République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université SAAD DAHLEB BLIDA 1

Faculté des sciences

Département d'Informatique

Mémoire présenté par :

Agguini Razika

Yallaoui Zineb Yasmine

En vue de l'obtention du diplôme de master

Domaine : Mathématique et Informatique

Filière : Informatique

Spécialité : Informatique

Option : Ingénierie du Logiciels

Thème

Collecte de données de performances via un accès SFTP pour la réalisation d'un BIGDATA pour des raisons d'analyse.

Promoteur : Mr. BALA Mahfoud **Encadreur :** Mr. BELHADJ Lakhdar

Soutenu le : 30 Octobre 2017

Devant le jury composé de :

Mr OULED KHAOUA Mohamed Président.

M^me^ TOUBALINE Nesrine Membre.

Promotion : 2016/2017

Remerciements

Suite à la clôture de notre cursus universitaire, et à la présentation de notre mémoire je tenais à remercier :

En premier lieu ALLAH LE TOUT PUISSANT, de nous avoir donné la volonté et le courage afin d'arriver à la finalité de ce modeste travail.

Nos familles qui nous ont beaucoup soutenus pendant toute notre formation, et qui continueront à nous aider dans tous les projets de l'avenir.

Par ailleurs nous souhaiterons manifester notre reconnaissance particulièrement à notre promoteur Mr BALA.M pour tout le savoir qu'il nous a apporté ainsi que pour nous avoir encadré et dirigé au cours de notre projet de fin d'études.

Nous tenons également à remercier notre encadreur Mr Belhadj.L pour nous avoir acceptés et suivis tout au long de ce projet de fin d'études.

Nous souhaitons remercier aussi Mr Guendouz.M pour sa précieuseaide durant notre stage.

Nous tenons aussi à remercier toute personne qui a participé de près ou de loin pour la réalisation de ce travail.

Plus généralement tout le personnel enseignant du département d'informatique de l'université BLIDA - 1 - qui ont participé à notre formation ainsi qu'à tous les étudiants.

Notre gratitude va aussi à tous les enseignants de nos années précédentes.

Résumé

L'analyse des données a prouvé son importance dans la découverte des connaissances, les prévisions et l'aide à la décision. À l'ère du Big Data, la question se pose souvent de savoir quelles sont les technologies et les architectures les mieux adaptées pour soutenir des processus analytiques à grande échelle. En raison de cette grande taille de données, il devient très difficile d'effectuer une analyse efficace en utilisant les techniques et les architectures traditionnelles. A cet effet, il y'a eu l'apparition des applications d'analyse de données ou Big Data Analytics Applications (BDA Apps) qui constitue un nouveau type d'applications logicielles qui analysent de grandes quantités de données à l'aide de Framework de traitement parallèle (par exemple, Hadoop).

Ce travail s'insère dans une vision à long terme qui débute par une approche légère, qui commence tout d'abord par le développement d'un système distribué de stockage et d'analyse des indicateurs de performance, à partir de données de type Big data (structuré et semi structuré) exploitant de grande quantité de données dans cluster.

Mots clés

Big Data, Applications d'analyse Big data, analyse , stockage, Hadoop, indicateurs de performance, cluster.

Abstract

Data analysis has provedits importance in knowledgediscovery, prevision and the decision support. In the era of Big Data, the question often arises as to what technologies and architectures are best suited to support analyticalprocesses on a large scale. Because of thislarge data, itbecomesverydifficult to make an effective analysisusingtraditional techniques and architectures. For thispurpose, thereis an emergenceof data analysis or Big Data Analytics Applications (Apps BDA), whichis a new type of software applications thatanalyse large amounts of data in applications usingparallelProcessing Framework (Eg Hadoop).

This workis part of a long-term vision thatbeginswith a light approach, which involves first the development of an distributed system for the storage and analysis of KPI (key Performance Indicator) from Big data type data (structured and semi-structured), carrying large amountsof data, and thatwillbe all done in a cluster.

Keywords

Big data analysis applications, Analysis, prevision, Storage, Hadoop, KPI (key Performance Indicator), Cluster.

Sommaire

ملخص

تحليل البيانات قد أثبت أهميتها في اكتشاف المعرفة ودعم القرار. وفي عصر البيانات الضخمة، غالبا ما ينشأ السؤال عن التكنولوجيات الأنسب لدعم العمليات التحليلية الواسعة النطاق. ونظرا لهذا الحجم الكبير من البيانات، يصبح من الصعب جدا إجراء تحليل فعال باستخدام التقنيات التقليدية. تحقيقا لهذه الغاية، كان هناك ظهور تطبيقات تحليل البيانات أو تطبيقات تحليلات البيانات الكبيرة (BDA Apps) وهو نوع جديد من تطبيقات البرمجيات التي تحلل كميات كبيرة من البيانات باستخدام الإطار من المعالجة المتوازية (على سبيل المثال، Hadoop).

هذا العمل هو جزء من رؤية طويلة الأجل تبدأ بنهج خفيف، بدءا من تطوير نظام موزع لتخزين وتحليل مؤشرات الأداء، وذلك باستخدام البيانات الضخمة البيانات (منظمة وشبه منظمة) مع كميات كبيرة من البيانات في نظام موزع.

الكلمات الرئيسية

البيانات الكبيرة، تطبيقات تحليل البيانات الكبيرة، التحليل، التخزين، Hadoop، مؤشرات الأداء،النظام موزع.

Liste des figures

Liste des figures

Liste destableaux

Introduction Générale

Nous avons assisté durant cette décennie à l'émergence de données à grande échelle.Celle-ci est dû essentiellement à l'utilisation de nouvelles technologies et applications telles que les capteurs digitaux, Internet, les réseaux sociaux mais aussi à l'évolution qu'a connu l'entreprisecaractérisée par l'interaction de plusieurs systèmes. L'augmentation en taille des bases de données et les nouveaux besoins d'analyse ont engendré l'essor du phénomène : **Big Data.**

Le but du Big Data est l'extraction d'informations à partir des données. Hadoop, son principal Framework, est un environnement d'exécution distribuée, performant et scalable, dont la vocation est de traiter des volumes de données considérables pour en extraire des informations utiles.

L'exploitation et l'intégration de ces données dans les processus de l'organisation peut permettre à l'entreprise d'améliorer les processus de prise de décision pour adapter en temps réel son approche Marketing et sa relation client, valoriser son image sur le marché, optimiser ses processus de gestion logistique et concevoir de nouveaux produits et services. Les solutions de stockage Big data commencent à être adoptées par les entreprises, qui les utilisent sur leurs périmètres stratégiques pour en tirer **des gains de performance** là où les solutions standards étaient limitantes. **L'intégration de ces solutions reste l'étape clé**.

Les opérateurs de téléphonie sont les principaux acteurs de la communication dans la téléphonie mobile. Ce sont des fournisseurs de réseaux téléphoniquesétablissant le contact entre les consommateurs. Depuis quelques années, les opérateurs de téléphonie ne cessent de progresser dans leurs prestationsLes réseaux mobiles étant soumisinstabilités dues aux types d'équipement radio qu'ils utilisent, ils se retrouvent dans l'obligation d'assurer le bon fonctionnement de cet équipement d'une manière permanente. En effet, avec l'importance de la masse d'informations à gérer et les exigences croissantes de prise de décision dans le domaine de la gestion des données de performance, l'opérateur de téléphonie mobile Ooredoo se trouve confrontée à un contexte qui l'oblige à intégrer les technologies Big Data dans son système d'information.

Pour cela, notre stage qui s'est déroulé à la Direction Ooredoo, et plus précisément dans le service Gestion des performances a pour but de concevoir une solutionqui permettra **la**

collecte, le stockage et l'intégration des données semi structurées qui représentent les indicateurs de performances pour des fin d'analyse et de prise de décision.

Problématique

Les réseaux de télécommunications ont pris de plus en plus d'importance dans notre société. Pour satisfaire au mieux les besoins et les intérêts des clients, les opérateurs doivent pouvoir offrir, au meilleur prix, des services d'excellente qualité.

C'est dans ce cadre que le service Gestion des performances réseau de la direction technologique Ooredoo Algérie qui est responsable de l'analyse des performances du réseau qui a pour but d'optimiser et améliorer son efficacité se trouve confronté au besoin de stockage et d'analyse des indicateurs de performance.

Cette analyse ne se fera que sur un système de stockage de tous les indicateurs de performance provenant de tout l'équipement de l'infrastructure de l'opérateur téléphonique à travers différents axes (temps topologie, région, fournisseur. . .). Le volume de ces indicateurs dans un lapse de temps minimale est considérables.

Objectifs

L'objectif de ce projet est de réaliser un système qui permet la collection, l'intégration, le stockage et l'analysedes données de performance qui représentent les indicateurs de performance, performance management, fault management et configuration management du réseau de télécommunication afin de corréler les différents indicateurs. Dans le but de faciliter l'analyse de ces indicateurs

L'utilisation des Big Data pourrait impacter fortement l'entreprise Ooredoo et ce de façon méliorative, ainsi l'entreprises pourra :

- Améliorer la prise de décision

- Réduire les coûts d'infrastructures informatiques via l'utilisation des serveurs standards et des logiciels open source

- Améliorer les performances opérationnelles

Tout ceci orientera l'entreprise vers une économie centrée sur la donnée.

Organisation du mémoire

Le présent mémoire est organisé en trois parties structurées de la manière suivante : La partie I est dédiée à la définition des concepts de base ayant trait avec les réseaux GSM, lestechnologies du Big Data et la notion d'entreposage de données qui seront présentés respectivement dans les chapitres I, II et III. La partie II, quant à elle, traitera sur la conception de notre système. Le chapitre IVest consacré pour la migration des données sources vers un environnement distribué. Nous présenterons l'analyse des données de l'infrastructure télécoms d'Ooredoo dans le chapitre V. La partie III est consacrée à l'implémentation de notre système. Le déploiement de l'environnement distribué est traité dans le chapitre VI. Le chapitre VII est consacré à la description des applications.

Enfin, la dernière partie conclut ce mémoire en dressant le bilan général de nos contributions et présente les perspectives de recherche que nous envisageons d'étudier à l'avenir.

Chapitre I

Généralités sur les réseaux GSM

I.1. Introduction

La téléphonie révolutionna nos moyens de communication permettant enfin de dialoguer à longue distance. Malgré des débuts difficiles, la téléphonie était devenue, au même titre que l'eau courante ou l'électricité, un service de base. Avec les progrès de l'informatique et des codages numériques, une nouvelle génération se profile ; la télécommunication mobile devenant ainsi un service de masse.

Le GSM (Global System for Mobile communications) est un système cellulaire et numérique de télécommunication mobile. Il a été rapidement accepté et a vite gagné des parts de marché. L'utilisation du numérique pour transmettre les données permet des services et des possibilités élaborées par rapport à tout ce qui a existé.

Le développement des réseaux mobiles n'a pas cessé d'accroitre ; plusieurs générations ont vu le jour (1G, 2G, 3G, 4G et prochainement la 5G) et connu une évolution remarquable, en apportant un débit exceptionnel qui ne cesse d'augmenter, une bande passante de plus en plus large et par conséquent un nombre plus important d'utilisateurs pouvant être supportés.

Tout au long de ce chapitre, nous allons essayer de présenter les caractéristiques principales du système GSM.

I.2. Principe de base d'un réseau mobile :

Le principe de fonctionnement du réseau mobile est basé sur un système cellulaire, c'est-à-dire que les stations de bases sont réparties sur le territoire selon un schéma qui permet à une cellule d'utiliser plusieurs fréquences qui seront différentes de celles des cellules voisines, ces mêmesfréquences seront réutilisées par des cellules suffisamment éloignées de façonàéviter les interférences.

Les systèmes mobiles sont standardisés pour être compatibles entre les réseaux des différents pays et s'interconnecter avec les réseaux de téléphonie fixe. Il existe dans le monde deux grands standards de systèmes mobiles, le standard IS41 d'origine américaine (norme ANSI-41) et le standard GSM, défini dans l'Europe par l'ETSI qui est le plus répandu.

I.3. Evolution des réseaux mobiles :

Les réseaux mobiles ont beaucoup évolué depuis leur apparition dans les années 1970 à nos jours. Cette évolution, de la première a la quatrième génération des réseaux cellulaires, est illustrée à la Figure 1.1.

La première génération des réseaux cellulaires (1G) est apparue vers le début des années 1970 avec un mode de transmission analogique et des appareils de taille relativement volumineuse. Les standards les plus utilisés à l'époque étaient l'AMPS (Advanced Mobile Phone System), le TACS (Total Access Communication System) et le NMT (Nordic Mobile Téléphone).

Le mode de transmission numérique est apparu au début des années 90 avec la deuxième génération des réseaux mobiles (2G). Il devient ainsi possible de transmettre, en plus de la voix, des données numériques de faible volume telles que les SMS (Short Message Service) et les MMS (MultiMedia Message Service). Les standards 2G les plus utilisés sont le GSM, l'IS-95 (Interim Standard-95) qui est base sur le codage CDMA (Code Division Multiple Access) et l'IS-136 (Interim Standard-136) qui se base sur le codage TDMA (Time Division Multiple Access). Le GSM est cependant le standard ayant connu la plus grande percée avec l'utilisation de la bande des 1900MHz en Amérique du Nord et au Japon et de la bande des 900MHz et 1800Mhz sur les autres continents. C'est d'ailleurs sur ce standard que se basent les réseaux GPRS (General Pocket Radio Service : 2.5G) et EDGE (Enhanced Data for GSM Evolution : 2.75G) qui sont venus corriger les faibles débits du GSM (environ 9,6 kbps). Le GPRS propose un débit théorique de 114 kbps permettant ainsi la transmission simultanée de la voix et de données. L'utilisation des applications multimédias est rendue possible parEDGE qui offre des débits allant jusqu'_a 384 kbps.

La troisième génération des réseaux mobiles (3G) est apparue pour établir des normes internationales afin de garantir une compatibilité mondiale, une mobilité globale, la compatibilité avec les réseaux 2G et des débits de 2 Mbps pour une mobilité faible et allant jusqu'à 144 kbps pour une mobilité forte. Les principales normes 3G sont le CDMA2000 et l'UMTS (Universal Mobile Telecommunication System). La norme CDMA2000 est une amélioration de la norme IS-95 et n'est pas compatible avec le GSM. D'autres améliorations ont été apportées plus tard en termes de débits a l'UMTS donnant lieu aux normes HSDPA (High Speed DownlinkPacket Access : 3.5G) qui offre un débits théorique maximum de 14.4

Mbps en ligne descendante et HSUPA (High Speed UplinkPacket Access : 3.75G) offrant un débit théorique maximum en ligne ascendante de 5.76 Mbps **[1]**. Ces deux normes sont regroupées sous le nom de HSPA (High Speed Packet Access).

La quatrième génération (4G) des réseaux sans fil est caractérisée par une mobilité accrue, des services diverses et des débits plus élevés. Elle projette des débits théoriques de 100 Mbps pour une mobilité forte et jusqu'_a 1 Gbps pour une faible mobilité **[2]**. Les principales normes 4G sont le LTE (Long Term Evolution) et le WiMAX. Le LTE a été développe par le groupe 3GPP (ThirdGeneration Partnership Project) et constitue une extension du HSPA.

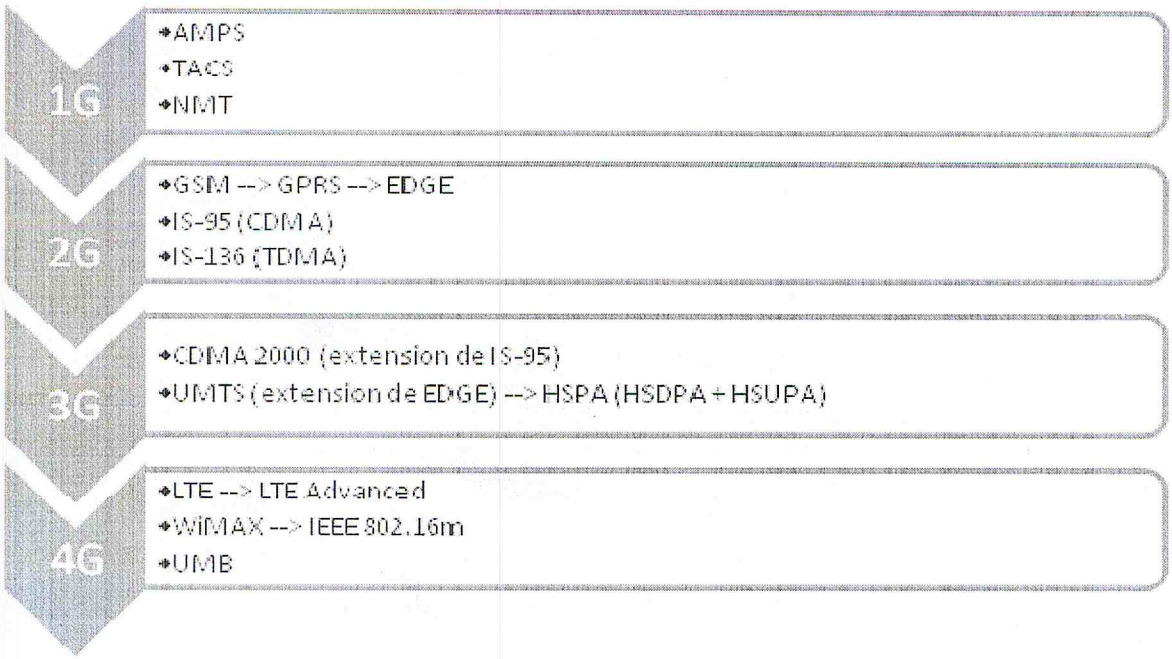


Figure I.1 :*Evolution des réseaux cellulaires***[2]**.

I.4. Présentation de l'infrastructure d'un réseau : [3]

Le réseau GSM a pour premier rôle de permettre des communications entre abonnés mobiles (GSM) et abonnés du réseau téléphonique commuté (RTC – réseau fixe). Le réseau GSM s'interface avec le réseau RTC et comprend des commutateurs.

Le réseau GSM se distingue par un accès spécifique : la liaison radio. Le réseau GSM est composé de trois sous-ensembles :

- Le sous-système radio – BSS (Base Station Sub-system) assure et gère les transmissions radios.
- Le sous-système d'acheminement – NSS (Network Sub System),ou SMSS (Switching and Management Sub-System) pour parler du sous-système d'acheminement. Le NSS comprend l'ensemble des fonctions nécessaires pour appels et gestion de la mobilité.
- Le sous-système d'exploitation et de maintenance – OSS (OperationSub-System) qui permet à l'opérateur d'exploiter son réseau.Ces trois sous-systèmes sont illustrés dans **la figure I.2**.

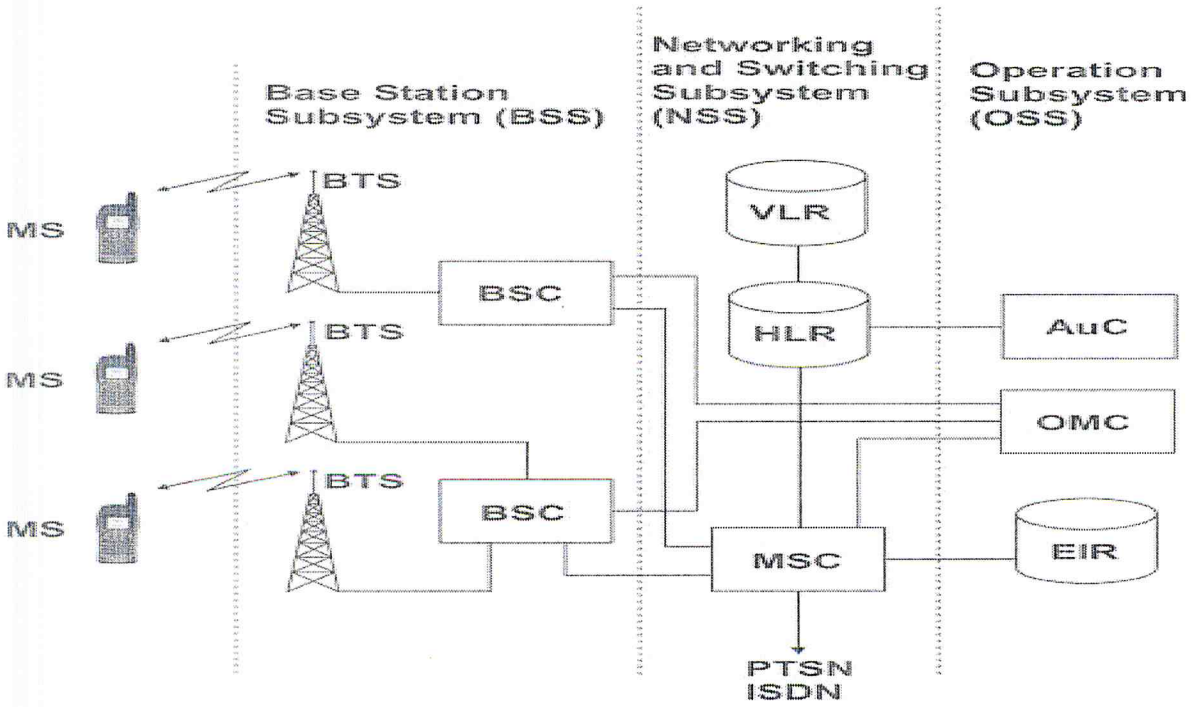


Figure I.2 :*Les trois sous-systèmes du réseau GSM.*

I.4.1.Architecture matérielle du sous-système radio BSS :

Le BSS comprend les BTS qui sont des émetteurs-récepteurs ayant un minimum d'intelligence et les BSC qui contrôlent un ensemble de BTS et permettent une première concentration des circuits. **[2]**

I.4.2.Architecture matérielle du sous-système fixe NSS :

Le NSS comprend des bases de données et des commutateurs.

- **Fonctions du HLR :**

Le HLR (Home Location Register) est une base de données de localisation et de caractéristiques des abonnes. Un réseau peut posséder plusieurs HLR selon des critères de capacité de machines, de fiabilité et d'exploitation. Le HLR est l'enregistreur de localisation nominale par opposition au VLR qui est l'enregistreur de délocalisation des visiteurs **[4]**.

- **Fonctions du VLR :**

L'enregistreur de localisation des visiteurs est une base de données associée à un commutateur MSC. Le VLR (Visitor Location Register)a pour mission d'enregistrer des informations dynamiques relatives aux abonnes de passage dans le réseau, ainsi l'opérateur peut savoir à tout instant dans quelle cellule se trouve chacun de ses abonnés. Les données mémorisées par le VLR sont similaires aux données du HLR mais concernent les abonnés présents dans la zone concernée.

A chaque déplacement d'un abonné le réseau doit mettre à jour le VLR du réseau visite et le HLR de l'abonné afin d'être en mesure d'acheminer un appel vers l'abonné concerné ou d'établir une communication demandée par un abonné visiteur.

Pour ce faire un dialogue permanent est établit entre les bases de données du réseau.

La mise à jour du HLR est très importante puisque lorsque le réseau cherche à joindre un abonné, il interroge toujours le HLR de l'abonné pour connaître la dernière localisation de ce dernier, le VLR concerné est ensuite consultés afin de tracer le chemin entre le demandeur et les demandés pour acheminer l'appel.

- **Fonction du MSC :**

Les MSC sont des commutateurs de mobiles généralement associes aux bases de données VLR. Le MSC assure une interconnexion entre le réseau mobile et le réseau fixe public. Le MSC gère l'établissement des communications entre un mobile et un autre MSC, la transmission des messages courts et l'exécution du handover si le MSC concerné est impliqué. (Le handover est un mécanisme grâce auquel un mobile peut transférer sa connexion d'une BTS vers une autre (handover inter BTS) ou, sur la même BTS d'un canal radio vers un autre (handover intra BTS). On parle de transfert automatique inter/intra cellule Le commutateur est un nœud important du réseau, il donne un accès vers les bases de données du réseau et vers le centre d'authentification qui vérifie les droits des abonnes. En connexion avec le VLR le MSC

contribue à la gestion de la mobilité des abonnés (à la localisation des abonnés sur le réseau) mais aussi à la fourniture de tous les télé services offerts par le réseau : voix, données, messageries ... Le MSC peut également posséder une fonction de passerelle, GMSC (Gateway MSC) qui est activée au début de chaque appel d'un abonné fixe vers un abonné mobile.

Un couple MSC / VLR gère généralement une centaine de milliers d'abonnés. Les commutateurs MSC sont souvent des commutateurs de transit des réseaux téléphoniques fixes sur lesquels ont été implants des fonctionnalités spécifiques au réseau GSM.

I.4.3.Sous système d'exploitation et de maintenance OSS :

Son rôle est L'administration de réseau qui comprend toutes les activités qui permettent de mémoriser et de contrôler les performances d'utilisation et les ressources de manière à offrir un niveau correct de qualité aux usagers. On distingue 5 fonctions d'administrations :

- **L'administration commerciale** : La déclaration des abonnés et des terminaux, la facturation, les statistiques ...
- **La gestion de la sécurité :** La détection des intrusions, le niveau d'habilitation ...
- **L'exploitation et la gestion des performances :** L'observation du trafic et de la qualité (performance), les changements de configuration pour s'adapter à la charge du réseau, la surveillance des mobiles de maintenance ...
- **Le contrôle de configuration du système :** Les mises à niveau de logiciels, les introductions de nouveaux équipements ou de nouvelles fonctionnalités ...
- **La maintenance :** Les détections de défauts, les tests d'équipements ...Le système d'administration du réseau GSM est proche du concept TMN qui a pour objet de rationaliser l'organisation des opérations de communication et de maintenance et de définir les conditions techniques d'une supervision économique et efficace de la qualité de service.

- **Fonctions de l'EIR (Equipement Identity register)**

L'EIR est une base de données annexe contenant les identités des terminaux. Un terminal est identifié par un numéro de série dénommé IMEI (IMEI = numéro d'homologation (série). Numéro d'identifiant. Numéro du terminal). La base EIR est consulté lors des demandes de services d'un abonné pour vérifier si le terminal utilise est autorisé à fonctionner sur le réseau.

- **Fonctions de l'AUC :**

Le centre d'authentification AUC (AUthentification Center) mémorise pour chaque abonné une clé secrète utilisée pour authentifier les demandes de services et pour chiffrer (crypter) les communications. L'AUC de chaque abonne est associe au HLR. Pour autant le HLR fait partie du « sous-système fixe » alors que l'AUC est attaché au « sous-système d'exploitation et de maintenance ».

- **Présentation de l'OMC et du NMC :**

Deux niveaux de hiérarchie sont définis dans la norme GSM. Les OMC (Operations and Maintenance Center) et le NMC (Network and Management Centre). Cette organisation a été définie afin de permettre aux opérateurs télécoms de gérer la multiplicité des équipements (émetteurs, récepteurs, bases de données, commutateurs ...) et des fournisseurs. Le NMC permet l'administration générale de l'ensemble du réseau par un contrôle centralisé.

Les OMC permettent une supervision locale des équipements (BSC /MSC / VLR) et transmettent au NMC les incidents majeurs survenus sur le réseau. Les différents OMC assurent une fonction de médiation.

I.5. Les indicateurs de performance :

Afin de permettre aux opérateurs d'obtenir des informations sur la qualité du service offert par leur réseau et de l'optimiser, des indicateurs de performance appelés KPIs (*Key Performance Indicators*) qui spécifient le fonctionnement radio des cellules ont été également définis.

En effet, un KPI est une valeur représentative permettant d'évaluer la performance de système. Cette valeur est obtenue à partir d'une ou de plusieurs mesures brutes relevées par des compteurs spécifiques. Ces indicateurs permettent la localisation des anomalies de réseau et par suite, l'identification et le diagnostic des causes de ces problèmes afin de réagir avec des actions correctives adéquates.

Dans le but d'offrir une qualité de service acceptable il faut que certains problèmes doivent être résolus. Ces problèmes sont principalement liés à :

a) La couverture :

Ce problème ne peut pas être détecté par le système mais évalué par les plaintes des abonnées et par les mesures radio. Les causes probables de ce problème sont les suivants :

- Mauvaise configuration du réseau c'est-à-dire problème lié à la position des sites, ou les types d'antennes.
- Problème d'installation qui peut être due à la perte des puissances dans les câbles.
- Problème de maintenance.

b) La disponibilité du réseau :

C'est la probabilité d'obtention d'un nouvel appel. La diminution du taux d'appels aboutis implique que les abonnées ne peuvent pas établir une communication. Les actions de l'échec d'établissement d'appel s'expliquent par :

- Le niveau d'accès minimum dans la cellule.
- L'interférence et la mauvaise couverture radio.

c) La qualité de voix :

L'opérateur agit contre le problème de la mauvaise qualité de communication, par les mesures système et par les analyseurs de la qualité vocale. Les causes de dégradation de la qualité de la voix sont :

- La hors couverture.
- La mauvaise installation.
- La qualité des terminaux.

d) Les coupures d'appels :

La coupure de communication peut être engendré par :

- La mauvaise couverture.
- Les interférences.

I.6. Conclusion :

Le département de Gestion de performance réseau contribue à l'amélioration de la qualité du réseau et son optimisation grâce à la gestion et l'analyse des indicateurs de performance ainsi que des caractéristiques de qualité.

Un volume considérable d'indicateurs est généré continuellement, l'analyse de ces données massives nécessite une solution de stockage Big data. Le chapitre suivant est consacré au concept Big data est ses différentes caractéristiques.

Chapitre II

Le Concept Big Data

II.1. Introduction :

Pour stocker et analyser des données issues de bases relationnelles (données structurées), les sociétés avaient recouruà l'entreposage de données (data warehousing).

Face à l'explosion du volume des données non structurées qui représente 80%des données de l'entreprise, la variété des données ainsi que la fréquence avec laquelle ces données sont générées, les SGBD traditionnels et les entrepôts de donnéessont considérés aujourd'hui inadaptés vu les contraintes en matière de cohérence les performances très faibles lors des processus d'analyse .

Pour répondre à cette problématique, un nouveau concept est apparu : le Big Data.Il s'agit d'un environnement distribué et scalable pour le traitement et le stockage de données à grande échelle.Il a pour principal but d'améliorer les temps d'exécution des traitements. **[5]**

II.2. Le concept Big Data :

II.2.1. Définition :

De multiples définitions existent, cependant, **aucune définition précise ou universelle ne peut être donnée au Big Data.** Etant un objet complexe polymorphe, sa définition varie selon les communautés.

« Big data » est un terme qui décrit l'évolution de tout montant exponentiel et la disponibilité de données structurées, les données semi-structurées et non structurées qui ont le potentiel pour être exploitées afin d'avoir des informations spécifiques **[6]**.

Selon M. Lessard, « Big Data » est une expression qui circule depuis quelque temps dans la niche hi-tech de l'informatique dématérialisée (informatique dans les nuages ou cloud computing) et qui fait référence aux outils, processus et procédures permettant à une entreprise de créer, manipuler et gérer de très larges quantités de données **[7]**.

Selon O'Reilly2, « Big Data sont des données qui dépassent la capacité de traitement des systèmes de bases de données classiques. Les données sont trop grosses, se déplacent trop rapidement, ou ne correspondent pas au rétrécissement des architectures de bases de données. Pour gagner de la valeur à partir de ces données, une autre façon de traitement s'impose. » **[8]**.

David Kellogg, quant à lui, définit simplement les Big Data comme étant « trop grandes pour être raisonnablement traitées par les technologies traditionnelles. » **[6]**.

David Kellogg, quant à lui, définit simplement les Big Data comme étant « trop grandes pour être raisonnablement traitées par les technologies traditionnelles. » **[6]**.

En résumé de ce qui a été dit concernant les Big Data, « Il s'agitd'une nouvelle technologie qui est apparue et s'est développée au cours de la dernière décennie exprimant la grande évolution dans le monde Hi-Tech et qui a poussé les chercheurs à mettre à niveau les anciens environnement et outils de traitement ».

II.2.2. Caractéristiques du BIG DATA :

Le big data peut être défini par ses cinq V qui font référence à cinq éléments clés à prendre en compte et à optimiser dans le cadre d'une démarche d'optimisation de la gestion du big data. : Le **Volume, la Vélocité**, la **Variété,** la **Véracité** et la**Valeur** des données.

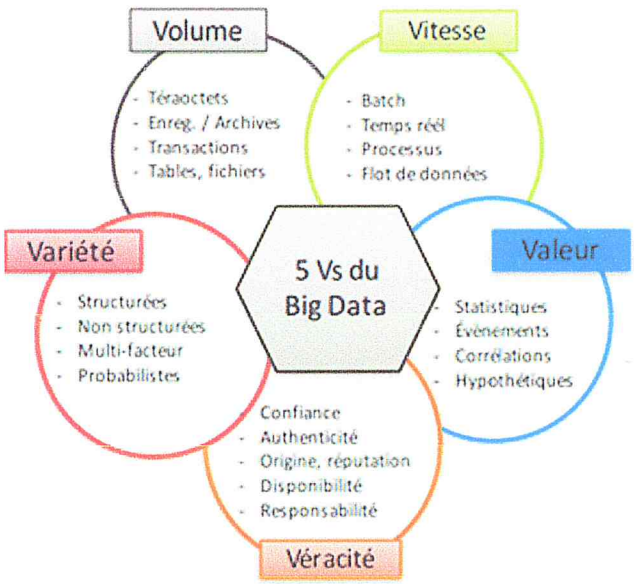


Figure II.1 :*Les 5V du Big Data.*

- **Volume**

Le Volume fait référence à la masse des données d'une part auxquelles nous avons accès, mais également que nous générons. Ce volume augmente à un rythme exponentiel. Ainsi desMéga-octets sommes-nous passés aux Giga-octets, puis aux Téraoctets, aujourd'hui aux Péta-octets, , demain aux Zéttaoctetset bientôt au Yottaoctets. Le Big data offre des outils pour stocker, accéder, et surtout analyser les données à grande échelle.

- **Vitesse**

applications en temps réel.Le Big data se doit d'être performant pour analyser la donnée, même si elle n'est pas dans nos bases de données. **[8]**.

- **Variété**

La variété fait référence à la diversité des formats des données. Le format classique étant celui de la base de données relationnelle, dans laquelle l'information est stockée selon un schéma rigide et organiséesous une forme tabulaire. La donnée est alors qualifiée de 'structurée'. Cependant, aujourd'hui plus de 80 % (certains analystes évoquent 95 à 99% !) de la donnée est qualifiée de 'non-structurée'oùfigurent le texte, le courriel, la photo, la vidéo, la voix, la messagerie, etc. Le big data offre la capacité de réunir toutes ces données et de les analyser.**[8].**

- **Véracité**

La véracité fait référence à la faible fiabilité et au désordre qui règne dans la donnée. Celle-ci manque trop souvent de qualité et de précision, ce qui la rend peu contrôlable. L'une des missions du big data est d'apporter un peu d'ordre à tout cela non pas en organisant la donnée, mais plutôt en organisant son accès et en permettant d'y associer les analytiques qui correspondent aux besoins des utilisateurs.

- **Valeur**

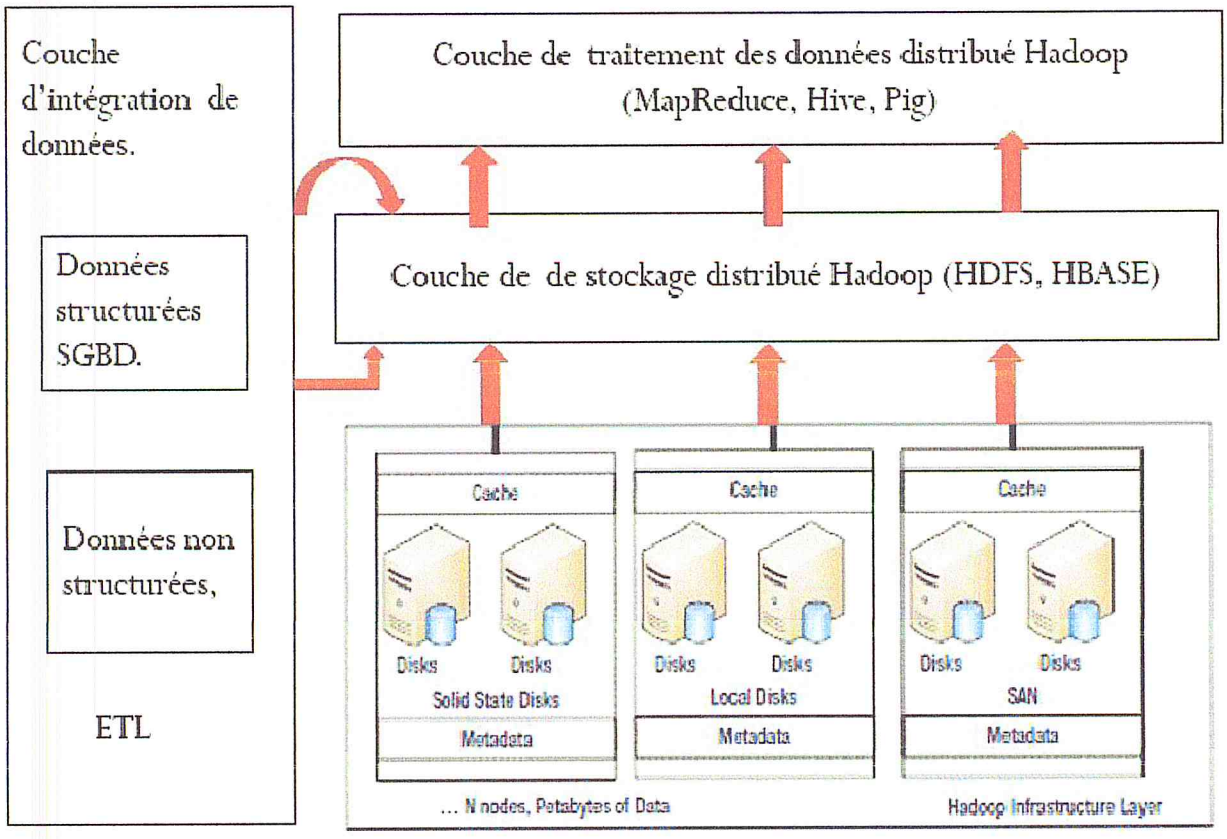
Expriment le besoin de la disposition de données pertinentes et significatives, pour donner suffisamment de sens et d'intérêt économique des analyses menées. La valeur recouvre en effet plusieurs spectres nécessitants chacun une analyse spécifique. On parlera ainsi de valeur d'impact sur un contexte, de valeur de modélisation, de valeur de prédiction, de valeur de management, de valeur économique ou de revente...etc. **[9]**.

Au final, ces 5 V permettent de définir le Big Data. Un outil qui permet de gérer des données qualitatives et volumineuses, très diverses, et traitées en temps réel. Surtout, un outil qui assure à l'utilisateur une visibilité des données, via des outils de reporting, qui lui permette de prendre de bonnes décisions.

II.2.3. Architecture Big Data :

La figure II.2 décrit les composants de l'architecture qui devraient faire partie de toute solution big data de pointe. On peut choisir dans chaque couche des Framework open source ou des produits packagés. Dans cette architecture on distingue principalement les couches suivantes : **[6]**

- **Couche matériel (infrastructure layer) :**comprend des serveurs virtuel VMware, ou des serveurs lame blade.
- **Couche stockage (storage layer) :** les données seront stockées soit dans une base
- NoSQL, ou bien directement dans le système de fichier distribué (DFS :Distributed File System)**Couche management et traitement :** on trouve dans cette couche les outils de traitement et analyse des données comme MapReduce ou Pig.
- **Couche visualisation :**Offre des outils de restitution et de visualisation des résultats



II.3. Le paradigme Map Reduce :

II.3.1. Définition :

MapReduce est un paradigme (un modèle) attrayant pour le traitement des données en parallèle, dans le calcul de haute performance dans un environnement en cluster **[11]**.

L'évolutivité de MapReduce s'est avérée élevée du fait que le travail est partitionné en de nombreuses petites tâches, en cours d'exécution sur plusieurs ordinateurs, dans un cluster à grande échelle.

Le modèle vise aussi à généraliser les approches existantes pour produire une approche unique, applicable à tous les problèmes. Il est conçu pour traiter de grands volumes de données en parallèle et cela en divisant le travail en un ensemble de tâches indépendantes.

MapReduce existait déjà depuis longtemps, dans les langages fonctionnels (Lisp, Scheme)**[12]**, mais la présentation du paradigme sous une forme rigoureuse, généralisable à tous les problèmes et orientée calcul distribué, est attribuable au département de recherche de Google qui a publié en 2004 un article sous le thème : « *MapReduce : Simplified Data Processing on Large Clusters* ».

Un des objectifs du modèle MapReduce est la répartition de charge de calcul sur les machines qui constituent le cluster. Le but est d'utiliser suffisamment de ressources tout en optimisant le temps de calcul et maintenir la fiabilité du système. MapReduce permet de :

— Traiter de grands volumes de données.

— Gérer plusieurs processeurs.

— La parallélisation automatique.

— L'équilibrage de charge.

— L'optimisation sur les transferts disques et réseaux.

— L'ordonnancement des entrées / sorties.

— La surveillance des processus.

— La tolérance aux pannes.

II.3.2. Principe de fonctionnement de Map Reduce :

Représentés sur la Figure II.3 MapReduce définit deux opérations différentes à effectuer sur les données d'entrée :

Mappage :

La première opération « MAP », écrite par l'utilisateur, dans un premier lieu transforme les données d'entrée en une série de couples « clef, valeur ». Ensuite elle regroupe les données en

Mappage :

La première opération « MAP », écrite par l'utilisateur, dans un premier lieu transforme les données d'entrée en une série de couples « clef, valeur ». Ensuite elle regroupe les données en les associant à des clefs, choisies de manière à ce que les couples « clef, valeur » aient une signification par rapport au problème à résoudre. En outre, l'opération « MAP » doit être parallélisable, les données d'entrée sont découpées en plusieurs fragments, et cette dernière est exécutée par chaque machine du cluster sur un fragment distinct **[12],[11]**.

Réduction :

La seconde opération « REDUCE », également écrite par l'utilisateur applique un traitement à toutes les valeurs de chacune des clefs différentes produite par l'opération « MAP ». À la fin de l'opération « REDUCE », on aura un résultat pour chacune des clefs différentes.
L'ensemble de valeurs pour une clef donnée est fusionné pour former un plus petit ensemble de valeurs. Habituellement, juste zéro ou une valeur de sortie est produite par l'invocation de la réduction. Cela permet de gérer des listes de valeurs qui sont trop volumineuses pour tenir dans la mémoire **[12] ,[11]**.

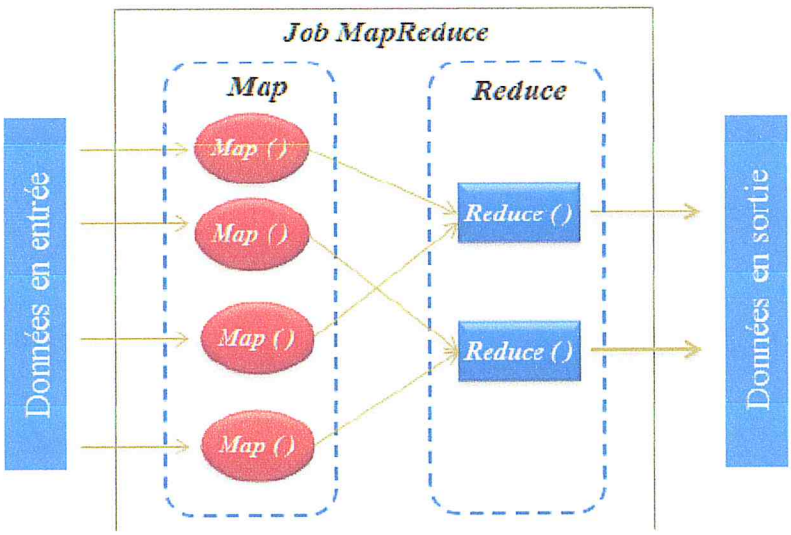


Figure II.3 :*Les deux opérations essentielles dans le modèle MapReduce [13].*

II.4. Modèle sur MapReduce :

L'exemple du compteur de motsdans un fichier texte, le programme compte le nombre de mots, le fichier input est exemple.txt. **[14][15]**.

LafonctionMap() s'écrit de la manière suivante : Map(clé1, valeur1)→List(clé2,valeur2). À partir d'un couple clé/valeur, la fonction Map() retourne un ensemble de nouveaux couples clé/valeur, cet ensemble peut être vide, d'une cardinalité un ou plusieurs. Dans notre exemple sur la Figure II.4ça donne :

Retour de la fonction Map() :(Apple,1) | (Orange,1) | (Apple,1) | (Peach,1)|(Orange,1) | (Apple,1) | (Peach.1)|(Apple,1) Lafonction Reduce() s'écrit de la manière suivante: Reduce(clé2, List(valeur2))→Lis, (valeur2).

À partir des groupesde valeurs associées à une clé, la fonction Reduce retourne généralement une valeur ourien,bien qu'il soit possible de retourner plus d'une valeurSuite à l'appel de la fonction Reduce(),résultat de l'exemple est le suivant : Retour de la fonction Reduce : (Apple,4) | (Orange,2) | (Peach,2)

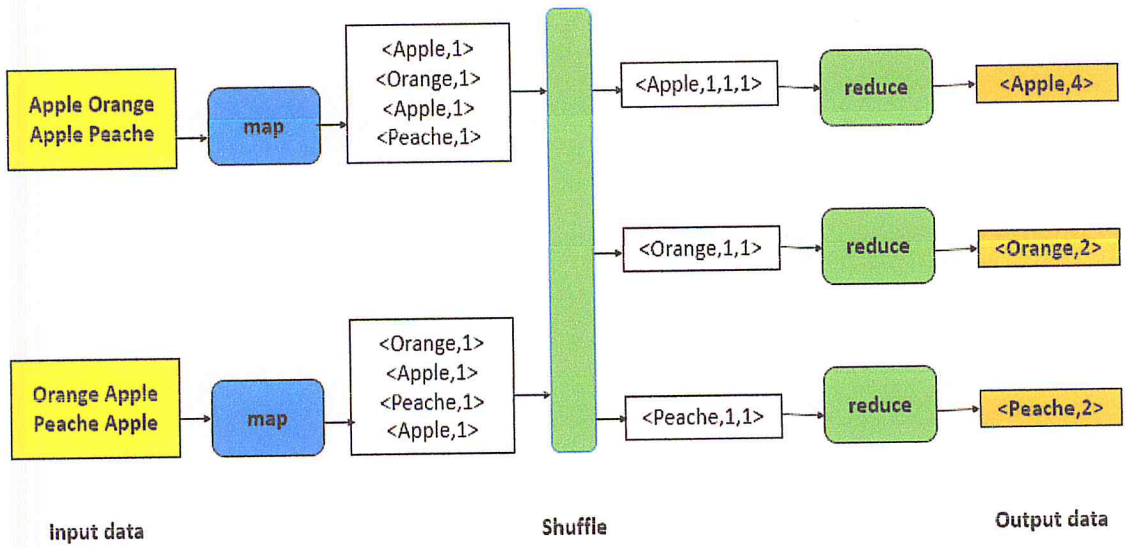


Figure II.4: *Exemple d'un programme MapReduce (WorldCount).* **[12].**

II.5. Le modèle de donnéesNoSQL :

II.5.1. Définition :

Les bases de données No-SQL sont une solution de stockage de données indexées qui contrairement à ses prédécesseurs relationnels de type Oracle, Sybase, MySQL etc., ne répondent pas à une définition précise. En somme, elles prennent des libertés avec le paradigme ACID (atornicity, consistency, isolation, durability) si cher aux administrateurs de bases de données relationnelles pour offrir en contrepartie des performances inédites et des factures énergétiques réduites. **[16]**

II.5.2. Les avantages du modèle No-SQL :

Les principaux avantages des bases de données No-SQL sont : **[16]**

- Leurs performances ne s'écroulent jamais quel que soit le volume traité. Leur temps de réponse est proportionnel au volume (on observe une dérive quadratique dans les SGBDR classiques).
- Elles se migrent facilement. En effet, contrairement aux SGBDR classiques, il n'est pas nécessaire de procéder à une interruption de service pour effectuer le déploiement d'une fonctionnalité impactant les modèles de données.
- Elles sont facilement scalables. A titre d'exemple, le plus gros cluster de No-SQL ait 400 To tandis qu'Oracle sait traiter jusqu'à une vingtaine de Téraoctet (pour es temps de réponse raisonnable).
- Elles sont consistantes de manière pratique (pour l'utilisateur une requête aura toujours la même réponse quel que soit le nœud du cluster).
- Elles s'intègrent facilement aux SI déployés dans les Clouds du marché.
- Elles possèdent un modèle extensible (le nombre de colonne d'une table n'est pas défini)
- Des solutions open source et gratuite.

Les principales bases de données No-SQL sont les suivantes : **[16]**

- **MangoDB :**

La plus populaire des bases No-SQL est écrite en C et n'utilise pas de machine virtuelle JAVA. Cette base est idéale pour débuter car elle est à la fois polyvalente et simple. Aussi à

l'aise pour le stockage massif de données que pour les développements rapides orienté web. Elle possède également une documentation de premier ordre.

- **CASANDRA :**

Cassandra est le projet open source qui découle de la technologie de stockage Facebook, à l'origine il a été écrit spécifiquement pour répondre à la croissance explosive de cette entreprise. Il est assez complexe à configurer, mais il per- met d'adresser toutes les situations où la performance et le traitement de la volumétrie est critique. Cassandra est écrite en JAVA.

- **Hbase :**

Hbase est inspirée des publications de Google sur BigTable. Comme BigTable, elle est une base de données orientée colonnes. Basées sur une architecture maître/esclave, les bases de données de ce type sont capables de gérer d'énormes quantités d'informations (plusieurs milliards de lignes par table.

II.5.3. Types des bases de données NoSQL :

a. Les entrepôts clé-valeur :

Un entrepôt clé-valeur (ECV) peut être envisagé comme une collection de tables de hachage persistantes c'est-à-dire comme une collection de couples clé-valeur persistées sur disque. La valeur en question peut être un morceau d'information sans aucune structure a priori. Il peut s'agit d'un nombre, d'un fichier XML, d'une vidéo, d'une grappe d'objets métiers sérialisés ou encore d'un fichier texte sans structure particulière **[5]**.

Les bases de données NoSQL fonctionnant sur le principe clé / valeur sont les plus basiques que l'on retrouve. On peut les apparenter à une sorte de HashMap, c'est-à-dire qu'une valeur, un nombre ou du texte est stocké grâce à une clé, qui sera le seul moyen d'y accéder. Leurs fonctionnalités sont tout autant basiques, car elles ne contiennent que

Les opérations que l'on peut effectuer sur un tel entrepôt sont :

- la récupération de la valeur pour une clé donnée.

- la mise à jour.

- la création ou la suppression d'une valeur pour une certaine clé.

- stocker toutes les données dans un seul espace est généralement considéré comme une mauvaise pratique car, chaque agrégat ayant potentiellement une structure différente.

- les risques de conflits entre clés augmentent.

- **Usage :**

Le stockage de données de session utilisateur est le cas typique d'utilisation d'un entrepôt clé valeur. De manière similaire, le stockage des préférences ou des profils des utilisateurs ou encore les associations entre clients et paniers d'achat des sites d'e- commerce conviennent bien aux ECV. La structure des données en table de hachage les rend cependant inadaptés pour retrouver des données à partir d'une valeur plutôt que d'une clé.

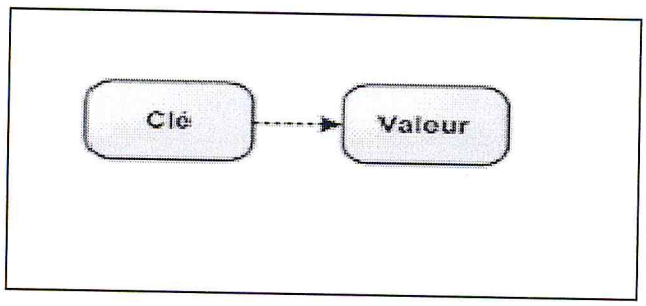


Figure II.5: *Schéma des entrepôts « clé/valeur »*[33] .

b. Les bases orientées documents :

Sur un plan conceptuel, les bases de données orientées document (BDOD) différent peu des ECV. Une BDOD peut schématiquement se décrire comme un ECV dont les valeurs sont des documents semi-structurés généralement écrits avec un format de type XML, JSON ou similaire. Par contraste avec les SGBDR qui exigent que chaque enregistrement d'une table ait les mêmes colonnes, spécifiées par un schéma. Rien n'exige que les documents stockés dans une BDOD aient tous le même format **[5]**.

La valeur non définie d'une colonne d'un enregistrement d'un SGBDR est représentée par la valeur NULL, un document d'une BDOD ne fera tout simplement pas figurer l'attribut en question.

Dans une BDOD chaque document constitue un élément atomique qui délimite une frontière transactionnelle naturelle.

BDOD	SGBDR
Base de données	Schéma
Collection (de documents)	Table
Document	Enregistrement
Id de document	Id d'enregistrement
DBRef (référence entre documents)	Jointure (entre tables)

Table II.1 :*Equivalences entre SGBD et BDOD*[33] .

- **Usage :**

Les applications d'e-commerce dont les produits varient pour être décrits au moyen d'un schéma stable, bénéficieront de l'usage d'une BDOD.

Les applications qui manipulent naturellement des documents comme les systèmes de gestion de contenu ou les plateformes de blogs pourront elles aussi utiliser avec profit une

BDOD.

Mentionnons encore les systèmes de logs qui, par définition, ont à stocker des associations entre des événements et des contenus sans structure prédéfinie.

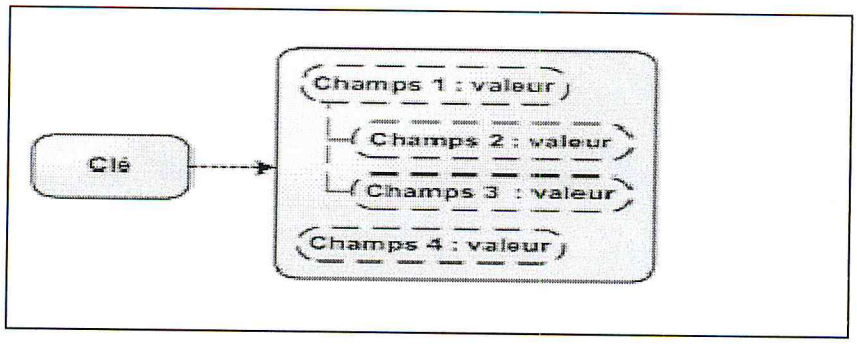


Figure II.6: *Schéma des entrepôts orienté document*[33] .

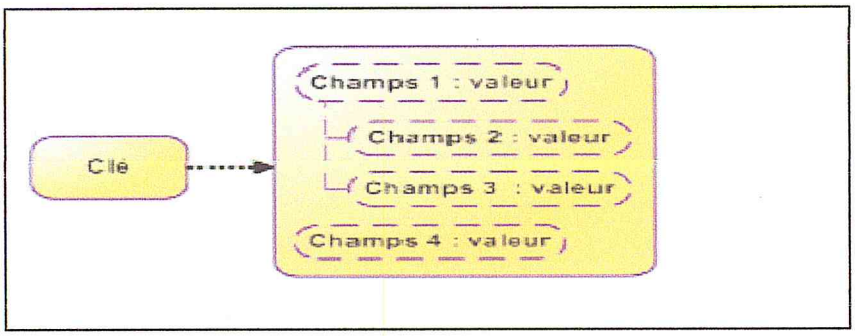


Figure II.6: *Schéma des entrepôts orienté document*[33] .

Les bases fonctionnant avecle principe dit des documents sont un peu plus élaborées que les "Clé/Valeur", bien qu'elles stockent des valeurs toujours liées à une clé d'accès. Mais contrairement à ces précédentes, une clé nous permet d'y entreposer des données complexes, qui peuvent très bien elles-mêmes contenir d'autres documents et ainsi de suite, ce qui permet de structurer ces données. Par exemple, on peut très bien penser stocker un article, qui contiendrait le nom de l'auteur, la date ainsi que le corps de l'article, mais également les commentaires liés à cet article, qui seraient eux-mêmes composés du commentaire en lui-même ainsi que du nom de l'auteur. Bien que l'on puisse structurer les données stockées, elles n'ont pas besoin de suivre un modèle de données, les documents stockés pouvant avoir des types très hétérogènes entre eux. La plupart des bases de données documentaires enregistrent les données sous format JSON.

c. Les bases orientées colonnes :

Un SGBD orienté colonnes (également appelé stockage en colonnes) réoriente le focus des données de la ligne sur la colonne, en stockant les données en colonnes et non en lignes.

Dans les bases de données relationnelles traditionnelles, les données sont modélisées sous forme de lignes de colonnes, l'accès se faisant toujours par les lignes. Le stockage NoSQL en colonnes gère les enregistrements dans des familles de colonnes qui peuvent contenir un grand nombre de colonnes dynamiques. Il n'existe pas de schéma fixe ; autrement dit les noms et les clés des colonnes peuvent varier. Une base de données orientée colonnes convient aux données faisant l'objet de peu d'écritures pour lesquelles les attributs ACID (atomicité, cohérence, isolation, durabilité) ne sont pas impératifs et dont le schéma est variable **[5].**

Le concept de colonnes est le plus simple à saisir, car l'analogie avec les bases relationnelles est proche. Dans les concepts à appréhender il existe des tables. Ce qui permet de bien comprendre comment les données sont organisées. L'analogie est simple, contrairement à une base relationnelle où les enregistrements ont des propriétés fixes. Dans cette dernière le schéma de la table établit le nombre de colonnes que contiendra une table.

- **Usage :**

La flexibilité des BDOC en fait un choix idéal, par exemple, pour stocker des rapports d'erreur issus de plusieurs applications dont chacune pourra utiliser le groupe de colonnes qui lui convient.

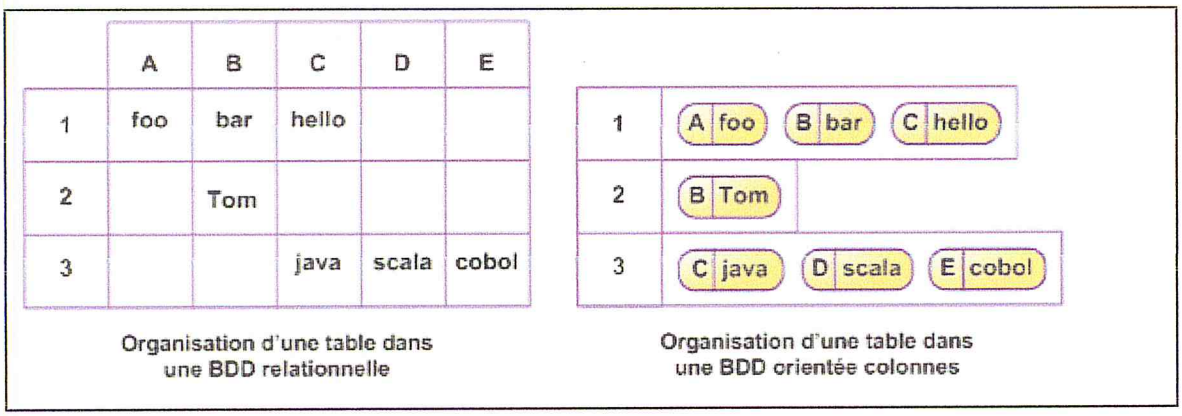


Figure II.7: *La Différence entre l'organisation d'une table dans une BDD relationnelle et l'organisation d'une table dans une BDD orientée colonnes*[33] .

d. Les bases de données orientée graphes :

Une base de données orientée graphes privilégie les relations entre les valeurs et stocke les données en utilisant la théorie mathématique des graphes. Pour représenter et stocker les données, ce type de base de données utilise la structure des graphes en noeuds, en relations et en propriétés. Dans ces bases de données, chaque élément contient un pointeur direct vers l'élément adjacent et aucun recours à un index n'est nécessaire.

Les bases de données en graphe sont peut-être les moins connues parmi les quatre concepts abordés. Elles s'appuient sur deux concepts : les entités et les relations entre elles. Dans le premier, les entités sont représentées comme les bases de données documentaires.

Dans le second, les relations entre deux entités peuvent s'apparenter à une table d'association dans les bases de données relationnelles, mais ces relations contiennent aussi leurs propres propriétés qui leur servent à se personnaliser. Le meilleur exemple que l'on pourrait donner pour l'illustrer serait un réseau social. Chaque entité représente une personne et l'on peut concevoir les relations qui s'opèrent entre elles.

En outre, les bases de données en graphes ont un fort lien de parenté avec les bases de données en réseau. Mais ces dernières n'ont de ressemblance que la structure des données, bien qu'elles-mêmes se différencient fortement par rapport aux différents types de relations pouvant exister. Elles possèdent différents algorithmes, basés sur la théorie des graphes, pour parcourir les données.

- **Usage :**

A priori, tous les domaines métiers qui s'expriment naturellement à l'aide de graphes sont éligibles à l'utilisation d'une BDOG. Ils sont particulièrement utiles dans les situations ou différentes catégories de liens expriment des appartenances à certains groupes sociaux ou géographiques.

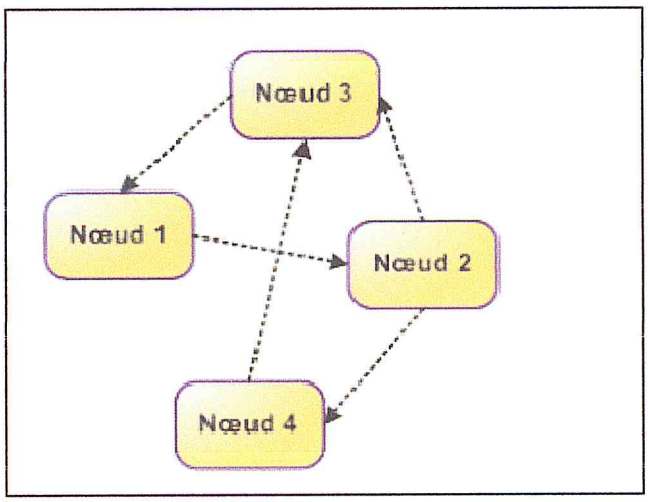


Figure II.8: *Schéma des entrepôts orienté graphe*[33] .

II.6. Conclusion :

Chapitre III

Le concept des entrepôts de données

III.1. Introduction :

Les entrepôts de données « Data Warehouse » sont la base des solutions analytiques dans les entreprises. En effet, avec un entrepôt de données alimenté d'une ou de plusieurs sources de données, une multitude de services est offerte aux analystes pour les aider à prendre des décisions. Parmi ces outils, on trouve les rapports, les tableaux de bord, les outils basés sur « OLAP » et l'historisation des données.

Utiliser un entrepôt de données, pour des raisons analytiques, vient souvent du problème que les systèmes opérationnels n'offrent pas la flexibilité d'analyser facilement les données à cause de leur structure qui est plus optimisée pour un bon fonctionnement des opérations et non pour faire des analyses.

Depuis plusieurs années, le passage aux entrepôts de données était inévitable **[17]**. Les outils basés sur des schémas orientés informationnels sont disponibles sur le marché et la création d'un ETC (programme qui automatise l'extraction des données d'une ou de plusieurs sources, la transformation et le chargement vers l'entrepôt de données) est devenue un élément populaire dans le domaine d'intelligence d'affaires.

L'analyse des données ainsi que les forages se font sur la base des données obtenues du système opérationnel. Avec l'arrivée des concepts CRM(Customer Relationship Management ; l'ensemble d'outils de gestion de la relation client), les entreprises commencent à essayer de comprendre les comportements et besoins de leurs clients, car celles-ci doivent bien servir le client et toute information sur ce dernier doit être considérée.

Une nouvelle dimension en termes de données s'est installée, surtout avec l'apparition des réseaux sociaux tels que Facebook, Twitter et Google+. L'enregistrement d'une transaction d'achat ne se limite plus à l'information des produits achetés et du montant payé. D'autres informations sont maintenant récoltées, même si elles ne sont pas utilisées dans l'immédiat, telles que l'adresse IP, les informations sur le navigateur et le système de l'utilisateur, le temps passé sur une page web et les intérêts de l'utilisateur.

Des mines de données sont enregistrées chaque jour, et les entrepôts de données deviennent de plus en plus volumineux. Cette situation commence à générer quelques inquiétudes en termes de performance d'exécution des ETC, d'accès aux données des entrepôts, et aux coûts générés par les licences des applications utilisées ainsi qu'au matériel en relation. En

parallèle, les demandes en termes de données ne cessent d'augmenter et les analystes sont de plus en plus gourmands pour tout ce qui est « Data » **[18] [19] [20].**

III.2. Les entrepôts de données :

III.2.1. Définition :

Un entrepôt de données, ou « *Data Warehouse* » est une base de données qui aide à la décision et qui contient des informations provenant d'une ou de plusieurs sources de données opérationnelles. Des rapports et des statistiques peuvent être générés sur la base d'un entrepôt de données. Un entrepôt de données peut être constitué de plusieurs sous entrepôts qu'on appelle « *Data Mart* », ce dernier est généralement destiné à un département spécifique.

III.2.2. L'entreposage de données :

L'entreposage de données est un processus qui est défini généralement selon trois niveaux : (a) acquisition des données, (b) stockage des données et (c) exploitation des données, tels que représentés dans la figure III.1.

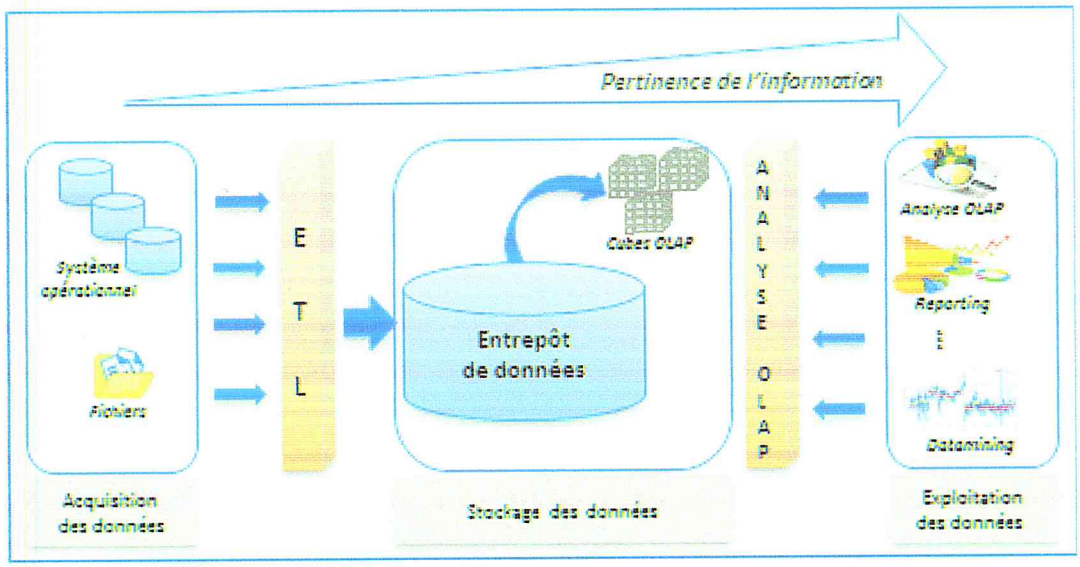


FIGURE III.1 :*Processus d'entreposage de données.*

a) Acquisition des données :

Elle consiste à ramener, à partir des systèmes d'informationopérationnels ou autres sources de données, les données jugées utiles pour l'aide à la décision. Celles-ci sont destinées à alimenter l'entrepôt de données. Cette opération est assurée au moyen d'un processus dit d'extraction, transformation etde chargement (ETL : Extract-Transform-Load).

b) Stockage des données :

Dans cette étape, les données sont structurées en contexte d'analyse décisionnelle et orientées vers l'utilisateur [Muckenhirn 2003]. Cependant, pour que l'utilisateur ne se noie pas dans le volume de données important caractérisant l'entrepôt, les données correspondant au contexte d'analyse de l'utilisateur sont générées dans une structure appelée magasin de données. Un magasin de données est un sous ensemble des données de l'entrepôt relatif à un domaine fonctionnel particulier (métier). C'est à partir del'entrepôt de données ou d'un magasin de données (*datamart*) que l'utilisateur peut créer des contextes d'analyse, appelés cubes de données, hypercubes et souvent cubes OLAP. Celles-ci sont des structures multidimensionnelles répondant aux besoins spécifiques d'un ou plusieurs utilisateurs.

c) Exploitation des données :

C'est la restitution des données décisionnelles pardes outils d'analyses appliqués sur des cubes OLAP. Grâce aux opérateurs OLAP, l'utilisateur peut explorer et naviguer à l'intérieur d'un cube OLAP afin d'aller vers les informations considérées les plus pertinentes pour la prise de décision.

III.3. Outil d'extraction, transformation et chargement ETL :

Aussi connus sous le terme ExtractTransformLoad (ETL), ces outils sont conçus afinde faciliter l'intégration de données hétérogènes, leur normalisation puis ils les rendent cohérentes entre elles **[21]**, pour qu'elles puissent être utilisées conjointement. Les données sont présentées dans un format permettant une exploitation immédiate sans recalcules par les décideurs et les analystes.

Un outil ETL est beaucoup plus que la plomberie pour obtenir des données depuis les sources vers l'entrepôt, mais il ajoute une valeur significative aux données.

Plus précisément, un outil ETL **[22]** :

- Supprime les erreurs et corrige les données manquantes.

- Fournit des mesures documentées de confiance dans les données.

- Capture le flux de données transactionnelles pour la garde.

- Ajuste les données provenant de sources multiples pour être utilisées ensemble.

- Structure les données pour être utilisables par les outils de l'utilisateur final.

Les outils ETL passent par trois grandes étapes :

1. **Extraction :**

Lors de l'extraction, les données sont identifiées et extraites de différentes sources, y compris les systèmes et les applications des bases de données. Très souvent, il est impossible d'identifier le sous-ensemble spécifique d'intérêt, donc plus de données que nécessaire doivent êtres extraites, de sorte que l'identification des données pertinentes se fera à un moment ultérieur. En fonction des capacités du système de la source, certaines transformations peuvent avoir lieu au cours de ce processus d'extraction. La taille des données extraites varie de quelques centaines de kilo-octets jusqu'à giga-octets, en fonction du système source et la situation de l'entreprise. La même chose est vraie pour le delta de temps entre deux extractions (logiquement) identiques : l'intervalle de temps peut varier entre jours / heures et minutes jusqu'à temps quasi réel.

Le processus d'extraction est constitué de deux phases, l'extraction initiale etl'extraction des données modifiées **[23]**. Dans l'extraction initiale les données sont obtenues pourla première fois depuis les différentes sources opérationnelles. Ce processus se fait qu'une seule fois après la construction de l'entrepôt de données. Par contre, dans l'extraction des données modifiées (en anglais, CDC pour Changed Data Capture), les outils ETL rafraîchissent l'Entrepôt avec les données modifiées et ajoutées dans les systèmes sources depuis la dernière extraction. Ce processus est périodique selon le cycle de rafraîchissement et les besoins des entreprises.

Après que les données sont extraites, elles doivent être physiquement transportées vers le système cible ou à un système intermédiaire pour un traitement ultérieur. Comme déjà dit, certaines transformations peuvent être effectuées au cours de ce processus de transportation.

2. **Transformation :**

L'étape de transformation implique l'application d'une série de règles et de fonctions sur les données extraites. Elle comprend la validation des enregistrements et leur rejet si elles ne sont pas acceptables. Cette seconde phase du processus assure la fiabilité des données et leur qualité en exécutant certaine tâche, car les données extraites de la source ne sont pas forcément dans l'état dans lequel elles seront stockées. Elles doivent être vérifiées, reformatées et nettoyées afin d'éliminer les valeurs incohérentes ainsi que les doublons. Il s'agit donc d'une suite d'opérations

s'agit donc d'une suite d'opérations qui a pour but de rendre les données cibles homogènes et puissent être traitées de façon cohérente.

2.1. Les taches de transformation de données :

L'ensemble de manipulations nécessaire au processus de transformation dépend des données sources. Certaines d'eux exigentpeu de transformations, tandis que d'autres peuvent nécessiter une ou plusieurs techniques de transformation pour répondre aux exigences d'entreprise. Les procédés les plus couramment utilisés pour la transformation sont :

- **Nettoyage de données :**

Le nettoyage des données s'occupe de détecter et de corriger ou de supprimer les erreurs et les incohérences présentes sur les données afin d'améliorer leur qualité.

- **Jointure** : faire joindre les données provenant de sources multiples.
- **Filtrage** : sélectionner seulement les données à charger (par exemple, sélectionner que certaines colonnes).
- **Eclatement** : Fractionner une colonne en plusieurs colonnes (ex. éclater la colonne qui contient une adresse composée de plusieurs champs en différentes colonnes).
- **Validation** : appliquer des règles afin de chercher les données pertinentes à partir des tables ou des fichiers référentiels pour les dimensions à variation lente (par exemple, rejeter une ligne si elle contient 3 colonnes vides).
- **Agrégation** : c'est un cumul résumant plusieurs lignes de données (par exemple, effectif des employés par âge, type contrat), ce qui permet d'avoir des temps de réponse très courts. Le niveau d'agrégation est choisi au moment de la construction de l'ED.
- **Tri et arrangement.**
- **Eliminer les doubles.**
- **La conversion :** Elle sert à transformer les données provenant des différentes sources dans un format cible.
- **La normalisation :** Cette tâche consiste à normaliser les données provenant des sources hétérogènes pour les rendre homogènes. Elle nécessite des règles précises servant de référentiel qui sont mémorisées sous forme de métadonnées.

3. Chargement des données :

C'est l'opération qui consiste à charger les données nettoyées et préparées dans l'entrepôt et à gérer les changements aux données existantes en définissant des stratégies d'historisation et de rafraîchissement.

Elle est une phase plutôt mécanique et la moins complexe mais elle risque d'être assez longue lorsque les données sont volumineuses, donc il est nécessaire de veiller à ce que la charge est exécutée correctement et avec le moins de ressources que possible. Pour cela et afin d'améliorer les performances de chargement dans une base de données –l'ED est souvent une BD-, il est utile de désactiver les indexes de la base de données et les contraintes préalablement au chargement et leur permettre de retour postérieurement.

Étant donné que toutes les tables de faits doivent préserver l'intégrité référentielle, la clé de dimension primaire est reliée à une clé étrangère correspondante dans la table de faits. C'est pour cette raison qu'il est important d'alimenter les tables de dimensions, puis les tables de faits.

Les tables de dimensions sont généralement alimentées à partir d'une ou plusieurs sources, elles contiennent autant que possible des attributs permettant de qualifier ou d'expliquer l'activité. Donc elles représentent des entités complémentaires à la conception de la table de faits qui contiennent les mesures d'activité de l'entreprise.

III.4. Traitement analytique en ligne OLAP :

En 1993, Edgar Frank Codd a introduit le terme On-Line AnalyticalProcessing (OLAP) qui à la différence de l'OLTP, introduit une technologie optimisée pour les requêtes humaines plutôt que pour les transactions **[24]**.

Il a aussi introduit 12 « règles de base » permettant de qualifier l'OLAP :

1.Vue conceptuelle multidimensionnelle : permet d'avoir une vision multidimensionnelle des données.

2.Transparence : l'utilisateur doit pouvoir accéder aux données, sans se préoccuper des sources de données.

3.Accessibilité : les données doivent toutes être accessibles, sans ambiguïté.

4.Performance consistante des rapports : les performances ne doivent pas être diminuées lors de l'augmentation du nombre de dimension ou de la taille de la base de données.

5.Architecture Client/Serveur : il est essentiel que le produit soit Client-Serveur mais aussi que les composants serveurs d'un produit OLAP intègrent facilement ses différents clients.

6.Dimensionnalité générique : toutes les dimensions doivent être équivalentes par rapport à leur structure et leurs capacités opérationnelles.

7.Ajustement automatique du niveau physique : le système OLAP ajuste automatiquement son schéma physique pour s'adapter au type du modèle et au volume des données.

8.Support multiutilisateur : l'outil doit fournir des accès concurrents, l'intégrité et la sécurité.

9.Opérations cross-dimensionnel libre : les calculs doivent être possibles à travers toutes les dimensions.

10.Manipulation intuitive des données : la manipulation des données se fait directement à travers les cellules dans le model analytique, sans recourir aux menus ou aux actions multiples.

11.Rapports flexibles : lors de la création de rapports, les dimensions peuvent être présentées de n'importe quelle manière.

12.Dimensions et niveaux d'agrégation illimités.

Entre entrepôt et OLAP, il n'y a qu'un pas. En effet, l'entrepôt est le lieu de stockage physique des données, tandis que l'OLAP est l'outil permettant leur analysemultidimensionnelle. Afin de rendre l'analyse la moins contraignante et la plus souple possible, l'OLAP propose des opérateurs. Il s'agit de mécanismes servant à naviguer dans les hiérarchies et les dimensions.

III.4.1. Environnement OLAP :

L'environnement OLAP est un ensemble d'outils permettant d'accéder aux entrepôtsde données pour permettre la création de contextes d'analyse appelés « cubes », et de permettre à plusieurs applications d'accéder à ces cubes de données à des fins d'analyse. Dans ce cas, une requête OLAP est une requête multidimensionnelle permettant d'agréger les données d'une ou de plusieurs mesures suivant les attributs d'une ou plusieurs dimensions.

III.4.2. Cube OLAP :

Un cube OLAP est une collection de données agrégées et consolidées pourrésumer l'information et expliquer la pertinence d'une observation. Il permet de représenter le fait à observer selon plusieurs axes d'observation (dimensions) [Gray 1997]. Les données du cube sont dites multidimensionnelles, car l'utilisateur les manipule suivant différents critères, dits attributs de dimensions. Durant le processus d'analyse, le décideur a besoin d'appréhender ses données suivant plusieurs combinaisons de dimensions. Pour ce faire, il va devoir effectuer des requêtes agrégeant les données par rapport à l'ensemble des critères choisis. Typiquement, ces requêtes mettent aussi en jeu des fonctions d'agrégation appliquées sur des attributs mesures selon les diverses dimensions retenues.

Exemple : la figure **III-1** représente un exemple de cube de données qui permet l'analyse des publications scientifiques.

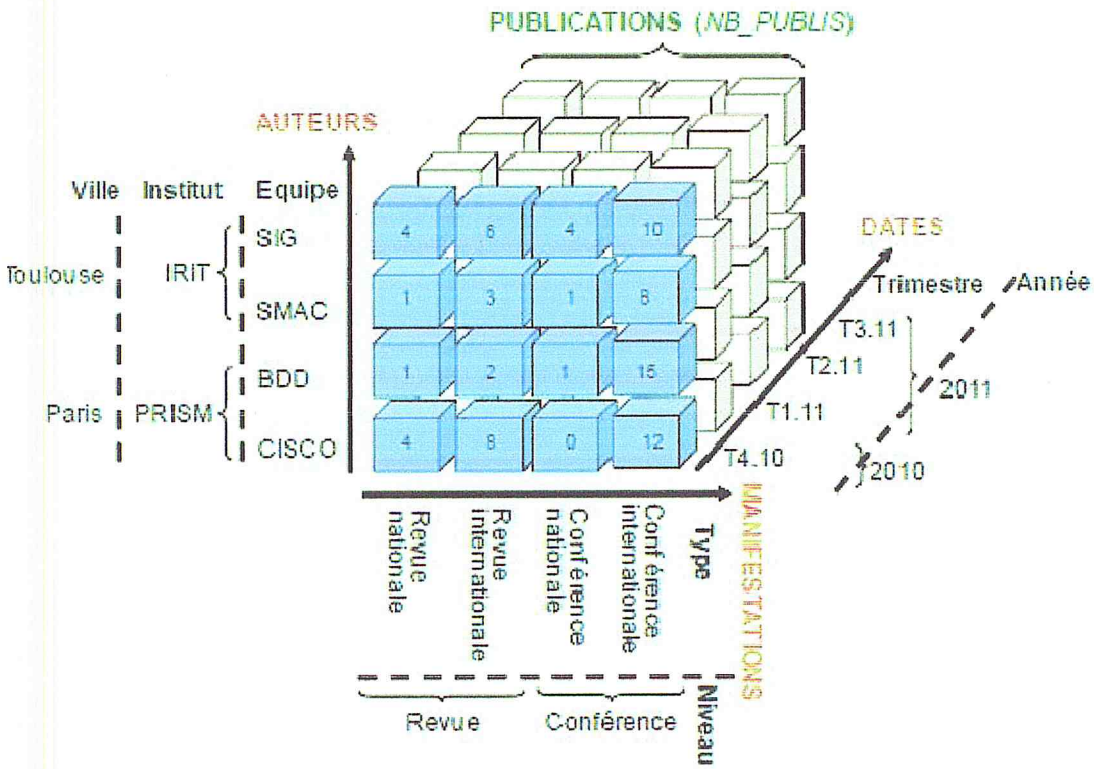


Figure III-1 : Exemple d'un cube de données **[25].**

III.4.3. Opérations OLAP :

C'est un ensemble d'opérateurs permettant d'explorer et de naviguer dans un cube de données. Ces opérateurs permettent de mettre en évidence une analyse particulière des données. Les plus emblématiques sont :

– Forage vers le bas (*drill-down*), qui consiste à descendre dans une hiérarchie de dimension vers un niveau plus détaillé

– Forage vers le haut (*roll-up*), qui consiste à remonter dans une hiérarchie de dimension vers un niveau plus agrégé

– Rotation (*rotate*), qui réoriente une analyse en changeant l'axe d'analyse en cours (rotation de dimension)

– Sélection de tranches : (*slice*), qui sélectionne un sous-ensemble réduit de membres sur une ou plusieurs dimensions, et (*dice*) qui réduit un cube d'une ou plusieurs dimensions.

III.5. Conclusion :

L'intelligence d'affaires ou le « *Business Intelligence* », est un domaine qui existe depuis plusieurs années et qui est en train de vivre des défis plus intéressants par rapport au début de son apparition dans les années 1990. Un de ces défis est la performance des entrepôts de données dans un contexte de mégadonnées.

La quantité de données qu'on traite aujourd'hui dans les modèles *ROLAP* est beaucoup plus importante que celle de la précédente décennie **[26]**. Les problèmes liés à la performance, aux coûts, à la lenteur de traitement, à la complexité d'échelonnage, et à la haute disponibilité ont poussé sans arrêt à chercher des solutions qui répondent aux besoins analytiques de nos jours avec des coûts acceptables. Tels que la migration des données vers un environnement distribué qui sera abordés dans le prochain chapitre.

Chapitre IV

Migration des données vers un environnement distribué

IV.1. Introduction :

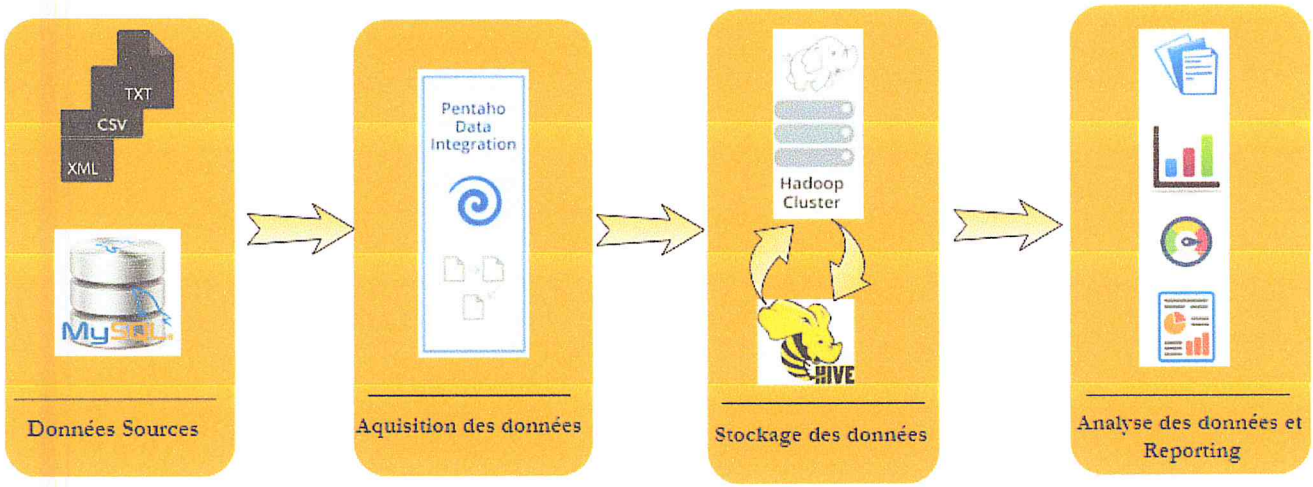
La plupart des plateformes Big Data sont basées sur le stockage distribué de données. Cette solution repose sur une architecture de stockage répartie sur plusieurs machines.

Contrairement aux systèmes traditionnels, ils stockent des blocks de fichiers très volumineux sur plusieurs nœuds. Ils sont conçus pour fonctionner sur du matériel peu coûteux et offrent un accès rapide aux ensembles de données massives.

Dans ce chapitre, nous allons présenter l'architecture du cluster de notre système ainsi que le schémade données capturées lors de l'étudedes besoins et le schéma des données cible.

IV.2. Architecture mise en place :

Pour aboutir ànotre objectif nous avons mis en place une architecture big data pour la migration de nos données semi-structurées et structuré vers un environnement complétement distribué(cluster Hadoop) pour des fins d'analyse et de visualisationLa figure IV.1représente l'architecture mise en place.



FigureIV.1 : *Architecture de migration des données vers un environnement distribué.*

IV.3. Capture des besoins :

Nous avons effectué notre stage au sein du département Gestion de performance réseaude l'opérateur téléphonique Ooredoo ; ce département est très important car il est responsable de développer des solutions qui facilitent la tâche de l'analyse des performances, ainsi queles

algorithmes de collecte des formules KPI (Key Performance Indicator) et cela pour chaque technologie 2G, 3G et 4G.

Ces solutions vont assurer la corrélation entre les phénomènes du réseau ainsi que la surveillance et l'analyse des performances et des caractéristiques de qualité dans le but de fournir des paramètres de configuration de réseau efficaces afin d'améliorer continuellement les performances du réseau Ooredoo.

IV.4. Données sources :

Les indicateurs de performance (KPI) Se représentent sous forme de trois types :

- **Données structurées**

Ce sont les données associées à des bases de données classiques, tels que les transactions relationnelles où l'information est organisée en lignes et en colonnes dans les tables. Presque tous les systèmes de gestion de base de données compris (SGBD) sont conçus pour des données structurelles **[27]**.

- **Données semi-structurées**

Les données semi-structurées sont organisées en entités sémantiques, ce qui les caractérise c'est que les entités similaires sont regroupées, les entités du même groupe n'ont pas forcément les même attributs, l'ordre des attributs n'est pas nécessairement important et les attributs ne sont pas forcément tous nécessaires, de plus la taille ou le type du même attribut du groupe peuvent différer.

Pour être organisé et recherché, les données semi-structurées doivent être fournies par voie électronique à partir des systèmes de bases de données, systèmes de fichiers (par exemple, les données bibliographiques, des données sur le Web) ou via des formats d'échange de données (par exemple, l'EDI, les données scientifiques, XML)**[27]**.Dans le cas de notre système les données semi structuré sont en forme de fichiers plats XML et CSV.

- **Fichiers CSV :**

Un fichier CSV (Comma Separated Values) est un simple fichier (texte) dans lequel les valeurs sont séparées généralement par une virgule, ce qui permet de sauvegarder les données dans unformat de tableau dont chaque ligne comporte le même nombre des champs. Ce type

de fichier esttrès facile à manipuler (par exemple, il est utilisé pour envoyer des données (par FTP, par email,etc.).

```
Huawei_Sharing_Template_Cell_Hourly_Num;;;
Save Time :12/02/2017 06:31:39;;;
User Name :quality;;;
;;;
Time;RNC;Integrity;Cell Unavaialble Time Ratio(%)
02/02/2017 00:00;Bejaia RNC;100%;2.2852
02/02/2017 00:00;CNE RNC02;100%;0.2804
02/02/2017 00:00;ORAN RNC 2;100%;0.652
02/02/2017 00:00;Rouiba RNC;100%;2.8587
02/02/2017 00:00;Blida RNC;100%;1.2938
02/02/2017 00:00;Tiziouzou RNC;100%;2.0942
02/02/2017 00:00;Skikda RNC;100%;4.7248
02/02/2017 00:00;Delly Brahim;100%;3.0885
02/02/2017 00:00;CNE RNC;100%;0.8296
02/02/2017 00:00;Setif RNC;100%;3.3035
02/02/2017 00:00;Mostaganem RNC;100%;0.7547
02/02/2017 00:00;Oran RNC;100%;1.0323
02/02/2017 00:00;Chlef RNC;100%;1.9231
02/02/2017 00:00;Rouiba RNC 2;100%;0
02/02/2017 00:00;Annaba RNC;100%;0
02/02/2017 01:00;ORAN RNC 2;100%;0.652
```

Figure IV.2 :*Données en format CSV.*

- **Fichiers XML :**

Un fichier XML (eXtensibleMarkupLanguage) décrit tout simplement un fichier texte contenant àla fois des données et des métadonnées, permettant de stocker des informations enrespectant unestructure donnée en utilisant des balises qui sont délimitées par les caractères « inférieur » et « supérieur ». Un fichier XML est un moyen extrêmement efficace pour établir la communication entre les systèmes, puisque XML (et les schémas XML qui suivent) fournit suffisamment d'informations pour échanger les données entre eux, si par exemple l'entrepôt de données comprend des données qui viennent de sources externes de l'entreprise, ces sources seront fournies en XML.

```
<?xml version="1.0"?>
<OMeS>
  <PMSetup  startTime="2017-08-26T15:45:00.000+01:00:00"
interval="15">
    <PMMOResult>
      <MO>
        <DN><![CDATA[PLMN-PLMN/MSC-224804]]></DN>
      </MO>
      <MO>
        <DN><![CDATA[PLMN-PLMN/ANN-300]]></DN>
      </MO>
      <PMTarget  measurementType="ANN">
        <M26B2C2>3</M26B2C2>
        <M26B2C3>0</M26B2C3>
        <M26B2C4>12</M26B2C4>
      </PMTarget>
    </PMMOResult>
    <PMMOResult>
      <MO>
        <DN><![CDATA[PLMN-PLMN/MSC-224804]]></DN>
      </MO>
      <MO>
        <DN><![CDATA[PLMN-PLMN/ANN-301]]></DN>
      </MO>
      <PMTarget  measurementType="ANN">
        <M26B2C2>20</M26B2C2>
        <M26B2C3>0</M26B2C3>
        <M26B2C4>74</M26B2C4>
      </PMTarget>
    </PMMOResult>
```

Figure IV.3 :*Données en format XML.*

- **Données non-structurées**

Ce sont les données qui ne suivent aucun format défini, elles peuvent être de tout type, elles ne suivent pas les règles, ne sont pas prévisibles, et peuvent généralement être décrites comme « Forme libre ». Les données non-structurées sont comme par exemple : (texte, images, vidéo ou son).Généralement les moteurs de recherches qui récupèrent les données non-structurées **[27]**.

Les organisations nécessitent d'intégrer et d'analyser des données à partir d'un ensemble complexe de deux sources d'information traditionnelles et non traditionnelles, de l'intérieur et de l'extérieur de l'entreprise. Avec l'explosion des capteurs, appareils intelligents et les technologies de collaboration sociale, les données sont générées d'innombrables formes, y compris : texte, données web, tweets, les données des capteurs, audio, vidéo, les fichiers journaux et d'autres. Dans le cas des indicateurs de performance ils sont de type texte.

IV.5. Schémadesdonnéescible :

Pour notre système nous avons proposé un schéma conceptuel pour le modèle de données cibles qui répond le plus aux besoins de l'entreprise conçu pour des fins d'analyse. Il est répresenté dans la Figure IV.4

MSS					DateTime		measurementType						
Topologie							Compteurs						
PLMN	MSC	CGR	CGRNAME	CGRDIR	startTime	interval	M16B2C10	M16B2C11	M16B2C16	M16B2C17	M16B2C24	M16B2C25	M16B2C3

Figure IV.4 : Schéma conceptuelle des données cibles

Pour concevoir ce modèle de données nous utilisons le schéma en étoile qu'on va présenter dans le chapitre V la figure V.2. Et il s'agit d'un schéma représentant l'analyse du trafficdes appelsentrant et sortant de l'opérateur ooredoo par rapport à une topologie, une date, et un type de compteur qu'on va détaillé dans les prochains chapitres.

IV.6. Partitionnementdesdonnées :

Dans une base ou un entrepôt de données, une partition est une division logique d'une table stockée en plusieurs parties indépendantes. Le partitionnement de tables est généralement effectué pour :

- Améliorer la gestion.
- Améliorer la performance.
- Améliorer la disponibilité.

Chaque partition se retrouve sur des répertoires, des serveurs ou des disques différents. Cela permet d'effectuer des requêtes en parallèle sur plusieurs partitions bien organisées, pour des fins d'analyse.

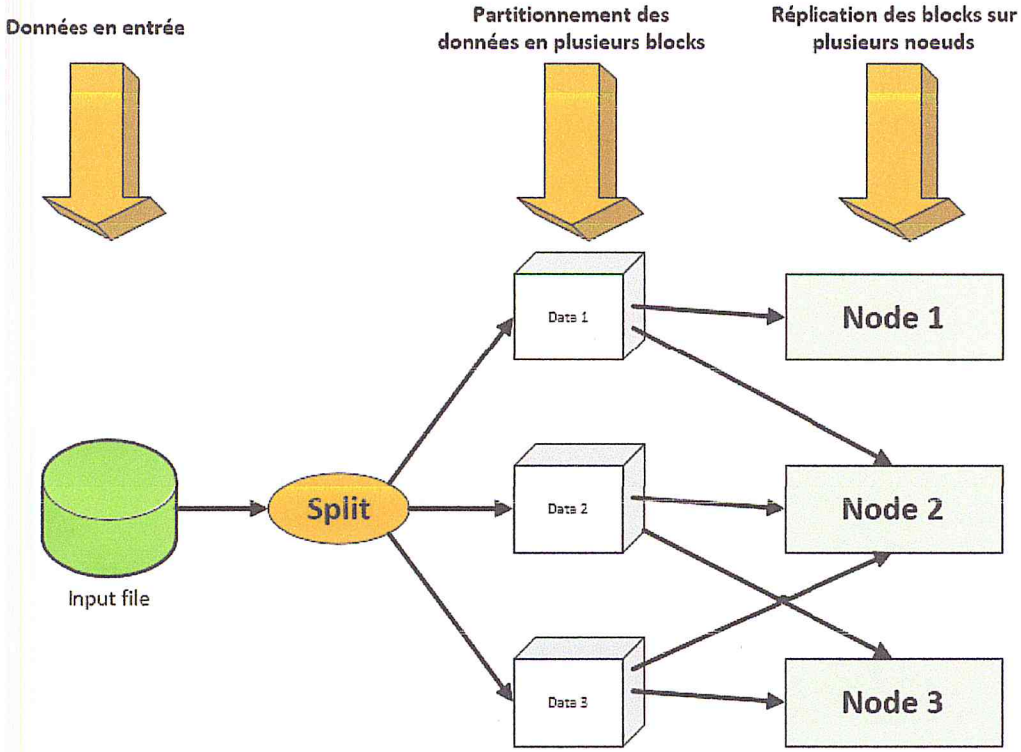


Figure IV.2 :*Schéma de partitionnement et de réplication des données dans HDFS.*

IV.7. Conclusion :

Le principal avantage de l'utilisation du cluster est qu'il est idéal pour analyser les grands volumes de données. Les big data tendent àêtre largement distribués et non structurés. La raison pour laquelle l'environnement distribué est bien adapté à ce type de données est parce qu'il fonctionne en divisant les données en pièces et en assignant chaque « pièce » à un nœud de cluster spécifique pour l'analyse. Les données ne nécessitent pas d'être uniformes car chaque donnée est gérée par un processus distinct sur un nœud de cluster distinct. Le chapitre suivant est consacré à la description détaillée de l'analyse des données.

Chapitre V

Analyse Des Données

V.1. Introduction :

Contrairement aux bases de données relationnelles, l'intérêt des bases de données décisionnelles ne se situe pas au niveau de l'individu (enregistrement) mais plutôt au niveau de l'identification des tendances dans un ensemble ou un groupe.

Les données à analyser doivent refléter la vision des analystes, c'est-à-dire apparaître sous une forme facilitant les prises de décision. Cette vision correspond à une structuration des données selon plusieurs axes d'analyse représentant des notions diverses telles que le temps, la localisation géographique, etc. On parle d'analyse multidimensionnelle.

V.2. Définition :

La modélisation multidimensionnelle, aussi appelée modélisation OLAP **[24]**, est une technique qui vise à organiser les données de telle sorte que les applications OLAP soient performantes et efficaces **[28]**. Elle permet aux décideurs de faire des analyses sur les mesures commerciales de l'entreprise de différentes façons **[29]**.

V.3. Modélisation multidimensionnelle des données :

V.3.1. Modélisation conceptuelle :

Trois types de modèles pour les entrepôts sont fréquemment utilisés : le modèle en étoile, en flocon de neige et en constellation. En effet, à partir du fait et des dimensions, il est possible d'établir une structure de données simple qui correspond au besoin de la modélisation multidimensionnelle. Cette structure est constituée du fait central et des dimensions.

- **Fait :**

Le fait représente le sujet d'analyse. Il est composé d'un ensemble de mesures qui représentent les différentes valeurs de l'activité analysée.

La table de faits est la clef de voûte du modèle dimensionnel où sont stockés les indicateurs de performances (mesures). Le concepteur s'efforce de considérer comme indicateurs les informations d'un processus d'entreprise dans un système d'information [8].

- **Dimension :**

Une dimension est une table qui représente un axe d'analyse selon lequel on veut étudier des données observables(les faits) qui, soumises à une analyse multidimensionnelle, donnent aux utilisateurs des renseignements nécessaires à la prise de décision [2].

Le modèle en étoile est le modèle le plus utilisé dans les entrepôts de données. Ce modèle est représenté dans la figure V.1.

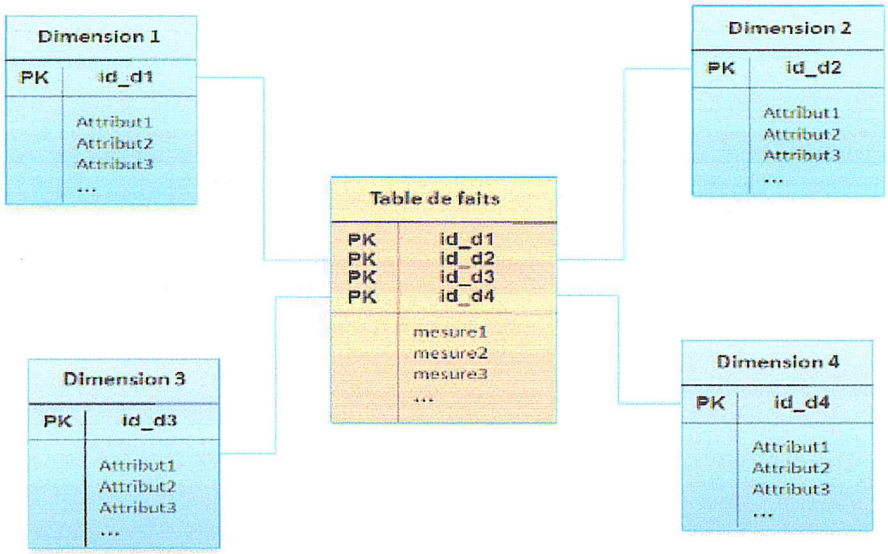


Figure V.1 : *Modèle en étoile.*

- **Avantage :**

– Facilité la navigation.

– Nombre de jointures limité.

- **Inconvénients :**

– Redondance dans les tables de dimensions.

– Toutes les dimensions ne concernent pas les mesures.

Le schéma en étoile de notre étude décrit une série de mesures sur lesquels notre analyse a été faite et qui consiste de calculer le pourcentage du nombre d'appels connectés avec succès (ASR) par rapport au nombre d'appels tentés et cela pour les appels entrant et sortant, puis calculer la somme des ASR total de l'opérateur téléphonique Ooredoo .

Une deuxième analyse a été faite sur le trafic des appels entrant et sortant et qui représente le volume total des appels pour le réseau Ooredoo , par la suite on a calculé le ROU (Rate of Use) qui est le taux d'utilisation (d'occupation) du réseau et toutes ses mesures sont schématisées dans notre table de fait FACT_Traffic et seront analysées par rapport à une date (DIM Temps),une topologie (DIM_Topologie) et un type de compteur bien précis (DIM_Compteurs) comme le montre la figure **V.2.**

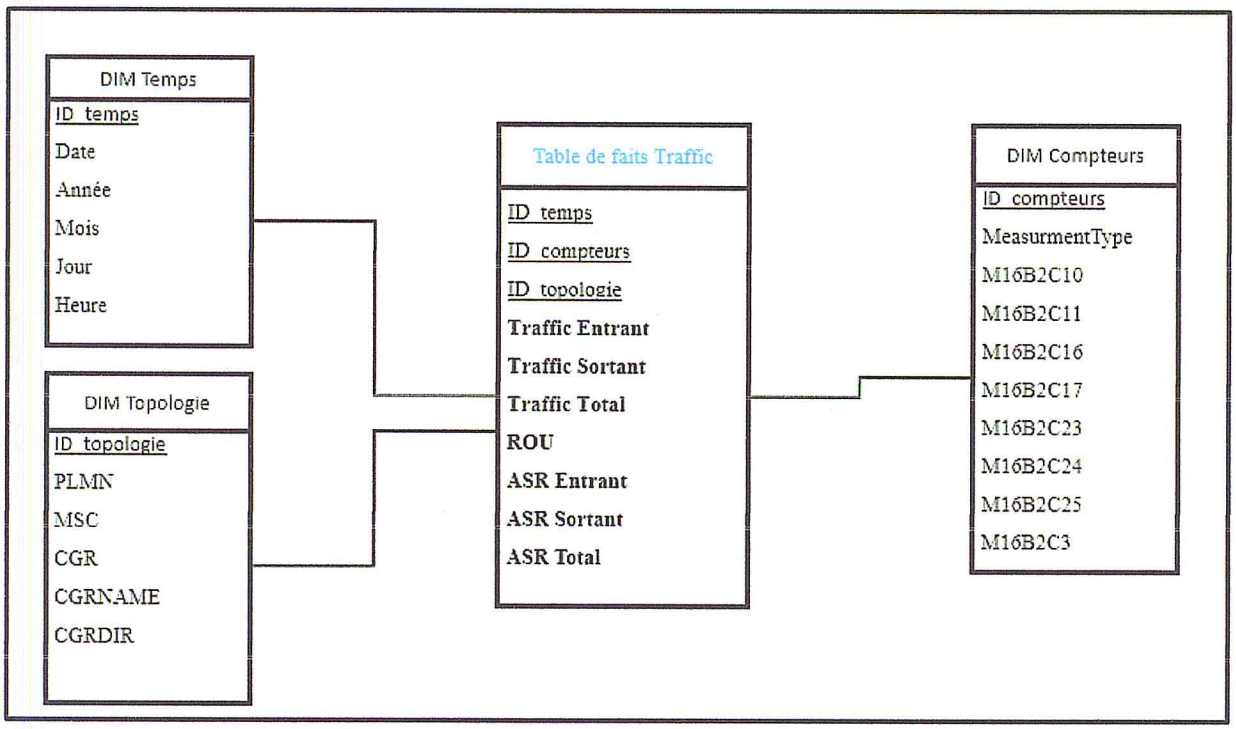


FIGURE V.2 : *Schéma en Etoile de notre étude.*

Il existe d'autres modèles multidimensionnels, notamment celui en flocon de neige (snowflake). Une modélisation en flocon de neige consiste à décomposer, par normalisation, les dimensions en les organisant en hiérarchies. La modélisation en flocon de neige est donc une émanation de celui en étoile ; le fait est conservé et les dimensions sont éclatées conformément à des hiérarchies des paramètres. L'avantage de cette modélisation est de formaliser une hiérarchie au sein d'une dimension. Ainsi, la modélisation en flocon de neige induit une normalisation des dimensions générant une plus grande complexité en termes de lisibilité et de gestion, mais rajoutant de nouveaux paliers d'analyse. Le modèle en flocon est représenté dans la figure V.3.

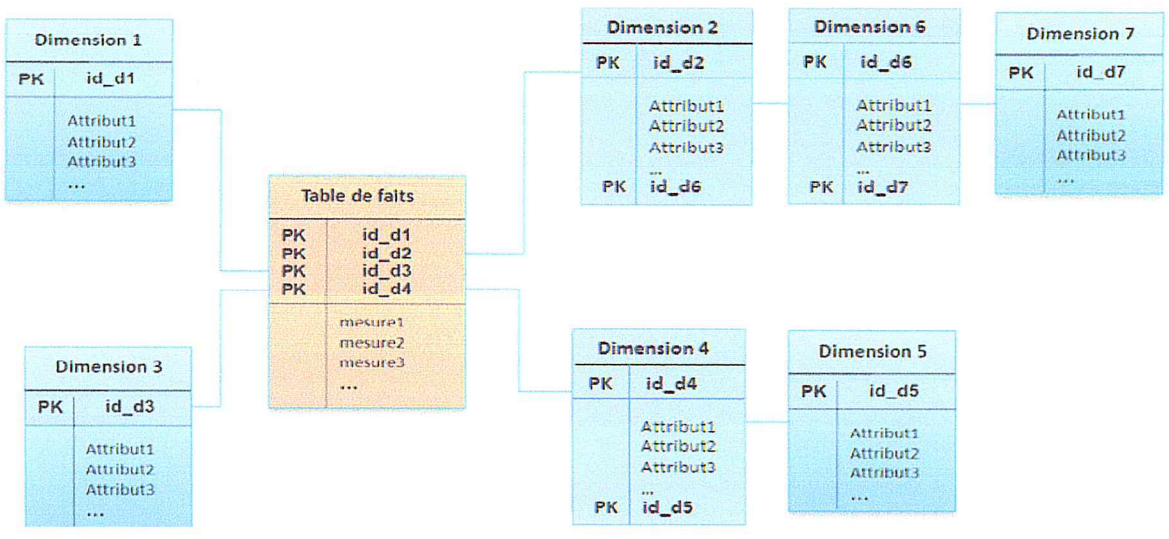


Figure V.3 : *Modèle en flocon de neige.*

- **Avantage :**
  - Normalisation des dimensions ;
  - Economie d'espace disque (réduction du volume) ;
- **Inconvénients :**
  - Modèle plus complexe (nombreuses jointures) ;
  - Requêtes moins performantes ;
  - Navigation difficile.

Le dernier type, issu du modèle en étoile, est celui en constellation. Il s'agit de fusionner plusieurs modèles en étoile sans duplication des dimensions communes.

Un modèle en constellation comprend donc plusieurs faits reliés à des dimensions où chacune des dimensions peut être associée à un ou plusieurs faits [Kimball 2002].Ce modèle est représenté dans la figure V.4.

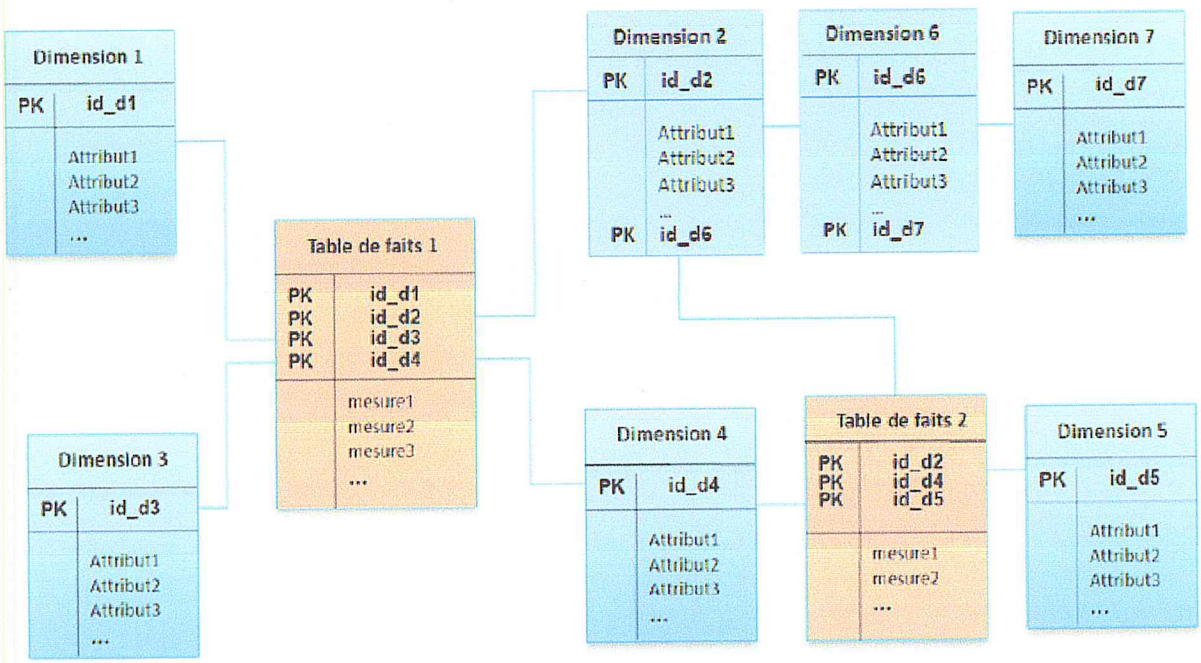


Figure V.4 : *Modèle en constellation.*

V.3.2. Modélisation logique :

Au niveau logique, plusieurs possibilités sont envisageables pour la modélisation multidimensionnelle. En effet, il est possible d'utiliser soit, un système de gestion de bases de données (SGBD) existant tel que les SGBD relationnels. Dans ce cas, l'environnement OLAP utilisé est le R-OLAP (Relational On-Line AnalyticalProcessing), soit un système de gestion de bases de données multidimensionnelles MOLAP (Multidimensional On-Line AnalyticalProcessing), soit un système hybride HOLAP (Hybrid On-Line AnalyticalProcessing) qui est la combinaison des deux systèmes relationnel et multidimensionnel.

V.3.3. Modélisation physique :

La modélisation physique consiste à choisir un mode d'implantation particulier dépendant du logiciel utilisé (notamment le SGBD) et à définir des scripts SQL pour la création et l'alimentation de l'entrepôt **[30]**.

V.4. Conclusion :

Dans ce chapitre nous avons introduit la modélisation multidimensionnelle desdonnées qui repose sur les concepts de fait et de dimension. L'association de fait et dedimension compose un schéma qui peut être soit en étoile, en flocon ou en constellation.

Plusieurs opérations permettent la navigation et la manipulation multidimensionnelle des données stockées.

Cependant, dans un monde constamment connecté, les données sont devenues de plus en plus massives, connues sous l'appellation big data. Ceci implique que les systèmes d'information décisionnels doivent s'adapter et faire face aux nouvelles exigences de stockage et d'analyse des données massives. Dans ce contexte, nous présentons dans le chapitre suivant le déploiement de l'environnement distribué pour le stockage et l'analyse des données massives.

Chapitre VI

Déploiement de l'environnement distribué

VI.1. Introduction :

Dans ce chapitre, nous allons aborder la partie pratique qui consiste à implémenter l'architecture proposée et la mise en œuvre complète de l'environnement distribué Hadoop.

Nous commencerons par présenter l'environnement de travail et des données sur lesquelles les tests se sont faits, par la suite l'installation et la mise en place de notre cluster Hadoop en mode totalement distribué avec ses interfaces de gestion et d'administrations, et au final l'implémentation des deux solutions proposées et les résultats obtenus à l'issu des différents tests d'évaluation.

VI.2. Présentation de l'environnement de travail :

VI.2.1. Systèmes d'exploitation :

Nous utilisons une plateforme Unix de type linux : « **Ubuntu linux 14.04 LTS Trusty** » sur les trois nœuds du cluster.

VI.2.2. Prérequis :

- PC portable, HP Probook 4740s ;
- Processeur : Intel(R) Core i3-2310M CPU@2.10GHz 2.10 GHz
- Type OS : 64 Bits
- RAM : 8 Go
- Disque : 732 Go 48
- VMWARE : Workstation 12 Pro
- Hadoop 2.7.3
- HBASE 0.94.8
- HIVE 2.1.1
- SQOOP 1.4.6

VI.2.3. Miseàjoursystème :

Avant toutes installations de nouveaux paquets il faut mettre à jour le cache des paquets sur votre machine. Les commandessuivantestéléchargeront la nouvelle liste des paquets proposés par le dépôt.

```
sudo add-apt-repository ppa:webupd8team/java
sudo apt-get update
sudo apt-get install oracle-java8-installer
sudo apt-get install oracle-java8-set-default
```

FigureVI.1 :*Installation de java 8.*

VI.2.4. Java :

Le langage Java est un langage de programmation orienté objet, développé par *Sun Microsystems*, Outre son orientation objet le langage Java a l'avantage d'être modulaire le faite qu'on peut écrire des portions de code génériques c.à.d. utilisables par plusieurs applications, rigoureux puisque la plupart des erreurs se produisent à la compilation et non à l'exécution et portable dont un même programme compilé peut s'exécuter sur différents environnements.

Sa particularité principale est que les logiciels écrits avec ce dernier sont très facilement portables sur plusieurs systèmes d'exploitation tels que Unix, Microsoft Windows, Mac OS ou Linux avec peu ou pas de modifications, c'est pour cette raison que nous avons l'utilisé.

Le Java Development Kit (JDK) désigne un ensemble de bibliothèques logicielles de base du langage de programmation Java, ainsi que les outils avec lesquels le code Java peut être compilé. Pour l'exécution de Hadoop l'environnement java et prérequis, nécessitant installation d'une version 5 ou plus, nous avons choisi d'installer la version (Java 8) présenté sur la figureVI.2.

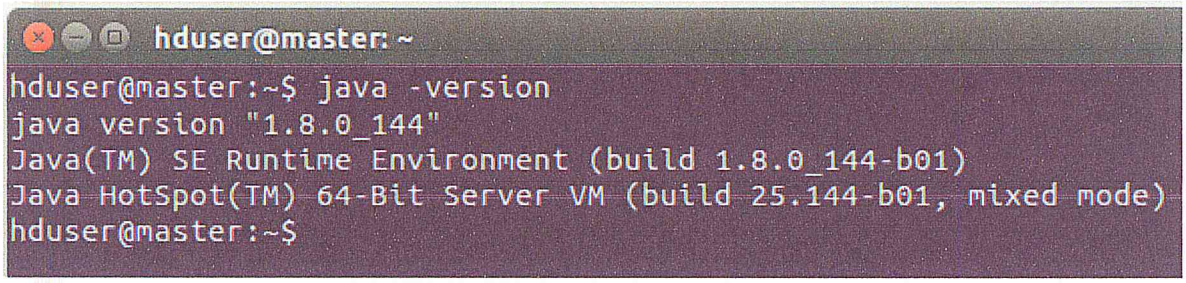


Figure VI.2 : *version de java.*

VI.2.5. Configuration SSH :

Hadoop nécessite un accès SSH pour gérer les différents nœuds. Bien que nous soyons dans une configuration simple nœud, nous avons besoin de configurer l'accès vers localhost la figure ci-dessous montre l'accès à un nœud esclave (slave-1) à partir d'un nœud maitre (Master).

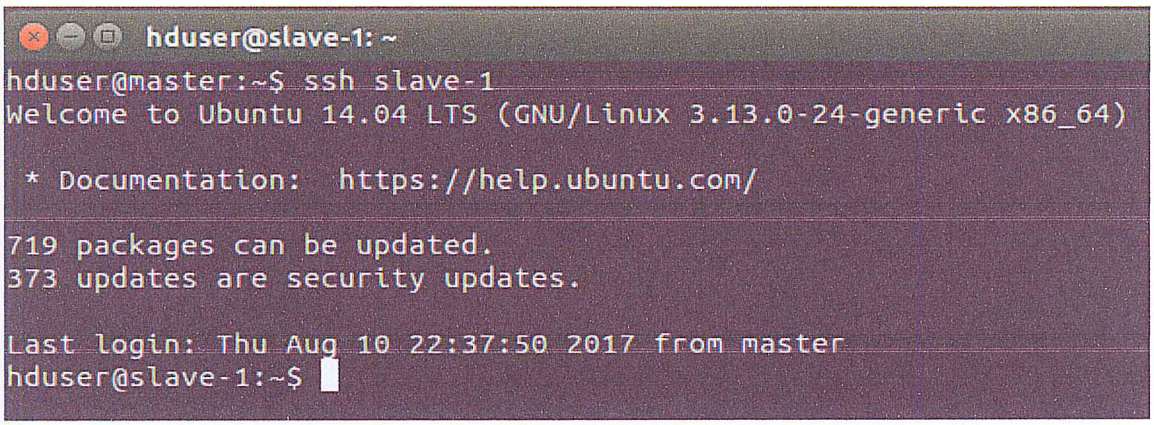


Figure VI.3 : *Configuration de l'accès SSH.*

VI.3. Déploiement de l'environnement de test :

Pour évaluer notre travail, nous avons configurer et créer un cluster de Trois Nœud **« Master »** ,**« Slave-1 »** et **« Slave-2 »** sur lesquels nous avons fait les configurations nécessaires qui inclurent essentiellement l'installation du framework Apache Hadoop chargé de la distribution.

VI.3.1Architecture du cluster Mise en place :

Le schéma de la figure VI.4 présente l'architecture du cluster hadoop qu'on a mis en place dans le cadre de ce travail.

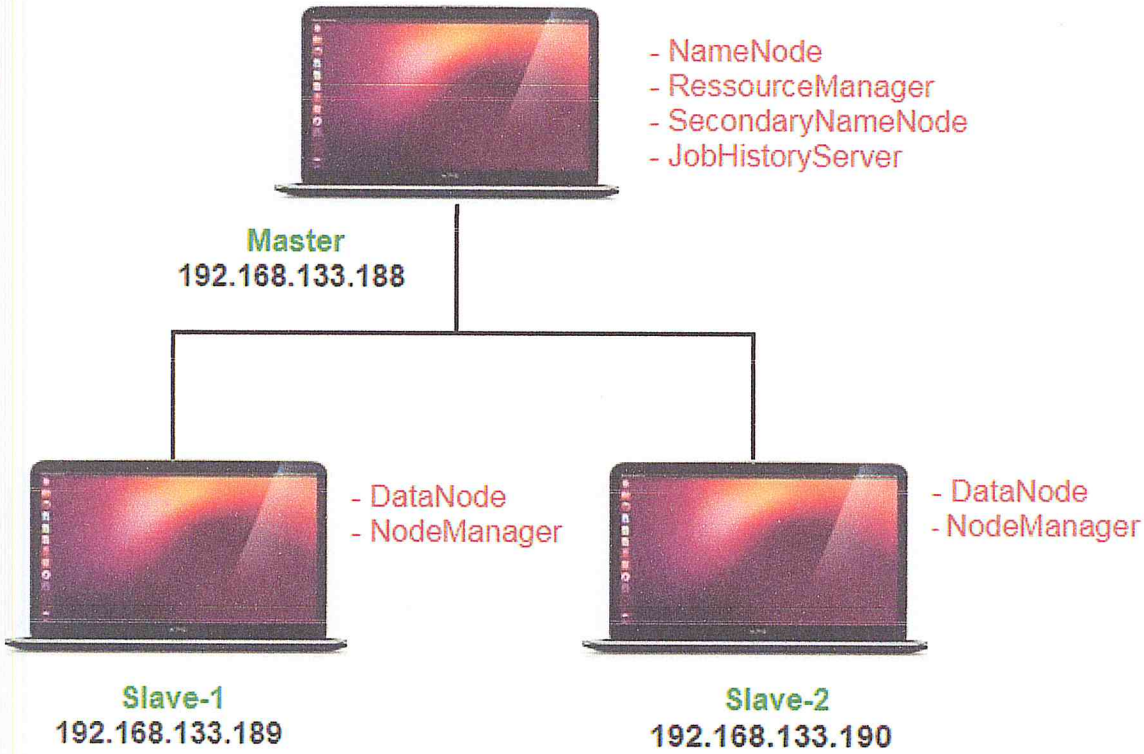
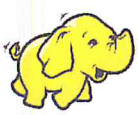


Figure VI.4:*Architecture du cluster Hadoop mis en place.*

Ce cluster est constitué de postes standards équipés de système d'exploitation Ubuntu (version 14.04). Cette architecture est hébergée dans un environnement virtuel, ce qui nous a permis de tester la virtualisation d'un cluster Hadoop, solution incontournable pour faire du Big Data sur le cloud. Ce schéma présente les différentes machines (maître et esclave) du cluster et les rôles qui leurs sont associés dans le cadre d'une architecture Hadoop.

VI.4. Ecosystème Hadoop :



Hadoop est un ensemble de logiciels et d'outils qui permettent de créer des applications distribuées, C'est une plate-forme logicielle Open-Source, écrite en java et fondée sur le modèle MapReduce de Google et les systèmes de fichiers distribués(HDFS). Elle permet de prendre en charge les applications distribuées en analysant de très grands ensembles de données.

Hadoop est utilisé particulièrement dans l'indexation et le tri de grands ensembles de données, le Data Mining, l'analyse de logs et le traitement d'images, le succès de Google lui est en partie imputable, en 2001, alors qu'il n'en est encore qu'à ses balbutiements sur le marché des moteurs de recherche, le futur géant développe ce qui inspira les composants phares d'Hadoop :

MapReduce, Google Big Table et Google BigFiles (futur Google file System), ces deux points forment l'écosystème Hadoop, écosystème fortement convoité et qui se trouve au centre de l'univers de Big data. Le schéma ci-après présente les différents éléments de l'écosystème Hadoop en fonction du type d'opération.

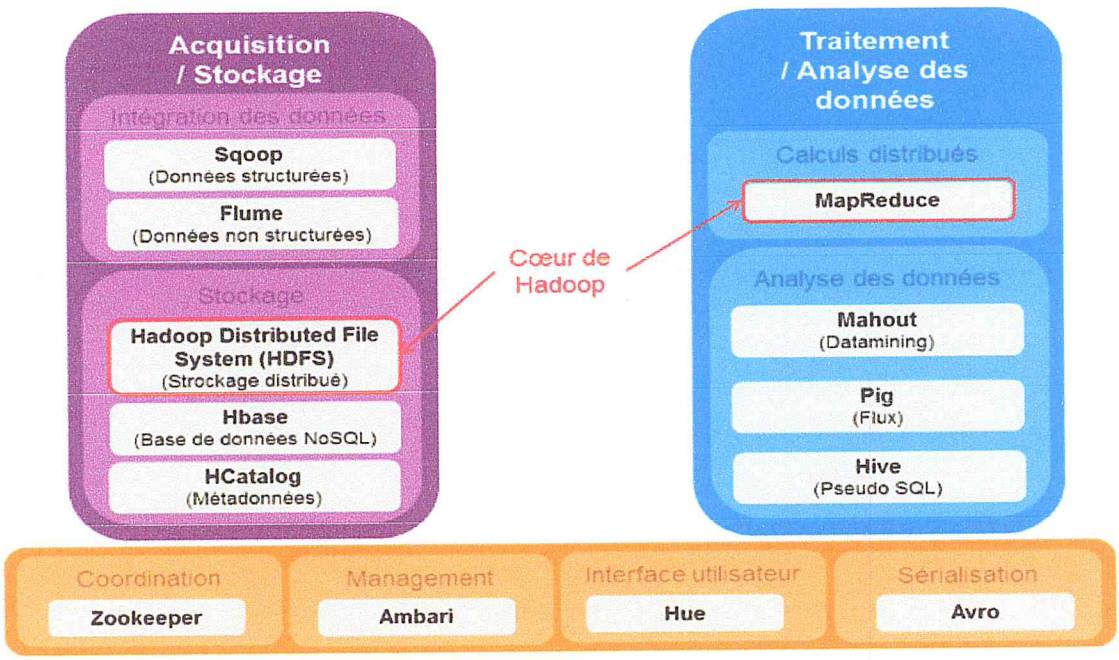


Figure VI.5 : *EcosystèmeHadoop.*

Hadoop est :

- Évolutif, car il utilise plus de ressources physiques, selon les besoins, et de manière transparente.
- Rentable, car il optimise les coûts via une meilleure utilisation des ressources présentes.
- Souple, car il répond à la caractéristique de variété des données en étant capable de traiter différents types de données.

- Et enfin, résilient, car il ne perd pas l'information et est capable de poursuivre le traitement si un nœud du système tombe en panne.

VI.4.1. Hadoop distributed File System (HDFS):

HDFS est le système de fichiers java, permettant de gérer le stockage des données sur des machines d'une architecture Hadoop. Il s'appuie sur le système de fichier natif de l'OS (unix) pour présenter un système de stockage unifié reposant sur un ensemble de disques et de systèmes de fichiers. La consistance des données réside sur la redondance ; une donnée est stockée sur au moins n volumes différents.

HDFS (Hadoop Distributed File System) est un système inspiré du système GFS développé par Google. Il se démarque des autres systèmes de fichier distribués par sa grande tolérance aux fautes**[31]** et le fait qu'il soit conçu pour être déployé sur des machines à faible coût. HDFS fournit un haut débit d'accès aux données et est adapté pour les applications qui nécessitent de grands groupes de données. Il a une architecture de type maître/esclave. Un Cluster HDFS est constitué d'un unique *NameNode*, un serveur maître qui gère le système de fichier et notamment les droits d'accès aux fichiers. A cela s'ajoute des *DataNodes*, en général un par noeud dans le Cluster, qui gère le stockage des données affectés au noeud sur lequel elle se trouve. Lafigure VI.6 présente l'architecture de HDFS.

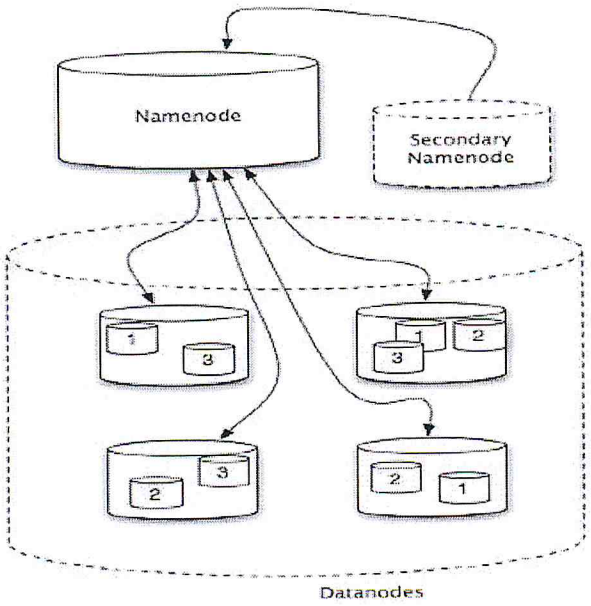
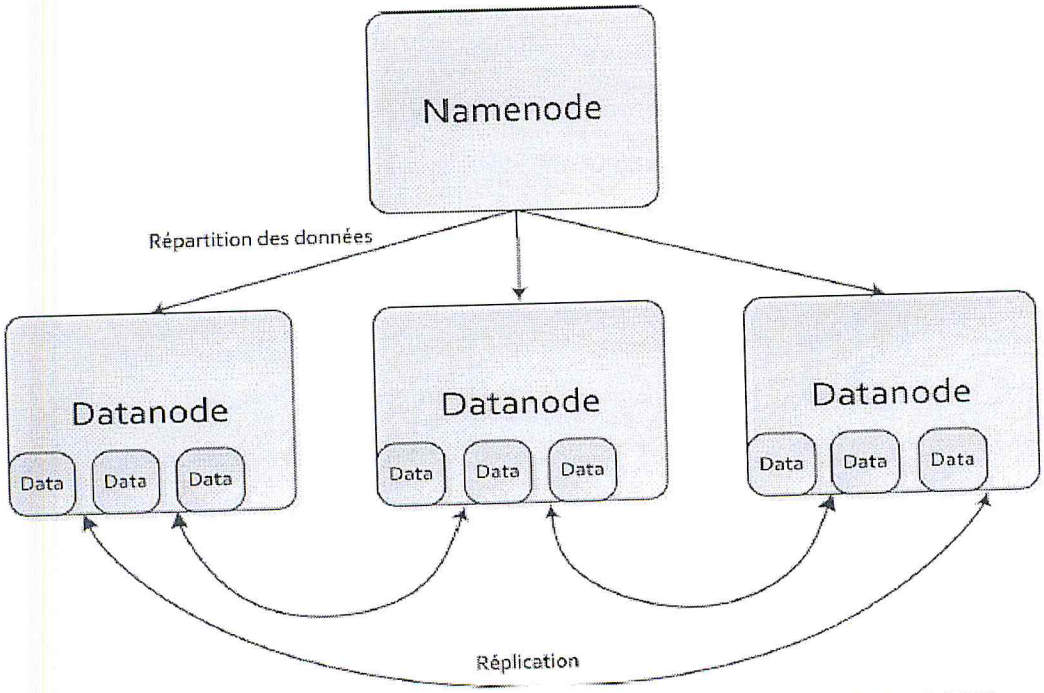


Figure VI.6: *Architecture de HDFS.*

Dans Hadoop, les différents types de données, qu'elles soient structurées ou non, sont stockées à l'aide de HDFS. Le HDFS va prendre les données en entrée et va ensuite les partitionner en plusieurs blocs de données. Afin de garantir une disponibilité des données en cas de panne d'un nœud, le système fera un réplica des données. Par défaut les données sont répliquées sur trois noeuds différents, deux sur le même support et un sur un support différent. Les différents nœuds de données peuvent communiquer entre eux pour rééquilibrer les



données sur la figure VI.7 on voit la réplication des données avec HDFS.**[32]**.

Figure VI.7:*Réplication des données à l'aide de HDFS.*

VI.4.2. Le principe maître/esclave :

Il est primordial de savoir qu'une architecture Hadoop est basée sur le principe maître/esclave, représentant les deux principaux rôles des machines. Les sous rôles relatifs au système de fichiers et à l'exécution des tâches distribuées sont associés à chaque machine de l'architecture.

Les machines maîtres ont trois principaux rôles qui leur sont associées :

- **JobTracker :** c'est le rôle qui permet à la machine maître de lancer des tâches distribuées, en coordonnant les esclaves. Il planifie les exécutions, gère l'état des machines esclaves et agrège les résultats des calculs.
- **NameNode :** ce rôle assure la répartition des données sur les machines esclaves et la gestion de l'espace de nom du cluster. La machine qui joue ce rôle contient des métadonnées qui lui permettent de savoir sur quelle machine chaque fichier est hébergé.
- **SecondaryNameNode** : ce rôle intervient pour la redondance du NameNode. Normalement, il doit être assuré par une autre machine physique autre que le NameNode car il permet en cas de panne de ce dernier, d'assurer la continuité de fonctionnement du cluster.

Deux rôles sont associés aux machines esclaves :

- **TaskTracker :** ce rôle permet à un esclave d'exécuter une tâche MapReduce sur les données qu'elle héberge. Le TaskTracker est piloté par JobTrackerd'une machine maître qui lui envoie la tâche à exécuter.
- **DataNode :** dans le cluster, c'est une machine qui héberge une partie des données. Les nœuds de données sont généralement répliqués dans le cadre d'une architecture Hadoop dans l'optique d'assurer la haute disponibilité des données.

Lorsqu'un client veut accéder aux données ou exécuter une tâche distribuée, il fait appel à la machine maître qui joue le rôle de JobTracker et de Namenode.

VI.4.3. Modes d'utilisation et d'installation :

Hadoop peut être sous trois modes différents :

a. Mode Standalone :

C'est le mode local et a pour objectif de tester le fonctionnement d'une tâche MapReduce. Ici, la tâche est exécutée sur le poste client dans la seule machine virtuelle Java (JVM), pas besoin d'une configuration particulière car c'est le mode de fonctionnement de base de Hadoop.

b. Mode Pseudo distributed :

Ce mode permettra de tester l'exécution d'une tâche MapReduce sur une seule machine tout en simulant le fonctionnement d'un cluster Hadoop. Je job est exécuté sur la machine et les opérations de stockage et de traitement du job seront gérées par des processus Java différents. L'objectif de ce mode est de tester le bon fonctionnement d'un job sans besoin de mobiliser toutes les ressources du cluster.

c. Mode Fullydistributed :

C'est le mode réel d'exécution d'Hadoop. Il permet de mobiliser le système de fichier distribué et les jobs MapReduce sur un ensemble de machines ; ceci nécessite de disposer de plusieurs postes pour héberger les données et exécuter les tâches.

Dans la suite, d'autres composants qui entrent dans l'écosystème Hahoop sont présentés.

VI.4.4. Composants Apache Hadoop :

- **Hbase**

Hbase est un système de gestion de bases de données non relationnelles distribuées, écrit en Java, disposant d'un stockage structuré pour les grandes tables. C'est une base de données NoSQL, orientée colonnes. Utilisé conjointement avec HDFS, ce dernier facilite la distribution des données de Hbase sur plusieurs nœuds. Contrairement à HDFS, Hbase permet de gérer les accès aléatoires read/write pour des applications de type temps réel.

- **HCatalog**

HCatalog permet l'interopérabilité d'un cluster de données Hadoop avec d'autres systèmes (Hive, Pig, ...). C'est un service de gestion de tables et de schéma Hadoop. Il permet :

- D'attaquer les données HDFS via des schémas de type tables de données en lecture/écriture.
- D'opérer sur des données issues de MapReduce, Pig ou Hive.

- **Hive : « Requêtage des données »**

Hive est un outil de requêtage des données, il permet l'exécution de requêtes SQL sur le cluster Hadoop en vue d'analyser et d'agréger les données. Le langage utilisé par Hive est nommé HiveQL. C'est un langage de visualisation uniquement, raison pour laquelle seules les

instructions de type « Select » sont supportées pour la manipulation des données.Hive proposent des fonctions prédéfinies (calcul de la somme, du maximum, de la moyenne), il permet également à l'utilisateur de définir ses propres fonctions qui peuvent être de 3 types :

- **UDF (User Defined Function) :** qui prennent une ligne en entrée et retournent une ligne en sortie. Exemple : mettre une chaîne de caractère en minuscule et inversement
- **UDAF (User Defined Aggregate Function) :** qui prennent plusieurs lignes en entrée et retournent une ligne en sortie. Exemple : somme, moyenne, max....
- **UDTF (User Defined Table Function) :** qui prennent une ligne en entrée et retournent plusieurs lignes en sortie. Exemple : découper une chaîne de caractère en plusieurs mots.

Hive utilise un connecteur jdbc/odbc, ce qui permet de le connecter à des outils de création de rapport comme QlikView.

- **Pig : « Scripting sur les données »**

Pig est une brique qui permet le requêtage des données Hadoop à partir d'un langage de script (langage qui interprète le code ligne par ligne au lieu de faire une compilation). Pig est basé sur un langage de haut niveau appelé PigLatin. Il transforme étape par étape des flux de données en exécutant des programmes MapReduce successivement ou en utilisant des méthodes prédéfinies du type calcul de la moyenne, de la valeur minimale, ou en permettant à l'utilisateur de définir ses propres méthode appelées User Defined Functions (UDF).

- **Sqoop : « Intégration SGBD-R »**

Sqoop est créé par Cloudera, Sqoop est un outil Hadoop en ligne de commande permettant d'échanger les données entre Hadoop et les SGBDR.

Sqoop gère les données stockées dans HDFS, Hive et HBase, et fonctionne dans les deux sens de Hadoop vers les SGBDR et vice versa. Il permet :

- ✓ d'importer des tables individuellement ou des schémas entiers vers HDFS ;
- ✓ de générer des classes Java qui permettent d'interagir avec les données importées ;
- ✓ d'exporter les données de HDFS vers les SGBDR.

- **Flume**

Flume permet la collecte et l'agrégation des fichiers logs, destinés à être stockés et traités par Hadoop. Il s'interface directement avec HDFS au moyen d'une API native.

- **Oozie : « Ordonnanceur »**

Oozie est utilisé pour gérer et coordonner les tâches de traitement de données à destination de Hadoop. Il supporte des jobs Mapreduce, Pig, Hive, Sqoop, etc.

- **Zookeeper**

Zookeeper est une solution de gestion de cluster Hadoop. Il permet de coordonner les tâches des services d'un cluster Hadoop. Il fournit au composants Hadoop les fonctionnalités de distribution.

- **Ambari**

Ambari est une solution de supervision et d'administration de clusters Hadoop. Il propose un tableau de bord qui permet de visualiser rapidement l'état d'un cluster. Ambari inclut un système de gestion de configuration permettant de déployer des services d'Hadoop ou de son écosystème sur des clusters de machines. Il ne se limite pas à Hadoop mais permet de gérer également tous les outils de l'écosystème.

- **Mahout**

Mahout est un projet de la fondation Apache visant à créer des implémentations d'algorithmes d'apprentissage automatique et de datamining.

- **Avro**

Avro est un format utilisé pour la sérialisation des données.

Le caractère open source de Hadoop a permis à des entreprises de développer leur propre distribution en ajoutant des spécificités.

La vue d'ensemble de la plateforme hadoop et ses différents composants est présenté sur la figure VI.8

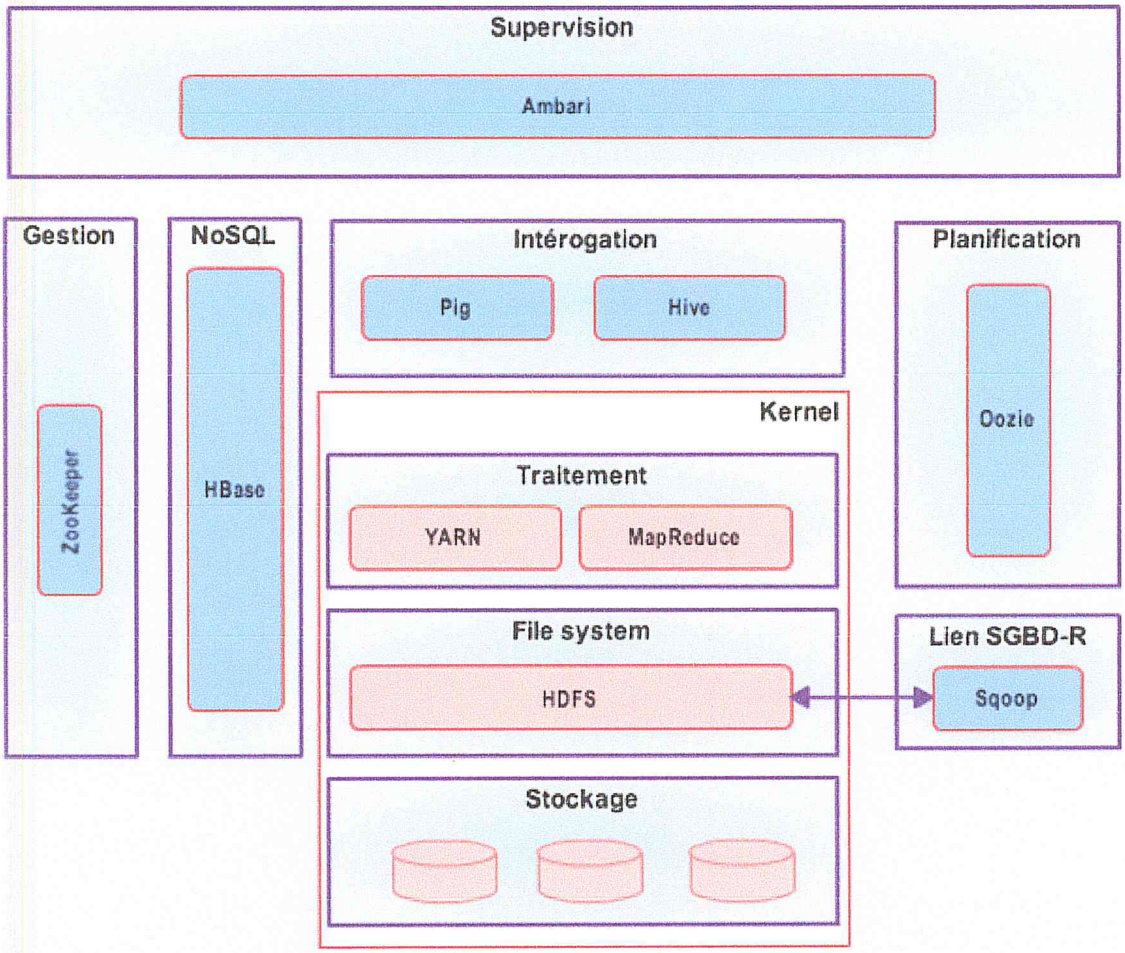


Figure VI.8:*La Plateforme hadoop.*

VI.5. Conclusion :

Dans ce chapitre, nous nous sommes intéressés à l'écosystème Hadoop. Nous avons d'abord présenté l'architecture du cluster Hadoop mise en place, nous avons détaillé par la suite le fonctionnement de ces deux technologies HDFS et son architecture maitre/esclave. Ensuite, nous avons décrit les différents composants du framework Hadoop et illustré une plateforme générale pour cet écosystème

L'implémentation et l'évaluation du système feront l'objet des prochains chapitres.

Chapitre VII

Description des applications

VII.1. Introduction

Dans ce chapitre nous allons présenter le résultat des expérimentations que nous avons pu faire afin de voir si nous avons abouti nos objectifs.

VII.2. L'environnement de test :

L'environnement de travail est un choix décisif pour **l'ingestion** et **le stockage**, notre choix s'est porté sur plusieurs critères dont nous citerons :

- ✓ Hadoop, qui permet de traiter plusieurs TeraOctets de données.
- ✓ Pentaho data integration (PDI), pour l'extraction, le nettoyage, les transformations et le chargement dans un environnement distribué.
- ✓ Pentaho Business Analytic (BA), pour l'analyse et la restitution des données.
- ✓ Hive, qui facilite les agrégations, le requêtage ad-hoc et l'analyse de gros volumes de données stockés dans le système de fichiers distribués
- ✓ Sqoop qui permet d'importer et de transférer efficacement une masse de données de ou vers une base de données.
- ✓ Une base de données Mysql qui permet l'interaction simple et efficace des données.

VII.2.1. Données de test :

Pour le test nous avons utilisé des données Structurées provenant d'une **base de données Mysql** ainsi que des données **Semi-structurées** de type **XML** et **CSV**, les figures VII.1 etVII.2 représentent la structure des fichiers utilisées.

```
<OMeS>
 <PMSetup  startTime="2017-08-26T16:00:00.000+01:00:00" interval="15">
    <PMMOResult>
      <MO>
        <DN><![CDATA[PLMN-PLMN/MSC-224802]]></DN>
      </MO>
      <MO>
        <DN><![CDATA[PLMN-PLMN/CGR-132/CGRNAME-BSOK24]]></DN>
      </MO>
      <MO>
        <DN><![CDATA[PLMN-PLMN/CGRDIR-3]]></DN>
      </MO>
      <PMTarget  measurementType="CGR">
        <M16B2C10>766</M16B2C10>
        <M16B2C11>465</M16B2C11>
        <M16B2C12>746</M16B2C12>
        <M16B2C13>456</M16B2C13>
        <M16B2C16>263</M16B2C16>
        <M16B2C17>244</M16B2C17>
        <M16B2C18>0</M16B2C18>
        <M16B2C19>0</M16B2C19>
        <M16B2C20>0</M16B2C20>
        <M16B2C21>0</M16B2C21>
        <M16B2C22>20</M16B2C22>
        <M16B2C23>9</M16B2C23>
        <M16B2C24>938</M16B2C24>
        <M16B2C25>772</M16B2C25>
        <M16B2C27>62</M16B2C27>
        <M16B2C3>93</M16B2C3>
```

Figure VII.1 : *Fichier de test « MSS.xml ».*

Le fichier représente une capture sur l'état d'une celluledans un **intervalle** donnée et une date donnée **« startTime »**pour un Type de mesure données **« measurementType = CGR »** ici on a pris l'exemple du type CGRavec la topologie de son réseau **PLMN**(MSC, CGRNAME, CGR, CGRDIR) et ses compteurs.

	A	B	C	D
1	Time,Integrity,Cell Unavaialble Time Ratio(%),RNC			
2	02/02/2017 00:00,100,2.2852,Bejaia RNC			
3	02/02/2017 00:00,100,0.2804,CNE RNC02			
4	02/02/2017 00:00,100,0.652,ORAN RNC 2			
5	02/02/2017 00:00,100,2.8587,Rouiba RNC			
6	02/02/2017 00:00,100,1.2938,Blida RNC			
7	02/02/2017 00:00,100,2.0942,Tiziouzou RNC			
8	02/02/2017 00:00,100,4.7248,Skikda RNC			
9	02/02/2017 00:00,100,3.0885,Delly Brahim			
10	02/02/2017 00:00,100,0.8296,CNE RNC			
11	02/02/2017 00:00,100,3.3035,Setif RNC			
12	02/02/2017 00:00,100,0.7547,Mostaganem RNC			
13	02/02/2017 00:00,100,1.0323,Oran RNC			
14	02/02/2017 00:00,100,1.9231,Chlef RNC			
15	02/02/2017 00:00,100,0,Rouiba RNC 2			
16	02/02/2017 00:00,100,0,Annaba RNC			
17	02/02/2017 01:00,100,0.652,ORAN RNC 2			
18	02/02/2017 01:00,100,2.2699,Bejaia RNC			
19	02/02/2017 01:00,100,0.2804,CNE RNC02			
20	02/02/2017 01:00,100,0.7908,CNE RNC			
21	02/02/2017 01:00,100,3.3035,Setif RNC			
22	02/02/2017 01:00,100,1.915,Tiziouzou RNC			

Huawei_Sharing

Figure VII.2 :*Fichier de test « Huawei_Sharing.csv ».*

VII.3. Installations et configuration du Cluster Hadoop :

VII.3.1. Accès aux interfaces utilisateurs :

La distribution Hadoop par Apache fournit des interfaces utilisateurs pour l'administration du cluster. Ceux-ci sont accessibles via des applications Web.

La première concerne l'état du cluster et est accessible via l'adresse :

« Http : //master:8088 ». Il vous est possible d'avoir une vue globale sur les nœuds du cluster et sur lesjobsencoursd'exécution.Unecaptured'écranestdonnéesur la figure VII.3.

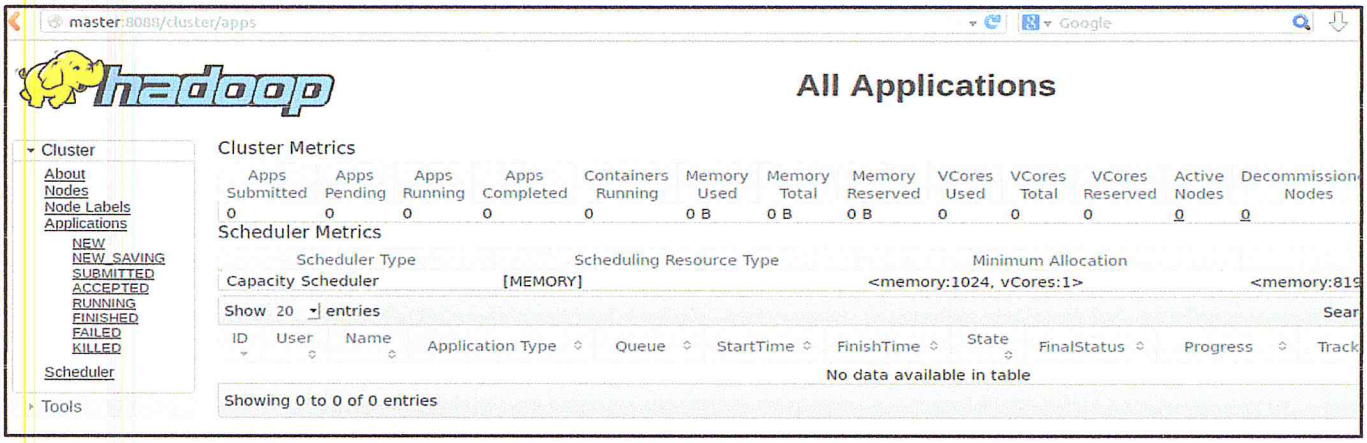


Figure VII.3 : *Interface des applications installées.*

La deuxième interface utilisateur présenté sur la figure VII.4concerne l'accès aux données contenues dans le nœud NameNode et est accessible via l'adresse **« http : //master :50070 ».** Elle permet d'obtenir des informations sur la capacité totale et connaître l'état de disponibilité des nœuds. Elle permet également d'avoir des informations sur les fichiers stockés dans HDFS.

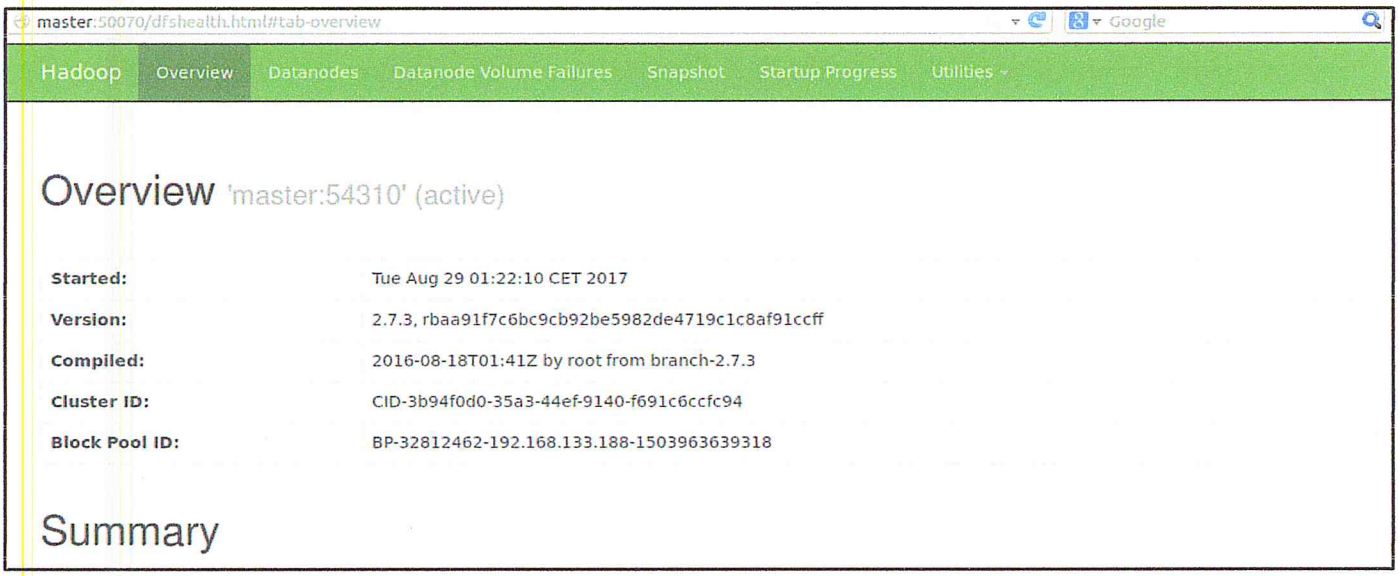


Figure VII.4 : *Interface du nœud Maitre « Master ».*

La figure VII.5 montre l'interface suivante est celle des nœuds esclave du cluster 'slave-1 ' et 'slave-2' et sur laquelle on peut voir si les nœuds sont actifs ou non, et est accessible via l'adresse **« http : //master :50070 ».**

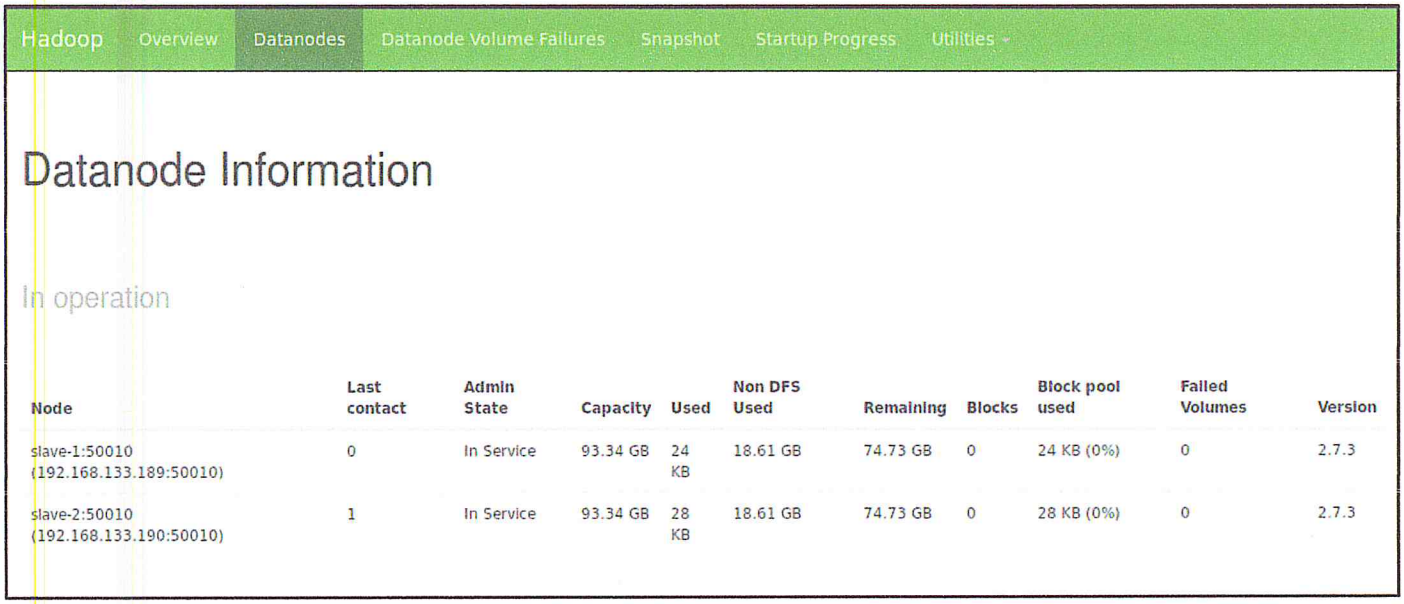
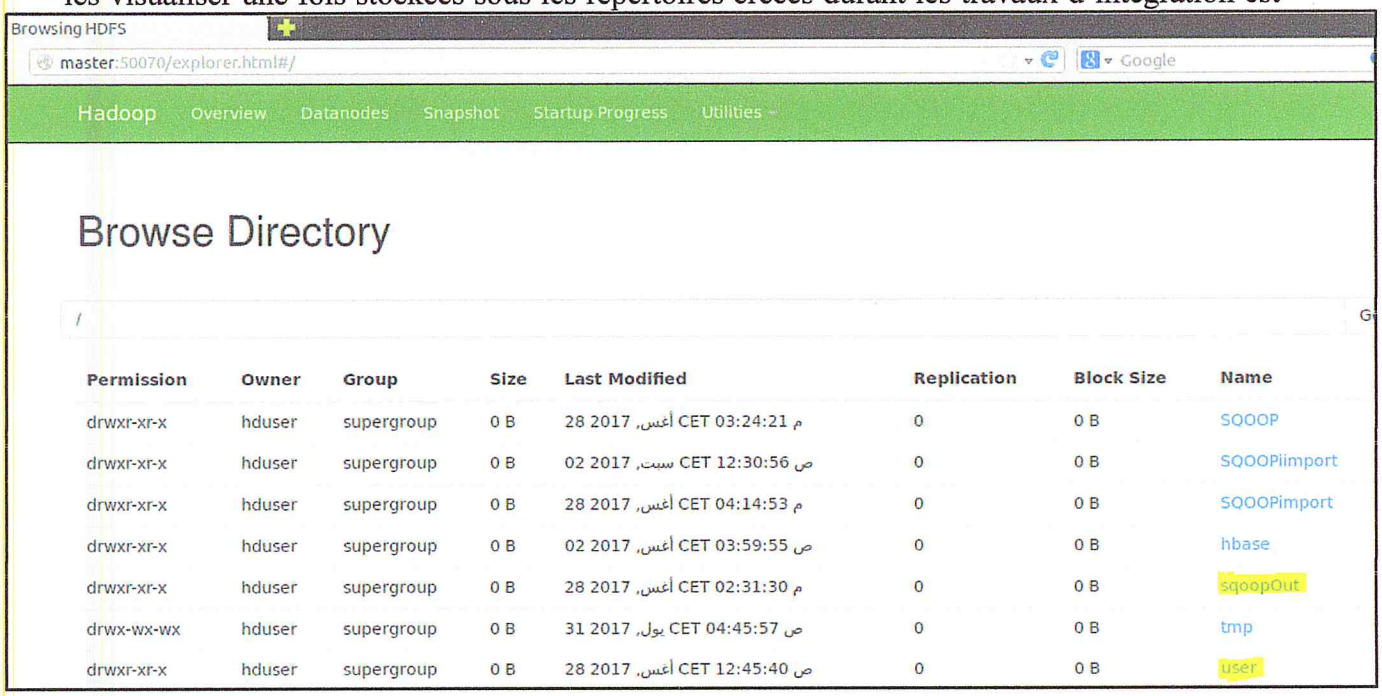


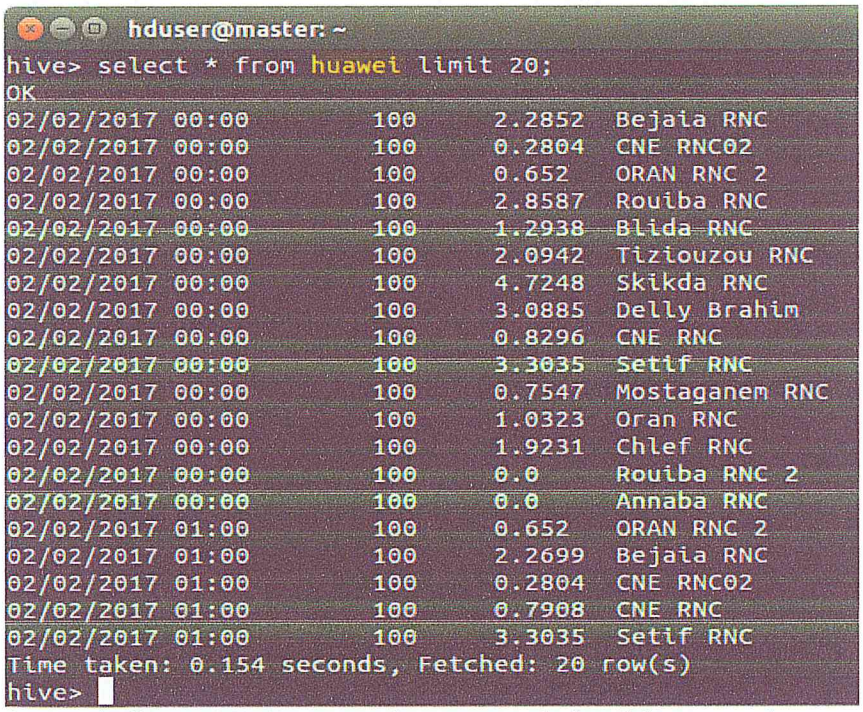
Figure VII.5 : *Interface des nœuds esclaves du cluster (slave-1/slave-2).*

Et l'interface du système de fichier distribué HDFS qui permet d'accéder aux données et de les visualiser une fois stockées sous les répertoires créées durant les travaux d'intégration est



sur la figure VII.5.

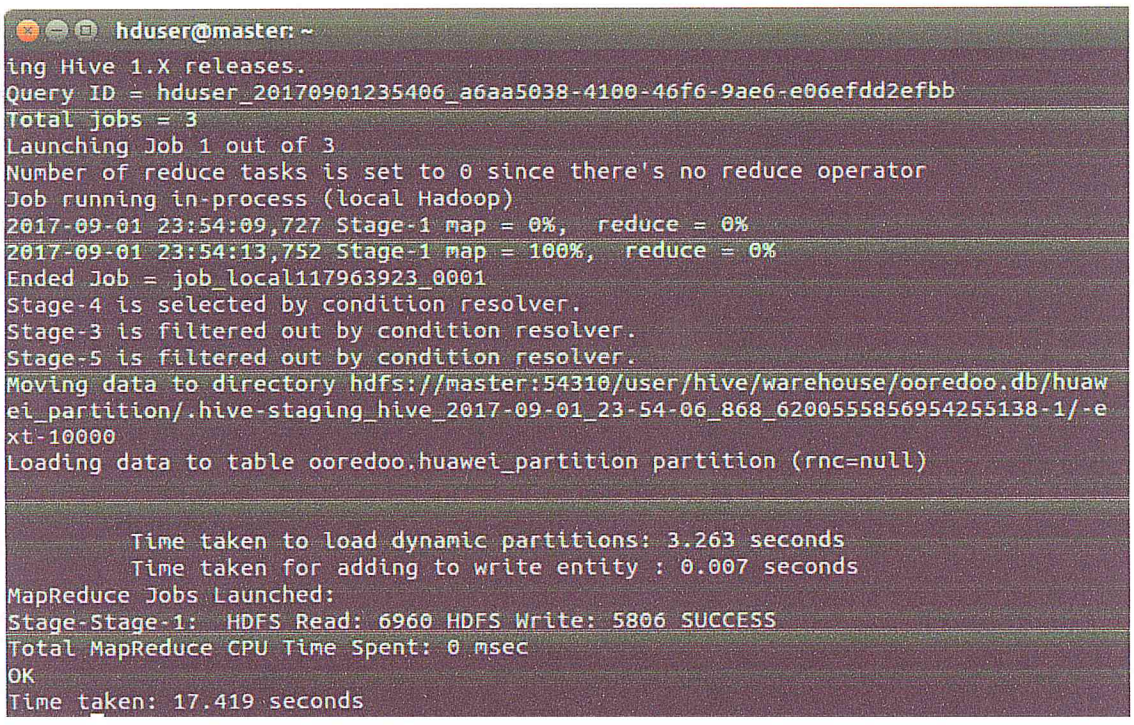
Nous avons créé une table dans Hive que nous avons chargé aves les données de notre



fichier ' huawei.csv', la figure **VII.7** montre le contenu de la table en entré.

Figure VII.7 : *Stockage des données dans la table « huawei ».*

Une fois la table remplie avec nos données de test on entapmera la phase de création des partitions dans le système de fichier distribué HDFS, on lance un job Map/Reduce pour partitionner la table selon le format voulu comme le montre la figure VII.8.



VII.4. Tests technique réalisés :

- ✓ **Test 1 :** Nous allons faire un exemple de partitionnement de données dans un environnement complétement distribué.
- ✓ **Test 2 :** Intégration, transformation et analyse des données de test.

VII.4.1. But des tests :

Lorsqu'on réalise un test, il est important de d'utiliser une démarche précise et de définir les objectifs à atteindre à la fin du test, ce qui permettra de faire un contrôle à la fin, par rapport aux résultats obtenus pour savoir si le test a été satisfaisant ou non.

La démarche a consisté à tester les composants séparément ou conjointement, à identifier les fonctionnalités de chacun.

Les objectifs des différents tests étaient :

- D'acquérir une meilleure connaissance de chaque composant
- Pouvoir présenter le fonctionnement de chaque composant
- Faire des recommandations par rapport à l'utilisation d'un composant
- Montrer la mise en œuvre de certains concepts
- Répondre aux besoins de l'entreprise

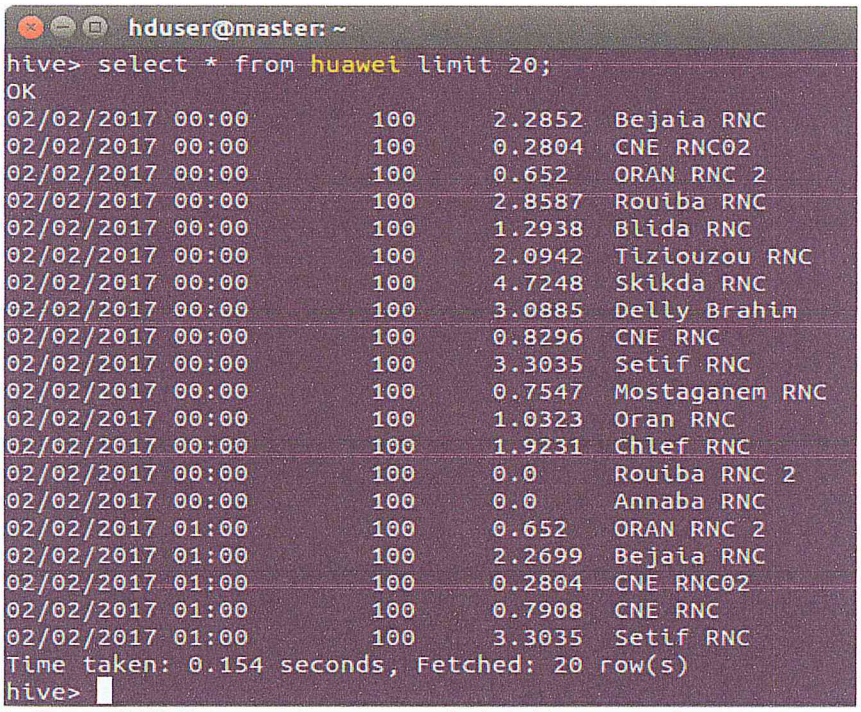
VII.5. Implémentation de la solution :

✓ Test 1 :

- **Tests de partitionnement :**

Pour les tests de partition, nous avons notre fichier « Huawei_sharing.csv » présenté précédemment, on nous a demandé de partitionner les données par apport au champ RNC, une fois stocké dans HIVE nous effectuerons les changements nécessaires.

Nous avons créé une table dans Hive que nous avons chargé aves les données de notre fichier



‘ huawei.csv’, la figure **VII.7** montre le contenu de la table en entré.

Figure VII.7 : *Stockage des données dans la table « huawei ».*

Une fois la table remplie avec nos données de test on entamera la phase de création des partitions dans le système de fichier distribué HDFS, on lance un job Map/Reduce pour partitionner la table selon le format voulu comme le montre la figure VII.8.

Figure VII.8 : *Job Map/Reduce pour la création des partitions dans hdfs.*

L'étape est exécutée avec succès, et les partitions sont créée dans HDFS, les figures VII.9, VII.10,VII.11 présentent cette exécution.

```
hduser@master: ~
hive> select * from huawei_partition limit 22;
OK
02/02/2017 00:00        100     0.0     Annaba RNC
02/02/2017 01:00        100     0.0     Annaba RNC
02/02/2017 00:00        100     0.0     Annaba RNC
02/02/2017 01:00        100     0.0     Annaba RNC
02/02/2017 00:00        100     2.2852  Bejaia RNC
02/02/2017 01:00        100     2.2699  Bejaia RNC
02/02/2017 02:00        100     2.3353  Bejaia RNC
02/02/2017 00:00        100     2.2852  Bejaia RNC
02/02/2017 01:00        100     2.2699  Bejaia RNC
02/02/2017 02:00        100     2.3353  Bejaia RNC
02/02/2017 00:00        100     1.2938  Blida RNC
02/02/2017 01:00        100     1.2938  Blida RNC
02/02/2017 00:00        100     1.2938  Blida RNC
02/02/2017 01:00        100     1.2938  Blida RNC
02/02/2017 00:00        100     0.8296  CNE RNC
02/02/2017 01:00        100     0.7908  CNE RNC
02/02/2017 00:00        100     0.8296  CNE RNC
02/02/2017 01:00        100     0.7908  CNE RNC
02/02/2017 00:00        100     0.2804  CNE RNC02
02/02/2017 01:00        100     0.2804  CNE RNC02
02/02/2017 02:00        100     0.2804  CNE RNC02
02/02/2017 00:00        100     0.2804  CNE RNC02
Time taken: 0.177 seconds, Fetched: 22 row(s)
hive>
```

Figure VII.9 : *Partitionnement des données.*

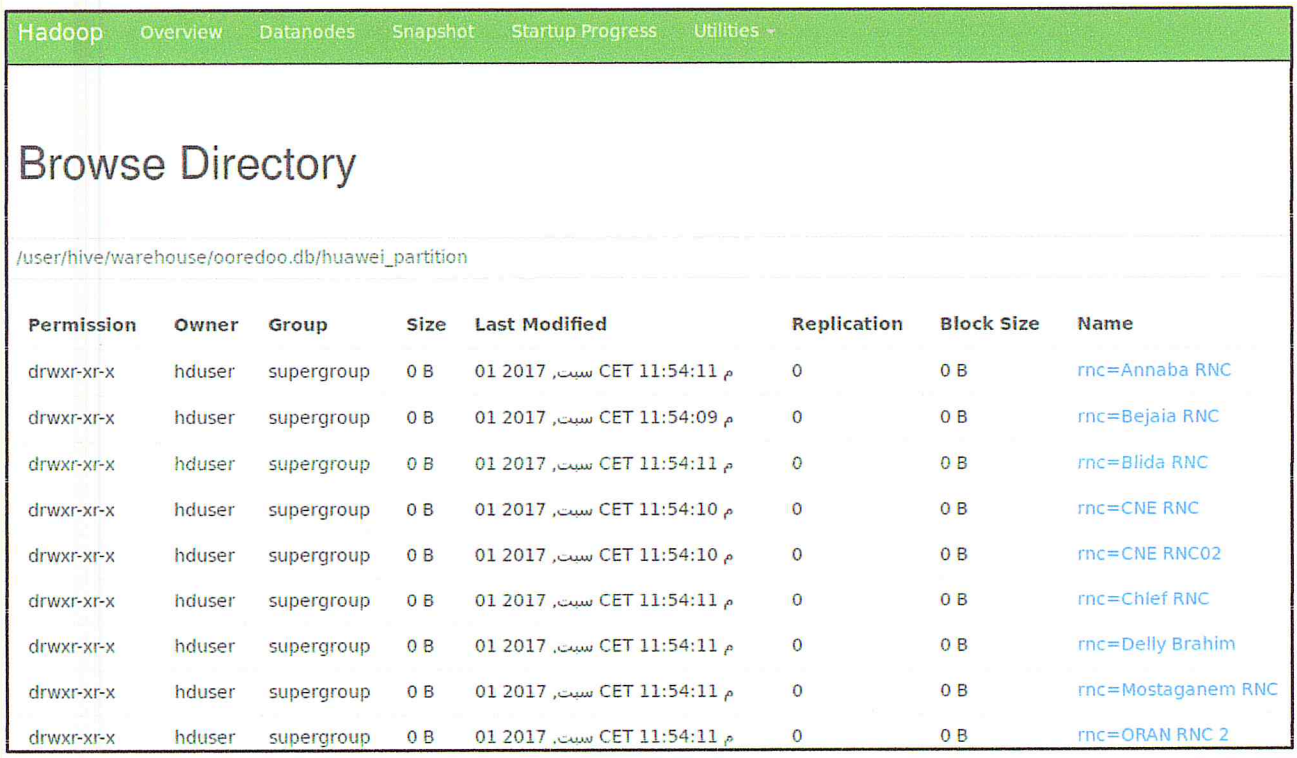


Figure VII.10 :*Aperçue des partitions créées dans HDFS.*

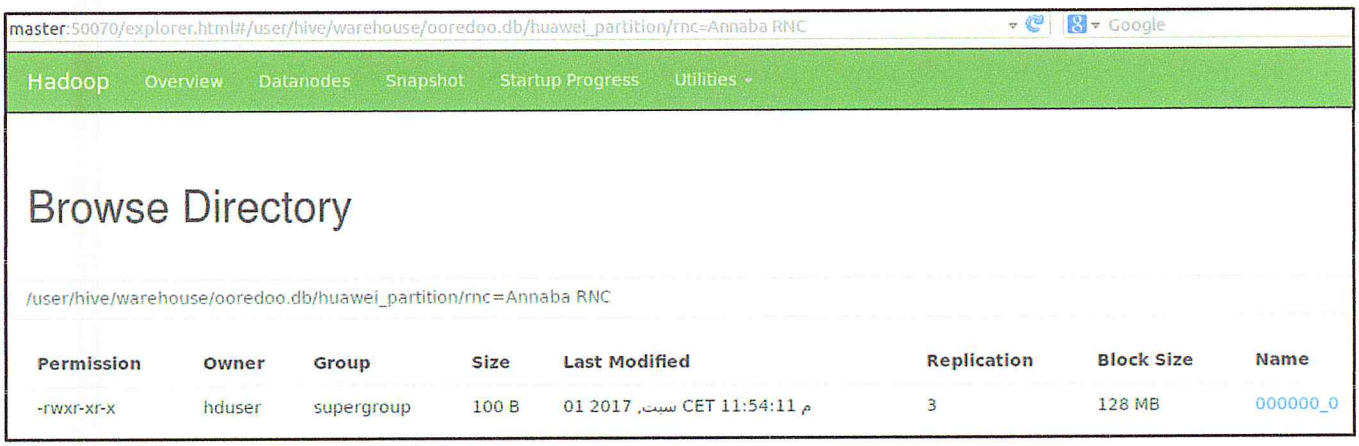


Figure VII.11 :*Aperçue du fichier créé de la partition d'exemple « RNC =Annaba RNC ».*

✓ **Test 2 :**

- **Alimentation de la dimension Temps :**

Nous allons illustrer les étapes d'extraction faites sur nos données de test, on commence par extraire le champ startTime qui représente la date de la capture des données illustré sur la figure VII.12.

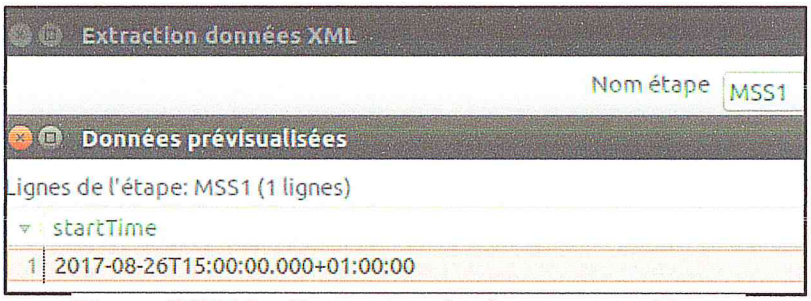


Figure VII.12 : *Extraction du champ « startTime ».*

Cette date et sous le format DATETIME on doit faire un éclatement de champ pour séparer la date et le temps comme le montre la figure VII.13.

Figure VII.13 : *Eclatement du champ « startTime ».*

On doit supprimer les champs inutiles comme **startTime** qu'on a décomposé en Date, Année, Mois, Jour et Heure, on garde seulement ces 5 champs présents sur la Figure VII.14.

startTime	Date	Année	Mois	Jour	Heure
2017-08-26T16:00:00.000+01:00:00	2017-08-26	2017	08	26	16:00
2017-08-26T16:00:00.000+01:00:00	2017-08-26	2017	08	26	16:00
2017-08-26T16:00:00.000+01:00:00	2017-08-26	2017	08	26	16:00

Figure VII.14 : *Suppression du champ « startTime » après l'éclatement.*

Supprimer les doublons pour enfin les stocké dans une base de données mysql, dans une table de dimension qu'on a nommé « TD_Temps » comme le montrent les Figure VII.15 et VII.26.

Date	Année	Mois	Jour	Heure
2017-08-26	2017	08	26	10:00
2017-08-26	2017	08	26	15:00
2017-08-26	2017	08	26	16:00

Figure VII.15 :*Alimentation de la table « TD_Temps ».*

- **Alimentation de la dimension Compteurs :**

1. **Extraction des compteurs et du type de mesure :**

La table TD_compteurs contient les compteurs du type de mesure CGR qu'on a choisie d'analyser et sur lequel on s'est basé pour atteindre nos objectifs. Présenté sur les figure VII.16, VII.17 et VII.27.

Données prévisualisées

Lignes de l'étape: MSS22 (3 lignes)

	measurementType	M16B2C10	M16B2C11	M16B2C16	M16B2C17	M16B2C24	M16B2C25	M16B2C3
1	CGR	766	465	263	244	938	772	93
2	CGR	686	420	241	218	871	925	93
3	CGR	684	331	153	126	581	428	93

Figure VII.16 : *Extraction des compteurs et du type de mesure.*

measurementType	M16B2C10	M16B2C11	M16B2C16	M16B2C17	M16B2C24	M16B2C25	M16B2C3
CGR	1089	687	424	418	1429	1477	93
CGR	1095	716	475	416	1561	1318	92
CGR	4296	4271	2016	2571	6747	8162	589
CGR	713	443	263	210	938	710	92
CGR	679	426	256	221	803	733	91
CGR	766	465	263	244	938	772	93
CGR	686	420	241	218	871	925	93
CGR	684	331	153	126	581	428	93

Figure VII.17 :*Alimentation de la table « TD_compteurs ».*

- **Alimentation de la dimension Topologie :**

1. **Extraction de la topologie du réseau :**

La topologie du réseau est très importante dans notre cas, c'est pour cela qu'on a récupérer ses champs ainsi que leurs valeurs pour les transformer selon le schéma voulu comme le montrent les figure VII.18, VII.22 et VII.25.

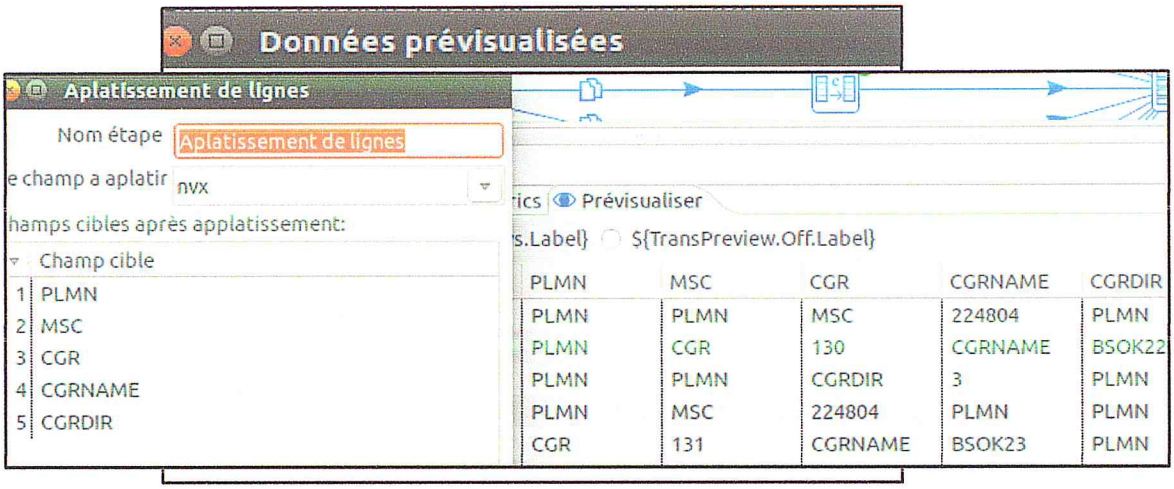


Figure VII.18 : *Extraction de la topologie du type « CGR ».*

2. Transformations et Chargement :

Pour schématiser la topologie extraite selon le format voulu on a dû faire plusieurs transformations :

✓ **Décomposition de champ en ligne :** sert à éclater un champ en plusieurs lignes dans notre cas on veut obtenir les champs suivants : (PLMN, MSC, CGR, CGRDIR, CGRNAME) qui sont séparés par des '/' et des tirets '-'. Présenté sur la figure VII.19.

DN	nv	nvx
PLMN-PLMN/MSC-224804	PLMN	PLMN
PLMN-PLMN/MSC-224804	PLMN/MSC	PLMN
PLMN-PLMN/MSC-224804	PLMN/MSC	MSC
PLMN-PLMN/MSC-224804	224804	224804
PLMN-PLMN/CGR-130/CGRNAME-BSOK22	PLMN	PLMN
PLMN-PLMN/CGR-130/CGRNAME-BSOK22	PLMN/CGR	PLMN
PLMN-PLMN/CGR-130/CGRNAME-BSOK22	PLMN/CGR	CGR
PLMN-PLMN/CGR-130/CGRNAME-BSOK22	130/CGRNAME	130

Figure VII.19 : *Décomposition des champs de la topologie en lignes.*

✓ **Aplatissement de ligne :** Sert à aplatir les lignes en colonnes comme le montre la figure VII.20.

Figure VII.20 :*Aplatissement des champs de la topologie.*

- ✓ **Remplacer Valeur d'un champ par constante :** Sert à nettoyer les champs, à enlever les valeurs inutiles et les remplacer par les valeurs voulus. Tel la figure VII.21.

PLMN	MSC	CGR	CGRNAME	CGRDIR
PLMN	224804	130	BSOK22	3

Figure VII.21 :*Récupération des valeurs voulues.*

- ✓ **Dé doublonnage de ligne (HashSet) :** Sert à supprimer les doublons de chaque colonne.
- ✓ **Altération structure flux :** Sert à enlever, modifier et à insérer des colonnesselon le besoin. Comme le montre la Figure VII.22.

PLMN	MSC	CGR	CGRNAME	CGRDIR
PLMN	224804	130	BSOK22	3
PLMN	224804	131	BSOK23	3
PLMN	206322	129	BSOK21	3
PLMN	224802	132	BSOK24	3
PLMN	224802	328	BBCH24	3
PLMN	206322	303	BAL100	3
PLMN	206322	128	BSOK20	3
PLMN	224802	133	BSOK25	3

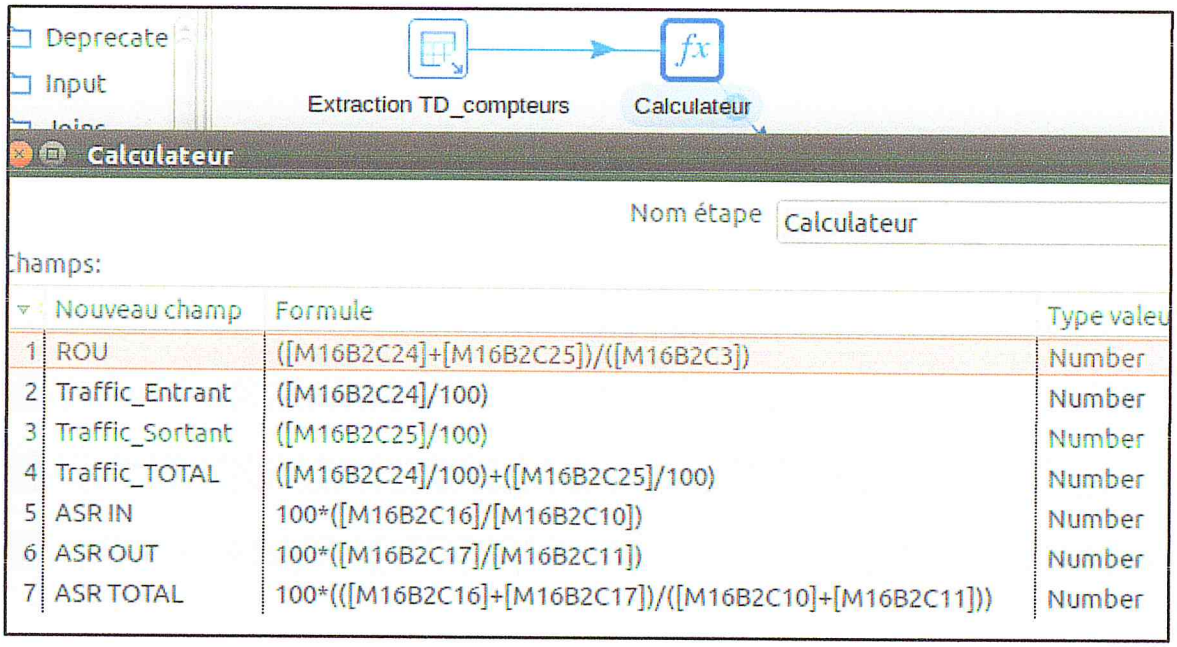
Figure VII.22 : *Alimentation de la table « TD_topologie ».*

- **Alimentation de la table de FAIT Traffic**

1. **Récupération des données à partir des tables de dimension :**

Afin d'alimenter notre table de fait on a récupérer les clés primaire des tables de dimension Temps, Topologie et Compteurs (id_temp, id_topologie et id_compteur) , et nous avons calculé les mesure sur les quelles notre analyse se porte , on doit connaitre le traffic entrant et le traffic sortant pour chaque type de mesure ,connaitre aussi ASR IN, ASR OUT,

On se basant sur les compteurs du type de mesure CGR on a pu les calculé avec les formules présentées sur la figure VII.23.



	Nouveau champ	Formule	Type valeu
1	ROU	([M16B2C24]+[M16B2C25])/([M16B2C3])	Number
2	Traffic_Entrant	([M16B2C24]/100)	Number
3	Traffic_Sortant	([M16B2C25]/100)	Number
4	Traffic_TOTAL	([M16B2C24]/100)+([M16B2C25]/100)	Number
5	ASR IN	100*([M16B2C16]/[M16B2C10])	Number
6	ASR OUT	100*([M16B2C17]/[M16B2C11])	Number
7	ASR TOTAL	100*(([M16B2C16]+[M16B2C17])/([M16B2C10]+[M16B2C11]))	Number

Figure VII.23 : *Les formules de mesures et d'analyse.*

- **Traffic Entrant :** C'est quand l'opérateur ooredoo reçoit un appel d'un opérateur extérieur (ex : Djezzy , Mobilis) .
- **Traffic Sortant :** C'est quand l'opérateur ooredoo contact les opérateurs extérieurs.
- **Traffic Total** : C'est la somme des deux trafique Entrant et Sortant.
- **ROU :** c'est le taux d'utilisationou d'occupation de la ligne
- **ASR** :c'est le pourcentage du nombre d'appels connectés avec succès au nombre d'appels tentés (c'est aussi appelé taux d'achèvement de l'appel) :

  ASR% = (nombre total d'appels répondués / nombre total d'appels) × 100

Par exemple, s'il y avait 156 appels composés dont 62 étaient connectés avec succès, alors:

ASR (%) = (62 [appels réussis] / 156 [appels composés]) x 100 = 39,74%

Par la suite on a fait une jointure des tables pour conclure le schéma final suivant qui comporte les clés primaire des tables de dimension comme clés étrangères ainsi que les mesures calculées comme le montrent les figures VII.24 et VII.28.

id_compteurs	ROU	Traffic_Entrant	Traffic_Sortant	Traffic_TOTAL	ASR IN	ASR OUT	ASR TOTAL	id_topologie
1	31,247311828	14,29	14,77	29,06	38,9348025712	60,8442503639	47,4099099099	1
2	31,2934782609	15,61	13,18	28,79	43,3789954338	58,1005586592	49,1993373827	2
3	25,3123938879	67,47	81,62	149,09	46,9273743017	60,1966752517	53,5426637096	3
4	18,3870967742	9,38	7,72	17,1	34,3342036554	52,4731182796	41,1860276198	4
5	19,311827957	8,71	9,25	17,96	35,1311953353	51,9047619048	41,5009041591	5
6	10,8494623656	5,81	4,28	10,09	22,3684210526	38,0664652568	27,4876847291	6
7	17,9130434783	9,38	7,1	16,48	36,8863955119	47,4040632054	40,9169550173	7
8	16,8791208791	8,03	7,33	15,36	37,7025036819	51,8779342723	43,1674208145	8

Figure VII.24 :*Alimentation de la table de fait « TF_Traffic ».*

- **Description des transformations de données :**

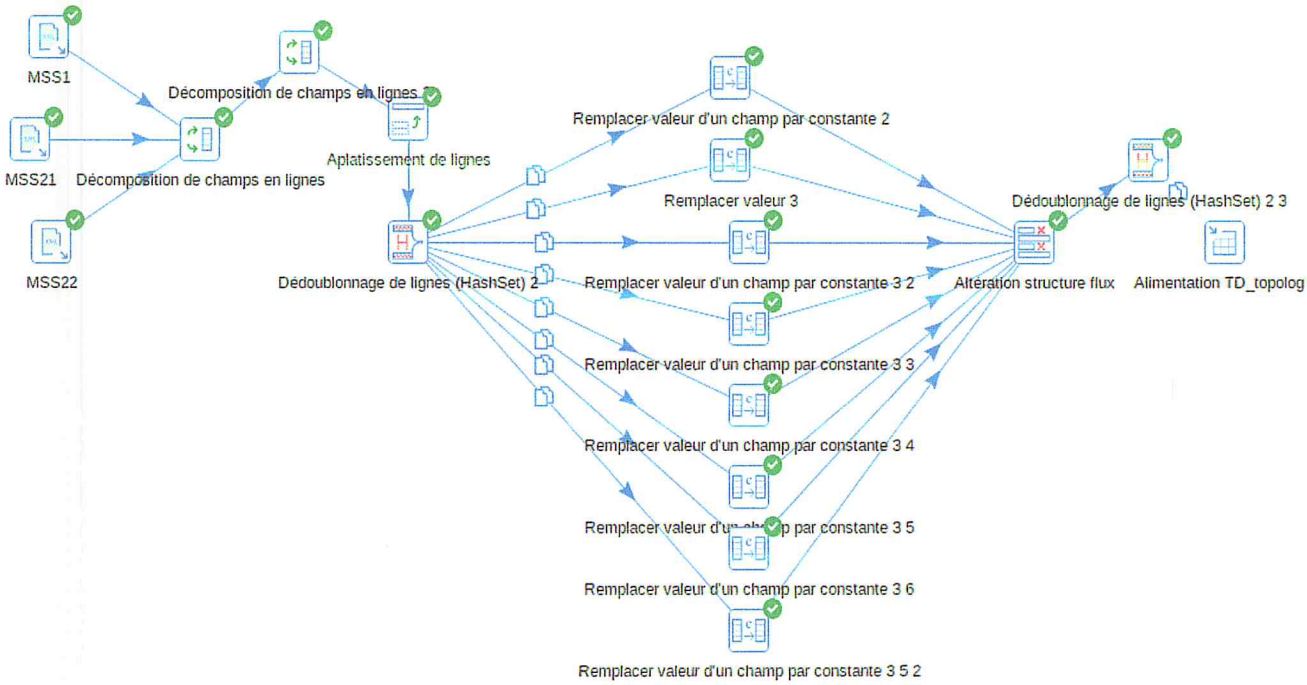
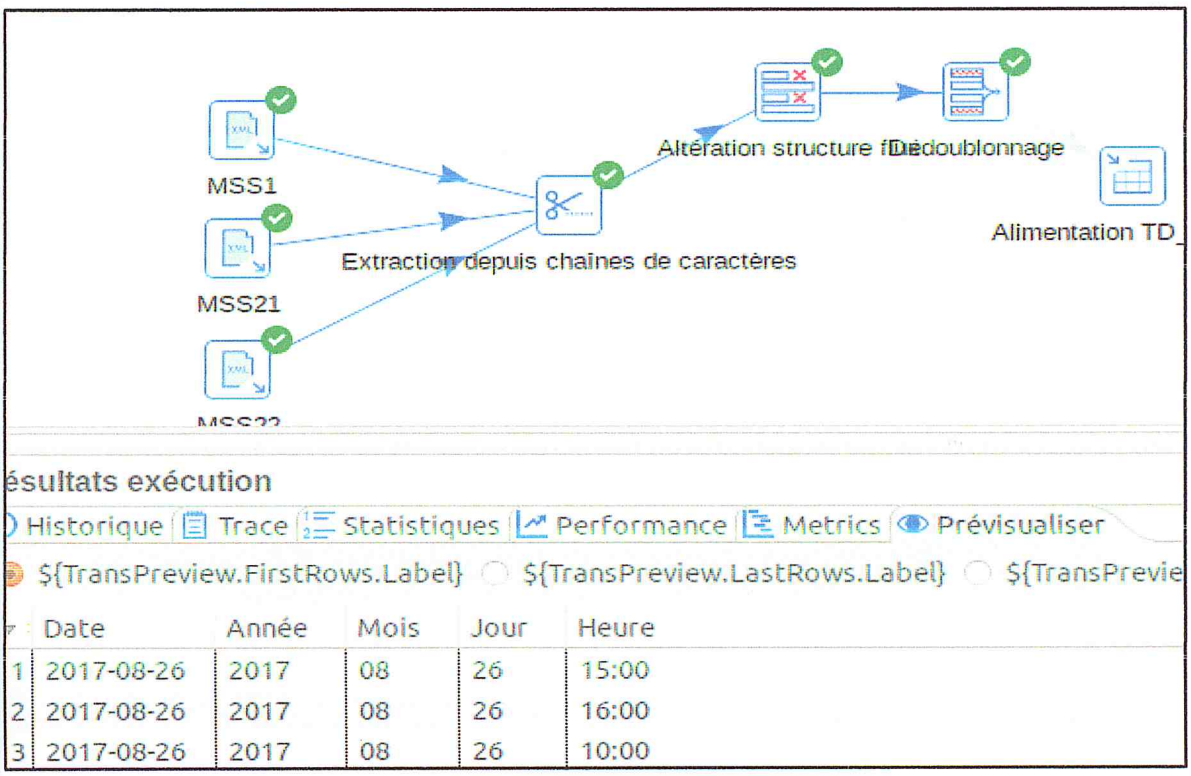
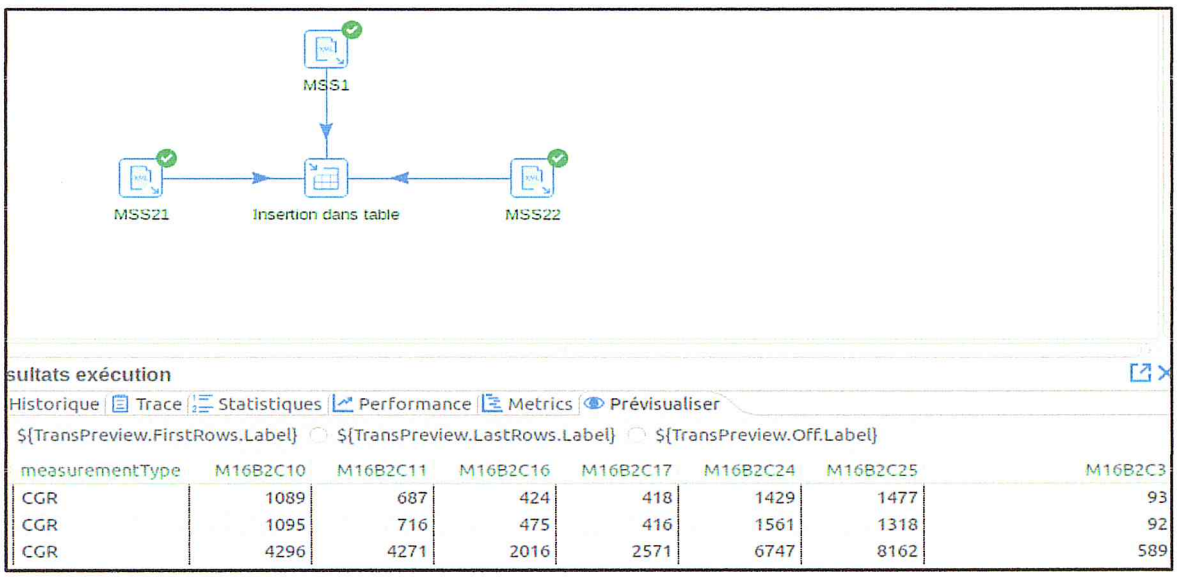


Figure VII.25 :*Exécution de la transformation « Alimentation de la table TD_topologie ».*



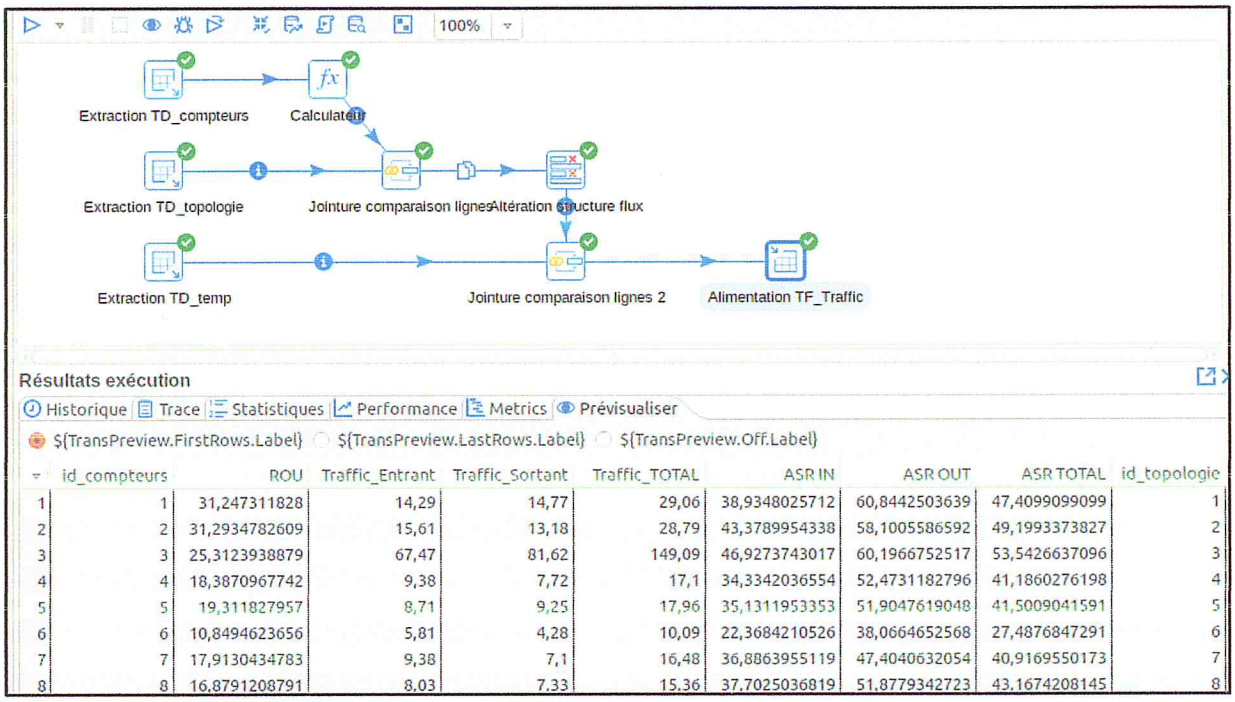
	Date	Année	Mois	Jour	Heure
1	2017-08-26	2017	08	26	15:00
2	2017-08-26	2017	08	26	16:00
3	2017-08-26	2017	08	26	10:00

Figure VII.26 :*Exécution de la transformation « Alimentation de la table TD_temp »*



measurementType	M16B2C10	M16B2C11	M16B2C16	M16B2C17	M16B2C24	M16B2C25	M16B2C3
CGR	1089	687	424	418	1429	1477	93
CGR	1095	716	475	416	1561	1318	92
CGR	4296	4271	2016	2571	6747	8162	589

Figure VII.27 :*Exécution de la transformation « Alimentation de la table TD_compteurs »*



	id_compteurs	ROU	Traffic_Entrant	Traffic_Sortant	Traffic_TOTAL	ASR IN	ASR OUT	ASR TOTAL	id_topologie
1	1	31,247311828	14,29	14,77	29,06	38,9348025712	60,8442503639	47,4099099099	1
2	2	31,2934782609	15,61	13,18	28,79	43,3789954338	58,1005586592	49,1993373827	2
3	3	25,3123938879	67,47	81,62	149,09	46,9273743017	60,1966752517	53,5426637096	3
4	4	18,3870967742	9,38	7,72	17,1	34,3342036554	52,4731182796	41,1860276198	4
5	5	19,311827957	8,71	9,25	17,96	35,1311953353	51,9047619048	41,5009041591	5
6	6	10,8494623656	5,81	4,28	10,09	22,3684210526	38,0664652568	27,4876847291	6
7	7	17,9130434783	9,38	7,1	16,48	36,8863955119	47,4040632054	40,9169550173	7
8	8	16,8791208791	8,03	7,33	15,36	37,7025036819	51,8779342723	43,1674208145	8

Figure VII.28 :*Exécution de la transformation « Alimentation de la table TF_Traffic ».*

Les transformations qui ont été faites devraient s'exécutaient en parallèles pour l'alimentation

de la table de fait, c'est pour cela qu'on a programmé un Job pour le faire il est présenté sur la figure VII.29.

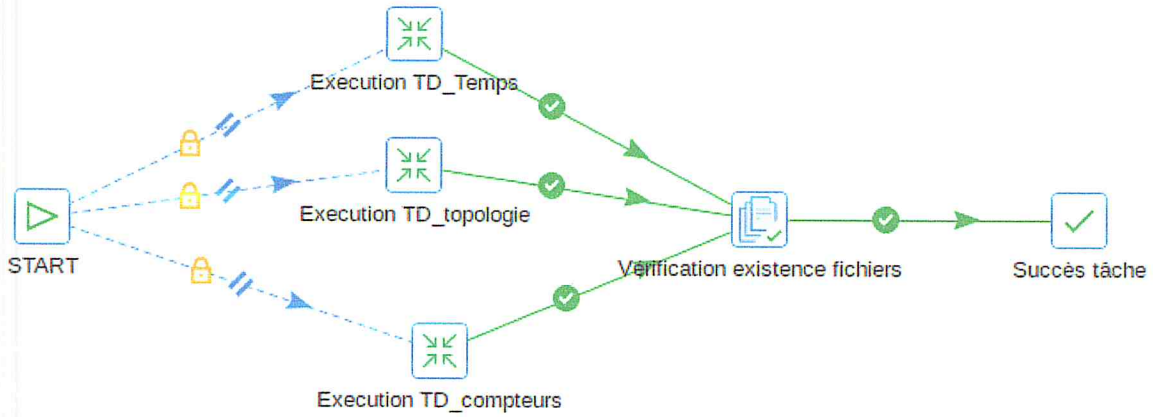


Figure VII.29 : Lancement d'un job pour lancer les transformations en parallèles.

Nous avons construit la hiérarchie de notre cube « CubeTestOOREDOO » avec notre table de fait FACT_traffic et nos dimensions avec les mesures appropriées comme le montrent les figures VII.30, VII.31.

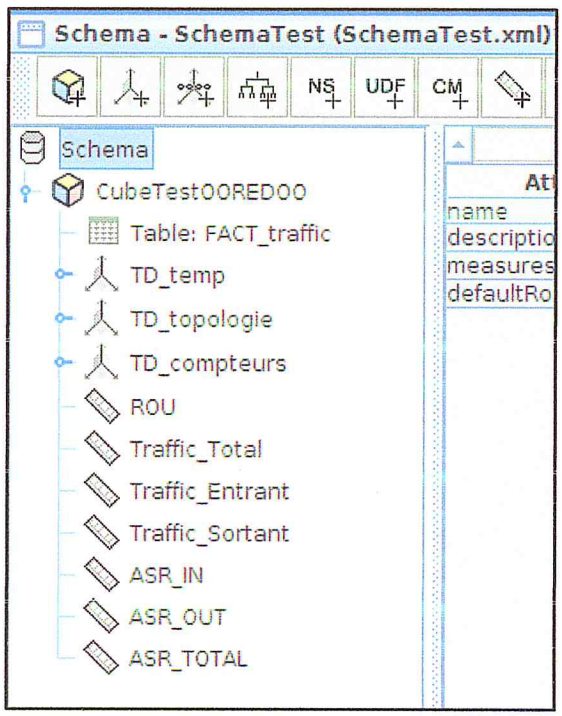


Figure VII.30 :*Schéma du cube crées.*

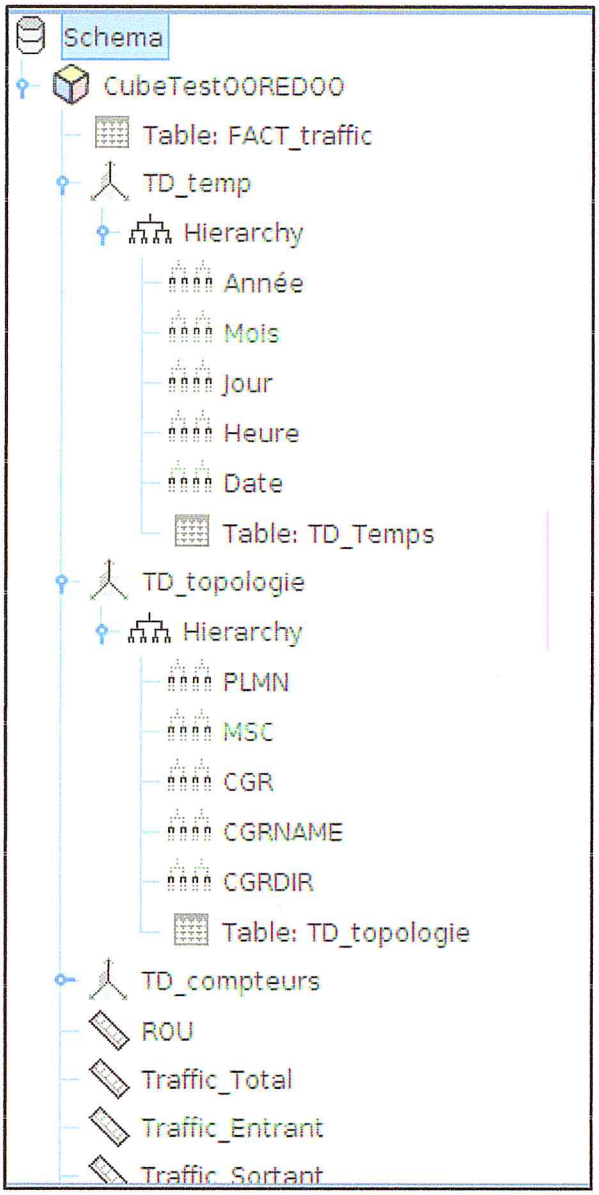


Figure VII.31 :*Schéma des dimensions dans le cube.*

VII.6. Stockage des données dans HDFS avec SQOOP :

Après le traitement, nous allons stocké les données dans notre environnement distribué pour des fins d'analyses, nous avons importé ses donnée de notre base de données mysql vers le système de fichier distribué à l'aide de l'outil Sqoop , comme le montre la figure **VII.32** .

```
hduser@master: ~
hduser@master:~$ sqoop import --connect jdbc:mysql://localhost:3306/ooredoo --table
FACT_traffic --username root -P --target-dir /OOREDOO/FACT  -m 1 --append
```

FigureVII.32 :*Importation des données de Mysql vers HDFS.*

On a importé notre table de fait « TF_Traffic »et la stocké dans le répertoire créé /OOREDOO/FACT dans HDFS, la figure VII.33 montre cela.

```
17/10/02 04:05:18 INFO mapred.LocalJobRunner: map task executor complete.
17/10/02 04:05:19 INFO mapreduce.Job:  map 100% reduce 0%
17/10/02 04:05:19 INFO mapreduce.Job: Job job_local1489869135_0001 completed successfully
17/10/02 04:05:19 INFO mapreduce.Job: Counters: 20
        File System Counters
                FILE: Number of bytes read=17978120
                FILE: Number of bytes written=18422775
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=0
                HDFS: Number of bytes written=411
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
        Map-Reduce Framework
                Map input records=8
                Map output records=8
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=985
                Total committed heap usage (bytes)=226492416
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=411
17/10/02 04:05:19 INFO mapreduce.ImportJobBase: Transferred 411 bytes in 14,3692 seconds (2
17/10/02 04:05:19 INFO mapreduce.ImportJobBase: Retrieved 8 records.
17/10/02 04:05:19 INFO util.AppendUtils: Creating missing output directory - FACT
```

Figure VII.33 :*Job Map/Reduce pour l'importation des données vers Hdfs.*

Et comme la réplication est un processus de partage d'informations pour assurer la cohérence

de données entre plusieurs sources de données redondantes, la figureVII.34 montre la réplication de la table partitionné sur les nœuds esclave «slaves 1» et «slave2» du cluster.

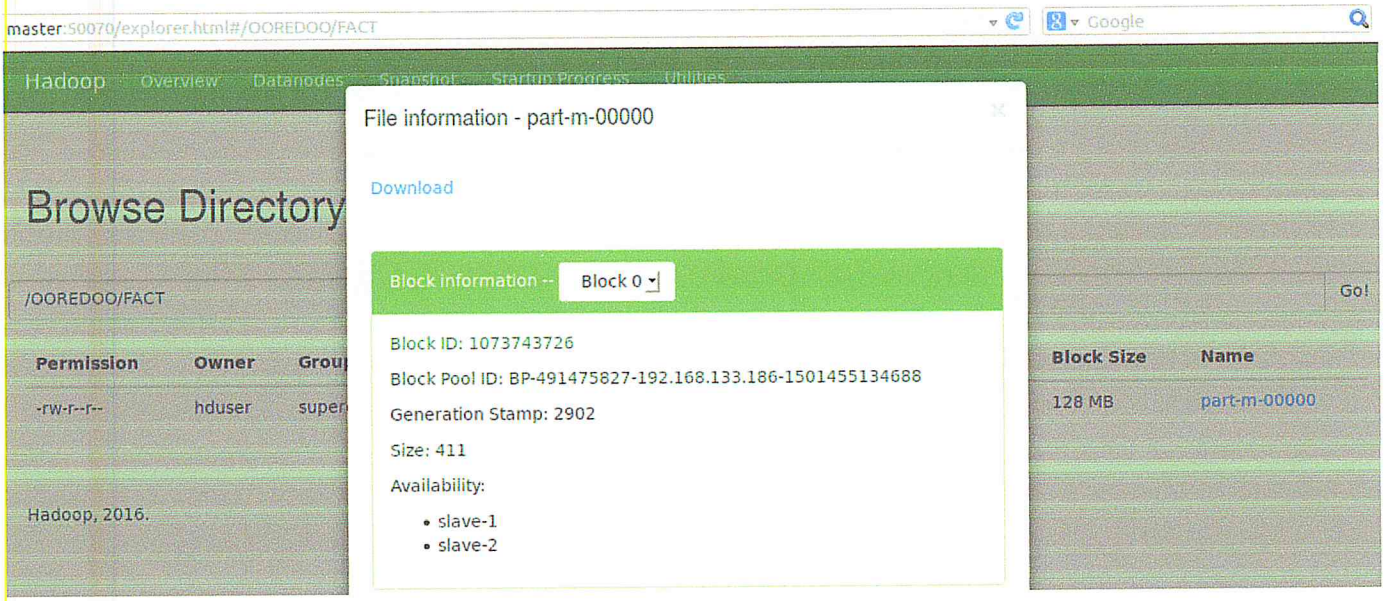


Figure VII.34 :*Réplication des données dans les Slaves (slave-1/slave-2).*

VII.7. Analyse expérimentale :

Afin de mettre en valeur notre étude nous avons établi un rapport d'analyse exprimant les mesures calculées dans le chapitre précèdent. Nous avons utilisé l'outil Pentaho Business Analytics (BA) pour analyser le trafic entrant et sortant de chaque type de mesure présentées sur les figures VII.35, VII.36, VII.37.

MeasurementType - MeasurementType						CGR							
Date	PLMN	MSC	CGR	CGRNAME	CGRDIR	ASR IN	ASR OUT	ASR TOTAL	ROU	Traffic Entrant	Traffic Sortant	Traffic Total	
2017-08-26	PLMN	206322	128	BSOK20	3	37.703	51.878	43.167	16.879	8.03	7.33	15.36	
			129	BSOK21	3	35.131	51.905	41.501	19.312	8.71	9.25	17.96	
			303	BAL100	3	36.886	47.404	40.917	17.913	9.38	7.1	16.48	
		224802	132	BSOK24	3	22.368	38.066	27.488	10.849	5.81	4.28	10.09	
			133	BSOK25	3	38.935	60.844	47.41	31.247	14.29	14.77	29.06	
			328	BBCH24	3	43.379	58.101	49.199	31.293	15.61	13.18	28.79	
		224804	130	BSOK22	3	46.927	60.197	53.543	25.312	67.47	81.62	149.09	
			131	BSOK23	3	34.334	52.473	41.186	18.387	9.38	7.72	17.1	

Figure VII.35 : *Tableau d'analyse final du type de mesure CGR.*

L'ASR devrait être d'au moins 40-50%, et tout ce qui dépasse 60% indiquera un excellent service de qualité, et en comparant les valeurs des ASR du type de compteur CGR on constate que le taux de réponse répondu est dans les normes pour le ASR sortant car il touche les 60% mais ne les dépasse pas par contre il est moyen pour le ASR entrant

Un ASR élevé indique un réseau fiable, car la plupart des appels tentés sont répondu.

Les valeurs ASR faibles peuvent être causées par:

- Comportement de l'utilisateur
- Lignes de destination occupées
- Congestion du commutateur lointain (problèmes de capacité au transporteur)

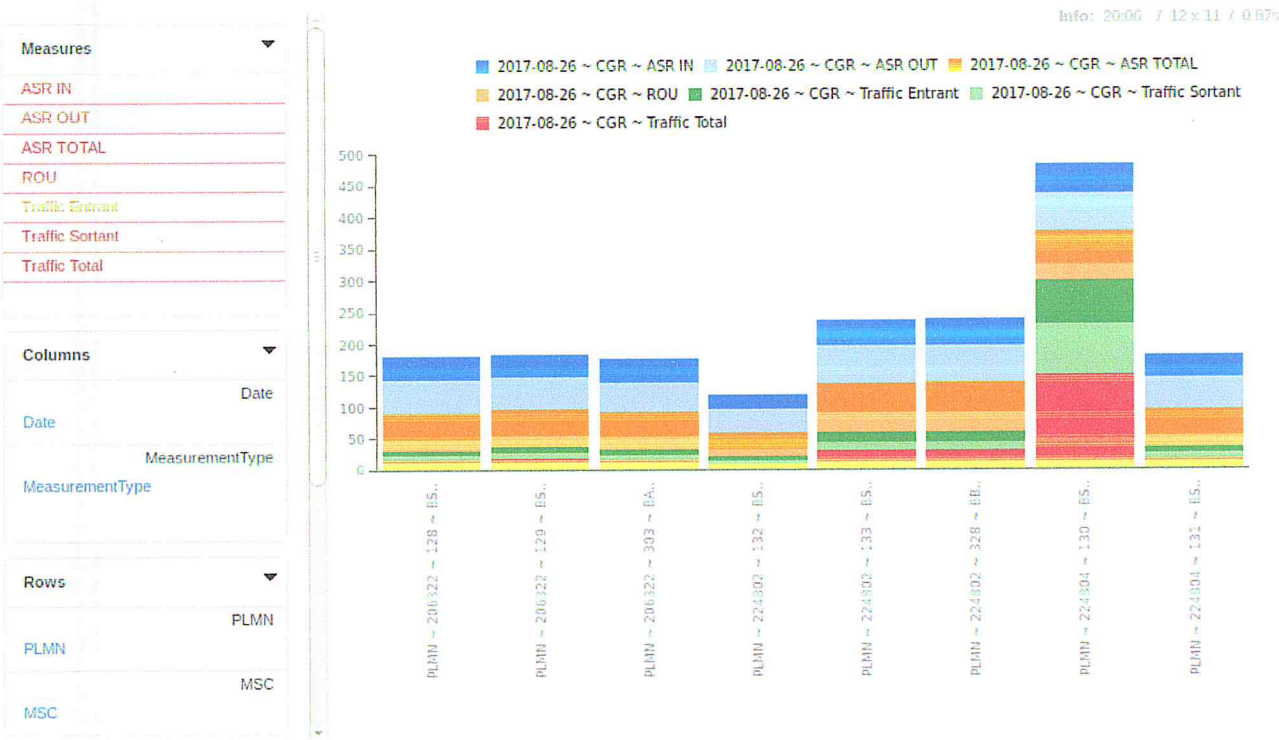
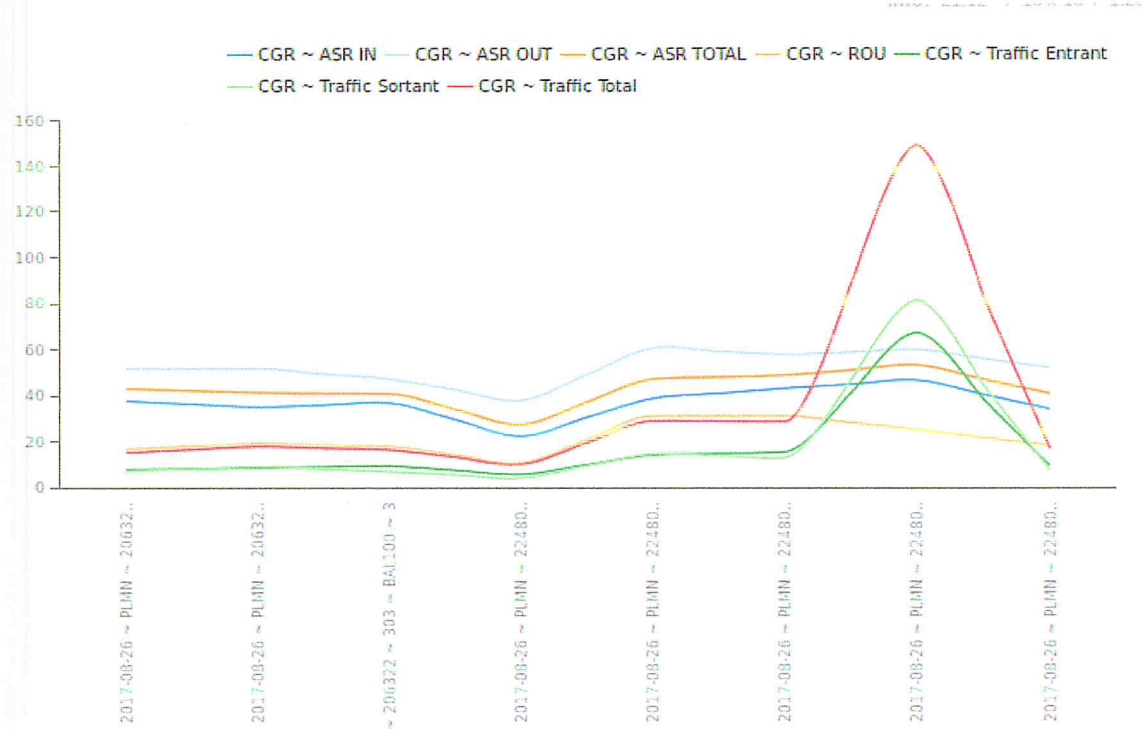


Figure VII.36: *Variation des mesures calculées pour chaque topologie*



Figue VII.37 : *Graphique en courbe représentant la variation des mesures pour chaque topologie.*

VII.8. Conclusion :

Dans ce chapitre, nous avons montré comment mettre en place et configurer un cluster Hadoop dans un environnement totalement distribué, nous avons montré le fonctionnement des composant hadoop (Hive, Sqoop) et leur utilité dans une architecture Big Data à travers des tests de partitionnement et des jobs Map/Reduce effectués pour répondre aux besoins.

À travers ce chapitre, on constate que la structure de données d'un système HDFS avec sa réplication entre les nœuds ainsi que la taille des blocs de données qui est supérieure à un système de fichier traditionnel donnent **plus de flexibilité** pour traiter des données **plus rapidement** d'une façon distribuée et parallèle

Conclusion Générale

L'informatique à la base était le traitement de grande quantité de données, mais cequi a changé avec la venue du Big Data est la formalisation de l'évolution des volumes, de la vitesse et de la variété des données qui crée de la valeur ajoutée.

Le Big Data a déjà marqué de son empreinte le marché IT (Information Technologie) et commence à s'imposer comme un standard en matière de traitement de données volumineuses. Toutefois, dans ce contexte il est difficile de savoir quel fournisseur propose l'implémentation la plus complète d'Apache Hadoop dans un Cloud public.

Le Big Data est avant tout une démarche stratégique, il faut se poser la question : Comment est-ce qu'à partir des données qu'on a ou qu'on peut collecter dans l'entreprise, on peut créer de la valeur ? Autrement dit, qu'est-ce qu'on peut créer à partir de nos données. C'est ça la stratégie, donner de la valeur à vos données en les transformant en valeur ajoutée. C'est grâce à cette stratégie que le Big Data fonctionnera à l'intérieur d'une entreprise.

L'objectif de ce mémoire se divise entre la phase d'intégration des données massives, celle-ci est basée sur un processus d'extraction, de transformation et de chargement (ETL). Particulièrement, dans un environnement distribué, et la phase d'analyse de données.

Nous avons tout d'abord parlé des réseaux GSMet on a vu que leurs développement ne cesse d'accroitre avec l'évolution desgénérations qui ont vu le jour (2G ,3G ,4G , 5G) ,on a aussi définit les caractéristiques d'un réseaux GSM, son architecture et les indicateurs de performance KPI qui assurent la qualité de service .Ensuite on a cerné notre champ d'étude sur le concept des big data en citant ses caractéristiques , son architecture , on a aussi vu ce que c'est le mappage et la réduction avec le paradigme mapreduce en donnant un exemple du processus. Après nous avons expliqué la migration des données vers un environnement distribué en proposant une architecture de travail, et comme ce dernier se situe dans un contexte décisionnel, nous avons exposé les fondamentaux sur les systèmes décisionnels : les entrepôts de données, les environnements OLAP et OLTP, les différents modèles de modélisation à un niveau conceptuel et logique et plus particulièrement, le processus ETL et ses différentes phases :Extraction (E), Transformation (T) et Chargement (L) ainsi en termes de ses fonctionnalités de base pour des besoins d'analyse et de visualisation.

Nous avons réalisé l'objectif attendu qui était la compréhension du nouveau paradigme qui est le Big data et implémenté le plus répondu de ses outils qui est Hadoop. Le plus important et que l'application d'Hadoop nous a permis de nous familiariser avec le concept du

Big data et surtout d'élaborer une architecture de stockage et d'analyse de données scalable et élastique, c'est-à-dire qu'il suffit d'ajouter un serveur Hadoop pour augmenter les performances de stockage et de traitement.

Nous avons pu atteindre nos objectifs préalablement fixés dans ce projet, et pour la vision à long terme dans laquelle des améliorations peuvent avoir lieu comme : Étaler l'ensemble des sources de données sur d'autres nouvelles sources, Sauvegarder l'ensemble des données collectées sur des périphériques de stockages distribués, rédiger des rapports d'analyse avec les technologies de traitement et de stockage.

Perspectives

Pour finir cette formidable expérience, nous pouvons dire que ce travail nous a permis d'acquérir une très bonne expérience professionnelle et d'évoluer dans un domaine intéressant et actuel. Comme un projet dans le cadre de PFE n'est jamais complètement terminé, nous pouvons citer les perspectives suivantes :

- Proposer des architectures d'autres fonctions d'ETL importantes afin de prendre en charge des processus d'ETL consistants et complexes.
- Prendre en charge d'autres formats de données sources telles que les bases de données NoSQL (Not Only SQL), les ERP, les logs, etc.
- Enfin, doter l'application par des interfaces graphiques pour faciliter son utilisation.

Pour conclure, on dirait que ce stage nous a été d'un apport indéniable en matière de connaissances acquises sur les technologies du Big Data. C'est une expérience enrichissante qui nous a permis de comprendre les enjeux d'un projet car nous ne possédons pas de connaissance concernant les technologies du Big Data. A l'issu de notre stage, la montée en compétence était bien perceptible au travers du travail réalisé.

Bibliographie :

[1] : ISOTALO, T. et LEMPIAINEN, J. (2010). Measurements on HSUPA withuplink diversityreception in indoor environment. European Wireless Conference (EW) 2010. IEEE, 523{527.

[2] : GAVRILOVSKA, L. et ATANASOVSKI, V. (2007). Interoperability in future wireless communications systems : A roadmap to 4G. MicrowaveReview, 13, 19{28.

[3] : Emmanuel, T., & Landry, E. (n.d.). PLANNIFICATION ET INGENIEURIE DES RESEAUX DE TELECOMS. UNIVERSITE DE YAOUNDE I, CAMEROUN.

[4] : PIERRE, S. (2007). Reseaux et systemes informatiques mobiles : fondements, architectures et applications. Presses internationales Polytechnique, Montreal.

[5] : ISOTALO, T. et LEMPIAINEN, J. (2010). Measurements on HSUPA withuplink diversityreception in indoor environment. European Wireless Conference (EW) 2010. IEEE, 523{527.

[6] : GAVRILOVSKA, L. et ATANASOVSKI, V. (2007). Interoperability in future wireless communications systems : A roadmap to 4G. MicrowaveReview, 13, 19{28.

[7] :Emmanuel, T., & Landry, E. (n.d.). PLANNIFICATION ET INGENIEURIE DES RESEAUX DE TELECOMS. UNIVERSITE DE YAOUNDE I, CAMEROUN.

[8] : J. Wiley, "Big Data For Dummies", 111 River Street Hoboken Hoboken, New Jersey USA, NJ 07030-5774, 2013.

[9] : M:LessardQu0est sequelebigData : http : ==zeroseconde:blogspot:com=2010=12=que le big **data:html?m = 0.**

[10]: H Shah,N. Swant. "Big data application and architecture".Edition.157,2012

[11] :ShamilHumbetov. *Data-Intensive ComputingwithMap-Reduce and Hadoop*. Rapp.tech. Department of Computer Engineering QafqazUniversity, p. 5.

[12] : Benjamin Renaut. *Hadoop / Big Data*. Rapp. tech. MBDS Université de Nice SOPHIA ANTIPOLIS, 2013/2014.

[13] :LudovicDenoyer et Sylvain Lamprier. *Introduction à MapReduce/Hadoop et Spark*. Rapp. tech. UPMC, p. 36.

[**14**] : J. Lejeune, "une plate-formed'exécutionde programmes Map-Reduce", Coursdel'écoledesMines deNantes.Zû1S.

[**15**] : MDomoulin."lntroductionaux algorithmesMapReduce".Courdel'universitéde france.2014

[16] :c.Soullard .SQLI ENTREPRISE. http :/www.technologies-ebusiness.com/langages/les-bases-no-sql.Consulté le 28/03/2015.

[17] : Kimball, R. (1996). The data warehousetoolkit:Practical techniques for building dimensional data warehouses.(ed) johnwiley& sons. *Inc., New York, NY, USA.*

[18] : Costa, M., & Madeira, H. (2004). Handling big dimensions in distributed data warehousesusing the DWS technique. *Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP,* pp. 31-37.

[19] : Lin, B., Hong, Y., & Lee, Z. (2009). *Data Warehouse Performance.*

[20] : IDG Entreprise. (2015). *An IDG Communcations Company, http://www.idgenterprise.com/,* http://core0.staticworld.net/assets/2015/03/16/2015-data-and-analytics-survey.pdf consulté le 24 janvier 2016.

[21] :Badard, T and E. Dubé. 2009. *EnablingGeospatial Business Intelligence*. Technology Innovation Management Review. Récupéré Mai 2016 depuis http://timreview.ca/node/289

[22]. Kimball, R and J. Caserta. 2011. *The Data WarehouseETLToolkit:Practical Techniques forExtracting, Cleaning, Conforming, and Delivering Data.* New Jersey: John Wiley& Sons. ISBN 0-764-57923-1

[23] : Kimball, R. Reeves, L. Ross, M and W Thornthwaite. 1998. *The Data Warehouse Lifecycle*
Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses. New Jersey: John Wiley& Sons. ISBN 0471255475, 9780471255475

[24] : Codd, E. F. Codd, S. B and C. T. Salley. 1993. *Providing OLAP to User-Analysts: An IT*
Mandate. Computerworld. Récupéré Mai 2016 depuis http://www.minet.unijena. de/dbis/lehre/ss2005/sem_dwh/lit/Cod93.pdf

[25] : Teste, O. 2009. *Modélisation et manipulation des systèmes OLAP : de l'intégration des documents à l'usager*. HDR : Université Paul.

[26] Lin, J., &Ryaboy, D. (2013). Scaling big data mininginfrastructure: The twitter experience. *ACM SIGKDD Explorations Newsletter, 14*(2), 6-19.

[27] : C:Scyphers ; BigData : AnOverview; sur http ://fr:Slideshare.net /cscyphers /bd – 101-slideshare; 2013:

[28] : Midouni, S. A. D. Darmount, J and F. Bentayeb. 2009. *Approche de modélisation multidimensionnelle des données complexes : Application aux données médicales*. Récupéré Mai 2016 depuis https://hal.archives-ouvertes.fr/hal-00411237/document

[29] :Ponniah, P. 2001. *Data WarehousingFundamentals: A Comprehensive Guide for IT Professionals.* The University of Michigan:Wiley. ISBN 0-471-22162-7.

[30] :Grim, Yazid. OLAP, les fondamentaux. www.developpez.com. [En ligne] 07/ 05/ 2008. [Citation : 24/ 03/ 2016.] URL :http://grim.developpez.com/articles/concepts/olap/#LIV-B.

[31] :Intel Corporation, Maîtriser les technologies Big Data pour obtenir des résultats en quasi-temps réel, 2013.

[32] :Adriano Girolamo PIAZZA, NoSQL Etat de l'art et benchmark Travail, 2013.

[33] : Marc Batty(2015), Bigdata Et Machine Learning, Paris : DUNOD, 235 p.