UNIVERSITE KASDI MERBAH OUARGLA

Faculté des Nouvelles Technologies de l'Information et de la Communication

Département d'Informatique et Technologie de l'information

Mémoire de fin d'études

Pour l'obtention du diplôme de Master en Informatique

Domaine : Mathématiques et Informatique

Filière: Informatique

Spécialité : Informatique Fondamentale

Présenté par: **BELGHOUL BADREDDINE**

Thème:

Qualité des données dans un Data warehouse

Soutenu publiquement
le : / /

Devant le jury :

Mme LAALLAM FATIMA ZOHRA	*Président*	*UKM OUARGLA*
Melle KORICHI WASSILA	*Examinateur*	*UKM OUARGLA*
Mme. BENKHEROUROU CHAFIKA	*Encadreur*	*UKM OUARGLA*

Année Universitaire : 2014 /2015

Résumé

Un Entrepôt de données est utilisé pour la prise de décisions dans l'entreprise par le biais de statistiques et de rapports réalisés via des outils de reporting. Son but est de fournir un ensemble de données servant à stocker d'énormes quantités de données, et qui sont mémorisées à partir de différentes sources. Pour cette raison, le problème de la non-qualité des données est posé. Il y aura un déduplication des données, Des données incomplètes, ou des valeurs nulles, ou des données ne sont pas efficaces …etc.

Notre travail consiste à la suppression des données dupliquées. C'est une étape très importante dans le processus d'intégration de données hétérogènes. Nous allons intégrer l'Algorithme de déduplication des données similaires dans l'application "Talend Open Studio", qui est un produit d'intégration de données open source conçu pour combiner, convertir et mettre à jour des données dans divers endroits à travers une entreprise. Grâce à ce travail, nous allons pouvoir trouver une solution au problème de la duplication des données.

Mots clés :

Entrepôt de données, qualité des données, algorithme de déduplication des données, Talend

Abstract

A data warehouse for decision-making in the company through statistics and reports made through reporting tools, its purpose is to provide a set of data providing a single reference, For storing the data, For its ability to absorb and store huge amounts of data, And which are stored from different sources, For this reason, be non-quality data, there will be a data deduplication, Incomplete data, Or null values, Or data are not effective….

Our work for the process of removing duplicate data, And that is a very important step in the process of integration of heterogeneous data. We will integrate deduplication algorithm similar data in the application "Talend Open Studio", which is an open source integration product designed to combine data, convert and update data in various locations across an enterprise. with this process, we can solve the problem of data duplication.

Key words:

Datawarehouse, Data quality, data deduplication , Talend

ملخص

تستعمل مستودع البيانات في اتخاذ القرارات في الشركة، من خلال الإحصائيات والتقارير التي تتم من خلال أدوات إعداد التقارير، والغرض منه هو توفير مجموعة من البيانات وتوفير مرجعية واحدة، لتخزين بياناتها، لقدرته على استيعاب وتخزين كميات هائلة من البيانات، والتي يتم تخزينها من مصادر مختلفة، لهذا السبب، تكون لا جودة للبيانات، وسوف يكون هناك بيانات مكررة، اوغير مكتملة، أو الخالية، أو ناقصة، أو بيانات عير فعالة ...الخ

سنعمل على إزالة البيانات المكررة، وهي خطوة هامة جدا في عملية تكامل البيانات غير المتجانسة.

سنقوم بدمج خوارزمية إلغاء بيانات المتماثلة المكررة في تطبيق "Talend Open Studio"، وهو منتج يعمل على تكامل البيانات وهو مفتوح المصدر يهدف إلى جمع وتحويل وتحديث البيانات في مواقع مختلفة في المؤسسات (شركات). مع هذه العملية، يمكننا حل مشكلة البيانات المتكررة.

الكلمات المفتاحية :

مستودع البيانات، جودة البيانات، حذف البيانات المتكررة، Talend (تالند)

Remerciements

Nous remercions Allah le tout puissant, qui nous a donné la force et la patience pour l'accomplissement de ce travail.

je tiens à exprimer remerciements et ma profonde gratitude à ma encadreur :

Madame Benkhrourou Chafika

pour son encadrement, son suivi et ces conseils tout au long de cette période.

je tiens aussi à remercier

Mme Laallam Fatima Zohra président du jury,

Melle korichi Wassila membre du jury

pour leur précieux temps accordé à la révision de mon mémoire.

Bien entendu, je tiens surtout à remercier

Mes parents

pour leurs sacrifices et leur patience,

tout au long de leurs vies.

Que toute personne ayant œuvré de près ou de loin à la réalisation de ce projet par une quelconque forme de contribution, trouve ici le témoignage de ma plus profonde reconnaissance.

Dédicaces

Grace à Dieu voilà mon travail terminé et il est temps
pour moi de partager ma joie
avec tous ceux qui m'ont soutenu et encouragé.
A travers cette modeste thèse
je tiens à présenter mes sincères dédicaces
à mes parents qui ont consacré leurs vie à notre
éducation et à faire notre bonheur et qui m'encouragent
toujours pour achever mes études tout en espérant de voir
les fruits de leurs sacrifices.
A mes chères sœurs As, Kh, Ah ,
et mes chers frères Zi, Ma
A mes Oncles, Tantes et cousins et alliés de la famille.
A ma fiancée
A l'ensemble des amis que j'ai connu
pendant mes études et à ceux qui m'ont prodigué
leurs vifs conseils, encouragements et témoigné de leur amitié.

TABLE DES MATIERES

Sommaire

List de Figure

List des tableaux

Introduction générale

Introduction Générale

Les entrepôts de données sont des outils conçus pour l'aide à la prise de décision, fréquemment utilisés en lecture, avec pour objectifs principaux de regrouper, d'organiser les informations provenant de sources diverses, les intégrer et les stocker pour donner à l'utilisateur une vue orientée métier, retrouver et analyser l'information avec facilité et rapidité.

Ils contribuent au développement des entreprises grâce au bon fonctionnement des services de marketing et de commercialisation, et plus encore constituant une source appréciable dans le traitement et dans l'analyse des données.

Puisque les données qui alimentent les entrepôts de données proviennent de différentes sources, ceci conduit à de nombreux problèmes parmi lesquels la non-qualité qui entrave la prise de décision.

Le problème de la non-qualité se présente sous diverses formes, par exemple la présence de données en double ou inutiles ou incomplètes ... etc., qui viennent de nombreuses sources.

Dans notre travail, nous allons nous intéresser au problème des données en doubles, et la façon de résoudre ce problème.

Le mémoire est organisé comme suit :

Dans le premier chapitre, nous Parlerons de l'intégration des données: nous allons définir et indiquer les principales caractéristiques de l'intégration. Nous parlerons aussi de l'intégration de données hétérogènes, et des types de l'hétérogénéité des données.

Dans le deuxième chapitre, nous allons présenter la qualité des données et les entrepôts de données. Dans la première partie, nous donnerons un aperçu sur la qualité et nous parlerons sur les critères qui définissent la qualité des données, Les objectifs et les problèmes de la non-qualité des données. Dans la deuxième partie,nous allons présenter les entrepôts de données ainsi que leurs avantages pour les entreprises afin de les exploiter pour la prise de décision.

Dans le troisième chapitre, nous allons présenter notre application. Pour cette raison, nous avons utilisé Talend Open studio et un algorithme de déduplication des données similaires pour résoudre le problème de la non-qualité des données dans les entrepôts.

Chapitre *I*.

L'intégration de donné

Introduction

On assiste depuis quelques années à l'émergence de nouvelles applications qui ont besoin de partager des informations entre différents systèmes. De ce fait, l'interopérabilité entre ces systèmes d'information est complexe puisque les applications doivent être adaptées pour pouvoir combiner les fragments de résultats issus de chaque source en vue de construire le résultat final. Ce processus d'adaptation peut être plus ou moins complexe : exploitation des résultats d'une source pour interroger une autre, élimination des redondances, etc.

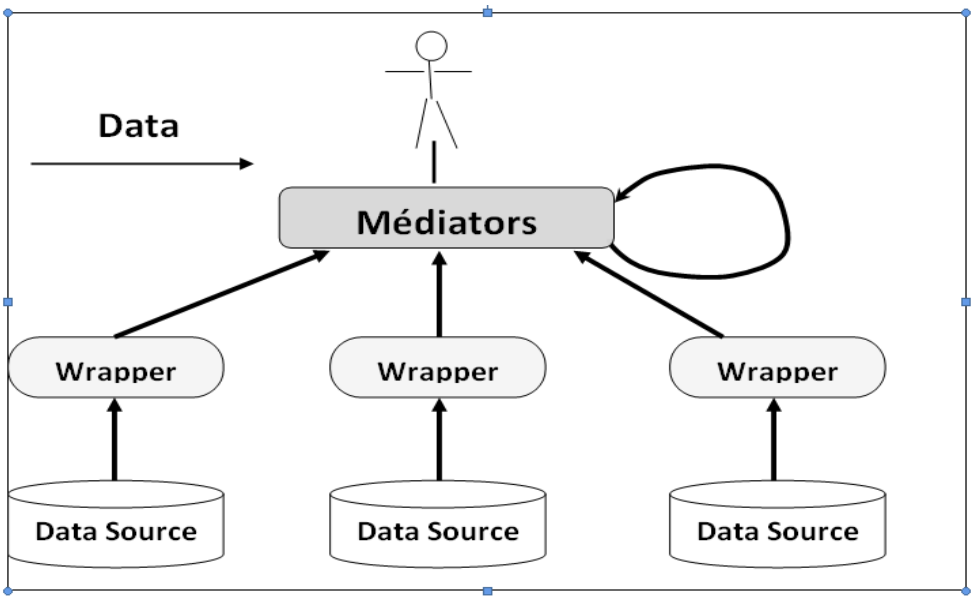
L'intégration des sources de données hétérogènes, autonomes et réparties est une solution pour l'interopérabilité entre différents systèmes d'information, puisqu'elle simplifie l'accès aux données.

Dans ce chapitre, nous allons présenter le domaine de l'intégration des données sur lequel nous allons travailler.

I. Intégration des données

1. Définition

L'intégration des données est le processus de récupération et de combinaison des données hétérogènes provenant de différentes sources pour fournir aux utilisateurs une interface uniforme et transparente sur un ensemble de données.



Figur(1-1): présente un schéma d'intégration ou sources de donnée

2. Principales caractéristique de l'intégration des données

Tous les systèmes d'intégration des données sont caractérisés par :

2.1.Distribution des données :

La distribution des données concerne des données stockées sur des supports répartis sur plusieurs sites.

L'avantage de la distribution des données est leur disponibilité car les bases de données reparties ne tombent pas en panne en même temps ainsi que le temps d'accès est amélioré (partage de la charge parallélisme)

2.2.Hétérogénéité des sources :

L'hétérogénéité dans les sources concerne les données, les modèles, les langages, etc.

Les Système homogènes utilisent :

- ✓ même logiciel gérant les données sur tous les sites.
- ✓ même modèle de données/ langage d'accès.
- ✓ même univers de discours/ sémantique.

II. L'intégration de données Hétérogènes

1. Problématique de l'intégration de données

Du fait du développement important de l'Internet, la recherche d'informations issues des sources de données réparties sur le réseau devient de plus en plus difficile. En effet, grâce à la révolution de nouvelles technologies de l'information, les entreprises aussi bien que les individus disposent d'une grande quantité de données. Ces données sont stockées dans des sources hétérogènes et autonomes.

- La *localisation* d'une source de donnée englobe tout aussi bien la référence du site sur lequel se situe la source (URI, adresse IP + port, annuaire **LDAP**), que le protocole de communication utilisé (TCP/IP,IPX, Appletalk), les moyens d'accès à la base (OOBC, JOBC) ainsi que le support (pages Web, SGBO).
- Le type de données géré par une source peut être structuré (base de données relationnelles), semi structuré (sources XML, OEm) ou non structuré (images, multimédia, texte libre).

- Les possibilités d'interrogation sont aussi nombreuses, et vont des langages de requêtes évolués et standardisés (SQL, OQL) ou propriétaires (Lorel) à de simples interfaces de programmation ou encore des recherches par motifs (moteur de recherche Web). Nous ferons abstraction des interfaces graphiques de requêtes utilisateurs, trop spécifiques et inadaptables dans le cadre d'une intégration.
- Enfin, les formats des résultats comblèrent les disparités qui peuvent exister les différentes sources de données. Celles-ci peuvent être formatées suivant divers modèles standards (XML, HTML, relationnel) ...[01]

2. Systèmes d'Intégration de données

Les Systèmes d'Intégration de données offrent des architectures d'interopérabilité sur une fédération de sources distribuées, autonomes et hétérogènes. Les entrepôts de données, les systèmes de médiation et les architectures P2P sont des exemples d'infrastructures permettant l'intégration de données c'est-à-dire l'accès à des données produites par des sources autonomes. A travers des schémas virtuels, des métadonnées et des correspondances sémantiques, ils permettent d'accèdes à ces sources de données de façon uniforme et transparente, en transformant par réécriture les requêtes d'un utilisateur en sous requêtes envoyées aux sources de données les plus appropriées. Leur réconciliation, en d'autres termes, leur mise en correspondance par rapport au schéma global, avant de les présenter à l'utilisateur.

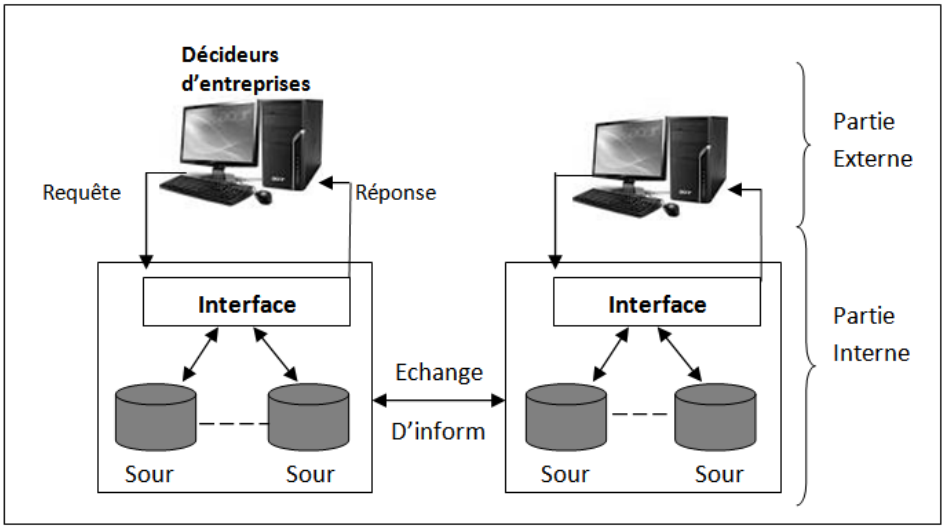
3. Définition & Composante

Un système d'intégration de données fournit une vue unifiées de données provenant de sources multiples et hétérogènes. Il permet d'accéder à ces données à travers une interface uniforme, sans se soucier de leur structure ni de leur localisation.[02]

Formellement, un système d'intégration de données est un tripléI:<G, S, M>, où: G représente le schéma global (définis sur un alphabet AG) modélisant le schéma intégré, S est l'ensemble des schémas des sources (définis sur un alphabet AS) décrivant la structure des sources participantes au processus d'intégration, M est une correspondance entre G et S qui établit la connexion entre les éléments du schéma global et ceux des sources.

Un système d'Intégration se compose de deux parties :

- Une partie (1) externe correspond aux utilisateurs du système intégré ou autres systèmes.
- Une partie (2) interne et comprend des sources d'informations et une interface uniforme qui permet à la partie externe d'interroger d'une manière transparente les sources de données, comme s'il n'y avait qu'une seule source.[03]



Figur(1-2): Système d'intégration d'information.

III. Types de l'hétérogénéité des données :

L'hétérogénéité des sources de données est généralement classée en deux types :

1. L'hétérogénéité des schémas :

La présence des variations et des conflits est inévitable puisque les humains pensent différemment et que les sources de données sont conçues pour des besoins applicatifs distincts. Ce type d'hétérogénéité peut se produire, lorsque des modèles de données différents sont utilisés pour décrire les données, mais aussi lorsque des schémas décrits dans un même modèle, peuvent être issues de conflits de noms (termes), conflits des structures, ou de conflits de classification.

2. L'hétérogénéité des données :

Ce type d'hétérogénéité est également inéluctable lorsqu'on souhaite intégrer des données provenant de différentes sources. Ce type d'hétérogénéité apparait lorsque différents vocabulaires et référentiels sont utilisés pour représenter les données. Certains attributs ne sont pas renseignés, ou certaines données contiennent des erreurs.[04]

3. Les étapes du processus de l'intégration des données

Etant donné un ensemble de sources de données hétérogènes {S1, S2, …,Sn}, le problème d'intégration consiste à construire un schéma intégré (ou schéma global) qui sera utilisé comme interface d'accès aux sources de données. La construction du schéma global à partir des schémas locaux est une tâche difficile. Cette difficulté est liée au fait que les sources stockent différentes sortes de données, en différents formats, ayant différentes significations et associées aux différents noms [05].

Il convient d'abord d'indiquer qu'il existe plusieurs méthodologies permettant l'intégration de bases de données classiques.

Le processus d'intégration est ainsi décomposé en trois phases distinctes :

3.1.Pré-intégration

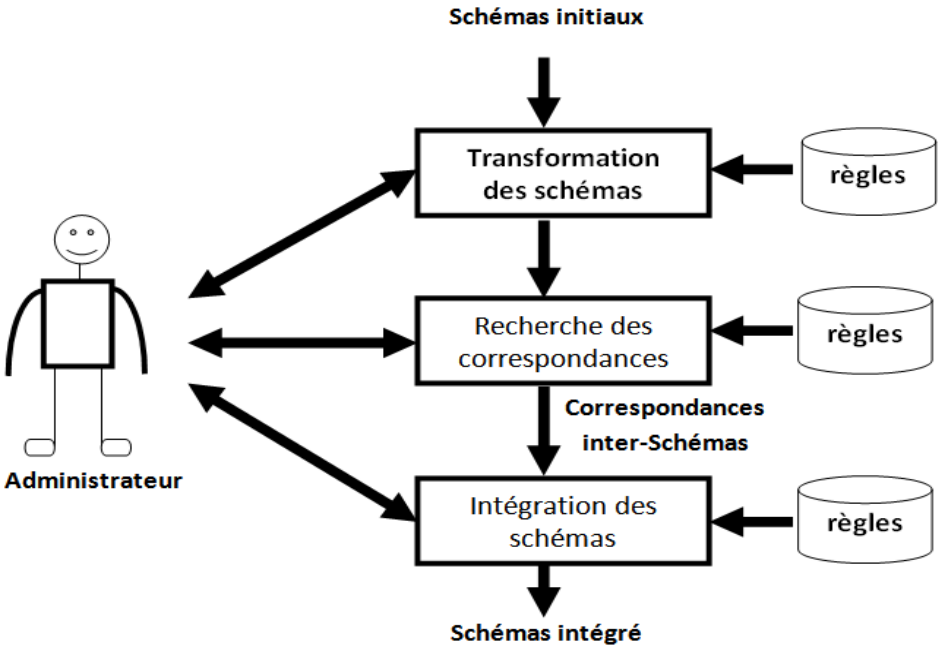
Cette étape consiste à transformer les schémas des bases de données existantes de différentes manières pour les rendre plus homogènes (sur les plans sémantique et syntaxique) par la définition des équivalences entre les domaines et les dépendances entre les schémas, et la correction des conflits logique (désignation, structure …etc.) et physiques (langage de requête et la restitution des résultats.[06]

3.2.Recherche des correspondances

Une étape consacrée à l'identification des éléments semblables dans les schémas initiaux et à la description précise des liens inter-schémas. Les correspondances doivent être définies à deux niveaux des données (population de la base de données) et au niveau du schéma de la base de données [06] .

3.3.Intégration

Une fois que les correspondances ont été décrites, l'intégration, proprement dite, peut commencer. L'intégration est l'étape finale qui unifie les types en correspondance en un schéma intégré et produit les règles de traduction associées entre le schéma intégré et les schémas initiaux. Ces règles seront utilisées pour transformer chaque requête définie sur le schéma intégré en ensemble de requêtes locales pour récupérer toutes les informations locales [06].



Figur(1-3): le processus global d'intégration. (Schéma intégré + Mappings)

4. Classification des approches d'intégration

Plusieurs approches et systèmes d'intégration ont été proposés, nous pouvons les distinguer selon :

4.1.Architectures d'intégration des données

Nous distinguons deux types de classes selon la représentation des données intégrées :

a) L'intégration matérialisée

Des supports spécifiques (entrepôts des données) stockent les informations provenant des sources, et l'intégration s'effectue directement sur ces entrepôts, ce qui produit une performance élevée mais exige de mémoires et de maintenance supplémentaires.

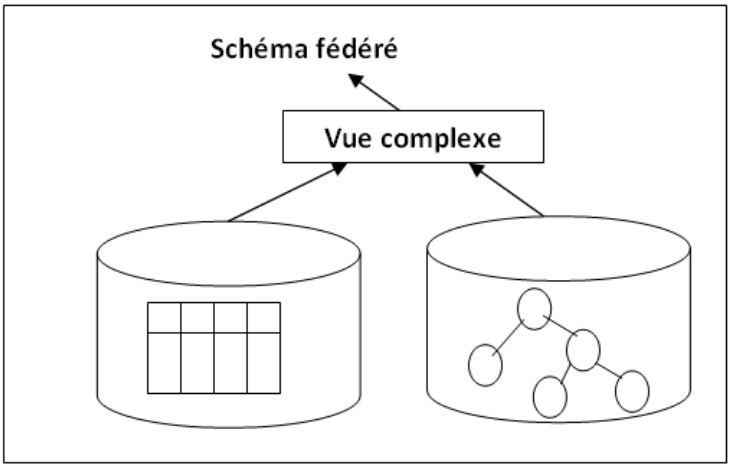
b) Schéma d'intégration virtuelle

Il s'appelle aussi le « mapping », qui représenta le lien entre le schéma global et les schémas locaux. Nous pouvons classifier les systèmes d'intégration de données en deux classes :

- **L'approche Global as View (GaV)**

L'approche GaV ou l'approche ascendante a été la première à être proposée pour intégrer des informations. Elle consiste à fournir un schéma global en fonction des schémas locaux par la

définition à la main (ou de façon semi-automatique) du schéma global en fonction des schémas des sources de données à intégrer puis à le connecter aux différentes sources.

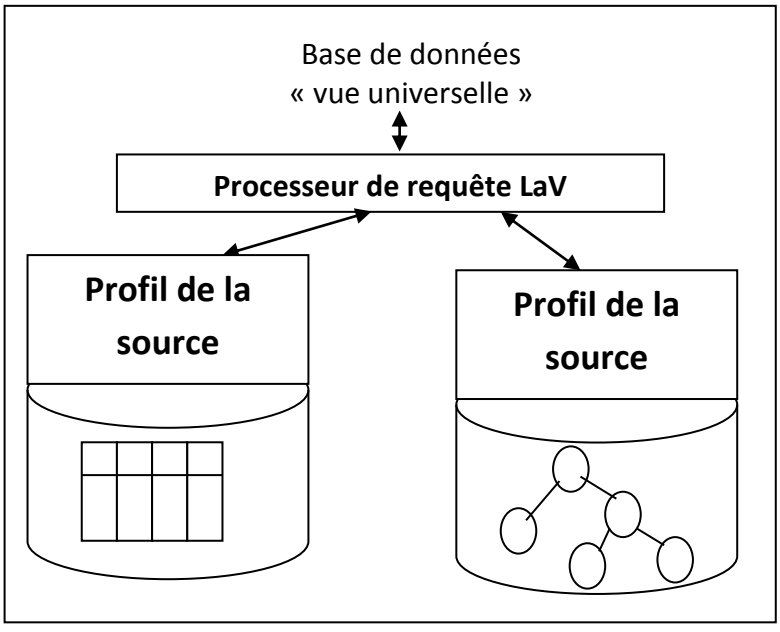


Figur(1-4): L'approche GaV

L'avantage de l'approche GaV est la simplicité de la traduction des requêtes venant des utilisateurs.

- **Approche Local as View (LaV)**

L'approche LaV est l'approche descendant, elle suppose l'existence d'un schéma global et elle consiste à décrire les schémas locaux selon le schéma global fixe. Donc les schémas locaux sont exprimés en fonction du schéma global, et cela par la définition des schémas des sources de données à intégrer comme des vues du schéma global.



Figur(1-5): représente l'approche LaV

L'un des points positifs de cette approche par rapports à l'approche GaV est la flexibilité. Mais cette flexibilité créerait une complexité dans la réécriture de la requête de l'utilisateur pour interroger les schémas locaux.

4.2. Méthodes d'intégration des données :

Il existe plusieurs choix pour diriger le processus d'intégration selon plusieurs critères :

a) Automaticité d'intégration

Un critère important qui permet de spécifier le degré d'automaticité de génération du système intégré s'il est manuel. Ce critère devient essentiel lorsque l'on veut intégrer un nombre important de sources de données indépendantes.

b) Manuelle :

Les méthodes d'intégration manuelles de données correspondent à la première génération de systèmes d'intégration. La structure des schémas locaux et ses correspondances avec le schéma global sont conçu manuellement, aucun traitement automatique ne peut être effectué en raison de l'explicité des concepts utilisés, tant au niveau global qu'au niveau d'une nouvelle source. [07]

c) Semi-automatique :

La deuxième génération des systèmes d'intégration utilisent des thésaurus (relations linguistiques telles que la synonymie, l'antinomie et l'hyperonymie). Ils donnent la possibilité d'un premier niveau d'automatisation par l'identification de quelques relations sémantiques entre les termes utilisées les sources de données.

Une telle intégration est qualifiée de semi-automatique lorsque le domaine visé par l'intégration est suffisamment limité et formalisé. [08]

d) Automatique

Dans les deux types de mapping précédents, il suffit que la nouvelle source soit explicitement exprimée en fonction de l'ontologie globale pour que l'intégration automatique soit possible. [08]

5. Objectif de l'intégration :

Plus particulièrement, l'intégration de données doit fournir [11] :

- ✓ Un accès (requêtes, éventuellement mis-à-jour)
- ✓ Uniforme (comme si c'était une seule BD homogène)
- ✓ A des sources (pas seulement des BD)

- ✓ Multiples (déjà deux est un problème)
- ✓ Automnes (sans affecter leur comportement, indépendant des autres ou de système d'intégration)
- ✓ Hétérogènes (différents modèles de données, schémas)
- ✓ Structurées (ou semi-structurées)

Conclusion

Le présent chapitre a donc pour but de présenter la problématique de l'intégration de données, à savoir l'hétérogénéité de données. Cette hétérogénéité provient de choix différents que sont faits pour représenter des faits du monde réel dans un format informatique. Nous avons aussi exposé les différentes architectures des systèmes d'intégration de données et les étapes du processus d'intégration.

Dans le chapitre nous présenterons la qualité des données dans un entrepôt de données.

Chapitre *II.*

La Qualité De

Introduction

La qualité des données peut faire la différence entre le succès et l'échec d'une entreprise. La non-gestion des données de mauvaise qualité accroît leur détérioration et nuit à la flexibilité de l'entreprise. De plus, dans les secteurs réglementés, les données de mauvaise qualité peuvent entraîner des violations de politiques de sécurité entrainant des sanctions.

Dans ce chapitre, nous allons présenter les entrepôts de données. Nous nous intéressons à la qualité des données dans les entrepôts de données et nous citerons les différents outils qui permettent de l'améliorer.

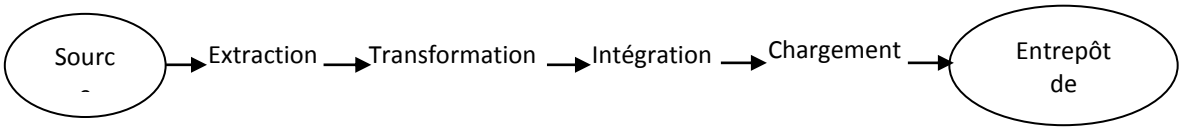
I. Entrepôt de données :

1. Présentation

Pour répondre à la demande excessive des informations, les entreprises se sont investies dans la construction d'un nouveau concept de bases de données séparées des systèmes de production, pour le stockage et la consultation des données. La firme IBM a nommé ce nouveau concept de base de données séparée « le dépôt des informations ». Plus tard, Ce concept a été baptisé par Bill Inmon « data warehouse » en 1990. [20]

Pour faire face aux nouveaux enjeux, l'entreprise doit collecter, traiter, analyser les informations de son environnement pour anticiper. Mais cette information produite par l'entreprise est surabondante, non organisée et éparpillée dans de multiples systèmes opérationnels hétérogènes et peut provenir de toutes les sources.

Il devient fondamental de rassembler et d'homogénéiser les données afin de permettre l'analyse des indicateurs pertinents pour faciliter la prise de décisions. L'objet de l'entrepôt de données est de définir et d'intégrer une architecture qui sert de fondation aux applications décisionnelles. [21]

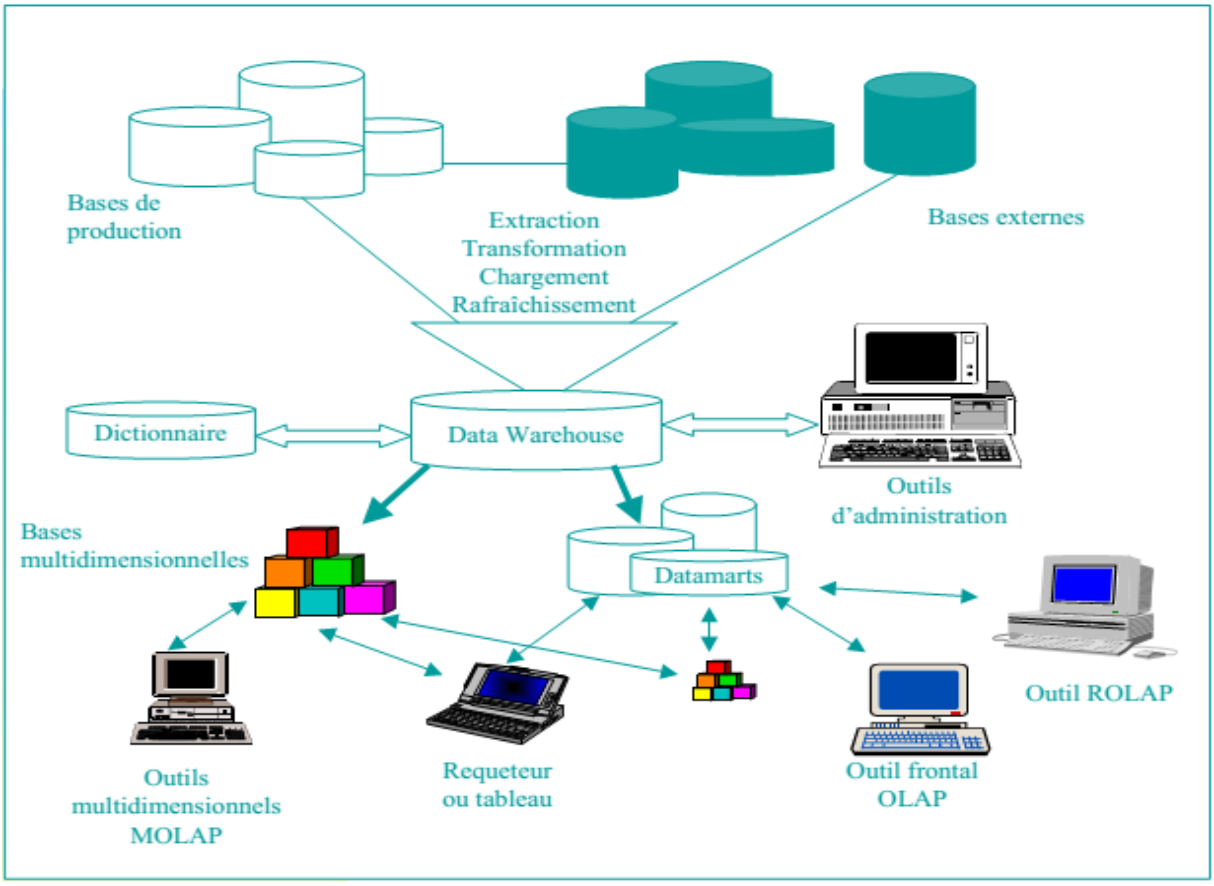


2. Un entrepôt de données (DataWarehouse):

Un entrepôt de données est une collection de données thématiques, orientées sujet, intégrées, non volatiles et historiées pour la prise de décisions (Bill Inmon) [22].

C'est un moyen utilisé par plusieurs entreprises pour avoir et maintenir un avantage compétitif. En rassemblant les informations des sources internes et externes, l'entrepôt moderne est une collection d'informations stockées d'une manière qui améliore l'accès aux données. Une fois les informations rassemblées, quelques types de logiciel de recherche seront utilisés sur les PC pour retirer les données de l'entrepôt où elles sont analysées, manipulées et reportées. [22]

L'infrastructure technique mise en œuvre est capable d'intégrer, d'organiser, de stocker et de coordonner de manière intelligible des données produites au sein du Système d'Information (issues des applications de production) ou importées depuis l'extérieur du SI dans lesquelles les utilisateurs finaux puisent des informations pertinentes à l'aide d'outils de restitution et d'analyse (OLAP "On-Line AnalyticalProcessing", Datamining).[22]



Figur(2-1): Architecture d'un Data warehouse

Les points clefs garantissant le succès d'un entrepôt de données sont les suivants :

- Les informations d'un entrepôt de données doivent être accessibles et fiables (de qualité).
- La conception d'un entrepôt de données doit répondre à un besoin de ROI (Return On Investment) élevé.
- La réponse aux demandes très diverses des utilisateurs.
- L'entrepôt de données doit évoluer avec les besoins des utilisateurs et du système d'information

Les systèmes sources de l'entrepôt peuvent contenir :

- Les gros et mini systèmes
- Les fichiers de données et bases de données client/serveur
- Les feuilles de calculs
- Les données obtenues des sources externes.
- Les données obtenues des PC personnels (sources internes)

Le Middleware est un ensemble d'outils utilisé pour alimenter l'entrepôt, il est responsable de l'entretien et de la transformation des données du système source. Ces instruments peuvent être désignés et codés à la compagnie ou achetés dans les magasins. Le nettoyage risque de prendre du temps et de coûter chère à l'entreprise si les données dans le système source ne sont pas bien maintenues. Le processus de transformation des données varie selon la complexité d'un système à l'autre et peut inclure :

- La translation de champs (du PC aux formats gros systèmes).
- Le changement de formatage des données.
- Le reformatage de champs.
- Le reformatage des structures de données (réordonner les colonnes des tableaux)
- L'agrégation des données de niveau bas dans un résumé d'informations.

Contrairement aux opérations de nettoyage, qui sont faites une fois seulement lors du premier chargement de l'entrepôt, la transformation est un processus continuel. Il est le processus continuel le plus difficile et le plus cher requis pour maintenir l'entrepôt. [23]

3. Les avantages entrepôt de données

Le Data Warehouse (entrepôt de données) offre de multiples avantages aux entreprises qui décident de l'utiliser pour exploiter leur data. En effet, les entrepôts de données permettent de résoudre des problèmes en établissant des liens entre des informations contradictoires, par exemple, ou en pointant des endroits qui nécessitent des corrections ou améliorations. On peut entre autres se servir de ces outils informatiques pour détecter les fraudes en croisant plusieurs données.

L'entrepôt de données permet en général aux gestionnaires de gagner du temps et de l'argent en produisant des analyses et des rapports statistiques rapides et précis, d'avoir accès aux données rapidement et d'effectuer des prévisions justes et réalistes des ventes ou du prochain chiffre d'affaires.

II. Définition de la qualité

Dans une démarche de qualité, il est important de définir clairement les caractéristiques attendues ainsi que les critères d'évaluation de la qualité des données. Il est ensuite plus facile de mettre en œuvre les mesures de suivi et les plans d'actions de correction. [12]

Une donnée est une description élémentaire, souvent codée, d'une chose, d'une transaction d'affaire, d'un événement, etc. Les données peuvent être conservées et classées sous différentes formes: papier, numérique, alphabétique, images, sons, etc.

L'information représente les données transformées sous une forme significative pour la personne qui les reçoit: elle a une valeur pour ses décisions et ses actions bien que la définition de la connaissance fasse encore débat parmi les philosophes, dans le monde de l'entreprise c'est le traitement des données et des informations qui permet de générer des connaissances: un moyen de compréhension ou d'apprentissage d'un problème ou d'une activité. [15,14]

L'idée générale est de gérer les données comme un actif de l'entreprise au même titre que ses produits, ses employés, ses clients. Il faut donc comprendre les besoins des clients [13]

La qualité est une préoccupation que l'on trouve dans beaucoup de domaine. De ce fait, la première difficulté réside dans l'absence de consensus sur la notion de qualité. Comme la communauté aujourd'hui préconise également l'application dès le début des normes et standards internationaux, nous nous intéressons ici aux définitions données par l'organisation

internationale de standardisation (ISO: International Standard Organization) et par Organisation de Coopération et de Développement Economiques (OECD: Organisation for Economic Cooperation and Development). [15]

La Définition de ISO:

La norme ISO définit la qualité comme «L'ensemble des propriétés et caractéristiques d'un produit ou d'un service qui lui confère l'aptitude à satisfaire des besoins exprimés ou implicites». Pratiquement, la qualité d'un produit signifie qu'il est adapté au besoin qu'il est censé satisfaire. La notion de qualité s'applique aussi bien à des produits qu'à des services. [15,16]

La Définition de OECD:

La qualité est vue comme un concept à facettes multiples. Les caractéristiques de qualité dépendent des perspectives, des besoins et des priorités d'utilisateur, qui changent à travers des groupes d'utilisateurs. Ainsi cette définition est complémentaire à la définition ISO en y ajoutant le contexte d'utilisation et le domaine de l'application c.à.d. que les besoins sont définis par l'utilisateur dans le cadre d'une application donnée. [15]

III. QUALITÉ DES DONNÉES

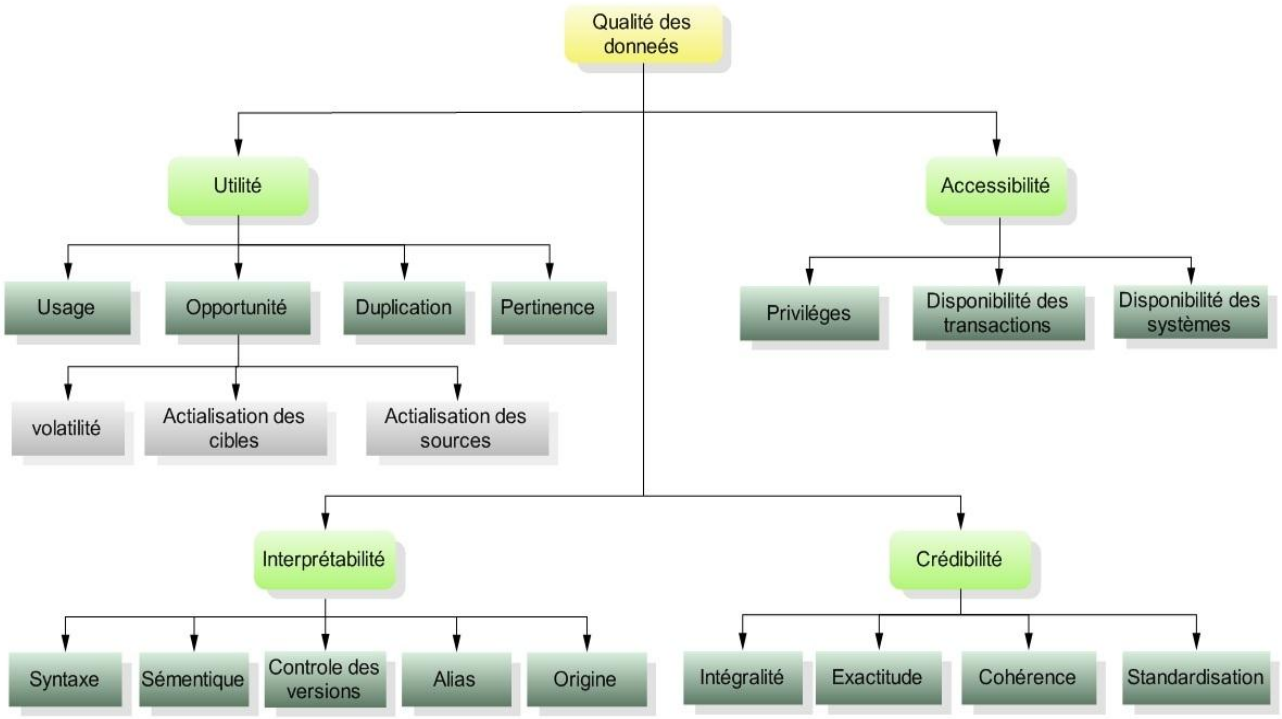
1. Principaux concepts

Les données jouent un rôle croissant dans les processus opérationnels et dans les choix stratégiques des entreprises. Les entreprises, qui ont souvent fortement amélioré leurs processus, doivent maintenant agir directement sur les données sous-jacentes qui les structurent pour espérer de nouveaux gains. [Site 1]

De même pour le pilotage, la qualité des résultats fournis par les systèmes décisionnels dépend directement de la qualité des données qui leur sont fournies. Les traitements en « redressement qualitatif » au niveau des systèmes décisionnels sont peu efficaces. Une amélioration n'est réellement envisageable qu'en ayant une meilleure maîtrise de la chaîne de l'information, c'est-à-dire en se rapprochant du point où l'information est créée ou garantie [17]

2. Définition Qualité Des Données

Si nous ne mettons pas en place aucune gestion de la qualité des données, le système pourra rapidement être saturé de données manquantes, superflues voir incorrectes. Le problème de la qualité des données se pose avec d'autant plus de difficulté que les volumes à traiter augmentent et que les applications tendent à se diversifier. Outre cela, les pressions réglementaires et les exigences de contrôle interne obligent les entreprises à s'intéresser de plus en plus à la qualité de leurs données. [15]



Figur(2-2): Les dimensions de la qualité des données

1. Les critères de la qualité des données

1.1. Relativité de la qualité des données

Pour faire une offre promotionnelle à un client à l'occasion de son anniversaire, il est indispensable de connaître sa date de naissance alors que pour mener une étude marketing sa tranche d'âge suffira.

Imaginons en effet un référentiel client contenant les âges de n clients. Supposons que les âges soient connus à un an près. Cet ensemble référentiel client ne pourra pas servir à souhaiter à chaque client son anniversaire à la bonne date. Pour cet usage, on peut considérer

que les données sont de très mauvaise qualité. Par contre, les données sont suffisantes pour une étude marketing.

Cet exemple illustre pourquoi des données jugées de mauvaise qualité par les personnels en charge de la relation client peuvent être jugées de bonne qualité pour effectuer des études marketing.

On notera également l'intérêt de l'établissement de métadonnées par le système d'information décrivant la précision de celles-ci, afin d'aider les utilisateurs à pouvoir donner eux-mêmes la précision du résultat de leurs études. [17]

1.2.Les critères intrinsèques

a) L'unicité

L'unicité des données est un aspect de la qualité des données qui désigne le résultat des processus visant à résoudre et à éviter les problèmes de duplication indésirable des données.

Un objectif phare de la gestion des données de référence est d'obtenir une vue unique des clients.

L'unicité des données sert aussi à n'avoir qu'une seule description d'un produit donné. Elle contribue alors à l'amélioration de la qualité des données produit. [site2][17]

b) L'exactitude

Une donnée est « exacte » si la valeur des attributs de l'entité concernée est égale à la grandeur qu'elle est censée représenter dans le monde réel. Cette notion englobe donc deux aspects : la précision et la validité. [17]

c) La complétude

Ces dernières années, l'apparition de normes de mise en conformité des données informatiques a bouleversé l'industrie informatique. [17]

Les normes réglementaires affectent des domaines aussi larges que le respect de la vie privée, la sécurité, la rétention, la protection et la responsabilité des données. Des systèmes de vérification sont mis en places pour préserver l'information et les données. Des procédures d'investigation permettent de vérifier l'intégrité de la confidentialité, sécurité et protection des données. [Site 3] [17]

Des mesures légales et commerciales protègent les sociétés en cas d'investigation, mais elles garantissent également la protection des données des consommateurs et des patients. Voici une liste des lois de mise en conformité les plus communes. [Site 3] [17]

d) La conformité

La gestion d'une base de données est réglementée par des contraintes juridiques : loi Informatique et Libertés, LCEN (Loi pour la Confiance dans l'Economie Numérique), etc.

La loi Informatique et Libertés s'applique sur différents domaines :

- fichiers clients et prospects
- phoning
- envoi de mailing, d'e-mailing
- information personnelle autorisée ou non
- etc.

Chaque entreprise doit réaliser une procédure légale afin d'assurer la conformité de ses bases de données. [17][site4]

La conformité aux réglementations s'avère coûteuse, complexe et en constante évolution. Cependant, une telle conformité ne suffit pas à protéger vos données confidentielles. À cela s'ajoutent les menaces ciblées, les environnements distribués, la mobilité des employés et le progrès des technologies, y compris le cloud computing et la consumérisation des ressources informatiques.

De nombreux défis à relever sont communs à toutes les entreprises. Au niveau le plus basique, la confidentialité des données et de la stratégie de conformité doivent s'appuyer sur tout un ensemble de contrôles de conformité, y compris [17] :

- Gestion des risques en temps réel
- Surveillance globale de la sécurité du réseau
- Réactivité automatique aux intrusions/incidents
- Prévention multicouche des pertes de données (DLP)
- Chiffrement du point final au cloud
- Contrôle des dispositifs pour les supports amovibles
- Protection anti spam et anti-programmes malveillants de bout en bout.

e) La cohérence

Cette notion est relative à l'absence d'informations conflictuelles au sein d'un même objet (par exemple, une incohérence serait détectée si un « prix actuel » d'un produit est supérieur au « prix maximum » de ce même produit). Mais cette notion existe aussi au niveau service : les valeurs d'une instance d'un objet métier ne sont pas en conflit avec les valeurs d'une autre instance ou d'une instance d'un autre objet [17] [site5].

f) L'intégrité

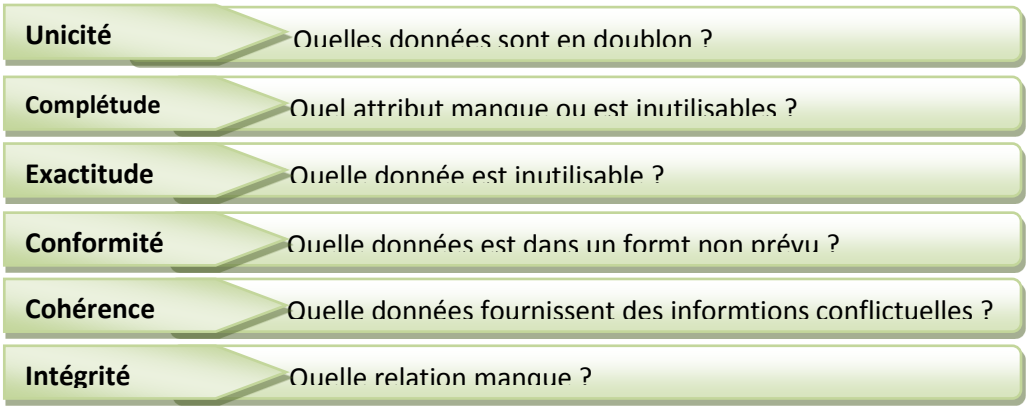
Les données, issues de différentes applications de production, peuvent exister sous toutes formes différentes.

Il faut les intégrer afin de les homogénéiser et de leur donner un sens unique, compréhensible par tous les utilisateurs.

Elles doivent posséder un codage et une description unique.

La phase d'intégration est longue et pose souvent des problèmes de qualification sémantique des données à intégrer (synonymie, homonymie, etc.…).

Ce problème est amplifié lorsque des données externes sont à intégrer avec les données du SIO. [17][Site 6]



Figur(2-3): Principaux critères de qualité

1.3.Les critères de sécurité

a) L'actualité

De nombreuses dimensions qualité concernent le rapport entre les données et le temps (actualité):

- L'obsolescence est le fait que la valeur de la donnée, autrefois exacte, ne l'est plus suite à un changement (dans le monde réel) de l'objet représenté.
- L'obsolescence peut aussi porter sur la représentation d'une donnée qui a été modifiée.
- Une valeur de donnée est à jour si elle est correcte en dépit d'un écart possible avec la valeur exacte, due à des changements liés au temps ; une donnée est périmée à la date si elle est incorrecte à cette date mais était correcte aux instants précédant.

L'actualisation est le degré mesurant à quel point une donnée en question est à jour (par exemple, l'âge ne devient obsolète qu'à la date anniversaire). []

Les utilisateurs présentent une grande sensibilité aux données mises à jour trop tardivement, situation qui leur impose l'utilisation d'historiques incomplets ou le recours à des données trop anciennes. Ce constat met en évidence le besoin de [17] :

- créer et mettre à jour les données suffisamment souvent (fine granularité temporelle) ;
- mettre à disposition des utilisateurs le plus vite possible.

b) L'accessibilité

L'accessibilité est la dimension qualité qui concerne la facilité d'accès aux données. Cela signifie que les services de données sont calibrés en fonction de leur utilisation et qu'ils existent souvent aussi bien en mode événement (déclenché à chaque mise à jour), qu'en mode requête (à la demande d'un processus consommateur) ou en mode batch pour des synchronisations en masse (pour le décisionnel par exemple). [17]

c) La pertinence

La pertinence est la dimension qualité qui définit l'utilité d'une donnée. Une donnée peut être accessible mais tellement détaillée que de nombreux attributs de l'objet proposé sont inutiles aux processus consommateurs. Une donnée doit être adéquate à son usage. Les services de donnée seront d'autant mieux utilisés que la granularité d'information dispensée correspondra aux besoins. [17]

d) La compréhensibilité

La compréhensibilité est la dimension qualité associée à la question : « cette donnée est-elle compréhensible ? ».

Une donnée est compréhensible si chaque utilisateur, chaque processus, chaque application trouve facilement la bonne information parmi les attributs disponibles d'un objet. C'est le cas si celui-ci est clair et que l'alignement sémantique de l'ensemble des concepts entre tous les dépositaires (humains ou informatiques) a été réalisé et documenté. [17]

3. Les objectifs de qualité des données :

L'objectif principal de qualité des données est d'assurer l'exactitude, la pérennité, la pertinence et la consistance des données à travers une organisation ou à travers les différentes divisions d'une organisation et des lors assurer que les décisions prises le sont sur des informations consistantes et justes.

En fonction des utilisations de la donnée, on détermine les critères qualité primordiaux à contrôler. Ensuite, on détermine les attributs de la donnée permettant de mesurer les critères qualité. Enfin, on spécifie le niveau minimal de qualité requise pour chaque critère retenu.

Les objectifs sont plus nombreux et plus diversifiés que dans le simple univers décisionnel. La granularité d'information et de suivi qualité doit répondre à tous les objectifs des processus consommateurs.

4. Principaux problèmes de la non qualité des données

4.1.Les problèmes de la qualité des données

Les problèmes de qualité des données se répandent de façon endémique à tous les types de données (structurées ou non) et dans tous les domaines d'applications. Les conséquences des données de mauvaise qualité sur les prises de décision et les coûts financiers qu'elle engendre sont considérables. [12]

Pour définir les problèmes de qualité dans l'entreprise, il est recommandé de définir les dimensions possibles et leur importance [12]:

- Duplication: les données sont répétées. L'entité est gérée par plusieurs systèmes d'informations sous des identifiants différents et donc sa vue n'est pas unifiée.

- Standards: les valeurs sont correctes par rapport à un intervalle de répartition ou à un domaine.

- Par manque de standards de codification, l'entreprise « Les chantiers Techniques de Marseille" peut apparaître comme <c Ets CTM », <c C.T.M. " ou « CTM SA »

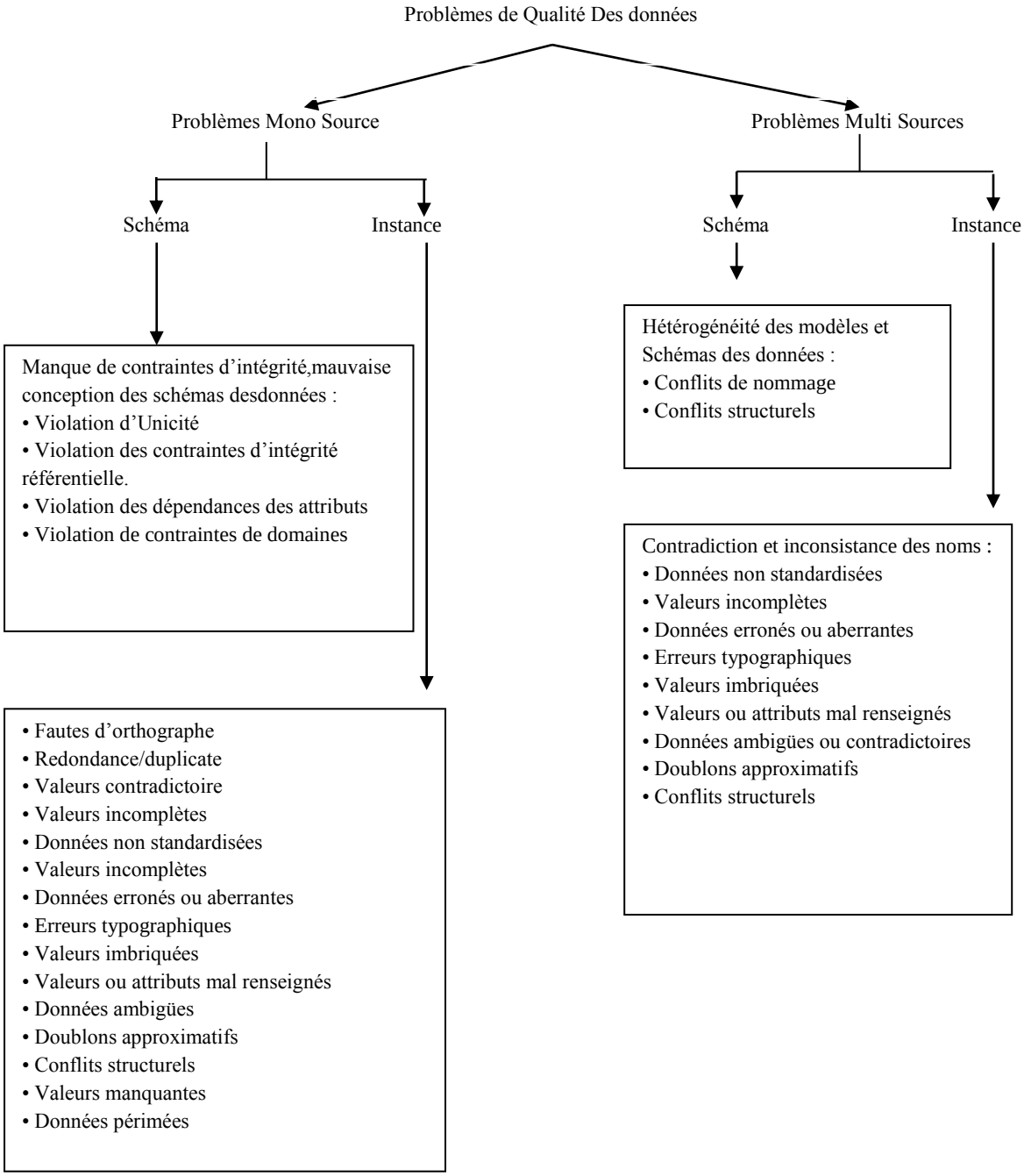
- Intégralité: toutes les données nécessaires sont disponibles pour le besoin métier. Il est impossible d'effectuer une campagne d'e-mailing avec une base de données clients ne contenant pas l'adresse email.

- Exactitude: les données représentent la réalité ou sont vérifiables à partir d'une source externe.
    - Le code postal ne correspond pas à la localité, le téléphone a changé ou le SIRET n'a pas été mis à jour lors du déménagement de l'entreprise.

- Interprétabilité : une donnée doit être représentée sous un format cohérent et sans ambiguïté. Par exemple, affichée sous la forme 02/01/1991 sur l'écran du responsable du personnel d'Ouargla, la date de naissance d'un employé est exacte, mais doit être affichée 01/02/1991 sur l'écran de son collègue d'Oum El Bouaghi.

- Opportunité: les données sont à jour au moment de leur utilisation. Le rapport mensuel des ventes doit inclure tous les résultats actualisés du mois pour toutes les régions commerciales.

Les données doivent avoir la qualité nécessaire pour supporter le type d'utilisation. En d'autres termes, la demande de qualité est aussi importante sur les données nécessaires à l'évaluation d'un risque que sur celles utilisées dans une opération de marketing de masse [12].



Figur(2-4): Classification des problèmes de la qualité des données.

4.1. Problème de rencontres

ÉTAPES DE TRAITEMENT	SOURCES DE PROBLEMES DE QUALITE DES DONNEES
Création des données	• Les Entrée manuelle : absence de vérifications systématiques des formulaires de saisie. • Entrée automatique : problèmes de capture OCR, de reconnaissance de la parole Incomplétude, absence de normalisation ou inadéquation de la modélisation conceptuelle des données : attributs peu structurés, absence de contraintes d'intégrité pour maintenir la cohérence des données. • Entrée de doublons. • Approximations. • Contraintes matérielles ou logicielles. • Erreurs de mesure. • Corruption des données : faille de sécurité physique et logique des données.
Collecte Import des données	• Destruction ou mutilation d'information par des prétraitements inappropriés. • Perte de données : buffer overflows, problèmes de transmission. • Absence de vérification dans les procédures d'import massif. • Introduction d'erreurs par les programmes de conversion de données.
Stockage des données	• Absence de méta-données. • Absence de mise à jour et de rafraîchissement des données obsolètes ou répliquées. • Modèles et structures de données inappropriés, spécifications incomplètes ou évolution des besoins dans l'analyse et conception du système. • Modifications ad hoc. • Contraintes matérielles ou logicielles.
Intégration des données	• Problèmes d'intégration de multiples sources de données ayant des niveaux de qualité et d'agrégation divers. • Problèmes de synchronisation temporelle.

	• Systèmes de données non conventionnels. • Facteurs sociologiques conduisant à des problèmes d'interprétations et d'intégration des données. • Jointures ad hoc. • Appariements aléatoires. • Heuristiques d'appariements des données inappropriées.
Recherche et analyse des données	• Erreur humaine. • Contraintes liées à la complexité de calcul. • Contraintes logicielles, incompatibilité. • Problèmes de passage à l'échelle, de performances et de confiance dans les résultats. • Approximations dues aux techniques de réduction des grandes dimensions. • Utilisation de boîtes noires pour l'analyse. • Attachement à une famille de modèles statistiques. • Expertise insuffisante d'un domaine. • Manque de familiarité avec les données.

Tableau 2-1 : Les problèmes de qualité des données [18,19]

5. Amélioration de la qualité des données :

5.1. Approche globale

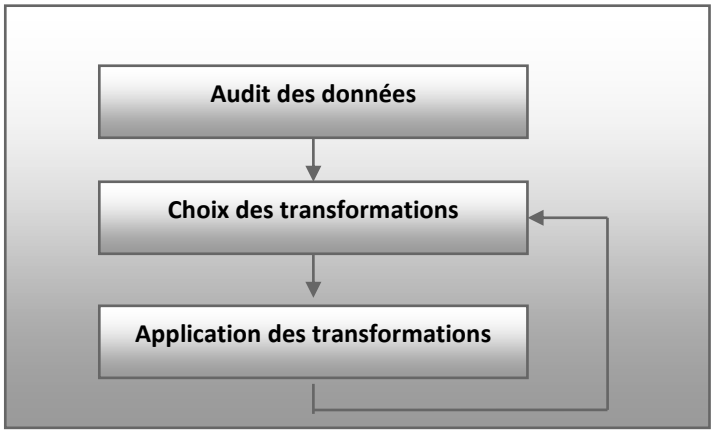
L'amélioration de la qualité des données est une démarche continue. Elle commence dès l'analyse des sources de données, et se poursuit avec la préparation du chargement du référentiel, et consiste enfin en un suivi régulier de l'activité [17].

Les étapes d'améliorer la qualité des données :

- Valider le niveau de qualité sur l'existant.
- Définir le niveau de qualité cible.
- Atteindre le niveau de qualité cible.
- Rester à ce niveau.
- Surveiller la qualité.

5.2. Approche « nettoyage », ou data cleansing

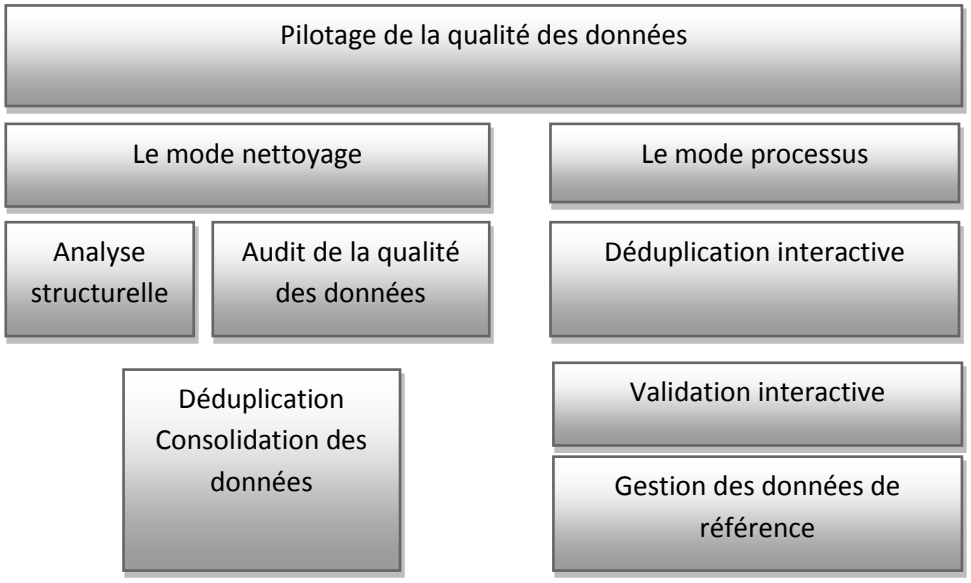
Le nettoyage des données fait partie des stratégies d'amélioration automatique de la qualité des données. Le problème de nettoyage des données qui consiste à détecter et éventuellement corriger des incohérences et des erreurs trouvées dans des jeux des données originaux, est bien connu dans le domaine de l'aide à la décision et des bases des données . Le nettoyage des données est un processus itératif et interactif qui comporte trois phases. [17]



Figur(2-5): Processus du nettoyage des données.

5.3. Approche « processus »

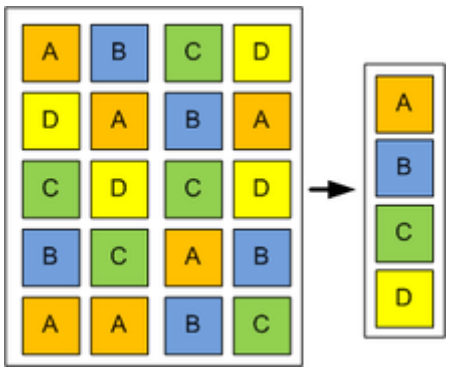
L'approche « processus » a pour objectif de prévenir l'introduction de données erronées dans un système d'information. On entend par « processus » toute la chaîne de traitements et d'opérations, de la création des données à leur destruction, en passant éventuellement par des modifications de leurs valeurs [17].



Figur(2-6): Approches « nettoyage » et « processus » de la qualité des données.

6. La déduplication des données :

La déduplication consiste, comme la compression, à identifier et factoriser les données redondantes. Elle est réalisée, non plus au niveau d'un fichier ou d'une base de données, mais sur un ensemble de fichiers voire sur une baie de stockage entière. Sur ce principe très général, se cachent de multiples concepts et produits dont le succès est dopé par celui de la virtualisation. Les machines virtuelles génèrent en effet d'importantes redondances de données et posent de gros problèmes d'espace disque de stockage [site 7].



Figur(2-7): Exemple de la déduplication des données

La technologie de déduplication de données offrant les meilleures performances en présence de segments de données qui se répètent, elle est surtout utilisée de nos jours pour le stockage des données de sauvegarde. La méthodologie permet au disque de prendre en charge

la rétention de jeux de données de sauvegarde pendant une période prolongée et peut être utilisée pour restaurer des fichiers ou des jeux de données complets à partir de n'importe quel événement de sauvegarde. Parce qu'elle fonctionne souvent sur des flux de données créés pendant le processus de sauvegarde, la déduplication a été conçue pour pouvoir identifier des blocs de données récurrents stockés à des emplacements différents dans un jeu de données transmis. Les blocs de taille fixe étant mal adaptés à ces conditions, la méthodologie de déduplication de Quantum est conçue autour d'un système de segments de données de longueur variable [24].

6.1. Pourquoi utiliser la déduplication des données

La suppression des données redondantes permet de réduire sensiblement les besoins en matière de stockage tout en améliorant l'efficacité de la bande passante. En effet, comme les coûts du stockage primaire ont diminué avec le temps, les entreprises stockent souvent plusieurs fois la même information de façon à ce que les nouveaux employés peuvent réutiliser des travaux antérieurs. Or, certaines opérations, comme la sauvegarde, stockent des informations extrêmement redondantes [site 8].

Le processus de déduplication des données permet alors de réduire les coûts de stockage en limitant le nombre de disques nécessaires. Il contribue également à améliorer la reprise après sinistre en raison du volume de données bien moins important à transférer. Les données de sauvegarde et d'archivage comportent habituellement beaucoup d'informations en double [site 8].

Les mêmes données sont stockées encore et encore. Elles occupent de l'espace de stockage sur les disques ou les bandes, elles consomment de l'électricité pour l'alimentation et le refroidissement des lecteurs de disques et de bande, et elles utilisent de la bande passante lors des opérations de réplication. Ce gaspillage génère une chaîne de coûts et un manque d'efficacité dans la gestion des ressources au sein de l'entreprise [site 8].

6.2. Niveaux de déduplication des données

La déduplication est née dans le monde de la sauvegarde, domaine dans lequel elle peut être réalisée au niveau de l'outil (on parle de sauvegarde à la source) ou de celui du support (à la cible).

a. La déduplication au niveau de la sauvegarde

L'adoption de la déduplication s'accélère dans les entreprises, car c'est aujourd'hui une solution fiable pour gérer la croissance des données. En effet, l'objectif de la déduplication est de réduire de façon significative la taille des sauvegardes, ce qui permet pour les entreprises d'économiser sur les infrastructures (matériel et logiciel) de stockage. La majorité des éditeurs de solutions de sauvegarde ont ajouté aujourd'hui à leur offre un module de déduplication qui fonctionne soit à la source, soit à la cible avec chacun leurs avantages et leurs inconvénients [site 9].

Si la déduplication reste essentiellement exploitée lors des opérations de sauvegarde, elle l'est beaucoup moins pour les données primaires faute d'outils techniques matures. Une situation qui devrait changer dans les années à venir avec l'arrivée de solutions performantes associant baies SSD (Solid State Storage) et déduplication. Enfin, ne pouvant pas fonctionner sur la technologie des bandes, la déduplication s'effectue sur des disques, la bande reste donc de plus en plus confinée à l'archivage [site 9].

b. La Déduplication à la source ou à la cible :

La déduplication consiste à éliminer les données redondantes lors de leur sauvegarde : au niveau de la source, ce qui réduit les besoins en bande passante, et au niveau de la cible, ce qui réduit les besoins en espace de stockage. Cette phrase résume bien la différence entre les déduplications à la cible et à la source. À la source pour une analyse plus fine « Dans le cas d'une déduplication à la source, donc du côté serveur, un tri est réalisé directement sur le serveur d'application avec un niveau de granularité très fin, c'est un travail qui prend plus de temps, mais qui permet de supprimer les goulots d'étranglement »[site 10].

Conclusion :

Dans ce chapitre, nous avons présenté un état de l'art sur la qualité des données et nous avons présente les entrepôts de données et leur problèmes. Après l'identification des limitations des différents travaux et approches dans les différents domaines nécessaires à l'amélioration de la qualité des données.

L'état de l'art que nous avons présenté dans ce chapitre sur la qualité des données et les entrepôts de données et de leur qualité est très utile dans notre travail, dans l'amélioration de leur qualité ainsi que le nettoyage des données.

Chapitre III.

La Conception

Introduction

Dans ce chapitre, nous allons présenter les méthodes et les outils dont nous aurons besoin dans notre travail. Tout d'abord, nous allons parler de l'application TALEND et de l'algorithme fonctionnant sur la déduplication des données similaires. Nous allons montrer les différentes étapes de l'utilisation de cet algorithme dans l'application TALEND.

I. Talend Open Studio

1. Définition

Talend Open Studio est un ETL (Extract Transform Load), développé par la société Frances TALEND. Cette société de services en ingénierie informatique, dont le siège social est basé à Redwood City (États-Unis), compte plus de 400 personnes. [Site 6]

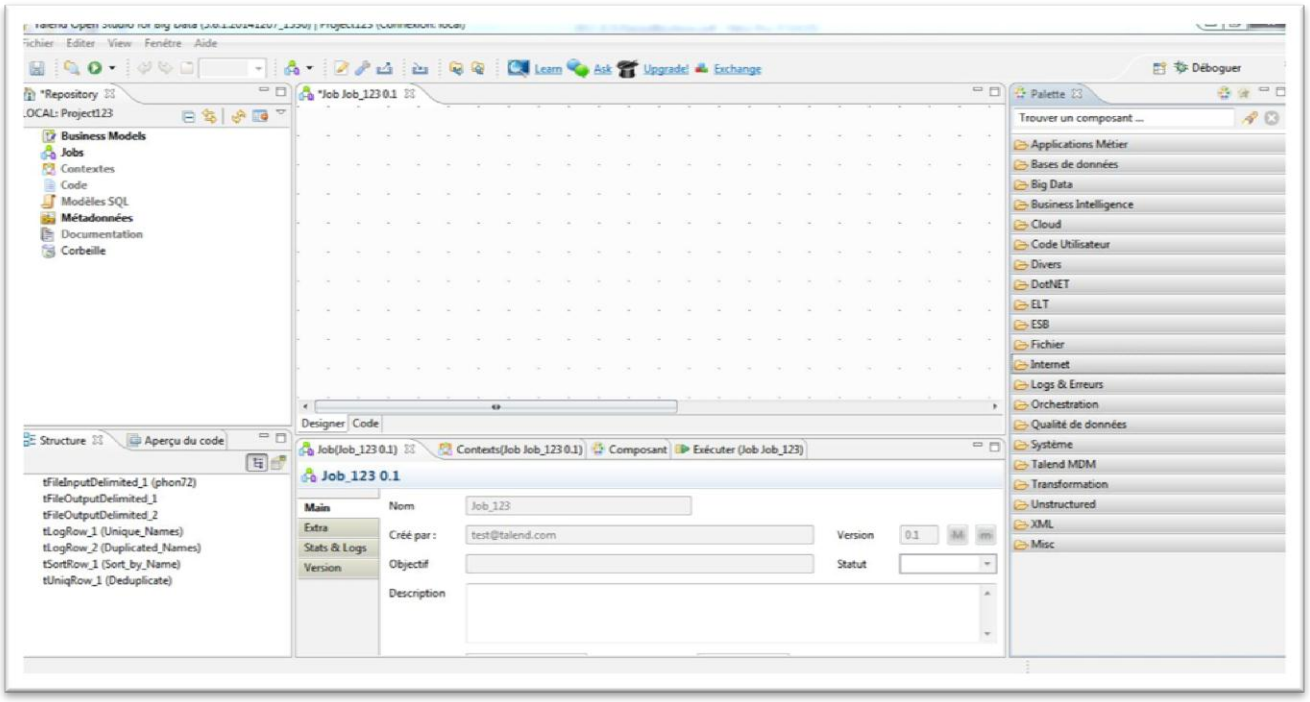
Créé en 2005, Talend Open Studio (TOS) est une plate-forme d'intégration de données Open Source, basée sur le langage Java. TOS permet de répondre à toutes les problématiques liées au traitement des données dans la chaîne décisionnelle :

- ETL : Extraction, Transformation, et Chargement des données.
- EAI : Echange de données Inter-Application.
- Synchronisation des données.

L'une des grandes forces de TOS réside dans la fait de pouvoir se connecter à quasiment toutes les sources de données, applications métier et type de fichier existant. Et ce grâce à plus de 250 composants utilisables par les développeurs. Parmi ses composants, on trouve différentes familles.

Dans notretravail, nous allons utiliser la version 5.6 de l'ETL Talend.

Figur(3-1): La version de l'application TALEND utilisée



Figur(3-2): l'interface de l'application TALEND Open Studio.

2. Avantages de TALEND Open Studio

- Portabilité de l'espace de travail optimisé grâce au référentiel sous forme de fichier.
- TOS tire parti des avantages de Java : portabilité, puissance,…
- Interface intuitive basée sur Eclipse.
- Vue graphique des jobs grâce aux interfaces graphiques élaborées des composants.
- Possibilité de créer de nouveaux composants.
- Communauté active.
- TALEND Open Studio est gratuit pour un utilisateur sur le référentiel. Son usage pour les PME d'un ou deux informaticiens permet de se doter d'un outil puissant sans engager de frais de licence.

3. Pourquoi choisir Talend ?

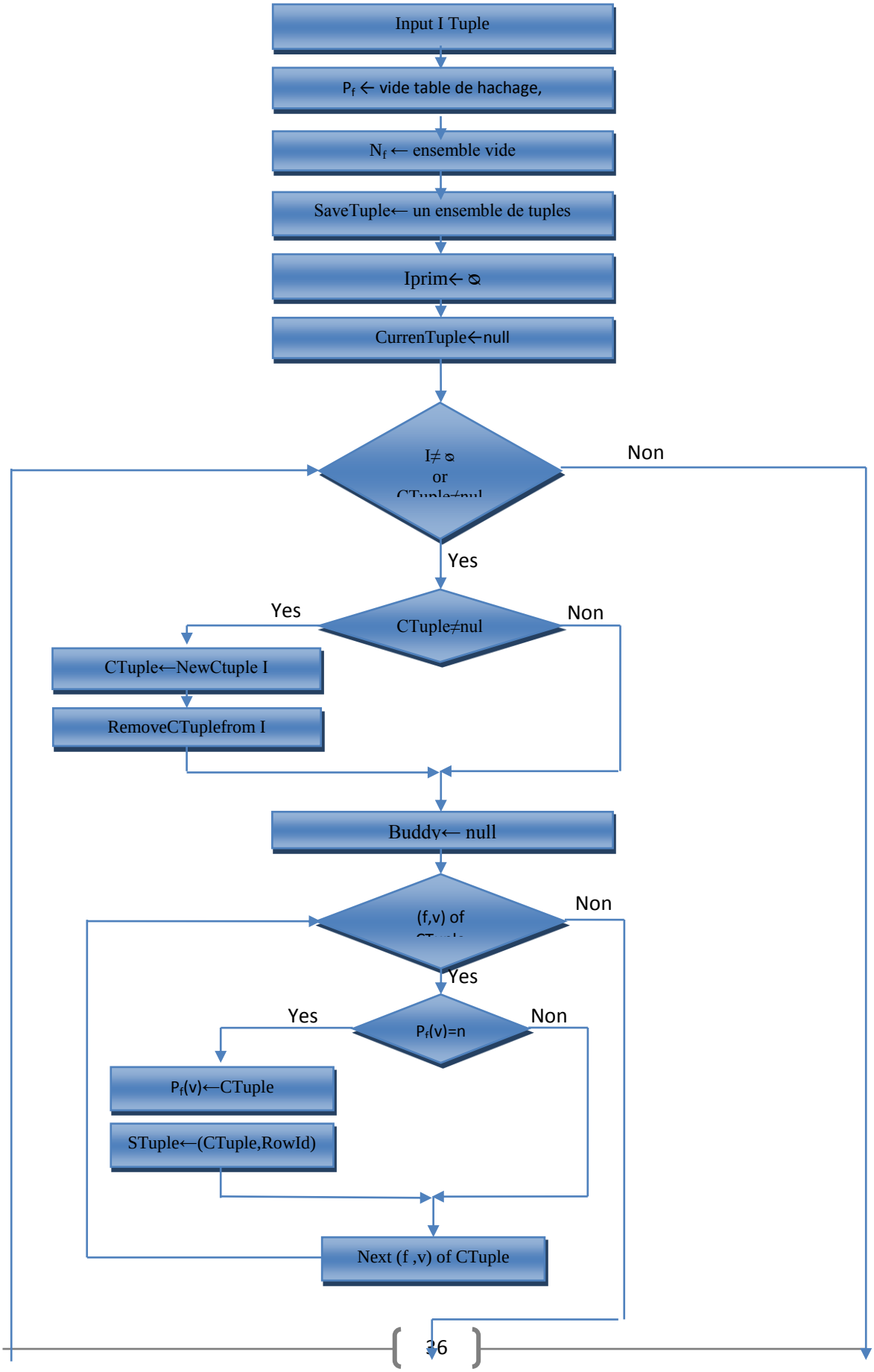
Talend Open Studio est un générateur de code Java approprié pour traiter les données dans un format facile, simple, rapide, et efficace, pour traiter gros volumes de données indépendamment de leur taille, dans un minimum de temps. C'est un éditeur compatible et intégré avec beaucoup de langages de programmation.

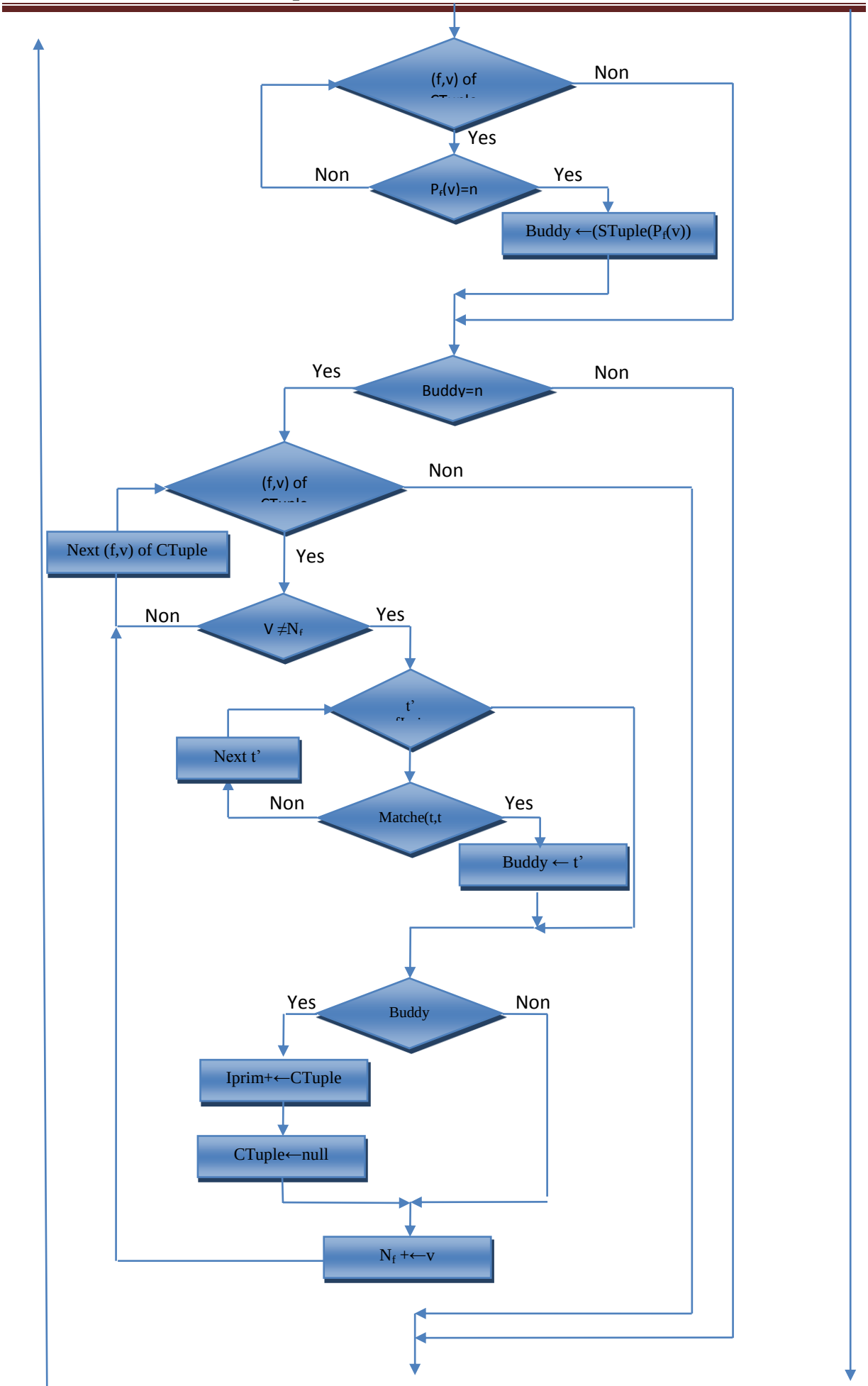
Toutes ces caractéristiques nous aident à résoudre le problème de la qualité des données, en particulier le problème que nous traitons sur la déduplication dans les grands entrepôts de données.

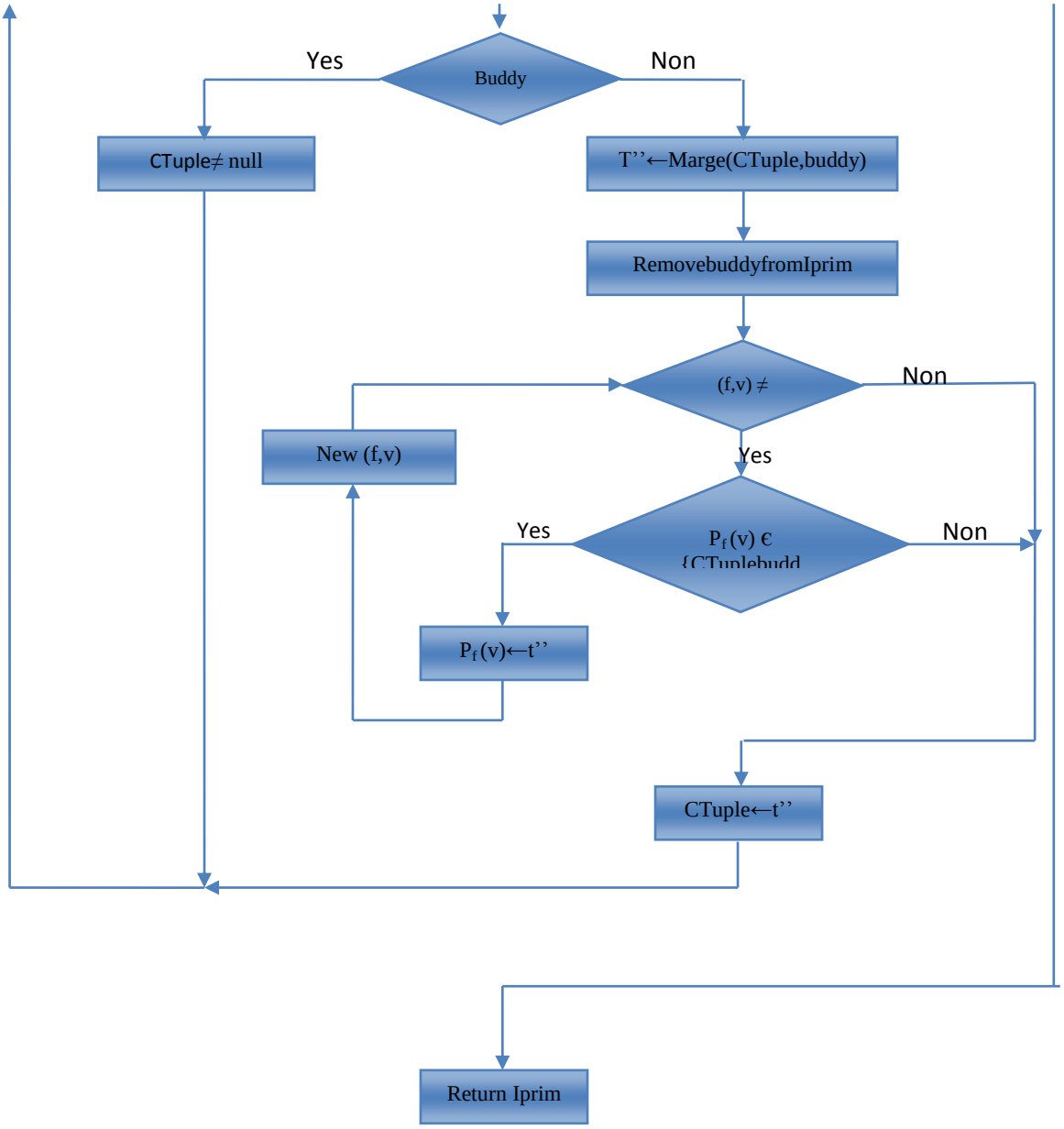
II. Algorithme de déduplication pour les bases et entrepôts de données :

La déduplication des données est une étape très importante dans le processus d'intégration de données hétérogènes. Elle assure une meilleure qualité pour les données résultats. Nous présentons dans cette partie, l'algorithme séquentiel pour l'élimination des données similaires utilisé dans notre projet.

1. **L'organigramme suivant représenté la démarche de l'algorithme :**




(f,v) of
Non
Yes
Non
$P_f(v)=n$
Yes
Buddy ←(STuple($P_f(v)$))
Yes
Buddy=n
Non
(f,v) of
Non
Next (f,v) of CTuple
Yes
Non
$V \neq N_f$
Yes
t'
Next t'
Non
Matche(t,t
Yes
Buddy ← t'
Yes
Buddy
Non
Iprim+←CTuple
CTuple←null
N_f +←v



Organigramme de : l'algorithme de la déduplication des données similaires.

2. Algorithme d'élimination de similaires

input : a set I of tuples,

output : a set Iprim of tuples, Iprim

P ⟵ empty hash table, for each feature f ,

N_f⟵ empty set, for each feature f,

SaveTuples⟵ a set of tuples, Iprim⟵∅,

currentTuple⟵ null

while I ≠∅ or currentTuple≠ null

ifcurrentTuple = null then

currentTuple⟵ a Tuple from I

removecurrentTuple from I

endif

buddy⟵null

for all(f, v) of currentTupledo

if $P_f(v)$ = null then

$P_f(v)$ ⟵currentTuple

SaveTuples⟵ (currentTuple, rowid)

endif

endfor

for all (f, v) of currentTupledo /*was any value previously encountered*/

if $P_f(v)$ ≠currentTuplethen /*we compare the value and the rowid*/

buddy⟵SaveTuples($P_f(v)$); exitfor

endif

endfor

if buddy = null then /*If not look for Matches*/

for all (f, v) of currentTupledo

if v ∉ $N_f$then/* if a value never made it to N_f*/

for all t' of Iprim do

```
ifMatch (t,t') then  buddy ⟵ t'
exitforendif
endfor
if  buddy≠ null thenexitforendif
add v to N  endif
endfor
endif
if  buddy = null then add currentTuple to Iprim
currentTuple⟵ null
else
t'' ⟵Merge (currentTuple, buddy)
remove buddy from Iprim
    /*update P*/
for all (f, v) where P f(v)Є{currentTuple, buddy}
do  Pf(v) ⟵ t'' endfor
/*update N*/
for all (f, v) where N f(v)Є{currentTuple, buddy}
doNf(v) ⟵ t''  endfor
currentTuple⟵  t''
endif
endwhile
returnIprim
```

L'algorithme de déduplication des données similaires

2.1. L'utilisation de l'algorithme au sein de l'application :

Ici, nous allons parler des étapes pour améliorer la qualité des données.

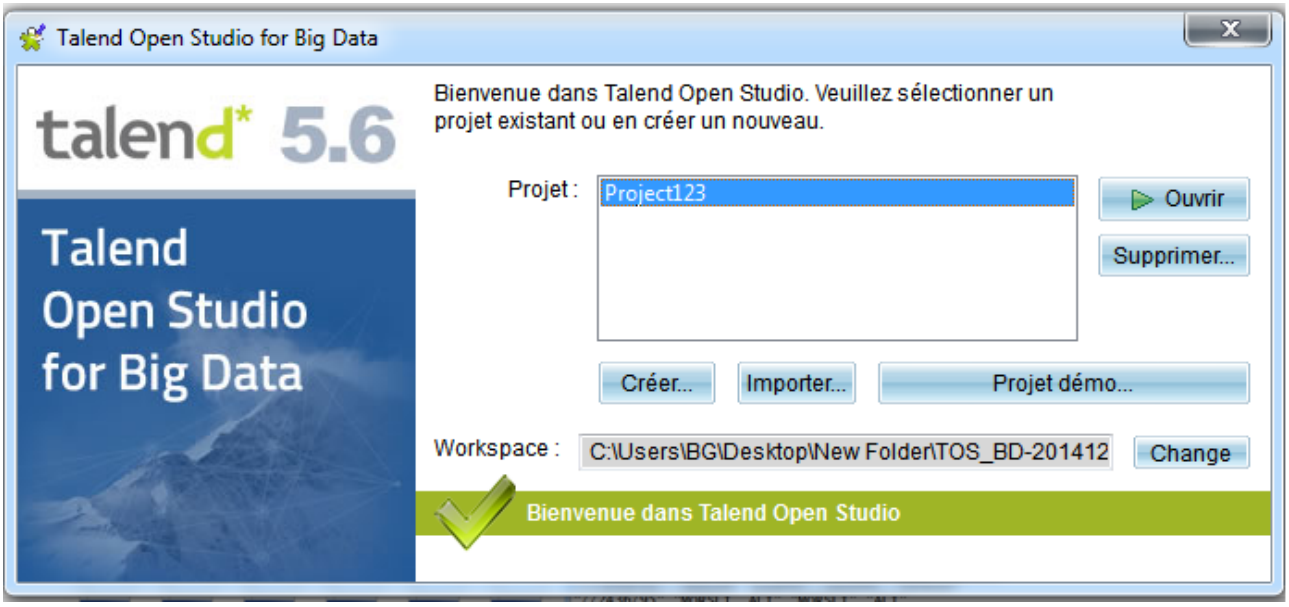
- Nous utilisons l'Algorithme précédent dans l'éditeur Talend, sur la liste de numéros de téléphone Djezzy.
- Nous allons suivre les étapes suivantes :

❖Avant d'exécuter Talend, on doit installer JDK.

Quel est le JDK ?

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, transformé en bytecode destiné à la machine virtuelle Java.

❖Lorsque vous ouvrez l'application TALEND, Une fenêtre s'affiche pour créer ou ouvrir un projet. Nous avons créé une PROJET (Par exemple' Project123').

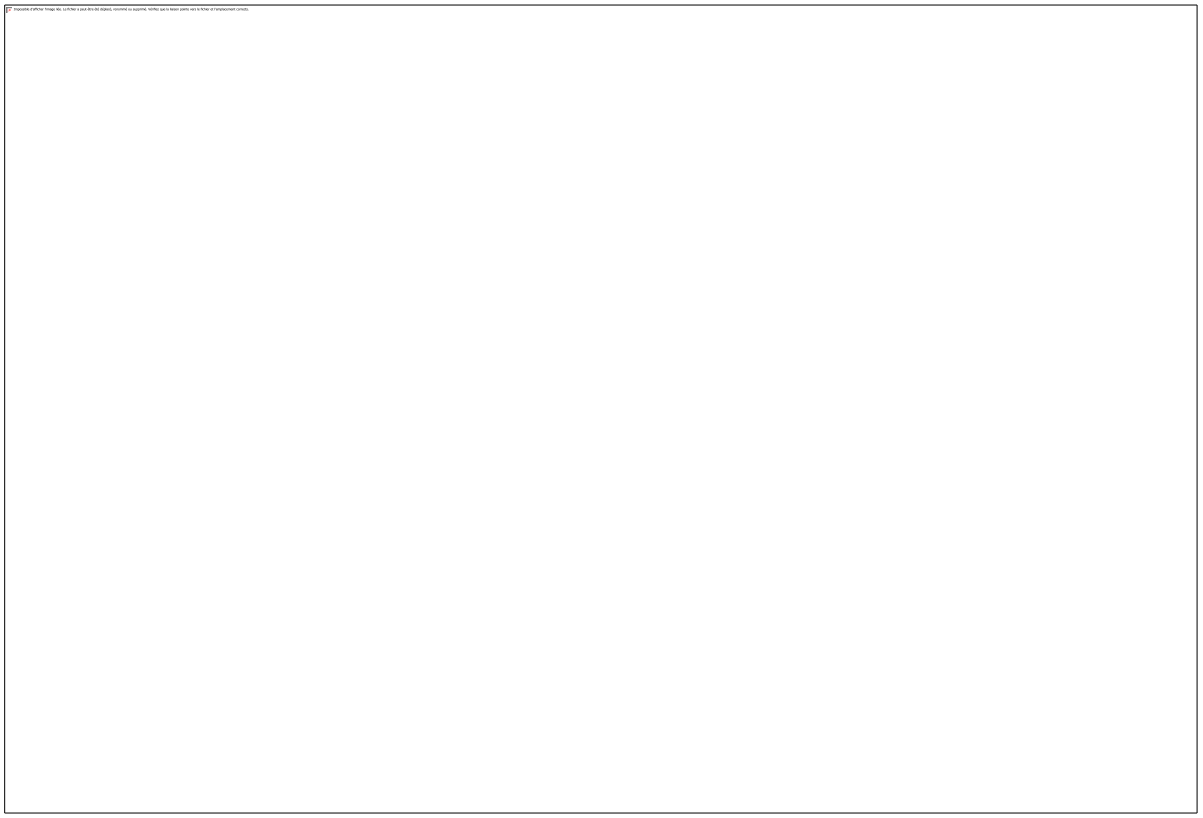


Figur(3-3): Fenêtre pour créer un projet ou sélectionner un projet existant.

- Ensuite, nous avons ouvert ce Projet.

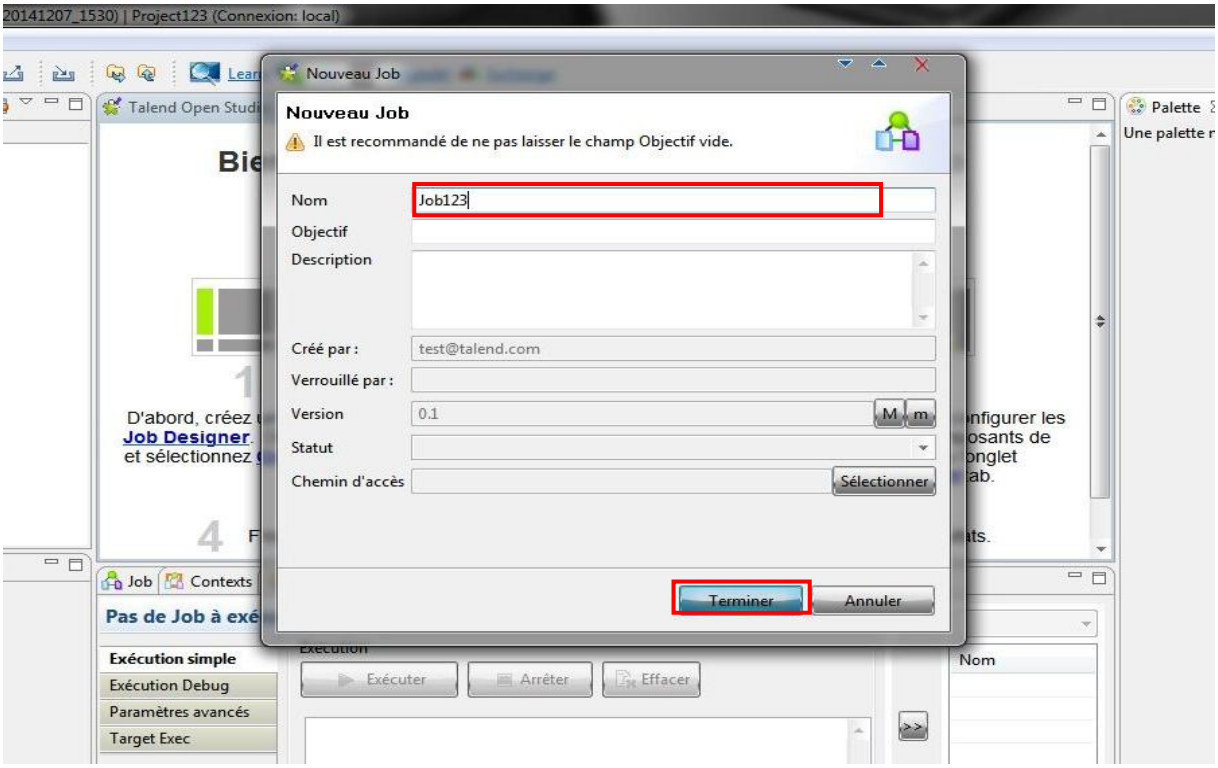
❖ Pour créer un JOB :

- Dans le référentiel à gauche (Repositiry), Cliquez-droit sur Job Designs.
- Dans le menu, cliquez sur Create Job pour ouvrir l'assistant Nouveau Job.



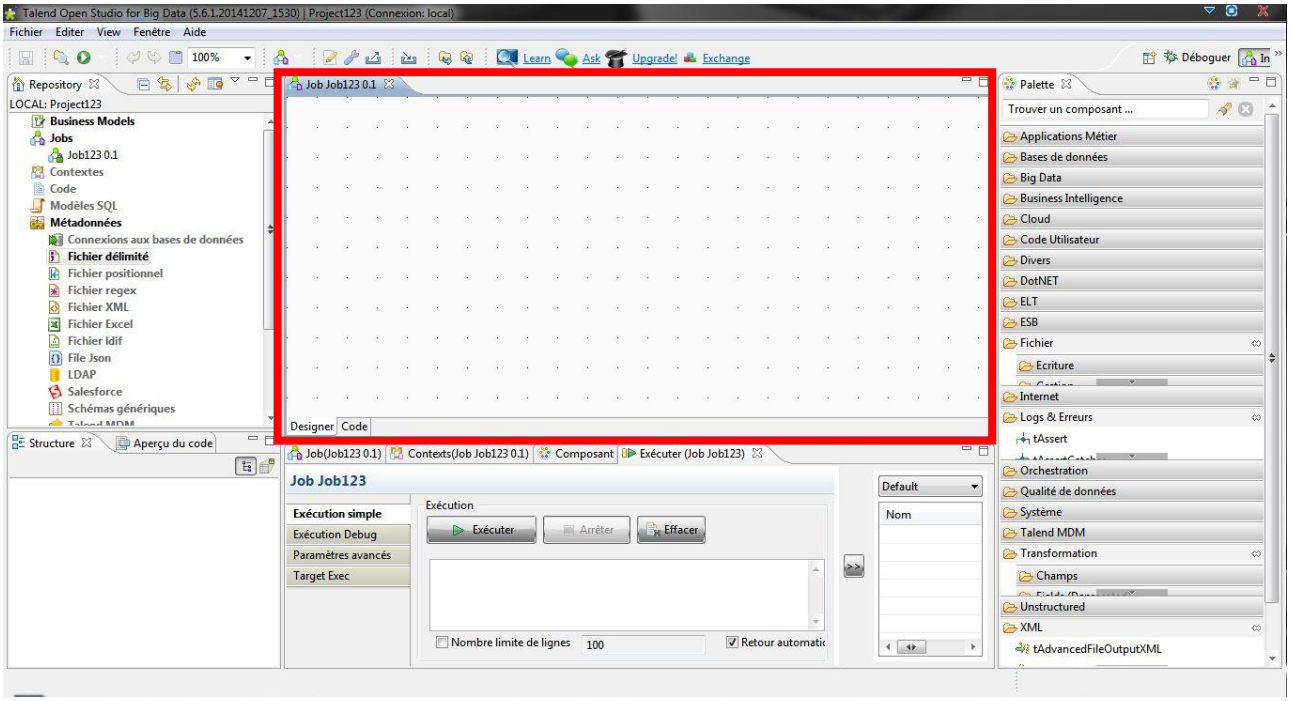
Figur(3-4): La Création de Job

- Dans le champ Nom, écrire le nom de job par exemple Project123



Figur(3-5): La Création de Job (Nommage et création)

- Cliquez sur Terminer pour fermer l'assistant et créer votre Job.
- Le Job Designer ouvre un Job vide

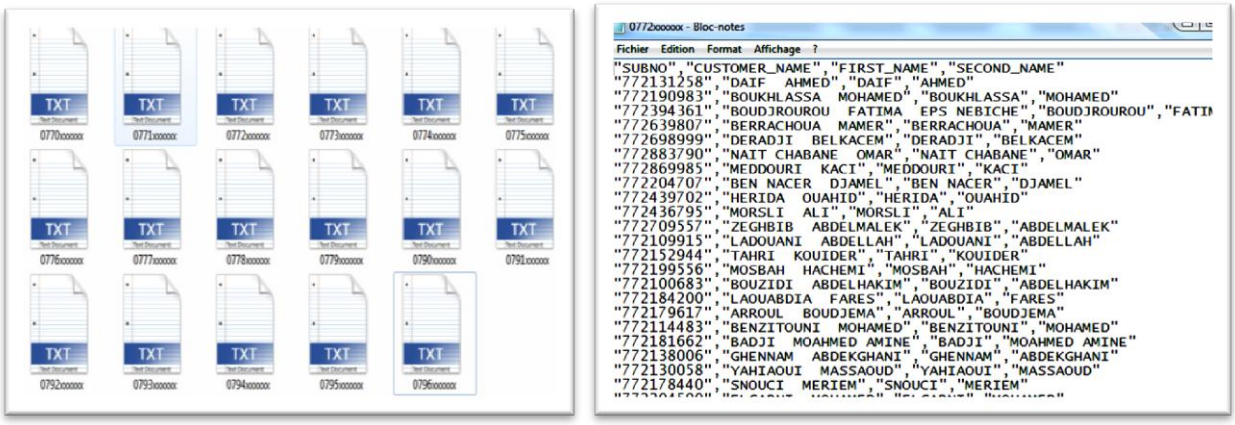


Figur(3-6): un Job Designer vide.

❖ Connecteur de lecture d'un fichier délimité

- Ensuite, on identifie les données que nous nous voulons nettoyer ou améliorer.

Par exemple, nous allons travailler ici sur les numéros de téléphone de l'entreprise de Djezzy, qui est sous le format txt.



Figur(3-7): Les fichiers utilisés

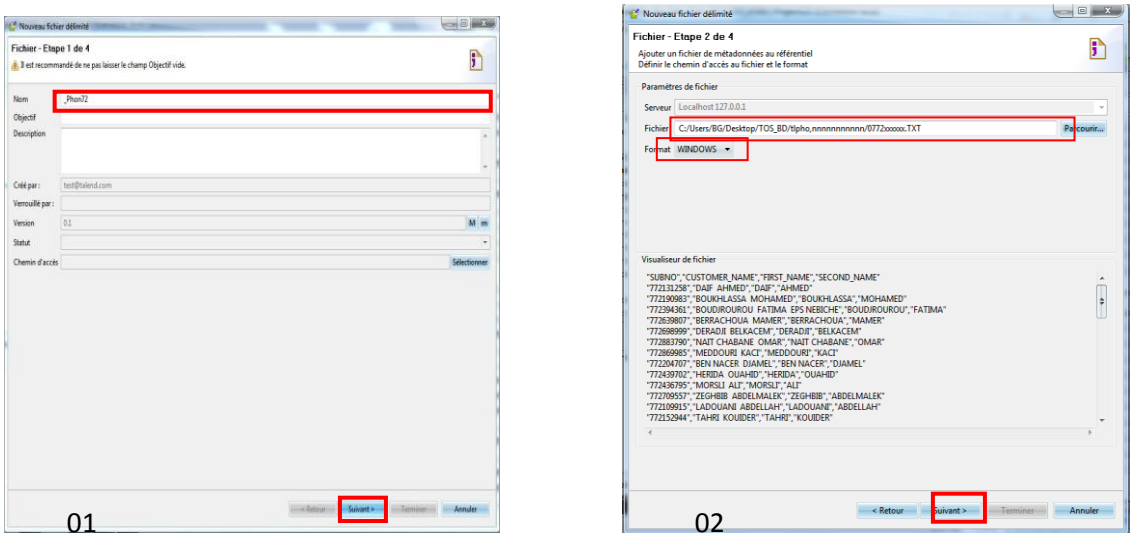
- On ajoute le composant d'entrée, dans le référentiel à gauche (Repositiry), Cliquez-gauche sur Métadonnées , Cliquez-droit sur Ficher délimité.

Dans le menu, cliquez sur « Create »un Fichier délimité pour ouvrir l'assistant Nouveau Fichier délimité.

- Dans la fenêtre on fait:

01 - choisir le nom de fichier (par exemple Phon73).

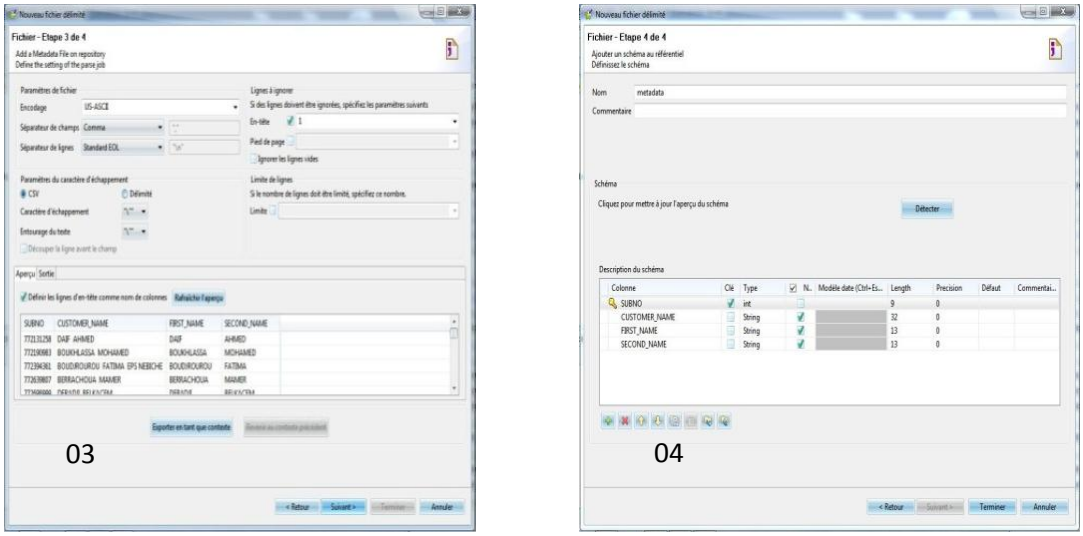
02 - Nous choisissons le fichier et le système, qui a été créé.



Figur(3-8): connecteur de lecture d'un fichier délimité (Les étapes 01 et 02)

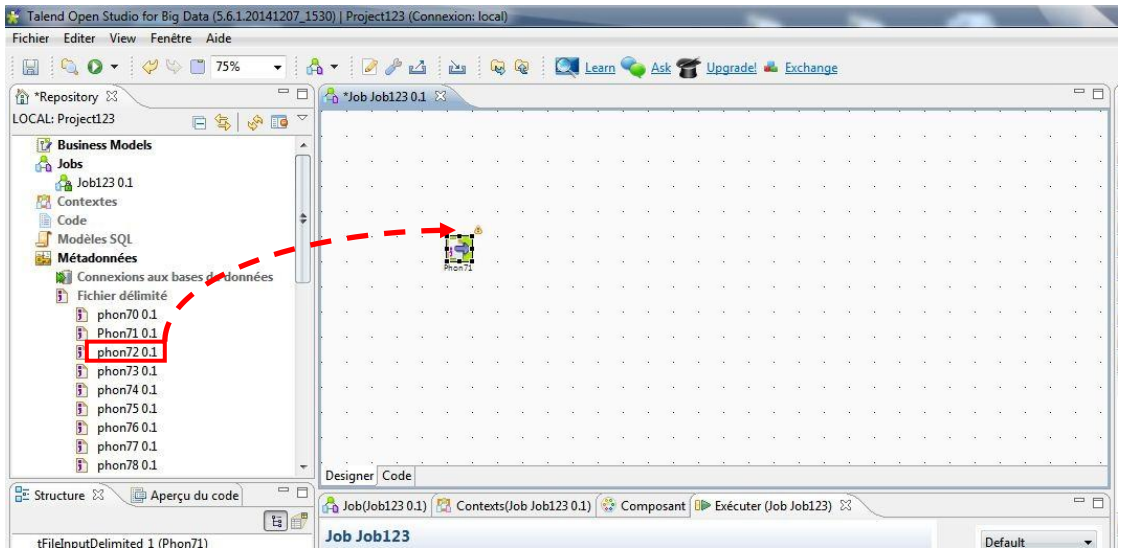
03 - définir le paramètre de la tâche d'analyse.

04 - identifier la clé.



Figur(3-9): connecteur de lecture d'un fichier délimité (Les étapes 03 et 04)

- Et on dépose sur le Job Designer

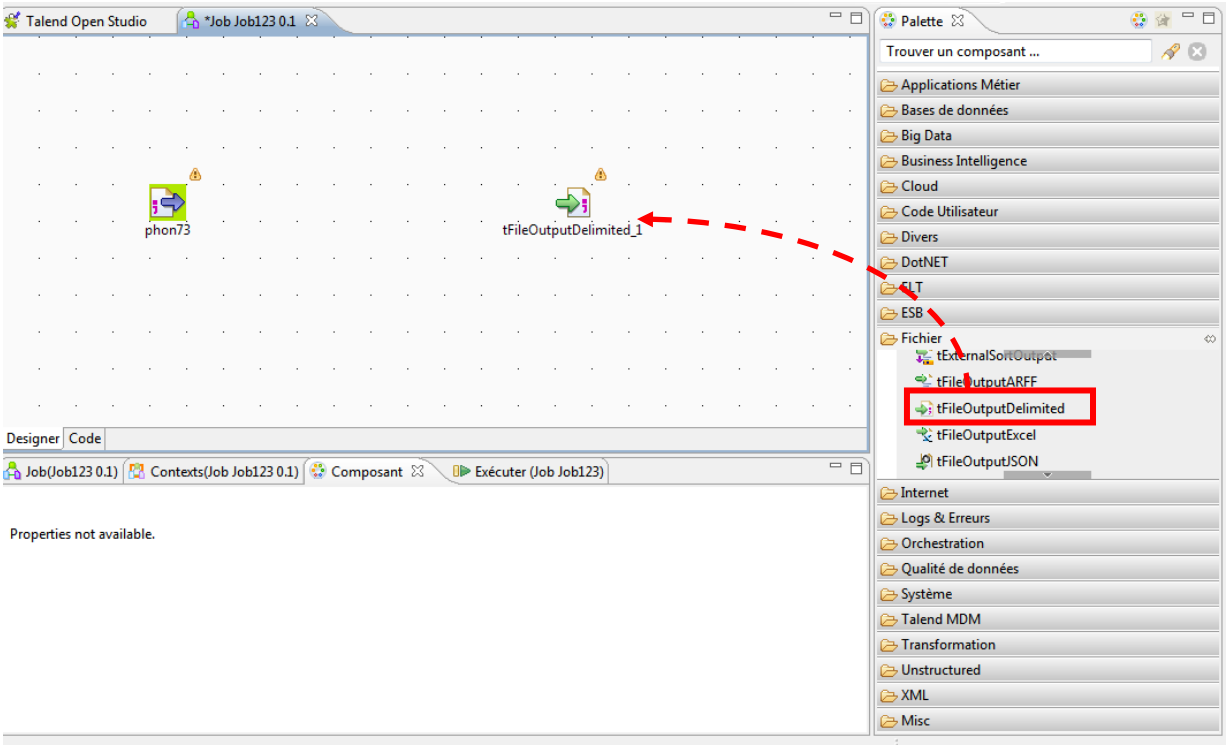


Figur(3-10): Déposer le ficher délimité de lecture sur le Job Designer

❖ Le connecteur d'écrire dans le fichier délimité

Pour ajouter le composant de sortie, dans la Palette à droite, cliquez sur la famille Fichier et sous-famille de l'écriture.

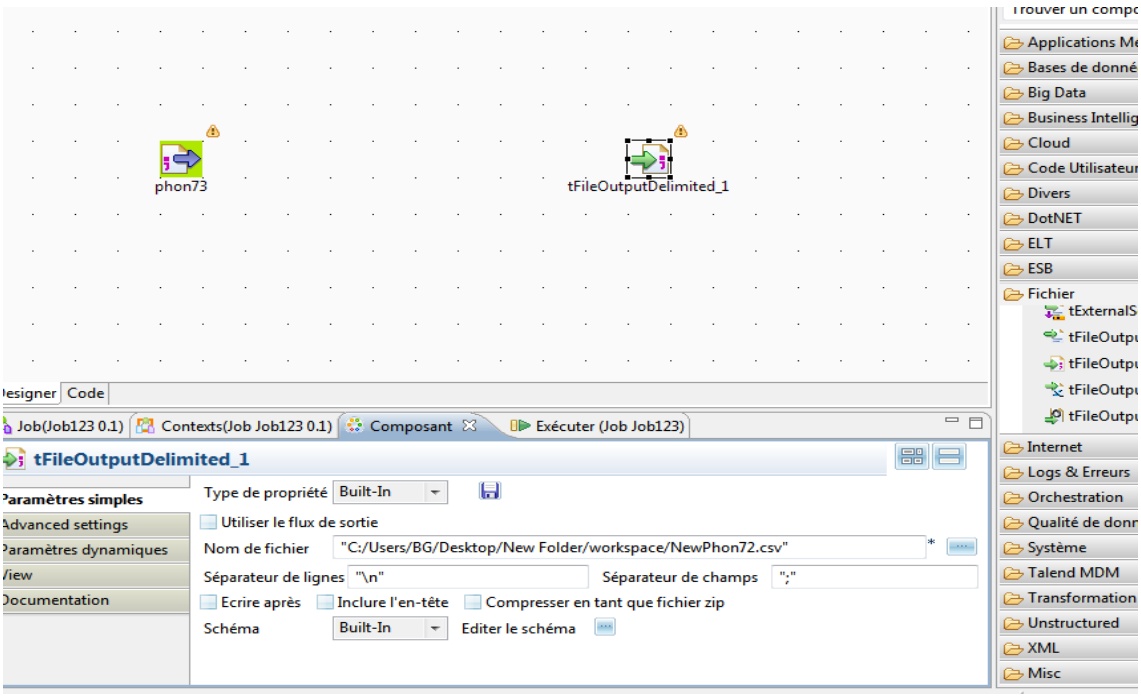
Cliquez sur le composant *tFileOutputDelimited*et le déposer sur le Job Designer



Figur(3-11): Déposer le fichier délimité d'écrire sur le Job Designer

❖ configurer les composants *tFileOutputDelimited*

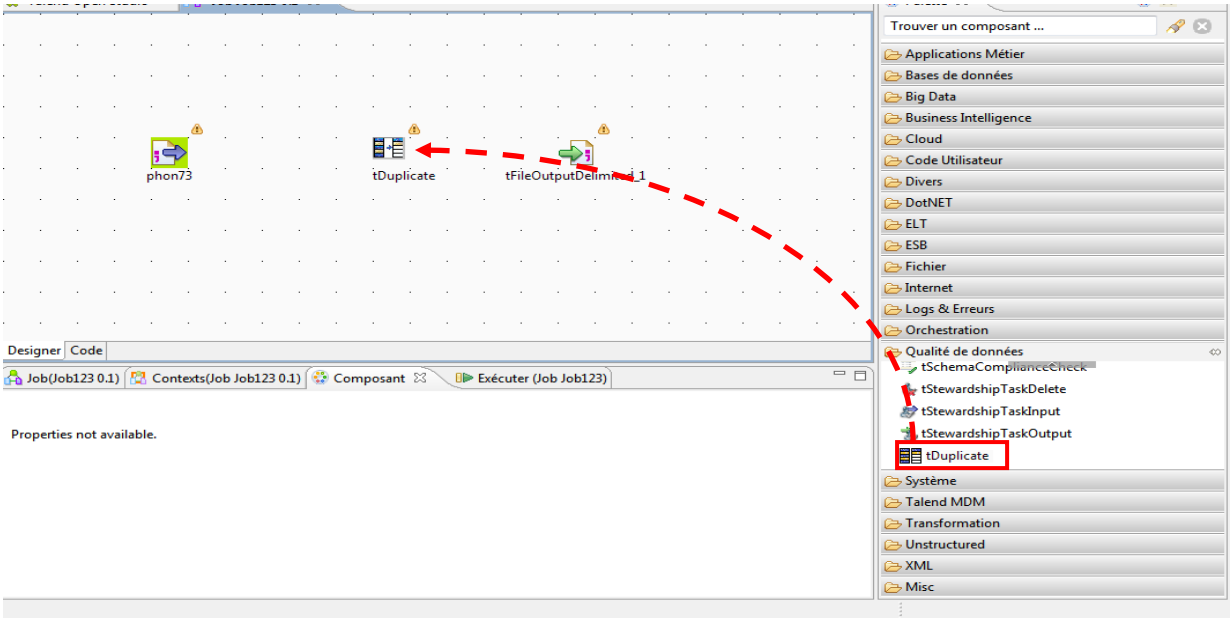
Double-cliquez sur le composant *tFileOutputDelimited*pour afficher sa vue Basic settings.



Figur(3-12): configurer les composants *tFileOutputDelimited*

Cliquez sur le bouton [...] à côté du champ Nom de fichier à sélectionner le nom et accédez à votre fichier de sortie (par exemple NewPhon73, dans le Desktop).

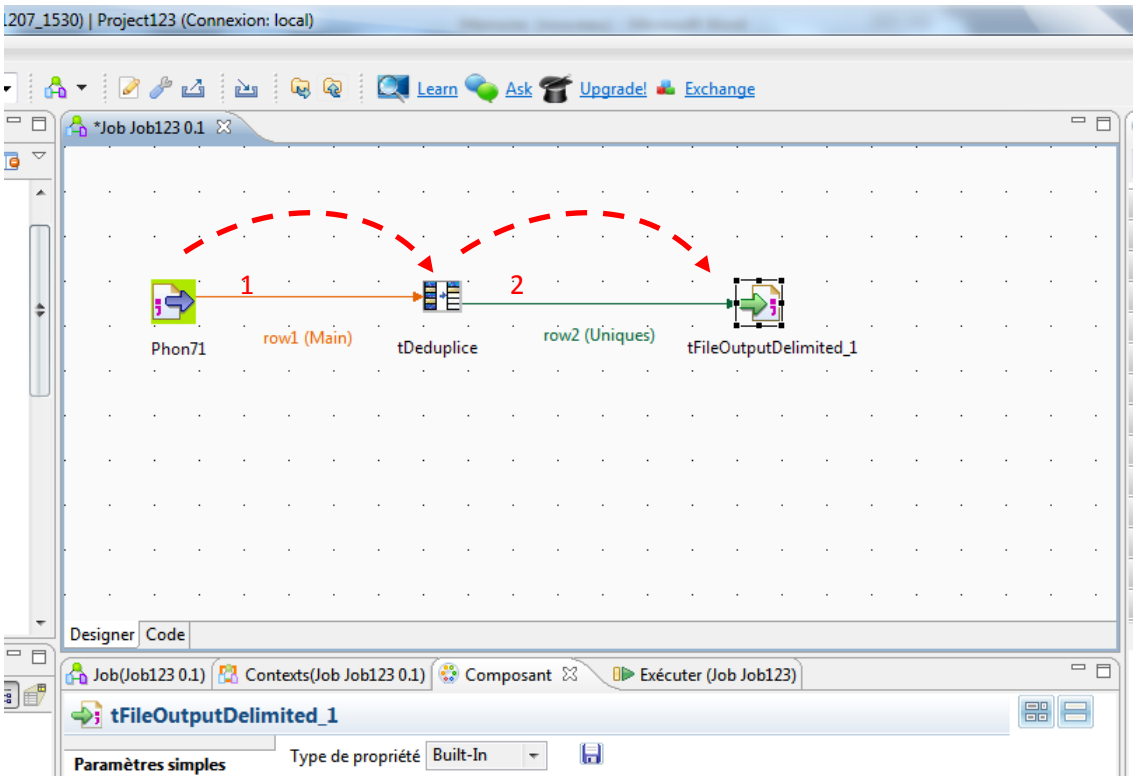
Pour ajouter le composant de traitement, dans la Palette à droite, cliquez sur la famille Qualité de données, Cliquez sur le composant *tDeduplicate*et le déposer sur le Job Designer



Figur(3-13): Ajout *tDeduplicate* dans le Job Designer

❖ Connectez les composants entre eux.

- Connectez le composant Phon73 (*tFileInputDelimited*) avec le composant *tDeduplicate* (En utilisant le bouton droit)
- Connectez le composant *tDeduplicate* avec le composant *tFileOutputDelimited* (En utilisant le bouton droit)

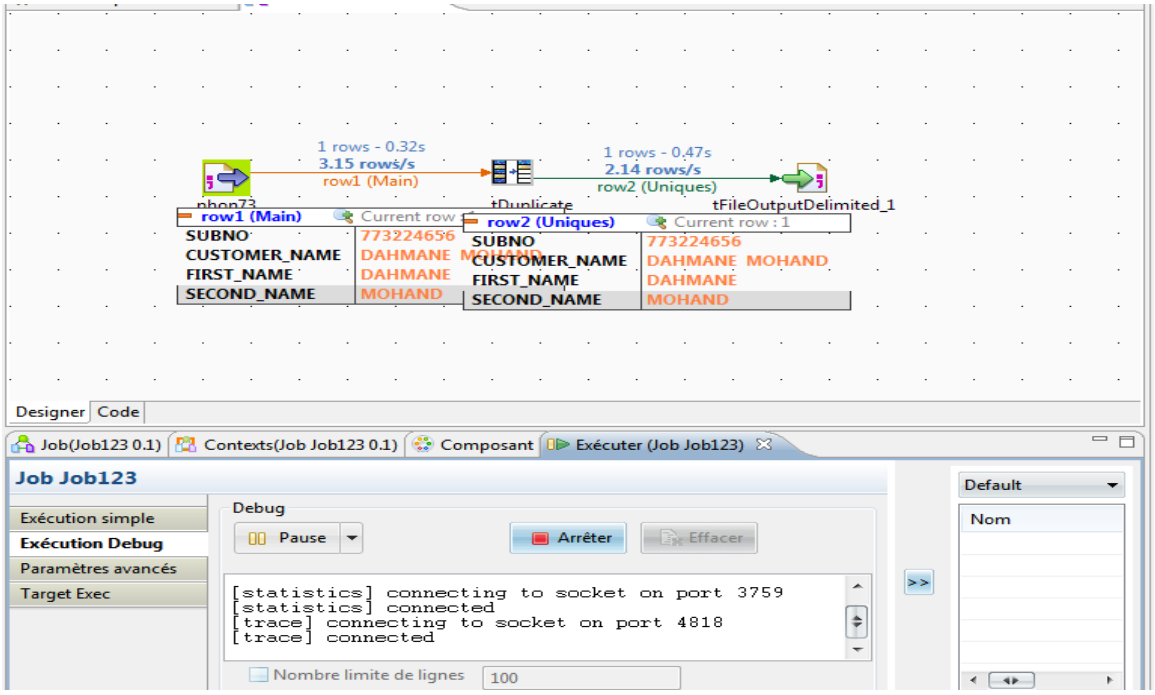


Figur(3-14): Connecter les composants entre eux.

❖ La Vérification et l'Exécution

Pour la vérification :

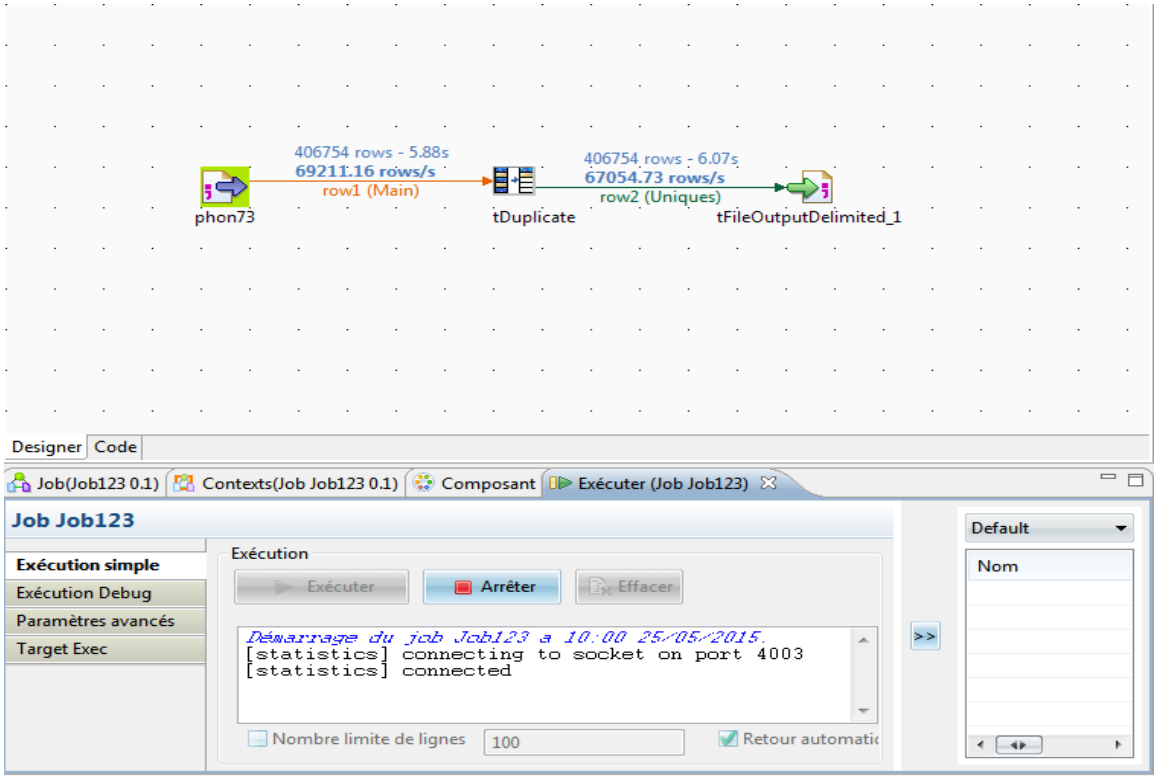
Dans la fenêtre Exécution on choisis Exécution debug, Et puis en cliquant sur Débogage des Traces Pour vérifier les statistiques d'exécution.



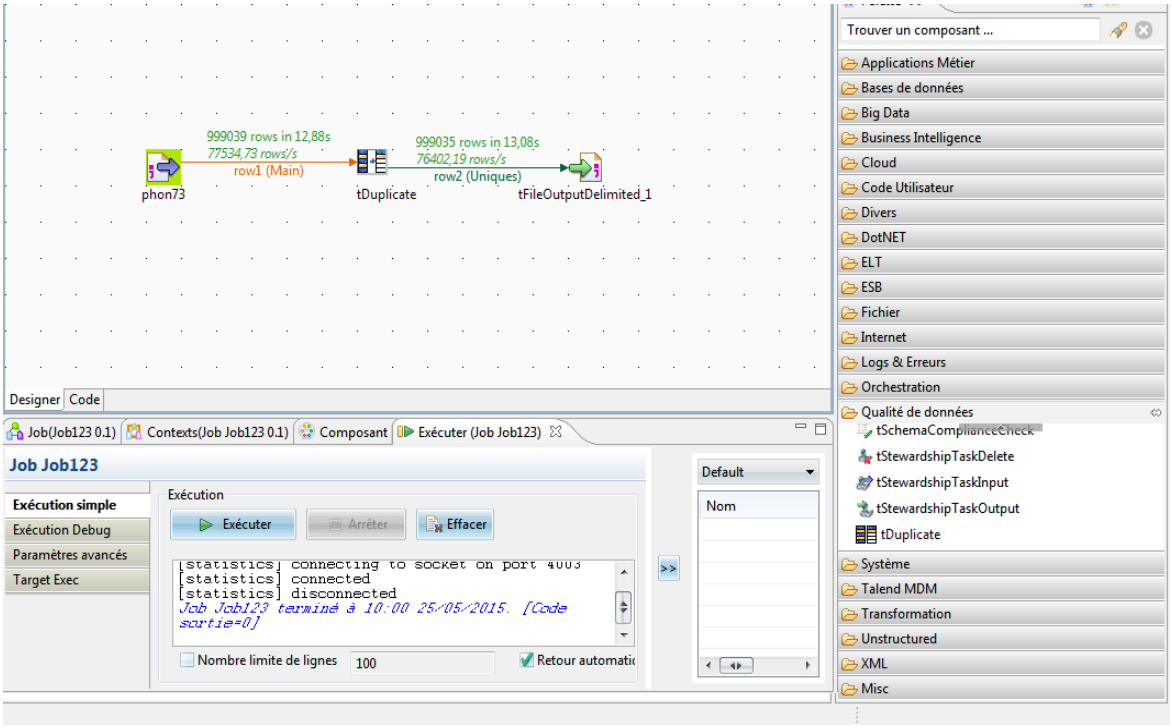
Figur(3-15): La Vérification

Et pour l'Exécution :

Dans la fenêtre Exécution on choisit Exécution simple, et puis en cliquant sur Exécuter pour lancer l'exécution.

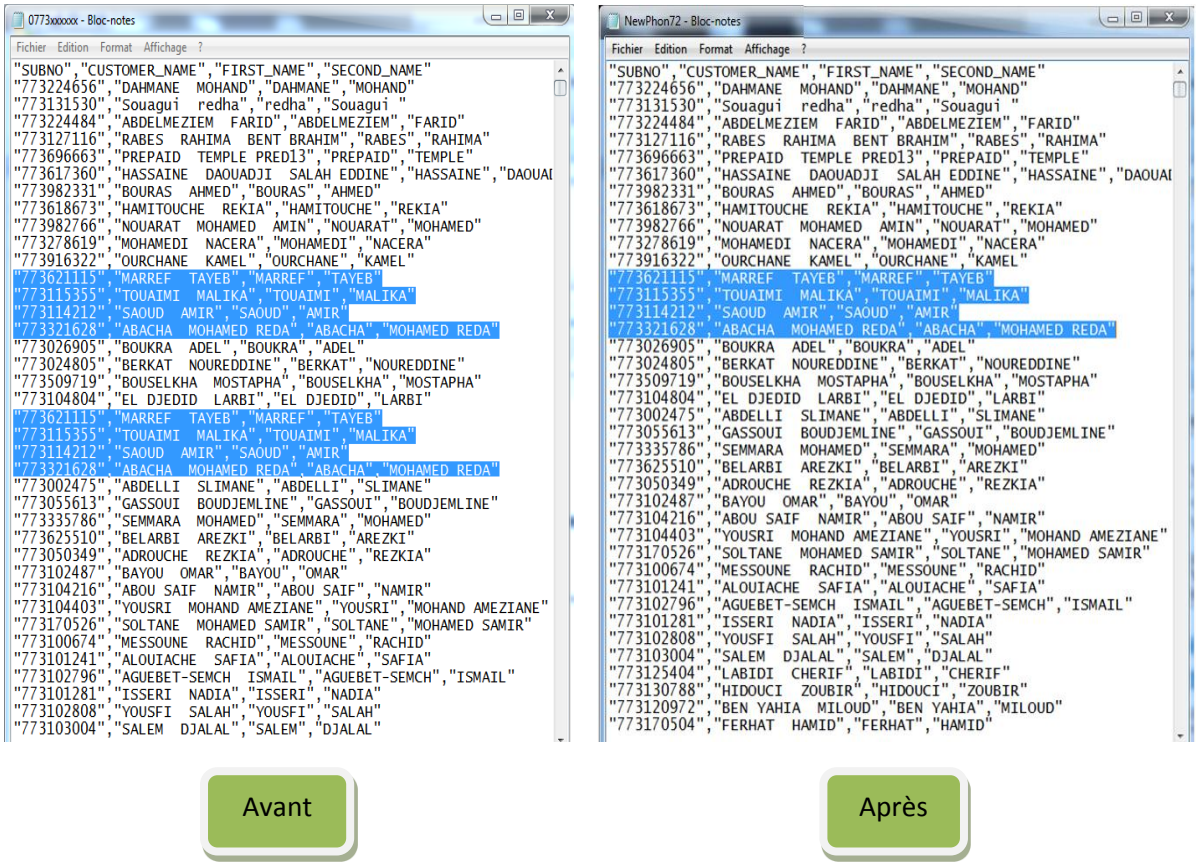


Figur(3-16): Pendant l'exécution



Figur(3-17): La fin de l'exécution de travail.

- Enfin, nous obtenons un nouveau fichier amélioré et sans éléments en double



Figur(3-18): Le résultat

Conclusion:

La qualité des données est une étape importante pour le développement des entreprises, mais on trouve que les données dupliquées sont le plus gros problème de la non-qualité. Voilà pourquoi notre travail a porté sur la recherche d'un moyen de résoudre ce problème.

Notre inclusion cet algorithme dans l'application Talend que et open source et gratuite, que permet d'améliorer la qualité des données. Et nous avons travaillé sur la recherche de la meilleure façon de supprimer les données en double et d'améliorer la qualité des données dans les grands entrepôts de données.

Cet algorithme améliorer les anciennes bases de données en forme nouvelle et moderne, et en plus de supprimer des données double.

CONCLUSION

Conclusion Générale :

Dans ce mémoire, nous avons essayé de résoudre le problème de la non-qualité des données dans les entrepôts de données. Pour ce faire, nous avons créé une application de déduplication des données en utilisant un algorithme de déduplication. Cet algorithme a été intégré dans la Platform de Talend Open Studio. Nous avons effectué des tests sur des données d'un entrepôt contenant des informations sur les abonnés d'un opérateur téléphonique.

Cependant, nous avons rencontré pendant notre travail sur la plate-forme de nombreuses difficultés par exemple la recherche de l'algorithme qui élimine les données similaires a été longue. Mais également l'opération d'intégrer cet algorithme dans la plate-forme a été très difficile faute de temps pour faire les tests sur des données.

Le travail effectué n'est pas complet car nous n'avons pas mis en œuvre un moyen pour comparer le temps d'exécution avec d'autres systèmes existants. Cette tâche reste à faire dans nos prochains travaux afin d'améliorer la qualité des données.

Bibliographie

Bibliographiques

[01]*TûyetTrâm DANG NGDOC. « Fédération de données semi-structurées avec XML ». Thèse de Doctorat en Informatique, de l'Université de Versailles Saint-Quentin-en-Yvelines. Juin 2003*

[02]*Xavier BARIL. « Un modèle de vues pour l'intégration de sources de données XML : VIMIX ». Thèse de Doctorat en Informatique, de l'Université des sciences et techniques du Languedoc. Décembre 2003*

[03]*Ahmed RAHNI. AMIDHA « Une approche médiatrice d'intégration de sources de données hétérogènes et autonomes ». Mémoire de stage effectué à l'ENSMA , Université de Poitiers. Juillet 2005*

[04]*Mohand-SaïdHacid, Chantal Reynaud. « L'intégration de sources de données » .LIRIS, UFR Informatique. Université Paris-Sud. novembre 1918*

[05]*L. Bellatreche, G. Pierra, D. Nguyen Xuan, D. Hondjack, « Intégration de sources de données autonomes par articulation a priori d'ontologies ». A paraître dans : Actes du XXIIème - Congrés INFORSID, Biarritz. Mai 2004*

[06]*Mme Amel BOUSSIS. « L'Intégration De Sources De Données A Base Ontologique Dans Un Environnement P2P ». Magister en Informatique. L'institut National d'informatique I.N.I.2007/2008*

----Fabrice Jouanot, « DILEMMA : vers une coopération de systèmes d'informations basée sur la médiation sémantique et la fusion d'objets », université de bourgougne, novembre 2001

[07]*L. Bellatreche, G. Pierra, D. Nguyen Xuan, D. Hondjack. « Intégration de sources de données autonomes par articulation a priori d'ontologies ». A paraître dans : Actes du XXIIème - Congrés INFORSID, Biarritz. Mai 2004*

[08]*Ilyes Boukhari. « Intégration et exploitation de besoins en entreprise étendue fondées sur la sémantique ». Thèse de Doctorat. ISAE-ENSMA Ecole Nationale Supérieure de Mécanique et d'Aérotechique - Poitiers, French. août 2006*

[09] *Christine Parent, Stefano Spaccapietr. « Intégration de bases de données: Panorama des problèmes et des approches ». Ecole Polytechnique Fédérale de Lausanne. Vol.4, N°3.1996*

[11]*Dan VODISLAV. « Cours IED. Architectures d'intégration de données ». Master Informatique M1.*

[12]*JEMM research. « DES DONNÉES QUALITÉ :Exploitez le capital de votre organisation ». livre blanc .janvier 2008*

[13] *Elie KABRE . « L'utilisation de l'information sanitaire comme un outil managérial des services hospitaliers: cas du Centre Hospitalier Régional de Lomé commune au Togo ». Diplôme d'études supérieures spécialisées en management .2006*

[14] *Christoher J., Murray L., Evans DB (Eds), Health systems performance assessment : Dabates, methods and empirism, Genève, Organisation Mondiale de la Santé (OMS), 2003*

[15] *Louardi BRADJI . « Adaptation des techniques de l'Extraction des Connaissances à partir des Données (ECD) pour prendre en charge la qualité des données » . Thèse de Doctorat en Informatique. Université Mentouri Constantine. Mars 2012*

[16] *C. Guerra-Garcia, I. Caballero, L. Berti-Equille, M. Piattini. « "DAQ_UWE: A Framework doeDesigning Data QualityAware Web Applications "« In Proceedings of the Conference on Information Quality (ICIQ), Adelaide, Australia, November 2011*

[17]*Franck Régnier-Pécastaing, Michel Gabassi, Jacques Finet. « Enjeux et méthodes de la gestion des données » , livre .2008*

[18] *Laure Berti-Équille .* « *Qualité des données ». Maître de Conférences*

[19]*R. Wang, V. Storey et C. Firth . "A framework for analysis of data quality research ; IEEE Transactions on Knowledge and Data Engineering"*

[20] *R.A. Moeller . Distributed data warehousing using Web technology. 2001*

[21] *Séraphin LOHAMBA OMATOKO . « Analyse et détection de l'attrition dans une entreprise de télécommunication ». Licencié en sciences informatique/Génie Logiciel ; Université Notre Dame du Kasayi 2011*

[22] *Ralph Kimball. Laura Reeves Margy Ross.* « *Concevoir et déployer un datawarehouse ». Warre nThornthwaite, 2001*

[23] *CHOUDER LAMRI . « Entrepôt Distribué de Données » . 2007*

[24] *Le bulletin technique TB00017 « Donnees De Reference Sur La Deduplication De Donnees »Livre Blanc. A Propos De Quantum. Juillet 2014*

Les sites web :

[site 1] Jean-Pierre BENOIT. https://www.linkedin.com/pulse/la-mod%C3%A9lisation-des-processus-peut-elle-contribuer-%C3%A0-de-benoit visité le 04 mai 2015

[site2] http://www.stibosystems.fr/french/solutions/glossaire/u/unicit%C3%A9-des-donn%C3%A9es.aspx visité le 04 mai 2015

[site 3] http://www.ontrack.fr/conformite-donnees visité le 05 mai 2015

[site 4] http://www.experian.fr/marketing-services/ressources/glossaire/ visité le 15 mai 2015

[site 5] http://www.trendmicro.fr/grandes-entreprises/conformite-aux-reglementations/ visité le 15 mai 2015

[site 6] http://fr.wikipedia.org/ visité le 10 mai 2015

[site7] http://www.ouestdecision.fr/contenu/les_editeurs/fiches/fiche-produit-talend.htm visité le 16 mai 2015

[site 7] http://www.zdnet.fr/actualites/deduplication-de-donnees-au-dela-des-economies-d-espace-disque-39389132.htm visité le 30 mai 2015

[site 8] http://france.emc.com/corporate/glossary/data-deduplication.htm visité le 30 mai 2015

[site 9] http://www.lemondeinformatique.fr/les-dossiers/sommaire-lire-deduplication-optimiser-le-stockage-pour-accelerer-la-sauvegarde-113.html visité le 30 mai 2015

[site 10] http://www.lemondeinformatique.fr/les-dossiers/lire-deduplication-optimiser-le-stockage-pour-accelerer-la-sauvegarde-520.html visité le 30 mai 2015