REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahlab de Blida 1

Faculté des sciences

Département Informatique

Mémoire de fin d'étude pour l'obtention du diplôme de Master 2 en Informatique

Option : Génie Logiciel

Thème

Conception & Réalisation d'un Système d'Information Décisionnel pour l'analyse du trafic web sur un réseau intranet

Proposé par :

Mr. Khider Mohamed

Promotrice: Mlle Attaf Sarah

Réalisé par :

Mr. Bouatmane Abdessamed

Mr. Hadj Mohammed Ayoub

Présidente du Jury : N. BOUSTIA

Examinateurs : K. AMEUR

M. MEZZI

Organisme d'Accueil :

Office National des Statistiques

Promotion 2014


بسم الله الرحمن الرحيم

Nous tenons à remercier ALLAH qui nous a aidé et nous a donné la patience et le courage durant ces longues années d'études, et nous avoir données la force à accomplir ce travail.

A la mémoire de Mr RAHMOUN BOUALEM (رحمه الله)

Qui a toujours servi la science, il a tout donné pour les étudiants à l'université de Saad Dahleb Blida, il n'est pas parmi nous aujourd'hui mais il vit avec nous dans notre mémoire et toujours dans nos cœurs.

Remerciement

Mes plus profonds remerciements vont tout spécialement à mes parents, mes frères et mes chères sœurs qui ont su me supporter et m'encourager tout au long de ma vie, ainsi que pour leur aide inestimable, leur patience et leur soutien indéfectible.

Je tiens à remercier tous les enseignants qui ont contribué de près ou de loin à ma formation.

Je remercie également notre chef de département d'informatique Mr Massied Mohamed ainsi que tout le personnel et les enseignants du département pour leur soutien.

Je souhaite remercier Mr : Khider Mohamed qui nous a proposé le thème, pour son entière disponibilité, son aide et ses conseils.

Je tiens aussi à remercier notre promotrice Mlle Sarah Attaf pour sa confiance, ses remarques et sa bienveillance.

Je voudrais également remercier les membres du jury pour avoir accepté d'évaluer ce travail et pour toutes leurs remarques et critiques.

« Something that has always pleased me is when I am in special need of help, the good deed is usually done by you, you supported me, and you encouraged me, thank you so much... »

J'exprime ma gratitude et mes personnelles remerciements à mon cher ami, mon frère, mon binôme Ayoub avec qui j'ai partagé ce travail et qui m'a accompagnée durant tout le déroulement de ce projet, ainsi que ces parents qui m'ont considéré comme un membre de famille et qui m'ont toujours soutenu.

Je souhaite aussi remercier tous mes très chers amis Hicham, Oussama, Abdou, Muss, Kouceila et mon voisin Youcef pour leur soutien, sans oublier la promotion de Master 2 GL et spécialement El Arbi Rabah Mohamed pour sa disponibilité et son aide.

Mes remerciements vont enfin à Rédha et Fouad, Merci à vous pour vos aides ! Merci à vous pour vos soutiens et vos gestes généreux, c'est une preuve d'une belle amitié, Mille mercis à vous mes amis que j'aime et à qui je tiens énormément.

Merci à vous tous

Bouatmane Abdessamed

Remerciement

Je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire spécialement à l'ensemble de profs et administrateurs de l'université Saad Dahlab Blida et son département informatique

Je tiens à remercier sincèrement Mr. Taleb Mohamed, qui, en tant que Directeur de répertoire informatique au sein de l'ONS, s'est toujours montré à l'écoute et très disponible tout au long de la réalisation de ce travail, ainsi pour l'inspiration, l'aide et le temps qu'il a bien voulu me consacrer.

Je tiens à remercier vivement :

Mon encadreur Mr. Khider Mohamed pour ces conseils et son suivi durant la réalisation de mon projet.

Ma promotrice Mlle Attaf Sarah pour m'avoir honoré en acceptant de diriger ce travail et pour ses conseils.

J'exprime ma gratitude, mes remerciements à mes parents qui ont fait de leur mieux pour m'aider.

J'adresse mes plus sincères remerciements à tous mes proches et amis, qui m'ont toujours soutenue et encouragée au cours de la réalisation de ce mémoire

Je remercie mon binôme Bouatmane Abdessamed qui m'a accompagné durant tout le déroulement de ce projet avec beaucoup de sagesse, perfection et patience

Pour finir, j'adresse mes remerciements aux membres du jury.

Hadj Mohammed Ayoub

Résumé : Ce travail consiste à concevoir et réaliser un outil d'aide à la décision, en se basant sur l'ECD (Extraction de la Connaissance des bases de données), en utilisant les concepts du Web Usage Mining, pour offrir aux administrateurs d'un serveur web apache l'ensemble de connaissances, y inclut les statistiques sur leurs sites. Il s'agit en fait, d'extraire de l'information à partir du fichier journal du serveur Web apache, et découvrir les habitudes et le comportement des visiteurs, et de répondre à leurs besoins en adaptant le contenu et la forme des pages web et du contenu d'une façon générale.

Mots Clés : Analyse du trafic web sur le réseau intranet, web usage mining, Fichier journal, Système décisionnel.

Summary: This job is to design and implement a tool for the decision, based on EKD (Extraction of Knowledge databases), using the concepts of Web Usage Mining, to provide to apache web server administrators all knowledge, it includes statistics on their websites. This is in fact, to extract information from the log file of apache web server, to discover visitors patterns and behavior, and to respond to their needs by adapting the content and form of web pages and the contents in general.

Keywords: Analysis of web traffic on the intranet, La fouille de l'usage du web, log file, Decision System.

ملخص : هذا العمل هو تصميم و تحقيق أداة مساعدة لإتخاذ القرار . مبنية على ال (إ ق ب م) " إستخراج قواعد-البيانات المعرفية " مستعملا مفاهيم – التنقيب في استعمال الشبكة – لتوفر لإداريي خادم شبكة آباتش مجمع المعارف ... ما يشمل الإحصائيات الخاصة بموقعهم... و يتجلى في :استخراج المعلومات من ملف التسجيل لخادم الشبكة آباتش , و اكتشاف اعتياد و سلوك الزوار , و الإجابة عن احتياجاتهم مكيفين محتوى و شكل صفحات الشبكة و المحتوى بصفة عامة .

الكلمات الدلالية : تحليل تدفق معلومات الشبكة في الشبكة الداخلية (الإنترانات) , تنقيب استعمال الشبكة , ملف التسجيل, نظام صنع القرار .

6. Magasin de données 39
7. Modélisation multidimensionnelle 39
7.1 Table de faits.....39
7.2 Table de dimension.....40
7.3 Schéma relationnels.....40
7.4 Schéma multidimensionnelle 43
8. Le concept OLAP.....43
8.1 Généralités.....43
8.2 Serveurs OLAP 44
9. Extraction,Transformation,Chargement « ETL ».....47
9.1 Définition ETL.....47
9.2 Décomposition ETL 47
9.3 Application des outils ETL a la BI.....48
10.Conclusion.....49

CHAPITRE 4 : Architecture et modèle pour un entrepôt de données fichier journal

1. Introduction 51
2. Le rôle de l'application 51
3. L'analyse de la source de données 51
4. Etat du décisionnel au sein de l'entreprise.....52
5. Définition des besoins.....52
6. Modélisation multidimensionnelle 53
6.1 Processus de modélisation.....53
6.2 Modélisation multidimensionnelle du processus d'activité.....54
7. Construction de la zone ETL 56
7.1 Architecture de l'alimentation.....56
7.2 La fréquence de l'alimentation 60
8. Conception des cubes OLAP 60
8.1 Définition des niveaux et des hiérarchies.....60
8.2 Présentation des cubes OLAP 61
9. Conclusion 62

CHAPITRE 5 : Mise en œuvre et implémentation

1. Introduction 64
2. Architecture technique de la solution 64
3. Architecture global de la solution 64
4. Construction de la zone de stockage.....65
5. Construction de la zone d'alimentation 65
6. Zone de restitution.....68
7. Fonctionnement de la solution conçue.....69
8. Les avantages et les caractéristiques de notre système.....71
9. Conclusion.....72

Conclusion generale.....74
Bibliographie 75
Webographie 78

Liste des figures

Figure 1.1 : Organigramme de l'ONS 13
Figure 2.1 : Le schéma générale d'un intranet 17
Figure 2.2 : Le processus du WUM 23
Figure 2.3 : Exemple d'un fichier log 24
Figure 2.4 : Exemple de log Squid 26
Figure 2.5 : Exemple de log Analog 26
Figure 2.6 : Log caritig 27
Figure 3.1 : Architecture d'un système d'information décisionnelle 34
Figure 3.2 : Structure des données d'un DW 37
Figure 3.3 : La relation entre DW et DM 39
Figure 3.4 : Modèle d'un système SID 40
Figure 3.5 : Exemple de modélisation en étoile 41
Figure 3.6 : Exemple de modélisation en flocon de neige 41
Figure 3.7 : Exemple de modélisation en constellation 42
Figure 3.8 : Exemple de schéma multidimensionnel 43
Figure 3.9 : Architecture ROLAP 44
Figure 3.10 : Architecture MOLAP 45
Figure 3.11 : Architecture HOLAP 46
Figure 3.12 : Schéma explicatif de l'utilisation d'outils ETL dans le SID 48
Figure 4.1 : Schéma en étoile de l'activité « Surveillance » 57
Figure 4.2 : Architecture du processus d'alimentation 57
Figure 4.3 : Fichier texte journal 58
Figure 4.4 : La table « logfile » 58
Figure 4.5 : Diagramme d'activité global du processus de chargement 60
Figure 4.6 : Cube multidimensionnelle « fichier log » 62
Figure 5.1 : Architecture technique de la solution 65
Figure 5.2 : Architecture globale du système décisionnelle 65
Figure 5.3 : Méthode de connexion a une base de données 67
Figure 5.4 : Méthode d'alimentation de la table de dimension Visiteur 68
Figure 5.5 : Alimentation de table de fait « FactLogFile » 68
Figure 5.6 : Construction du cube multidimensionnelle 69
Figure 5.7 : Exemple d'un rapport 70
Figure 5.8 : L'analyse multidimensionnel OLAP 71

Liste des tableaux

Tableau 2.1 : Quelques outils de mesure d'audience 20

Tableau 3.1 : Comparaison des deux systèmes 35

Tableau 3.2 : Différence entre SGBD et DW.....38

Tableau 4.1 : Descriptif des attributs de la dimension « DimDate » 55

Tableau 4.2 : Descriptif des attributs de la dimension « DimPage » 56

Tableau 4.3 : Descriptif des attributs de la dimension « DimVisiteur » 56

Tableau 4.4 : Niveaux des hiérarchies des dimensions 61

Liste des abréviations

BI	Business Intelligence
OLAP	On-Line Analysis Processing
SI	Système d'Information
SID	Système d'Information Décisionnel
DW	Data Warehouse
DM	Data Mart
SGBD	Système de Gestion de Base de Données
OLTP	On-Line Transactional Processing
MOLAP	Multidimensional On-Line Analysis Processing
ROLAP	Relational On-Line Analysis Processing
HOLAP	Hybrid On-Line Analysis Processing
SQL	Structured Query Language.
ETL	Extract, Transform, Load
JDBC	Java Data Base Connectivity
ECD	Extraction de la Connaissance des bases de données
BD	Base de données

« *Le succès consiste d'aller d'échec en échec*

sans perdre son enthousiasme »

-Winston Churchill-

INTRODUCTION GENERALE

INTRODUCTION GÉNÉRALE

1. Contexte :

Aujourd'hui, l'Intranet a pris une place considérable dans le bon fonctionnement de l'entreprise, la circulation de l'information est devenue une stratégie de communication interne. En effet, lorsqu'on a une bonne circulation de l'information, on favorise la communication et ça devient, de ce fait, facteur de cohésion, de motivation, de décision efficace et de créativité.

L'importance grandissante du rôle d'intranet dans le succès des entreprises amène les chercheurs et les responsables à s'intéresser à l'analyse des données générées par les différents types de serveurs, ils s'intéressent à connaitre et comprendre le comportement des intra-nautes et de l'adaptation du contenu.

L'analyse des données est essentielle pour connaître l'évolution et les tendances de visite d'un site Web. En effet, pour la majorité des sites, c'est la seule façon d'obtenir des données fiables en utilisant des outils d'analyse du trafic web.

L'analyse du trafic web sur le réseau intranet (intranet analyse) tout comme l'analyse du web est devenu une technique incontournable pour l'optimisation des sites web. Elle permet notamment de connaitre l'origine de son trafic et de prendre des décisions intelligentes.

Notre travail consiste à réaliser un outil d'aide à la décision « *intranet analytics* » qui consiste à collecter, mesurer, rapporter et analyser des données quantitatives provenant du web dans le but d'optimiser les sites et permettra aux analystes d'une entreprise utilisant un réseau intranet d'avoir une vue précise sur le trafic Web dans le réseau.

En effet, cet outil utilisera les données extraites à partir des fichiers journaux des serveurs Web Apache. Mais, ceux-ci doivent être traités selon les règles du Web Usage Mining afin d'extraire un maximum d'informations utiles.

Ainsi que les données obtenues seront stockées dans une base multidimensionnelle sous forme de cubes de données, ce type de bases offre l'avantage d'utiliser la technologie OLAP qui nous fournit des fonctionnalités très avancées pour la visualisation et l'analyse.

2. Problématique :

L'utilisation de l'intranet augmente s'accentue de plus en plus, ce qui oblige les administrateurs de vouloir analyser le trafic web sur leur réseau, cela permet de bien comprendre le comportement des utilisateurs et les actions vers le serveur, chaque action produit un fichier journal, ces fichiers textes listent chronologiquement les évènements exécutés.

Les fichiers journaux peuvent contenir des informations confidentielles (adresses IP, configuration du système, liste de processus...etc.), leur accès devrait être sécurisé. Pour cela, il faut limiter les droits d'accès à ces fichiers, ainsi que leur analyse peut se faire manuellement mais la plupart des administrateurs n'osent pas à les ouvrir car les logs ne sont pas aisés à déchiffrer.

3. Les objectifs :

Notre objectif principal consiste à réaliser un outil pour aider les administrateurs du serveur web apache pour analyser le trafic web sur leur réseau intranet, cela peut se faire en analysant le comportement des utilisateurs du site étudié. Plus en détail, mesurer le nombre de visiteurs, le nombre de visiteurs qui ont fait des changements sur le site ainsi que le nombre de visiteurs uniques et les pages visitées afin de pouvoir identifier les failles dans la sécurité et les accès non autorisés et optimiser le site. Il permet également d'établir des statistiques de connexions au serveur avec une interface ergonomique.

4. Organisation de notre mémoire :

Notre travail s'organise autour de cinq chapitres principaux :

Les trois premiers chapitres permettent de faire une présentation avec un tour d'horizon sur les différents concepts théoriques liés à notre travail.

- Le premier chapitre est consacré à présenter l'organisme d'accueil.
- Dans le deuxième chapitre, nous allons présenter la fonctionnalité de l'intranet et une présentation des règles du Web Usage Mining, ainsi que les travaux de recherche sur ce domaine.
- Le troisième chapitre, nous allons présenter le système d'information décisionnel.

Au niveau des deux derniers chapitres, nous essayons à expliquer les différentes étapes nécessaires à la mise en œuvre de notre solution.

- En ce qui concerne le quatrième chapitre, nous allons appliquer le système décisionnel pour l'analyse du fichier journal, nous présentons ainsi la conception de notre travail, les différentes étapes nécessaires à l'implémentation de notre conception comme (le prétraitement, le nettoyage, l'exploration et l'analyse du fichier journal), sauvegarder et restituer des données dans une base de données multidimensionnelle et enfin l'analyse OLAP (création et utilisation des cubes).
- Et dans le dernier chapitre nous présentons l'implémentation et la mise en œuvre de notre application avec les différents outils utilisés.

CHAPITRE 1

PRESENTATION DE L'ORGANISME D'ACCEUIL

CHAPITRE 1 : PRÉSENTATION DE L'ORGANISME D'ACCEUIL

Office National des Statistiques

1. Introduction :

L'Office national des statistiques (ONS) est le service officiel des statistiques en Algérie, créé au lendemain de l'indépendance, en 1964. Il a le statut d'établissement public à caractère administratif.

L'Office National des Statistiques est l'Institution Centrale des Statistiques de l'Algérie. C'est un établissement public à caractère administratif chargé de la collecte, du traitement et de la diffusion de l'information statistique socio-économique (tel que recensement de la population et de l'habitat, enquête sur la main d'œuvre, enquête sur les entreprises industrielles, etc...).

L'Office National des Statistiques est placé sous la tutelle du ministère de la Prospective et des Statistiques [2].

2. Historique :

L'Office National des Statistiques fut créé au lendemain de l'indépendance, en 1964, sous l'appellation de Commissariat National pour le Recensement de la Population (C.N.R.P) et ceci afin de réaliser le premier recensement de la population de l'Algérie indépendante en 1966. En 1971, il change de dénomination et devient Commissariat National aux Recensements et Enquêtes Statistiques (C.N.R.E.S). De grands travaux ont été réalisés pendant cette période tels que : le deuxième recensement de la population et de l'habitat en 1977 ; l'enquête démographique en 1972-1973 ; l'enquête cartographique en 1972-1975 qui devrait servir de base à la réalisation du recensement, et l'enquête sur la consommation des ménages en 1979-1980. Par ailleurs, une réorganisation de l'appareil statistique a donné naissance à l'actuel Office National des Statistiques par le biais du décret législatif **N° 82-484 du 18/12/1982** complété et modifié par le décret **N° 85-311 du 17/12/1985**. L'O.N.S est alors chargé de l'organisation et la coordination des travaux statistiques. De grandes enquêtes ont été réalisées, parmi ces dernières on citera le recensement de la population et de l'habitat de 1987, les enquêtes annuelles auprès des ménages de 1982 à 1992, les enquêtes annuelles auprès des entreprises,.... Enfin, le décret **N° 95-159 du 03/06/1995** a donné lieu à une nouvelle réorganisation de l'Office National des Statistiques [2].

3. Le système National statistique :

Le système national d'information statistique a été réorganisé par le décret législatif N° 94-01 du 15 Janvier 1994, qui définit les principes généraux et fixe le cadre organisationnel ainsi que les droits et obligations des personnes physiques et morales dans les domaines de la production, la conservation, l'utilisation et la diffusion de l'information statistique. Ainsi, toute information quantitative ou qualitative permettant la connaissance des faits économiques sociaux et culturels par des procédés numériques est considérée comme une information statistique. Suivant le principe de la liberté d'information, toute personne physique ou morale a la faculté de produire, traiter et diffuser l'information statistique.

Cependant ne relève du domaine public que l'information statistique qui aura été élaborée par les services de l'Etat ou qui aura bénéficié de l'enregistrement statistique. Au terme du décret législatif cité ci-dessus, "L'enregistrement statistique est la reconnaissance par l'Etat du caractère d'intérêt public des enquêtes, études et travaux statistiques.

A ce titre, elle est accessible à tout demandeur. Sans préjudice des procédures juridiques et administratives, sa rétention peut faire l'objet pour son obtention, d'un recours. Par ailleurs, dans le cadre du secret statistique, le décret législatif précise que les renseignements individuels figurant sur les questionnaires revêtus de l'enregistrement statistique et ayant trait à la vie personnelle et familiale ne peuvent faire l'objet de communication de la part du service dépositaire ou de publication que conformément à la loi sur les archives nationales. Les renseignements individuels ne peuvent en aucun cas être utilisés à des fins de contrôle fiscal, de répression économique, d'enquêtes judiciaires, d'atteinte à la vie privée des personnes, ou de concurrence.

4. Le conseil national de la statistique :

Le conseil national des statistiques est chargé de l'élaboration de la politique nationale la statistique et de l'information économique, de la coordination de l'élaboration et du contrôle d'exécution des programmes nationaux, sectoriels et spécifique de travaux statistiques conforme à la politique nationale arrêtée en la matière, de se prononcer et d'arrêter les méthodes, procédures et modalités de calcul et composition de tous les indices, indicateurs, agrégats et comptes servant de référence officielles. En plus, de veiller à la garantie effective du secret statistique ainsi qu'au strict respect de l'obligation statistique, de veiller à la promotion de la circulation de l'information statistique et au perfectionnement permanent des circuits assurant la disponibilité d'informations fiables, régulières et adaptées aux besoins des agents socio-économiques. Il peut être crée auprès du conseil un ou plusieurs comités permanents investis de missions définies par leur texte de création. Le conseil est habilité à recourir à toute compétence ou expertise extérieures au conseil.

5. Fonctions de l'Office National des Statistiques :

Aux termes du décret législatif 94-01 du 15/01/1994, les prérogatives de l'Office National des Statistiques ont été reconduites et élargies.
C'est ainsi que l'Office National des Statistiques veille à l'élaboration, la disponibilité et à la diffusion d'informations fiables, régulières et adaptées aux besoins des agents économiques et sociaux.
Il assure ou fait assurer la disponibilité régulière des données, analyses statistiques et études économiques nécessaires à l'élaboration et au suivi de la politique économique et sociale des pouvoirs publics.
Il élabore et diffuse régulièrement, en application du programme national statistique, indices, indicateurs de l'économie nationale ainsi que les comptes de la nation.
Il gère les enregistrements statistiques des enquêtes et travaux statistiques, tient et met à jour un répertoire des agents économiques et sociaux auxquels est attribué le **Numéro d'Identification Statistique (NIS)** [2].

6. Organisation interne de l'Office National des Statistiques :

Le Décret Législatif 94-01 du 15 Janvier 1994 relatif au système statistique identifie les organes de production et de coordination du système d'information statistique qui sont :

- Le Conseil National de la Statistique (C.N.S.) ;
- L'Institution Centrale des Statistiques,
- Les services statistiques des administrations et des collectivités territoriales,
- Les organes publics et privés spécialisés, dont les instituts de sondages statistiques.

L'Office National des Statistiques est l'Institution Centrale des Statistiques prévue à l'article 11 du Décret cité ci-dessus.

L'Office est un établissement public national disposant de services centraux et de structures régionales.

Le cadre d'intervention de l'O.N.S. s'étend de la participation à l'élaboration du rapport annuel sur l'exécution du plan national et aux projets de plans et programmes nationaux de travaux statistiques dont l'Office a la charge de superviser les travaux d'élaboration technique à la réalisation, l'exploitation et l'analyse des recensements et enquêtes statistiques, ainsi que la mise en place de fichiers et bases de données dont il assure la gestion.

En outre, l'ONS est chargé de la diffusion et de la promotion de l'information statistique. Cette mission, l'Office l'exerce à travers :

- Une série de publications diversifiées à savoir :

  - Annuaire statistique de l'Algérie
  - Algérie en quelques chiffres
  - Note de Conjoncture
  - Publications périodiques (Revues, Bulletins, Collections etc...)
  - Séries Statistiques (Revue Rétrospective)

Il aura notamment pour fonction de promouvoir le système national d'information statistique en veillant aux règles et méthodes générales d'élaboration, de révision et de mise à jour des codes, nomenclatures, fichiers et concepts statistiques, à la disponibilité et à la diffusion d'informations fiables, régulières et adaptées, et ce, pour les besoins des agents économiques et sociaux.

Comme, il est tenu de réaliser à la demande du Gouvernement ou de tout autre service de l'Etat, tous travaux entrant dans sa mission.

Dans le cadre de la mise en place des instruments et procédures, il est tenu de mettre à jour un répertoire des agents économiques et sociaux auxquels est attribué le numéro d'identification statistique.

En application des dispositions du Décret 95-159 du 3 Juin 1995 portant réaménagement des statuts de l'Office National des Statistiques et de l'Arrêté interministériel n°620 les 26.11.1995 portantes organisations interne de l'Office, l'administration de l'ONS sous l'autorité du Directeur Général qui est assisté dans l'ensemble de ses fonctions de :

- D'un Directeur Général Adjoint

- D'un Directeur Technique chargé de la Comptabilité Nationale

- D'un Directeur Technique chargé des Statistiques des Entreprises et du Suivi de la Conjoncture

- D'un Directeur Technique chargé des Statistiques Sociales et des Revenus

- D'un Directeur Technique chargé des Statistiques de la Population et de l'Emploi

- D'un Directeur Technique chargé des Statistiques Régionales et de la Cartographie

- D'un Directeur Technique chargé des Traitements Informatiques et des Répertoires Statistiques.

- D'un Directeur chargé de l'Inspection

- D'un Directeur Chargé du Secrétariat Technique du CNS.

- D'un Directeur des Publications de la Diffusion, de la Documentation et de l'Impression.

- D'un Directeur de l'Administration et des Moyens

- De quatre (04) Directeurs d'Annexes Régionales.

Le Directeur Général Adjoint est chargé d'assister le Directeur Général pour la conduite et la coordination des travaux techniques. Le Directeur Général peut lui déléguer en tant que de besoins de certaines missions sans préjudice des attributions des Directeurs Techniques et des Directeurs de l'Office.

Le Directeur Technique chargé de la comptabilité Nationale a pour mission :

- D'élaborer périodiquement les comptes économiques de la Nation. Dans ce cadre, il collecte toute information statistique concernant les secteurs réels, financier et social et toute donnée nécessaire à l'accomplissement de sa mission.

- De contribuer à l'élaboration des rapports d'évaluation de l'exécution des programmes et plans nationaux de développement.

- D'élaborer et de mettre à jour les instruments méthodologiques utilisés pour l'établissement des comptes nationaux.

Le Directeur est assisté de quatre (04) Chefs d'Etudes et de Cinq (05) Chefs de projets.

Le Directeur Technique chargé des Statistiques des Entreprises et du Suivi de la Conjoncture a pour mission :

- de collecter, traiter et analyser les informations statistiques relatives à la production de biens et services.

Dans ce cadre, il réalise les enquêtes statistiques appropriées auprès des entreprises.

A ce titre, il contribue au développement des méthodes statistiques en rapport avec son domaine d'activité et en relation avec les administrations et organismes concernés.

- De participer pour ce qui le concerne, à l'élaboration des comptes nationaux ;

- d'établir les différents indices de production, de coût et de prix des biens et services ;

- de suivre et d'analyser la conjoncture économique à partir d'indicateurs issus des résultats des enquêtes statistiques et de données de source administrative ;

- de participer à l'élaboration des publications de l'ONS en rapport avec ses attributions.

Le Directeur est assisté de trois (03) Chefs d'Etudes et trois (03) Chefs de projet.

Le Directeur Technique chargé des Statistiques Sociales et des Revenus a pour mission :

- De collecter, traiter et analyser les informations statistiques relatives aux revenus, aux statistiques sociales à la consommation et aux conditions de vie des ménages ;

Dans ce cadre, il réalise les enquêtes auprès des ménages et des agents économiques concernés ;

- de contribuer au développement des méthodes statistiques dans son domaine ;

- de participer à l'élaboration de comptes nationaux particulièrement le compte des ménages ;

- d'établir périodiquement les indicateurs sociaux appropriés.

Le Directeur est assisté de deux (02) Chefs d'Etudes et de trois (03) Chefs de Projet.

Le Directeur Technique chargé de la Population et de l'Emploi a pour mission :

- De préparer et de coordonner la réalisation et exploitation des recensements de la population et de l'habitat ;

- de collecter, traiter et analyser l'information démographique ;

- d'exploiter les données sur les faits d 'Etat-Civil pour les analyses démographiques et établir périodiquement la situation démographique du pays ;

- d'élaborer les prévisions démographiques ;

- contribuer à l'élaboration des publications sur les recensements de population et les données démographiques ;

- d'élaborer les prévisions démographiques ;

- contribuer à l'élaboration des publications sur les recensements de population et les données démographiques ;

- d'assurer la conservation spécifique des archives des recensements de population conformément à la réglementation en la matière ;

- de réaliser ou participer à la réalisation des enquêtes auprès des ménages et des établissements en vue d'élaborer les statistiques de population active et d'emploi ;

- de participer à l'élaboration des publications en rapport avec sa mission.

Le Directeur est assisté de trois (03) Chefs d'Etudes et de trois (03) Chefs de Projet.

Le Directeur Technique chargé des Statistiques Régionales, des Statistiques Agricoles et de la Cartographie a pour mission :

- de collecter ou recueillir auprès d'autres organisations ou institutions méthodologiques, les informations statistiques régionales ;

- de recueillir les statistiques agricoles et hydrauliques et de veiller au bon fonctionnement du système statistique agricole ;

- de réaliser les travaux cartographiques pour les recensements et enquêtes statistiques ;

- d'assurer la visualisation cartographique de l'information statistique.

Le Directeur est assisté de deux (02) Chefs d'Etudes et de deux (02) Chefs de projet.

Le Directeur Technique chargé des Traitements Informatiques et des Répertoires Statistiques a pour mission :

- de développer les applications et logiciels informatiques pour les besoins de l'ONS ;

- de gérer le répertoire national des entreprises et des établissements ;

- de gérer le numéro d'identification statistique (NIS) ;

- de promouvoir le développement des banques de données statistiques ;

- d'assurer ou faire assurer la maintenance des matériels informatiques de l'Office et veiller à leur utilisation rationnelle, en relation avec les autres directeurs ou directeurs d'annexes régionales.

Le Directeur est assisté de trois (03) Chefs d 'Etudes et de trois (03) Chefs de Projet.

Le Directeur des Publications, de la Diffusion, de la Documentation et de l'Impression a pour mission :

- d'élaborer et diffuser les publications de l'Office; notamment les indices et indicateurs économiques et sociaux, les comptes de la nation, annuaires statistiques et les publications du Recensement Général de la Population et de l'Habitat ;

- de gérer le centre de documentation du siège de l'Office et de contribuer au développement de la documentation statistique dans les annexes régionales, en liaison avec les centres de documentation nationaux et régionaux des administrations et organismes publics ;

- d'assurer l'impression des documents et publications de l'Office ;

- de coordonner, conformément à la réglementation, les échanges d'information avec les organismes et institutions internationaux ;

- d'assurer la gestion des archives de l'Office relatives aux publications.

Le Directeur est assisté de trois (03) Sous-Directeurs :

- Le Sous-Directeur de la publication, de l'annuaire et des revues statistiques, assisté d'un Chef de Projet ;

- Le Sous-Directeur de la diffusion, de la documentation et des archives, assisté d'un Chef de Projet ;

- Le Sous-Directeur de l'Impression, assisté d'un Chef de Projet.

Le Directeur de l'Administration et des Moyens a pour mission :

- d'élaborer en liaison avec les autres structures le projet de budget de l'Office ;

- d'exécuter par délégation du Directeur Général le budget et tenir la comptabilité de l'Office ;

- d'établir les projets de marchés et les exécuter après approbation ;

- d'assurer ou faire assurer la maintenance, l'entretien et l'approvisionnement en matériel, fourniture et équipement de l'Office ;

- de gérer le personnel, établir et mettre en œuvre les plans de recrutement et de formation du personnel en relation avec les autres structures de l'Office ;

- de veiller à la discipline du travail et à la sécurité des biens et des personnes de l'O.N.S.;

- de veiller à l'application de la réglementation en matière d'archives.

Le Directeur est assisté de trois (03) Sous-Directeurs :

- Le Sous-Directeur du personnel et de la Formation, assisté :

- d'un Chef de Bureau de la gestion du personnel et de l'action sociale ;

- d'un Chef de bureau de la formation et du perfectionnement.

Le Sous-Directeur du budget et des marchés, assistés :

- d'un Chef de Bureau de l'exécution et du suivi des opérations de fonctionnement ;

- d'un Chef de Bureau des marchés et du suivi des opérations d'équipement.

- Le Sous-Directeur des Moyens Généraux, assisté :

- d'un Chef de Bureau des inventaires, et des stocks ;

- d'un Chef de Bureau des moyens matériels.

Le Directeur Chargé du Secrétariat Technique du Conseil National de la Statistique (CNS) a pour mission de :

- de préparer les réunions du CNS et de ses formations spécialisées le cas échéant notamment d'assurer la préparation des projets de programmes nationaux de travaux statistiques soumis à son examen ;

- de coordonner l'établissement des règles et instruments de normalisation et de méthodologie statistique à soumettre au CNS ;

- de suivre l'application des décisions et avis du CNS ;

- d'assurer le secrétariat technique des réunions du CNS ;

- de gérer l'attribution du numéro d'enregistrement des enquêtes statistiques considérées d'utilité publique par le CNS.

Le Directeur est assisté d'un (01) Chef d'Etudes.

Le Directeur Chargé de l'Inspection a pour mission :

- d'effectuer conformément à un programme approuvé par le Directeur général des inspections dans les structures de l'Office et de ses annexes pour vérifier l'état d'application de la réglementation et des décisions ;

- de contribuer à prévenir et à résoudre les conflits de travail en relation avec le Directeur de l'Administration et des Moyens et les représentants des travailleurs ;

- d'évaluer le degré d'efficience dans l'utilisation des moyens mis à la disposition de l'Office ;

- de faire rapport au Directeur Général sur le résultat des inspections effectuées.

Le Directeur chargé de l'inspection doit établir au moins un rapport annuel ; il est assisté d'un Chef d'Etudes.

Le Directeur Général est en outre assisté, d'un Chef d'Etudes chargé :

- du suivi des programmes de coopération avec les institutions similaires étrangères et internationales ;

- de la préparation des réunions du Conseil d'Orientation ;

- du suivi des relations avec les organes de communication.

Les Quatre (04) annexes régionales de l'Office Nationale des Statistiques ont leur siège respectivement à Alger, Constantine, Oran et Ouargla.

Le Directeur d'Annexe Régionale a pour mission :

- d'exécuter au niveau régional concerné le programme de travail de l'ONS, notamment des enquêtes et recensements prévus ;

- de diffuser au niveau régional les publications de l'ONS ;

- de contribuer à la réalisation des publications de l'ONS relatives à la région concernée ;

- d'assister les administrations locales en matière de travaux statistiques dans le cadre d'un programme approuvé par la Direction Générale et conformément aux dispositions de l'article 6 du décret n° 95-159 du 03 Juin 1995 susvisé ;

- de veiller à la discipline du travail et à la sécurité des biens et des personnes affectées à l'annexe régionale.

Le Directeur d'annexe est assisté d'un (01) Chef d'Etudes, de deux Chefs de Projet et d'un (01) Chef de Bureau des Moyens.

7. Organigramme de l'ONS [2] :

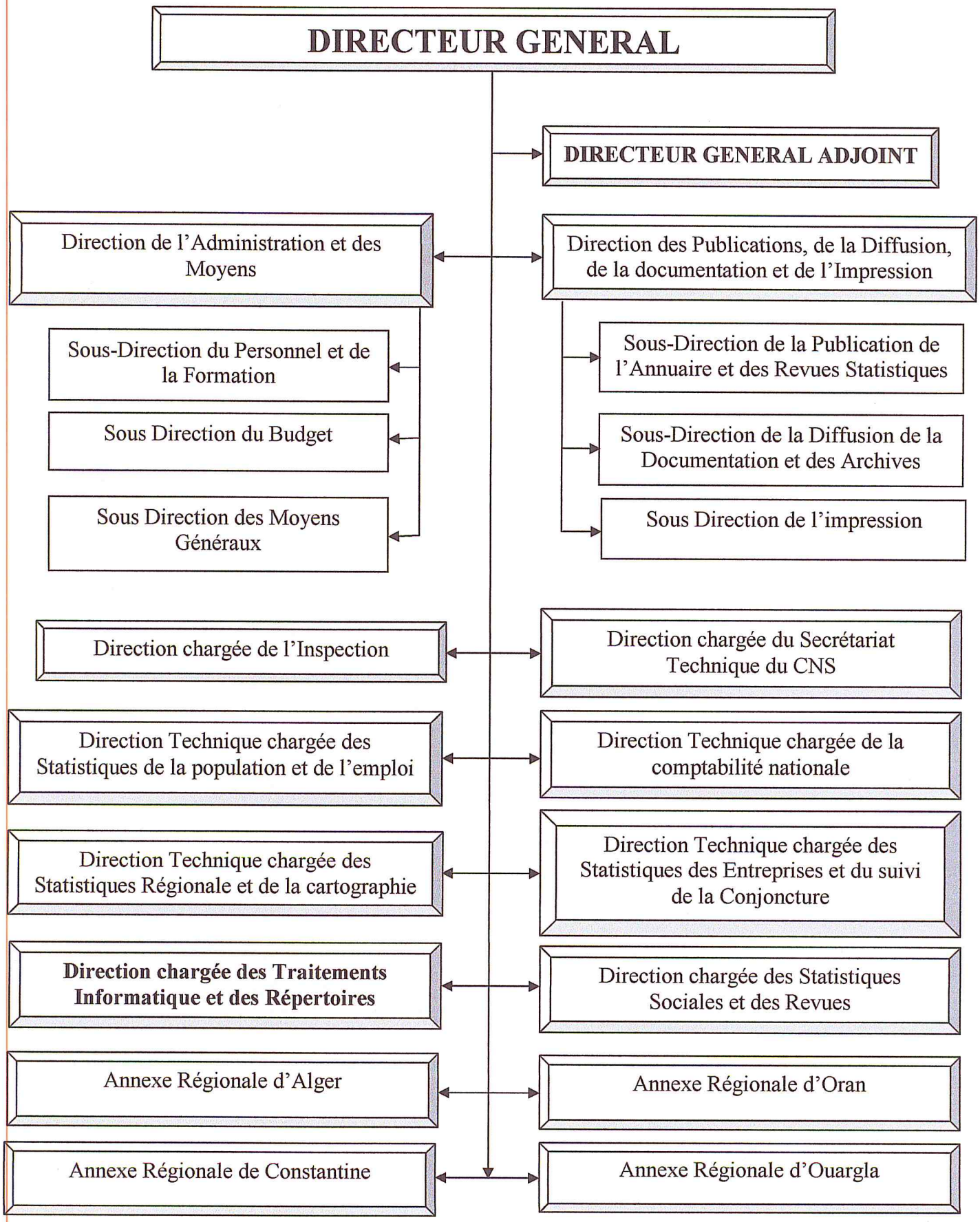


Figure 1.1 : Organigramme de l'ONS

8. Conclusion :

Ce chapitre a été consacré à donner une présentation précise de cet important office national ainsi de certaines de ses fonctions et celle de la **Direction chargée des Traitements Informatique et des Répertoires.**

CHAPITRE 2

ANALYSE DU TRAFIC WEB SUR UN RESEAU INTRANET

➢ **L'intranet pour communiquer :**

En matière de communication, les besoins se sont également précisés :

- Rechercher une personne sur un annuaire par son nom.
- Communiquer par messagerie avec tout le personnel sans exception.
- Pouvoir gérer ses congés en ligne : demande de congé, ou de récupération, obtenir une réponse, consulter son congé, valider la demande.
- Pouvoir s'exprimer et échanger sur un sujet dans un forum interne.

➢ **Le schéma général** [5] **:**

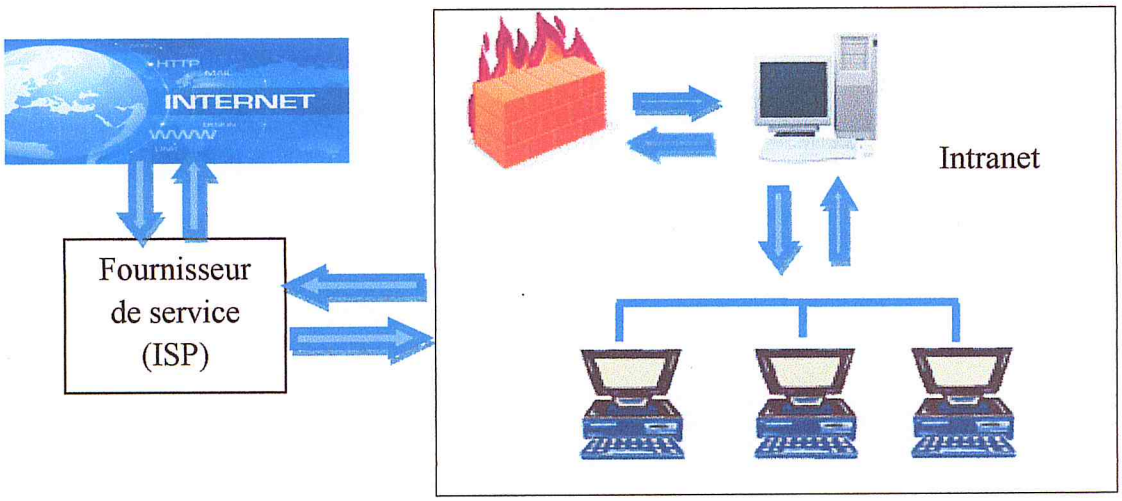


Figure 2.1 : Le schéma général d'un intranet

4. Les fonctionnalités de l'intranet :

Selon **Frédéric.C et Thomas.J** les outils essentiels de l'intranet sont : Portail, gestion de contenu, partage documentaire et travail collaboratif [8].

➢ **Portail :**

La dimension « portail » fait de l'intranet la clé de voûte, la porte d'entrée du système d'information de l'entreprise.

➢ **La gestion de contenu :**

La gestion de contenu regroupe l'ensemble des fonctionnalités permettant la création, la validation, la mise à jour et la présentation d'informations sous forme de pages Web.

➢ **La gestion documentaire :**

La gestion documentaire regroupe l'ensemble des fonctionnalités permettant l'acquisition, le stockage, le classement, le partage, la consultation et l'archivage de documents sous forme de fichiers informatiques (le plus généralement au format PDF, Microsoft Office, OpenOffice).

➢ **Les espaces de travail collaboratif :**

Les espaces collaboratifs ont pour cible des communautés. Une communauté peut se créer soit autour d'un projet, soit autour d'un thème, d'un métier ou d'une pratique. Les espaces de projet offrent aux membres d'une équipe de projet une plateforme de collaboration, d'échanges et de capitalisation.

5. Les avantages de l'intranet [5] :

Mais, pourquoi utiliser un Intranet d'entreprise ? Quels sont les avantages de la mise en place d'une telle solution ?

- **Gérer l'information**

Un Intranet nous permettra de gérer et de stocker nos informations. Celles-ci sont ainsi accessibles et structurées automatiquement.

- **Collaboration**

Le but principal reste la collaboration entre les différentes équipes de l'entreprise. Elles partagent donc la même information au même moment.

- **Centraliser l'information**

L'information étant centralisée dans un même espace, on ne perde plus une minute à chercher un dossier.

- **Temps réel**

Tout se passe en temps réel, tout le monde travaille ainsi sur des données à jour évitant donc les pertes de temps.

- **Mobilité**

Les informations étant accessibles depuis n'importe quel poste.

- **Sécurité**

La sécurité ainsi que l'intégrité des données sont entièrement prises en charge. On se consacre ainsi pleinement à notre travail.

- **Illimité**

L'utilisation étant illimitée, on travaille sans limite. Plus besoin de se soucier de la place que prennent nos dossiers.

- **Structure**

L'information étant immédiatement structurée, on y accède simplement et rapidement.

- **Les emails**

Nous recevons tous un nombre considérable d'email chaque jour. On les limite puisque les collaborateurs accèdent à l'information directement en ligne.

- **Discussion**

L'intranet offre un espace de discussion nous permet ainsi de mieux s'informer et de collaborer. Donc, un Intranet d'entreprise apporte de nombreux avantages.

6. L'analyse du trafic web sur le réseau intranet :

Le terme trafic web fait référence à la circulation des flux d'information sur un réseau informatique [40].

6.1 Définition de l'analyse du trafic web :

L'analyse du trafic web ou mesure de l'audience permet de quantifier la fréquentation d'un site web, réseau intranet ou un réseau social en fonction d'indicateurs tels que le nombre de visiteurs uniques, les pages vues, les visites etc. Elle regroupe la mesure, la collecte, l'analyse et la présentation de données provenant d'Internet utilisées afin de comprendre et d'optimiser l'utilisation du Web [39].

6.2 Pourquoi la mesure de l'audience

Pour tout site professionnel, l'analyse de trafic est un impératif. Il est vivement recommandé de prendre quelques dizaines de minutes par jour pour consulter les statistiques de fréquentations : nombre de visites, nombre de pages vues, taux de rebonds, pages vues par visites …etc.

La mesure d'audience intranet répond à la nécessité d'identifier précisément quels sont les services dont l'utilisation ne décolle pas. Au-delà, il s'agit pour l'entreprise de concentrer ses efforts sur les services intranet qui répondent véritablement à un besoin.

Sans même parler d'objectifs marketings et financiers à contrôler, il y'a plusieurs raisons pour justifier un suivi quotidien :

- Suivre l'évolution globale du trafic du site, d'un point de vue quantitatif et qualitatif. Nombre de visites (visiteurs uniques), origine du trafic (sources), nouvelles visites, etc.
- Optimiser le contenu du site : en analysant certains indicateurs, comme le taux de rebond et le nombre de pages vues/visiteur.
- Assigner des objectifs à certaines pages du site : il est possible de surveiller à quel endroit les visiteurs entrent dans le chemin menant vers l'objectif et à quel endroit ils le quittent, afin de corriger éventuellement ce chemin.

6.3 Comment mesurer l'audience ?

Deux approches technologiques sont aujourd'hui utilisées :

- La technologie dite des marqueurs TAG : Ce système est différent et nécessite une intervention technique sur les pages du site web. Il s'agit de placer un petit programme sur les pages pour lesquelles on souhaite mesurer le nombre de visites [39].
- L'analyse selon les règles de la fouille de données d'usage du Web : extraire les données à partir des fichiers journaux.

La mesure passe habituellement par l'utilisation d'outils de mesure spécifiques. Ces outils étaient à l'origine destinés aux sites Web. Ils utilisent des marqueurs ou les logs.

6.4 Les outils existent d'analyse d'audience :

Les outils utilisés pour l'analyse d'audience sur un intranet sont les mêmes que ceux utilisés pour les sites web. Analyse du navigateur utilisé, identification rapide des liens affichés, la plupart de ces outils son cher, ils sont basé sur le marquage et les tags, ils ne mesurent pas tous la même chose.

Outil	Principale caractéristique	Lien
Woorank web Analytics	Visibilité sur une page des performances globales du site.	http://woorank.com/
Alexa	Analyse du trafic entrant.	http://lexa.com/
Compete web analytics	Spécialisé dans l'analyse du trafic des campagnes publicitaires. Analyse le trafic en provenance des États-Unis uniquement.	http://compete.com/
Piwik	Nombreuses fonctionnalités d'analyse. À installer sur son poste	http://piwik.com/

	de travail.	
Crazy egg	Approche atypique. Analyse des zones de l'écran qui génèrent des clics.	http://crazyegg.com/
Optimizely	Spécialisé dans l'étude comparative des performances des *landing pages* (page d'entrées) des sites de e-commerce.	http://optimizely.com/
Google Analytics	*Made by Google.*	www.google.com/intl/fr/analytics/
KISSinsights	Insertion et analyse de questionnaires visiteurs dans les pages du site.	http://kissinsights.com/
Clicktale	Outil complet. Analyse en temps réel.	http://www.clicktale.com/
Wordstream	Analyse du trafic moteur. Utile pour l'optimisation de campagnes type *adWords*.	www.wordstream.com/

Tableau 2.1 : Quelques outils de mesure d'audience [41]

7. la fouille de données d'usage du Web (WUM) :

Notre solution pour mesurer l'audience est basée sur les règles de la fouille de données, cette approche essentiellement présente de nombreux avantages de cette technologie qui rend attrayant pour les sociétés, compris les organismes gouvernementaux, qui utilisent cette technologie pour classer les menaces et la lutte contre le terrorisme, et l'identification des activités criminelles.

Les entreprises peuvent établir de meilleures relations avec la clientèle en leur offrant exactement ce dont ils ont besoin. Ils peuvent comprendre les besoins de la clientèle et mieux ils peuvent réagir plus rapidement aux besoins des clients (salariés). Ils peuvent attirer et retenir les clients, ils peuvent économiser sur les coûts de production en utilisant la connaissance acquise des besoins des clients. Ils peuvent augmenter la rentabilité de la cible de tarification basée sur les profils créés. Ils peuvent même trouver le client qui pourrait à défaut d'un concurrent de l'entreprise va essayer de garder le client en fournissant à des offres promotionnelles spécifiques du client, réduisant ainsi le risque de perdre un client [22].

7.1 Définition du WUM :

La fouille de données d'usage du Web (Web Usage Mining (WUM), en anglais) désigne l'ensemble de techniques basées sur la fouille de données pour analyser l'usage d'un site Web [10]. En d'autres termes, le WUM correspond au processus d'Extraction de Connaissances dans les Bases de Données (ECD) appliqué aux données d'usage du Web [11].

7.2 Motifs du WUM :

D'après **C.Michel** [12], il y a cinq motifs du WUM :

- Évaluation et caractérisation générale de l'activité sur un site Web : l'objectif est l'observation et non pas la modélisation. Les techniques d'analyse utilisées sont souvent simples. Elles relèvent, en effet, du dénombrement et des statistiques simples (moyennes, histogramme, indices, tris croisés).
- Amélioration des modes d'accès aux informations : le WUM permet de comprendre comment les utilisateurs se servent d'un site, d'identifier les failles dans la sécurité et les accès non autorisés.
- Modification de la structure : le WUM peut révéler le besoin de restructurer des pages et des liens afin d'améliorer la structure du site Web. En effet, les pages considérées comme similaires par des techniques de classification peuvent être reliées de manière hypertextuelle.
- Personnalisation de la consultation : cet enjeu important pour de nombreuses applications Internet ou sites de e-commerce consiste à proposer des recommandations dynamiques à un utilisateur en se basant sur son profil et une base de connaissances d'usages connus.
- Mise en œuvre de l'intelligence économique : cet objectif concerne en particulier les sites marchands. Il s'agit de comprendre quand, comment et pourquoi l'utilisateur est attiré par ce site, les produits qu'il faut lui proposer à la vente...etc.

7.3 Les phases du WUM :

Selon **C.Michel** [12] le processus général de WUM est constitué de trois phases principales : prétraitement de données (Preprocessing), Extraction des données (Usage mining) et analyse du modèle (Pattern Analysis), comme le montre le schéma suivant :

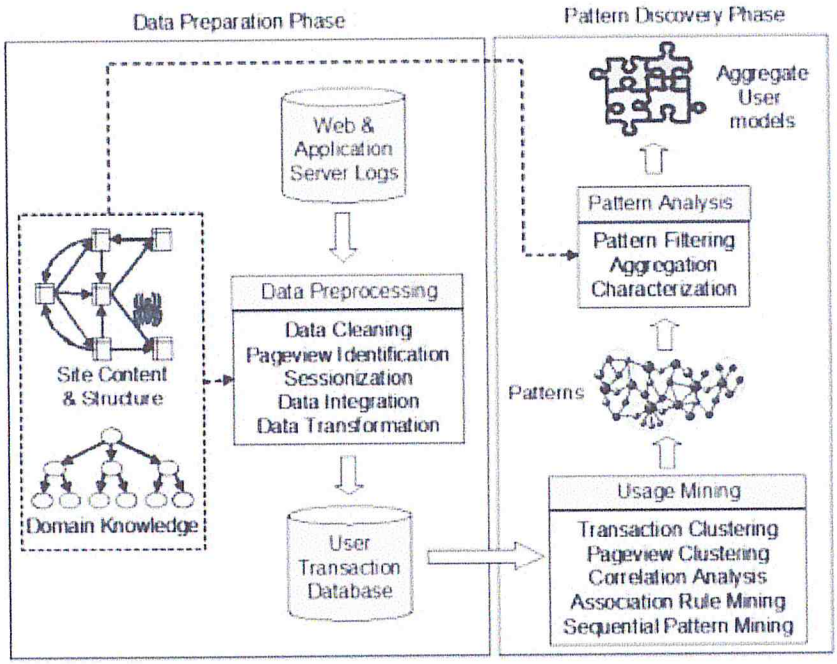


Figure 2.2 : Le processus du WUM [13]

➢ **Le prétraitement** [14] :

La première étape d'un processus WUM se compose principalement de deux types de tâches :

- Tâches classiques de prétraitement : fusion des fichiers logs web, nettoyage et structuration de données.
- Tâches avancées de prétraitement : stockage des données structurées dans une base de données (notée BD par la suite), généralisation et agrégation des données. A la fin de l'étape de prétraitement les données initiales sont donc nettoyées et structurées, en général, dans une BD.

➢ **Fouille (Extraction) de données** [14] :

Une fois les données transformées en séquences d'URLs (sessions, visites ou épisodes) ou actions, les méthodes classiques de la fouille des données peuvent être appliquées.

➢ **Analyse et interprétation** [14] :

Les résultats d'un processus de fouille de données sont, par définition, impossibles à estimer à l'avance, car l'objectif d'un tel processus est de découvrir des "nouvelles connaissances". De même, la quantité et la qualité des résultats ne peuvent être contrôlées que dans une certaine mesure.

Dans l'étape d'analyse et d'interprétation qui clôt le processus WUM, les patrons de visite découverts par la fouille de données doivent être analysés (idéalement de façon automatique) afin de garder les plus pertinents.

7.4 Les problèmes du WUM :

L'utilisation du web usage mining est un domaine de recherche récent de Data Mining. Le développement rapide et dynamique du WUM a laissé un peu de temps pour les analystes de comprendre et de résoudre les problèmes qui se posent dans ce domaine. Selon **Doru Tanasa** [19] les principaux problèmes sont :

- La quantité de données ne cesse d'augmenter ;
- L'étape de prétraitement ne reçoit pas assez d'efforts d'analyse ;
- Les sites Web ont peu ou pas de définitions sémantiques pour leurs pages Web ;
- Les techniques d'extraction de motifs séquentiels pour WUM ne sont pas appropriées pour faire face avec les spécificités de données sur l'utilisation du Web, principalement avec la quantité énorme ;
- Les trois étapes du processus de WUM ne sont pas coordonnées pour créer un ensemble cohérent et processus unique.

7.5 Fichier journal (Log) :

Les données exploitées dans le WUM proviennent des fichiers Logs [15].
Les analyses globales sont basées sur le fichier «journal» ou « log », qui est un fichier texte enregistré sur le site serveur, dans lequel une ligne est écrite pour chaque demande de l'utilisateur (par exemple pour un changement de la page, ou pour le téléchargement d'un fichier). Il existe plusieurs formats de fichiers log, y compris journal commun Format (CLF), le format de journal étendu (XLF), le format de journal commun contient l'adresse IP de l'ordinateur client, la date et l'heure de la demande, du type de requête, l'URL demandée, le protocole HTTP, le retour du serveur [16].

7.5.1 Extrait d'un fichier log :

```
127.0.0.1 - - [23/Jun/2009:15:12:57 +0100] "GET /Ministeres.html HTTP/1.1" 200 20284
127.0.0.1 - - [23/Jun/2009:15:12:57 +0100] "GET /IMG/siteon0.png HTTP/1.1" 200 341242
192.168.4.38 - - [24/May/2011:11:37:10 +0100] "POST /ecrire/?exec=legender HTTP/1.1"
302 375
```

Figure 2.3 : Exemple d'un fichier log

Selon ce format six informations sont enregistrées :

- Le nom du domaine ou l'adresse de Protocole Internet (IP) de la machine appelante ;
- Le nom et le login HTTP de l'utilisateur (en cas d'accès par mot de passe) ;
- La date et l'heure de la requête ;

- La méthode utilisée dans la requête (GET, POST, etc.) et le nom de la ressource Web demandée (l'URL de la page demandée) ;
- Le statut de la requête i.e. le résultat de la requête (succès, échec, erreur, etc.) ;
- La taille de la page demandée en octets.

Quelques explications sont nécessaires concernant le type de demande et le code de retour. Les principales valeurs du type de demande sont [16] :

- GET : télécharger un objet à partir du serveur ;
- PUT : stocker un élément sur le serveur ;
- DELETE : supprimer un élément à partir du serveur ;
- HEAD : très similaire à la méthode GET. Cependant le serveur ne retourne que l'entête de la ressource demandée sans les données. Il n'y a donc pas de corps de message ;
- POST : est utilisée pour envoyer des données au serveur.

Les principales valeurs du code de retour sont :

- 200 : demande complètement réussi ;
- 2xx : demande partiellement réussi ;
- 3xx : redirection ;
- 401 : accès refusé ;
- 404 : URL introuvable ;
- 4xx : d'autres erreurs ;
- 5xx : les erreurs du serveur.

7.5.2 Les types des fichiers Logs :

Il existe plusieurs fichiers logs :

➢ **Fichier journal du serveur apache :**

Un log apache peut renseigner sur plusieurs paramètres comme l'octet envoyé, l'adresse IP distante, l'hôte distant, le nom utilisateur distant, le port du serveur, le statut de la requête, l'heure, l'url demandé, l'hôte virtuel du serveur etc. L'ensemble de ces informations permet d'avoir une idée générale sur toutes les requêtes qui étaient envoyées au serveur et les id des machines correspondant à ces requêtes.

Exemple de log apache (Voir figure 2.3).

➢ **Fichier journal du serveur Squid :**

Tous les logs de Squid se trouvent dans le répertoire /var/log/squid. Il y a des logs pour le cache, les accès et l'utilisation du disque. Le fichier access.log garde la trace des requêtes des clients, de leur activité, et fournit une ligne pour chaque requête HTTP& ICP reçue par le

serveur Proxy, l'adresse IP du client, la méthode d'interrogation, l'URL demandée, etc. Les données de ce fichier peuvent être analysées pour disposer d'information sur les accès.
Des programmes comme sarg, calamaris, Squid-Log- Analyzer sont disponibles pour analyser ces données et génèrent des rapports (au format HTML). Ces rapports peuvent être établis par les utilisateurs, les adresses IP, les sites visités, etc… [17].

Exemples de log Squid :

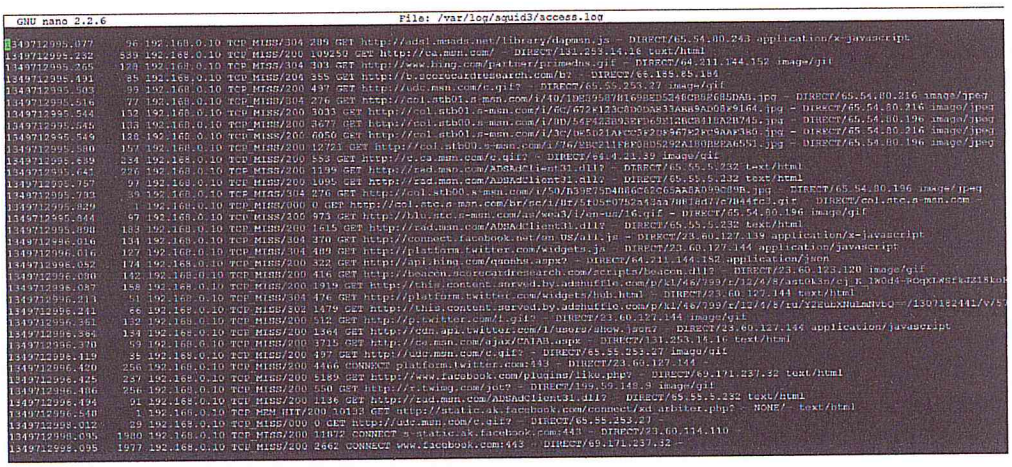


Figure 2.4 : Exemples de log Squid [18]

- **Fichier journal d'Analog :**

Analog est un programme d'analyse des fichiers log. Il présente de nombreux avantages : Rapide, flexible, il est facile à installer et à utiliser. Il donne des statistiques très précises sur l'heure, le domaine géographique, l'organisation, les termes recherchés, le système d'exploitation de l'utilisateur connecté, le code statut (requête incorrecte), le fichier demandé par l'utilisateur etc...

Exemples de log :

host.analog.cx - - [31/Dec/1999:22:11:12 +0000] "GET /sample.html HTTP/1.0" 200 1234 "http://referrer.com/" "Mozilla/4.0 (compatible; MSIE 4.01; Windows 98)"

host.analog.cx - - [31/Dec/1999:23:11:12 +0000] "GET /sample.html HTTP/1.0" 200 1234 "http://referrer.com/" "Mozilla/4.0 (compatible; MSIE 4.01; Windows 98)"

host.analog.cx - - [01/Jan/2000:02:11:12 +0000] "GET /sample.html HTTP/1.0" 200 1234 "http://google.com/search?q=sample%20search" "Mozilla/4.0 (compatible; MSIE 4.01; Windows 98)"

Figure 2.5 : Exemples de log Analog

➢ **Fichier journal du Caritig :**

Caritig est un des sites qui dispose de statistiques permettant d'avoir une idée sur le nombre des visiteurs, les pages consultées et la durée de la recherche sur le site. Les sociétés améliorent ainsi leurs sites pour le rendre plus attractif. Ces statistiques sont extraites en général d'un fichier log, les informations publiées sur le site sont : l'agent (adresse), l'heure, l'adresse IP, l'hôte etc.

Exemple de statistique :

Internet Explorer 5.0 - Win 98 195.83.96. --- [Srv] station.adm.ac-versailles.fr 08/01 15:38

Internet Explorer 5.5 - Win NT 80.9.159. --- [Srv] mix-toulouseabo.wanadoo.fr 08/01 15:37

Internet Explorer 4.0 - Win NT 193.249.234. --- [Srv] mix-lagnyabo.wanadoo.fr 08/01 15:33

Internet Explorer 5.5 - Win NT 80.9.158. --- [Srv] mix-toulouseabo.wanadoo.fr 08/01 15:30

Internet Explorer 5.0 - Win 98 80.8.169. --- [Srv] ca-marseilleabo.wanadoo.fr 08/01 15:19

Yahoo ! (Annuaire)

Figure 2.6 : Log caritig

7.5.3 Problèmes spécifiques aux données des fichiers logs :

Bien que les données fournies par les fichiers Logs soient utiles, il importe de prendre en compte les limites inhérentes à ces données lors de leur analyse et de leur interprétation.

Parmi les difficultés qui peuvent survenir :

➢ **Les requêtes inutiles** [22] :

Chaque fois qu'il reçoit une requête, le serveur enregistre une ligne dans le fichier Log. Ainsi, pour charger une page, il y'aura autant de lignes dans le fichier que d'objets contenus sur cette page (les éléments graphiques). Un prétraitement est donc indispensable pour supprimer les requêtes inutiles.

➢ **Les firewalls** [22] **:**

Ces protections d'accès à un réseau masquent l'adresse IP des utilisateurs. Toute requête de connexion provenant d'un serveur doté d'une telle protection aura la même adresse et ce, quel que soit l'utilisateur. Il est donc impossible, dans ce cas, d'identifier et de distinguer les visiteurs provenant de ce réseau.

➢ **Le Web caching** [22] :

Afin de faciliter le trafic sur le Web, une copie de certaines pages est sauvegardée au niveau du navigateur local de l'utilisateur ou au niveau du serveur Proxy afin de ne pas les télécharger chaque fois qu'un utilisateur les demande. Dans ce cas, une page peut être consultée plusieurs fois sans qu'il y' ait autant d'accès au serveur. Il en résulte que les requêtes correspondantes ne sont pas enregistrées dans le fichier Log.

➢ **L'utilisation des robots** [22] :

Les annuaires du Web, connus sous le nom de moteurs de recherche, utilisent des robots qui parcourent tous les sites Web afin de mettre à jour leur index de recherche. Ce faisant, ils déclenchent des requêtes qui sont enregistrées dans tous les fichiers Logs des différents sites, faussant ainsi leurs statistiques.

➢ **L'identification des utilisateurs** :

L'identification des utilisateurs à partir du fichier Log n'est pas une tâche simple. En effet, en employant le fichier Log, l'unique identifiant disponible est l'adresse IP et « l'agent» de l'utilisateur. Cet identifiant présente plusieurs limites [20] :

- **Adresse IP unique / Plusieurs sessions serveurs** :

  La même adresse IP peut être attribuée à plusieurs utilisateurs accédant aux services du Web à travers un unique serveur Proxy.

- **Plusieurs adresses IP / Utilisateur unique** :

  Un utilisateur peut accéder au Web à partir de plusieurs machines.

- **Plusieurs agents / Utilisateur unique** :

  Un internaute qui utilise plus d'un navigateur, même si la machine est unique, est aperçu comme plusieurs utilisateurs.

➢ **L'identification des sessions** [21] :

Toutes les requêtes provenant d'un utilisateur identifié constituent sa session. Le début de la session est défini par le fait que l'URL de provenance de l'utilisateur est extérieure au site. Par contre, aucun signal n'indique la déconnexion du site et par suite la fin de la session.

➢ **Le manque d'information** [22] :

Le fichier Log n'apporte rien sur le comportement de l'utilisateur entre deux requêtes : Que fait ce dernier ? Est-il vraiment en train de lire la page affichée ? De plus, le nombre de visites d'une page ne reflète pas nécessairement l'intérêt de celle-ci. En effet, un nombre élevé de visites peut simplement être attribué à l'organisation d'un site.

8. Présentation des principaux travaux de recherche existants (Etat de l'art) :

Le nombre d'accès aux pages Web ne cesse d'augmenter. Le Web est devenu l'une des plates-formes les plus répandues pour la diffusion et la recherche d'information. Par conséquence, beaucoup d'opérateurs de sites Web sont incités à analyser l'usage de leurs sites afin d'améliorer leur réponse vis-à-vis des attentes des internautes. Or, la manière dont un site Web est visité peut changer en fonction de divers facteurs. Les modèles d'usage doivent ainsi être mis à jour continuellement afin de refléter fidèlement le comportement des visiteurs.
Beaucoup de recherches sur l'analyse du web, nous allons présenter, les principaux travaux existants sur le Web Usage Mining présentés dans ce chapitre selon les étapes du déroulement de son processus.

- **Doru TANASA 2005** [19] **:** l'auteur a proposé une méthodologie générale de prétraitement des logs Web et une méthodologie générale divisive avec trois approches (ainsi que des méthodes concrètes associées) pour la découverte des motifs séquentiels ayant un faible support. Sa première contribution concerne le prétraitement des données d'usage Web, domaine encore très peu abordé dans la littérature. L'originalité de la méthodologie de prétraitement proposée consiste dans le fait qu'elle prend en compte l'aspect multi-sites du WUM, indispensable pour appréhender les pratiques des internautes qui naviguent de façon transparente, par exemple, sur plusieurs sites Web d'une même organisation.
Sa deuxième contribution vise la découverte à partir d'un fichier log prétraités de grande taille, des comportements minoritaires correspondant à des motifs séquentiels de très faible support. Il a proposé une méthodologie générale visant à diviser le fichier log prétraités en sous-logs, se déclinant selon trois approches d'extraction de motifs séquentiels au support faible (Séquentielle, Itérative et Hiérarchique). Celles-ci ont été implémentées dans des méthodes concrètes hybrides mettant en jeu des algorithmes de classification et d'extraction de motifs séquentiels.

- **Malika Charrad 2005** [42] **:** L'approche qu'elle a proposé afin d'aider à comprendre le comportement des internautes comporte trois phases : prétraitement des fichiers Logs, classification des pages et classification des internautes. Dans la phase de prétraitement, les requêtes sont organisées en visites qui représentent les unités d'interaction entre les utilisateurs du Web et le serveur web. Dans la phase de classification des pages, une représentation interne du site Web est créée à partir des fichiers Logs afin d'extraire des chemins de navigation. Des paramètres introduits à partir des statistiques sur les accès aux pages sont utilisés pour la catégorisation des pages Web en pages auxiliaires et pages de contenu. Les requêtes aux pages de contenu servent à la découverte des motifs de navigation. Afin de construire des segments d'utilisateurs, deux méthodes hybrides de classification automatiques basées sur l'analyse en composantes principales, l'analyse des correspondances multiples et les cartes topologiques de Kohonen sont appliquées aux visites.

- **Da Silva 2006** [44] **:** l'auteur a proposé une méthodologie pour la génération automatique des données artificielles permettant la simulation des changements. Guidés par les pistes nées des analyses exploratoires, il a proposé une nouvelle approche basée sur des fenêtres non recouvrante pour la détection et le suivi des changements sur des données évolutives. Cette approche caractérise le type de changement subi par les groupes de comportement (apparition, disparition, fusion, scission) et applique deux indices de validation basés sur l'extension de la classification pour mesurer le niveau des changements repérés à chaque pas de temps.

- **Nabila Mezoug et Hanane Bessa 2009** [43] **:** L'approche présentée : est une étude de cas en fouille de données d'usage de web qui consiste à analyser les données (les fichiers log ou bien le journal des connexions) afin de transformer ces données en des connaissances utiles pour l'identification d'éventuels comportements typiques d'internautes selon leur profil, s'est déroulée en trois étapes :
Elle consiste dans un premier temps à un prétraitement des données qui servent à la récupération et la concaténation des fichiers log afin que les requêtes soient organisées en navigations.

Des paramètres introduits à partir des statistiques sur les accès aux pages sont utilisés pour la catégorisation des pages Web afin de sauvegarder les pages de contenu et d'éliminer les pages auxiliaires qui ne présentent aucun intérêt, c'est la classification des pages qui est basée sur deux méthodes hybrides à savoir l'analyse en composante principale et le clustering k_means, ces pages de contenu permettent l'extraction des profils.

➢ **Etude critique :**

Nous avons conclus après cette présentation que le processus du Web Usage Mining est un sujet de recherche très important, et objet de plusieurs articles, thèses et travaux de recherche scientifique.
Nous ne sommes pas expérimentés dans les différentes thématiques de recherche pour que nous puissions comparer les techniques existantes en termes d'efficacité, mais nous pouvons remarquer que la plupart des travaux sont basé sur le domaine de classification des données et les auteurs sont en face à des grand problèmes tel que les données sont de plus en plus volumineuses et les comportements des utilisateurs restent difficiles à détecter à cause de la quantité de données à analyser.

9. Conclusion :

Ce chapitre a servi d'introduction aux domaines liés à notre étude. Nous avons défini certaines notions relatives aux fonctionnalités et les avantages de l'intranet, l'analyse du trafic web sur ce réseau et les différentes méthodes utilisées plus particulièrement le Web Usage Mining et les principaux travaux de recherches sur lequel porte notre étude. Le chapitre suivant est consacré à la présentation d'un domaine qui sera appliqué pour notre solution qui est le système décisionnel, ainsi que les différentes étapes de ce domaine.

CHAPITRE 3

SYSTEME D'INFORMATION DECESIONNEL

CHAPITRE 3 : SYSTEME D'INFORMATION DECISIONNEL

1. Introduction :

Après l'analyse des fichiers journaux des serveurs web apache nous obtenons des données qui seront exploitées selon l'informatique décisionnelle.
Face à ce besoin nous allons voir une présentation sur le système décisionnel.

2. Qu'est-ce que le Système d'Information Décisionnel :

Afin de mieux comprendre le système décisionnel, il faut rappeler ce qu'est un système d'information.

Un système d'information (SI) est un ensemble organisé de ressources (matériels, logiciels, personnel, données et procédures) qui permet de collecter, regrouper, classifier, traiter et diffuser de l'information dans un environnement donné [23].

Le traitement transactionnel en ligne (Online transaction processing OLTP en anglais) est un type d'application informatique qui sert à effectuer des modifications d'informations en temps réel. Ce type d'application est utilisé dans des activités opérationnelles, typiquement des transactions commerciales (opérations bancaires, achats de biens, billets, réservations) [32], contrairement **le traitement analytique en ligne** (online analytical processing, OLAP en anglais) qu'est un type d'application informatique orienté vers l'analyse sur-le-champ d'informations selon plusieurs axes, dans le but d'obtenir des rapports de synthèse tels que ceux utilisés en analyse financière. Les applications de type OLAP sont couramment utilisées en informatique décisionnelle, dans le but d'aider la direction à avoir une vue transversale de l'activité d'une entreprise [33].

L'informatique décisionnelle ou Système d'Information Décisionnel (SID) (Business Intelligence BI en anglais) parfois appelé tout simplement « le Décisionnel » est une branche de l'architecture des systèmes d'informations. Elle permet de mettre en œuvre des moyens pour collecter, consolider et restituer des données afin d'offrir à une entreprise une aide à la décision [24], c'est-à-dire l'exploitation des données de l'entreprise dans le but de faciliter la prise de décision par les décideurs [25].

3. L'architecture d'un SID :

Les entrepôts de données intègrent les informations en provenance de différentes sources, souvent réparties et hétérogènes et qui ont pour objectif de fournir une vue globale de l'information aux analystes et aux décideurs. Ces applications d'aide à la décision sont de type OLAP (On-Line Analytical Processing ou Analyse en ligne).

La construction et la mise en œuvre d'un entrepôt de données représentent une tâche complexe qui se compose de plusieurs étapes. La première consiste à l'analyse des sources de données et à l'identification des besoins des utilisateurs.

La deuxième correspond à l'organisation des données à l'intérieur de l'entrepôt. Finalement, la troisième consiste à établir divers outils d'interrogation (d'analyse, de fouille de données ou d'interrogation). Chaque étape présente des problématiques spécifiques. Ainsi, par exemple, lors de la première étape, la difficulté principale consiste en l'intégration des données, de manière à qu'elles soient de qualité pour leur stockage. Pour l'organisation, ils existent plusieurs problèmes comme : la sélection des vues à matérialiser, le rafraîchissement de l'entrepôt, la gestion de l'ensemble de données (courantes et historisées), entre autres. En ce qui concerne le processus d'interrogation, nous avons besoin des outils performants et conviviaux pour l'accès et l'analyse de l'information.

Nous résumons l'architecture à trois niveaux principaux [26] :

- Les sources d'information qui correspondent à l'ensemble des bases de données de production et sites dont sont extraites les informations décisionnelles ;
- L'entrepôt qui contient l'ensemble des données extraites de ces sources ;
- Les magasins extraits de l'entrepôt et dédiés aux différentes classes de décideurs.

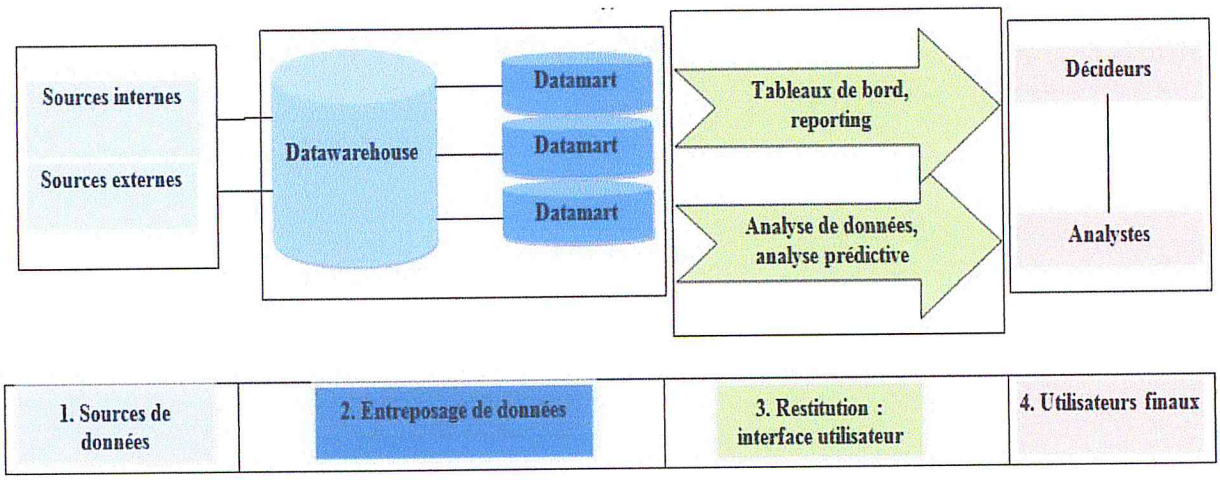


Figure 3.1 : Architecture d'un système d'information décisionnel [27].

4. Systèmes transactionnels et systèmes décisionnels :

Les SGBD ont été créés pour gérer de grands volumes d'information contenus dans les différents systèmes opérationnels qui appartiennent à l'entreprise. Ces données sont manipulées en utilisant des processus transactionnels en ligne [36].

Parallèlement à l'exploitation de l'information contenue dans ces systèmes opérationnels, les dirigeants des entreprises ont besoin d'avoir une vision globale concernant toute cette information pour faire des calculs prévisionnels, des statistiques ou pour établir des stratégies de développement et d'analyses des tendances.

Le tableau 3.1 compare les caractéristiques des systèmes transactionnels et décisionnels par rapport aux données et aux utilisateurs [30].

	S.Transactionnel	SID
Utilisateurs	- Nombreux - Variés (employés, directeurs,...) - Concurrents - Mises à jour et Interrogations - Requêtes prédéfinies - Réponses immédiates - Accès à peu d'information	- Peu - Décideurs, analystes - Non concurrents - Interrogations - Requêtes imprévisibles et complexes - Réponses moins rapides - Accès à de nombreuses informations
Données	- Exhaustives - Courantes - Dynamiques - Orientées applications	- Résumées - Historiques - Statiques - Orientées sujets (d'analyse)

Tableau 3.1 : Comparaison des deux systèmes

5. Entrepôt de données :

Entrepôt de données (Data Warehouse (DW) en anglais) définit par **Bill Inmon** [28] comme suit : « Le DW est une collection de données orientées sujet, intégrées, non volatiles et évolutives dans le temps, organisées pour le support d'un processus d'aide à la décision. »

Un entrepôt de données est le lieu de stockage des données provenant généralement des bases de productions, mais aussi d'autres sources de données hétérogènes externes [29].

5.1 Les caractéristiques d'un entrepôt de données :

Les caractéristiques d'un entrepôt de données selon **Bill Inmon** [28] sont :

- **Orienté sujet :**

Le DW est organisé autour des sujets majeurs de l'entreprise, contrairement à l'approche transactionnelle utilisée dans les systèmes opérationnels, qui sont conçus autour d'applications et de fonctions telles que : cartes bancaires, solvabilité client…, les DW sont organisés autour de sujets majeurs de l'entreprise tels que : clientèle, ventes, produits…. Cette organisation affecte forcément la conception et l'implémentation des données contenues dans le DW. Le contenu en données et en relations entre elles diffère aussi. Dans un système opérationnel, les données sont essentiellement destinées à satisfaire un processus fonctionnel et obéit à des règles de gestion, alors que celles d'un DW sont destinées à un processus analytique.

- **Intégrée :**

Le DW va intégrer des données en provenance de différentes sources. Cela nécessite la gestion de toute incohérence.

- **Evolutives dans le temps :**

Dans un système décisionnel il est important de conserver les différentes valeurs d'une donnée, cela permet les comparaisons et le suivi de l'évolution des valeurs dans le temps, alors que dans un système opérationnel la valeur d'une donnée est simplement mise à jour. Dans un DW chaque valeur est associée à un moment.

- **Non volatiles :**

C'est ce qui est, en quelque sorte la conséquence de l'historisation décrite précédemment. Une donnée dans un environnement opérationnel peut être mise à jour ou supprimée, de telles opérations n'existent pas dans un environnement DW.

- **Organisées pour le support d'un processus d'aide à la décision :**

Les données du DW sont organisées de manière à permettre l'exécution des processus d'aide à la décision (Reporting, Data Mining…).

5.2 La structure des données d'un DW [28] :

Le schéma suivant présente la structure des données dans un DW

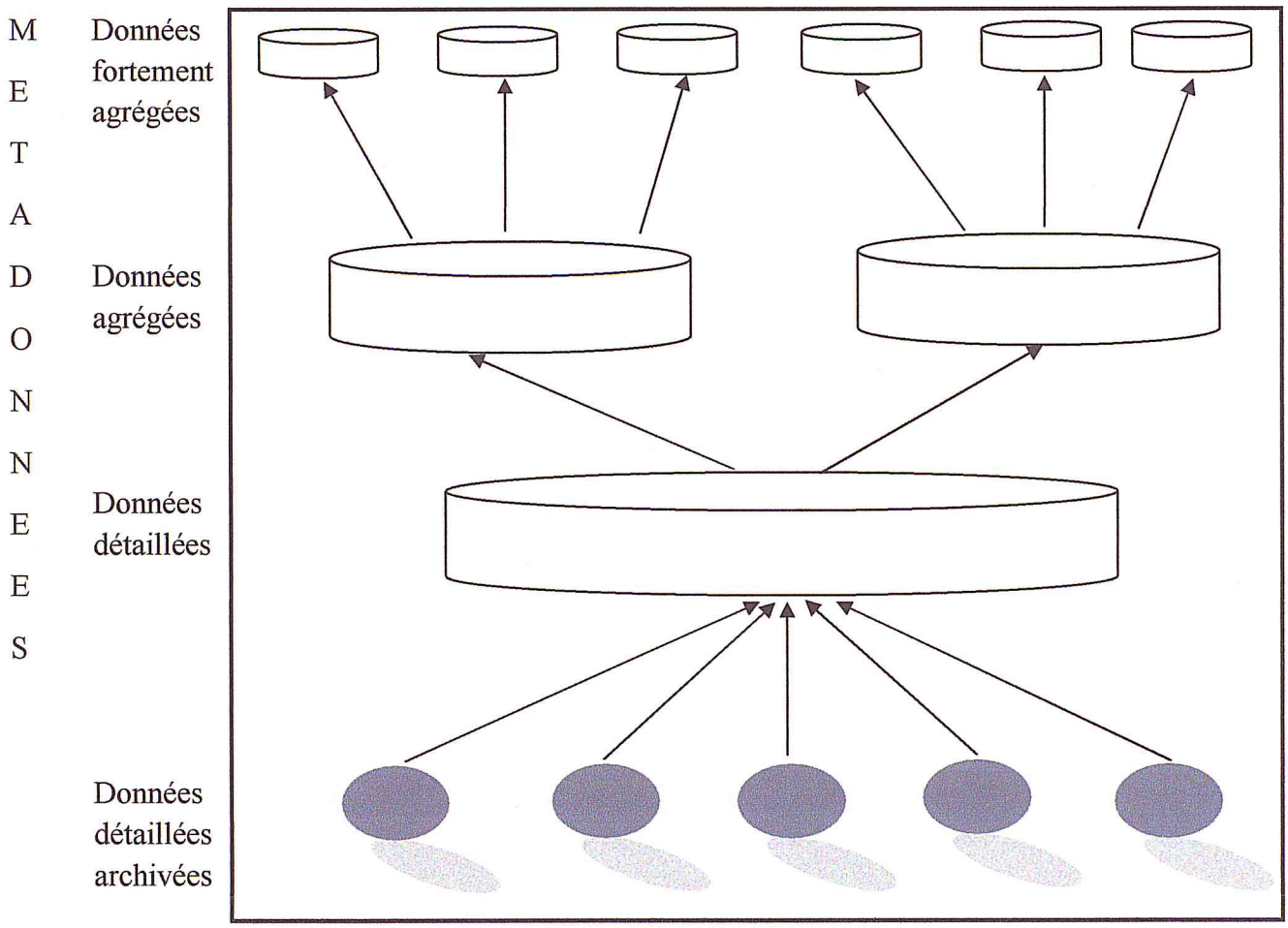


Figure 3.2 : Structure des données d'un DW [28].

➢ **Les données d'un DW :**

- **Données détaillées :**

Ce sont les données qui reflètent les événements les plus récents, fréquemment consultées, généralement volumineuses car elles sont d'un niveau détaillé.

- **Données détaillées archivées :**

Anciennes données rarement sollicitées, généralement stockées dans un disque de stockage de masse, peu coûteux, à un même niveau de détail que les données détaillées.

- **Données agrégées :**

Données agrégées à partir des données détaillées.

- **Données fortement agrégées :**

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

- **Meta données :**

Ce sont les informations relatives à la structure des données, les méthodes d'agrégation et le lien entre les données opérationnelles et celles du DW. Les métadonnées doivent renseigner sur :

- Le modèle de données,
- La structure des données telle qu'elle est vue par les développeurs,
- La structure des données telle qu'elle est vue par les utilisateurs,
- Les sources des données,
- Les transformations nécessaires,
- Suivi des alimentations.

5.3 DW versus SGBD :

D'après la présentation sur le DW et notre connaissance sur le système de gestion de base de données (SGBD) nous avons pu faire la différence entre les deux. Le tableau suivant récapitule cette différence selon différents critères [30] :

	SGBD	Entrepôt de données
Objectifs	- Gestionnaire de production	- Consultation et analyse
Utilisateurs	- Gestionnaire de production	- Décideurs, analystes
Taille de la base	- Plusieurs Giga-octets	- Plusieurs Téraoctets
Organisation des données	- Par traitement	- Par métier
Type de données	- Données de gestion (courantes)	- Données d'analyse (résumées, historisées)
Requêtes	- Simples, prédéterminées, données détaillées	- Complexes, spécifiques, agrégations et group by
Transactions	- Courtes et nombreuses, temps réel	- Longues, peu nombreuses

Tableau 3.2 : Différence entre SGBD et DW

6. Magasin de données :

Le magasin de données (Data Mart DM en anglais) est un modèle miniature de l'entrepôt de données au niveau départemental, alimentées par le DW et basé sur les besoins départementaux en informations [28]. La relation entre les deux représenté sur le schéma suivant :

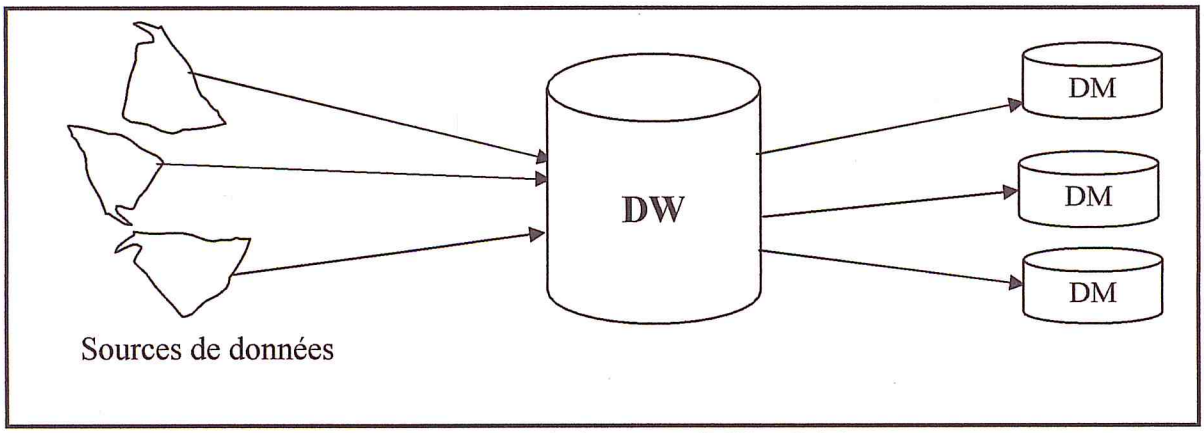


Figure 3.3 : La relation entre DW et DM [28]

7. Modélisation multidimensionnelle :

Pour arriver à construire un modèle approprié pour un entrepôt de données, nous pouvons choisir, soit un schéma relationnel (le schéma en étoile, en flocon de neige ou en constellation) ; soit un schéma multidimensionnel. Avant de décrire les différents schémas, nous commençons par quelques concepts de base.

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

7.1 Table de faits :

Une table de faits représente l'objet de l'analyse. Elle contient principalement des mesures sous forme d'attributs représentant les éléments d'analyse. Les faits les plus utilisables sont les numériques, les valeurs continues et additives. Les mesures peuvent être par exemple, une quantité, une vente, etc. ; qui sont résumées ou représentées par une moyenne. Ces mesures sont reliées chacune à une table de dimension avec des clés étrangères.

La granularité des tables de faits est une caractéristique importante expliquée par le niveau de détail des mesures représentées.

La table de faits peut ne pas être sujette à des analyses relatives à l'agrégation. Elle représente la réalisation d'un événement sans le mesurer. Ces tables sont des tables sans faits.

7.2 Table de dimension :

Une table de dimension est un objet qui inclut un ensemble d'attributs permettant à l'utilisateur d'avoir des mesures suivant différentes perspectives d'analyse. Les attributs sont des indicateurs pour les différentes vues d'analyses possibles. Il est commun d'avoir plus que cent attributs dans une application réelle. Pour cette raison, les tables de dimension sont considérées comme étant grandes.

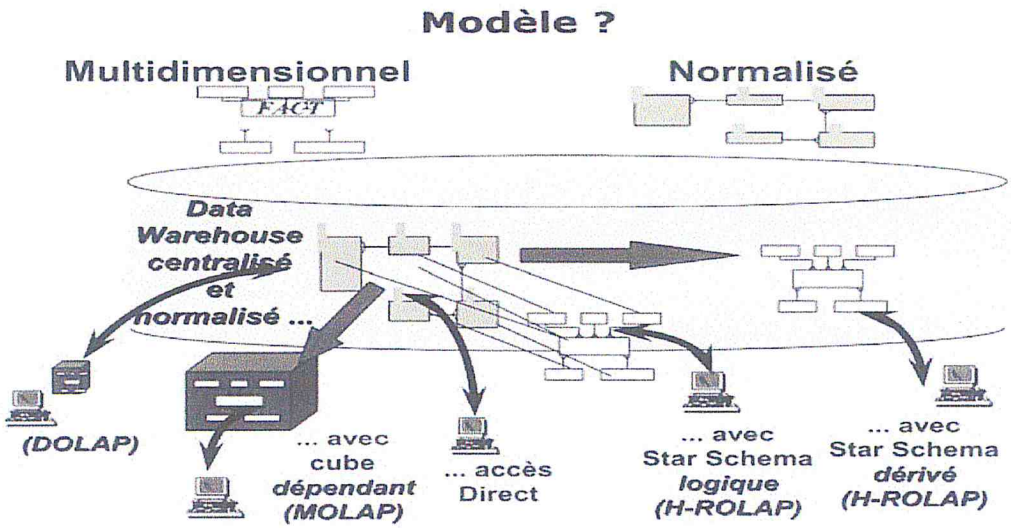


Figure 3.4 : Modèle d'un SID [31]

7.3 Schémas relationnels :

Dans les schémas relationnels nous trouvons deux types de schémas. Les premiers sont des schémas qui répondent fort bien aux processus de type OLTP qui ont été décrits précédemment, alors que les deuxièmes, que nous appelons des schémas pour le décisionnel, ont pour but de proposer des schémas adaptés pour des applications de type OLAP.

Nous décrivons les différents types des schémas relationnels pour le décisionnel.

- **Le schéma en étoile** [34] :

Il se compose du fait central et de leurs dimensions. Dans ce schéma il existe une relation pour les faits et plusieurs pour les différentes dimensions autour de la relation centrale. La relation de faits contient les différentes mesures et une clé étrangère pour faire référence à chacune de leurs dimensions.

La figure 3.5 montre le schéma en étoile en décrivant les ventes réalisées dans les différents magasins de l'entreprise au cours d'un jour. Dans ce cas, nous avons une étoile centrale avec une table de faits appelée **Ventes** et autour leurs diverses dimensions : **Temps, Produit, Client et Magasin.**

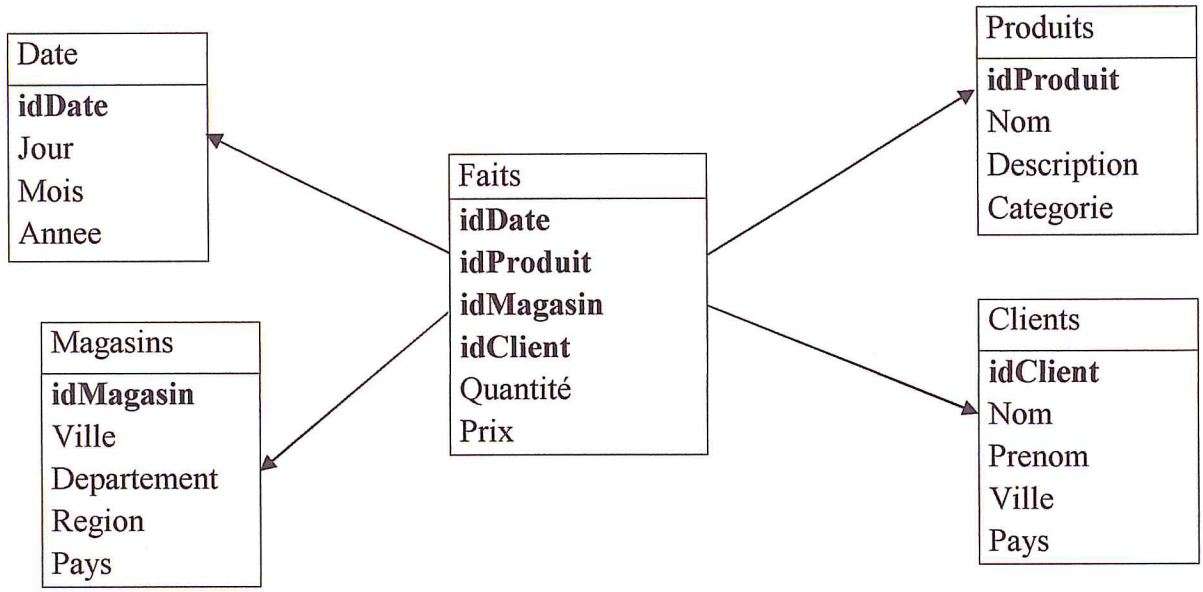


Figure 3.5 : Exemple de modélisation en étoile [34]

➢ **Le schéma en flocon de neige (Snowflake) :**

Il dérive du schéma précédent avec une relation centrale et autour d'elle les différentes dimensions, qui sont éclatées ou décomposées en sous hiérarchies. L'avantage du schéma en flocon de neige est de formaliser une hiérarchie au sein d'une dimension, ce qui peut faciliter l'analyse. Un autre avantage est représenté par la normalisation des dimensions, car nous réduisons leur taille [35].

Les hiérarchies pour le schéma en flocon de neige de l'exemple de la figure 3.6 sont :

Dimension Temps = Jour →Mois → Année
Dimension Magasin = Ville → Département → Région → Pays

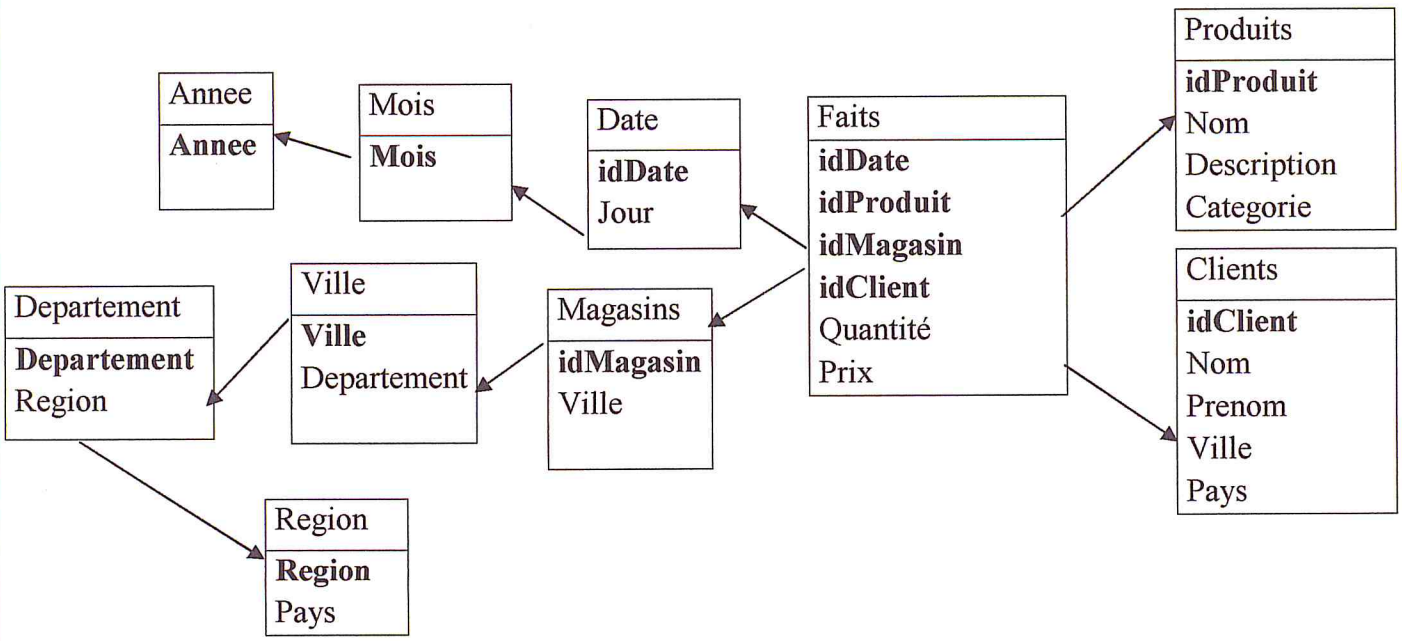


Figure 3.6 : Exemple de modélisation en flocon de neige

Dans l'exemple ci-dessus, la dimension Temps a été éclatée en trois, **Temps**, **Mois** et **Annee**. La deuxième dimension **Magasin**, a été décomposée en quatre : **Magasin**, **Ville, Département** et **Région.**

➢ **Le schéma en constellation :**

Le schéma en constellation représente plusieurs relations de faits qui partagent des dimensions communes. Ces différentes relations de faits composent une famille qui partage les dimensions mais où chaque relation de faits a ses propres dimensions [35].

La figure 3.7 présenté par **T. María et E. Serna** [30] montre le schéma en constellation qui est composé de deux relations de faits. La première s'appelle Ventes et enregistre les quantités de produits qui ont été vendus dans les différents magasins pendant un certain jour. La deuxième relation gère les différents produits achetés aux fournisseurs pendant un certain temps.

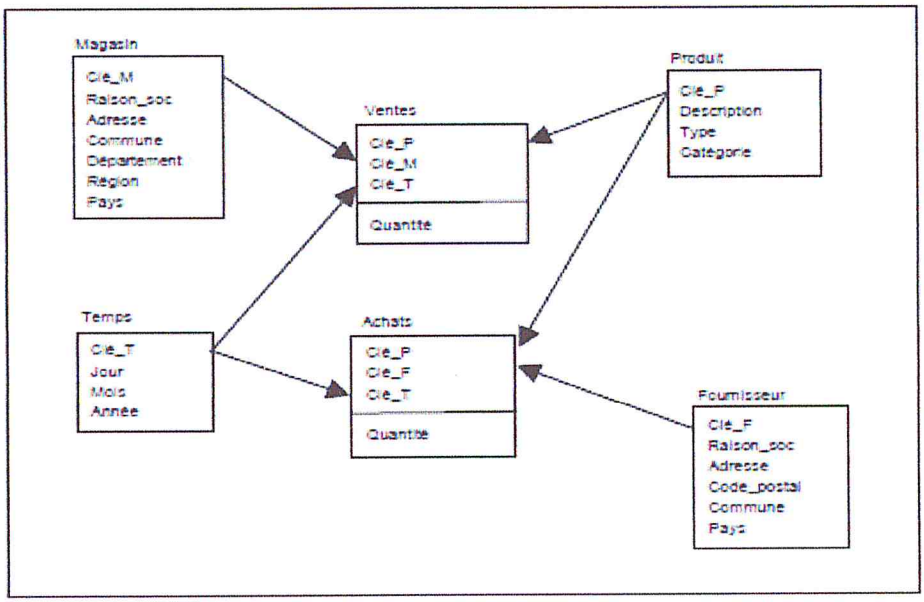


Figure 3.7 : Exemple de modélisation en constellation

La relation de faits Ventes partage leurs dimensions **Temps** et **Produits** avec la table **Achats**. Néanmoins, la dimension **Magasin** appartient seulement à **Ventes**. Egalement, la dimension **Fournisseur** est liée seulement à la relation **Achats.**

7.4 Schéma multidimensionnel (Cube) :

Dans le modèle multidimensionnel, le concept central est le cube, lequel est constitué des éléments appelés cellules qui peuvent contenir une ou plusieurs mesures. La localisation de la cellule est faite à travers les axes, qui correspondent chacun à une dimension. La dimension est composée de membres qui représentent les différentes valeurs [30].

En reprenant une partie du schéma en étoile de la figure 3.5, nous pouvons construire le schéma multidimensionnel suivant.

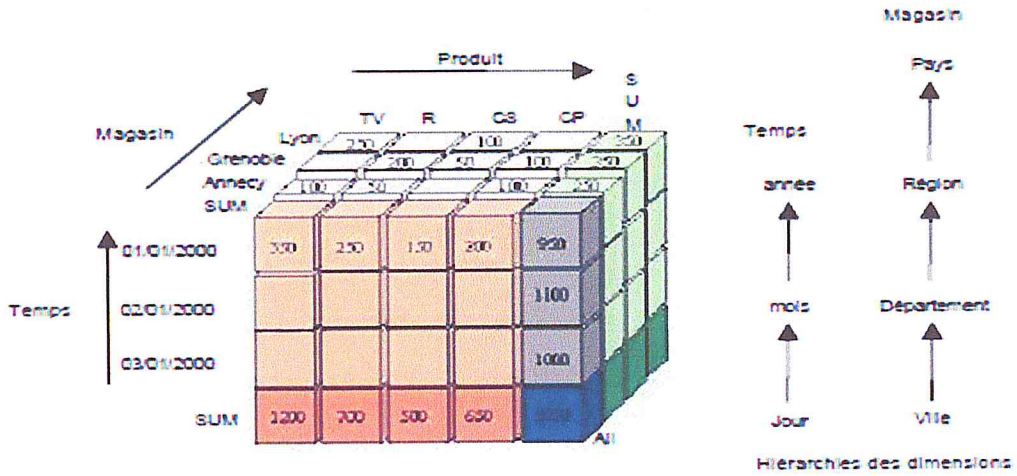


Figure 3.8 : Exemple de schéma multidimensionnel [30]

La figure 3.8, présente un schéma multidimensionnel pour les ventes qui ont été réalisées dans les magasins pour les différents produits au cours d'un temps donné (jour). Par exemple, nous avons la quantité de 100 Téléviseurs vendus dans le magasin d'Annecy pendant le 1er janvier 2000.

8. Le concept OLAP :

Aujourd'hui, OLAP permet aux décideurs, d'une entreprise d'avoir accès rapidement et de manière interactive à une information pertinente présentée sous des angles divers et multiples, selon leurs besoins particuliers.

8.1 Généralités :

Le terme OLAP (On-Line Analytical Processing) désigne une classe de technologies conçue pour l'accès aux données et pour une analyse instantanée de ces dernières, dans le but de répondre aux besoins d'analyse [36].

8.2 Serveurs OLAP (On-Line Analytical Processing) [30] :

Les données opérationnelles constituent la source principale d'un système d'information décisionnel. Les systèmes décisionnels complets reposent sur la technologie OLAP, conçue pour répondre aux besoins d'analyse des applications de gestion.

➢ ROLAP (Relational OLAP) :

Dans les systèmes relationnels OLAP, l'entrepôt de données utilise une base de données relationnelle. Le stockage et la gestion de données sont relationnels. Toutefois, le modèle relationnel requiert des extensions pour supporter les requêtes d'analyses multidimensionnelles du niveau d'application. Le moteur ROLAP traduit dynamiquement le modèle logique de données multidimensionnel M en modèle de stockage relationnel R (la plupart des outils requièrent que la donnée soit structurée en utilisant un schéma en étoile ou un schéma en flocon de neige). Techniquement, le moteur ROLAP exécute une transformation à partir d'une requête multidimensionnelle m contre M vers une requête relationnelle r contre R. L'efficacité du résultat de la requête est le facteur dominant pour la performance et le passage à l'échelle global du système. Ainsi, les stratégies d'optimisation représentent le point principal qui distingue les systèmes ROLAP existants.

La figure 3.9 montre une architecture pour le serveur ROLAP.

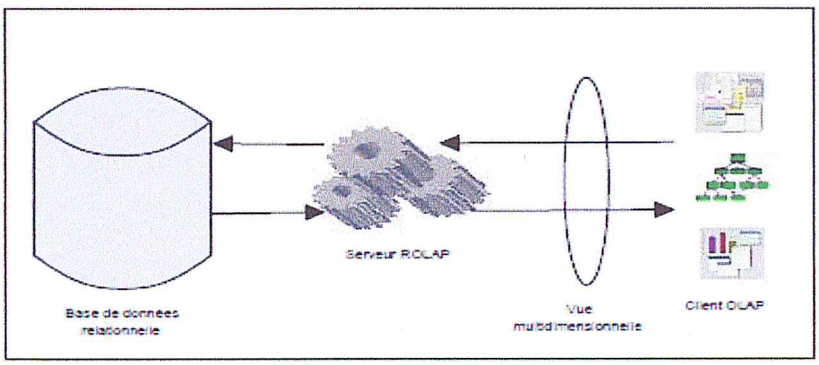


Figure 3.9 : Architecture ROLAP

La technologie ROLAP a deux avantages principaux :

1) elle permet la définition de données complexes et multidimensionnelles en utilisant un modèle relativement simple,

2) elle réduit le nombre de jointures à réaliser dans l'exécution d'une requête. Le désavantage est que le langage de requêtes tel qu'il existe, n'est pas assez puisant ou n'est pas assez flexible pour supporter de vraies capacités d'OLAP.

➢ **MOLAP (Multidimensional OLAP) :**

Les systèmes multidimensionnels OLAP utilisent une base de données multidimensionnelle pour stocker les données de l'entrepôt et les applications analytiques sont construites directement sur elle. Dans cette architecture, le système de base de données multidimensionnel sert tant au niveau de stockage qu'au niveau de gestion des données. Les données des sources sont conformes au modèle multidimensionnel, et dans toutes les dimensions, les différentes agrégations sont pré calculées pour des raisons de performance.
La figure 3.10 montre une architecture pour les systèmes MOLAP.

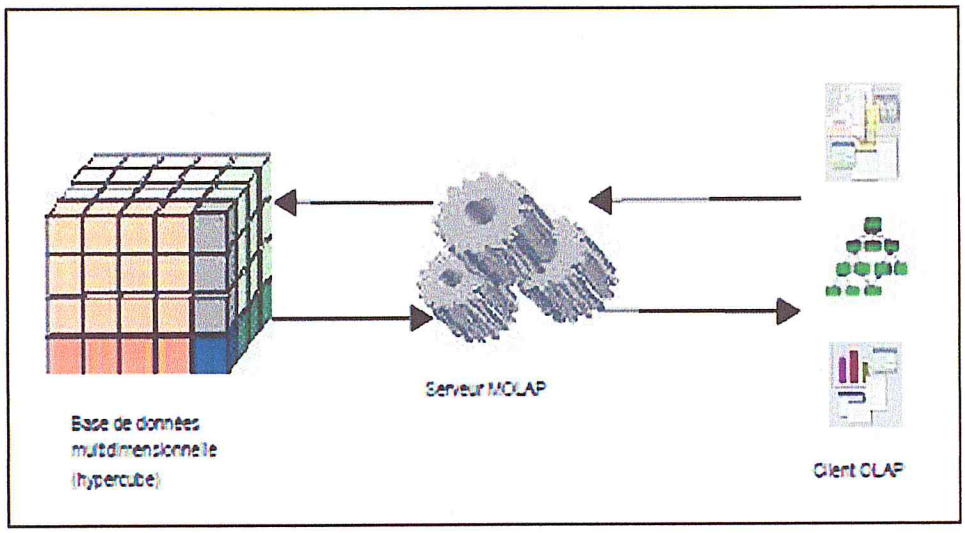


Figure 3.10 : Architecture MOLAP

Les avantages des systèmes MOLAP sont basés sur les désavantages des systèmes ROLAP et elles représentent la raison de leur création. D'un côté, les requêtes MOLAP sont très puissantes et flexibles en termes du processus OLAP, tandis que, d'un autre côté, le modèle physique correspond plus étroitement au modèle multidimensionnel. Néanmoins, il existe des désavantages au modèle physique MOLAP. Le plus important, à notre avis, c'est qu'il n'existe pas de standard du modèle physique.

➢ **HOLAP (Hybrid OLAP) :**

Un système HOLAP est un système qui supporte et intègre un stockage des données multidimensionnel et relationnel d'une manière équivalente pour profiter des caractéristiques de correspondance et des techniques d'optimisation.

Un système HOLAP doit fournir :

La transparence du système : Pour la localisation et l'accès aux données, sans connaître si elles sont stockées dans un SGBD relationnel ou dimensionnel. Pour la transparence de la fragmentation,...

Un modèle de données général et un schéma multidimensionnel global : Pour aboutir à la transparence du premier point, tant le modèle de données général que le langage de requête uniforme doivent être fournis. Etant donné qu'il n'existe pas un modèle standard, cette condition est difficile à réaliser.

Une allocation optimale dans le système de stockage : Le système HOLAP doit bénéficier des stratégies d'allocation qui existent dans les systèmes distribués tels que : le profil de requêtes, le temps d'accès, l'équilibrage de chargement,...

Actuellement, la plupart des systèmes commerciaux utilisent une approche hybride. Cette approche permet de manipuler des informations de l'entrepôt de données avec un moteur ROLAP, tandis que pour la gestion des data marts, ils utilisent l'approche multidimensionnelle. Dans la figure 3.11, nous montrons une architecture en utilisant les types de serveurs ROLAP et MOLAP pour le stockage de données.

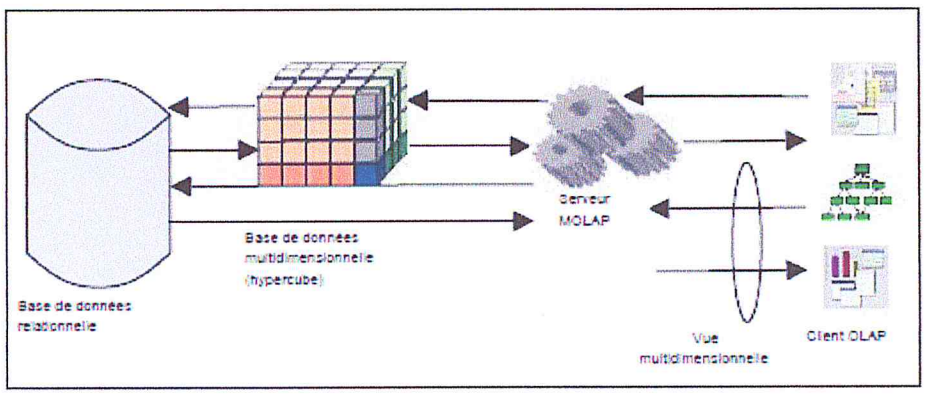


Figure 3.11 : Architecture HOLAP

9. Extraction, Transformation, Chargement « ETL » [37] :

Afin d'alimenter le DWH à partir des différentes applications de l'entreprise, on utilise une gamme d'outils appelés ETL.

9.1 Définition ETL :

La notion d'ETL (Extract Transform Loading), recouvre à la fois des outils et des processus d'alimentation. Il s'agit d'un élément clé dans l'intégration d'applications, en particulier dans le monde de la Business Intelligence et du DataWarehousing. Outils et processus ETL sont des briques d'une infrastructure de SI, dont la valeur ajoutée et le retour sur investissement s'expriment dans le temps en accompagnant l'évolution du système d'information global ou du système d'information décisionnel.

Les outils ETL gèrent toutes les étapes de la collecte des données au sein des systèmes d'information hétérogènes : SGBD, ERP, applications spécifiques, fichiers plats, bases hiérarchiques... depuis le nettoyage des données collectées, la consolidation et la mise en concordance des données éparses jusqu'à leur distribution auprès des applications cibles ou des systèmes décisionnels (analyse, tableau de bord...).

Le processus ETL est une opération de migration de données qui consiste aussi à la rendre aisément consommable. Ce processus représente une part majeure des traitements et nécessite une attention régulière tout au long du cycle de vie du système, dans la mesure où il est garant de la qualité des données.

9.2 Décomposition ETL :

Un processus ETL se décompose en trois phases : l'extraction, la transformation le chargement.

- **L'extraction des données :**

Il s'agit en premier lieu d'aller chercher les données là où elles se trouvent. L'outil ETL a la capacité de se connecter aux différentes applications, bases de données ou fichiers.

Pour cela, plusieurs technologies sont utilisables :

- Les passerelles fournies par les éditeurs de logiciels de gestion de bases de données.
- Les utilitaires de réplication, utilisables si les systèmes de production et décisionnels, sources et cibles, sont homogènes.
- Les outils spécifiques d'extraction.

L'outil doit être à même de lire sélectivement les données sources, et donc de filtrer les données en lecture afin de n'extraire que l'information pertinente.

- **La transformation des données :**

Les ETL sont des ateliers spécialisés dans la migration de données. La transformation des données est leur fonctionnalité principale. Ils doivent disposer d'une fonction permettant de vérifier qu'une donnée est cohérente par rapport aux données déjà existantes dans la base cible. Ils doivent aussi fournir des outils pour convertir les données (par exemple un langage ou une interface graphique de description de transformation). Enfin, ils doivent être conçus pour manipuler de gros volumes de données.

- **Le chargement et le transfert des données :**

Le chargement prend en compte la gestion du format final des données. Pour la mise en œuvre du transfert de données, on distingue deux approches possibles :

- Le transfert de fichiers : l'ETL transporte les données du système source vers le système cible via un moteur.
- Le transfert de base à base. Dans ce cas, les outils travaillent en mode connecté, d'une source de données à une cible. Les données sont extraites ensemble à la source, puis transférées à la cible en y appliquant éventuellement des transformations à la volée. Un seul processus, plus rapide, a ainsi l'avantage de pouvoir effectuer, sans rupture, les transferts et toutes les autres opérations d'alimentation.

9.3 Application des outils ETL à la BI :

La plupart du temps les ETL sont utilisés dans le domaine de la Business Intelligence (BI) décrit auparavant. La majorité des outils ETL mettent à dispositions des outils spécifiques pour alimenter des entrepôts de données. Les clés de substitutions, l'alimentation de cubes OLAP.

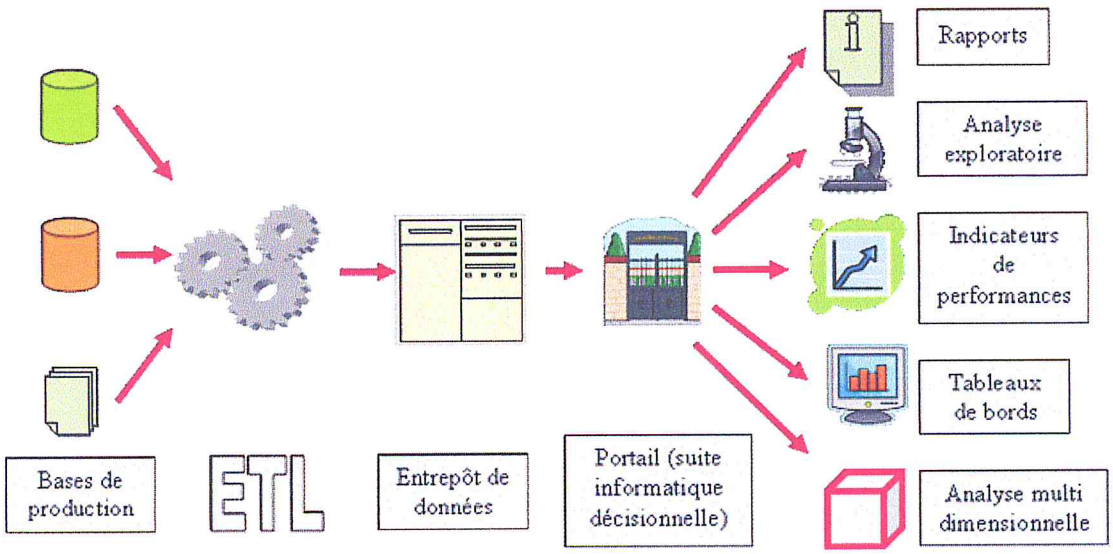


Figure 3.12 : Schéma explicatif de l'utilisation d'outils ETL dans le SID [24]

10. Conclusion :

Dans ce chapitre, nous avons fait présenter le SID, plus précisément nous avons traité le sujet des entrepôts de données qui étendent les concepts et la technologie traditionnelle des bases de données pour créer des systèmes d'aide à la décision.

En utilisant l'architecture d'un entrepôt de données, nous avons expliqué les différents composants qu'il intègre, comme les diverses sources. Nous avons aussi décrit les différents modèles multidimensionnels pour la construction d'un entrepôt de données.

L'avant dernière partie a été consacré aux types de serveurs décisionnels. Dans un premier temps, nous avons décrit le serveur ROLAP, le serveur MOLAP et le serveur HOLAP. Nous avons définit la phase ETL qui est une phase principale dans le SD.

Dans Les prochains chapitres nous allons voir comment appliquer le SD pour l'analyse du fichier journal.

CHAPITRE 4

ARCHITECTURE ET MODELE POUR UN ENTREPOT DE DONNEES FICHIER JOURNAL

1. Introduction :

Tout DW doit être en mesure de répondre aux attentes des utilisateurs. Cela ne peut, évidemment, se faire sans une étude approfondie de leurs besoins.

Et selon la nature des besoins exprimés lors de cette analyse du besoin, une solution nommée « Intranet Analytics » analyse les fichiers journaux selon le système décisionnel pourra être privilégiée. L'application doit présenter des interfaces conviviales et ergonomiques afin de faciliter l'utilisation par un utilisateur qu'il soit spécialiste ou non.
Dans ce chapitre nous allons déterminer l'ensemble des besoins des décideurs, c'est une étape primordiale qui procède la modélisation du DW.

Une fois les besoins identifiés, nous voilà prêts à lancer la conception de notre DW.
La conception d'un entrepôt de données est une tâche complexe et délicate. Elle se compose de plusieurs étapes. Dans un premier temps, nous devons analyser la source de données. Cette analyse sert aussi bien à la sélection de l'ensemble de données à stocker dans l'entrepôt, qu'à la sélection des outils requis pour l'extraction et la transformation de ces données avant leur stockage. La deuxième étape consiste à organiser ces données à l'intérieur de l'entrepôt. Pour ce faire, nous devons concevoir le modèle multidimensionnel à utiliser.

2. Le rôle de l'application :

- L'application consiste à générer des rapports statistiques et à faire des analyses en temps réel.
- Les informations sont consolidées et centralisées dans un seul entrepôt de données.
- L'application pourra par la suite contribuer à d'autres recherches dans le domaine du trafic web.

3. L'analyse de la source de données :

L'étude des sources de données nous a permis d'identifier :

Le fichier Log Access : fichier présent sur le serveur qui liste chronologiquement les évènements exécutés. Chaque requête génère une ligne sur le fichier journal et son analyse donne indication sur le trafic du web.

➢ L'analyse des fichiers logs :

En 2008, l'ONS a mis en place un réseau informatique INTRANET qui utilise un serveur web apache ce dernier génère un fichier log Access pour l'analyse du trafic web. Lors de notre étude nous avons remarqué l'absence d'un outil qui analyse automatiquement ces fichiers.

Notre solution que nous présentons est un outil d'aide à la décision, qui permettra le prétraitement des données inclus dans la phase ETL et stocker les informations extraites dans une base de données multidimensionnelle. Cette solution sera appliquer pour la première fois dans cette organisation dans le but d'avoir une vue précise sur le trafic web dans le réseau.

- **Les fonctionnalités de fichier log Access :**

Le journal des accès au serveur enregistre toutes les requêtes que traite ce dernier.
Bien évidemment, le stockage des informations dans le journal des accès n'est que le point de départ de la gestion de la journalisation. L'étape suivante consiste à préparer et analyser ces informations de façon à pouvoir en extraire des statistiques utiles.

4. Etat du décisionnel au sein de l'entreprise :

Lors de notre étude de l'existant, nous avons constaté l'absence d'un système décisionnel au niveau de l'ONS, tout processus d'analyse et de prise de décision se base essentiellement sur des rapports dont les données sont extraites et consolidées à partir des systèmes transactionnels ou des fichiers.

5. Définition des besoins :

Chaque site dispose d'une problématique et d'enjeux qui lui sont propres. La première étape d'une démarche de l'analyse de l'intranet consiste donc à déterminer les objectifs business liés à son activité. A partir de ces objectifs, il est alors possible de définir les indicateurs clés de la performance de l'activité online et ainsi de pouvoir juger de la performance du site.

L'intranet de l'ONS peut, avec notre solution de mesure d'audience, être surveillé. Comme sur un site web, le repérage des pages les plus visitées, des liens les plus utilisés et des fonctions les plus prisées constitue le meilleur moyen pour analyser l'utilisation réelle de l'intranet. Les informations recueillies ne serviront pas uniquement à améliorer l'ordre des pages ou la présentation, mais permettront d'imaginer de nouvelles fonctions et de faire évoluer l'intranet.

- ✓ **Les mesures :**

- Nombre de pages vues : nombre de pages affichées par le visiteur et dans une période donné (mois par exemple). (Par rapport aux visiteurs)
- Nombre de visiteur unique : qui ne voient qu'une seule page.
- Nombre de visiteurs : nombre d'accès à une page dans une période donné (mois).
- Nombre de changement : Nombre de visiteur qui ont envoyé des données au serveur.
- Volumétrie de données.

✓ **Les indicateurs :**

- Le taux de rebond : est le pourcentage de visiteurs qui quittent le site juste après y être arrivé. Autrement dit, c'est le pourcentage visiteur unique.

- Le taux de transformation : le pourcentage de nombre de changement sur le nombre de visiteur.

- Taux d'intérêt : C'est tout simplement le pourcentage de nombre de page vue diviser par le nombre de visiteurs sur la période (heure, journée ou mois).

6. Modélisation multidimensionnelle :

Notre objectif est de proposer une modélisation multidimensionnelle des données permettant de fournir aux utilisateurs finaux des indicateurs et des états, et d'exploiter à mieux les données stockées au niveau du nouveau système de DW.

6.1 Processus de modélisation :

La modélisation multidimensionnelle est une méthode de conception spécifique aux entrepôts de données, elle se résume essentiellement en quatre étapes dont l'utilité est de **Kimball [34]** :

- **Sélectionner le processus d'activité à modéliser :** Un processus d'activité est une activité normale d'une organisation, généralement assisté par un système source collectant des informations.

- **Déclarer le grain du processus d'activité :** le grain est le plus bas détail auquel les mesures de la table de faits sont représentées, ainsi il aide à détecter les dimensions et les mesures principales qui vont contribuer à la modélisation, le grain répond généralement à la question « comment décrire une ligne unique de la table de faits ? ».

- **Choisir les dimensions :** Cela revient à choisir les dimensions qui s'appliquent à chaque ligne de la table de faits. Les dimensions résultent de la question : « Comment les gestionnaires décrivent-ils les données qui résultent du processus concerné ? ».

- **Identifier les faits numériques :** Les faits sont déterminés par la réponse à la question « Que mesurons-nous ? ».

6.2 Modélisation multidimensionnelle du processus d'activité :

Pour la modélisation de notre DW, nous allons regrouper les données selon les besoins des décideurs cités dans le chapitre précédent.

➢ **Fichier journal (access) :**

Le journal des accès au serveur enregistre toutes les requêtes que traite ce dernier, c'est une fonction très sensible qui présente une grande importance pour les décideurs de l'ONS, son activité principale la plus sensible : L'activité « Surveillance ».

➢ **Modélisation multidimensionnelle de l'activité « Surveillance » :**

- **Présentation de l'activité :**

Surveillance est une activité continue examinant les comportements des divers usagers.

- **Grain de l'activité :**

Tend à connaitre le nombre de pages vues, le nombre de visiteur, le nombre de changement ainsi la volumétrie de données par mois, pages et par visiteurs.

- **Les dimensions de l'activité :**

D'après le grain de l'activité on relève les dimensions « DimDate », « DimPage », « DimVisiteur ».

- **La dimension « DimDate » :**

Cette dimension, permet de faire des regroupements temporels selon le jour, mois et l'année, les utilisateurs ont besoin de suivre leur activité d'un mois à un autre et d'en garder l'historique.

La dimension « DimDate » est caractérisée par les attributs suivants :

Attributs	Désignation
IdDate=DateKey	L'identificateur de la dimension
Annee	Année
Mois	Numéro du mois
Jour	Jour du mois

Tableau 4.1 : descriptif des attributs de la dimension « DimDate »

- **La dimension « DimPage » :**

Cette dimension contient la méthode utilisée soit « GET » ou « POST », l'url et la taille de la page chargée.
La dimension « DimPage » est caractérisée par les attributs suivants :

Attributs	Désignation
IdPage	L'identificateur de la dimension
Methode	la méthode utilisée
Url	url de la page
Taille	Taille de la page chargée

Tableau 4.2 : descriptif des attributs de la dimension « DimPage »

- **La dimension « DimVisiteur » :**

Cette dimension contient l'adresse IP du visiteur.
La dimension « DimVisiteur » est caractérisée par les attributs suivants :

Attributs	Désignation
IdVisiteur	L'identificateur de la dimension
Adresse_Ip	L'adresse IP

Tableau 4.3 : descriptif des attributs de la dimension « DimVisiteur »

- **Les faits de l'activité :**

Les faits que nous avons enregistrés dans notre table des faits "logFile" sont :
« nbr_page_vue » Nombre de pages vues,
« nbr_visiteur » Nombre de visiteurs,
« nbr_chang » Nombre de changement,
« volumetrie_donnees » Volumétrie de données.

- **Le schéma en étoile :**

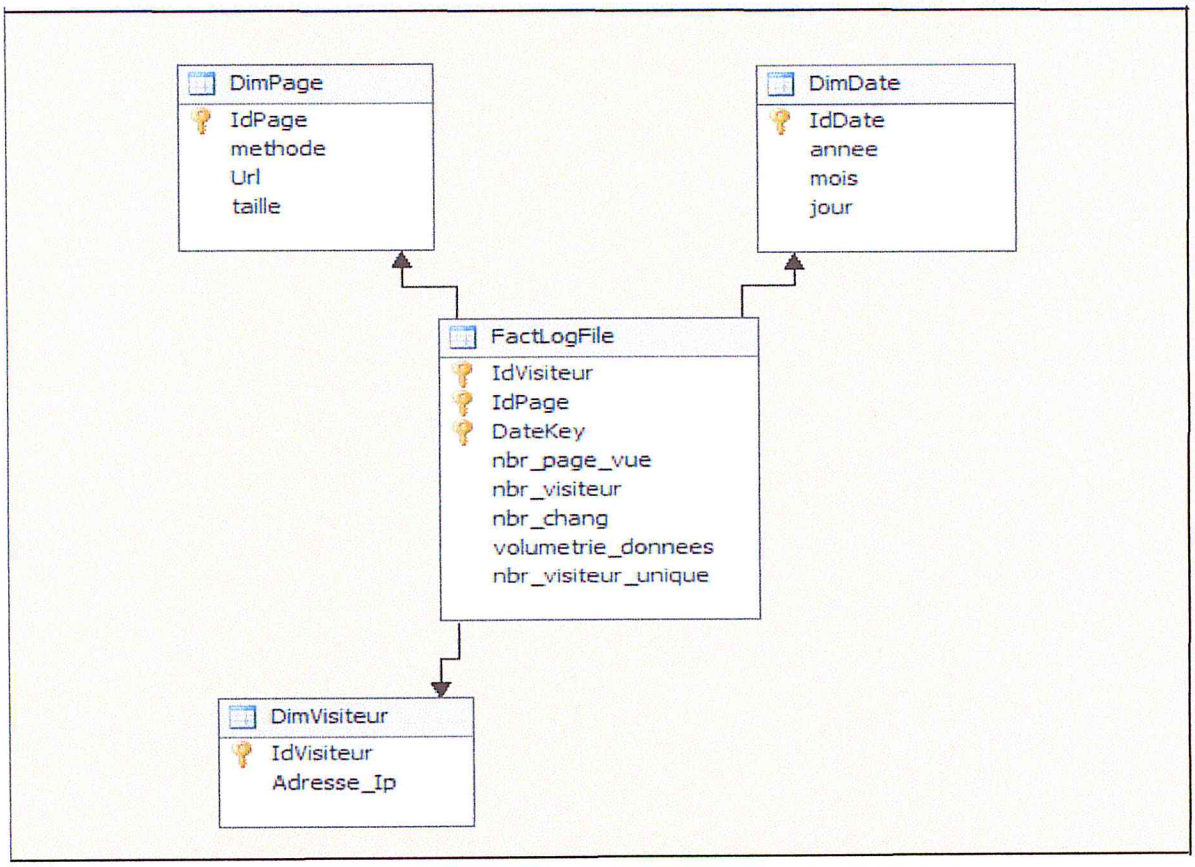


Figure 4.1 : Schéma en étoile de l'activité « Surveillance »

7. Construction de la zone d'alimentation (ETL) :

L'ETL est la phase principale dans le BI, elle est non seulement la plus difficile mais c'est aussi la plus chronophage, nous avons créé notre propre outil ETL compatible avec les fichiers journaux.

7.1 Architecture de l'alimentation :

Le schéma ci-dessous présente clairement l'architecture de la solution d'alimentation adoptée :

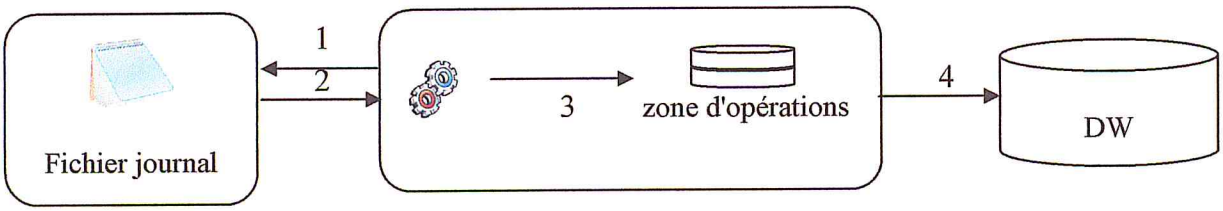


Figure 4.2 : Architecture du processus d'alimentation

1 : Demande de données.
2 : Acquisition de données.
3 : Transformation.
4 : Chargement.

Le processus ETL passe par trois étapes nécessaires qui sont :

- **L'extraction des données :**

Cette étape consiste à extraire les données à partir de fichier journal vers la table intermédiaire « logfile » en utilisant le langage java avec des requêtes sql, il s'agit de charger les données dans cette table pour l'étape suivante.

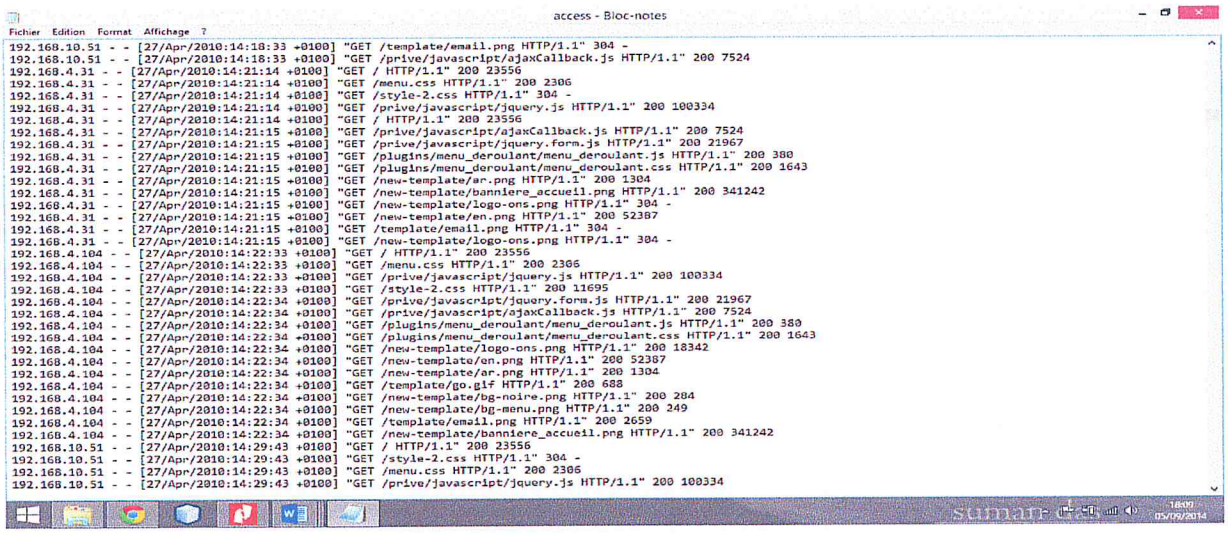


Figure 4.3 : fichier texte journal

Les différents champs de ce fichier vont être importés la table déterminée comme suit :

Résultats | Messages

	Ip	LoginClient	UtilsateurClient	Date_heure	Methode	Url	Protocole	CodeServer	Taille_charger
1	192.168.10.73	-	-	26-12-2013 13:52:56	GET	/ecrire/	HTTP/1.1"	200	35941
2	192.168.10.73	-	-	26-12-2013 13:52:56	GET	/plugins/fckeditor-spip-2/fckconfig.js	HTTP/1.1"	200	15604
3	192.168.10.73	-	-	26-12-2013 13:53:46	GET	/ecrire/	HTTP/1.1"	200	4097
4	192.168.10.73	-	-	26-12-2013 13:53:47	GET	/ecrire/	HTTP/1.1"	200	9980
5	192.168.10.73	-	-	26-12-2013 13:53:47	GET	/prive/javascript/spip_barre.js	HTTP/1.1"	200	5795
6	192.168.10.73	-	-	26-12-2013 13:53:53	GET	/ecrire/	HTTP/1.1"	200	9991
7	192.168.10.73	-	-	26-12-2013 13:53:53	GET	/prive/javascript/spip_barre.js	HTTP/1.1"	200	5795
8	192.168.10.73	-	-	26-12-2013 13:53:59	GET	/ecrire/	HTTP/1.1"	200	81407
9	192.168.10.73	-	-	26-12-2013 13:54:06	GET	/ecrire/	HTTP/1.1"	200	86707
10	192.168.10.73	-	-	26-12-2013 13:54:10	GET	/-Au-31-12-2012,272-html	HTTP/1.1"	200	21347
11	192.168.10.73	-	-	26-12-2013 13:54:17	GET	/IMG/xls/AGE_An_12_marc_page6.xls	HTTP/1.1"	200	32768
12	192.168.10.73	-	-	26-12-2013 13:54:24	GET	/IMG/pdf/AGE_An_12_marc_page6.pdf	HTTP/1.1"	200	44061
13	192.168.10.73	-	-	26-12-2013 13:55:05	GET	/ecrire/	HTTP/1.1"	200	41367
14	192.168.10.73	-	-	26-12-2013 13:55:05	GET	/ecrire/	HTTP/1.1"	200	40469
15	192.168.10.73	-	-	26-12-2013 13:55:07	GET	/ecrire/	HTTP/1.1"	200	35951
16	192.168.10.73	-	-	26-12-2013 13:55:07	GET	/ecrire/	HTTP/1.1"	200	35048
17	192.168.10.73	-	-	26-12-2013 13:55:08	GET	/plugins/fckeditor-spip-2/fckconfig.js	HTTP/1.1"	200	15604
18	192.168.10.73	-	-	26-12-2013 13:56:07	GET	/ecrire/	HTTP/1.1"	200	60848

Figure 4.4 : la table « logfile »

Le fichier log se transforme en une table composée de plusieurs colonnes, chaque colonne correspond à un champ spécifié du fichier LOG :

- La colonne « Ip» correspond aux adresses IP des visiteurs.
- La colonne « login_client » correspond au Nom du serveur utilisé par le visiteur
- La colonne « utilisateur_client » correspond au Nom de l'utilisateur (en cas d'accès par un mot de passe).
- La colonne « date_heure » correspond à la date d'accès.
- La colonne « methode » correspond à la méthode utilisée (GET/POST).
- La colonne « url» correspond au URL demandé.
- La colonne « protocole » correspond au protocole utilisé.
- La colonne « codeServer» correspond au code de retour du serveur (200).
- La colonne « taille_charger » correspond à la taille chargée.

- **La transformation des données :**

Les données du « logfile » vont subir des transformations nécessaires selon les besoins.

- **Le chargement de données :**

Les données préparées seront chargées dans le DW.

Le diagramme d'activité ci-dessous illustre le processus global du chargement du DW, qui débute par le chargement des dimensions, suivi par le chargement des faits.

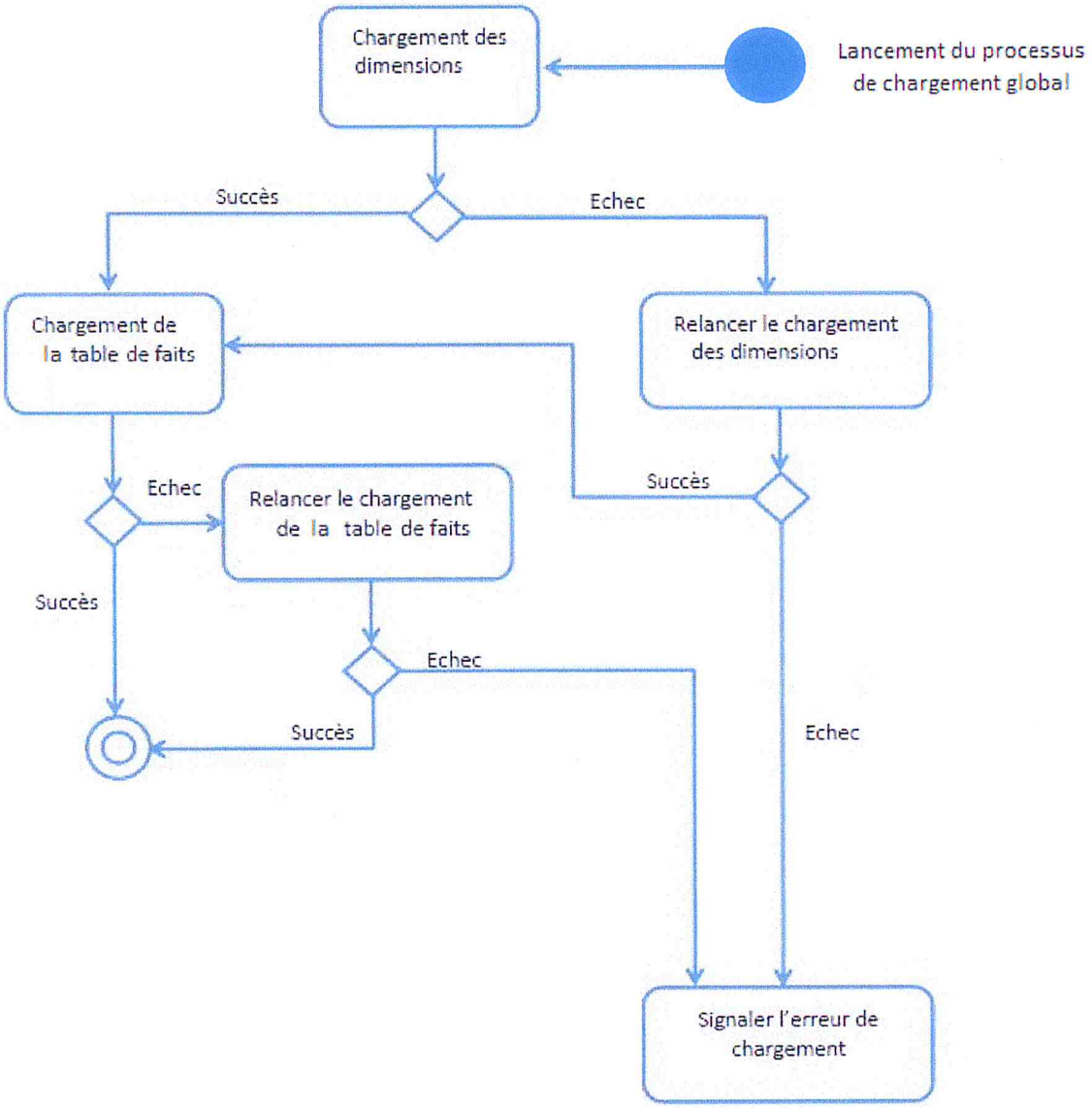


Figure 4.5 : diagramme d'activité global du processus de chargement

7.2 La fréquence d'alimentation :

Un autre aspect important du processus d'alimentation à souligner est la définition de la fréquence du mouvement des données.

L'opération de transformation de la source vers l'entrepôt de données sera exécutée à chaque fin du mois, ce choix est préféré par rapport aux autres périodes. En général, le plus détaillé en termes de période d'analyse est le mois. Notre analyse sur les périodes sera suivie par la hiérarchie suivante : Année → Mois → jour.

8. Conception des cubes OLAP :

Dans cette étape, nous allons créer notre base de données multidimensionnelle, qui est définit comme un cube de données qui permet de croiser des dimensions pour stocker des variables.

Une bonne conception des bases de données multidimensionnelle (Cubes OLAP), permet aux utilisateurs d'accéder de façon rapide, consistante et interactive à un grand volume de données.

La technique de modélisation utilisée pour la mise en œuvre des cubes OLAP est la modélisation MOLAP car elle offre des avantages comme l'efficacité et la rapidité.

8.1 Définition des niveaux et des hiérarchies :

Pour concevoir les cubes, nous allons commencer par définir les différentes hiérarchies des dimensions qui est une étape indispensable et se fait en deux phases :

- Identifier et grouper les attributs par niveau pour chaque dimension.
- Définir toutes les hiérarchies possibles dans une même dimension.

Ces deux phases sont résumées dans le tableau ci-dessous :

Dimension	Attributs	Niveaux	Hiérarchies
Dim_Date	Annee	Niveau 1 :N1	Hiérarchie_1 : H1=N1->N2->N3->N4
	Mois	Niveau 2 :N2	
	Jour	Niveau 3 :N3	
	IdDate	Niveau 4 :N4	
Dim_Visiteur	Adresse_Ip	Niveau 1 :N1	Hiérarchie_1 : H1=N1->N2
	IdVisiteur	Niveau 2 :N2	

DimPage	**Methode**	Niveau 1 :N1	Hiérarchie_1 : H1= N1->N2 ->N3->N4
	url	Niveau 2 :N2	
	Taille	Niveau 3 :N3	
	IdPage	Niveau 4 :N4	

Tableau 4.4 : Niveaux des hiérarchies des dimensions

8.2 Présentation des cubes OLAP :

Nous allons donner la présentation de cube OLAP spécifique au fichier journal.

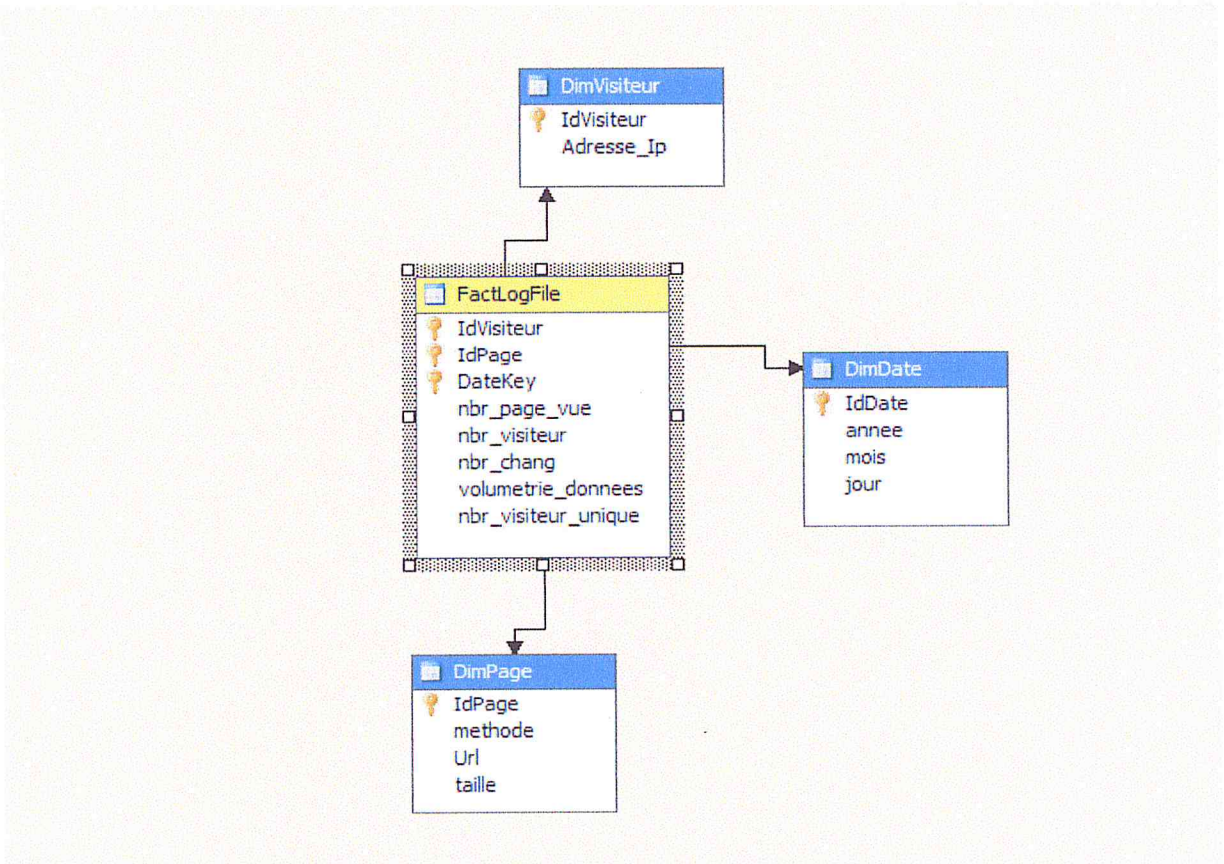


Figure 4.6 : Cube multidimensionnel « fichier log »

9. Conclusion :

Au cours de ce chapitre, nous avons défini les besoins des décideurs, nous avons pu extraire les indicateurs d'analyse à partir de leurs besoins, nous avons aussi présenté la modélisation multidimensionnelle de notre DW, en définissant les faits et les dimensions, ensuite leur implémentation relationnelle en utilisant le modèle en étoile.

Ensuite, nous avons exposé la solution proposée pour le chargement du DW, présenté par un diagramme d'activité.

Et enfin, nous avons mis en place la base de données multidimensionnelle qui se matérialise par l'ensemble des cubes OLAP, nous avons défini pour chaque dimension les hiérarchies possibles pour permettre la navigation facile et rapide.

Dans le chapitre suivant nous allons présenter le détail de la mise en œuvre de notre système, ainsi que les outils qui l'implémente.

CHAPITRE 5

MISE EN OEUVRE ET IMPLEMENTATION

1. Introduction :

Après avoir présenté, dans les chapitres précédents, les différents concepts théoriques liés à notre travail, ainsi que la conception de notre système. Nous nous intéresserons dans ce chapitre à la présentation de la solution réalisée, en s'appuyant sur des illustrations nous montrerons les étapes de création du DW, de son alimentation à l'aide de notre outil, l'étape de l'analyse multidimensionnelle ainsi que la création du tableau de bord sous SQL SERVER BI.

2. Architecture technique de la solution :

La figure suivante illustre la structure et l'architecture technique de la solution proposée :

Figure 5.1 : Architecture technique de la solution

3. Architecture globale de la solution :

Le schéma suivant présente l'architecture globale de notre système décisionnel :

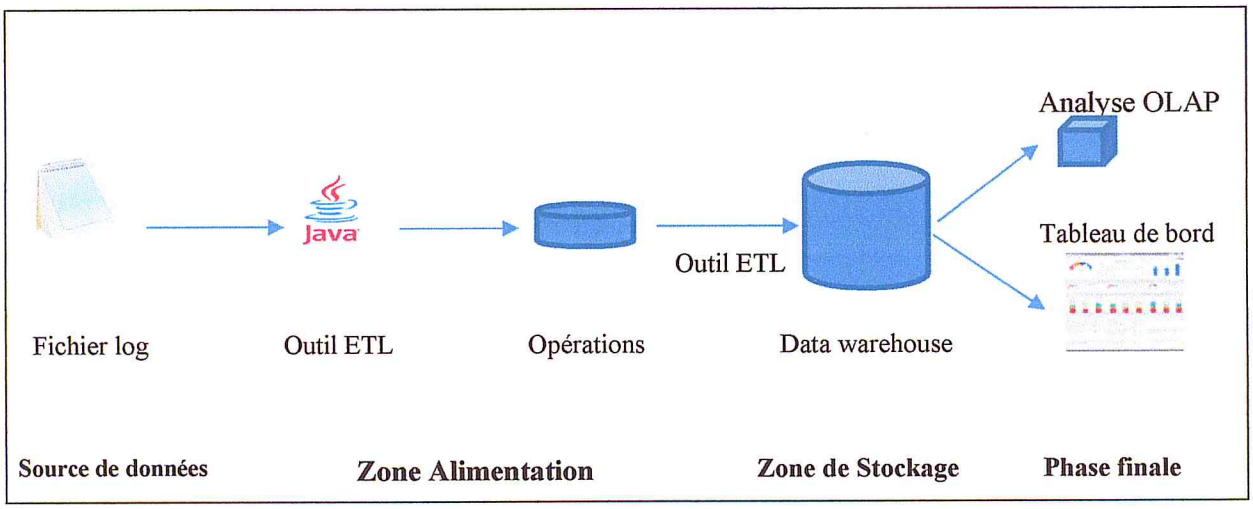


Figure 5.2 : Architecture globale du système décisionnel

4. Construction de la zone de stockage :

Dans la conception, il a été nécessaire de mettre en place deux types de bases de données à savoir : la base de données d'entreposage (DW), la base de données de la zone des opérations (table intermédiaire).

Ces deux bases de données ont été implémentées sous le SGBD **SQL SERVER**, avec une licence de Microsoft. Car SQL Server peut rassembler toutes les données dans un moteur de gestion puissant et supportant la montée en charge qui peut stocker un nombre impressionnant de données et qui supporte les grappes de serveur. Doté de services d'analyse et de *reporting*, il permet de gérer les données en toute sécurité, comme il permet la connexion avec Microsoft Excel.

MICROSOFT SQL Server Business Intelligence Services pour :

- Le *reporting* à la demande
- L'analyse en ligne
- L'entreposage de données
- Les tableaux de bord

5. Construction de la zone d'alimentation :

Parmi les outils ETL existants, notre choix s'est porté sur le développement d'un simple ETL qui permet :

- L'Extraction des données depuis les fichiers journaux.
- La Transformation et nettoyage des données.
- Le chargement (Loading en anglais) des données dans un entrepôt.

Les caractéristiques de cet outil sont :

- ✓ Traitement des données volumineuses (plusieurs lignes dans le fichier).
- ✓ Simple.
- ✓ Compatible avec les fichiers journaux.
- ✓ Le temps de l'exécution est un peu long par rapport aux linges du fichier journal.

- **Création de connexion server :**

La figure suivante nous montre comment nous avons créé une connexion JDBC SQL SERVER.

```
Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");
System.out.println("Driver  ok");
String url="jdbc:odbc:LogJDBC";

Connection cnx=DriverManager.getConnection(url);
```

Figure 5.3 : Méthode de connexion à une base de données

La connexion se fait pour la base de données de la table intermédiaire et le data warehouse de la même façon.

- **L'extraction des données :**

Dans cette étape nous avons chargé le fichier log dans la table « logfile ».

- **La transformation des données :**

  - **Nettoyage des graphiques, images ou scriptes :**

Les données concernant les pages possédant des graphiques, images ou des scripts, n'apporteront rien à l'analyse. Elles seront donc filtrées.

Pour cela nous somme amené à supprimer de notre base de données les URLs suivants :

Les urls correspondant aux images d'extension « .gif » par la requête
(delete from LOGUNIV where url_des_pages like '%.gif ')
Les urls correspondant aux images d'extension « .jpg » par la requête
(delete from LOGUNIV where url_des_pages like '%.jpg ')
Les urls correspondant aux images d'extension « .ico » par la requête
(delete from LOGUNIV where url_des_pages like '%.ico')
Les urls correspondant aux feuilles de styles d'extension « .css » par la requête
(delete from LOGUNIV where url_des_pages like '%.css')
Les urls correspondant aux images d'extension « .png » par la requête
(delete from LOGUNIV where url_des_pages like '%.png')

- **Résolution des pages uniques interprétées différemment par l'interpréteur du serveur :**

Pour les pages d'extension « .php », il fallait résoudre le problème de la page unique mais les interprétées différemment par l'interpréteur du serveur.
Alors l'idée était de modifier la colonne « url » et supprimer pour une même page la partie de sa url qui commence par ' ?' jusqu'à la fin.
Par exemple : supprimer après '?'

/spip.php|||?page=informer_auteur&var_login=admin&var_compteur=1296048710508|||

- **Modification du format date et heure :**

Il existe aussi le problème du format, date et heure. Le format utilisé par le fichier log est incompatible avec les bases de données, il faut donc changer la colonne date_heure en format DATE TIME.
Exemple :

[26/Jan/2011:14:32:07 +0100] → 26-01-2011 14:32:07

- **Alimentation des tables de dimension :**

L'alimentation d'une table de dimension est relativement simple, en utilisant le langage SQL.

La figure suivante montre un exemple d'alimentation d'une table de dimension.

```
PreparedStatement pst = cnt.prepareStatement("INSERT INTO [dw].[dbo].[DimVisiteur] "
 + " ([Adresse_Ip]) VALUES(?) SELECT SCOPE_IDENTITY()");
```

Figure 5.4 : méthode d'alimentation de la table de dimension Visiteur

- **Alimentation de la table de fait**

L'alimentation d'une table de fait se fait comme l'alimentation d'une table de dimension, en utilisant le langage SQL.

```
pst4 = cnt.prepareStatement("INSERT INTO [dw].[dbo].[FactLogFile] "
  + "([IdVisiteur],[IdPage],[DateKey]) VALUES (?,?,?)");
```

Figure 5.5 : Alimentation de la table de fait « **FactLogFile**»

6. Zone de restitution :

La zone de restitution est composée d'un ensemble d'outils qui permettent aux utilisateurs l'exploitation du DW mis en place.

Nous avons fait le choix d'utiliser les outils suivants et des serveurs ont été mis en place :

- Un moteur MOLAP **« visual studio analysis services»** : qui est un moteur de classe mondiale multidimensionnelle analytique intégrée dans SQL SERVER BI destiné pour l'implémentation des cubes conçus pour l'analyse multidimensionnelle, il permet une gestion centralisée des données et les règles métier dans une plateforme sécurisée, évolutive et prête pour l'entreprise.

- Une plateforme BI « **Visual studio reporting services** » : qui est aussi intégrée dans SQL SERVER BI, complète, et répond à l'ensemble des besoins décisionnels. Elle possède un ensemble de fonctionnalités développées en propre : Tableaux de bords, analyse OLAP, rapports de masse.

- Construction des cubes OLAP :

Le cube représente notre base de données multidimensionnelle, il est créé au niveau du serveur SQL. Voici un exemple de cube que nous avons réalisé.

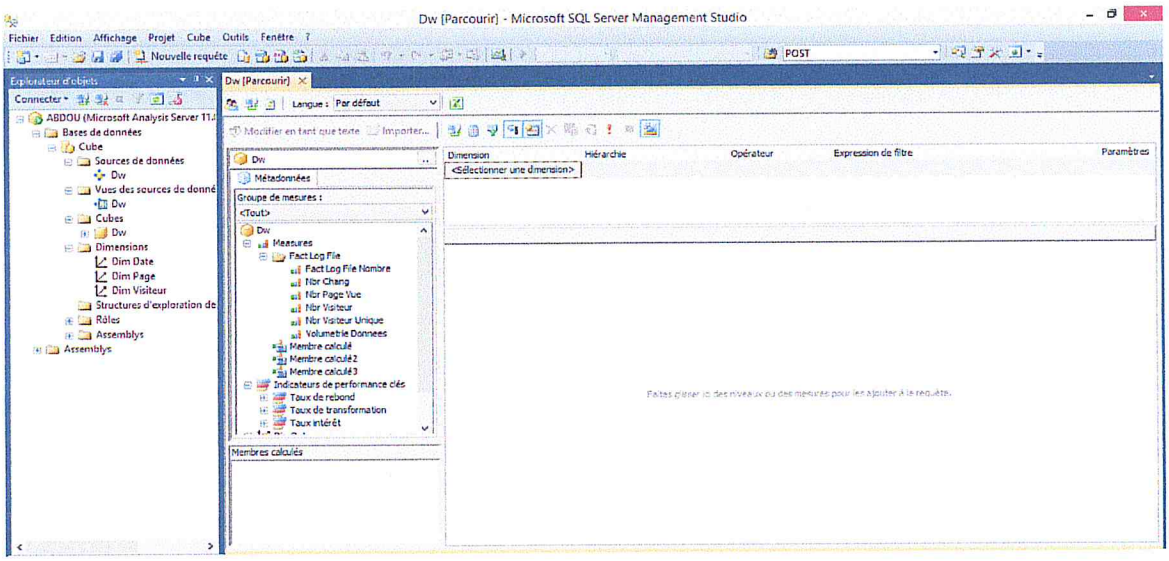
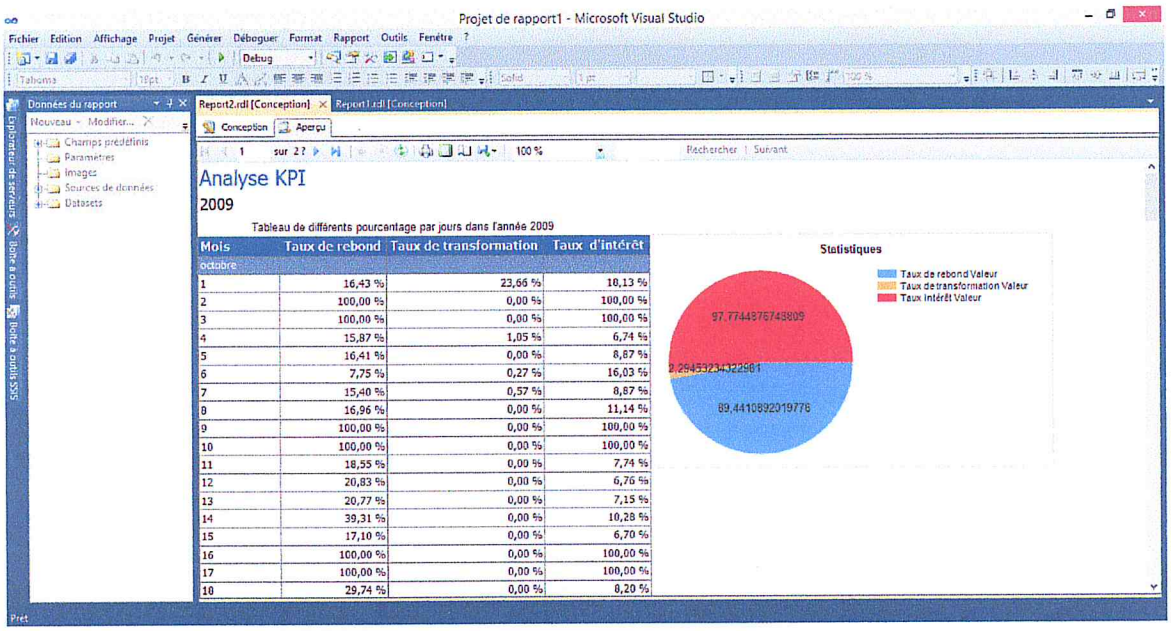


Figure 5.6 : Construction du cube multidimensionnel

- **Construction des rapports :**

La figure suivante montre la construction d'un rapport de de différents indicateurs selon l'année, mois et jour.



Mois	Taux de rebond	Taux de transformation	Taux d'intérêt
octobre			
1	16,43 %	23,66 %	18,13 %
2	100,00 %	0,00 %	100,00 %
3	100,00 %	0,00 %	100,00 %
4	15,87 %	1,05 %	6,74 %
5	16,41 %	0,00 %	8,87 %
6	7,75 %	0,27 %	16,03 %
7	15,40 %	0,57 %	8,87 %
8	16,96 %	0,00 %	11,14 %
9	100,00 %	0,00 %	100,00 %
10	100,00 %	0,00 %	100,00 %
11	18,55 %	0,00 %	7,74 %
12	20,83 %	0,00 %	6,76 %
13	20,77 %	0,00 %	7,15 %
14	39,31 %	0,00 %	10,28 %
15	17,10 %	0,00 %	6,70 %
16	100,00 %	0,00 %	100,00 %
17	100,00 %	0,00 %	100,00 %
18	29,74 %	0,00 %	8,20 %

Figure 5.7 : Un exemple d'un rapport

7. Fonctionnement de la solution conçue :

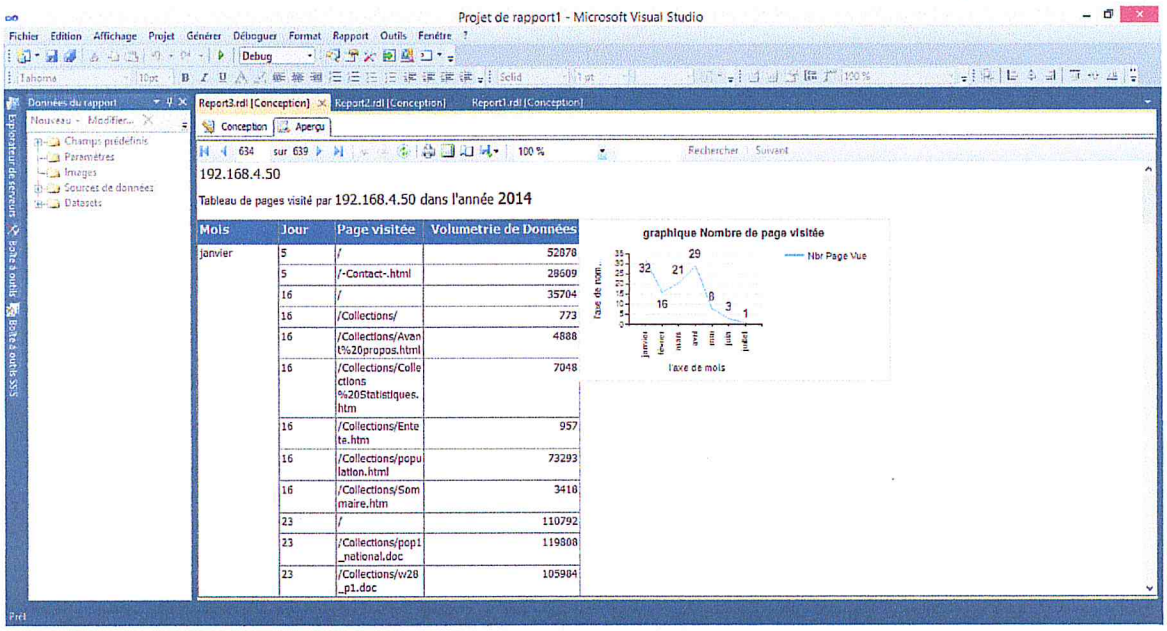
A travers les interfaces présentées ci-dessous, on vise à montrer quelques fonctionnalités proposées par la solution que nous avons mise en œuvre.

- **l'accès à l'application** :

Pour accéder à l'application, il suffit juste de faire entrer le login et le mot de passe de chaque utilisateur.

- **Analyse multidimensionnelle des données avec visual studio analysis services :**

Une fois le cube crée, nous pouvons exploiter les données de la base multidimensionnelle à des fins d'analyses. Par exemple, si nous voulons voir le nombre de page visité, selon les axes d'analyse : Visiteur, année, mois nous aurons comme résultat ce qui est indiqué dans la figure.



Mois	Jour	Page visitée	Volumetrie de Données
janvier	5	/	52878
	5	/-Contact-.html	28609
	16	/	35704
	16	/Collections/	773
	16	/Collections/Avant%20propos.html	4888
	16	/Collections/Collections%20Statistiques.htm	7048
	16	/Collections/Entete.htm	957
	16	/Collections/population.html	73293
	16	/Collections/Sommaire.htm	3418
	23	/	110792
	23	/Collections/pop1_national.doc	119808
	23	/Collections/w28_p1.doc	105904

Figure 5.8 : L'analyse multidimensionnelle OLAP

➢ **Utilisateurs :**

Pour l'utilisation de notre système décisionnel, les utilisateurs sont l'ensemble des décideurs et analystes de l'ONS, ils ont le droit d'accéder à la partie navigation et aux rapports prédéfinis, ainsi qu'aux tableaux de bord.

➢ **La sécurité :**

- **Niveau base de données** : pour assurer la sécurité de l'ensemble des bases de données, nous utilisons la sécurité liée au serveur de la base de données SQL SERVER, Mécanisme d'autorisation et d'authentification permettant de sécuriser l'accès des utilisateurs à la base de données.

- **Niveau utilisateurs** : pour sécuriser l'accès, nous avons utilisé le mécanisme d'autorisation et d'authentification, il s'agit d'attribuer à chaque utilisateur un login et un mot de passe, le droit d'accès est défini dans la plate-forme SQL SERVER BI.

8. Les Avantages et les caractéristiques de notre système :

Notre système permet de :

- Générer des rapports sur l'état du trafic et du comportement des visiteurs.
- La possibilité de surveillance en temps réel.
- Générer des rapports spécifiques à certains utilisateurs, et les résultats visualisables à l'aide d'une interface Web.
- Transférer les données en sécurité, interface mutualisée de consultation des tableaux de bord.
- Optimiser les performances d'un réseau intranet.
- Assurer que les fichiers journaux sont périodiquement revus et analysés.
- Protéger la confidentialité et l'intégrité des données d'un fichier journal.
- Protéger la disponibilité des fichiers journaux (assurer de ne pas perdre des informations).
- Stocker des informations pour une longue période.
- La Rapidité.

9. Conclusion :

Dans ce chapitre nous avons décrit la mise en œuvre de notre système et les outils utilisés, ainsi que la démonstration du système final qui est créé avec Microsoft SQL SERVER BI.

CONCLUSION GENERALE

CONCLUSION GÉNÉRALE

Conclusion générale :

Nous avons tenté dans ce mémoire d'étudier l'analyse du trafic web sur le réseau intranet ce qu'il permet de comprendre le comportement des utilisateurs d'un site intranet en exploitant l'information disponible dans les fichiers Logs. L'enjeu est ici important, à l'heure où le nombre d'utilisateurs de l'Intranet augmentent exponentiellement et par conséquent les données à analyser deviennent de plus en plus volumineuses. D'autre part, les administrateurs des réseaux intranet cherchent à comprendre le comportement des visiteurs de leurs sites pour leur offrir un contenu personnalisé répondant à leurs besoins. L'exploitation des données des fichiers Logs a vue plusieurs méthodes en particulier avec les méthodes de fouille des données adaptées, mais le volume est toujours important.

Dans le cadre de ce mémoire, nous avons développé une méthodologie de prétraitement des fichiers Logs selon le système décisionnel permettant de transformer l'ensemble de requêtes enregistrées dans les fichiers Logs à des données structurées, exploitables et utiles. Stocker les données dans une base de données multidimensionnelle nous a permis de surmonter l'obstacle de la quantité des données et de tirer profit du pouvoir faire le *reporting* pour afficher le tableau de bord. Ce dernier permet aux administrateurs du site étudié de connaître les profils des utilisateurs de leur site afin de personnaliser les services fournis par le site.

Ce travail aurait été plus intéressant si nous avions étudié l'extranet et le site web de l'ONS et disposé des données démo-géographiques sur les utilisateurs du site web étudié. Un visiteur pourra être étudié selon sa navigation sur le site et son profil (d'où il vient mon visiteur ?). Ainsi, des services pourront être personnalisés de façon adéquate à cet utilisateur. Il est aussi possible d'étudier les pages qu'ils n'ont pas pu être affichés car il y'a une erreur, avec la possibilité de déterminer cette erreur.

Bibliographie :

M : Mémoire ;

L : Livre ;

A : Article.

[3] M **Projet de « Claude Berne »** « Etude pour la conception et la mise en œuvre d'un intranet au SCD Lyon 1» Gestion de projet, p6.p32.

[4] M **Jean-Eric PELET & Stéphane MENET** Communication des Organisations « L'INTRANET dans une entreprise ? Pourquoi, comment… », P4.

[5] A Mr.**Gaabouri Youssef** un Project Manager chez **Millipore**. Disponible sur : http://web04.univlorraine.fr/ENSAIA/marie/web/ntic/pages/2004/gaabo.html

[6] A **Intranet** Écrit par Administrateur Mardi, 07 Août 2012 22:57 Disponible sur : http://www.nsdcorporation.com/index.php?option=com_content&view=article&id=56&Itemid=72

[7] A **Gaelle Pennetier** responsable de l'enseignement supérieur, responsable formation, Établissement : Lycée René-Descartes Champs-sur-Marne, « L'intranet Techniques et Enjeux ».

[8] L **Frédéric CRÉPLET | Thomas JACOB** « Réussir un projet Intranet 2.0 », p36.

[10] M **Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava** « Data Preparation for Mining World Wide Web Browsing Patterns ».

[11] L « fouille de données » cour de **Philippe Preux** Université de Lille 3. Disponible sur : http://www.grappa.univ-lille3.fr/~ppreux/fouille/ .

[12] A **C. Michel**, « Caractérisation d'usages et personnalisation d'un portail pédagogique. État de l'art et expérimentation de différentes méthodes d'analyse du Web Usage Mining».

[13] L **Bing Liu « Web Data Mining** Exploring Hyperlinks, Contents, and Usage Data », p450.

[14] M « Technologies du web mining appliquées au E-commerce », p11.

[15] M **Malika Charrad** « Une approche générique pour l'analyse croisant le contenu et l'usage du web WCUM » Disponible sur : *tel.archives-ouvertes.fr/docs/00/51/63/67/PDF/2010CNAM0694_0_0.pdf.*

[16] L **Stéphane Tufféry** « Data Mining and Statistics For Decision Making », 1st edition P638. P639.

[17] A **D.S. Oberoi** « Configurer Squid comme serveur proxy ». Disponible sur : http://www.linuxfocus.org/Francais/March2002/article235.html

[19] M **Doru Tanasa** « Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support », P10.

[20] M **Jaideep Srivastava, Robert Cooleyz, Mukund Deshpande, Pang-Ning Tan** « Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data», P14.

[21] M **Y. Lechevallier, D. Tonasa, B. Trousse, et R. Verde** « Classification automatique : Applications au Web Mining,», p157.

[22] M **Nassim et Mohamed ELARBi etTAHAR DJEBBAR** «Traitement et exploration du fichier Log du serveur web pour l'extraction des connaissances».

[24] M **Florian Francheteau** « Etude des ETL open source », p12.

[25] A **Comment ça marche** « Informatique décisionnelle (Business intelligence) » disponible sur : http://www.commentcamarche.net/contents/307-informatique-decisionnelle-business-intelligence

[26] A **Claude CHRISMENT, Geneviève PUJOLLE, Franck RAVAT, Olivier TESTE, Gilles ZURFLUH « Entrepôts de données »** Disponible sur : http://www.techniques-ingenieur.fr/base-documentaire/technologies-de-l-information-th9/bases-de-donnees-42309210/entrepots-de-donnees-h3870/

[27] A **Laurent CAPORIONDO** « Mise en place d'un système d'information décisionnel dans une entreprise » Disponible sur : http://eduscol.education.fr/ecogest/si/SID

[28] L **Bill Inmon** « Building the Data Warehouse Third edition», p31.

[29] L **Alain Venot, Anita Burgun et Catherine Quantin,** « Informatique Médicale, e-Santé – Fondements et applications » p251. documentation en ligne disponible sur : books.google.fr/books?isbn=2817803388

[30] M **María Trinidad, Serna Encinas** « Entrepôts de données pour l'aide à la décision médicale : conception et expérimentation », p18.

[31] L **Michel Bruley** « Propos sur les SI Décisionnel », Septembre 2011.

[34] L **Ralph Kimball, Margy Ross** « The Data Warehouse Toolkit The Defi nitive Guide to Dimensional Modeling Third Edition ».

[35] M **ELhoussaine ZIYATI** « Optimisation de requêtes OLAP en Entrepôts de Données Approche basée sur la fragmentation génétique ».

[36] L **E.F. Codd, S.B. Codd and C.T. Salley** « Providing OLAP to User-Analysts: An IT Mandate ».

[37] A **Le cahier des charges** « Outils Extract-Transform-Load (ETL) » Disponible sur : http://www.guidescomparatifs.com/etl-integration.asp

[38] M **Inès Gam El Golli** « Ingénierie des Exigences pour les Systèmes d'Information Décisionnels : Concepts, Modèles et Processus La méthode CADWE ».

Webographie :

@ : Site internet.

[1] @ Intranet Wikipedia : **http://fr.wikipedia.org/wiki/Intranet**

[2] @ Office National des Statistiques : **http://www.ons.dz/**

[9] @ Fouille du web Wikipedia : **http://fr.wikipedia.org/wiki/Fouille_du_web**

[18] @ Image publié sur : **http://squid-web-proxy-cache.1019090.n4.nabble.com/Help-with-Understanding-the-access-log-file-td4656930.html**

[23] @ SI Wikipedia : **http://fr.wikipedia.org/wiki/Syst%C3%A8me_d%27information#cite_note-1**

[32] @ OLTP Wikipedia: **http://fr.wikipedia.org/wiki/Traitement_transactionnel_en_ligne**

[33] @ OLAP Wikipedia: **http://fr.wikipedia.org/wiki/OLAP**

[39] @ **http://fr.wikipedia.org/wiki/Audience_d'un_site_Web**

[40] @ **http://fr.wikipedia.org/wiki/Trafic_Internet**

[41] @ **http://www.les-infostrateges.com/actu/12041414/10-outils-de-mesure-d-audience**

[44] M **DA SILVA** « Analyse des données évolutives : application aux données d'usage du Web ».

[42] M **Malika Charrad** « Techniques d'extraction de connaissances appliquées aux données du Web ».

[43] M **Nabila Merzoug et Hanane Bessa** « Application du processus de fouille de données d'usage du web sur les fichiers logs du site cubba ».