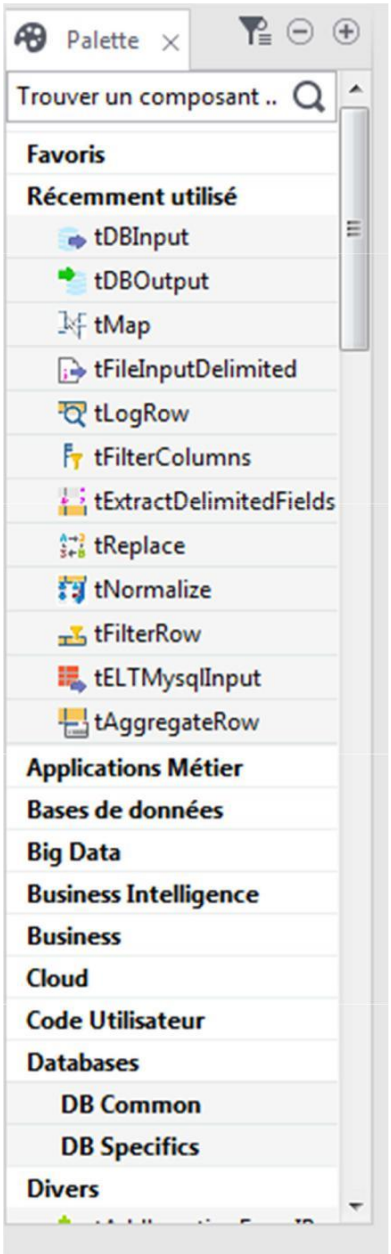

Figure 34: fenêtre de la palette

- **Connexion à la base de données :**

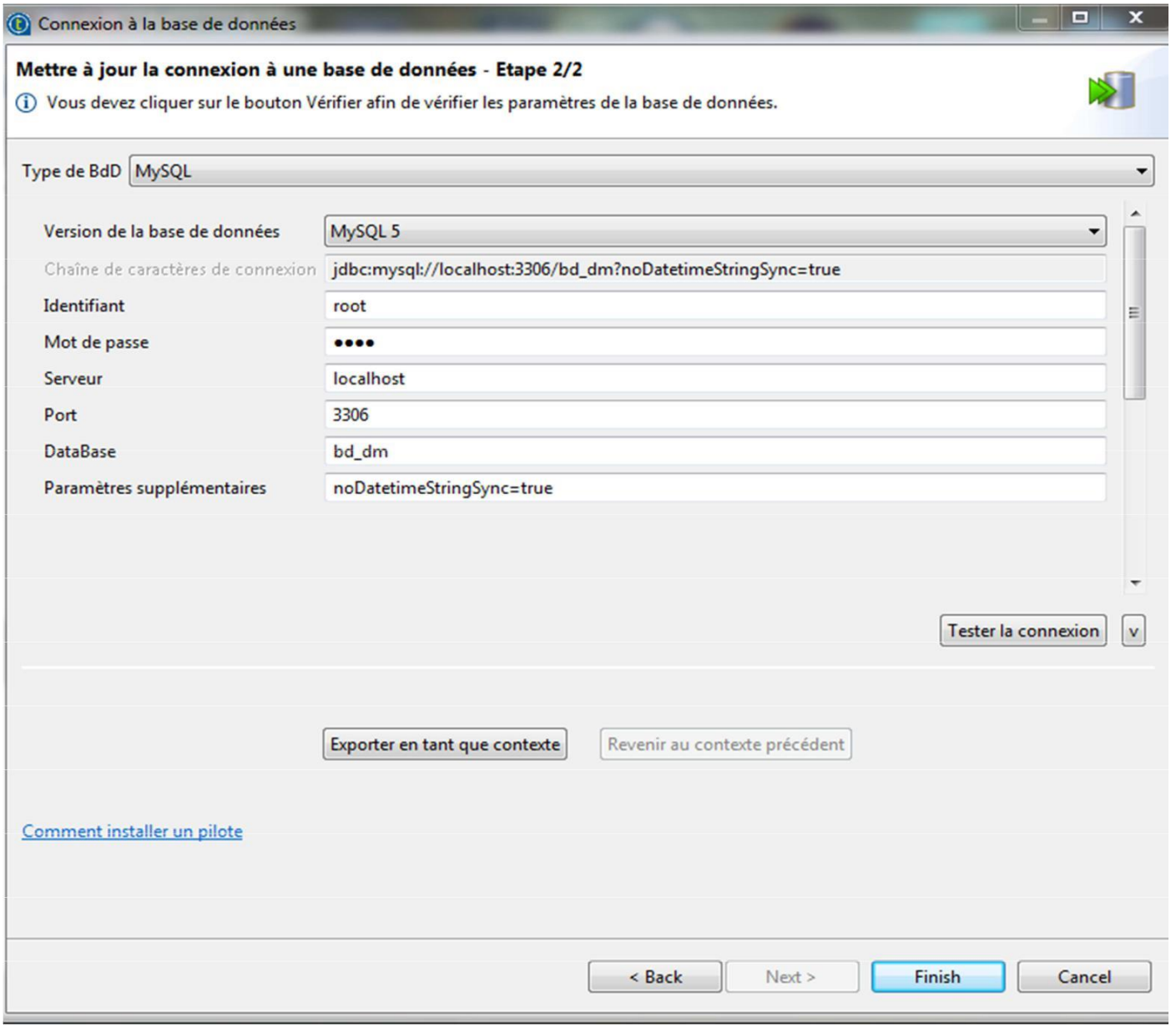

Figure 35: fenêtre de la connexion à bdd

- **Récupération du schéma de la base de données :**

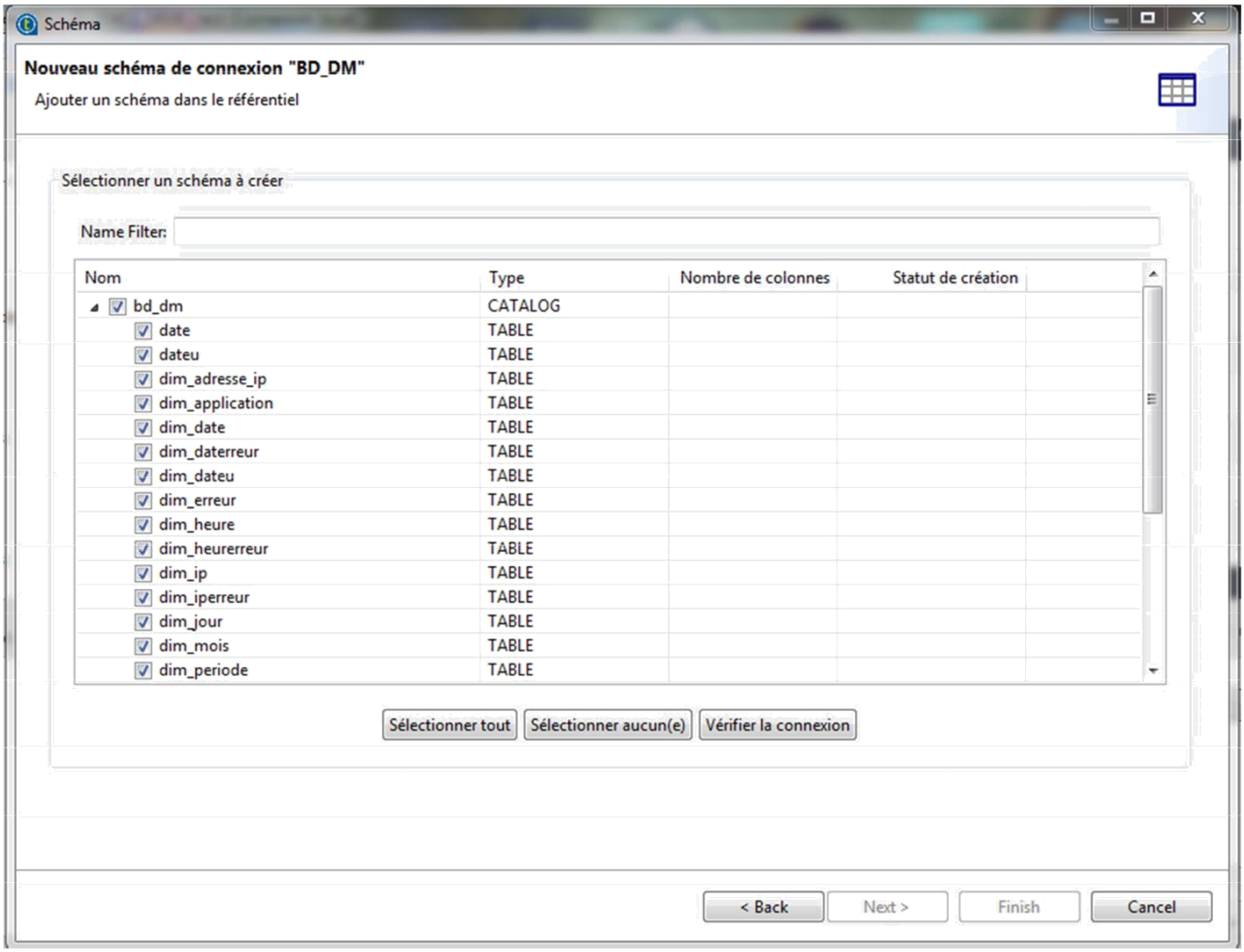

Figure 36: *fenêtre de récupération du schéma de la bdd*

- **Schéma de la base de données :**

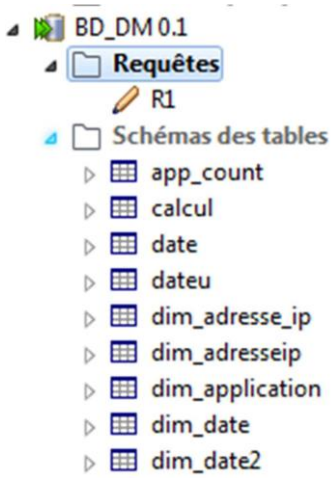

Figure 37: *fenêtre du schéma de la bdd*

- **Configuration de la base de données :**

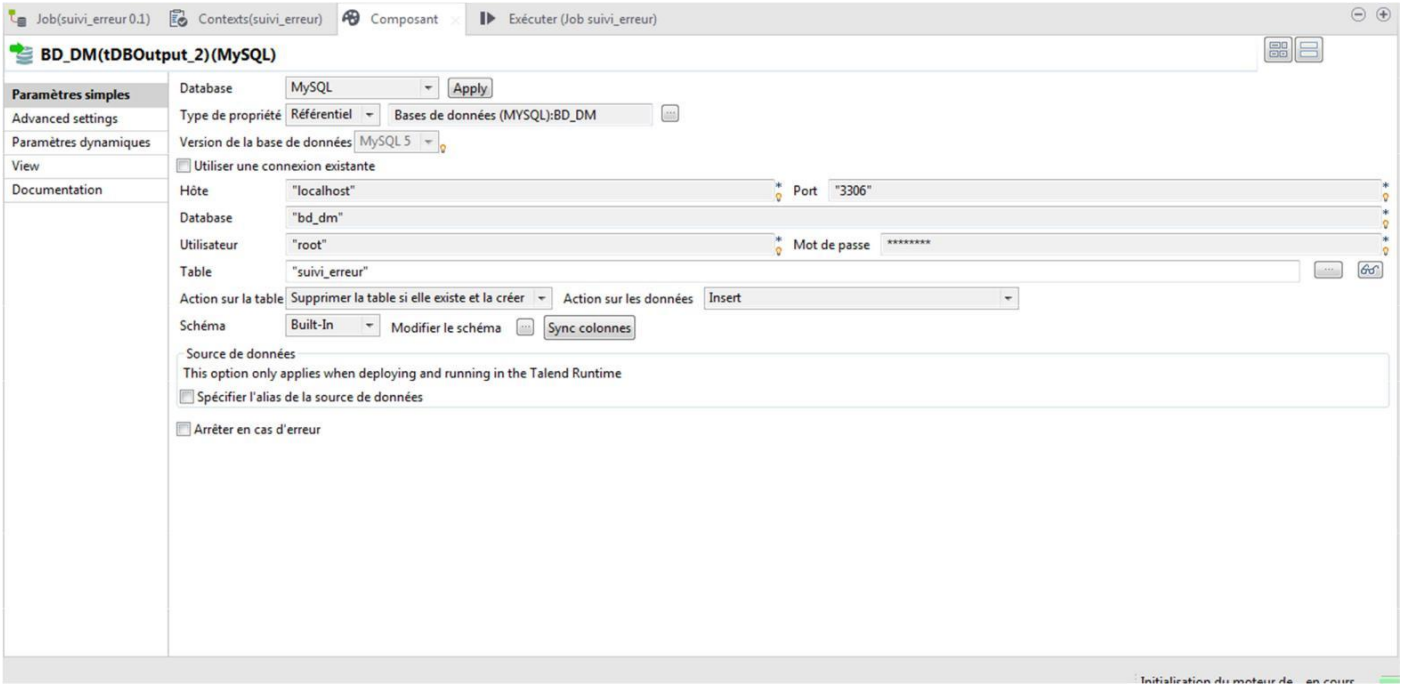

Figure 38: *Fenêtre de configuration de la bdd*

Le tableau ci-dessous présente les composants de Talend que nous avons utilisés dans le processus ETL.

Composant	Nom	Fonction
	tDBInput	• Lit une base de données et en extrait des champs à l'aide de requêtes .
	tDBOutput	• Écrit, met à jour, modifie ou supprime les données d'une base de données .
	tFileInputDelimited	• Lit un fichier ou un flux de données ligne par ligne, afin de le diviser en champs et d'envoyer ses champs

		au composant suivant
	tReplace	• Effectue un Rechercher/Remplacer dans les colonnes d'entrée spécifié es.
	tExtractDelimitedFields	• génère des colonnes multiples à partir d'une colonne.
 b	tFilterColumns	• Opère des modifications spécifiques, établies à partir d'un mapping du nom des colonnes, sur un schéma défin i.
	tLogRow	• Affiche les données ou les résultats dans la console Run.
	tMap	• Multiplexage et démultiplexage des données • transformation des données sur tout type de champs ; • concaténation et inversion de champs ; • filtrage de champs à l'aide de contraintes ; • gestion des rejets de données.
	tNormalize	• Normalise un flux entrant en fonction du standard SQL.

Tableau 12: *différents composants talend utilisés*

- **Chargement de la table de fait Suivi_erreur :**

La figure ci-dessus représente l'alimentation de la table de faits suivi_erreur.

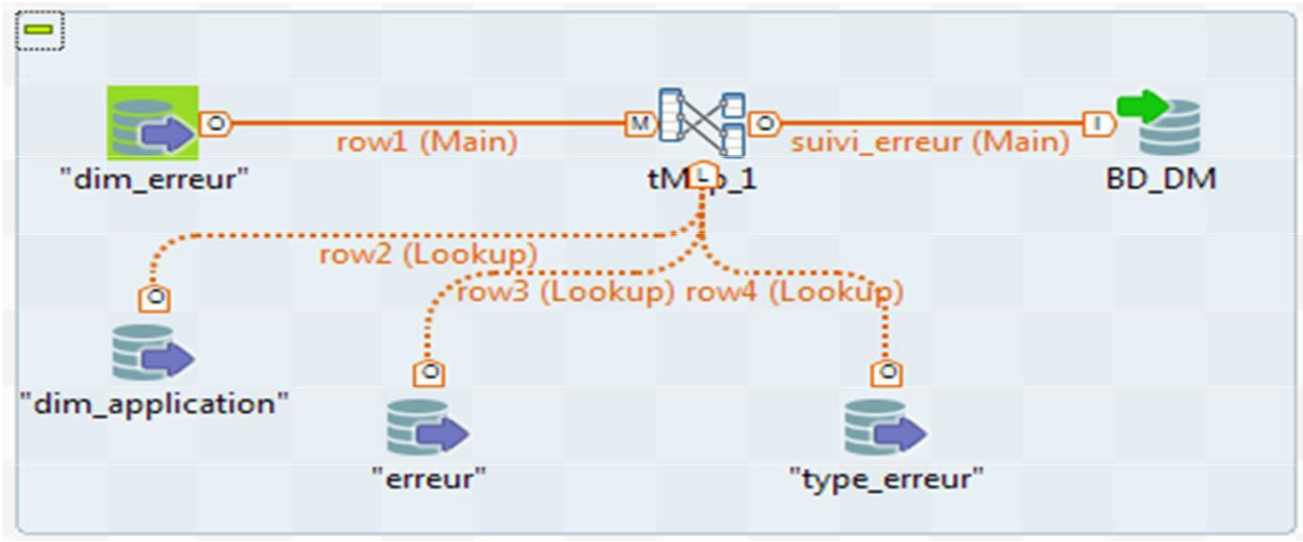

Figure 39: *alimentation de la table de fait suivi_erreur*

- **tMap du sivi_erreur :**

dans cette étape on effectue un mappage des donnéesen reliant toutes les dimensions associées à la table de faits suivi_erreur afin de l'alimenter.

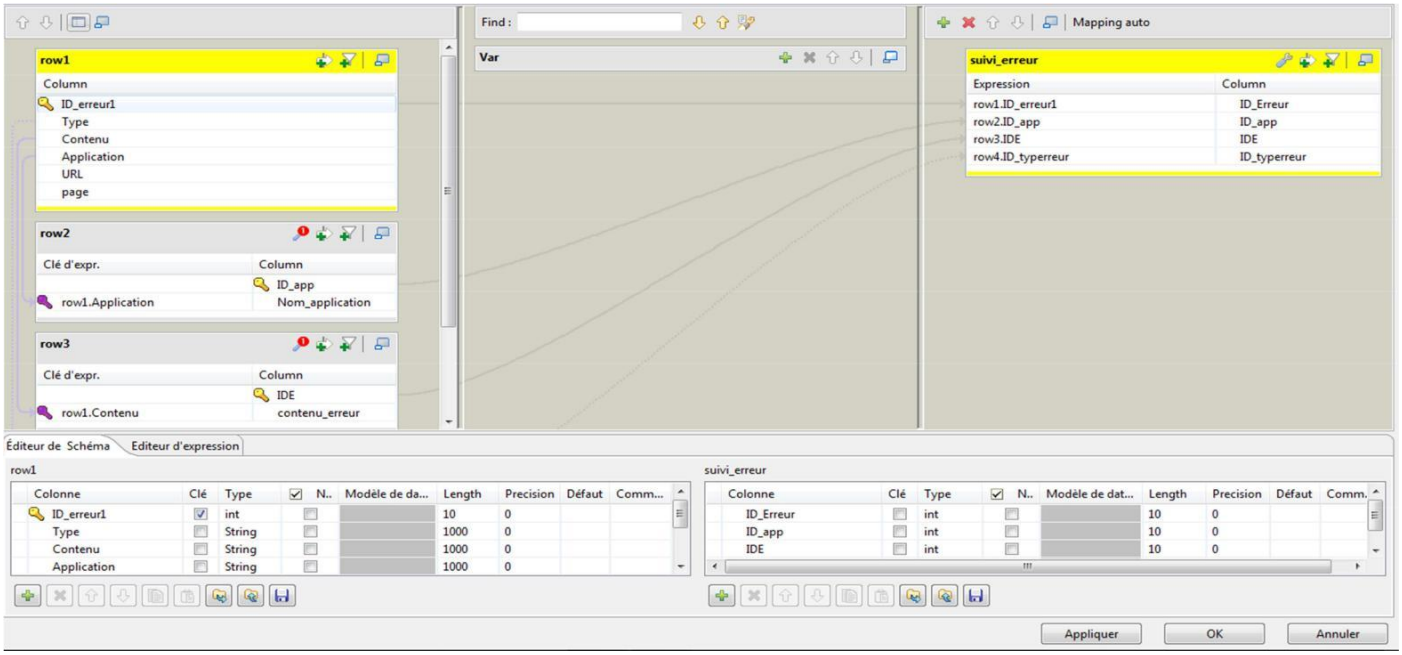

Figure 40: *tMap de la table de faitsuivi_erreur*

- **Filtrage access_log :**

Cette étape nous permet de filtrer les données en assantp par plusieurs étapes afin de les stocker dans une base de données intermédiaire(stagin area).

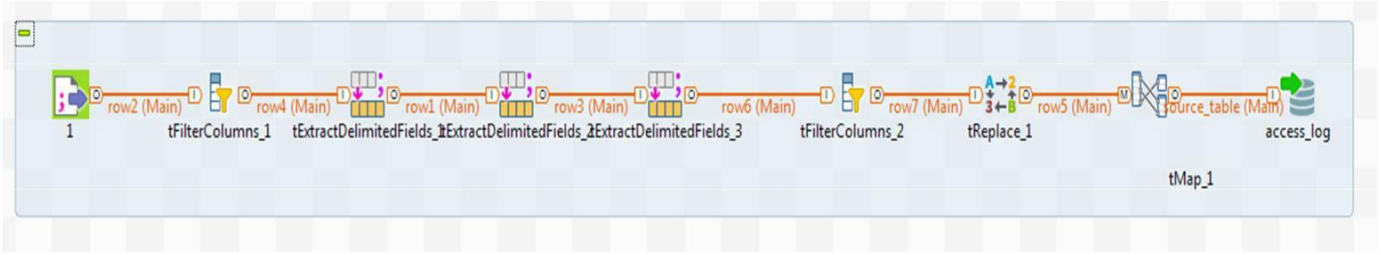

Figure 41: *filtrage du fichier log « access_log »*

Ajout des requêtes de nettoyage :

Cette étape nous permet le nettoyage des données auniveau de la staging area qui se fait avec des requêtes que nous avons définies au niveau de « SQLBuilder ».

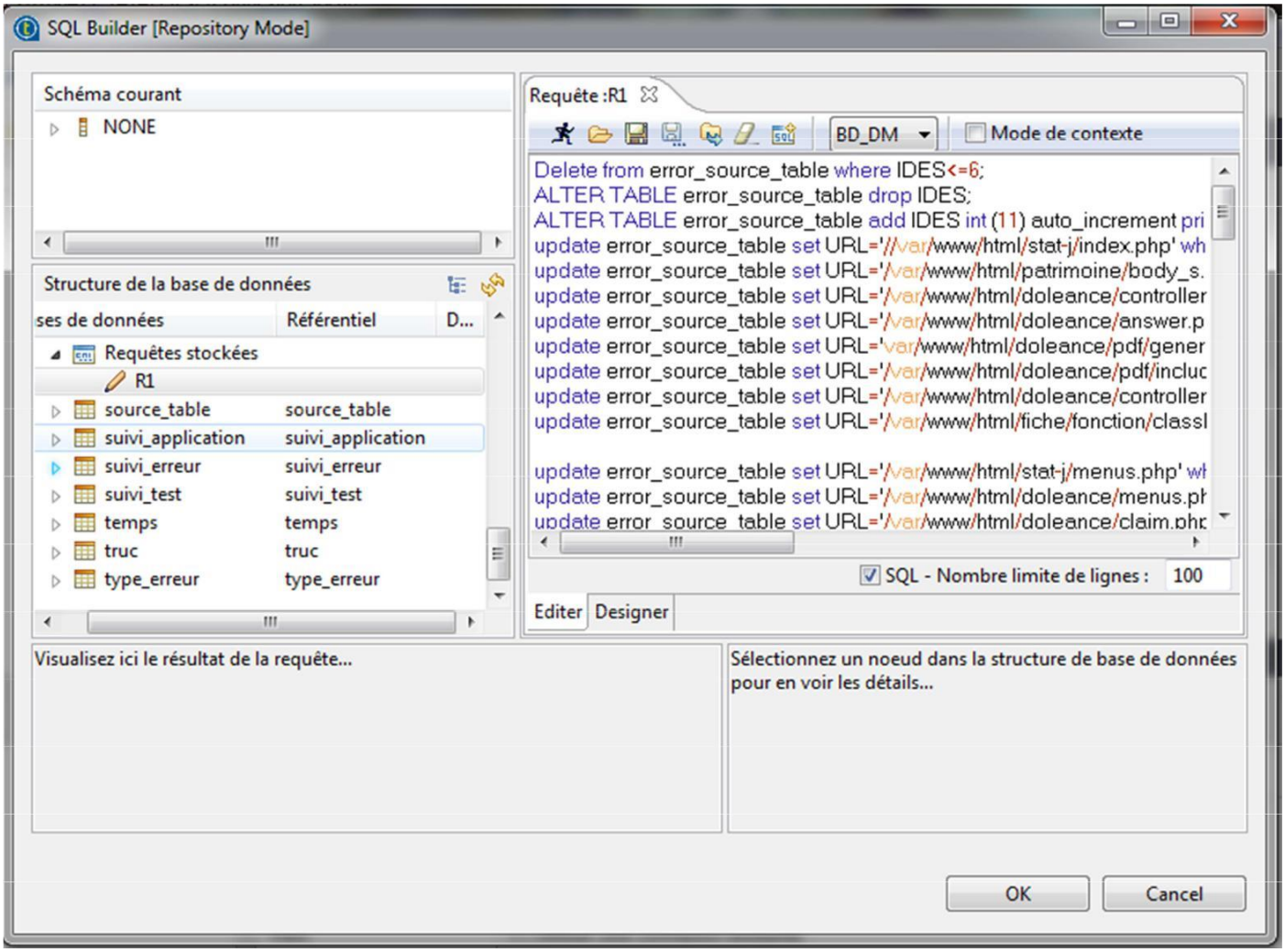

Figure 42: *ajout des requêtes de nettoyage*

- **Filtrage error_log :**

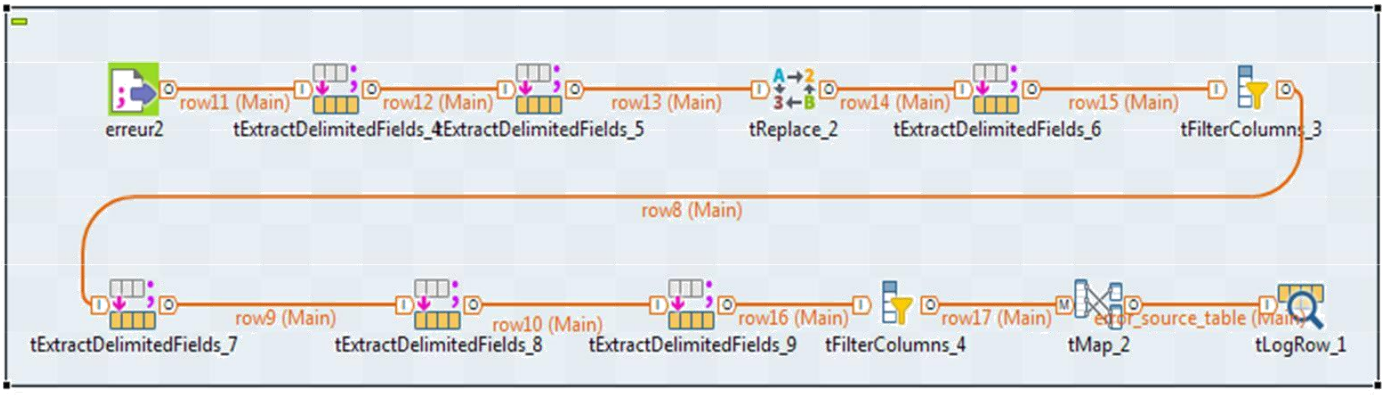

Figure 43: *filtrage du fichier log « error_log »*

2.3. Réalisation des cubes Olap :

On vous présente ci-dessous quelques captures du fonctionnement de schéma workbench pour la création des cubes olap :

- **Connexion à la base de données :**
  La figure ci-dessus représente la fenêtre de connexion à la base de données qui est nécessaire avant la création des cubes.

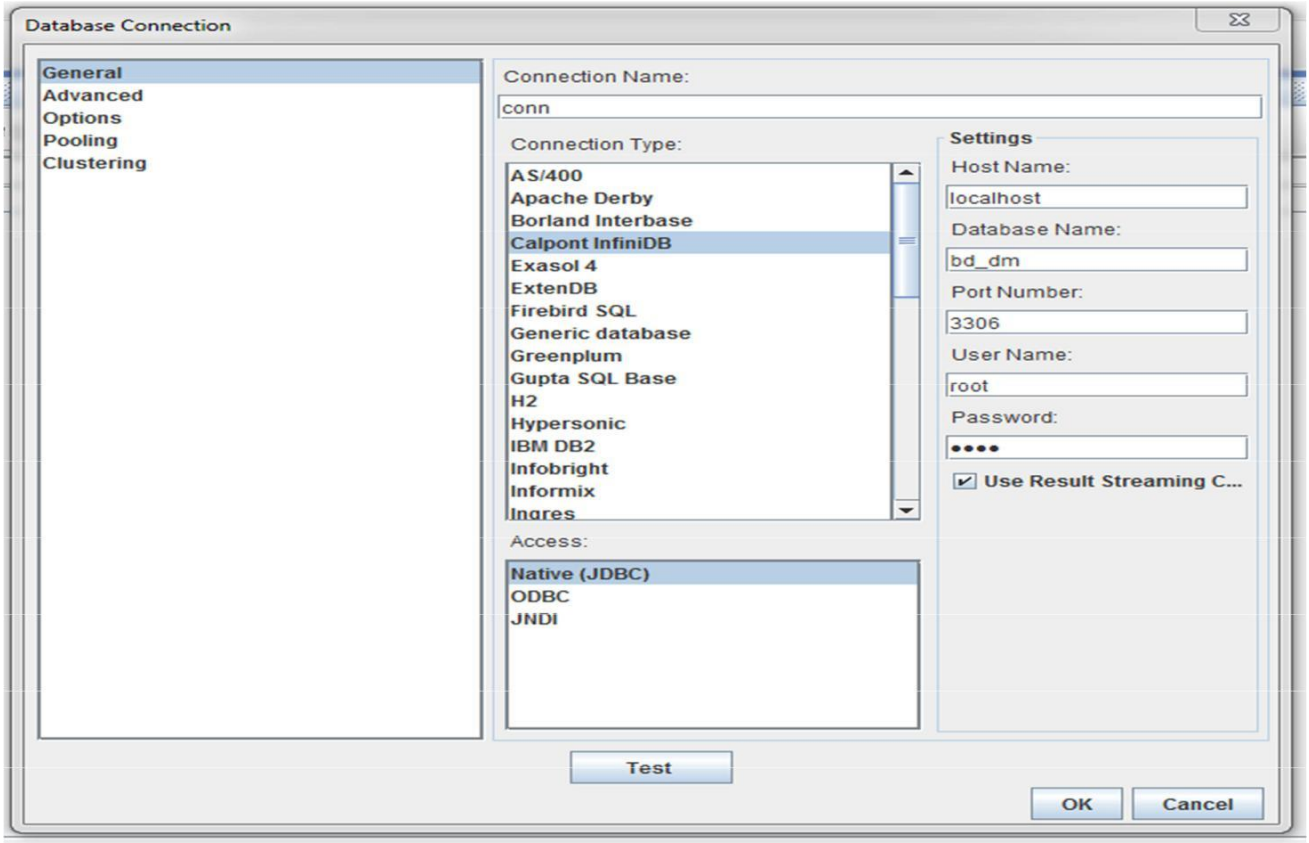

Figure 44: *fenêtre de connexion à la bdd*

Les figures ci-dessous représentent les étapes à suivre pour la création d'un cube de données.

- **Création du cube :**

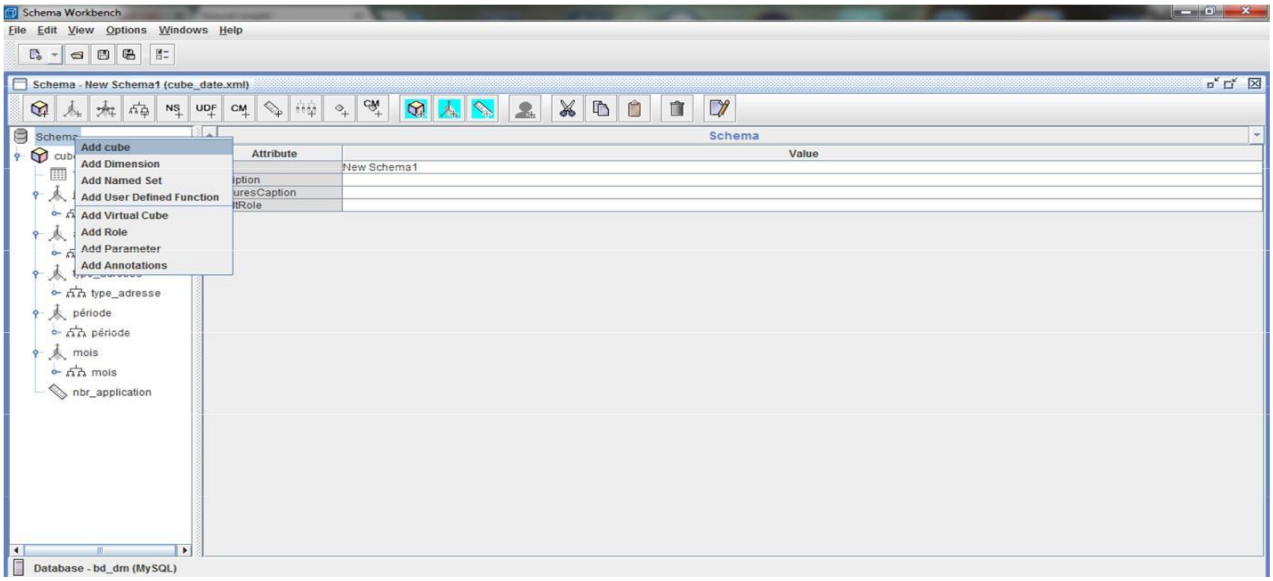

Figure 45: *création du cube*

- **Configuration du cube :**

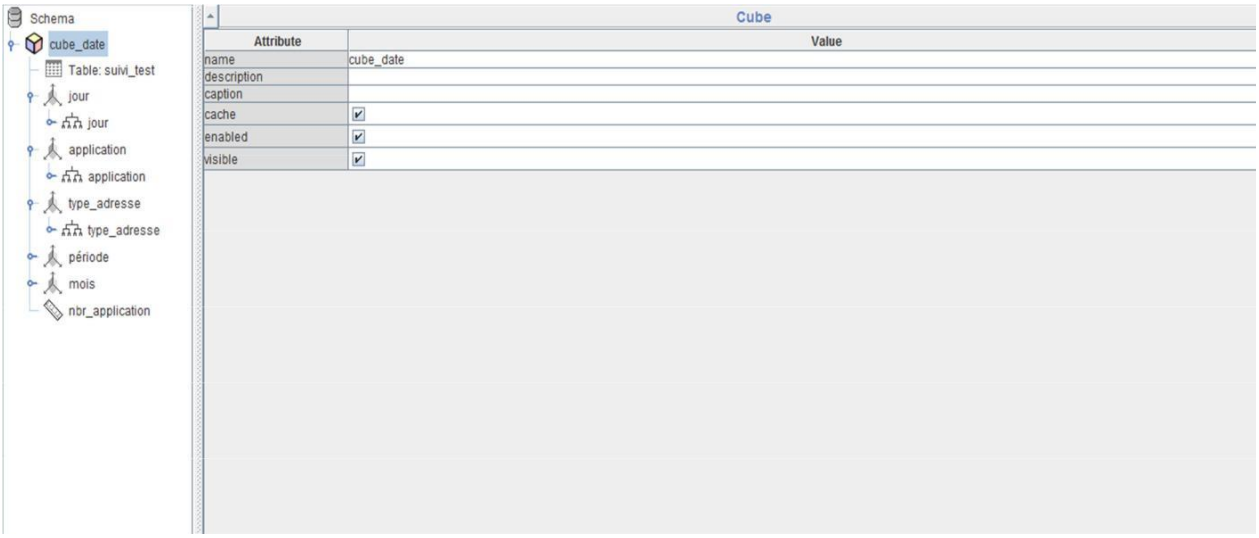

Figure 46: *fenêtre de configuration du cube*

Les figures ci-dessous décrivent comment ajouter la table de faits qui contient les clés étrangères des tables de dimensions qui représenteront nos axes d'analyse **.**

- **Ajout de la table de fait :**

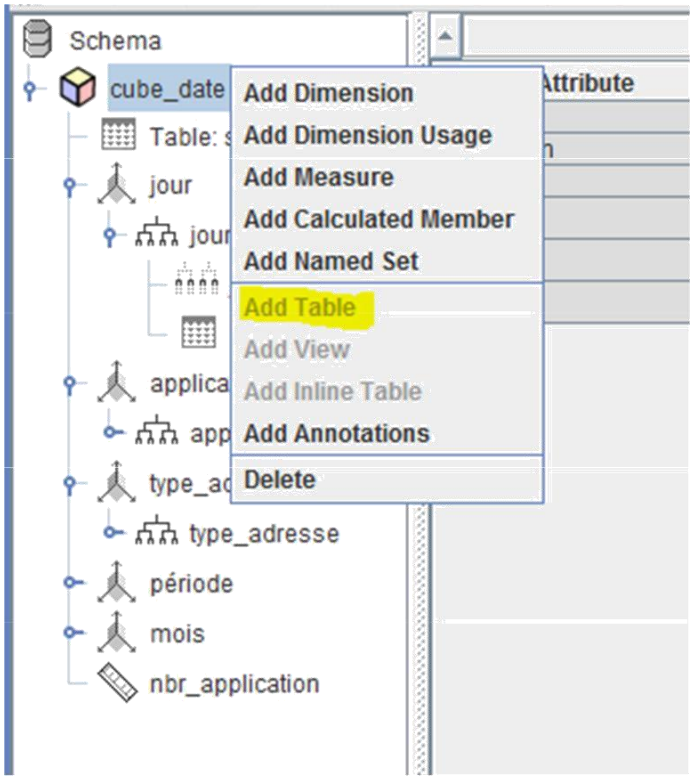

Figure 47: *ajout de la table de fait*

Les figures ci-dessous décrivent les étapes à suivre pour l'ajout d'une dimension au cube de données :

- **Ajout d'une dimension :**

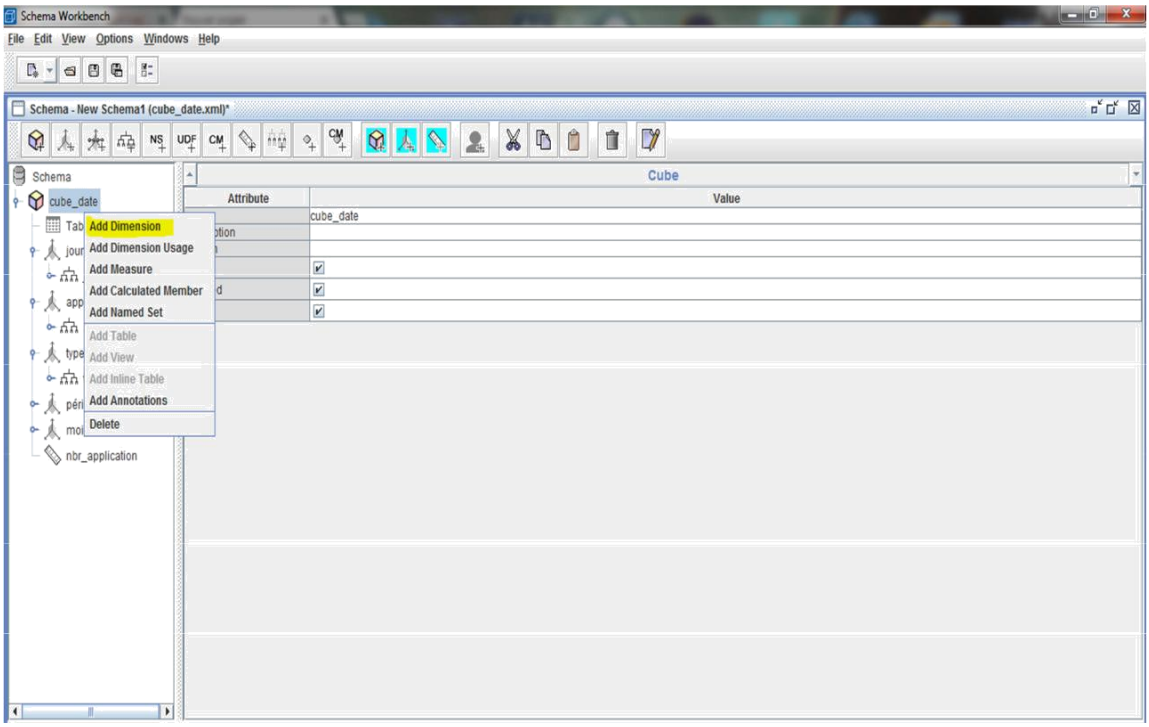

Figure 48: *ajout de la dimension*

Configuration de la dimension :

C'est à ce niveau que l'on décrit le nom de la dimension la table qui lui est associée ainsi que la clé primaire de la table qui représente une clé étrangère dans la table de faits ajoutée au cube.

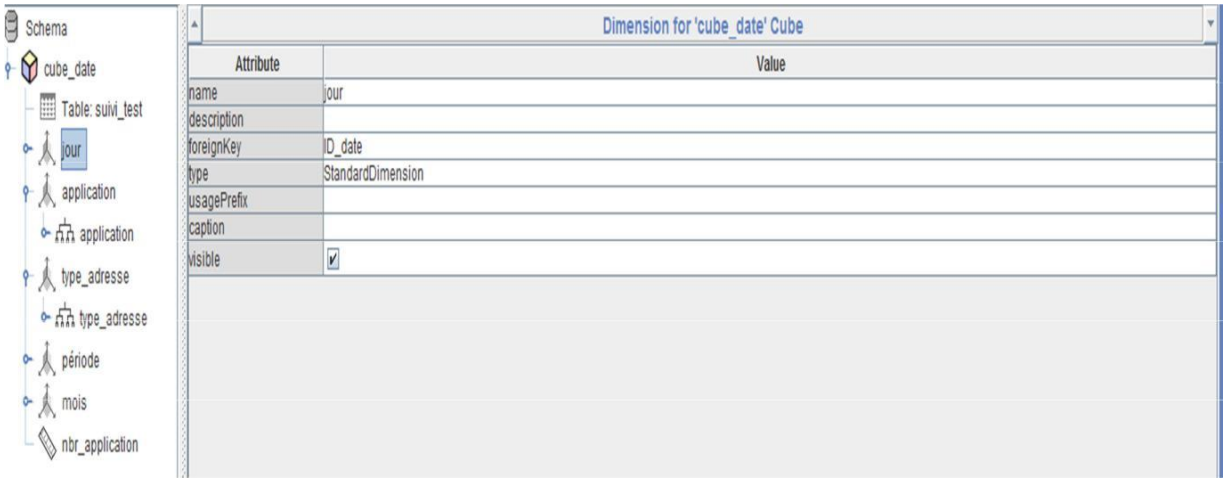

Figure 49: *fenêtre de configuration de la dimension*

Les figures ci-après décrivent les étapes à suivrepour l'ajout d'une hiérarchie aux dimensions.

Ajout de la table de dimension :

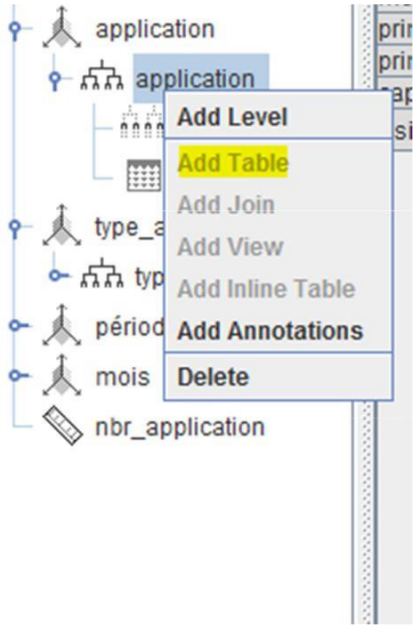

Figure 50: *ajout de la table de dimension*

Ajout des hiérarchies :

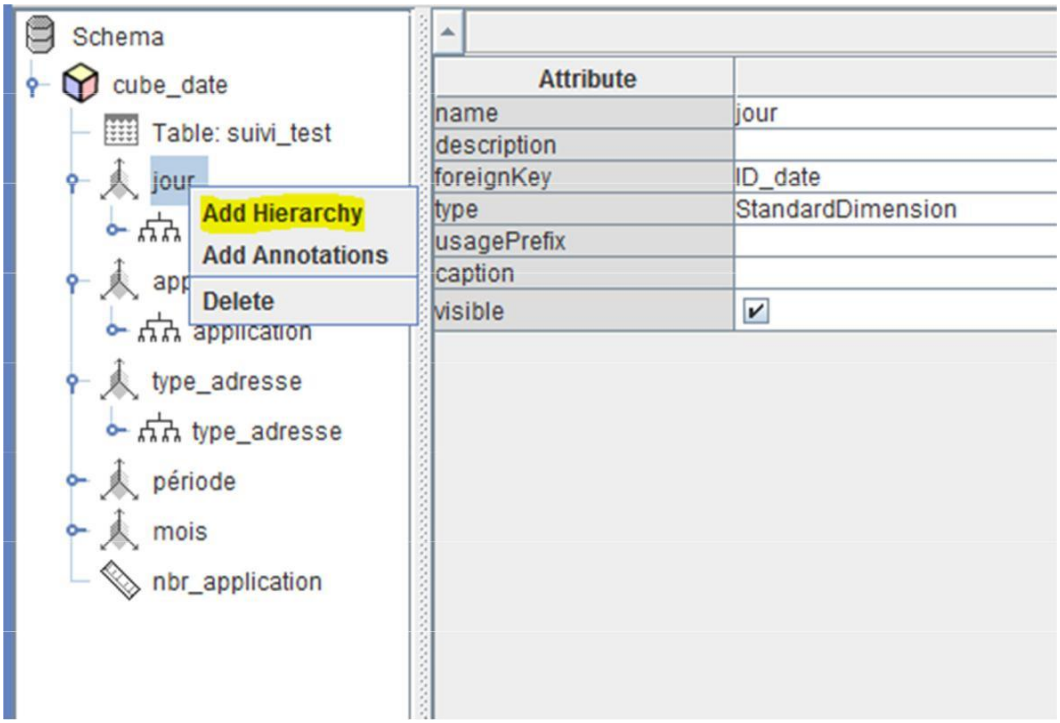

Figure 51: *ajout d'une hiérarchie*

- **Configuration de la hiérarchie :**

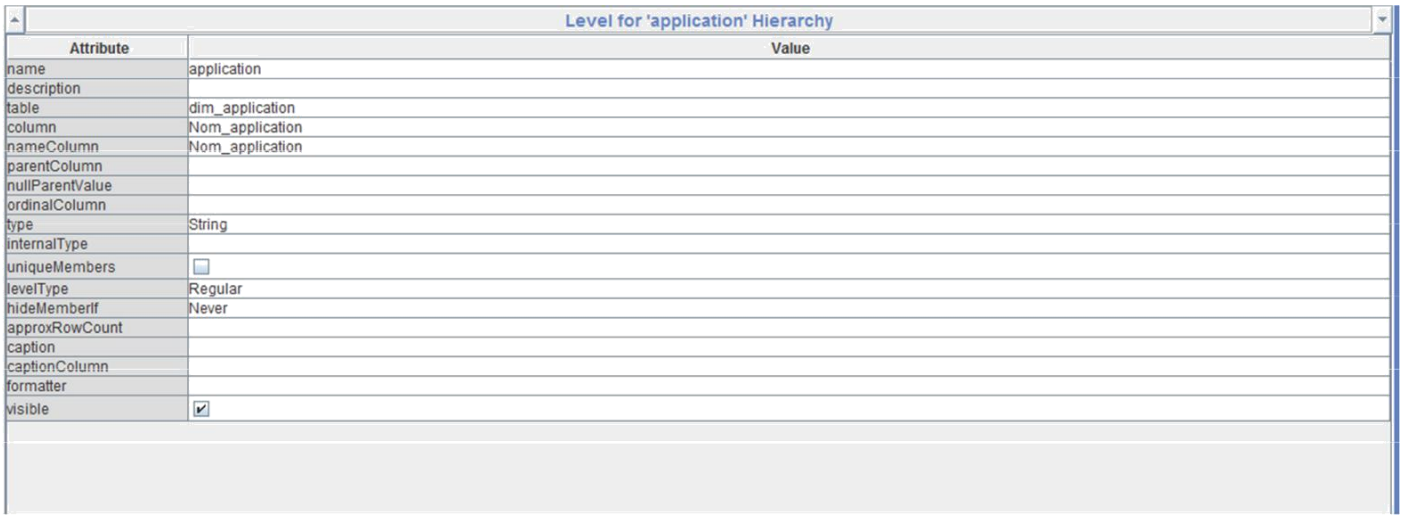

Figure 52: *fenêtre de configuration de la hiérarchie*

Les figures ci-dessous décrivent les étapes à suivre pour l'ajout des mesures au cube de données :

- **Ajout des mesures :**

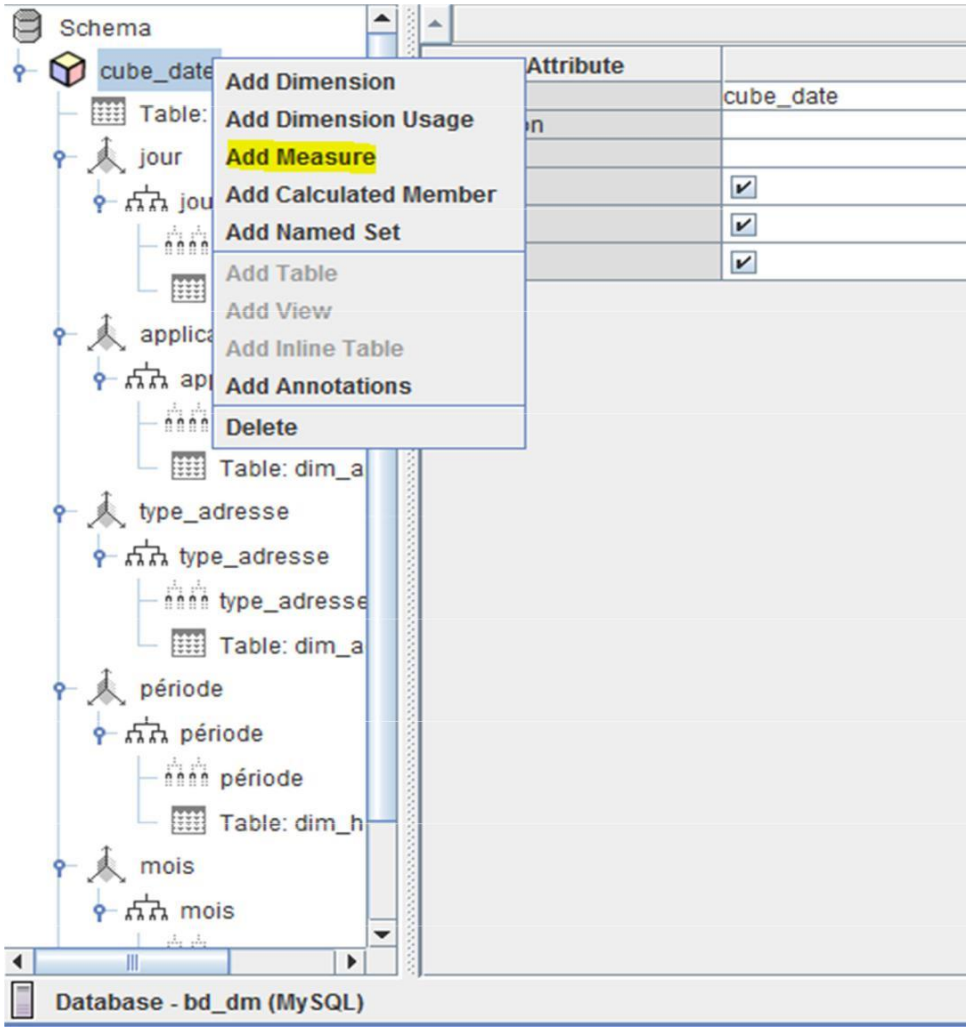

Figure 53: *ajout d'une mesure*

- **Configuration de la mesure :**

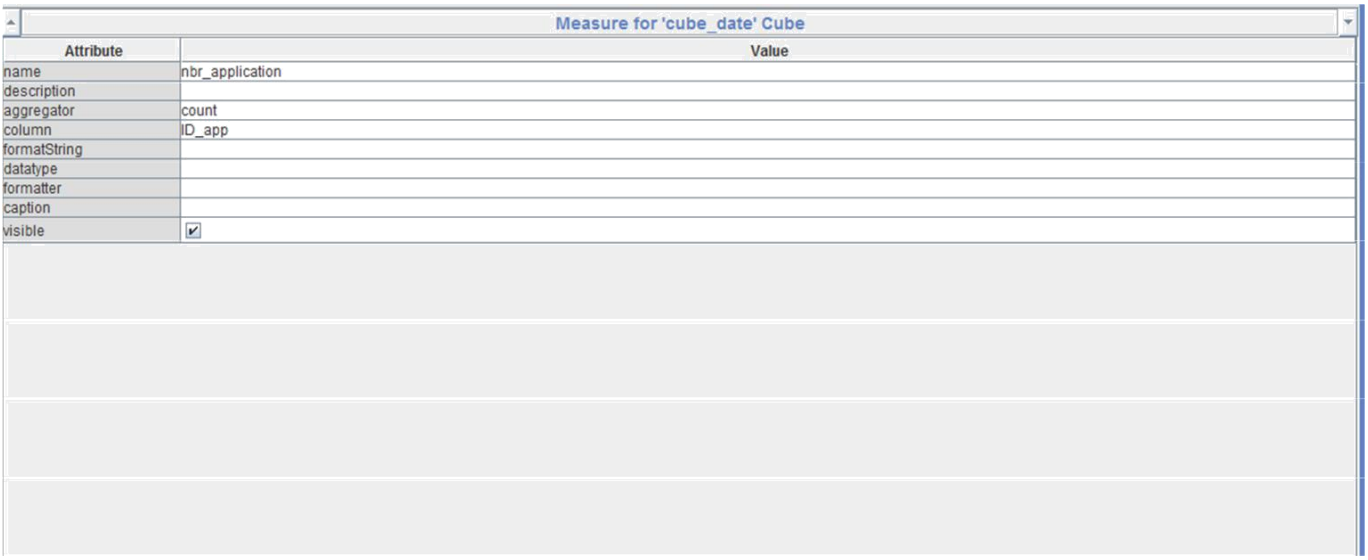

Figure 54: *fenêtre de configuration de la mesure*

Les figures ci-dessous représentent les schémas desdeux cubes que nous avons créés avec leurs tables de faits, dimensions et mesures.

- **Schéma du cube application :**

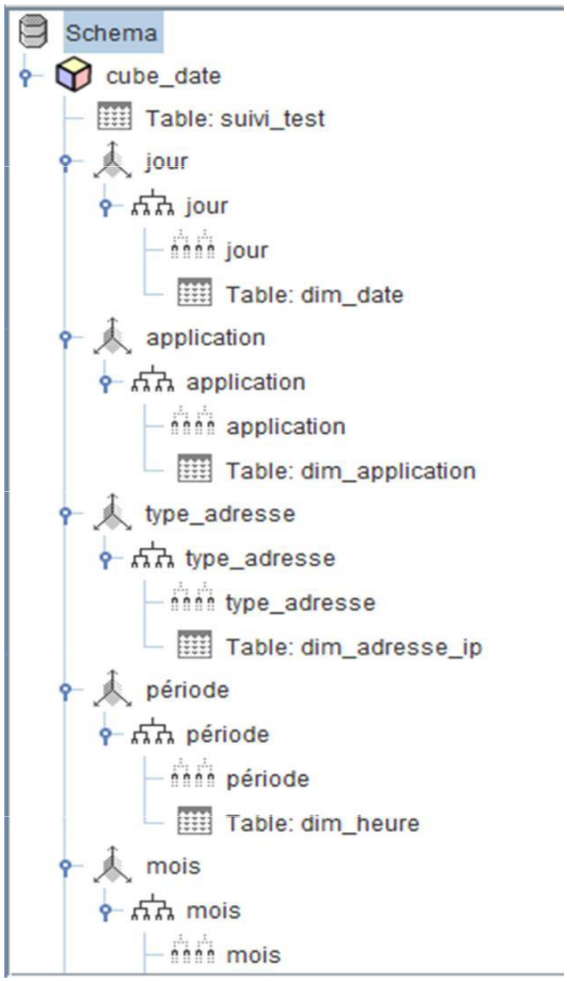

Figure 55: *schéma du cube application*

- **Schéma du cube erreur :**

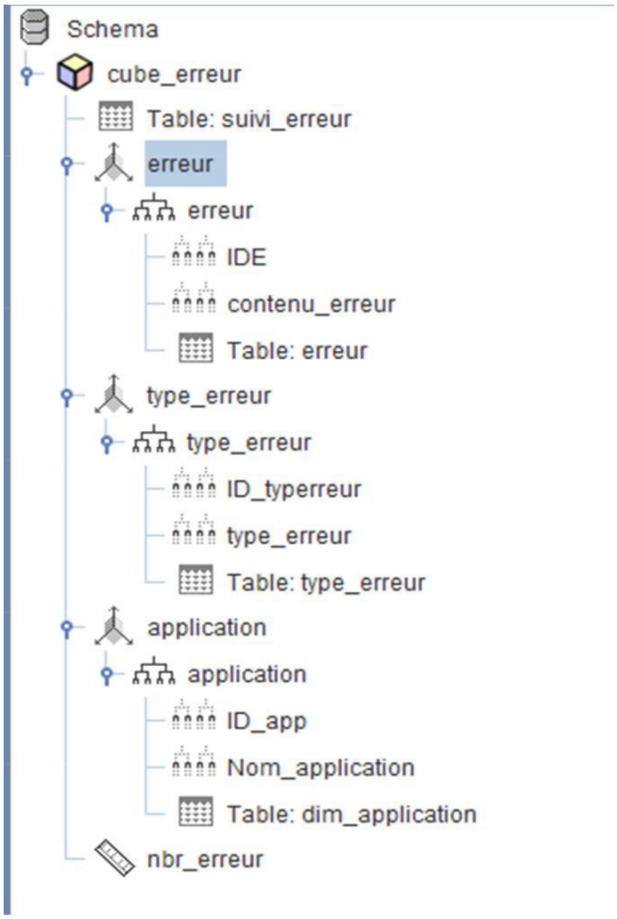

Figure 56: *schéma du cube erreur*

Schéma xml du cube application :

```
<Schema name="New Schema1">
  <Cube name="cube_date" visible="true" cache="true" enabled="true">
    <Table name="suivi_test">
    </Table>
    <Dimension type="StandardDimension" visible="true" foreignKey="ID_date" highCardinality="false" name="jour">
      <Hierarchy name="jour" visible="true" hasAll="true" primaryKey="ID_date">
        <Table name="dim_date">
        </Table>
        <Level name="jour" visible="true" table="dim_date" column="jour" nameColumn="jour" type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" visible="true" foreignKey="ID_app" highCardinality="false" name="application">
      <Hierarchy name="application" visible="true" hasAll="true" primaryKey="ID_app">
        <Table name="dim_application">
        </Table>
        <Level name="application" visible="true" table="dim_application" column="Nom_application" nameColumn="Nom_application" type="String" uniqueMembers="false" levelType="R
egular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" visible="true" foreignKey="ID_ip" highCardinality="false" name="type_adresse">
      <Hierarchy name="type_adresse" visible="true" hasAll="true" primaryKey="ID_ip">
        <Table name="dim_adresse_ip">
        </Table>
        <Level name="type_adresse" visible="true" table="dim_adresse_ip" column="type_adresse" nameColumn="type_adresse" type="String" uniqueMembers="false" levelType="Regu
lar" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" visible="true" foreignKey="id_heure" highCardinality="false" name="p&#233;riode">
      <Hierarchy name="p&#233;riode" visible="true" hasAll="true" primaryKey="id_heure">
```

Figure 57 : *schéma xml du cube*

2.4. Réalisation des rapports :

Les figures ci-dessous représentent les étapes quenous avons suivies pour la création des rapports :

La fenêtre d'authentification :

La figure ci-dessous représente la fenêtre d'authentification de l'utilisateur au serveur **Jaspersoft :**

Figure 58: *fenêtre d'authentification*

- **Fenêtre d'accueil :**
  La figure 59 représente la fenêtre d'accueil qui s'affiche après authentification :

Figure 59: *fenêtre d'accueil*

- **Création de la source de données :**

Avant l'ajout des cubes de données créés il est nécessaire de définir notre source de données qui est notre datawarehouse.

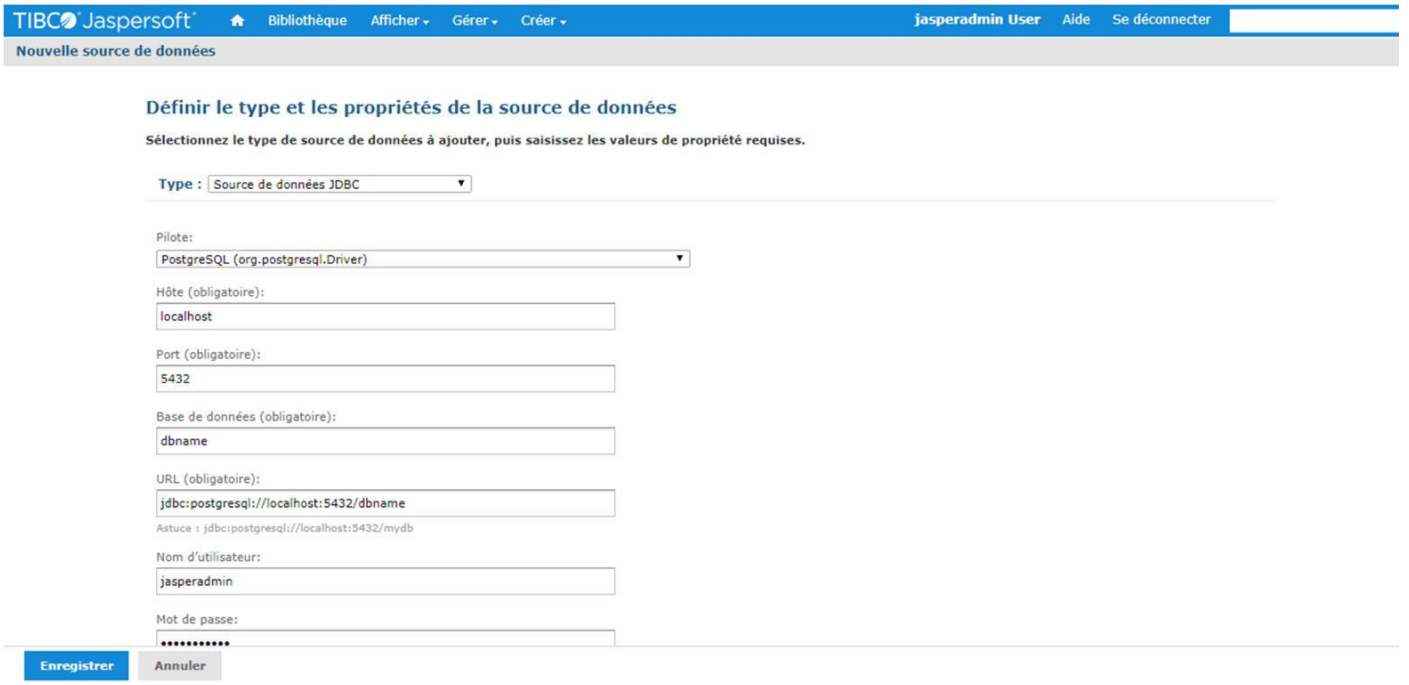

Figure 60: *création de la source de données*

- **Création de la connexion mondrian :**

  Cette étape consiste à créer une connexion entre **pentaho workbench mondrian** et **jaspersoft** pour l'affichage des cubes de données créés.

Figure 61: *création de la connexion mondrian*

- **Liste des sources de données :**
  La **figure 62** représente les sources de données créées au niveaudu serveur jaspersoft.

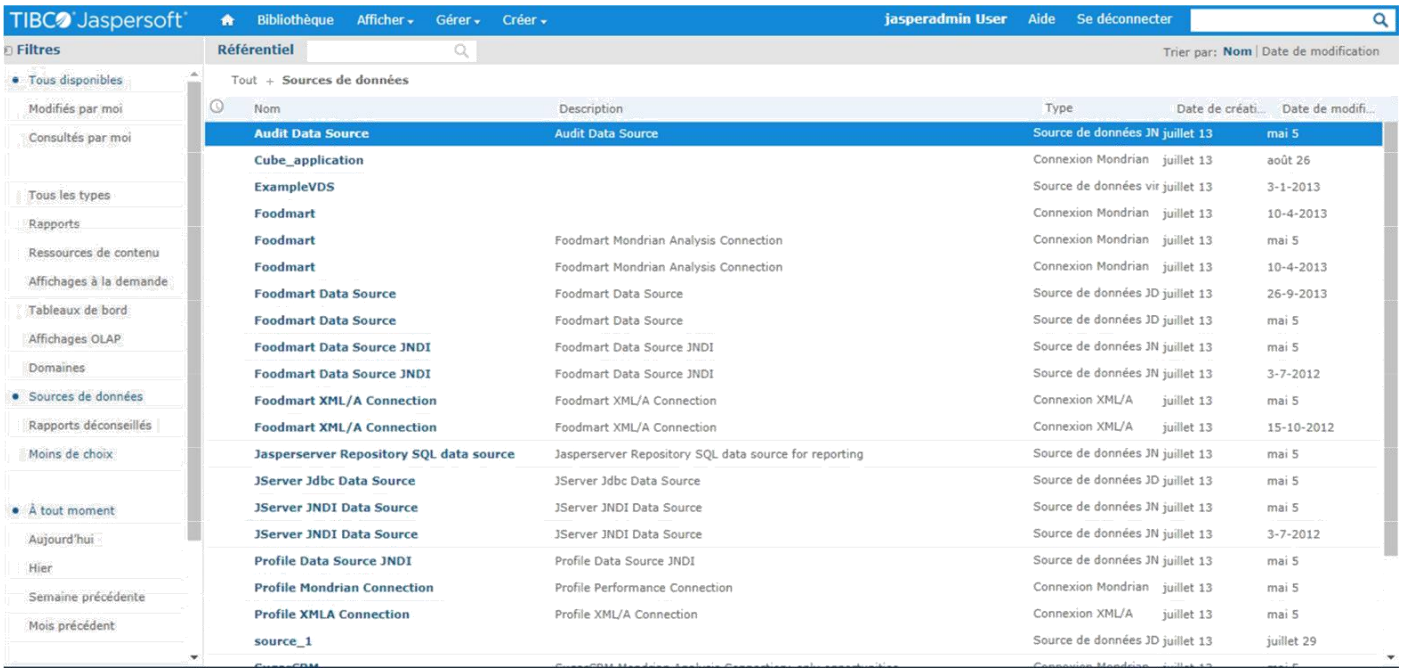

Figure 62: *liste des sources de données*

- **Sélection du cube à afficher**
  Cette étape consiste à sélectionner le cube de données à afficher.

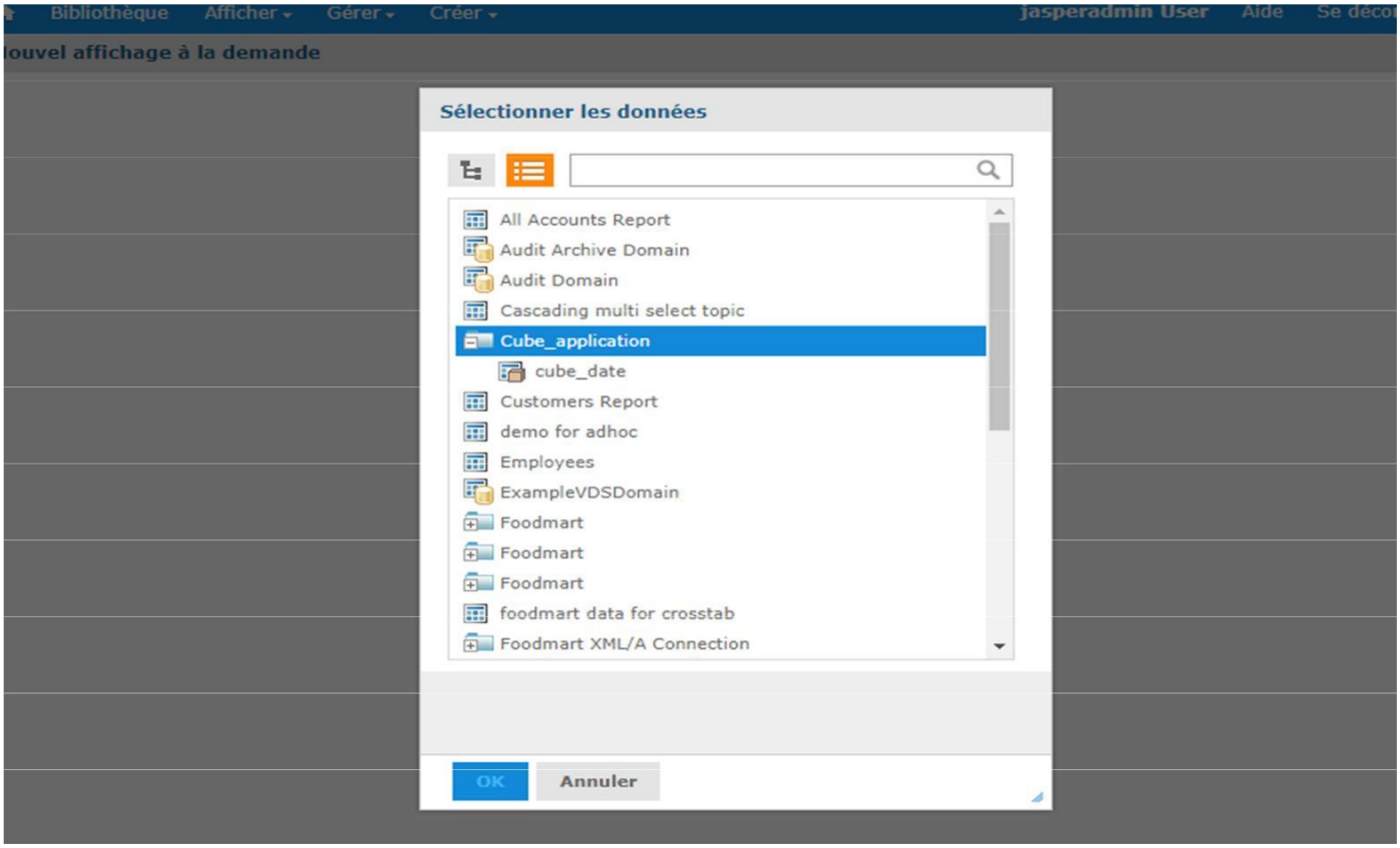

Figure 63: *sélection du cube à afficher*

Les figures ci-dessous représentent quelques rapports que nous avons créés avec**Jaspersoft.**

- **rapport suivi_application :**

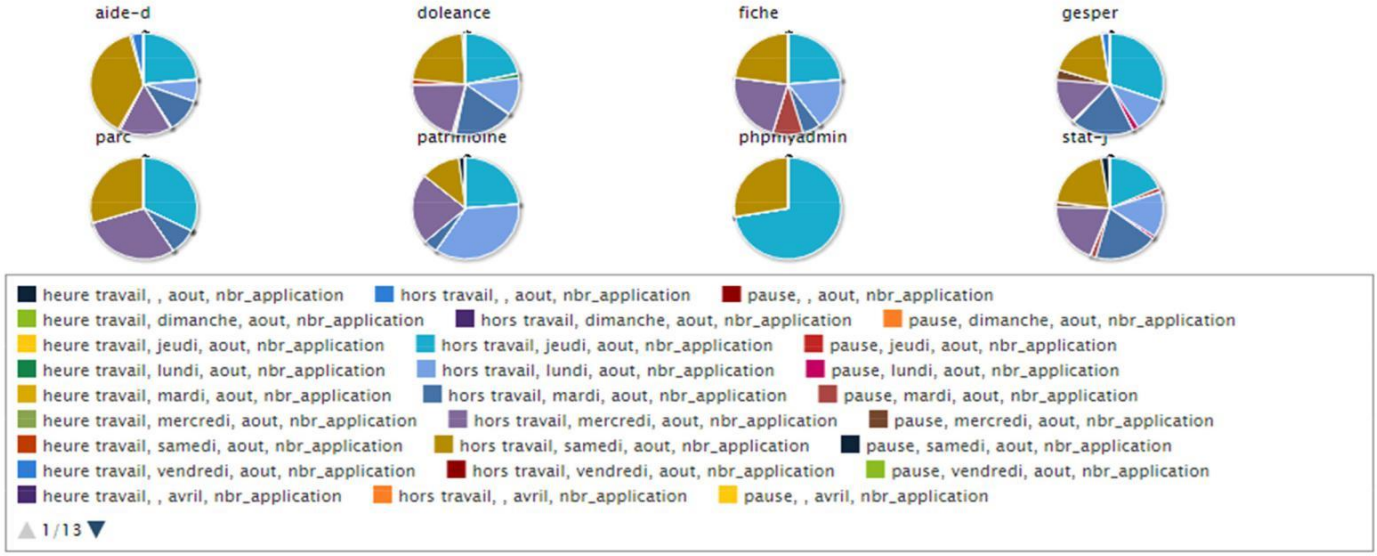

Figure 64: *nombre d'accès aux applications suivant la période,le jour et le mois*

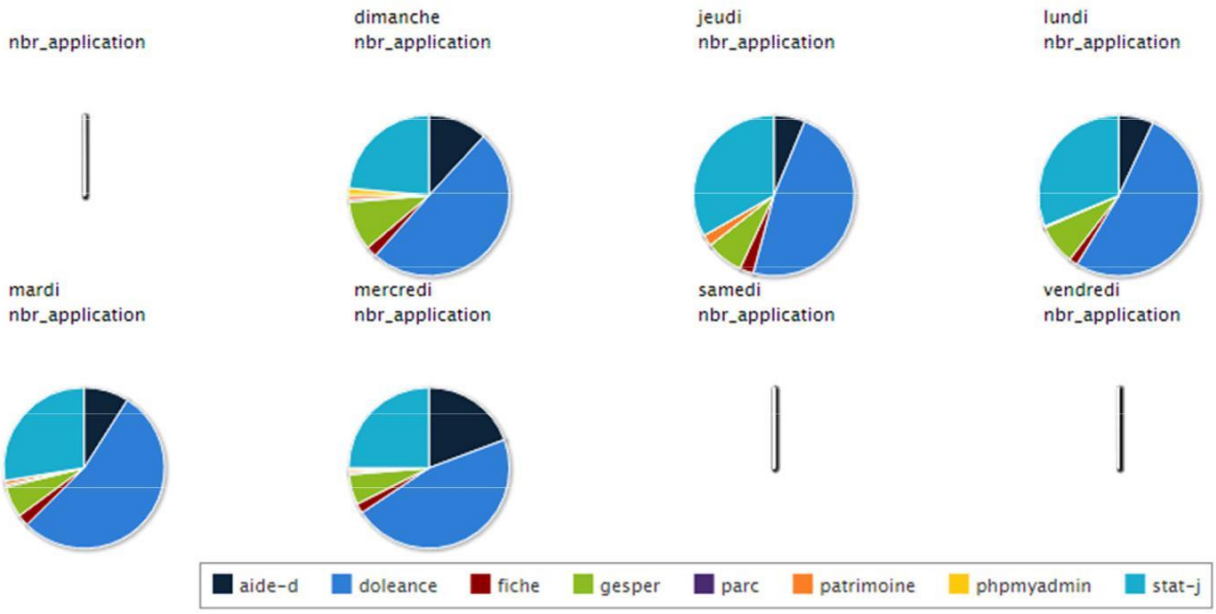

Figure 65: *nombre d'accès aux applications suivant le jour*

Conclusion :

A travers ce chapitre, nous avons présenté étape par étape la réalisation du système décisionnel, allant des outils et des technologies utilisés pour l'implémentation de l'entrepôt de données, et sa zone d'alimentation jusqu'à l'aboutissement du prod uit final.